Prepared in cooperation with the City of Houston

Regression Model Development and Computational Procedures to Support Estimation of Real-Time Concentrations and Loads of Selected Constituents in Two Tributaries to Lake Houston near Houston, Texas, 2005–9

Scientific Investigaitons Report 2012–5006

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

Regression Model Development and Computational Procedures to Support Estimation of Real-Time Concentrations and Loads of Selected Constituents in Two Tributaries to Lake Houston near Houston, Texas, 2005–9

By Michael T. Lee, William H. Asquith, and Timothy D. Oden

Prepared in cooperation with the City of Houston

Scientific Investigaitons Report 2012–5006

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

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

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

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

This and other USGS information products are available at http://store.usgs.gov/
U.S. Geological Survey
Box 25286, Denver Federal Center
Denver, CO 80225

To learn about the USGS and its information products visit http://www.usgs.gov/
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:
Lee, M.T., Asquith, W.H., and Oden, T.D., 2012, Regression model development and computational procedures to support estimation of real-time concentrations and loads of selected constituents in two tributaries to Lake Houston near Houston, Texas, 2005–9: U.S. Geological Survey Scientific Investigations Report 2012–5006, 40 p.

Contents

Figures

Tables

Conversion Factors, Datums, and Water-Quality Units

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)

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 are given in most-probable number per 100 milliliters (MPN/100 mL).

Turbidity is given in Formazine Nephelometric Units (FNU).

Regression Model Development and Computational Procedures to Support Estimation of Real-Time Concentrations and Loads of Selected Constituents in Two Tributaries to Lake Houston near Houston, Texas, 2005–9

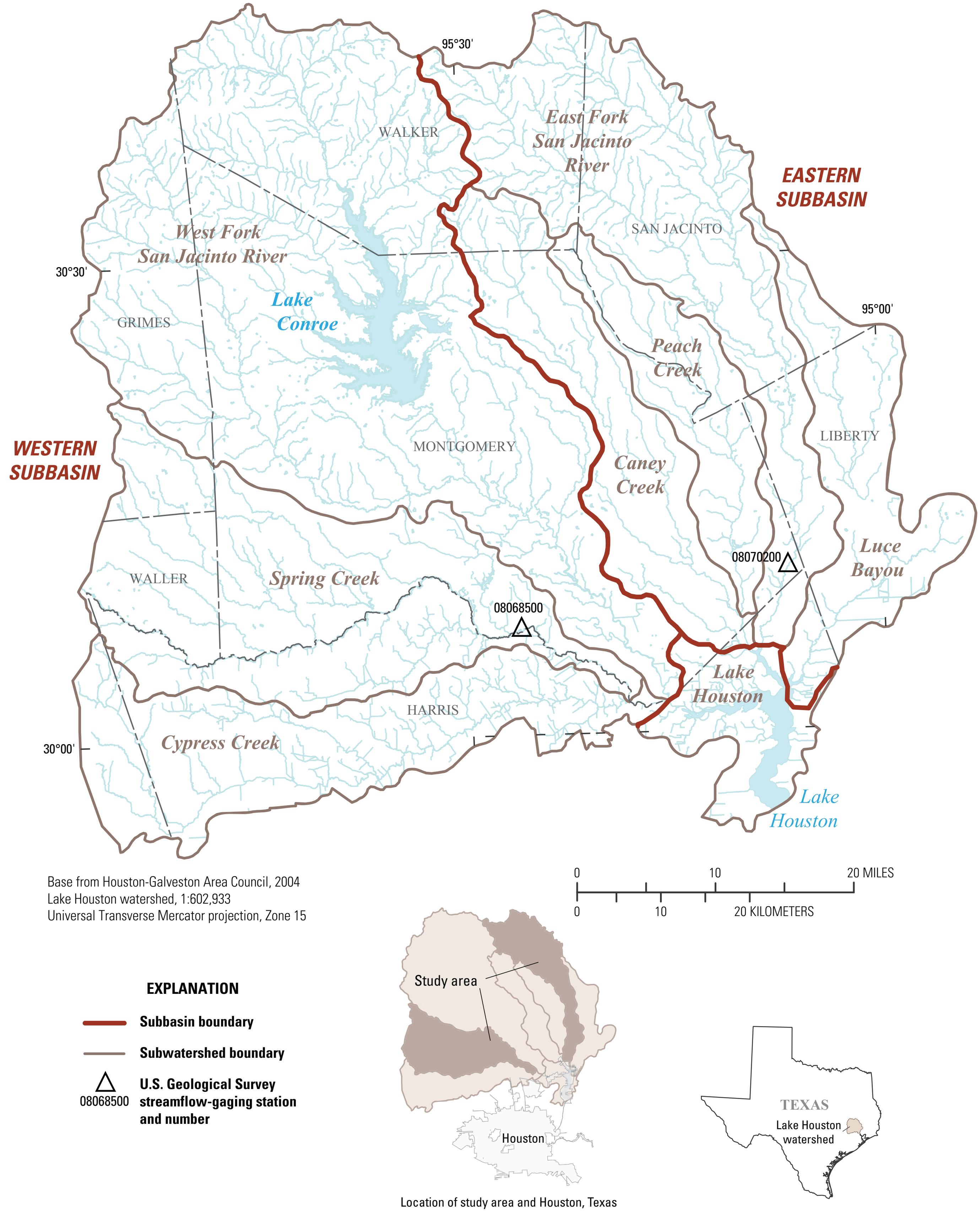
By Michael T. Lee, William H. Asquith, and Timothy D. Oden

