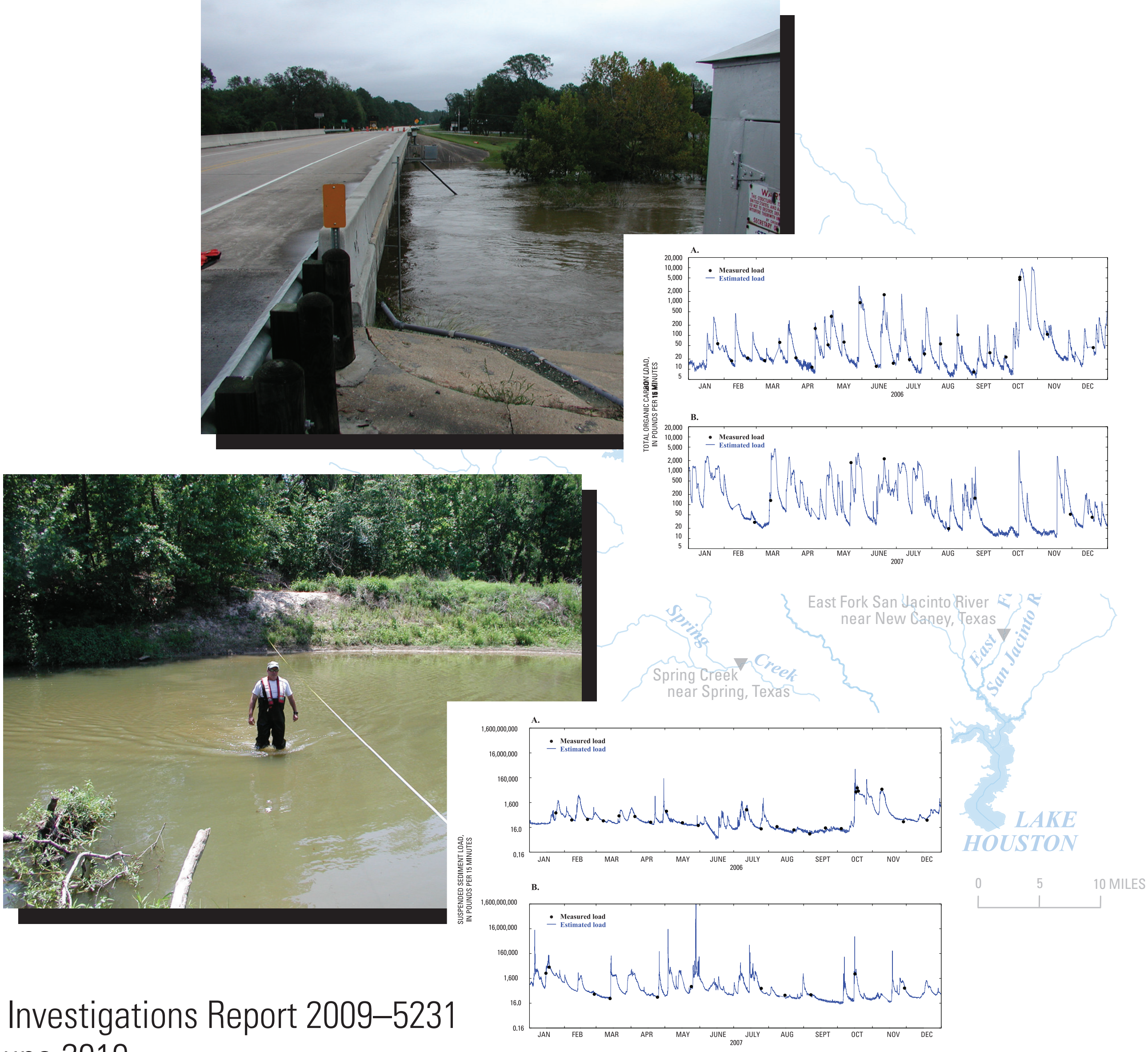
USGS
science for a changing world

In cooperation with the City of Houston

# Regression Models to Estimate Real-Time Concentrations of Selected Constituents in Two Tributaries to Lake Houston near Houston, Texas, 2005–07



Scientific Investigations Report 2009–5231
Revised June 2010

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

**Front cover:**

**Top,** East Fork San Jacinto River near New Caney, Texas (08070200) during high flow October 18, 2006.

**Bottom,** U.S. Geological Survey technician establishing a cross section at Spring Creek near Spring, Texas (08068500) during low-flow sampling.

# Regression Models to Estimate Real-Time Concentrations of Selected Constituents in Two Tributaries to Lake Houston near Houston, Texas, 2005–07

By Timothy D. Oden, William H. Asquith, and Matthew S. Milburn

In cooperation with the City of Houston

Scientific Investigations Report 2009–5231
Revised June 2010

**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: 2009
Revised 2010

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

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

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:
Oden, T.D., Asquith, W.H., and Milburn, M.S., 2009, Regression models to estimate real-time concentrations of selected constituents in two tributaries to Lake Houston near Houston, Texas, 2005–07: U.S. Geological Survey Scientific Investigations Report 2009–5231, 44 p. Revised June 2010.

# Contents

# Figures

# Tables

# Conversion Factors, Datums, Water-Quality Units, and Abbreviations

## Inch/Pound to SI

| Multiply | By | To obtain |
|---|---|---|
| | Length | |
| inch (in.) | 25.4 | millimeter (mm) |
| mile (mi) | 1.609 | kilometer (km) |
| | Area | |
| square mile ($mi^2$) | 2.590 | square kilometer ($km^2$) |
| | Flow rate | |
| cubic foot per second ($ft^3/s$) | 0.02832 | cubic meter per second ($m^3/s$) |
| | Mass | |
| pound, avoirdupois (lb) | 0.4536 | kilogram (kg) |
| | Volume | |
| ounce, fluid (fl. oz.) | 29.574 | milliliter (mL) |

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

$$°F=(1.8×°C)+32$$

## Datums

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.

## Water-Quality Units

Specific conductance is given in microsiemens per centimeter at 25 degrees Celsius (µS/cm at 25 °C).

Concentrations of chemical constituents in water are given in either milligrams per liter (mg/L) or micrograms per liter (µg/L).

Bacteria is given in most probable number per 100 milliliters (MPN/100 mL).

Turbidity is given in Formazine Nephelometric Units (FNU).

## Abbreviations (figures 3–13)

DF, degrees of freedom

lm, L-moment

Pr, probability

Signif., significance

sqrt, square root

Std., standard

# Regression Models to Estimate Real-Time Concentrations of Selected Constituents in Two Tributaries to Lake Houston near Houston, Texas, 2005–07

By Timothy D. Oden, William H. Asquith, and Matthew S. Milburn

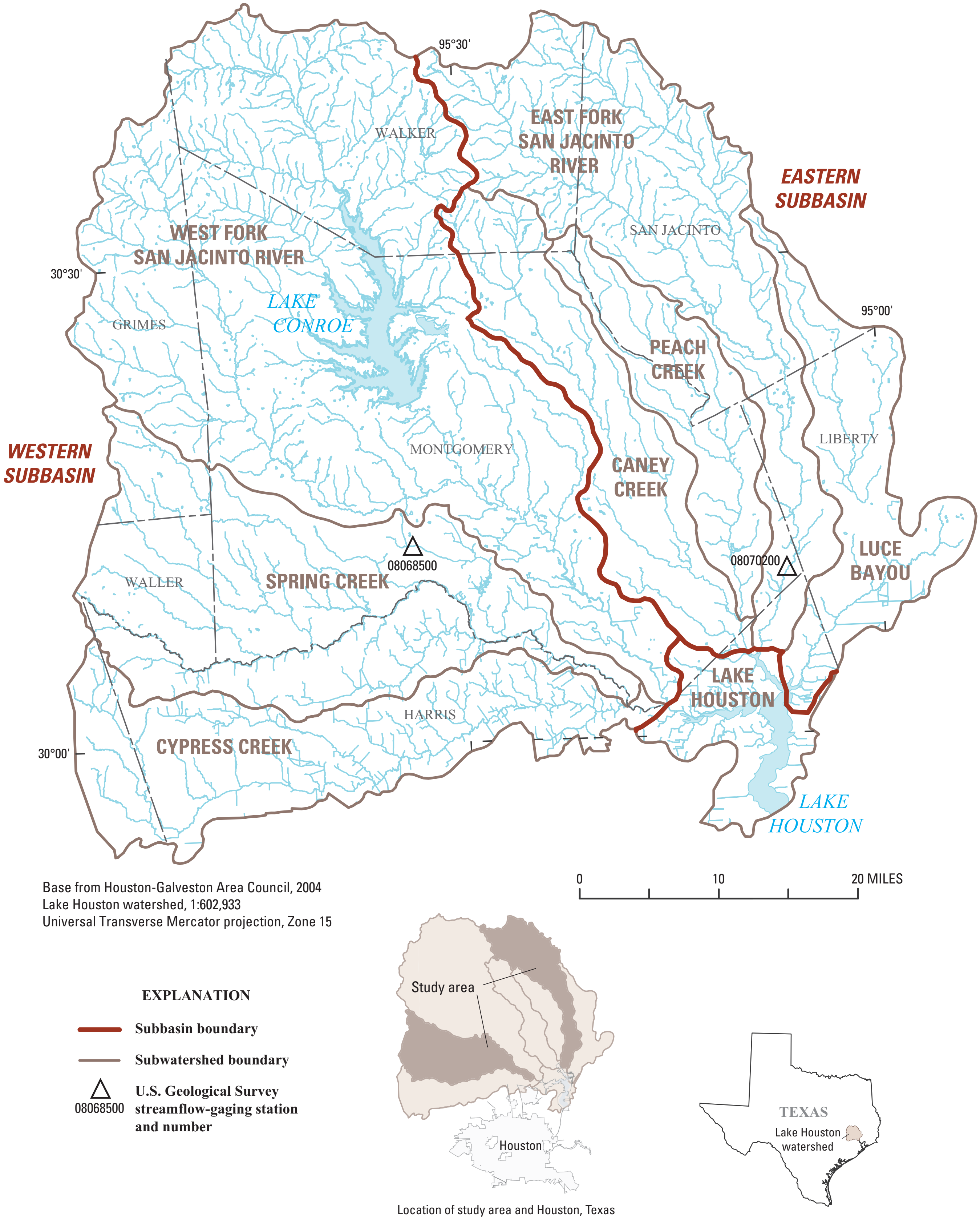
## Abstract

In December 2005, the U.S. Geological Survey in cooperation with the City of Houston, Texas, began collecting discrete water-quality samples for nutrients, total organic carbon, bacteria (total coliform and *Escherichia coli*), atrazine, and suspended sediment at two U.S. Geological Survey streamflow-gaging stations upstream from Lake Houston near Houston (08068500 Spring Creek near Spring, Texas, and 08070200 East Fork San Jacinto River near New Caney, Texas). The data from the discrete water-quality samples collected during 2005–07, in conjunction with monitored real-time data already being collected—physical properties (specific conductance, pH, water temperature, turbidity, and dissolved oxygen), streamflow, and rainfall—were used to develop regression models for predicting water-quality constituent concentrations for inflows to Lake Houston. Rainfall data were obtained from a rain gage monitored by Harris County Homeland Security and Emergency Management and colocated with the Spring Creek station. The leaps and bounds algorithm was used to find the best subsets of possible regression models (minimum residual sum of squares for a given number of variables). The potential explanatory or predictive variables included discharge (streamflow), specific conductance, pH, water temperature, turbidity, dissolved oxygen, rainfall, and time (to account for seasonal variations inherent in some water-quality data). The response variables at each site were nitrite plus nitrate nitrogen, total phosphorus, organic carbon, *Escherichia coli*, atrazine, and suspended sediment. The explanatory variables provide easily measured quantities as a means to estimate concentrations of the various constituents under investigation, with accompanying estimates of measurement uncertainty. Each regression equation can be used to estimate concentrations of a given constituent in real time. In conjunction with estimated concentrations, constituent loads were estimated by multiplying the estimated concentration by the corresponding streamflow and applying the appropriate conversion factor. By computing loads from estimated constituent concentrations, a continuous record of estimated loads can be available for comparison to total maximum daily loads. The regression equations presented in this report are site specific to the Spring Creek and East Fork San Jacinto River streamflow-gaging stations; however, the methods that were developed and documented could be applied to other tributaries to Lake Houston for estimating real-time water-quality data for streams entering Lake Houston.

## Introduction

Houston, Texas (fig. 1), is the fourth largest city in the Nation, with an estimated population of about 5.4 million people in 2006 (Texas State Data Center, 2007). Historically, groundwater has been the major source of supply for the City of Houston. However, development of groundwater resources has contributed to water-level declines and land-surface subsidence (Kasmarek and Strom, 2002; Kasmarek and Houston, 2008). In 2008, Lake Houston supplied about 20 percent of the total source-water supply for the City of Houston (Dannelle Belhateche, City of Houston Public Works and Engineering Department, oral commun., 2008). However, as a result of regulations to limit groundwater withdrawals to arrest land-surface subsidence, Lake Houston is expected to become the primary source of water for the city in the future; the overall goal is to increase the use of surface water to no less than 80 percent of the total demand by 2030 (Harris-Galveston Subsidence District, 1999). Because Lake Houston is a major source of potable water and also a recreation resource for the Houston area, the possible effects of urbanization on the water quality of tributaries to Lake Houston are of interest to water managers. Two of the seven tributaries to Lake Houston, Spring Creek and East Fork San Jacinto River (fig. 1), are the focus of this report.

In compliance with the Federal Clean Water Act, the Texas Commission on Environmental Quality compiles an inventory of water bodies that are either impaired (do not meet applicable water-quality standards) or threatened (are not expected to meet standards in the future) (Texas Commission on Environmental Quality, 2008). Lake Houston (segment 1002) first appeared in 2006 and again in 2008 on the State of Texas list of impaired or threatened water bodies (known as the 303[d] list) for bacteria. All of Spring Creek (segment 1008) has been listed for bacteria since 1996, and one segment of Spring Creek (1008_02) has been listed for



Base from Houston-Galveston Area Council, 2004
Lake Houston watershed, 1:602,933
Universal Transverse Mercator projection, Zone 15

**Figure 1.** Lake Houston watershed and tributary subwatersheds and location of U.S. Geological Survey streamflow-gaging stations 08068500 Spring Creek near Spring, Texas, and 08070200 East Fork San Jacinto River near New Caney, Texas.

depressed dissolved oxygen concentrations (not conducive to healthy ecosystems) since 1996. The East Fork San Jacinto River (segment 1003) also first appeared on the 303(d) list in 2006 for bacteria and is still listed at present (2009).

This report was done as a part of the Lake Houston Project, a cooperative project between the U.S. Geological Survey (USGS) and the City of Houston to monitor water quality in Lake Houston and its watershed. Watershed water-quality monitoring began in December 2005 and is ongoing in 2009; ongoing in-lake water-quality monitoring began in April 2006. Continuous, real-time monitoring of streamflow and water-quality properties (specific conductance, pH, water temperature, turbidity, and dissolved oxygen) in Spring Creek and East Fork San Jacinto River is done to alert drinking-water managers to potential changes in quality of water entering Lake Houston. The continuously monitored streamflow and water-quality properties, in conjunction with regression models using those data as surrogates for selected constituents (nitrite plus nitrate nitrogen, total phosphorus, total organic carbon, *Escherichia coli* [*E. coli*], atrazine, and suspended sediment) can be used to estimate concentrations for constituents that are lacking a continuous record; then the estimated concentrations can be used to compute estimated constituent loads. With near real-time knowledge of water quality of the tributaries, water managers will be able to identify potential effects of tributary inflows on the water quality of Lake Houston before they happen and to adjust drinking-water plant operations accordingly. In addition, over time the results of tributary water-quality monitoring will contribute to the understanding of watershed influences on Lake Houston and the effects of those influences on Lake Houston as a drinking-water and recreational resource.

## Purpose and Scope

The purpose of this report is to document regression models developed to estimate real-time concentrations of nitrite plus nitrate, total phosphorus, total organic carbon, *E. coli*, atrazine, and suspended sediment in two tributaries to Lake Houston, Spring Creek and East Fork San Jacinto River. The regression models were developed using real-time, continuously measured streamflow and water-quality properties (specific conductance, pH, water temperature, turbidity, and dissolved oxygen) as well as real-time and discrete water-quality samples analyzed for nitrite plus nitrate, total phosphorus, total organic carbon, *E. coli*, atrazine, and suspended sediment. Rainfall data and time were considered as additional explanatory variables in the regression models. The process used to develop and evaluate numerous possible regression models to obtain a best-fit regression model for each water-quality constituent (using significant water-quality properties as explanatory variables) for each site is explained. The data were collected at two USGS streamflow-gaging stations, 08068500 Spring Creek near Spring, Tex. (hereinafter, Spring Creek site), and 08070200 East Fork San Jacinto River near New Caney, Tex. (hereinafter, East Fork San Jacinto site). Although atrazine samples were collected at each site, a sufficient number of uncensored atrazine concentrations to construct a regression model were available only at the Spring Creek site. The best-fit regression models for each constituent are presented for each station. Lastly, estimated constituent loads for 2006 and 2007 computed from concentrations estimated using the best-fit regression models are presented and compared to loads computed from concentrations measured in discrete water-quality samples.

## Description of Study Area

Lake Houston is about 25 miles northeast of Houston, Tex. The watershed of Lake Houston comprises the subwatersheds of seven tributaries and the area immediately adjacent to the lake in parts of seven counties (fig. 1), including large areas of densely populated Harris and Montgomery Counties. Sneck-Fahrer and others (2005) divided the Lake Houston watershed into eastern and western subbasins, primarily on the basis of relative amounts of development, with the eastern subbasin being the less developed. The western subbasin encompasses three tributary subwatersheds and the eastern subbasin encompasses four tributary subwatersheds (table 1). The study area of this report comprises one subwatershed from each subbasin, Spring Creek in the western subbasin and East Fork San Jacinto River in the eastern subbasin.

The Spring Creek subwatershed in the western subbasin is the second most densely populated of the seven Lake Houston subwatersheds, with a population density in 2000 of about 390 people per square mile (U.S. Census Bureau, 2000). Urban and agricultural land together account for 41 percent of the 453 square miles of the subwatershed (Multi-Resolution Land Characteristics Consortium, 2003), more than

**Table 1.** Subwatershed drainage areas for tributaries to Lake Houston, near Houston, Texas (modified from Sneck-Fahrer and others, 2005).

| Subwatershed | Drainage area (square miles) |
|---|---|
| Western subbasin | |
| West Fork San Jacinto River | 998 |
| Spring Creek[1] | 453 |
| Cypress Creek | 305 |
| Eastern subbasin | |
| East Fork San Jacinto River[1] | 404 |
| Peach Creek | 151 |
| Caney Creek | 222 |
| Luce Bayou | 210 |

[1] Subwatershed for which regression analysis was used to develop predictive equations in this report.

twice the percentage of developed and cultivated land in the East Fork San Jacinto River subwatershed. The predominant land-use classification in the Spring Creek subwatershed is forest (31 percent). Wetland and rangeland account for 10 and 16 percent, respectively, of the Spring Creek subwatershed.

The East Fork San Jacinto River subwatershed in the eastern subbasin is the least densely populated of the seven subwatersheds that drain to Lake Houston, with a population density in 2000 of about 80 people per square mile (U.S. Census Bureau, 2000). Urban and agricultural land together account for 18 percent of the 404 square miles of the subwatershed (Multi-Resolution Land Characteristics Consortium, 2003). As in the Spring Creek subwatershed, the predominant land-use classification in the subwatershed is forest (47 percent), followed by wetland (19 percent) and rangeland (15 percent).

The climate in the study area is classified as humid subtropical (Texas State Climatologist, 2008), characterized by cool, temperate winters and long, hot summers and by high humidity. During 2005–07, annual rainfall ranged from 41.2 to 65.5 inches at Intercontinental Airport, Houston, Tex. (National Oceanic and Atmospheric Administration, 2008). Rainfall in 2005 was about 6.6 inches below normal (1971–2000 average of 47.8 inches) while 2006 and 2007 were wetter than normal; 2007 was the sixth wettest year on record for the Houston area (65.5 inches).

## Previous Investigations

Previous water-quality information for the study area is summarized in Sneck-Fahrer and others (2005). Sneck-Fahrer and others (2005) assessed relative contributions to the water quality of Lake Houston from the more-developed western subbasins and the less-developed eastern subbasins using analyses of water samples from Cypress Creek (western subbasin) and East Fork San Jacinto River (eastern subbasin). Constituent yields allowed direct comparison of loads from Cypress Creek and East Fork San Jacinto River. In Cypress Creek, storm yields of nitrite plus nitrate nitrogen for high flows ranged from 8 to 45 pounds per square mile per day. In East Fork San Jacinto River, the maximum storm yield of nitrite plus nitrate nitrogen for high flows was 1.47 pounds per square mile per day. At low flows, the median daily yield of dissolved phosphorus from Cypress Creek was 84 times larger than the median daily yield from East Fork San Jacinto River; at high flows, it was 16 times larger. At high flows, the maximum daily yield of atrazine from Cypress Creek was 460 times larger than the maximum daily yield from East Fork San Jacinto River.

Christensen and others (2000) developed regression equations to estimate constituent concentrations and loads for the Little Arkansas River in south-central Kansas. Ryberg (2006) also used continuous water-quality monitoring and regression analyses to estimate constituent loads in the Red River of the North in Fargo, N. Dak. The work by Christensen and others (2000) and Ryberg (2006) served as a guide for developing similar methods to suit the hydrologic setting of the study area for this report.

# Methods

This section describes how the data for this study were collected and analyzed and explains the methodology used to develop the regression equations. The R environment for statistical computing (R Development Core Team, 2006) was used to implement all multiple regression methods and associated diagnostic tests of multiple regression results described in this report.

## Streamflow Measurements

Streamflow is the volume of water passing an established reference point in a stream at a given time. Methods used to determine streamflow (discharge) are described in Buchanan and Somers (1969). Streamflow measurements during the course of the study were made about five times per year at the Spring Creek site and about five times per year at the East Fork San Jacinto site. Stage, or gage height, was measured every 15 minutes using submersible pressure transducers to the nearest 0.01 foot at the Spring Creek and East Fork San Jacinto sites. The data were electronically recorded and transmitted by satellite to a downlink site and then to the USGS Texas Water Science Center in Austin, Tex. A stage-discharge relation was developed on the basis of streamflow measurements and the stage of the stream at the time of measurement (Kennedy, 1984). These unique relations were used to compute a continuous record of streamflow (Kennedy, 1983) from the stage record at each site. Instantaneous stage and streamflow values are stored in the USGS National Water Information System (NWIS) database (U.S. Geological Survey, 2009).

## Continuous Water-Quality Monitoring

Continuous monitoring of four physical properties (specific conductance, pH, water temperature, and dissolved oxygen) began at the Spring Creek site in November 1999 using a YSI Environmental 600XL Sonde. In November 2005, a YSI Environmental 6600 Sonde was installed at the Spring Creek site to include turbidity. Continuous monitoring of specific conductance, pH, water temperature, turbidity, and dissolved oxygen began at the East Fork San Jacinto site in November 2005 using a YSI Environmental 6600 Sonde (monitor). Each of the five sensors on the sondes was calibrated as described in “National Field Manual for the Collection of Water-Quality Data” (U.S. Geological Survey, variously dated); the continuous monitor and record were maintained as outlined in Wagner and others (2006).

The Spring Creek and East Fork San Jacinto sites use a swinging well design to monitor real-time water-quality properties. Swinging wells are constructed of schedule 80 polyvinyl chloride pipe with holes in the lower 3 feet, allowing water to pass through wherever the sonde is located. Each monitor is located near the centroid of flow in each stream in a swinging well. The data from each sonde were electronically recorded and transmitted by satellite to a downlink site and then to the USGS Texas Water Science Center in Austin. Specific conductance, pH, water temperature, turbidity, and dissolved oxygen data are stored in the USGS NWIS database in 15-minute intervals. The Spring Creek and East Fork San Jacinto sites are still monitoring real-time water-quality properties at the present (2009).

## Discrete Water-Quality Sample Collection, Analysis, and Results

Discrete water-quality samples were manually collected at each sampling site. Thirty-nine samples were collected at the Spring Creek site and 38 samples were collected at the East Fork San Jacinto site. Samples were analyzed for nutrients, total organic carbon, bacteria, atrazine, and suspended sediment.

### Sample Design and Collection

Hydrologic conditions in the Spring Creek and East Fork San Jacinto River watersheds vary and might affect chemical constituent concentrations, so discrete water-quality samples were collected over a wide range of streamflow conditions (fig. 2). Discrete water-quality samples for the first year (December 2005–November 2006) of this study were collected about every 2 weeks to observe seasonal patterns in water quality. Samples at these fixed-frequency sample times were collected as scheduled without regard to hydrologic condition, such as rising, falling, or stable streamflows. During storms or periods of high flow, unscheduled samples were also periodically collected during the first year of the study. Discrete water-quality samples for the second year (December 2006–December 2007) of the study were collected once a month. As in the first year of the study, stormwater-runoff samples for the second year were collected whenever possible.

Discrete water-quality samples were collected either by wading, when accessible, or from bridges during higher flows. All samples were collected and processed as outlined in the USGS “National Field Manual for the Collection of Water-Quality Data” (U.S. Geological Survey, variously dated). Depth-integrated samples were collected, using a Teflon bottle and nozzle, either by multiple verticals when stream velocities were less than about 1.5 feet per second or by the flow-weighted, equal-width increment method when stream velocities were greater than about 1.5 feet per second. Samples from each vertical were combined in a Teflon churn, dispensed into appropriate sample containers, and shipped at 4 degrees Celsius (°C) by overnight courier to appropriate laboratories. Samples for bacteria analysis were collected directly from the centroid of flow in sterile, autoclaved bottles.

### Sample Analysis

Samples collected and analyzed for nutrients and total organic carbon were analyzed by the USGS National Water Quality Laboratory, Denver, Colo., using published methods. Methods for nutrient analysis are documented in Fishman (1993), U.S. Environmental Protection Agency (1993; method 365.1), and Patton and Truitt (2000). Total organic carbon analysis is documented in Wershaw and others (1987). Suspended-sediment samples were analyzed by the USGS Sediment Laboratory, Baton Rouge, La., using procedures described in Guy (1969) and Mathes and others (1992). Atrazine samples were analyzed by the USGS Organic Geochemistry Research Laboratory, Lawrence, Kans., using the Enzyme-Linked Immunosorbent Assay (ELISA) method documented in Aga and Thurman (1997). *E. coli* and total coliform bacteria were analyzed at the Houston office of the USGS Texas Water Science Center, using the defined substrate method documented in American Public Health Association and others (2005) and were reported as most probable number per 100 milliliters (MPN/100 mL) with confidence intervals.

Summary statistics of the discrete water-quality samples are summarized in table 2. The data for the Spring Creek and East Fork San Jacinto sites are stored in the USGS NWIS database and can be accessed online at *http://nwis.waterdata.usgs.gov/tx/nwis/qwdata?site_no=08068500* and *http://nwis.waterdata.usgs.gov/tx/nwis/qwdata?site_no=08070200*, respectively.

### Quality Control

Quality-control (QC) samples were collected as described in “National Field Manual for the Collection of Water-Quality Data” (U.S. Geological Survey, variously dated) and analyzed by the same laboratories and methods as the environmental samples. QC samples include equipment blanks (two), field blanks (five), and split replicate samples (nutrients and total organic carbon [six], bacteria [17], atrazine and suspended sediment [seven]). QC samples were collected to evaluate any contamination, as well as bias and variability of the water chemistry data, that might have resulted from sample collection, processing, transportation, and laboratory analysis. QC results are listed in table 3.

Equipment blanks were collected annually in a controlled environment to determine if the cleaning procedures for sample containers and the equipment for sample collection and sample processing were sufficient to produce contaminant-free samples. Orthophosphate, detected in the equipment blank in 2006 at less than the laboratory reporting level (LRL), was reported as estimated (Childress and others, 1999). The USGS uses two reporting conventions for the analytical data from the National Water Quality Laboratory, the

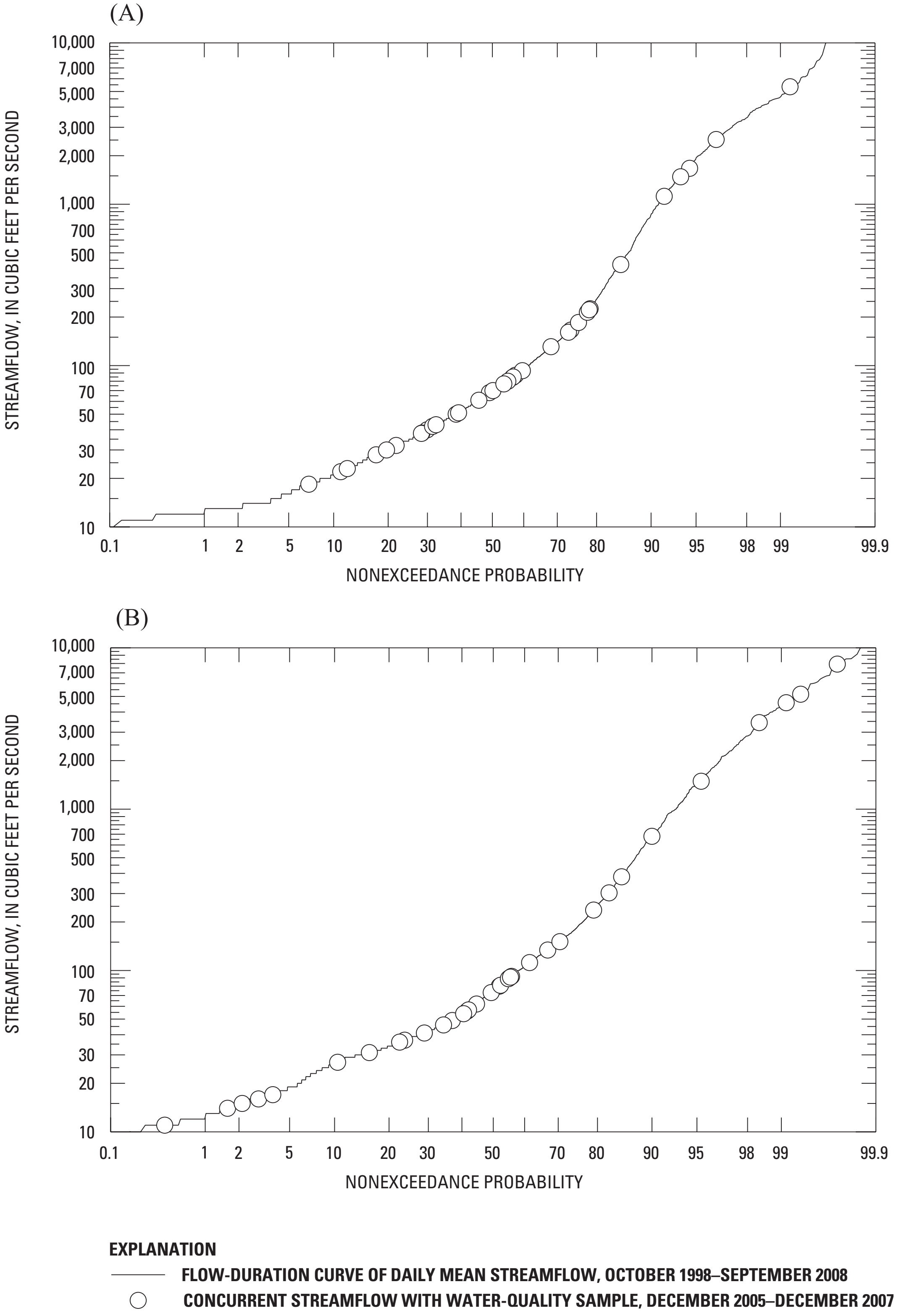


**Figure 2.** Flow duration curve and corresponding discrete water-quality samples, (A) Spring Creek near Spring, Texas, and (B) East Fork San Jacinto River near New Caney, Texas.

**Table 2.** Summary statistics for samples collected at two tributaries to Lake Houston near Houston, Texas, 2005–07.

[n, number of samples; <, less than[1]; E, estimated[1]]

| U.S. Geological Survey station name | Station number | Summary statistic | Ammonia plus organic nitrogen, water, filtered (milligrams per liter as nitrogen) | Ammonia plus organic nitrogen, water, unfiltered (milligrams per liter as nitrogen) | Ammonia, water, filtered (milligrams per liter as nitrogen) | Nitrite plus nitrate, water, filtered (milligrams per liter as nitrogen) | Nitrite, water, filtered (milligrams per liter as nitrogen) | Orthophosphate, water, filtered (milligrams per liter as phosphorus) | Phosphorus, water, filtered (milligrams per liter) |
|---|---|---|---|---|---|---|---|---|---|
| Spring Creek near Spring, Tex. (n=39) | 08068500 | Minimum | 0.48 | 0.64 | <0.01 | 0.35 | 0.013 | 0.128 | 0.145 |
| | | Maximum | 1.25 | 2.28 | .41 | 7.39 | .12 | 1.62 | 1.64 |
| | | Median | .82 | 1.26 | .06 | 2.02 | .038 | .634 | .663 |
| East Fork San Jacinto River near New Caney, Tex. (n=38) | 08070200 | Minimum | .14 | .23 | <.01 | <.06 | <.002 | <.02 | .013 |
| | | Maximum | .74 | .98 | .05 | .51 | .03 | .059 | .072 |
| | | Median | .37 | .44 | E.02 | .1 | .003 | .02 | .034 |

| U.S. Geological Survey station name | Station number | Summary statistic | Phosphorus, water, unfiltered (milligrams per liter) | Organic carbon, water, unfiltered (milligrams per liter) | *Escherichia coli*, Colilert Quantitray method, water (most-probable number per 100 milliliters) | Total coliform, Colilert Quantitray method, water (most probable number per 100 milliliters) | Atrazine, water, filtered, recoverable, immunoassay, unadjusted (micrograms per liter) | Suspended sediment (milligrams per liter) |
|---|---|---|---|---|---|---|---|---|
| Spring Creek near Spring, Tex. (n=39) | 08068500 | Minimum | 0.263 | 7.62 | 36 | 2,420 | <0.1 | 15 |
| | | Maximum | 1.81 | 21 | 41,000 | 1,300,000 | 14 | 987 |
| | | Median | .864 | 11.5 | 360 | 35,100 | .95 | 44.5 |
| East Fork San Jacinto River near New Caney, Tex. (n=38) | 08070200 | Minimum | .054 | 3.96 | 23 | 1,000 | <.1 | 6 |
| | | Maximum | .21 | 32.6 | 36,100 | 242,000 | .32 | 170 |
| | | Median | .098 | 7.74 | 94 | 8,200 | <.01 | 18 |

[1] Concentrations measured as less than the long-term method detection level (LT-MDL) are reported as less than the laboratory reporting level (LRL). Concentrations measured between the LT-MDL and LRL are reported but given an "E" remark code to indicate that they are semiquantitative (Mueller and Spahr, 2005).

**Table 3.** Results of quality-control samples collected at two tributaries to Lake Houston near Houston, Texas, 2005–07.

[Environ., Environmental; --, not analyzed; <, less than laboratory reporting level; E, estimated; *, value reviewed and rejected; >, greater than]

| U.S. Geological Survey station name | Station number | Sample date | Sample time | Sample type | Ammonia plus organic nitrogen, water, filtered (milligrams per liter as nitrogen) | Ammonia plus organic nitrogen, water, unfiltered (milligrams per liter as nitrogen) | Ammonia, water, filtered (milligrams per liter as nitrogen) | Nitrite plus nitrate, water, filtered (milligrams per liter as nitrogen) | Nitrite, water, filtered (milligrams per liter as nitrogen) | Orthophosphate, water, filtered (milligrams per liter as phosphorus) |
|---|---|---|---|---|---|---|---|---|---|---|
| Spring Creek near Spring, Tex. | 08068500 | 12/1/2005 | 1330 | Environ. | 0.85 | 1.3 | 0.08 | 6.87 | 0.028 | 1.56 |
| | | *12/1/2005* | *1331* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | 2/7/2006 | 1400 | Environ. | .75 | 1.1 | .06 | 4.52 | .02 | .95 |
| | | *2/7/2006* | *1401* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | 2/21/2006 | 1400 | Environ. | 1.1 | 1.4 | .39 | 4.01 | .054 | .91 |
| | | *2/21/2006* | *1401* | *Replicate* | *1.2* | *1.4* | *.38* | *4.04* | *.054* | *.91* |
| | | 5/16/2006 | 1115 | Environ. | .78 | 1.8 | <.04 | 1.33 | .069 | .63 |
| | | *5/16/2006* | *1116* | *Replicate* | *.78* | *1.9* | *<.04* | *1.32* | *.07* | *.64* |
| | | 6/20/2006 | 1000 | Environ. | .86 | 1.5 | .035 | .35 | .034 | .197 |
| | | *6/20/2006* | *1001* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | *6/28/2006* | *845* | *Blank* | *.12* | *<.10* | *.015* | *<.06* | *<.002* | *<.006* |
| | | 6/28/2006 | 930 | Environ. | 1 | 1.2 | .088 | 4.66 | .028 | 1.01 |
| | | 7/12/2006 | 1300 | Environ. | .83 | 1.1 | .058 | 3.36 | .057 | .719 |
| | | *7/12/2006* | *1346* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | *8/23/2006* | *1320* | *Blank* | *E.07* | *<.10* | *<.010* | *<.06* | *<.002* | *<.006* |
| | | 8/23/2006 | 1345 | Environ. | .77 | 1.8 | <.010 | 2.12 | .021 | .738 |
| | | 9/20/2006 | 1230 | Environ. | 1.1 | 1.5 | .166 | 2.01 | .101 | .891 |
| | | *9/20/2006* | *1231* | *Replicate* | *1.1* | *1.5* | *.17* | *2.02* | *.101* | *.901* |
| | | 8/15/2007 | 1420 | Environ. | .64 | .73 | .055 | 3.56 | .053 | .735 |
| | | *8/15/2007* | *1421* | *Replicate* | -- | -- | -- | -- | -- | -- |
| East Fork San Jacinto River near New Caney, Tex. | 08070200 | 12/1/2005 | 1020 | Environ. | .14 | .23 | <.04 | .1 | <.008 | .03 |
| | | *12/1/2005* | *1021* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | *12/21/2005* | *930* | *Blank* | *E.06* | *<.10* | *<.04* | *<.06* | *<.008* | *<.02* |
| | | 12/21/2005 | 1030 | Environ. | .45 | .38 | <.04 | .07 | <.008 | <.02 |
| | | 3/7/2006 | 1130 | Environ. | .46 | .44 | <.04 | .07 | <.008 | E.01 |
| | | *3/7/2006* | *1131* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | 4/4/2006 | 1100 | Environ. | .74 | .8 | .05 | .17 | E.006 | .02 |
| | | *4/4/2006* | *1101* | *Replicate* | *.65* | *.8* | *.05* | *.17* | *E.007* | *.02* |
| | | 7/25/2006 | 945 | Environ. | .35 | .47 | .027 | .51 | .007 | .059 |
| | | *7/25/2006* | *946* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | 8/8/2006 | 1000 | Environ. | .38 | .39 | .023 | .22 | .002 | .034 |
| | | *8/8/2006* | *1001* | *Replicate* | *.42* | *.39* | *.022* | *.23* | *.003* | *.032* |
| | | 9/6/2006 | 1100 | Environ. | .36 | .27 | .025 | .08 | E.001 | .024 |
| | | *9/6/2006* | *1130* | *Blank* | *.14* | *<.10* | *E.009* | *<.06* | *<.002* | *<.006* |
| | | 10/4/2006 | 945 | Environ. | .22 | .3 | E.016 | .15 | <.002 | .047 |
| | | *10/4/2006* | *946* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | 10/18/2006 | 1330 | Environ. | .65 | .98 | <.020 | E.05 | .003 | .013 |
| | | *10/18/2006* | *1331* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | 1/15/2007 | 1340 | Environ. | .5 | .76 | .038 | .09 | .003 | .021 |
| | | *1/15/2007* | *1341* | *Replicate* | *.52* | *.78* | *.035* | *.09* | *.003* | *.019* |
| | | 5/24/2007 | 1038 | Environ. | -- | -- | -- | -- | -- | -- |
| | | *5/24/2007* | *1039* | *Replicate* | -- | -- | -- | -- | -- | -- |
| | | *8/15/2007* | *1059* | *Blank* | *.35* | *<.10* | *.031* | *<.06* | *E.001* | *<.006* |
| | | 8/15/2007 | 1132 | Environ. | * | .33 | <.020 | .07 | .004 | .015 |
| *Equipment blank* | *08070200* | *8/22/2006* | *1330* | *Blank* | *<.10* | *<.10* | *<.010* | *<.06* | *<.002* | *E.003* |
| *Equipment blank* | *301056095265000* | *11/28/2007* | *1358* | *Blank* | *<.14* | *<.14* | *<.020* | *<.04* | *<.002* | *<.006* |
| *Equipment blank* | *301056095265000* | *11/28/2007* | *1359* | *Blank* | -- | -- | -- | -- | -- | -- |

**Table 3.** Results of quality-control samples collected at two tributaries to Lake Houston near Houston, Texas, 2005–07—Continued.

| U.S. Geological Survey station name | Station number | Sample date | Sample time | Sample type | Phosphorus, water, filtered (milligrams per liter) | Phosphorus, water, unfiltered (milligrams per liter) | Organic carbon, water, unfiltered (milligrams per liter) | *Escherichia coli*, Colilert Quantitray method, water (most probable number per 100 milliliters) | Total coliform, Colilert Quantitray method, water (most probable number per 100 milliliters) | Atrazine, water, filtered, recoverable, immunoassay, unadjusted (micrograms per liter) | Suspended sediment (milligrams per liter) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Spring Creek near Spring, Tex. | 08068500 | 12/1/2005 | 1330 | Environ. | 1.61 | 1.73 | 9.6 | 130 | >2,400 | 0.3 | 36 |
| | | 12/1/2005 | 1331 | Replicate | -- | -- | -- | 90 | >2,400 | -- | -- |
| | | 2/7/2006 | 1400 | Environ. | .97 | 1.11 | 9.8 | 60 | 3,400 | .95 | 33 |
| | | 2/7/2006 | 1401 | Replicate | -- | -- | -- | 55 | 3,300 | -- | -- |
| | | 2/21/2006 | 1400 | Environ. | .87 | 1.05 | 9.6 | 37 | 6,100 | 1.51 | 24 |
| | | 2/21/2006 | 1401 | Replicate | .86 | 1.02 | 9.3 | 27 | 6,500 | 1.34 | 23 |
| | | 5/16/2006 | 1115 | Environ. | .66 | .87 | 13.4 | 490 | 110,000 | 2.56 | 60 |
| | | 5/16/2006 | 1116 | Replicate | .68 | .85 | 15.3 | 690 | 98,000 | 2.68 | 61 |
| | | 6/20/2006 | 1000 | Environ. | .24 | .44 | 17 | 8,000 | 410,000 | 1.45 | 547 |
| | | 6/20/2006 | 1001 | Replicate | -- | -- | -- | 7,100 | 460,000 | -- | -- |
| | | 6/28/2006 | 845 | Blank | <.02 | <.02 | <.4 | -- | -- | <.10 | <1 |
| | | 6/28/2006 | 930 | Environ. | 1 | 1.25 | 8.9 | 43 | >2,400 | .62 | 31 |
| | | 7/12/2006 | 1300 | Environ. | .7 | .89 | 9 | 120 | 41,000 | .68 | 38 |
| | | 7/12/2006 | 1346 | Replicate | -- | -- | -- | 190 | 69,000 | -- | -- |
| | | 8/23/2006 | 1320 | Blank | <.02 | <.02 | <.4 | -- | -- | <.10 | <1 |
| | | 8/23/2006 | 1345 | Environ. | .78 | 1 | 14.1 | 6,800 | 440,000 | 1.75 | 147 |
| | | 9/20/2006 | 1230 | Environ. | .94 | 1.1 | 13.1 | 270 | 82,000 | 1.73 | 56 |
| | | 9/20/2006 | 1231 | Replicate | .95 | 1.09 | 10.3 | 360 | 49,000 | 1.77 | 46 |
| | | 8/15/2007 | 1420 | Environ. | .75 | .92 | 8.1 | 310 | 17,000 | .18 | 42 |
| | | 8/15/2007 | 1421 | Replicate | -- | -- | -- | 190 | 17,000 | -- | -- |
| East Fork San Jacinto River near New Caney, Tex. | 08070200 | 12/1/2005 | 1020 | Environ. | .039 | .086 | 4 | 43 | 1,700 | <.10 | 11 |
| | | 12/1/2005 | 1021 | Replicate | -- | -- | -- | 34 | 1,700 | -- | -- |
| | | 12/21/2005 | 930 | Blank | E.002 | <.004 | <.4 | <1 | <1 | <.10 | 1 |
| | | 12/21/2005 | 1030 | Environ. | .025 | .058 | 7.7 | 130 | 2,000 | .1 | 16 |
| | | 3/7/2006 | 1130 | Environ. | .025 | .076 | 7.3 | 34 | 1,600 | <.10 | 17 |
| | | 3/7/2006 | 1131 | Replicate | -- | -- | -- | 34 | 1,700 | -- | -- |
| | | 4/4/2006 | 1100 | Environ. | .037 | .09 | 14.2 | 43 | 5,200 | .13 | 19 |
| | | 4/4/2006 | 1101 | Replicate | .041 | .101 | 12.7 | 50 | 10,000 | <.10 | 14 |
| | | 7/25/2006 | 945 | Environ. | .071 | .158 | 7 | 93 | 9,200 | <.10 | 6 |
| | | 7/25/2006 | 946 | Replicate | -- | -- | -- | 84 | 12,000 | -- | -- |
| | | 8/8/2006 | 1000 | Environ. | .043 | .111 | 5.5 | 120 | 8,100 | <.10 | 12 |
| | | 8/8/2006 | 1001 | Replicate | .043 | .111 | 6.1 | 150 | 6,900 | <.10 | 13 |
| | | 9/6/2006 | 1100 | Environ. | .034 | .113 | 4 | 26 | 13,000 | <.10 | 6 |
| | | 9/6/2006 | 1130 | Blank | <.004 | <.004 | <.4 | -- | -- | <.10 | <1 |
| | | 10/4/2006 | 945 | Environ. | .056 | .124 | 4.2 | 63 | 4,900 | <.10 | 15 |
| | | 10/4/2006 | 946 | Replicate | -- | -- | -- | 52 | 3,900 | -- | -- |
| | | 10/18/2006 | 1330 | Environ. | .036 | .134 | 18.2 | 610 | 44,000 | .12 | 125 |
| | | 10/18/2006 | 1331 | Replicate | -- | -- | -- | 690 | 39,000 | -- | -- |
| | | 1/15/2007 | 1340 | Environ. | .029 | .122 | 16.1 | 2,100 | 18,000 | <.10 | 110 |
| | | 1/15/2007 | 1341 | Replicate | .028 | .123 | 13.3 | 2,000 | 17,000 | <.10 | 81 |
| | | 5/24/2007 | 1038 | Environ. | -- | -- | -- | 250 | 10,000 | <.10 | 39 |
| | | 5/24/2007 | 1039 | Replicate | -- | -- | -- | 240 | 9,900 | <.10 | 39 |
| | | 8/15/2007 | 1059 | Blank | <.006 | <.008 | | | | | 1 |
| | | 8/15/2007 | 1132 | Environ. | .017 | .081 | 6.5 | 71 | 15,000 | <.10 | 27 |
| Equipment blank | 08070200 | 8/22/2006 | 1330 | Blank | <.02 | <.02 | <.4 | | | <.10 | |
| Equipment blank | 301056095265000 | 11/28/2007 | 1358 | Blank | <.006 | E.006 | -- | -- | -- | -- | 1 |
| Equipment blank | 301056095265000 | 11/28/2007 | 1359 | Blank | -- | -- | .9 | -- | -- | <.10 | -- |

LRL and the long-term method detection level (LT-MDL). The LRL is two times the LT-MDL, and concentrations measured between the LRL and LT-MDL are reported as estimated concentrations (Childress and others, 1999). In 2007, total organic carbon was detected at 0.9 milligram per liter (mg/L) and most likely is attributed to the methanol used to clean the equipment. Methanol is a known contaminant for organic carbon; when equipment is not adequately rinsed, residual methanol will result in detections in blank samples.

Field blanks were collected and processed at sampling sites prior to the collection of environmental samples. Constituent concentrations in field blank samples mostly were less than the LRL except for dissolved ammonia plus organic nitrogen (filtered water in tables 2 and 3), which was detected in all five field blanks; two of the detections were estimated concentrations (less than the LRL). The concentration for the August 15, 2007, environmental sample at the East Fork San Jacinto site was rejected because the detected concentration of dissolved ammonia plus organic nitrogen in the field blank associated with this sample was higher than the reported concentration in the environmental sample.

Split replicate samples (referred to as replicate samples in this report) were collected during the study. Replicate samples are prepared by dividing a single volume of water into multiple samples to provide a measure of the variability of sample processing and analysis. Replicate samples were compared to the associated environmental samples by computing the relative percent difference (RPD) for each constituent. RPD was computed using the equation

$$\mathrm{RPD} = |C_1 - C_2|/((C_1 + C_2)/2) \text{ x } 100, \qquad (1)$$

where

$C_1$= concentration from environmental sample; and
$C_2$= concentration from replicate sample.

RPDs of 10 percent or less indicate good agreement between analytical results if the concentrations are sufficiently large compared to the LRL. The RPD exceeded 10 percent for five of the 48 sample pairs of nutrients, four of six for total organic carbon, two of seven for atrazine, and three of seven for suspended sediment. The RPD exceeded 10 percent primarily when constituent concentrations were at or near the LRL so that small variability in analysis caused large RPDs.

The 17 bacteriological replicate samples were analyzed in the same manner as the environmental samples. The acceptable RPD for bacteriological replicate samples was set at 30 percent. The Colilert method used for *E. coli* and total coliform allows the simultaneous detection of *E. coli* and total coliform and is reported as most probable number. Most probable number analyses result in a statistical estimate of the original number of cells in a known volume of water; results are reported with a 95-percent confidence interval and upper and lower confidence intervals (Stoeckel and others, 2005). The RPD exceeded 30 percent for eight of 34 bacteriological sample pairs. The confidence intervals for the eight replicate samples with RPDs exceeding 30 percent overlapped, indicating there were no statistically significant differences between replicate samples.

# Regression Models to Estimate Concentrations

## Development of Models

The R environment for statistical computing (R Development Core Team, 2006) was used to develop the regression models for estimating real-time concentrations for selected water-quality constituents. Most of the regression methodologies used in this report are described in either Furnival and Wilson (1974) or Helsel and Hirsch (2002). Multiple linear regression analyses were done using the leaps and bounds algorithm devised by Furnival and Wilson (1974), an exhaustive, all-subset method for selecting the preferred model for each constituent. The potential explanatory or predictive variables included discharge (streamflow), specific conductance, pH, water temperature, turbidity, dissolved oxygen, rainfall, and time. Rainfall data were obtained from a rain gage monitored by Harris County Homeland Security and Emergency Management (2009) colocated with the Spring Creek site. Time was investigated as a possible explanatory variable to account for seasonal variations inherent in some concentrations. The explanatory variables provide easily measured quantities as a means to estimate concentrations of the various constituents under investigation with accompanying estimates of measurement uncertainty. Each regression equation can be used to estimate concentrations of a given constituent in real time on the basis of explanatory variables also measured in real time. Corresponding 90-percent prediction intervals can be computed to display the uncertainty associated with the estimate. In conjunction with estimated concentrations, constituent loads also can be estimated by multiplying the estimated concentration by the corresponding streamflow and applying the appropriate conversion factor.

### Transformation

The development of multiple linear regression equations for concentration and load estimation is well documented in previous publications; for example, see Christensen and others (2000) and Ryberg (2006). Normally distributed response and explanatory variables with linear relations and constant variance are required for statistically valid multiple linear regression applications. Natural logarithmic transformations on the response and explanatory variables are commonly used to improve linearity and to compensate for non-normality (data that do not follow a bell-shaped continuous *probability distribution* centered on a mean) and heteroscedasticity (non-constant variance about the regression line) in model residuals. However, various power transformations can sometimes improve regression models more than logarithmic transformations. The theory of power transformations is discussed in Neyman and Scott (1960), Box and Cox (1964), and Hoyle (1968). An evaluation of appropriate variable transformations was made on all potential constituents, physical properties, rainfall, and time data. Exploratory data analyses that were done but not reported include scatter plots of all constituent concentrations in relation to each possible explanatory variable. The scatter plots assisted the authors in determining if transformations were necessary to increase linearity among response and explanatory variables to improve normality and reduce heteroscedasticity. In addition to scatter plots of concentrations in relation to time, boxplots of concentrations grouped by month were also prepared to evaluate possible seasonal variation. Two general methods were used to investigate transformations on the response and explanatory variables: inverse-response plots and Box-Cox procedures. Using the inverse-response plots and Box-Cox procedures three transformations were evaluated: transforming response and explanatory variables, transforming only the response variable, and transforming only the explanatory variables.

Box and Cox (1964) devised a maximum likelihood method for optimizing a transformation of a strictly positive response variable such that the residuals from the regression are as close to normally distributed as possible. The Box-Cox procedure is best summarized by a graph with the potential powers (power estimate) on the horizontal axis and values of the log-likelihood function on the vertical axis, along with a 95-percent confidence interval around the power estimate. The global maximum likelihood estimate is the point that maximizes the curve of the log-likelihood function. Standard likelihood theory can be used to derive the confidence interval (Weisberg, 2005).

Generally, transformations are necessary for the explanatory variables as well as the response variable. A multivariate extension of the Box-Cox procedure was proposed by Velilla (1993) which transforms explanatory variables toward joint normality. Once the explanatory variables have been transformed and are approximately linearly related, the Box-Cox transformation procedure is then used on the response variable. Inverse-response plots provide an additional method for optimizing transformations for the response variable. The inverse-response plots are created by plotting the observed response values on the horizontal axis and fitted response values on the vertical axis. Usually the explanatory variables must be transformed first to achieve linearity. After verifying that explanatory variables were approximately linearly related, an appropriate transformation of the response variable was investigated. Natural-base logarithmic transformations were used. The theory of inverse-response plots for estimating response transformations can be found in Cook and Weisberg (1994), based on results of Li and Duan (1989).

The alr3, car, and MASS packages of R provide functions that implement the inverse-response plots and Box-Cox procedures (R Development Core Team, 2006) used to determine the appropriate transformations. Application of these packages is discussed in Fox (2002), Faraway (2005), and Weisberg (2005). Once the transformation or transformations for the response and explanatory variables are selected, the analysis continues with the selection of explanatory variables that produce a preferable regression model; preferable is ascertained as the best-fit model through numerous statistical techniques described in the following sections.

## Selection of Variables for the Model

The goal of variable selection is to determine the "best" regression model, although when dealing with several independent (predictive) variables simultaneously in a regression analysis, it can be difficult to determine the best model choice; often there are several reasonable candidates from which to choose (Kleinbaum and Kupper, 1978). A thorough process would involve investigating all possible regression models, although this often is a formidable task. With $k$ variables, there are $2^k$ possible regression models. Standard stepwise procedures, including forward selection and backward elimination, only examine $k(k-1)/2$ of the $2^k$ possible subsets. Combining the forward and backward algorithms into a single stepwise method will inspect more subsets than either method alone, but still will not examine all $2^k$ possible models. Furnival and Wilson (1974) established algorithms that when joined form a simple leap and bound technique to find the best subsets of all $2^k$ possible models, without actually examining all subsets. In this case, "best" describes a model as having a minimum residual sum of squares (RSS), also called the error sum of squares (SSE), for a given number of variables. This method will obtain the $n$ best regression models for each subset size. This exhaustive variable search technique to develop the best-fit regression models was accomplished using the "leaps" package, described in detail in Fox (2002) and Faraway (2005).

The coefficient of determination, $R^2$, describes the proportion of the total sample variability in the response explained by the regression model. The coefficient will only increase as additional explanatory variables are added to the model, thus it might not be an appropriate criterion for determining the usefulness of a model that has numerous explanatory variables. The adjusted $R^2$ statistic, denoted as adjusted R-squared, compensates for this by assessing a "penalty" for the number of explanatory variables in the model; adding additional explanatory variables increases the value of adjusted R-squared only when the predictive capability of the model increases. Choosing a model with the highest adjusted R-squared value is equivalent to choosing a model with the lowest mean standard error (Helsel and Hirsch, 2002).

## Evaluation of Candidate Regression Models

When the response and the predictive variables are normally distributed, three criteria-based statistics based on likelihood theory can be used for evaluating regression models. The criteria-based statistics are Akaike's information criterion (*AIC*), corrected *AIC* ($AIC_C$), and the Bayesian information criterion (*BIC*). *AIC* is designed to balance model complexity (the number of predictive variables) and goodness of fit (how well the model probabilities "fit" the observed frequencies [Iman and Conover, 1983]) and is structured in such a way that models with smaller *AIC* values are identified as the models that fit the data the best. *AIC* is known for showing preference or bias for over-fitted models because of the weak penalty for model complexity (Good, 2005). Over-fitted models are numerically unstable and have too many predictive variables; the objective in modeling is predictive capability not goodness of fit (Good, 2005). $AIC_C$ was developed by Hurvich and Tsai (1989) and provides a bias-corrected version of *AIC* when sample size is small or when the number of variables estimated is a moderate to large fraction of the sample size. Burnham and Anderson (2004) recommend using $AIC_C$ unless $n/K$ >40, where $K = (k + 2)$, $n$ is the sample size, and $k$ is the number of variables in the model. In practice $AIC_C$ should be used if possible, and $AIC_C$ converges to *AIC* as $n$ increases. Although similar to *AIC*, *BIC* (proposed by Schwarz [1978]) has a stronger penalty term and thus favors smaller models with fewer predictive variables. In conjunction with the criteria-based statistics, the prediction error sum of squares (*PRESS*) statistic provides an excellent and general measure of the quality of a particular regression equation. An equation that produces the least error when making new predictions is obtained by minimizing the *PRESS* statistic (Helsel and Hirsch, 2002). The *PRESS* statistic is an excellent statistic to use for model-comparison purposes because it can be used to compare nested (models in which the various *factors* are contained within one another in a specific hierarchical order) as well as non-nested models. Only when two models have different response units (different transformation powers of the response variable) is the *PRESS* invalid for model comparison. In the case of prediction based on *PRESS*, the value of *PRESS* for a given model can be used to create an $R^2$-like statistic, denoted $R_p^2$, providing an indication of the relative predictive capability of the *PRESS*-based regression model. Conceptually one would expect *PRESS*-based regression models to explain a greater percentage of the variation (higher $R_p^2$) compared to the proportion of total sample variability explained by a regression model characterized by adjusted R-squared. For the purposes of this investigation, *PRESS* was given considerable weight in the final selection of regression models. In the case of model comparisons between different response transformations, various residual plots and marginal-model plots (Weisberg, 2005) were used to identify preferable models. *AIC*, $AIC_C$, *BIC*, adjusted R-squared, and $R_p^2$ were calculated for each model and used in conjunction with additional regression diagnostics, such as the Breusch-Pagan test, to help determine the best overall models. The Breusch-Pagan test tests a linear regression model to the residuals of the model and rejects the model if the excessive variance is based on extra explanatory variables (Hothorn and others, 2009).

## Normality of Residuals

Another component of regression analysis is the assessment residual normality. The student's $t$-distribution, or $t$-distribution, is the name given to a family of distributions indexed by a variable called degrees of freedom (Iman and Conover, 1983). To satisfy the $t$-distribution normal-population requirement for valid hypothesis tests, normality of the residuals is evaluated with a residual normality test. A powerful procedure for assessing the normality of residuals is the Shapiro-Wilk test (Helsel and Hirsch, 2002), in which the assumption of normality is rejected at an $\alpha$ level ($p$-value) of less than about .05.

The Mann-Kendall nonparametric test for monotonic trends was used in testing the models that appeared to have seasonal patterns. The Mann-Kendall test on residuals is a hybrid procedure—parametric removal of effects of the exogenous variables, followed by a nonparametric test for trend (Helsel and Hirsch, 2002). The test determines if the concentrations or residuals used are independent of time.

## Collinearity

Model selection methods require an investigation of collinearity between predictive variables. Collinearity refers to a linear relation among some or all of the predictive variables in a regression model. When there is collinearity, there is redundancy among predicative variables (Ott and Longnecker, 2001). In addition, many of the predictive variables will have insignificant $t$-values in a full model, where a full model involves all variables of interest, indicating their addition to the model does not improve the model in a statistically significant manner. Kleinbaum and Kupper (1978) note that partial $F$-tests make it possible to partition the regression sum of squares into three components and to determine if the addition of a given predictive variable improves the overall model, taking into account the contributions of other predictive variables already in the model. Partial $F$-tests should be used when many predictive variables have significant $t$-values and there is not high collinearity. Analysis of variance (*ANOVA*) tests use partial $F$-tests for comparison between nested models to determine if the additional variables provide a better explanation of the variation in the response, despite the loss in degrees of freedom. This test is a good check on the significance of additional variables in a model, as it is based on different statistical criteria than stepwise procedures that use *AIC* and *BIC* to assign significance.

Variance inflation factors (*VIF*) are used to check for high collinearity between explanatory variables (Stine, 1995). Explanatory variables carrying similar information about the response have a high collinearity and, when such variables are all included in the model, give rise to increased variance in the estimation of the regression coefficients. A *VIF* represents the increase in variance because of correlation between predictive variables, where a minimum value of 1 occurs when no correlation is present. Typically, *VIF* values greater than 10 are a cause of concern and indicate that a poor estimate of the associated regression coefficient has been produced by the model.

To further assist in model selection, *PRESS* and $R_p^2$ statistics also were computed using the R environment for statistical computing. To ensure predictions from the model are as accurate as possible, residuals of the final (candidate) model must be normally distributed. To ensure this, the Shapiro-Wilk test was used, as well as visual inspection of residual plots. A comprehensive discussion regarding the selection of predictive variables for multiple linear regression analyses using the R system are provided by Fox (2002), Faraway (2005), and Weisberg (2005).

## Graphical Analysis

Graphical analysis is a vital component of regression analysis; it facilitates visual inspection and verification of data patterns such as linearity and constant variance underlying linear regression model theory. Residual plots are used to check if regression models "fit" the observed data. Patterns in residual plots can be used to indicate when the data fail the requirements of a normal distribution with constant variance, warranting further investigation and possible application of transformation techniques. Sometimes, residual plots will indicate heteroscedasticity even when the errors have constant variance. Residuals can be modified (standardized) to assist in the evaluation process. The standardized residual is the ratio of the residual to the residual standard error (Helsel and Hirsch, 2002). Standardized residuals have a mean of 0 and a standard deviation of 1 (Iman and Conover, 1983). By using standardized residuals, patterns in residual plots can be evaluated to indicate an incorrect model fit. When the explanatory variables are approximately linearly related—they follow a multivariate normal distribution—residual patterns provide direct information on how the model has been miss-specified. In addition, plots of standardized residuals show how many estimated standard deviations any point is away from the fitted regression model. Studentized residuals also were used to aid in identifying extreme outliers (Helsel and Hirsch, 2002).

In addition to residual plots, a variety of other visual diagnostics are necessary before candidate models can be evaluated, including rigorous outlier tests. Testing for outliers is like performing *n* significance tests, one for each of the *n* values. To correct for multiple-comparison testing problems when investigating outliers, a Bonferroni correction was used. The Bonferroni correction (commonly referred to as a Bonferroni test) is a multiple-comparison correction used when several dependent or independent *statistical tests* are being performed simultaneously. The Bonferroni correction can be used to help identify outliers, and the R environment for statistical computing reports the Bonferroni *p*-value for the most extreme observation in a script referred to in R documentation as the Bonferroni Outlier Test (R Development Core Team, 2006). To avoid a great deal of spurious positives, the *alpha value* needs to be lowered to account for the number of comparisons being performed (Ott and Longnecker, 2001). This means that for *n* tests, each with size $\alpha$, the probability of falsely rejecting at least one value as an outlier is no greater than $n\alpha$. So, a level of $\alpha/n$ is used for each test to keep the overall level no more than $\alpha$. Implementing this technique is achieved by multiplying the *p*-value returned from a *t*-distribution based outlier test by the sample size (Weisberg, 2005).

Data points with considerable influence on the fit of a regression model are called high leverage points (Rousseeuw and Leroy, 2003). A common rule used to quantify an acceptable upper limit for the leverage and classify a point of high leverage is when a hat value is greater than $2(k + 1)/n$, where $n$ is the number of observations and $k$ is the number of explanatory variables in the model. If a particular value is a high leverage point but also an outlier, it is deemed a bad leverage point, whereas good leverage points have high leverage but are not outliers (Møller and others, 2006). The distinction between good and bad leverage points is necessary and provides valuable information regarding the legitimacy of a data value. Bad leverage points should be examined and additional model investigation attempted before removal of the value is considered. Only when significant statistical evidence has been compiled or previous information of an error associated with the value is known, should the value be declared invalid and deleted from the analysis.

Added-variable plots, sometimes called partial regression plots, provide an additional graphical technique for identifying influential points. Added-variable plots supplement formal testing procedures and provide the analyst with useful information for developing an understanding of data properties without the need for complex data manipulations (Haining, 1990). Effects of other explanatory variables on the response are removed, as discussed in Fox (2002) and Faraway (2005), so the marginal relation between a response and an explanatory variable can be directly and visually assessed. These plots are useful in identifying the influence and leverage of observations on each coefficient in the regression model.

Marginal-model plots provide a graphical method to assess how well a model fits data (Cook and Weisberg, 1997). In essence, marginal-model plots are graphical equivalents of goodness-of-fit tests. Generally, a nonparametric fit of the predicted concentrations based only on the model is compared to a nonparametric fit based only on the measured data. A nonparametric fit of the predicted concentrations is a distribution-free method that extracts information from the data by comparing each value with all others (ranking the data) rather than by computing parameters (Helsel and Hirsch, 2002). In nonparametric regression models the structure of the relation between variables is treated nonparametrically, but there might be parametric assumptions about the distribution of model residuals. When there are multiple explanatory variables, graphs of LOWESS smooths (Helsel and Hirsch, 2002) for each variable are created for comparison. If the two nonparametric estimates agree, then the data are said to be modeled correctly by the parametric model under investigation, as discussed in Weisberg (2005).

Quantile-quantile plots (Q-Q plots), constructed to compare the measured and predicted datasets, provide information on the relation between the two datasets. Q-Q plots (not shown) plot the quantiles of one dataset in relation to the quantiles of the other dataset. If the two datasets come from the same distribution, the quantile pairs will plot along a straight line (Helsel and Hirsch, 2002). These plots are useful to determine if data from the regression model are similar to the environmental data.

## Retransformation Bias Correction

Procedures to develop regression equations for estimating and predicting water-quality constituents have advanced substantially over the last 10 years. For example, regression equations have been developed with a transformed response to fit the observation data better (Christensen and others, 2000). For water-quality modeling, transformed response variables provide an estimate for the concentration value based on instantaneous values of the explanatory variables. When the response is transformed to develop a best-fit model, it must be retransformed to obtain an estimate in the original units. Estimators that are unbiased in the transformed scale will be biased once the retransformation has been made to return the response to the original scale. Retransformation bias corrections are made to remove bias; the form of the bias correction factor will depend on the transformation that was used.

The two response transformations most useful in this analysis were the natural logarithmic and the square-root transformations. The natural logarithmic transformation is useful for describing the relation between concentration and streamflow. Even with additional predictive variables, the concentration for many constituents is well modeled with systematic natural logarithmic transformation. A bias correction is necessary when using natural logarithms to transform the predictive variable if the retransformation yields a median estimate; median estimates tend to underestimate the actual arithmetic mean for water-quality data. Simply inverting a log-transformed response, called a rating curve estimator $L_{RC}$, will return a biased low, inconsistent estimate of the arithmetic mean. Bias increases with the degree of scatter about the rating curve. Extensive research has been done to find estimators that return the expected value of a water-quality streamflow load estimate in real time if the response was log-transformed, for example Duan (1983), Crawford (1991), and Cohn (2005).

A minimum variance unbiased estimator (MVUE), $L_{MVUE}$, derived by Finney (1941) adjusts for this bias and returns an efficient estimate of the mean. $L_{MVUE}$ has the desirable properties of being unbiased and having a function of sufficient statistics with a minimum variance for its expectation. This estimator is the best choice when the log-normal model is correct and the residual errors are approximately normally distributed. Bradu and Mundlak (1970) derive an unbiased estimator for the variance of the MVUE of the mean. Alternative expressions for the exact variance of the MVUE of the mean can be found in Mehran (1973). For an advanced theoretical discussion on the derivation and validity of these estimators, see Hoyle (1968), Likes (1980), and Cohn (2005).

Correcting for the bias associated with the square-root transformation is straightforward. The MVUE of the mean and an expression for the exact variance of the MVUE of the mean are derived from and discussed in Bartlett (1936), Neyman and Scott (1960), and Stuart and others (1999).

One method of representing the ability of a particular model to estimate constituent concentrations involves comparing measured concentrations or values to the predicted concentrations or values obtained by using the regression equation on the model data and computing the RPD.

Discussion so far has focused on the parametric estimators used in this study to correct the bias caused by retransformation of the estimated water-quality properties. For example, Finney's MVUE estimator requires the residuals to be normally distributed, an assumption not commonly met by water-quality data. Nonparametric estimators provided a useful alternative to the retransformation methods. Duan (1983) derived a "smearing" estimator that only requires the residuals to be independent and homoscedastic (constant variance about the regression line). In the case of a base-10 logarithmic transformation, the correction factor involved re-expressing the residuals in the original units and computing their mean. This "factor" was then multiplied by the geometric mean estimate derived when the model is directly inverted. The smearing estimator was generalized for use with other transformations, specifically the square-root transformation.

## Analysis of Censored Data

To avoid false-positive quantification of a constituent, very low concentrations are censored and reported as a "less than" value by the laboratory (Mueller and Spahr, 2005). To compute summary statistics on constituents that contain censored data, three different methods were investigated based on specific criteria relating to the total number of observations and percentage of censored values. These include, but not discussed in depth in this report, the nonparametric Kaplan-Meier (K-M) method, maximum likelihood estimation (MLE), and substitution of a value for the LRL. The mathematical theory underlying each method and the appropriate conditions for implementation are thoroughly discussed in "Nondetects and Data Analysis" (NADA) (Helsel, 2005). An R package is available, also called NADA, that contains numerous functions to handle censored data including the K-M method and MLE for computing summary statistics.

The K-M method, which does not depend on the distributional shape of the data, is generally recommended for data with up to 50-percent censoring and a single LRL. This was the method used at East Fork San Jacinto for nitrite plus nitrate concentrations, because the constituent censored values met this criterion. More than 50 percent of the atrazine concentrations at the East Fork San Jacinto site were less than the LRL and thus censored. Because of the large amount of censored atrazine data for the East Fork San Jacinto site, a regression model to predict atrazine concentrations and loads could not be developed for this site.

For datasets that contain censored values with one or more LRLs, MLE provides a parametric procedure for computing regression models and computing summary statistics such as mean and variance. MLE was used in cases where the constituents had multiple LRLs—for ammonia at both Spring Creek and East Fork San Jacinto sites and for nitrite and orthophosphate at the East Fork San Jacinto site. However, the use of MLE requires the data to approximate a normal distribution. MLE methods also provide poor estimates of the mean and variance when applied to small datasets (sample size less than 30) and should be used only on large datasets. Even though MLE was explored for the constituents listed above and models were developed, the models were not acceptable because of large errors.

For datasets with left-censored observations such as the water-quality constituents evaluated in this study, concentrations and values can be transposed to a right-censored format prior to the regression analysis. As a consequence, response variables will be modeled as if negative values are possible, biasing the response variables. However, this bias is overcome by applying a logarithmic transformation to the response variables. Many of the censored constituents evaluated in this study were skewed and have values spanning a few orders of magnitude; lognormal distribution and MLE regression assumptions (Helsel, 2005) were generally met.

The Spring Creek dataset contained one censored value (<0.10 microgram per liter [μg/L]) for atrazine. For this one data value, one-half the LRL was substituted for the data value in the development of the regression model. Although the substitution might result in a biased estimate of the trend slope, the presence of only a few nondetected values in a dataset (less than 5 percent) is not likely to affect the accuracy substantially (Helsel and Hirsch, 2002). Substitution for this single value did not induce a bias in the data and thus was considered an acceptable approach.

## Best-Fit Models, Spring Creek

Regression models for Spring Creek were developed for all constituents (table 2) analyzed for the study. Models developed and evaluated for total ammonia plus organic nitrogen, dissolved ammonia plus organic nitrogen, ammonia nitrogen, nitrite nitrogen, orthophosphate phosphorous, dissolved phosphorous, and total coliform bacteria were rejected because of large errors associated with the models. Best-fit models developed for nitrite plus nitrate nitrogen, total phosphorous, organic carbon, *E. coli* bacteria, atrazine, and suspended sediment are described in this section.

### Nitrite plus Nitrate Nitrogen

The significant explanatory variables in the best-fit model for estimating total nitrite plus nitrate at the Spring Creek site were specific conductance and pH. The Box-Cox procedure was used to transform response and explanatory variables simultaneously to approximate a normal distribution. A summary of the regression analysis is shown in figure 3.

Measured nitrite plus nitrate concentrations and estimated concentrations from the regression are shown in figure 3. Nearly identical fits in all marginal-model plots (not shown) concur with other regression model diagnostics and demonstrate that measured data are well modeled by the explanatory variables. The marginal-model plots did not show evidence of heteroscedasticity in any of the variables. Increased variability in the variable estimates caused by collinearity between explanatory variables in the model is almost negligible—all *VIF*s were no greater than 2.0.

In the graph of model residuals and estimated concentrations from the regression (fig. 3), samples 33 and 39 are labeled as possible outliers. However, a Bonferroni test failed to reject either concentration as an outlier, so all 39 sampled concentrations were used to construct the model. No discernable patterns are evident in the graph, indicating no inadequacies between the model and observed data. Significant evidence of residual normality is given by the Q-Q plot (not shown) and a large *p*-value (.95) from the Shapiro-Wilk test statistic.

The comparatively high adjusted R-squared of .925 of the best-fit model indicates the explanatory variables specific conductance and pH account for a substantial amount of the variation in the measured nitrite plus nitrate concentration data. The $R_p^2$ value was .916, a further indication that the model has good predictive capabilities. The appropriateness of the model was corroborated by residual analysis and additional diagnostics. The associated residual standard error was 0.091 mg/L (in log-transformed space), and the median RPD was only 4.00 percent.

### Total Phosphorus

The statistically significant explanatory variables in the best-fit model for estimating total phosphorus at the Spring Creek site were specific conductance, water temperature, and turbidity—all logarithmically transformed using the Box-Cox procedure. An inverse-response plot using the transformed explanatory variables returned a minimum RSS with a square-root transformation on the response. A summary of the regression analysis is shown in figure 4.

In the graph of measured and estimated total phosphorus concentrations (fig. 4), the estimated concentrations show a linear pattern. Samples 2 and 9, labeled on the graph, were collected on December 21, 2005, and April 21, 2006, respectively. Investigation of sample 2 revealed it to have the lowest temperature, 10.9 °C, of all samples. Added-variable plots (not shown) helped illustrate this fact. The added-variable plots also show the deviation of sample 9 from the main cluster of samples for all explanatory variables. However, the sample was included in model construction because significant evidence for rejection as an outlier was not determined using

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       pH                SC
 Min.   :6.850   Min.   :147.0
 1st Qu.:7.455   1st Qu.:264.0
 Median :7.740   Median :357.0
 Mean   :7.812   Mean   :366.8
 3rd Qu.:8.070   3rd Qu.:447.5
 Max.   :9.080   Max.   :671.0
```

### Summary of Regression Analysis for the Constituent of:

Nitrite plus Nitrate ($NO_2$ $NO_3$)

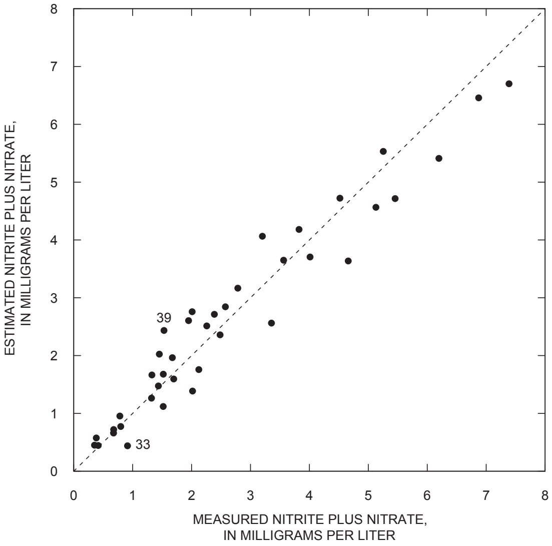
```
SUMMARY STATISTICS FOR NO2NO3, IN MILLIGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.353   1.382   2.017   2.628   3.693   7.388

REGRESSION EQUATION
Call:
lm(formula = (NO2NO3)^(1/3) ~ pH + sqrt(SC))

Residuals:
      Min        1Q    Median        3Q       Max
-0.186761 -0.054057 -0.007496  0.056360  0.219631

Coefficients:
             Estimate Std. Error t-value Pr(>|t|)
(Intercept)  0.366935   0.210589   1.742  0.08997 .
pH          -0.104028   0.030885  -3.368  0.00181 **
sqrt(SC)     0.092959   0.004629  20.082  < 2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.09091 on 36 degrees of freedom
Multiple R-Squared: 0.9288,     Adjusted R-squared: 0.9249
F-statistic: 234.9 on 2 and 36 DF,  p-value: < 2.2e-16
```

### Nomenclature (all potential variables)

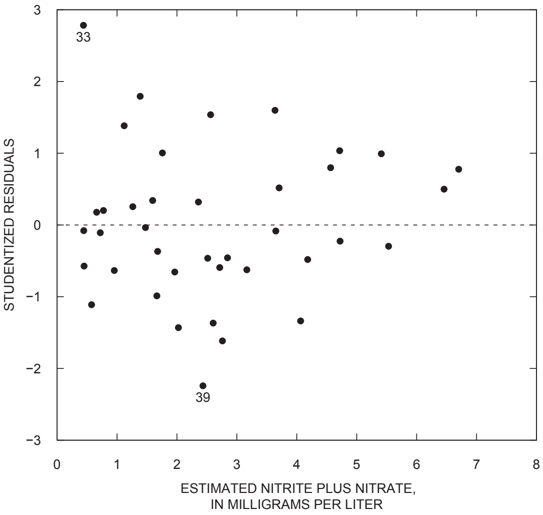
- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day *d* (days into year) divided by 365; and
- log($x$) is natural log of $x$.

Special note: the `poly()` function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The `poly()` function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

------ LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 3.** Summary of regression analysis for nitrite plus nitrate nitrogen for 08068500 Spring Creek near Spring, Texas, 2005–07.

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       SC              Turb              Temp
 Min.   :147.0   Min.   : 11.40   Min.   :10.92
 1st Qu.:261.5   1st Qu.: 20.12   1st Qu.:19.43
 Median :361.0   Median : 39.20   Median :24.11
 Mean   :368.6   Mean   : 72.84   Mean   :23.22
 3rd Qu.:450.8   3rd Qu.: 93.00   3rd Qu.:27.20
 Max.   :671.0   Max.   :303.00   Max.   :31.26
```

### Summary of Regression Analysis for the Constituent of: Total Phosphorous (Phos)

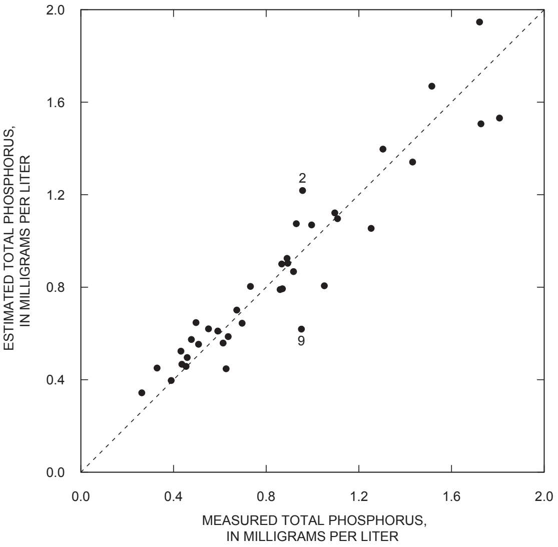
```
SUMMARY STATISTICS FOR Phos, IN MILLIGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2630  0.5188  0.8635  0.8556  1.0370  1.8070

REGRESSION EQUATION
Call:
lm(formula = sqrt(Phos) ~ poly(log(SC), 2) + log(Turb) +
    log(Temp))

Residuals:
     Min       1Q   Median       3Q      Max
-0.12325 -0.04116 -0.01203  0.03820  0.19206

Coefficients:
                    Estimate Std. Error t-value Pr(>|t|)
(Intercept)          0.08056    0.16817   0.479  0.63507
poly(log(SC), 2)1    1.58539    0.11865  13.362 7.25e-15 ***
poly(log(SC), 2)2    0.41861    0.07394   5.662 2.62e-06 ***
log(Turb)            0.10842    0.02064   5.253 8.74e-06 ***
log(Temp)            0.12953    0.04670   2.773  0.00905 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.0735 on 33 degrees of freedom
Multiple R-Squared: 0.8949,     Adjusted R-squared: 0.8821
F-statistic: 70.21 on 4 and 33 DF,  p-value: 1.142e-15
```

### Nomenclature (all potential variables)

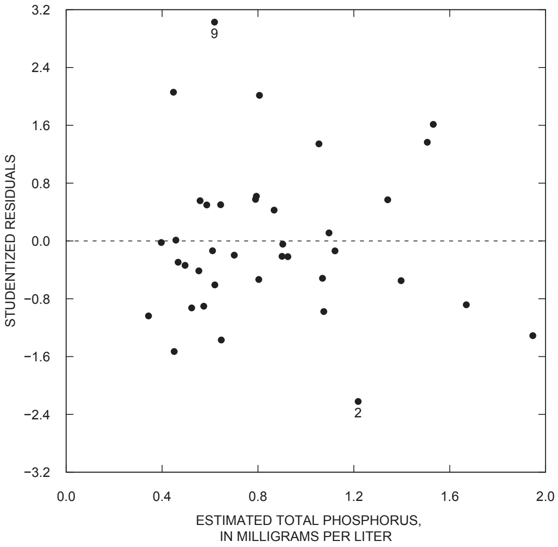
- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day *d* (days into year) divided by 365; and
- $\log(x)$ is natural log of *x*.

Special note: the poly() function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The poly() function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

- - - - - LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 4.** Summary of regression analysis for total phosphorus for 08068500 Spring Creek near Spring, Texas, 2005–07.

the Bonferroni outlier test in the R environment for statistical computing (R Development Core Team, 2006).

Marginal-model plots indicate phosphorus concentration is strongly correlated with specific conductance. The relation of water temperature with phosphorus concentration was highly variable compared to other explanatory variables, also illustrated by marginal-model plots.

The adjusted R-squared for the best-fit regression model was .882, with a corresponding residual standard error of 0.0735 mg/L (in log-transformed space) and a median RPD of 4.45 percent.

Results for the total phosphorus sample collected on January 26, 2006, were unavailable, leaving 38 of the 39 samples to develop the model. The missing value from January 26, 2006, was the only total phosphorus sample collected during January. Based on graphical evidence, periodic variables that include data for January will likely improve predictive capabilities of the regression model.

## Organic Carbon

Explanatory variables in the best-fit model for estimating organic carbon concentrations at the Spring Creek site were specific conductance, turbidity, and sine and cosine terms for seasonal fluctuations, where *Date* represents the Julian day of the year as a fraction of the year, normalized between 0 and 1. A summary of the regression analysis is shown in figure 5.

The marginal-model plots (not shown) show nearly identical fits for all explanatory variables used in the final model, and no discernable patterns that might have been of concern were detected in the graph of model residuals and estimated concentrations. Three samples were identified as possible outliers (samples 7, 14, and 31; fig. 5), as those samples show slight disagreement to the best-fit model.

Associated with the best-fit model is an adjusted R-squared of .756, a residual standard error of 0.0184 mg/L (in log-transformed space), and a median RPD of 4.41 percent. Additionally, all *VIF*s are less than 3.0. Organic carbon concentrations varied seasonally; seasonality effects were visually evident in boxplots of concentrations grouped by month (not shown).

Three samples were not used in model development and are not shown in figure 5. Sample 28, collected on November 28, 2006 (51.5 mg/L) and sample 35, collected on July 25, 2007 (17.4 mg/L) were identified as statistically significant outliers and were not used in the development of the model. Additionally, sample 32, collected on April 24, 2007, became contaminated and could not be analyzed, leaving 36 of the 39 samples to develop the model. Organic carbon concentrations ranged from 7.62 to 21.0 mg/L (table 2), with a median concentration of 11.5 mg/L.

Concentrations showed significant monotonic correlation (*y* generally increases or decreases as *x* increases) with streamflow, dissolved oxygen, specific conductance, and turbidity. Streamflow and turbidity exhibited positive correlation with concentrations, whereas specific conductance and dissolved oxygen exhibited negative correlation with concentrations. A similar, significant correlation between explanatory variables was observed. Also substantial intra-variable correlations between explanatory variables exist. The strong intra-variable relations between concentration and explanatory variables indicate additional samples (larger degrees of freedom in the regression) most likely will change significant explanatory variables.

## *Escherichia Coli*

The statistically significant explanatory variables included in the best-fit regression model for estimating *E. coli* were streamflow and rain. A binary variable was used to indicate when collection times either coincided with or were within 24 hours of a storm. A summary of the regression analysis is shown in figure 6.

A power transformation provided the best-fit model. The quadratic relation between logarithmically transformed *E. coli* and streamflow was visually confirmed with a scatter plot (not shown). Homoscedasticity of the residuals was demonstrated by the residual plots, whereas marginal-model plots indicate that the model adequately fits the measured data.

The largest residual value was 2.84 for sample 38, collected on December 18, 2007. *E. coli* for sample 38 was 3,255 MPN/100 mL. There was no statistically significant evidence to warrant discarding this value as an outlier on the basis of investigating the residuals. Rather, strong evidence in support of residual normality was provided by the Shapiro-Wilk *p*-value of .92 and an acceptable a Q-Q plot (not shown).

The regression equation has an adjusted R-squared of about .812, a residual standard error of 0.924 MPN/100 mL (in log-transformed space), and a median RPD of 9.78 percent. All 38 sampled *E. coli* values and corresponding explanatory variable values appear legitimate and thus were included in the development of the regression model. The sample collected on May 22, 2007, was 41,000 MPN/100 mL, a relatively large value when compared to other samples, but evidence for removing the sample as an outlier was not determined. The largest organic carbon concentration, 21.0 mg/L, also occurred on this date.

The sample collected on September 7, 2007, was contaminated, leaving no estimate to use with the available streamflow value for this date, 222 cubic feet per second. Two other *E. coli* values were associated with streamflow values of 214 and 225 cubic feet per second, respectively, indicating that the amount of information lost because of the missing *E. coli* value was not substantial.

## Atrazine

Streamflow, turbidity, and seasonal, periodic terms were the explanatory variables in the best-fit regression model for atrazine at the Spring Creek site. Logarithmic transformations were used on streamflow and turbidity, whereas the response required a one-third power transformation. All variable

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       SC              Turb              Date
 Min.   :147.0   Min.   : 11.40   Min.   :0.07123
 1st Qu.:266.5   1st Qu.: 20.18   1st Qu.:0.30206
 Median :361.0   Median : 44.60   Median :0.47945
 Mean   :371.8   Mean   : 75.69   Mean   :0.51792
 3rd Qu.:458.8   3rd Qu.: 99.55   3rd Qu.:0.73014
 Max.   :671.0   Max.   :303.00   Max.   :0.97260
```

### Summary of Regression Analysis for the Constituent of: Total Organic Carbon (OrgC)

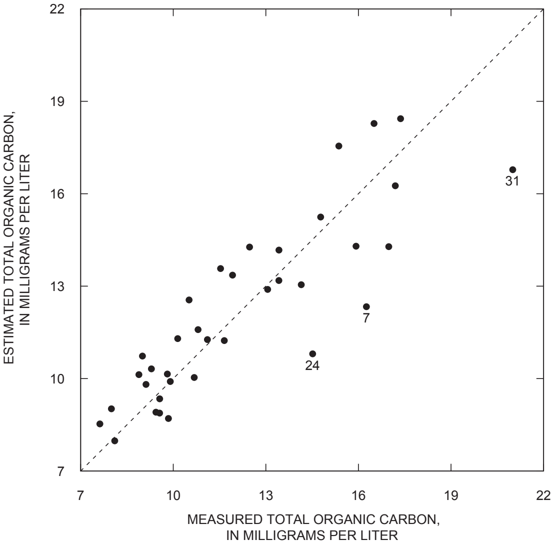
```
SUMMARY STATISTICS FOR OrgC, IN MILLIGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.622   9.559  11.320  12.190  14.580  21.000

REGRESSION EQUATION
Call:
lm(formula = Con3 ~ log(SC) + Turb3 + sin(4 * pi * Date) +
    cos(4 * pi * Date))

Residuals:
      Min        1Q    Median        3Q       Max
-0.043305 -0.010962  0.001392  0.014564  0.026522

Coefficients:
                   Estimate Std. Error t-value Pr(>|t|)
(Intercept)        0.120898   0.061275   1.973 0.057463 .
log(SC)            0.024127   0.011398   2.117 0.042406 *
Turb3              0.484436   0.115679   4.188 0.000216 ***
sin(4 * pi * Date) 0.009733   0.004725   2.060 0.047887 *
cos(4 * pi * Date) 0.018276   0.004458   4.100 0.000277 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.01845 on 31 degrees of freedom
Multiple R-Squared: 0.784,      Adjusted R-squared: 0.7561
F-statistic: 28.13 on 4 and 31 DF,  p-value: 6.352e-10
```

### Nomenclature (all potential variables)

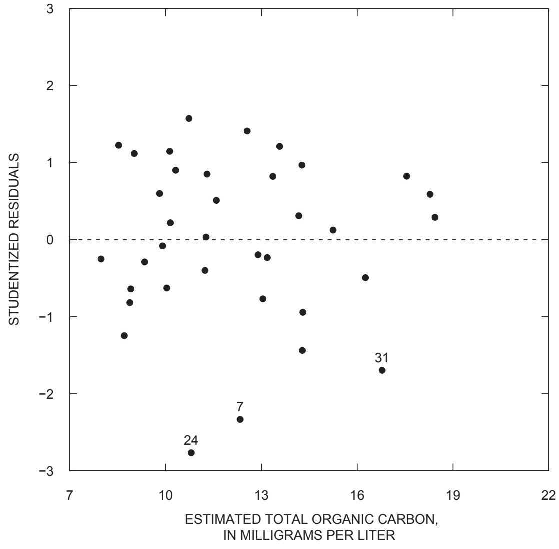
- $Q$ is streamflow, in cubic feet per second;
- $pH$ is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day $d$ (days into year) divided by 365; and
- $\log(x)$ is natural log of $x$.

Special note: the `poly()` function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The `poly()` function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

- - - - - LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 5.** Summary of regression analysis for total organic carbon for 08068500 Spring Creek near Spring, Texas, 2005–07.

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       Q             Rain
 Min.   :  18.40   Yes: 11
 1st Qu.:  38.64   No : 27
 Median :  73.50
 Mean   : 520.37
 3rd Qu.: 180.25
 Max.   :5350.00
```

### Summary of Regression Analysis for the Constituent of: Escherichia coli (ECB)

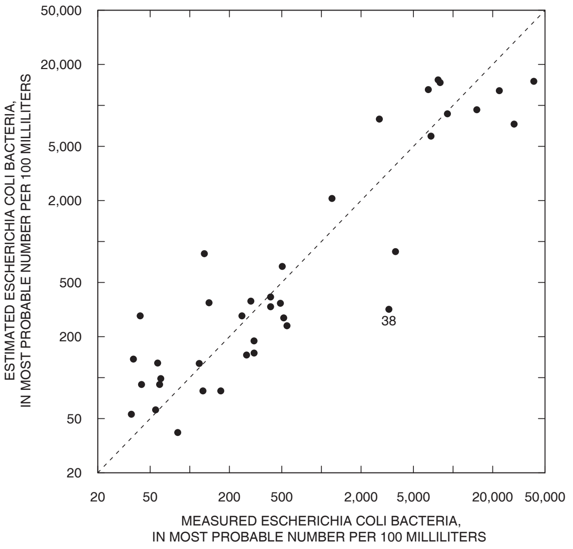
```
SUMMARY STATISTICS FOR ESCHERICHIA COLI (ECB), IN MOST-PROBABLE
    NUMBER (MPN) PER 100 MILLILITERS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   36.0   120.0   359.5  4266.0  3554.0 41060.0

REGRESSION EQUATION
Call:
lm(formula = log(ECB) ~ poly(log(Q), 2) + Rain)

Residuals:
      Min        1Q    Median        3Q       Max
-1.913527 -0.596184 -0.005551  0.597244  2.325746

Coefficients:
                Estimate Std. Error t-value Pr(>|t|)
(Intercept)       5.6688     0.2105  26.932  < 2e-16 ***
poly(log(Q), 2)1  5.3790     1.4275   3.768 0.000626 ***
poly(log(Q), 2)2 -3.1014     0.9240  -3.356 0.001955 **
Rain              2.3830     0.5107   4.666 4.63e-05 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.9236 on 34 degrees of freedom
Multiple R-squared: 0.8268,     Adjusted R-squared: 0.8115
F-statistic: 54.08 on 3 and 34 DF,  p-value: 4.962e-13
```

### Nomenclature (all potential variables)

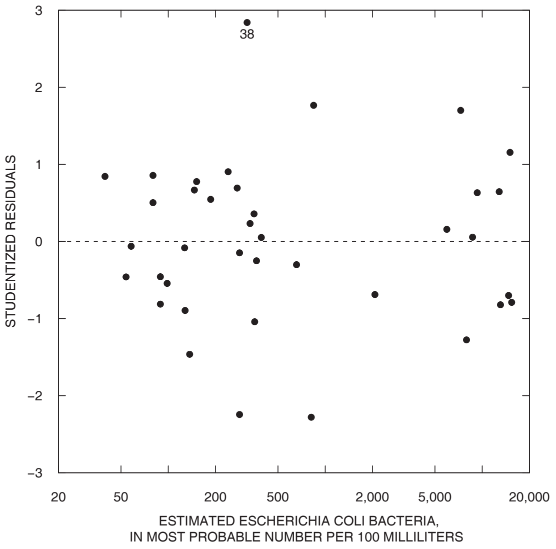
- $Q$ is streamflow, in cubic feet per second;
- $pH$ is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day $d$ (days into year) divided by 365; and
- $\log(x)$ is natural log of $x$.

Special note: the `poly()` function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The `poly()` function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

------ LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 6.** Summary of regression analysis for *Escherichia coli* for 08068500 Spring Creek near Spring, Texas, 2005–07.

transformation procedures discussed in the “Methods” section were investigated before a final regression model was selected. A variation in concentrations during the year is evident in a concentration-by-month boxplot (not shown). A period of 4π is used with the sine and cosine terms in the model to account for the apparent seasonality. A summary of the regression analysis is shown in figure 7.

The residuals for atrazine adhered to a normal distribution and were independent and homoscedastic; two residual plots are shown (fig. 7). Three samples are labeled in each graph; samples 9 and 11 are the two largest measured concentrations and do not provide any evidence against the best-fit model. The second largest turbidity of 268.7 Formazine Nephelometric Units (FNU), measured on May 5, 2006, corresponds to sample 11. The marginal-model plots provided additional evidence in favor of this model. There was no significant evidence to indicate any problems in the model associated with collinearity or heteroscedasticity. Residual normality is assumed based on the approximate linear relation in a Q-Q plot (not shown) and the Shapiro-Wilk test *p*-value of .21. The best-fit model has an associated adjusted R-squared of .745, a residual standard error of 0.175 µg/L (in log-transformed space), and a 9.95 percent median RPD.

Two samples, collected on March 21, 2006 (14.0 µg/L) and March 13, 2007 (11.0 µg/L), were discarded from the data and not used to construct the best-fit regression model. Atrazine concentrations for the two samples differed considerably from the rest of the atrazine data collected at the Spring Creek site. During construction of the regression model, it was determined that including the two extreme atrazine concentrations only reduced model accuracy, and these values were consistently identified as significant outliers.

The atrazine value for the sample collected on September 6, 2006, was censored (<0.1 µg/L). This one sample accounts for less than 3 percent of the total number of samples. More model investigation techniques are available for non-censored data compared to censored data. Therefore, model investigation techniques for censored data were not needed. A concentration of 0.05 µg/L was used in place of the censored value for model development.

## Suspended Sediment

Suspended sediment is an important indicator of the quality of surface water conveyed to Lake Houston. Constituents such as nitrogen and phosphorus are readily transported in streams by their adsorption to silt and clay particles (Mueller and Spahr, 2006). Suspended sediment conveyed to Lake Houston generally comes from seven major tributaries (Sneck-Fahrer and others, 2005). Suspended sediment is defined as sand, silt, or clay depending on particle diameter. Sand is defined as particles less than or equal to 4.00 millimeters and greater than 0.062 millimeter; silt is defined as particles less than or equal to 0.062 millimeter and greater than 0.004 millimeter; and clay is defined as particles less than or equal to 0.004 millimeter (Barlow, 1997).

Statistically significant explanatory variables in the best-fit model for estimating suspended-sediment concentrations at the Spring Creek site were streamflow, water temperature, and turbidity. Box-Cox procedures indicated logarithmic transformations were most suitable for the explanatory and response variables. A summary of the regression analysis is shown in figure 8.

A graph of the estimated suspended-sediment concentrations for the regression model designed to predict real-time concentrations, as well as measured suspended-sediment concentrations, is shown in figure 8. Of 39 suspended-sediment samples collected, 38 were available for developing the regression model (the sample collected on February 27, 2007, was ruined).

A graph of model residuals and estimated concentrations from the best-fit regression model is shown in figure 8. The relatively high adjusted R-squared value (.917) indicates streamflow, water temperature, and turbidity account for a substantial amount of the variation in the measured data. Residual plots, marginal-model plots used to graphically depict goodness of fit (not shown), and additional diagnostics confirm the goodness of fit of the best-fit model. The residual standard error for the model was 0.325 mg/L (in log-transformed space), the Shapiro-Wilk statistic had an associated *p*-value of .889, and the median RPD was 4.50 percent. The LOWESS smooths for all marginal-model plots (not shown) were nearly identical, indicting this model provides an adequate fit to the data. All *VIF*s were less than 4.0, indicating there was not a high degree of collinearity among the predictive variables included in the best-fit model.

The statistical significance of streamflow, water temperature, and turbidity as explanatory variables seems logical, as these are well-known indicators of suspended-sediment concentration. Turbidity is inversely proportional to transparency depth and is a measure of the scattering of light (American Society for Testing and Materials, 2003). Thus, a decrease in the amount of transparency in water seems appropriate as a possible indicator for an increase in the amount of suspended sediment in the water.

## Best-Fit Models, East Fork San Jacinto River

Regression models for East Fork San Jacinto River were developed for all constituents analyzed for the study (table 2), except atrazine. Models developed and evaluated for total ammonia plus organic nitrogen, dissolved ammonia plus organic nitrogen, ammonia nitrogen, nitrite nitrogen, orthophosphate phosphorous, dissolved phosphorous, and total coliform bacteria were rejected because of large errors associated with the models. An atrazine regression model was not developed for the study because more than 50 percent of the data were below the LRL (censored). Best-fit models developed for nitrite plus nitrate nitrogen, total phosphorous, organic carbon, *E. coli* bacteria, and suspended sediment are described in this section.

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       Q                 Turb              Date
 Min.   :  18.40   Min.   : 11.4   Min.   :0.07123
 1st Qu.:  38.09   1st Qu.: 19.3   1st Qu.:0.32329
 Median :  70.00   Median : 49.5   Median :0.49041
 Mean   : 558.46   Mean   : 75.5   Mean   :0.51824
 3rd Qu.: 199.50   3rd Qu.:102.8   3rd Qu.:0.70274
 Max.   :5350.00   Max.   :303.0   Max.   :0.97260
```

### Summary of Regression Analysis for the Constituent of: Atrazine (Atz)

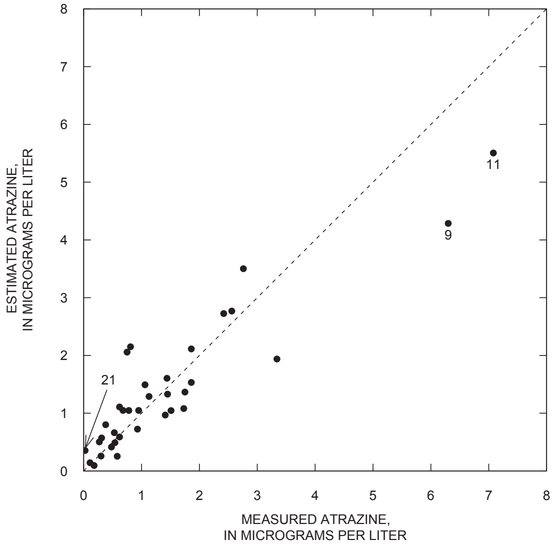
```
SUMMARY STATISTICS FOR Atz, IN MICROGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.025   0.535   0.930   1.414   1.740   7.080

REGRESSION EQUATION
Call:
lm(formula = (Atz)^(1/3) ~ log(Q) + log(Turb) + Date +
    sin(4 * pi * Date) + cos(4 * pi * Date))

Residuals:
      Min        1Q    Median        3Q       Max
-0.381460 -0.100473 -0.008568  0.111630  0.267816

Coefficients:
                    Estimate Std. Error t-value Pr(>|t|)
(Intercept)          0.80162    0.14035   5.712 3.52e-06 ***
log(Q)              -0.10968    0.03372  -3.253 0.002897 **
log(Turb)            0.29973    0.05552   5.398 8.39e-06 ***
Date                -0.84118    0.12125  -6.937 1.26e-07 ***
sin(4 * pi * Date) -0.19103    0.04448  -4.295 0.000179 ***
cos(4 * pi * Date) -0.12463    0.04387  -2.841 0.008142 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.175 on 29 degrees of freedom
Multiple R-squared: 0.7823,     Adjusted R-squared: 0.7447
F-statistic: 20.84 on 5 and 29 DF,  p-value: 8.379e-09
```

### Nomenclature (all potential variables)

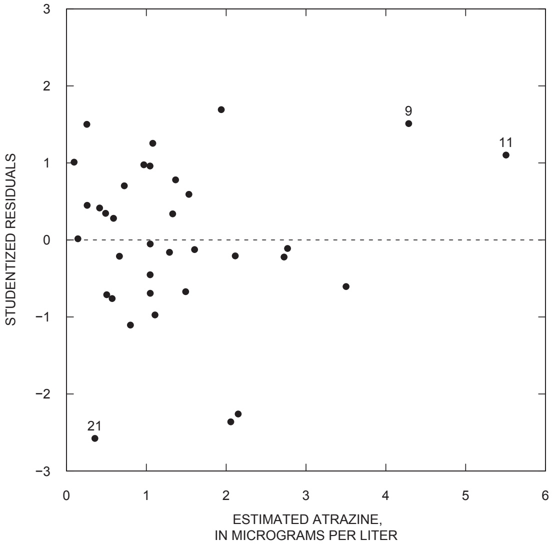
- $Q$ is streamflow, in cubic feet per second;
- $pH$ is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day $d$ (days into year) divided by 365; and
- $\log(x)$ is natural log of $x$.

Special note: the poly() function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The poly() function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

------ LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 7.** Summary of regression analysis for atrazine for 08068500 Spring Creek near Spring, Texas, 2005–07.

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       Q                 Turb             Temp
 Min.   :  18.40   Min.   : 11.60   Min.   :10.92
 1st Qu.:  38.64   1st Qu.: 20.45   1st Qu.:19.43
 Median :  78.50   Median : 44.60   Median :24.11
 Mean   : 524.37   Mean   : 74.12   Mean   :23.07
 3rd Qu.: 206.75   3rd Qu.: 93.00   3rd Qu.:27.20
 Max.   :5350.00   Max.   :303.00   Max.   :31.26
```

### Summary of Regression Analysis for the Constituent of: Suspended Sediment (SS)

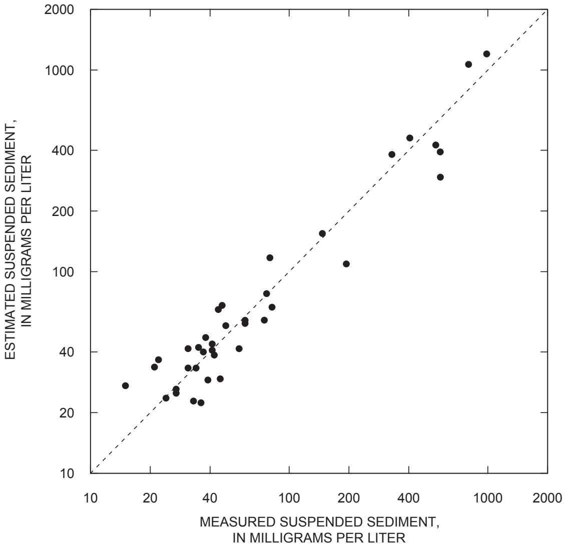
```
SUMMARY STATISTICS FOR SS, IN MILLIGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   34.25   44.50  152.90   81.50  987.00

REGRESSION EQUATION
Call:
lm(formula = log(SS) ~ log(Q) + poly(log(Turb), 2) + log(Temp))

Residuals:
      Min        1Q     Median        3Q       Max
-0.594926 -0.194488 -0.002587  0.241283  0.673800

Coefficients:
                     Estimate Std. Error t-value Pr(>|t|)
(Intercept)           0.68026    0.67540   1.007 0.321167
log(Q)                0.42009    0.06664   6.304 3.96e-07 ***
poly(log(Turb), 2)1   2.64107    0.60731   4.349 0.000124 ***
poly(log(Turb), 2)2   0.95660    0.35300   2.710 0.010592 *
log(Temp)             0.50494    0.19524   2.586 0.014303 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.3254 on 33 degrees of freedom
Multiple R-Squared: 0.9263,     Adjusted R-squared: 0.9174
F-statistic: 103.7 on 4 and 33 DF,  p-value: < 2.2e-16
```

### Nomenclature (all potential variables)

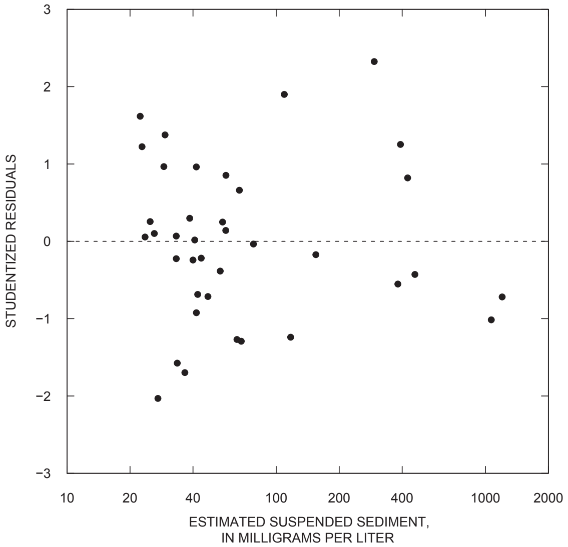
- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day *d* (days into year) divided by 365; and
- log(*x*) is natural log of *x*.

Special note: the `poly()` function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The `poly()` function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

------ LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if 18 present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 8.** Summary of regression analysis for suspended sediment for 08068500 Spring Creek near Spring, Texas, 2005–07.

## Nitrite plus Nitrate Nitrogen

Of the 38 nitrite plus nitrate samples collected, three resulted in censored values less than the nitrite plus nitrate LRL of 0.06 mg/L in 2007 or 0.04 mg/L in 2008. The collection dates of these three samples were October 19, 2006, November 9, 2006, and January 18, 2007, coinciding with the three highest streamflow values (7,930; 4,570; and 5,160 cubic feet per second, respectively). Evidence of dilution is provided by a negative quadratic relation of nitrite plus nitrate concentration with streamflow in the regression model, as well as by scatter plots (not shown) using transformed and untransformed variables. MLE regression was used for model development for the three censored values for nitrite plus nitrate.

Streamflow and a seasonal term were chosen as explanatory variables for estimating the concentration of nitrite plus nitrate. The best-fit regression model was determined by logarithmically transforming streamflow using the Box-Cox procedure and by applying a quadratic seasonality adjustment, where *Date* represents the Julian day of the year as a fraction of the year, normalized between 0 and 1. A summary of the regression analysis is shown in figure 9.

The graph of residuals and estimated concentrations (Helsel and Hirsch, 2002) shows a homoscedastic pattern, indicative of a valid model. Sample 7, collected on March 21, 2006, had a slightly elevated studentized residual value of 2.66 for the concentration of nitrite plus nitrate, so it was investigated further. A Bonferroni test on the studentized residuals did not reject sample 7 as an outlier. Including sample 7 in the analysis, evidence for residual normality was provided by acceptable residual plots (not shown) and a Shapiro-Wilk test *p*-value of .93.

Regression diagnostics and residual analysis provided evidence of model validity. Marginal-model plots provided visual evidence that nitrite plus nitrate data are well modeled by the explanatory variables of streamflow and a seasonal term. No evidence of collinearity between explanatory variables was determined; all *VIF*s are less than 2.0. An adjusted R-squared of .712, a residual standard error of 0.374 mg/L (in log-transformed space), and a median RPD of 9.98 percent are associated with this model.

The five largest streamflow values, ranging from 1,490 to 7,930 cubic feet per second, coincide with nitrite plus nitrate samples collected during the cooler months of October, November, and January. Corresponding streamflow values did not exceed 237 cubic feet per second for the nitrite plus nitrate samples collected during the warmer months (April through September). A boxplot of nitrite plus nitrate concentrations by month (not shown) revealed an approximate unimodal (one highest value in the annual distribution) shape, with larger concentrations occurring during the summer months. No samples were collected during June 2007, leaving only two concentrations for June 2006 to estimate the periodic component of the model. Both concentrations for June 2006 are lower than expected compared to measured concentrations during surrounding months.

## Total Phosphorus

The best-fit model for total phosphorus contained the explanatory variables streamflow, specific conductance, turbidity, and the periodic functions sine and cosine, with periods of $4\pi$, to adjust for seasonal effects. Both streamflow and specific conductance were logarithmically transformed, whereas turbidity and the modeled response did not require transformation. A summary of the regression analysis is shown in figure 10.

No discernable patterns are evident in the studentized residual plot (fig. 10). The total phosphorus sample collected on January 24, 2006, and labeled as sample 3 in the residual plot has a residual value of about -2.89, although a Bonferroni outlier test did not identify it as an outlier. The largest phosphorus concentration of 0.210 mg/L (measured sample 35, collected October 16, 2007) is greater than three times the interquartile range (IQR), 0.0302 mg/L, of the measured concentrations. Inspection of a Cook's distance plot (not shown) (measure of the influence of each observation on the regression coefficients [Dalgaard, 2008]), the residual plot, and the plot of measured and estimated concentrations identify this as a point of high leverage, but low influence.

Satisfactory residual plots (not shown) and a Shapiro-Wilk *p*-value of .68 indicate the normality assumption among the residuals is met. Acceptable marginal-model plots (not shown) for all explanatory variables provide additional evidence in favor of this model. Additionally, the marginal-model plots illustrate the dissimilarity sample 35 has with all other streamflow and specific conductance values.

The total phosphorus sample collected on July 25, 2006, with a high concentration of 0.158 mg/L, was discarded and not used in the regression analysis because the Bonferroni outlier test identified it as an outlier. One of the two lowest suspended-sediment concentrations (6 mg/L) was measured on July 25, 2006. The largest nitrite plus nitrate concentration also was measured on July 25, 2006, and had a relatively large deviation from the line of best fit in the graph of measured and estimated concentrations. The adjusted R-squared for the best-fit regression model was .719, with a corresponding residual standard error of 0.0159 mg/L (in log-transformed space) and a median RPD of 8.00 percent.

## Organic Carbon

Explanatory variables in the best-fit regression model for estimating organic carbon at the East Fork San Jacinto site were streamflow and turbidity. The Box-Cox procedure was used for the logarithmic transformations of explanatory and response variables. Results from the regression analysis are summarized in figure 11.

The residuals displayed the desired random pattern lacking heteroscedasticity (fig. 11). The samples collected on March 4, 2006 (sample 8) and June 13, 2006 (sample 13), labeled in figure 11, have the largest residual values (absolute), although neither was identified as an outlier and discarded

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       Q              Date
 Min.   :  11.0   Min.   :0.0411
 1st Qu.:  37.0   1st Qu.:0.2671
 Median :  67.5   Median :0.5644
 Mean   : 680.7   Mean   :0.5300
 3rd Qu.: 146.8   3rd Qu.:0.7938
 Max.   :7930.0   Max.   :0.9726
```

### Summary of Regression Analysis for the Constituent of:
Nitrite plus Nitrate ($NO_2$ $NO_3$)

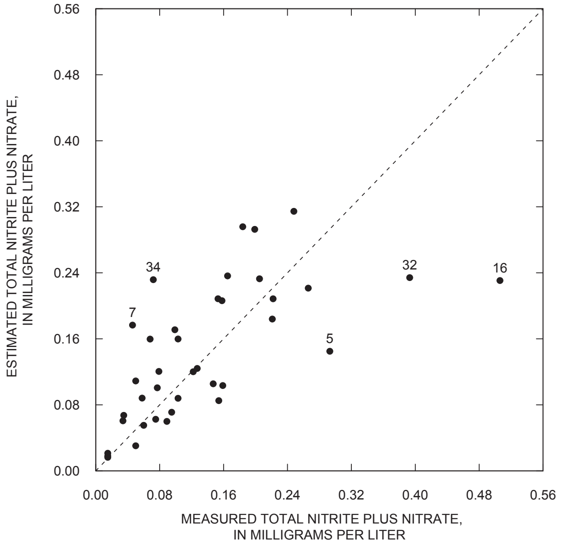
```
SUMMARY STATISTICS FOR NO2NO3, IN MILLIGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0150  0.0620  0.1030  0.1358  0.1793  0.5060

REGRESSION EQUATION
Call:
lm(formula = (NO2NO3)^(-1/3) ~ poly(log(Q), 2) + poly(Date, 2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.56114 -0.23378  0.00617  0.19357  0.88907

Coefficients:
                 Estimate Std. Error t-value Pr(>|t|)
(Intercept)       2.24555    0.06073  36.976  < 2e-16 ***
poly(log(Q), 2)1  2.06213    0.41318   4.991 1.89e-05 ***
poly(log(Q), 2)2  2.14896    0.40289   5.334 6.89e-06 ***
poly(Date, 2)1    0.59724    0.38014   1.571 0.125694
poly(Date, 2)2    1.75588    0.43421   4.044 0.000297 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.3744 on 33 degrees of freedom
Multiple R-Squared: 0.7434,     Adjusted R-squared: 0.7123
F-statistic:  23.9 on 4 and 33 DF,  p-value: 2.371e-09
```

### Nomenclature (all potential variables)

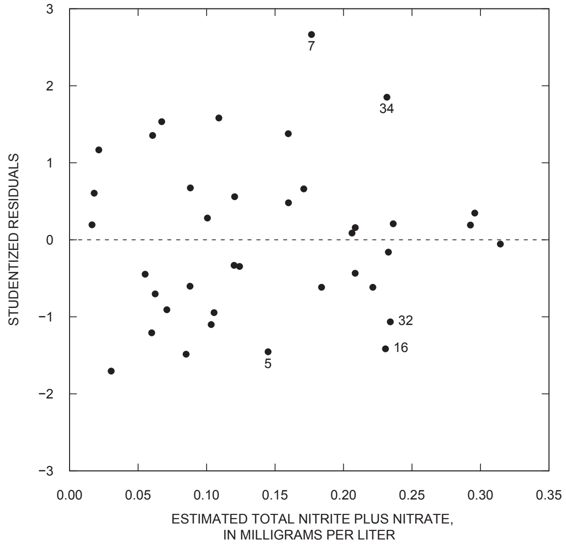
- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day *d* (days into year) divided by 365; and
- log(*x*) is natural log of *x*.

Special note: the poly() function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The poly() function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

------ LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 9.** Summary of regression analysis for nitrite plus nitrate nitrogen for 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–07.

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
          Q                     SC                  Turb                 Date
 Min.    :  11.0   Min.    : 49.0   Min.    : 7.30   Min.    :0.0411
 1st Qu.:  37.0   1st Qu.:146.0   1st Qu.:11.80   1st Qu.:0.2575
 Median :  73.0   Median :180.0   Median :15.60   Median :0.5644
 Mean    : 698.1   Mean    :168.1   Mean    :27.75   Mean    :0.5291
 3rd Qu.: 151.0   3rd Qu.:209.0   3rd Qu.:39.20   3rd Qu.:0.7945
 Max.    :7930.0   Max.    :277.0   Max.    :90.50   Max.    :0.9726
```

### Summary of Regression Analysis for the Constituent of:
Total Phosphorous (Phos)

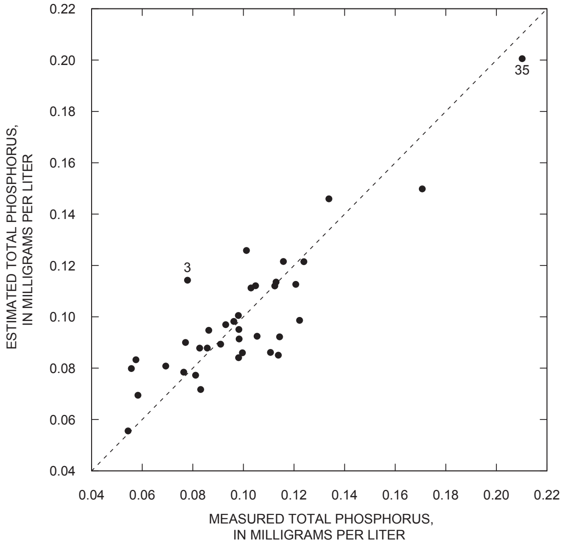
```
SUMMARY STATISTICS FOR Phos, IN MILLIGRAMS PER LITER
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05440 0.08270 0.09820 0.09983 0.11290 0.21020

REGRESSION EQUATION
Call:
lm(formula = Phos ~ log(Q) + log(SC) + Turb +
     sin(4 * pi * Date) + cos(4 * pi * Date))

Residuals:
       Min         1Q    Median         3Q       Max
-0.036360 -0.008247 -0.001157  0.009646  0.028747

Coefficients:
                         Estimate  Std. Error t-value Pr(>|t|)
(Intercept)          0.2781701  0.0594032   4.683 5.32e-05 ***
log(Q)              -0.0155395  0.0030207  -5.144 1.42e-05 ***
log(SC)             -0.0287821  0.0098910  -2.910  0.00664 **
Turb                 0.0013905  0.0001925   7.223 4.01e-08 ***
sin(4 * pi * Date) -0.0052041  0.0039071  -1.332  0.19258
cos(4 * pi * Date) -0.0121902  0.0037978  -3.210  0.00309 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.01593 on 31 degrees of freedom
Multiple R-Squared: 0.758,      Adjusted R-squared: 0.719
F-statistic: 19.42 on 5 and 31 DF,  p-value: 9.824e-09
```

### Nomenclature (all potential variables)

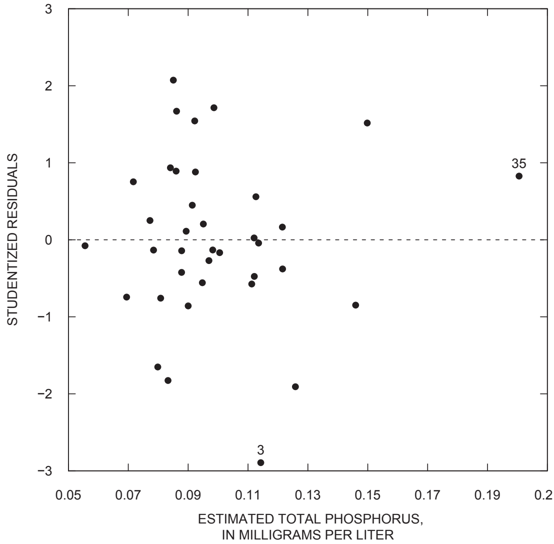
- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day *d* (days into year) divided by 365; and
- log($x$) is natural log of $x$.

Special note: the `poly()` function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The `poly()` function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

- - - - - LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 10.** Summary of regression analysis for total phosphorus for 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–07.

from the dataset. The sample collected on June 13, 2006, coincided with the highest specific conductance (277 microsiemens per centimeter at 25 °C) and the lowest streamflow (11 cubic feet per second) of all measurements. Although specific conductance is not included in the model as an explanatory variable, there was a strong inverse correlation between organic carbon and specific conductance. Regression models are most reliable when the range of predicted values is similar to the range of measured values; there is increased uncertainty associated with predictions near or beyond the limits defined by actual measured values.

Strong agreement between the LOWESS smooths in the marginal-model plots for all four explanatory variables (not shown) provide visual evidence that the model fits the measured data well. The Breusch-Pagan test (R documentation, R Development Core Team, 2006) indicates the data are homoscedastic (heteroscedasticity was rejected with an approximate *p*-value of .886), and *VIF*s were all less than 2.0, providing statistical evidence that assumptions regarding constant error variance have been satisfied and that collinearity among explanatory variables is not present.

A linear relation between measured and estimated concentrations is displayed in figure 11, and samples with large residuals (sample 8 collected on April 4, 2006, and sample 13 collected on June 13, 2006) are labeled for comparison. In addition to the results of the marginal-model plots, Breusch-Pagan test results, and *VIF*s, the goodness of fit of the final regression model was further substantiated by the adjusted R-squared of .878 with a residual standard error of 0.153 mg/L (in log-transformed space). The median RPD was 4.82 percent, and the *PRESS* statistic was about 1.10, which corresponded to an $R_p^2$ of .835.

The organic carbon concentration of 32.6 mg/L on September 20, 2006, was discarded as an outlier. An additional sample collected May 24, 2007, was ruined, leaving 36 of the original 38 samples to construct the regression model. The studentized residual from the questionable sample of 32.6 mg/L was about 13 when the best-fit model was developed. A Bonferroni adjusted *t*-test (*p*-value <.001) on the residuals rejected the September 20, 2006, concentration as an extreme outlier. The concentration of 32.6 mg/L is almost twice that of the next largest concentration of 18.2 mg/L on January 18, 2007, which had corresponding streamflow and turbidity values of 5,160 cubic feet per second and 81.7 FNU, respectively. Streamflow and turbidity for the organic carbon concentration of 32.6 mg/L are 17 cubic feet per second and 11.3 FNU, respectively, both below the first quartile in their respective classes. A scatter plot of organic carbon concentration relative to both streamflow and turbidity indicates the concentration of 32.6 mg/L plots differently than all other values; this value did not agree with the positive relation between streamflow and turbidity. With additional regression diagnostics, there is sufficient evidence to justify the exclusion of the concentration.

### *Escherichia Coli*

Logarithmically transformed streamflow and turbidity were the explanatory variables included in the best-fit regression model for estimating *E. coli*. A summary of the regression analysis is shown in figure 12.

Measured relative to estimated and residual graphs in figure 12 provide adequate, visual evidence for the adherence of the model to requisite normal distribution assumptions. Four values are labeled in both graphs because they show substantial departure from the rest of the data. The Shapiro-Wilk test did not provide sufficient evidence to reject the assumption of normally distributed residuals; approximate *p*-value was .799. All residual and marginal-model plots indicate the best-fit regression model is appropriate for the measured data.

The adjusted R-squared value for the best-fit model was .607—although the model did not explain as much variability, this adjusted R-squared value by itself should not be interpreted as evidence of a poor model. Other regression model diagnostics and statistical tests for determining strength and relevance of models must be considered. In log-transformed space the residual standard error for this model is 0.887 MPN/100 mL; the median RPD was 10.0 percent. As additional measurements are made and incorporated into an updated model, uncertainty will likely decrease, yielding an improved regression model for estimating *E. coli.*

Turbidity values ranged from 7.30 to 90.5 FNU, and the magnitude of the maximum turbidity value (sample 36) was more than three times the magnitude of the IQR determined from all measured values, indicating it was an extreme outlier. No model could efficiently describe the variation in measured *E. coli* concentrations when sample 36 collected on October 16, 2007, was included in the analysis. *E. coli* for sample 36 was 36,100 MPN/100 mL, more than six times as large as the next largest value of 5,475 MPN/100 mL measured on October 17, 2006. *E. coli* for sample 36 was about 138 times greater than the IQR of 261 MPN/100 mL. Regression analysis was performed on the rest of the data and the best-fit model was obtained. The best-fit model was fit to the data, including sample 36, where regression diagnostics identified it as a significant outlier. Additional leverage points were identified, but not discarded from the data. A median RPD of 9.78 percent was obtained for the model that included sample 36.

### Suspended Sediment

Streamflow and turbidity were the statistically significant explanatory variables for estimating suspended-sediment concentrations for the East Fork San Jacinto site. Logarithmic transformations proved suitable for both streamflow and the response, whereas turbidity required no transformation. A summary of the regression analysis is shown in figure 13.

The homoscedacity of residuals in the studentized residual plot (fig. 13) provides evidence of model adequacy. The graph of measured relative to estimated concentrations

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       Q                Turb
 Min.   :  11.0   Min.   : 7.30
 1st Qu.:  37.0   1st Qu.:12.18
 Median :  76.5   Median :16.90
 Mean   : 716.6   Mean   :28.27
 3rd Qu.: 172.5   3rd Qu.:39.50
 Max.   :7930.0   Max.   :90.50
```

### Summary of Regression Analysis for the Constituent of: Total Organic Carbon (OrgC)

```
SUMMARY STATISTICS FOR OrgC, IN MILLIGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.965   6.443   7.617   9.173  12.250  18.190

REGRESSION EQUATION
Call:
lm(formula = log(OrgC) ~ poly(log(Q), 2) + poly(log(Turb), 2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.25706 -0.09356 -0.01361  0.12333  0.34015

Coefficients:
                    Estimate Std. Error t-value Pr(>|t|)
(Intercept)          2.12342    0.02544  83.475  < 2e-16 ***
poly(log(Q), 2)1     2.20846    0.25132   8.787 6.39e-10 ***
poly(log(Q), 2)2    -0.66288    0.15533  -4.268 0.000173 ***
poly(log(Turb), 2)1  0.24825    0.24537   1.012 0.319486
poly(log(Turb), 2)2 -0.53624    0.16457  -3.258 0.002718 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.1526 on 31 degrees of freedom
Multiple R-Squared: 0.8915,     Adjusted R-squared: 0.8775
F-statistic: 63.66 on 4 and 31 DF,  p-value: 1.667e-14
```

### Nomenclature (all potential variables)

- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day *d* (days into year) divided by 365; and
- log($x$) is natural log of $x$.

Special note: the `poly()` function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The `poly()` function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression

### Residual Plot for Regression

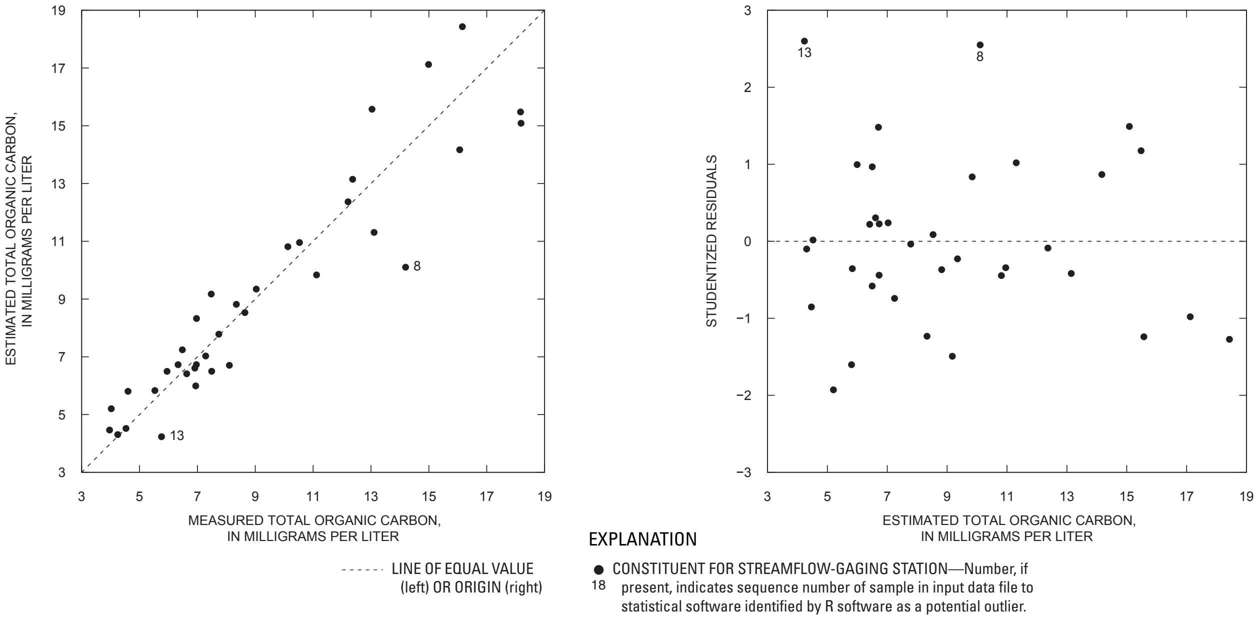


**Figure 11.** Summary of regression analysis for total organic carbon for 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–07.

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
      Q                Turb
 Min.   :  11.0   Min.   : 7.30
 1st Qu.:  37.0   1st Qu.:11.80
 Median :  73.0   Median :15.60
 Mean   : 691.6   Mean   :25.91
 3rd Qu.: 151.0   3rd Qu.:37.00
 Max.   :7930.0   Max.   :81.70
```

### Summary of Regression Analysis for the Constituent of: Escherichia coli (ECB)

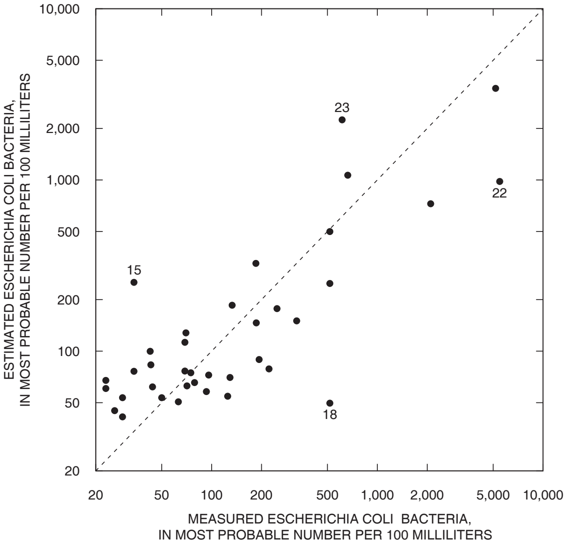
```
SUMMARY STATISTICS FOR ESCHERICHIA COLI (ECB), IN MOST-PROBABLE
    NUMBER (MPN) PER 100 MILLILITERS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   23.0    44.0    93.0   496.8   248.1  5475.0

REGRESSION EQUATION
Call:
lm(formula = log(ECB) ~ log(Q) + poly(log(Turb), 2))

Residuals:
       Min         1Q     Median         3Q        Max
-2.0047954 -0.5658744  0.0009804  0.4702329  2.3415816

Coefficients:
                    Estimate Std. Error t-value Pr(>|t|)
(Intercept)           3.5338     0.7457   4.739 3.97e-05 ***
log(Q)                0.2922     0.1577   1.853   0.0728 .
poly(log(Turb), 2)1   3.3014     1.5381   2.146   0.0393 *
poly(log(Turb), 2)2   2.4903     1.0203   2.441   0.0202 *
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.8872 on 33 degrees of freedom
Multiple R-squared: 0.6397,     Adjusted R-squared: 0.607
F-statistic: 19.53 on 3 and 33 DF,  p-value: 1.843e-07
```

### Nomenclature (all potential variables)

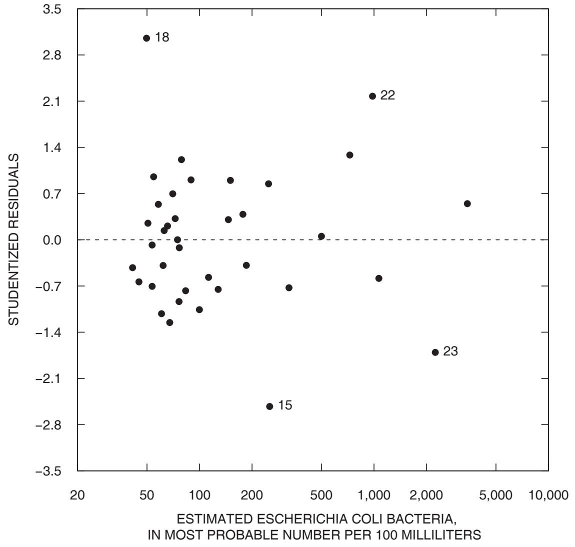
- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day *d* (days into year) divided by 365; and
- log($x$) is natural log of $x$.

Special note: the `poly()` function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The `poly()` function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

------ LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 12.** Summary of regression analysis for *Escherichia coli* for 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–07.

exhibits approximate linearity around the line of equal value (fig. 13). The suspended-sediment concentration on March 21, 2006 (sample 7), identified on the residual plot in figure 13, had the largest absolute studentized residual value, about 3.29. The corresponding Bonferroni adjusted *p*-value obtained using the Bonferroni outlier script in the R environment for statistical computing (R Development Core Team, 2006) is about .09. Such a small *p*-value associated with the Bonferroni test has important implications because this is a conservative test; two-sided tests based on bivariate normal variables (such as the Bonferroni test) are always conservative (Samuel-Cahn, 1996). The authors decided against discarding the March 21, 2006, suspended-sediment value as an outlier. The Shapiro-Wilk test statistic had an associated *p*-value of .36, indicating the requirement of residual normality was met.

Agreement between the LOWESS smooths for all marginal-model plots (not shown) indicates that this model provides an adequate fit to the data. A lack of collinearity was indicated by the *VIF*s, which were all less than 3.0. The model has an adjusted R-squared of .745, a residual standard error of 0.426 mg/L (in log-transformed space), and a median RPD of 6.69 percent. After adjusting the concentrations for seasonal and exogenous effects (ensuring each response variable is independent of all others), a Mann-Kendall test on the residuals did not find evidence of a statistically significant monotonic trend.

The suspended-sediment sample collected on June 13, 2006, was contaminated and could not be used for model construction. In addition, suspended-sediment samples collected on June 28, 2006 (sample 13) and December 18, 2007 (sample 38) were deemed unrepresentative and discarded from the model data, leaving 35 of 38 suspended-sediment concentrations to develop the regression equation. A thorough graphical comparison (not shown) combined with substantial statistical evidence justified excluding the June 28, 2006, and December 18, 2007, suspended-sediment samples. The June 28, 2006, suspended-sediment concentration was 109 mg/L, with corresponding streamflow and turbidity of 27 cubic feet per second and 15.6 FNU, respectively. This sample provided the sixth largest suspended-sediment concentration, whereas the associated streamflow value was less than the corresponding first quartile computed using all 38 streamflow samples. Although the suspended-sediment concentration was large for the June 28, 2006, sample, turbidity approximated the median turbidity of 15.85 FNU. Streamflow and turbidity were less than their respective medians, whereas the suspended-sediment concentration was considerably greater than the third-quartile concentration of 44.75 mg/L. The suspended-sediment concentration on December 18, 2007, was 695 mg/L, and the corresponding streamflow was 303 cubic feet per second, well below the mean streamflow of 685 cubic feet per second, yet still above the third quartile of 158.5 cubic feet per second. Positive skewness (few high values) associated with streamflow explains the large difference between these two values.

The quadratic and linear relation of streamflow and turbidity with suspended sediment was evident in a scatter plot (not shown) of all three transformed variables—suspended sediment, streamflow, and turbidity. The scatter plot also illustrates the approximate (and required) linear relation between streamflow and turbidity.

## Measured and Estimated Constituent Loads

Constituent estimates were obtained from selected best-fit regression equation models using real-time water-quality data for 2006 and 2007. Because the real-time water-quality data are collected at 15-minute intervals, a year with no missing values would have 35,040 data points. Missing 15-minute values were found randomly throughout the record and were simply ignored during the regression process. Instantaneous loads were obtained by multiplying the constituent value estimated from the regression model by the associated streamflow and a constant conversion factor.

Measured loads computed from the collected discrete water-quality data and loads estimated using regression models are shown for nitrite plus nitrate, total phosphorus, total organic carbon, atrazine (Spring Creek site only), and suspended sediment at each site (figs. 14–22). Graphical depictions of load estimates provide a visual reference for planning the collection of discrete water-quality samples, particularly with respect to periods of low and high flow.

Measured and estimated constituent loads generally match closely. For example, the measured and estimated nitrite plus nitrate loads from each site differ only slightly. Small, rapid fluctuations in nitrite plus nitrate loads occur within larger monthly fluctuations, especially at the Spring Creek site. The explanatory variables for the regression model for nitrite plus nitrate loads at the Spring Creek site are specific conductance and pH. pH follows a diurnal pattern because it is affected by photosynthetic processes and tends to increase during the daytime. Also, pH is negatively correlated with streamflow and turbidity (for the most part).

Estimates of loads for nitrite plus nitrate, total phosphorus, total organic carbon, atrazine, and suspended sediment computed from the real-time water-quality data are listed for the Spring Creek site (table 4) and the East Fork San Jacinto site (table 5). Load estimates include daily average by year (2006 and 2007), daily average for month by year (2006 and 2007), and total for each month and year. Two daily average estimated loads for each year are listed: the first was computed by averaging the estimated daily average loads for the months; the second was computed as the estimated total load for the year divided by 365. Generally, the first estimated daily load is less than the second because no adjustment for missing daily values was made when computing the monthly total. Real-time, continuously monitored data are collected every 15 minutes; sometimes a 15-minute interval or series of 15-minute intervals might be deleted because of fouling (if criteria described in Wagner and others [2006] are not met) or might be missing because of lost transmission. When the

### Inflow Statistics of Applicable Explanatory Variables: [Min., minimum, Qu., quartile, Max., maximum]

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
       Q                Turb
 Min.   :  14.0   Min.   : 7.30
 1st Qu.:  40.0   1st Qu.:11.78
 Median :  76.5   Median :15.65
 Mean   : 713.5   Mean   :27.95
 3rd Qu.: 153.5   3rd Qu.:37.55
 Max.   :7930.0   Max.   :90.50
```

### Summary of Regression Analysis for the Constituent of: Suspended Sediment (SS)

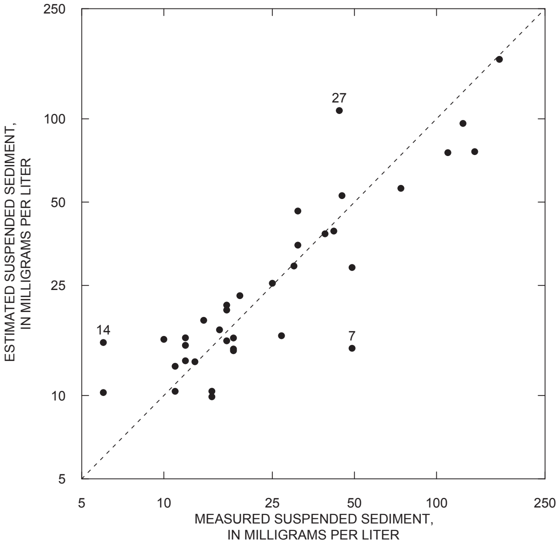
```
SUMMARY STATISTICS FOR SS, IN MILLIGRAMS PER LITER
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   13.75   18.00   36.28   42.50  170.00

REGRESSION EQUATION
Call:
lm(formula = log(SS) ~ poly(log(Q), 2) + Turb)

Residuals:
      Min        1Q    Median        3Q       Max
-0.951962 -0.197273 -0.001296  0.226476  1.196067

Coefficients:
                  Estimate Std. Error t-value Pr(>|t|)
(Intercept)       2.491914   0.161342  15.445  < 2e-16 ***
poly(log(Q), 2)1  0.907973   0.690301   1.315   0.1977
poly(log(Q), 2)2 -0.997305   0.437366  -2.280   0.0294 *
Turb              0.025235   0.005184   4.868 2.91e-05 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.426 on 32 degrees of freedom
Multiple R-Squared: 0.7669,     Adjusted R-squared: 0.7451
F-statistic:  35.1 on 3 and 32 DF,  p-value: 3.093e-10
```

### Nomenclature (all potential variables)

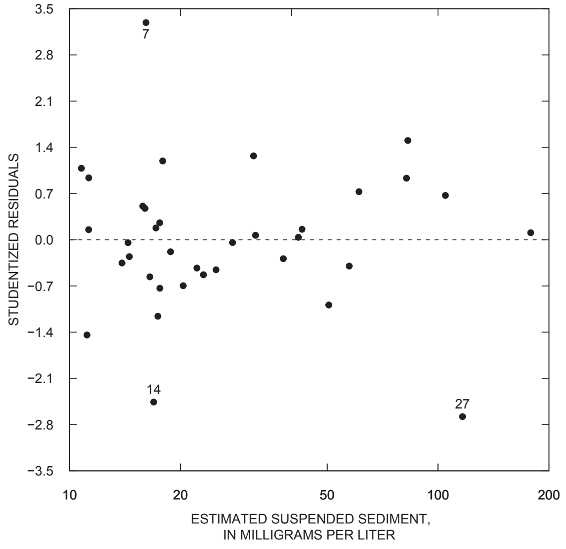
- $Q$ is streamflow, in cubic feet per second;
- $pH$ is pH, in standard units;
- *Rain* is binary: 1 if data collected within 24 hours of storm, otherwise 0;
- *SC* is specific conductance, in microsiemens per centimeter at 25°Celsius;
- *Turb* is turbidity in Formazine Nephelometric Units;
- *Temp* is water temperature, in °Celsius;
- *Date* is Julian day $d$ (days into year) divided by 365; and
- $\log(x)$ is natural log of $x$.

Special note: the poly() function, which is used in some regressions, creates orthogonal polynomials and not conventional polynomials of form $ax + bx^2$. The poly() function can be used to implement the regression [see documentation by R Development Core Team (2006)].

### Measured Relative to Estimated Plot for Regression



### Residual Plot for Regression



EXPLANATION

------ LINE OF EQUAL VALUE (left) OR ORIGIN (right)

● 18 CONSTITUENT FOR STREAMFLOW-GAGING STATION—Number, if present, indicates sequence number of sample in input data file to statistical software identified by R software as a potential outlier.

**Figure 13.** Summary of regression analysis for suspended sediment for 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–07.

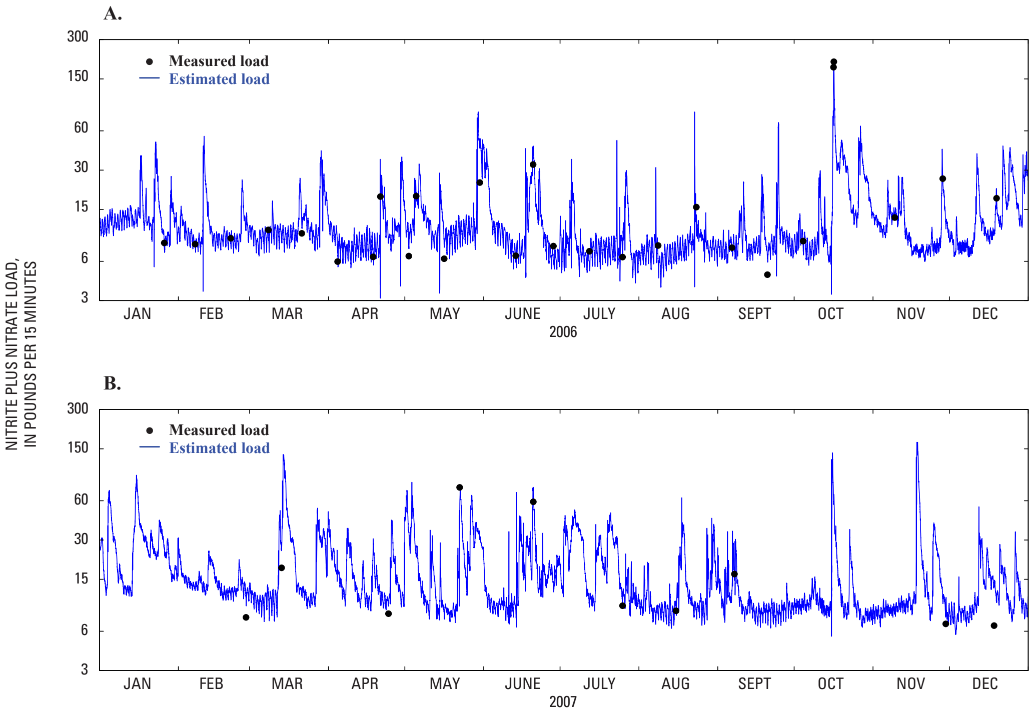


**Figure 14.** Measured and estimated nitrite plus nitrate nitrogen loads for 08068500 Spring Creek near Spring, Texas, (A) 2006 and (B) 2007.

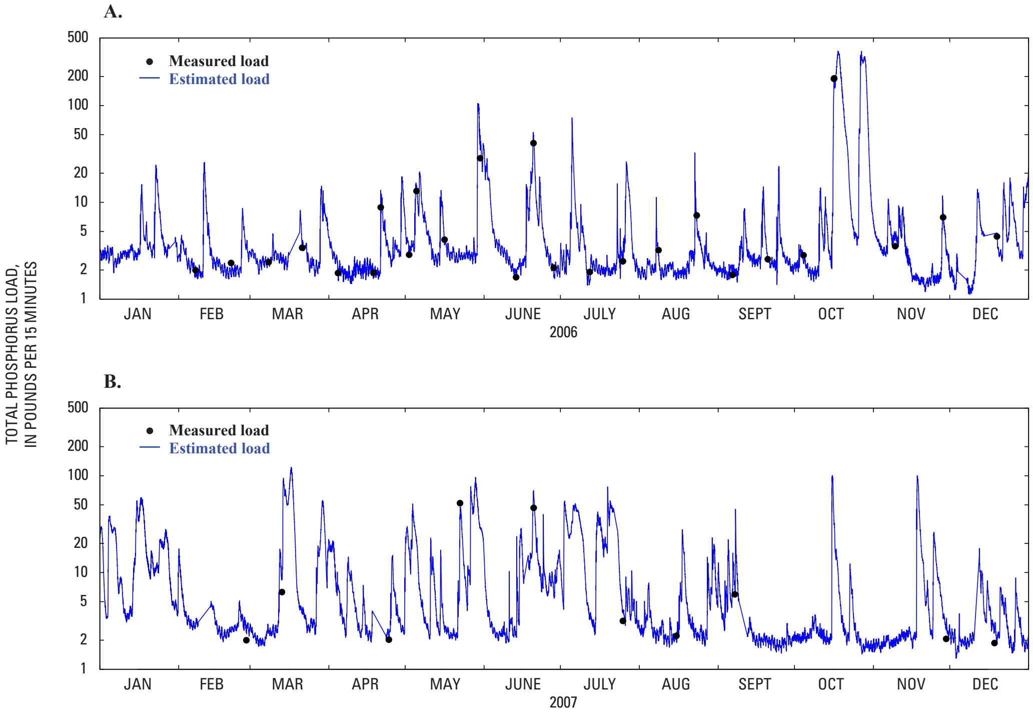


**Figure 15.** Measured and estimated total phosphorus loads for 08068500 Spring Creek near Spring, Texas, (A) 2006 and (B) 2007.

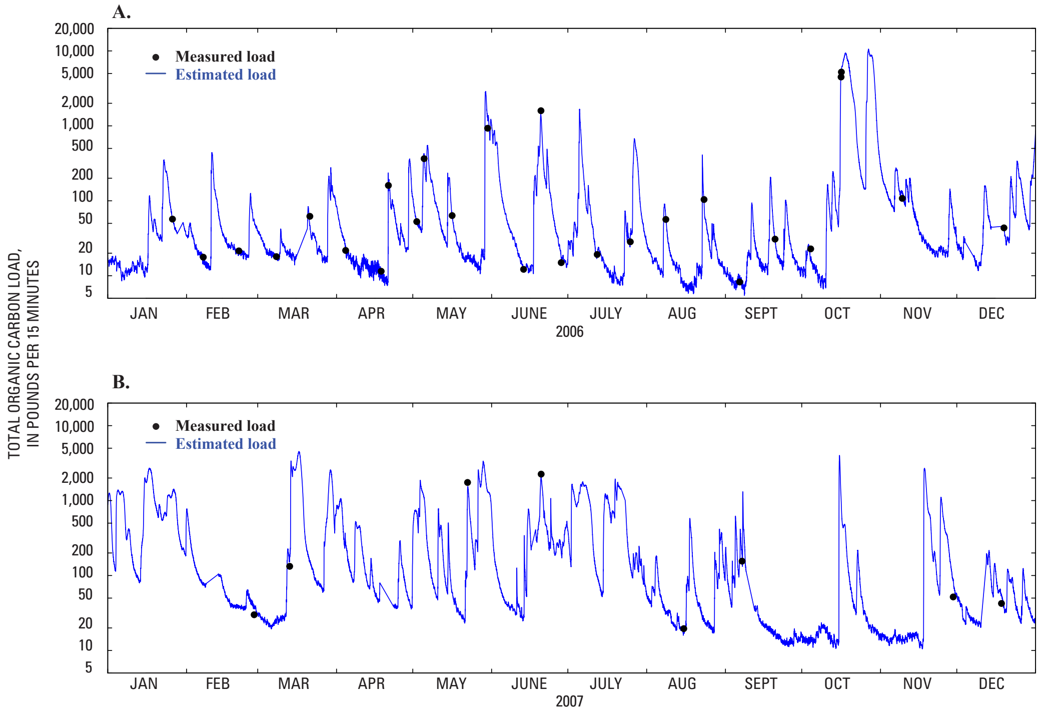


**Figure 16.** Measured and estimated total organic carbon loads for 08068500 Spring Creek near Spring, Texas, (A) 2006 and (B) 2007.

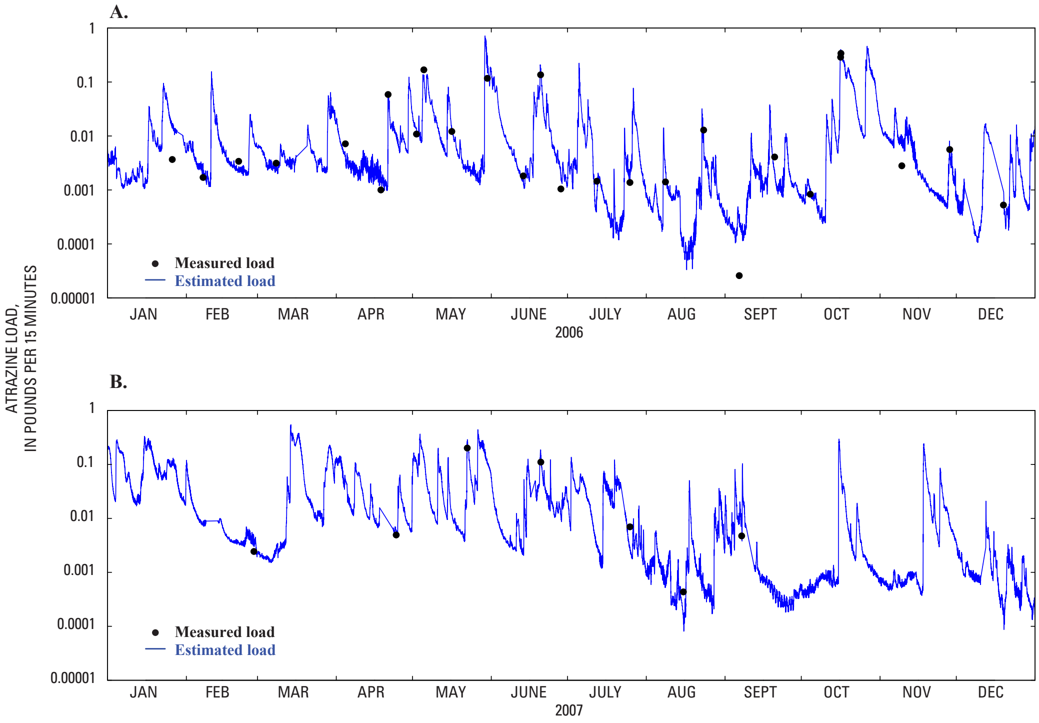


**Figure 17.** Measured and estimated atrazine loads for 08068500 Spring Creek near Spring, Texas, (A) 2006 and (B) 2007.

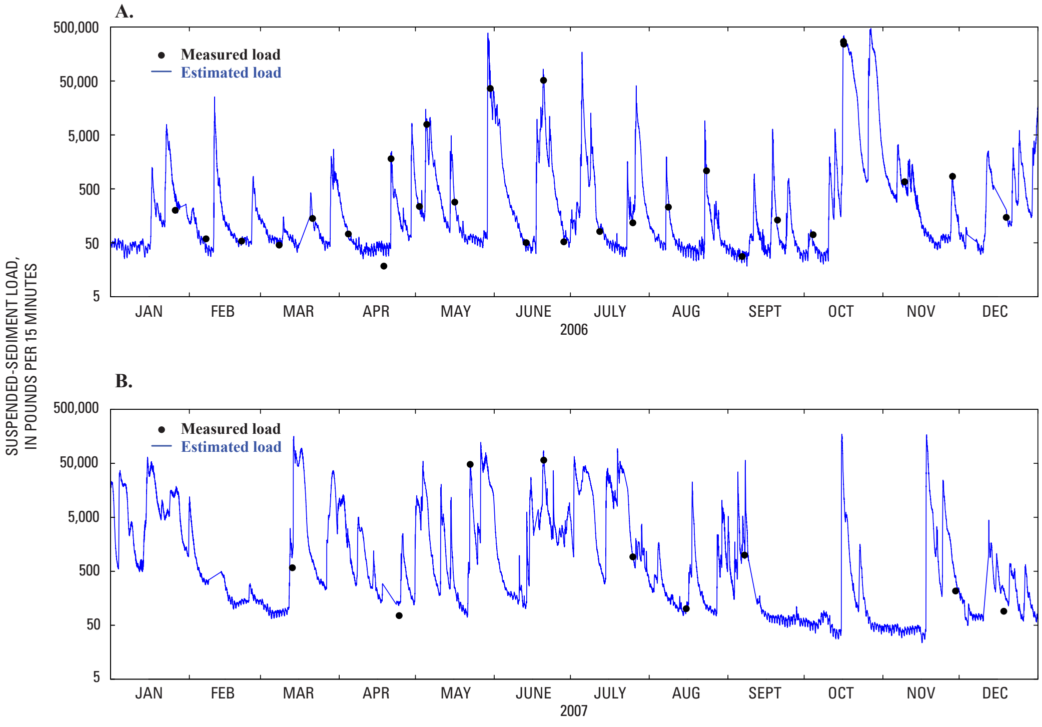


**Figure 18.** Measured and estimated suspended-sediment loads for 08068500 Spring Creek near Spring, Texas, (A) 2006 and (B) 2007.

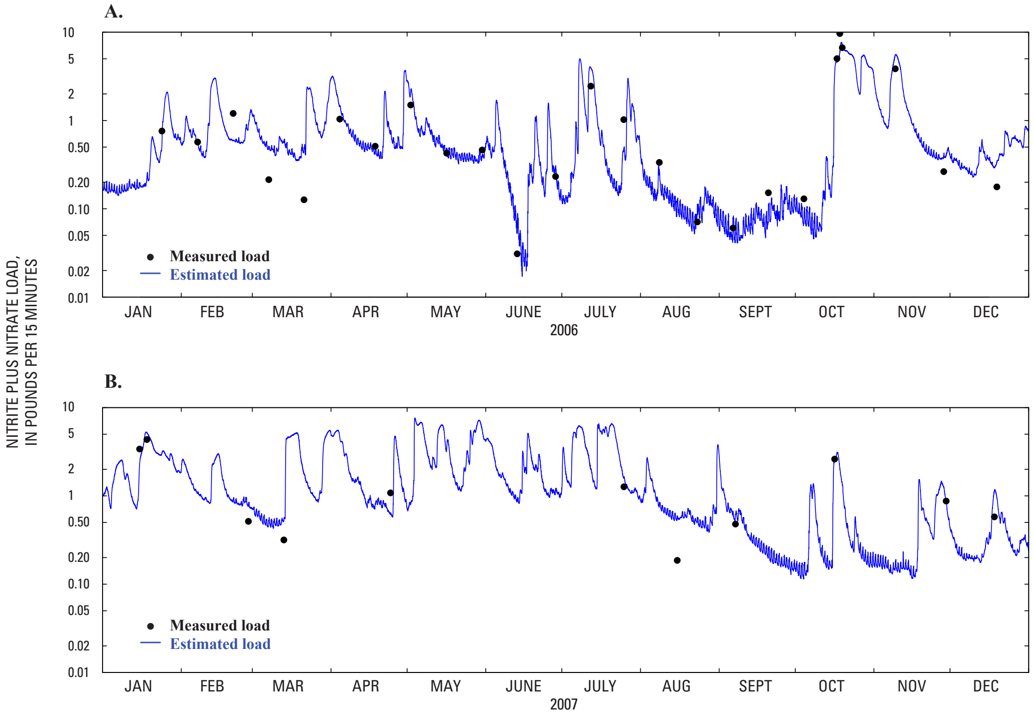


**Figure 19.** Measured and estimated nitrite plus nitrate nitrogen loads for 08070200 East Fork San Jacinto River near New Caney, Texas, (A) 2006 and (B) 2007.

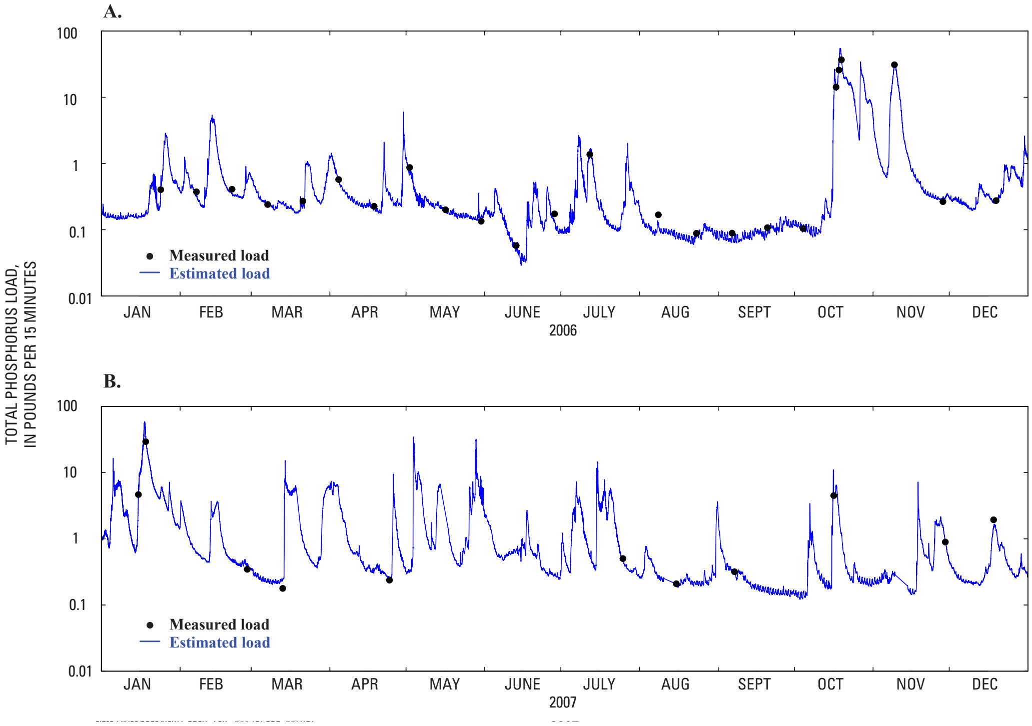


**Figure 20.** Measured and estimated total phosphorus loads for 08070200 East Fork San Jacinto River near New Caney, Texas, (A) 2006 and (B) 2007.

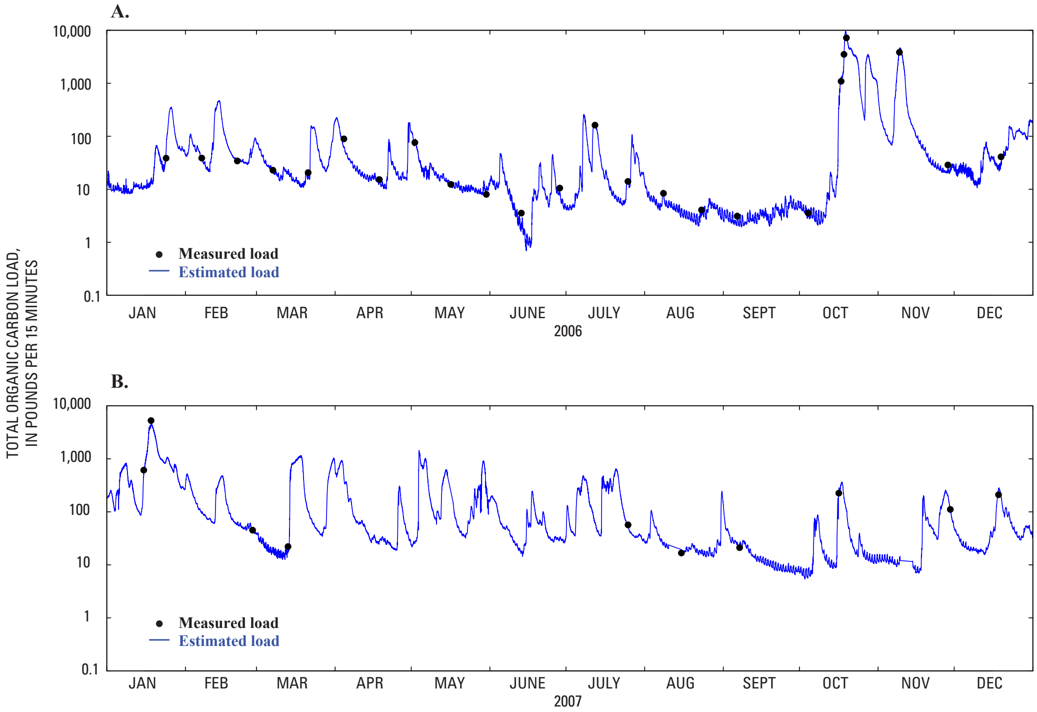


**Figure 21.** Measured and estimated total organic carbon loads for 08070200 East Fork San Jacinto River near New Caney, Texas, (A) 2006 and (B) 2007.

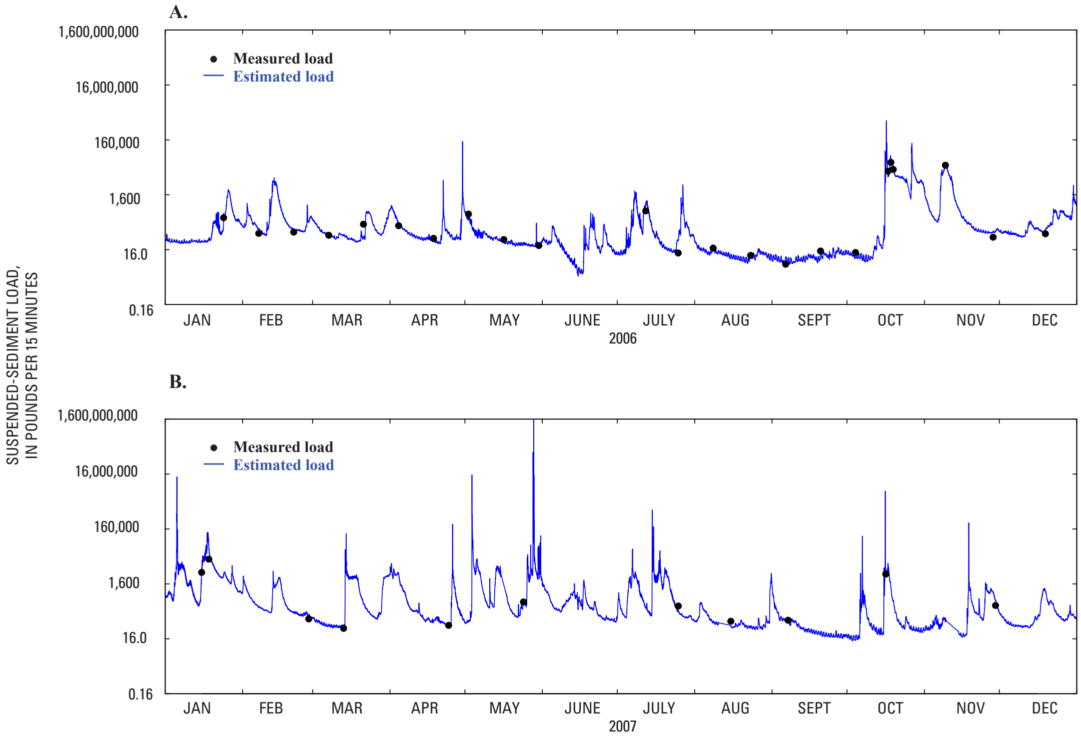


**Figure 22.** Measured and estimated suspended-sediment loads for 08070200 East Fork San Jacinto River near New Caney, Texas, (A) 2006 and (B) 2007.

continuous real-time data needed to compute the 15-minute load estimation were missing, no load value was computed for that interval. Interpolation of the missing real-time data was not done, and the daily load average was based on less than the 96 data points that typically are available for most days when data are collected for all 15-minute intervals. Days with missing continuous 15-minute interval data would have potentially low daily average load estimations.

**Table 4.** Estimated loads from regression models for 08068500 Spring Creek near Spring, Texas, 2006–07.

[In pounds]

| Month | 2006 Loads | | 2007 Loads | |
|---|---|---|---|---|
| | Estimated daily average for month | Total estimated for month | Estimated daily average for month | Total estimated for month |
| Nitrite plus nitrate nitrogen | | | | |
| January | 1,340 | 41,400 | 2,630 | 81,500 |
| February | 1,070 | 30,000 | 1,380 | 38,700 |
| March | 1,120 | 34,900 | 2,150 | 66,700 |
| April | 980 | 29,400 | 1,560 | 46,800 |
| May | 1,430 | 44,400 | 2,220 | 69,000 |
| June | 1,260 | 37,800 | 1,600 | 48,000 |
| July | 850 | 26,200 | 2,170 | 67,300 |
| August | 760 | 23,500 | 1,220 | 37,800 |
| September | 920 | 27,500 | 1,000 | 30,100 |
| October | 2,060 | 63,900 | 1,300 | 40,200 |
| November | 1,080 | 32,500 | 1,620 | 48,500 |
| December | 1,460 | 45,300 | 1,200 | 37,300 |
| **Total** | | 437,000 | | 612,000 |
| **Daily average** | 1,190 | 1,200 | 1,670 | 1,680 |
| Total phosphorus | | | | |
| January | 409 | 12,700 | 1,600 | 49,600 |
| February | 316 | 8,850 | 334 | 9,350 |
| March | 329 | 10,200 | 1,660 | 51,600 |
| April | 308 | 9,230 | 572 | 17,100 |
| May | 818 | 25,300 | 1,590 | 49,400 |
| June | 672 | 20,100 | 888 | 26,600 |
| July | 478 | 14,800 | 1,920 | 59,400 |
| August | 247 | 7,660 | 413 | 12,800 |
| September | 307 | 9,200 | 366 | 11,000 |
| October | 5,930 | 184,000 | 510 | 15,800 |
| November | 318 | 9,530 | 680 | 20,400 |
| December | 523 | 16,200 | 305 | 9,450 |
| **Total** | | 328,000 | | 333,000 |
| **Daily average** | 888 | 898 | 903 | 911 |
| Total organic carbon | | | | |
| January | 4,480 | 139,000 | 78,600 | 2,440,000 |
| February | 4,530 | 127,000 | 9,550 | 267,000 |
| March | 3,800 | 118,000 | 68,000 | 2,110,000 |
| April | 3,980 | 119,000 | 21,400 | 642,000 |
| May | 19,000 | 589,000 | 58,500 | 1,810,000 |
| June | 16,400 | 492,000 | 27,400 | 822,000 |
| July | 10,300 | 320,000 | 65,200 | 2,020,000 |
| August | 2,280 | 70,600 | 7,270 | 225,000 |
| September | 2,430 | 72,800 | 6,990 | 210,000 |
| October | 193,000 | 5,970,000 | 12,300 | 382,000 |
| November | 7,990 | 240,000 | 21,200 | 635,000 |
| December | 9,370 | 290,000 | 5,370 | 166,000 |
| **Total** | | 8,550,000 | | 11,700,000 |
| **Daily average** | 23,100 | 23,400 | 31,800 | 32,100 |

**Table 4.** Estimated loads from regression models for 08068500 Spring Creek near Spring, Texas, 2006–07—Continued.

| Month | 2006 Loads | | 2007 Loads | |
|---|---|---|---|---|
| | Estimated daily average for month | Total estimated for month | Estimated daily average for month | Total estimated for month |
| | Atrazine | | | |
| January | 0.993 | 30.8 | 9.13 | 283 |
| February | .815 | 22.8 | 1.00 | 28.1 |
| March | .705 | 21.8 | 6.80 | 211 |
| April | .868 | 26.0 | 3.21 | 96.3 |
| May | 3.74 | 116 | 7.62 | 236 |
| June | 2.15 | 64.5 | 2.23 | 66.8 |
| July | .777 | 24.1 | 2.22 | 68.9 |
| August | .133 | 4.11 | .289 | 8.94 |
| September | .206 | 6.19 | .427 | 12.8 |
| October | 6.36 | 197 | .901 | 27.9 |
| November | .548 | 16.4 | 1.25 | 37.6 |
| December | .310 | 9.62 | .121 | 3.76 |
| **Total** | | 539 | | 1,080 |
| **Daily average** | 1.47 | 1.48 | 2.93 | 2.96 |
| | Suspended sediment | | | |
| January | 39,000 | 1,210,000 | 1,070,000 | 33,200,000 |
| February | 39,400 | 1,100,000 | 58,600 | 1,640,000 |
| March | 17,300 | 537,000 | 1,190,000 | 36,900,000 |
| April | 24,600 | 737,000 | 167,000 | 5,000,000 |
| May | 646,000 | 20,000,000 | 1,040,000 | 32,400,000 |
| June | 313,000 | 9,390,000 | 461,000 | 13,800,000 |
| July | 224,000 | 6,960,000 | 1,380,000 | 42,600,000 |
| August | 16,300 | 506,000 | 75,300 | 2,330,000 |
| September | 15,000 | 449,000 | 95,400 | 2,860,000 |
| October | 5,060,000 | 157,000,000 | 277,000 | 8,590,000 |
| November | 42,300 | 1,270,000 | 442,000 | 13,200,000 |
| December | 94,500 | 2,930,000 | 25,600 | 792,000 |
| **Total** | | 202,000,000 | | 193,000,000 |
| **Daily average** | 545,000 | 554,000 | 523,000 | 530,000 |

**Table 5.** Estimated loads from regression models for 08070200 East Fork San Jacinto River near New Caney, Texas, 2006–07.

[In pounds]

| Month | 2006 Loads |  | 2007 Loads |  |
|---|---|---|---|---|
|  | Estimated daily average for month | Total estimated for month | Estimated daily average for month | Total estimated for month |
|  | Nitrite plus nitrate nitrogen |  |  |  |
| January | 42.3 | 1,310 | 226 | 7,010 |
| February | 91.3 | 2,560 | 130 | 3,630 |
| March | 79.7 | 2,470 | 191 | 5,930 |
| April | 96.2 | 2,890 | 189 | 5,680 |
| May | 65.5 | 2,030 | 358 | 11,100 |
| June | 36.3 | 1,090 | 185 | 5,550 |
| July | 103 | 3,190 | 321 | 9,960 |
| August | 13.0 | 403 | 81.3 | 2,520 |
| September | 7.81 | 234 | 41.9 | 1,260 |
| October | 241 | 7,480 | 56.1 | 1,740 |
| November | 136 | 4,070 | 41.5 | 1,240 |
| December | 41.7 | 1,290 | 33.4 | 1,040 |
| **Total** |  | 29,000 |  | 56,600 |
| **Daily average** | 79.5 | 79.5 | 155 | 155 |
|  | Total phosphorus |  |  |  |
| January | 38.7 | 1,200 | 595 | 18,400 |
| February | 80.6 | 2,260 | 99.9 | 2,800 |
| March | 34.6 | 1,070 | 166 | 5,130 |
| April | 43.9 | 1,320 | 142 | 4,250 |
| May | 25.2 | 780 | 324 | 10,000 |
| June | 15.3 | 460 | 68.1 | 2,040 |
| July | 41.2 | 1,280 | 171 | 5,300 |
| August | 9.05 | 281 | 35.6 | 1,100 |
| September | 8.91 | 267 | 26.0 | 780 |
| October | 818 | 25,400 | 70.6 | 2,190 |
| November | 329 | 9,880 | 63.5 | 1,910 |
| December | 43.9 | 1,360 | 40.7 | 1,260 |
| **Total** |  | 45,500 |  | 55,200 |
| **Daily average** | 124 | 125 | 150 | 151 |
|  | Total organic carbon |  |  |  |
| January | 4,430 | 137,000 | 81,500 | 2,530,000 |
| February | 8,420 | 236,000 | 14,400 | 402,000 |
| March | 3,870 | 120,000 | 24,900 | 772,000 |
| April | 4,110 | 123,000 | 14,200 | 425,000 |
| May | 2,020 | 62,600 | 25,400 | 789,000 |
| June | 1,010 | 30,300 | 6,060 | 182,000 |
| July | 3,650 | 113,000 | 17,200 | 534,000 |
| August | 452 | 14,000 | 3,130 | 97,100 |
| September | 332 | 9,960 | 1,750 | 52,500 |
| October | 131,000 | 4,070,000 | 4,300 | 133,000 |
| November | 43,200 | 1,300,000 | 5,520 | 166,000 |
| December | 5,890 | 183,000 | 5,060 | 157,000 |
| **Total** |  | 6,390,000 |  | 6,240,000 |
| **Daily average** | 17,400 | 17,500 | 17,000 | 17,100 |

**Table 5.** Estimated loads from regression models for 08070200 East Fork San Jacinto River near New Caney, Texas, 2006–07—Continued.

| Month | 2006 Loads | | 2007 Loads | |
|---|---|---|---|---|
| | Estimated daily average for month | Total estimated for month | Estimated daily average for month | Total estimated for month |
| | | Suspended sediment | | |
| January | 17,600 | 545,000 | 1,340,000 | 41,600,000 |
| February | 48,900 | 1,370,000 | 50,700 | 1,420,000 |
| March | 9,530 | 295,000 | 106,000 | 3,280,000 |
| April | 36,300 | 1,090,000 | 115,000 | 3,460,000 |
| May | 5,290 | 164,000 | 215,000,000 | 6,660,000,000 |
| June | 3,390 | 102,000 | 35,400 | 1,060,000 |
| July | 17,800 | 553,000 | 277,000 | 8,600,000 |
| August | 1,130 | 34,900 | 13,900 | 431,000 |
| September | 841 | 25,200 | 6,360 | 191,000 |
| October | 737,000 | 22,800,000 | 454,000 | 14,100,000 |
| November | 185,000 | 5,550,000 | 71,600 | 2,150,000 |
| December | 18,400 | 571,000 | 15,200 | 471,000 |
| **Total** | | 33,100,000 | | 6,740,000,000 |
| **Daily average** | 90,100 | 90,800 | 18,100,000 | 18,500,000 |

# Summary

In December 2005, the U.S. Geological Survey, in cooperation with the City of Houston, Texas, began collecting discrete water-quality samples for nutrients, total organic carbon, bacteria (*Escherichia coli* [*E. coli*] and total coliform), atrazine, and suspended sediment at two gaging stations upstream from Lake Houston near Houston. Data from discrete water-quality samples, in conjunction with monitored real-time data already being collected—physical properties (specific conductance, pH, water temperature, turbidity, and dissolved oxygen), streamflow, and rainfall—were used to develop regression models for predicting water-quality constituent concentrations for inflows to Lake Houston.

The regression equations presented in this report are site specific to streamflow-gaging stations on two tributaries to Lake Houston; however, the methods that were developed and documented could be applied to other tributaries to Lake Houston for estimating real-time water-quality data. The continuously monitored streamflow and water-quality properties, in conjunction with regression models using those data as surrogates for selected constituents (nitrite plus nitrate nitrogen, total phosphorus, total organic carbon, *E. coli*, atrazine, and suspended sediment) can be used to estimate concentrations for constituents that are lacking a continuous record. Used in conjunction with monitored real-time data, real-time modeled water-quality constituents will help water managers make critical near real-time decisions. Water managers will be able to assess the water quality of the tributaries of Lake Houston and identify effects on water quality in near real time. This information will help water managers make near real-time adjustments in drinking-water plant operations.

Streamflow, physical property, and water-quality constituent data were collected at two U.S. Geological Survey streamflow-gaging stations, 08068500 Spring Creek near Spring, Texas (Spring Creek site), and 08070200 East Fork San Jacinto River near New Caney, Texas (East Fork San Jacinto site). During 2005–07, discrete samples were collected at the Spring Creek site (39 samples) and at the East Fork San Jacinto site (38 samples). Hydrologic conditions within the Spring Creek and East Fork San Jacinto River watersheds vary and might affect chemical constituent concentrations, so discrete water-quality samples were collected over a wide range of streamflow conditions. Discrete water-quality samples for the first year (December 2005–November 2006) of this study were collected about every 2 weeks to observe seasonal patterns in water quality. Samples at these fixed-frequency sample times were collected as scheduled without regard to hydrologic condition, such as rising, falling, or stable streamflows. During storms or periods of high flow, unscheduled samples were also periodically collected during the first year of the study. Discrete water-quality samples for the second year (December 2006–December 2007) of the study were collected once a month. As in the first year of the study, stormwater-runoff samples for the second year were collected whenever possible.

Regression analyses were done using streamflow, continuous water-quality, and discrete water-quality data collected during 2005–07 at the Spring Creek and East Fork San Jacinto sites. Rainfall data obtained from a rain gage monitored by the Harris County Homeland Security and Emergency Management and colocated with the Spring Creek site were used in the regression analyses. The R environment for statistical computing was used to develop the regression models for estimating real-time concentrations for selected water-quality constituents. Multiple linear regression analyses were done using the leaps and bounds algorithm, an exhaustive, all-subset method for selecting the preferred model for each constituent. The leaps and bounds package uses a search technique to find the best subsets of possible models where "best" describes a model as having a minimum residual sum of squares for a given number of variables. Numerous possible regression models were evaluated to determine the best-fit relation between physical properties of the water and water-quality data. The potential explanatory or predictive variables included discharge (streamflow), specific conductance, pH, water temperature, turbidity, dissolved oxygen, rainfall, and time (to account for seasonal variations inherent in some water-quality data). The response variables at each site were nitrite plus nitrate nitrogen, total phosphorus, organic carbon, *E. coli*, atrazine, and suspended sediment. Because normally distributed response and explanatory variables with linear relations and constant variance were required for statistically valid multiple linear regression models, transformations were done when necessary to increase linearity among response and explanatory variables to improve normality and reduce heteroscedasticity.

The explanatory variables provide easily measured quantities as a means to estimate concentrations of the various constituents under investigation, with accompanying estimates of measurement uncertainty. Each regression equation can be used to estimate concentrations of a given constituent in real time, on the basis of explanatory variables also measured in real time. Corresponding 90-percent prediction intervals can be computed to display the uncertainty associated with the estimate.

The best-fit regression models at the Spring Creek and East Fork San Jacinto sites frequently had different statistically significant explanatory variables. The significant explanatory variables in the best-fit model for estimating total nitrite plus nitrate at the Spring Creek site were specific conductance and pH. In contrast, streamflow and a seasonal term were the explanatory variables for estimating the concentration of nitrite plus nitrate for the East Fork San Jacinto site. The statistically significant explanatory variables in the best-fit model for estimating total phosphorus at the Spring Creek site were specific conductance, water temperature, and turbidity—all logarithmically transformed using the Box-Cox procedure. At the East Fork San Jacinto site, streamflow, specific conductance, turbidity, and the periodic functions sine and cosine, with periods of 4π, to adjust for seasonal effects resulted in the best-fit model for estimating total phosphorus. Explanatory variables in the best-fit model for estimating organic carbon concentrations at the Spring Creek site were specific conductance, turbidity, and sine and cosine terms for seasonal fluctuations. Explanatory variables in the best-fit regression model for estimating organic carbon at the East Fork San Jacinto site were streamflow and turbidity. The statistically significant explanatory variables included in the best-fit regression model for estimating *E. coli* at the Spring Creek site were streamflow and rain. At the East Fork San Jacinto site, logarithmically transformed streamflow and turbidity were the explanatory variables included in the best-fit regression model for estimating *E. coli*. Although atrazine samples were collected at each site, a sufficient number of uncensored atrazine concentrations to construct a regression model were available only at the Spring Creek site. Streamflow, turbidity, and seasonal, periodic terms were the explanatory variables in the best-fit regression model for atrazine at the Spring Creek site. Statistically significant explanatory variables in the best-fit model for estimating suspended-sediment concentration at the Spring Creek site were streamflow, water temperature, and turbidity, whereas streamflow and turbidity were the explanatory variables in the best-fit model for the East Fork San Jacinto site.

Streamflow and turbidity were statistically significant estimators for most constituents at the Spring Creek and East Fork San Jacinto sites. Streamflow, turbidity, and all real-time monitored physical properties, except for dissolved oxygen, had statistical significance in at least one of the regression equations. Sine and cosine functions of time were used to explain seasonal variations in total organic carbon and atrazine at the Spring Creek site and total phosphorus at the East Fork San Jacinto site.

For each regression equation, the adjusted R-squared was evaluated as an indicator of the regression equation to explain variability in constituent concentrations. The adjusted R-squared for each best-fit regression equation at the Spring Creek site were .925 for nitrite plus nitrate, .882 for total phosphorus, .756 for total organic carbon, .812 for *E. coli*, .745 for atrazine, and .917 for suspended sediment. Also the median relative percent difference was computed for each equation. The median relative percent difference compares the measured concentrations to the concentrations estimated by the regression equations, so the smaller the relative percent difference the better the regression equation. The median relative percent difference for the Spring Creek site was 4.00 for nitrite plus nitrate, 4.45 for total phosphorus, 4.41 for total organic carbon, 9.78 for *E. coli*, 9.95 for atrazine, and 4.50 for suspended sediment. The adjusted R-squared for each equation at the East Fork San Jacinto site was .712 for nitrite plus nitrate, .719 for total phosphorus, .878 for total organic carbon, .607 for *E. coli*, and .745 for suspended sediment. The median relative percent difference for the East Fork San Jacinto site was 9.98 for nitrite plus nitrate, 8.00 for total phosphorus, 4.82 for total organic carbon, 10.0 for *E. coli*, and 6.69 for suspended sediment.

In conjunction with estimated concentrations, constituent loads were estimated by multiplying the estimated concentration by the corresponding streamflow and applying the appropriate conversion factor. By computing loads from estimated constituent concentrations, a continuous record of estimated loads can be available for comparison to total maximum daily loads.

## References

Aga, D.S., and Thurman, M.T., 1997, Environmental immunoassays—Alternative techniques for soil and water analysis [abs.]: American Chemical Society Symposium Series 657, p. 1–20.

American Public Health Association, American Water Works Association and Water Environment Federation, 2005, Standard methods for the examination of water and wastewater (21st ed.): Washington, D.C., American Public Health Association, p. 9–72 to 9–74.

American Society for Testing and Materials, 2003, D1889–00 Standard test method for turbidity of water, *in* Annual book of ASTM standards, water and environmental technology 2003, v. 11.01: West Conshohocken, Pa., American Society for Testing and Materials, 6 p.

Barlow, P.M., 1997, Particle-tracking analysis of contributing areas of public-supply wells in simple and complex flow systems, Cape Cod, Massachusetts: U.S. Geological Survey Water-Supply Paper 2434, 66 p.

Bartlett, M.S., 1936, The square root transformation in analysis of variance: Journal of the Royal Statistical Society, v. 3, no. 1, p. 68–78.

Box, G.E.P., and Cox, D.R., 1964, An analysis of transformations: Journal of the Royal Statistical Society, Service Bulletin 26, p. 211–246.

Bradu, D., and Mundlak, Y., 1970, Estimation in lognormal linear models: Journal of the American Statistical Association, v. 65, no. 329, p. 198–211.

Buchanan, T.J., and Somers, W.P., 1969, Discharge measurements at gaging stations: U.S. Geological Survey Techniques of Water-Resources Investigations, book 3, chapter A8, 65 p., accessed May 3, 2008, at *http://pubs.usgs.gov/twri/twri3a8/*.

Burnham, K.P., and Anderson, D.R., 2004, Understanding AIC and BIC in model selection: Sociological Methods and Research, v. 33, no. 2, p. 261–304.

Childress, C.J.O., Foreman, W.T., Connor, B.F., and Maloney, T.J., 1999, New reporting procedures based on long-term method detection levels and some considerations for interpretations of water-quality data provided by the U.S. Geological Survey National Water Quality Laboratory: U.S. Geological Survey Open-File Report 99–193, 19 p.

Christensen, V.G., Jian, Xiaodong, and Ziegler, A.C., 2000, Regression analysis and real-time water-quality monitoring to estimate constituent concentrations, loads and yields in the Little Arkansas River, south-central Kansas, 1995–99: U.S. Geological Survey Water-Resources Investigations Report 00–4126, 36 p.

Cohn, T.A., 2005, Estimating contaminant load in rivers—An application of adjusted maximum likelihood to type 1 censored data: Water Resources Research, v. 41, no. 8, p. 13.

Cook, R.D., and Weisberg, S., 1994, Transforming a response variable for linearity: Biometrika, v. 81, p. 731–737.

Cook, R.D., and Weisberg, S., 1997, Graphics for assessing the adequacy of regression models: Journal of the American Statistical Association, v. 92, p. 490–499.

Crawford, C., 1991, Estimation of suspended-sediment rating curves and mean suspended-sediment loads: Journal of Hydrology, v. 129, p. 331–348.

Dalgaard, P., 2008, Introductory statistics with R (2d ed.): New York, Springer Science and Business Media, 364 p.

Duan, N., 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, p. 605–610.

Faraway, J.J., 2005, Linear models with R: Boca Raton, Fla., Chapman and Hal, CRC press, 240 p.

Finney, D.J., 1941, On the distribution of a variate whose logarithm is normally distributed: Journal of the Royal Statistical Society Supplement, v. 7, p. 155–161.

Fishman, M.J., ed., 1993, Methods of analysis by the U.S. Geological Survey National Water Quality Laboratory—Determination of inorganic and organic constituents in water and fluvial sediments: U.S. Geological Survey Open-File Report 93–125, 217 p.

Fox, J., 2002, An R and S-Plus companion to applied regression: Thousand Oaks, Calif., Sage Publications, 294 p.

Furnival, G., and Wilson, R., 1974, Regression by leaps and bounds: Technometrics, v. 16, p. 499–511.

Good, P.I., 2005, Resampling methods—A practical guide to data analysis: Boston, Mass., Birkhäuser Books, ISBN 978–0-8176–4386–7, 218 p.

Guy, H.P., 1969, Laboratory theory and methods for sediment analysis: U.S. Geological Survey Techniques of Water-Resources Investigations, book 5, chapter C1, 58 p., accessed May 4, 2008 at *http://pubs.usgs.gov/twri/twri5c1/*.

Haining, R., 1990, The use of added variable plots in regression modeling with spatial date: The Professional Geographer, v. 42, no. 3, p. 336–344.

Harris County Homeland Security and Emergency Management, 2009, Harris County rain gage information: accessed August 19, 2009, at *http://www.hcoem.org/HCRainfall.aspx*.

Harris-Galveston Subsidence District, 1999 [amended 2001], District regulatory plan: accessed April 22, 2008, at *http://www.hgsubsidence.org/assets/pdfdocuments/HGRegPlan.pdf*.

Helsel, D.R., 2005, Nondetects and data analysis—Statistics for censored environmental data: New Jersey, Wiley, 250 p.

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources, in Hydrologic analysis and interpretation: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chapter A3, accessed June 2009 at *http://pubs.usgs.gov/twri/twri4a3*.

Hothorn, T., Zeileis, A., Millo, G., and Mitchell, D., 2009, Testing linear regression models: The Comprehensive R Archive Network, accessed July, 19, 2009 at *http://cran.r-project.org/web/packages/lmtest/lmtest.pdf*.

Hoyle, M.H., 1968, The estimation of variables after using a Gaussianating transformation: The Annals of Mathematical Statistics, v. 39, no. 4, p. 1,125–1,143.

Hurvich, C.M., and Tsai, C.H., 1989, Regression and time series model selection in small samples: Biometrika, v. 76, p. 297–307.

Iman, R.L., and Conover, W.J., 1983, Modern business statistics: New York, Wiley, ISBN 0471096687, 777 p.

Kasmarek, M.C., and Houston, N.A, 2008, Water-level altitudes 2008 and water-level changes in the Chicot, Evangeline, and Jasper aquifers and compaction 1973–2007 in the Chicot and Evangeline aquifers, Houston-Galveston region, Texas: U.S. Geological Survey Scientific Investigations Map 3031, 4 p., 17 sheets.

Kasmarek, M.C., and Strom, E.W., 2002, Hydrogeology and simulation of ground-water flow and land-surface subsidence in the Chicot and Evangeline aquifers, Houston area, Texas: U.S. Geological Survey Water-Resources Investigations Report 02–4022, 61 p.

Kennedy, E.J., 1983, Computation of continuous records of streamflow: U.S. Geological Survey Techniques of Water-Resources Investigations, book 3, chapter A13, 53 p., accessed May 3, 2008, at *http://pubs.usgs.gov/twri/twri3-a13/*.

Kennedy, E.J., 1984, Discharge ratings at gaging stations: U.S. Geological Survey Techniques of Water-Resources Investigations, book 3, chapter A10, 59 p., accessed May 3, 2008, at *http://pubs.usgs.gov/twri/twri3-a10/*.

Kleinbaum, D.G., and Kupper, L.K., 1978, Applied regression analysis and other multivariable methods: Boston, Mass., Duxbury Press, 131 p.

Li, K.C., and Duan, N., 1989, Regression analysis under link violation: Annuals of Statistics, v. 17, p. 1,009–1,052.

Likes, J., 1980, Variance of the MVUE for lognormal variance: Technometrics, v. 22, no. 2, p. 253–258.

Mathes, W.J., Sholar, C.J., and George, J.R., 1992, Quality-assurance plan for analysis of fluvial sediment: U.S. Geological Survey Open-File Report 91–467, 31 p., accessed May 4, 2008, at *http://pubs.er.usgs.gov/usgspubs/ofr/ofr91467*.

Mehran, F., 1973, Variance of the MVUE for the lognormal mean: Journal of the American Statistical Association, v. 68, no. 343, p. 726–727.

Møller, S.F., Frese, J.V., and Bro, R., 2006, Robust methods for multivariate data analysis: Journal of Chemometrics, v. 19, no. 10, p. 549–563, accessed July 17, 2009, at *http://dx.doi.org/10.1002/cem.962*.

Mueller, D.K., and Spahr, N.E., 2005, Water-quality, streamflow, and ancillary data for nutrients in stream and rivers across the nation, 1992–2001: U.S. Geological Survey Data Series 152, accessed July 15, 2009, at *http://pubs.usgs.gov/ds/2005/152/*.

Mueller, D.K., and Spahr, N.E., 2006, Nutrients in stream and rivers across the nation, 1992–2001: U.S. Geological Survey Scientific Investigations Report 2006–5107, 44 p., accessed October 28, 2009, at *http://pubs.usgs.gov/sir/2006/5107/*.

Multi-Resolution Land Characteristics Consortium, 2003, National land cover dataset 2001, zone 10: accessed October 17, 2008, at *http://www.mrlc.gov/multizone_download.php?zone=10*.

National Oceanic and Atmospheric Administration, 2008, Southeast Texas climate data: National Weather Service Forecast Office, Houston/Galveston, Texas, accessed April 30, 2008, at *http://www.srh.noaa.gov/hgx/climate.htm*.

Neyman, J., and Scott, E.L., 1960, Correction for bias introduced by a transformation of variables: The Annals of Mathematical Statistics, v. 31, p. 643–655.

Ott, R.L., and Longnecker, M., 2001, An introduction to statistical methods and data analysis (5th ed.): Pacific Grove, Calif., Duxbury, 1,152 p.

Patton, C.J., and Truitt, E.P., 2000, Methods of analysis by the U.S. Geological Survey National Water Quality Laboratory—Determination of ammonium plus organic nitrogen by a Kjeldahl digestion method and an automated photometric finish that includes digest cleanup by gas diffusion: U.S. Geological Survey Open-File Report 00–170, 31 p.

R Development Core Team, 2006, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, ISBN 3–900051–07–0, accessed at *http://www.R-project.org*.

Rousseeuw, P.J., and Leroy, A.M., 2003, Robust regression and outlier detection: John Wiley and Sons, ISBN 13–9780471488552, 356 p.

Ryberg, K.R., 2006, Continuous water-quality monitoring and regression analysis to estimate constituent concentrations and loads in the Red River of the North, Fargo, North Dakota, 2003–05: U.S. Geological Survey Scientific Investigations Report 2006–5241, 35 p.

Samuel-Cahn, E., 1996, Is the Simes improved Bonferroni procedure conservative?: Biometrika, v. 83, no. 4, p. 1.

Schwarz, G., 1978, Estimating the dimension of a model: Annals of Statistics, v. 6, no. 2, p. 461–464.

Sneck-Fahrer, D.A., Milburn, M.S., East, J.W., and Oden, J.H., 2005, Water-quality assessment of Lake Houston near Houston, Texas, 2000–2004: U.S. Geological Survey Scientific Investigations Report 2005–5241, 64 p.

Stine, R.A., 1995, Graphical interpretation of variance inflation factors: The American Statistician, v. 49, no. 1, p. 53–56.

Stoeckel, D.M., Bushon, R.N., Demcheck, D.K., Skrobialowski, S.C., Kephart, C.M., Bertke, E.E., Mailot, B.E., Mize, S.V., and Fendick, R.B., Jr., 2005, Bacteriological water quality in the Lake Pontchartrain Basin, Louisiana, following Hurricanes Katrina and Rita, September 2005: U.S. Geological Survey Data Series 143, 21 p.

Stuart, A., Ord, J.K., and Arnold, S., 1999, Kendall's advanced theory of statistics—Classical inference and the linear model, version 2A: Oxford, England, Oxford University Press.

Texas Commission on Environmental Quality, 2008, 2008 Texas water quality inventory and 303(d) list: accessed April 23, 2008, at *http://www.tceq.state.tx.us/compliance/monitoring/water/quality/data/08twqi/twqi08.html*.

Texas State Climatologist, 2008, Texas climatic bulletin: Office of the Texas State Climatologist, College of Geosciences, Department of Atmospheric Sciences, Texas A&M University, accessed April 30, 2008, at *http://www.met.tamu.edu/osc/TXclimat.htm*.

Texas State Data Center, 2007, 2006 Total population estimates for Texas metropolitan statistical areas: Office of the State Demographer, Texas Population Estimates and Projection Program, accessed April 22, 2008, at *http://txsdc.utsa.edu/tpepp/2006_txpopest_msa.php*.

U.S. Census Bureau, 2000, Census 2000 summary file 1 (SF 1) 100-percent data: accessed October 17, 2008, at *http://factfinder.census.gov/servlet/DCGeoSelectServlet?ds_name=DEC_2000_SF1_U*.

U.S. Environmental Protection Agency, 1993, Methods for the determination of inorganic substances in environmental samples: Cincinnati, Ohio, Environmental Monitoring Systems Laboratory, EPA/600/R–93/100, 79 p.

U.S. Geological Survey [variously dated], National field manual for the collection of water-quality data: U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chapters A1–A9, available online at *http://pubs.water.usgs.gov/twri9A*.

U.S. Geological Survey, 2009, National Water Information System (NWIS Web) data available on the World Wide Web: accessed May 2009 at *http://waterdata.usgs.gov/tx/nwis/*.

Velilla, S., 1993, A note on the multivariate Box-Cox transformation to normality: Statistics and Probability Letters, v. 17, no. 4, p. 259–263.

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., 8 attachments, accessed May 3, 2008, at *http://pubs.usgs.gov/tm/2006/tm1D3/*.

Weisberg, S., 2005, Applied linear regression (3d ed.): Hoboken, N.J., John Wiley and Sons, 310 p.

Wershaw, R.L., Fishman, M.J., Grabbe, R.R., and Lowe, L.E., eds., 1987, Methods for the determination of organic substances in water and fluvial sediments: U.S. Geological Survey Techniques of Water-Resources Investigations, book 5, chapter A3, 80 p.

Publishing support provided by
Lafayette Publishing Service Center

Information regarding water resources in Texas is available at
http://tx.usgs.gov/

Blank Page

Oden, Asquith, and Milburn—Regression Models to Estimate Real-Time Concentrations of Selected Constituents in Two Tributaries to Lake Houston near Houston, Texas, 2005–07—SIR 2009–5231

ISBN 978-1-4113-2628-6

9 781411 326286

Printed on recycled paper