Abstract

In December 2005, the U.S. Geological Survey (USGS), in cooperation with the City of Houston, Texas, began collecting discrete water-quality samples for nutrients, total organic carbon, bacteria (*Escherichia coli* and total coliform), atrazine, and suspended sediment at two USGS streamflow-gaging stations that represent watersheds contributing to Lake Houston (08068500 Spring Creek near Spring, Tex., and 08070200 East Fork San Jacinto River near New Caney, Tex.). Data from the discrete water-quality samples collected during 2005–9, in conjunction with continuously monitored real-time data that included streamflow and other physical water-quality properties (specific conductance, pH, water temperature, turbidity, and dissolved oxygen), were used to develop regression models for the estimation of concentrations of water-quality constituents of substantial source watersheds to Lake Houston. The potential explanatory variables included discharge (streamflow), specific conductance, pH, water temperature, turbidity, dissolved oxygen, and time (to account for seasonal variations inherent in some water-quality data). The response variables (the selected constituents) at each site were nitrite plus nitrate nitrogen, total phosphorus, total organic carbon, *E. coli*, atrazine, and suspended sediment. The explanatory variables provide easily measured quantities to serve as potential surrogate variables to estimate concentrations of the selected constituents through statistical regression. Statistical regression also facilitates accompanying estimates of uncertainty in the form of prediction intervals. Each regression model potentially can be used to estimate concentrations of a given constituent in real time. Among other regression diagnostics, the diagnostics used as indicators of general model reliability and reported herein include the adjusted R-squared, the residual standard error, residual plots, and p-values. Adjusted R-squared values for the Spring Creek models ranged from .582–.922 (dimensionless). The residual standard errors ranged from .073–.447 (base-10 logarithm). Adjusted R-squared values for the East Fork San Jacinto River models ranged from .253–.853 (dimensionless). The residual standard errors ranged from .076–.388 (base-10 logarithm). In conjunction with estimated concentrations, constituent loads can be estimated by multiplying the estimated concentration by the corresponding streamflow and by applying the appropriate conversion factor. The regression models presented in this report are site specific, that is, they are specific to the Spring Creek and East Fork San Jacinto River streamflow-gaging stations; however, the general methods that were developed and documented could be applied to most perennial streams for the purpose of estimating real-time water quality data.

Introduction

Houston, Texas, is the fourth largest city in the United States, with an estimated population of about 5.9 million people in 2009 (Texas State Data Center, 2011). Historically, groundwater has been the major source of supply for the City of Houston; however, development of groundwater resources has contributed to deleterious water-level declines and land-surface subsidence (Kasmarek and Strom, 2002; Kasmarek and others, 2010). Lake Houston is a surface-water-supply reservoir for the city of Houston and currently (2011) supplies between 10 and 20 percent of the total source-water supply (City of Houston, 2011). Furthermore, as a result of regulations restricting groundwater withdrawals for the purpose of mitigating or arresting 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 and planners. Two of the seven tributaries to Lake Houston, Spring Creek and East Fork San Jacinto River (fig. 1),



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

Location of study area and Houston, Texas

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, Tex.

representing approximately 31 percent of the drainage area for tributaries to Lake Houston, are the focus of this report.

In compliance with the Federal Clean Water Act, the Texas Commission on Environmental Quality compiles and maintains an inventory, commonly known as the “303(d) list,” of water bodies that are either impaired (do not meet applicable State water-quality standards) or threatened (are not expected to meet standards in the future) (Texas Commission on Environmental Quality, 2011). Lake Houston (segment 1002) first appeared in 2006 and again in 2008 on the inventory for bacteria. All of Spring Creek (segment 1008) has been listed on the 303(d) list for bacteria since 1996, and one portion of Spring Creek has been listed for depressed dissolved oxygen concentrations that are not conducive to healthy ecosystems since 1996. In addition, the East Fork San Jacinto River (segment 1003) first appeared on the 303(d) list in 2006 for bacteria and was still listed in the 2008 303(d) list.

The U.S. Geological Survey (USGS) and the City of Houston maintain a cooperative program to monitor water quality in Lake Houston and its contributing watersheds. Watershed water-quality monitoring began in December 2005 and is currently (2012) ongoing. 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 are collected 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-equation modeling using those data as surrogates for selected constituents (nitrite plus nitrate nitrogen, total phosphorus, total organic carbon, *Escherichia coli*, atrazine, and suspended sediment) can be used to estimate concentrations for constituents lacking continuous record. The estimated concentrations can be used to compute estimated constituent loads (a value proportional to the product of streamflow and concentration). Oden and others (2009) developed regression models to estimate constituent concentrations on Spring Creek and East Fork San Jacinto River based on data collected from 2005–7.

With near real-time water-quality data for the tributaries (every 15 minutes), water managers and planners will be able to identify potential effects of tributary inflows on the water quality of Lake Houston with sufficient alert time and adjust drinking-water plant operations accordingly. In addition, over time the results of tributary water-quality monitoring will contribute to the broader 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

This report documents updates that were made to previously published (Oden and others, 2009) regression models developed using data collected during 2005–7 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 updated using data collected during 2005–9. Real-time (every 15-minutes), continuously measured streamflow and water-quality properties (specific conductance, pH, water temperature, turbidity, and dissolved oxygen); discrete water-quality samples analyzed for nitrite plus nitrate, total phosphorus, total organic carbon, *E. coli*, atrazine, and suspended sediment; and time as an additional explanatory variable for seasonality were used in the models. The data were collected at two USGS streamflow-gaging stations, 08068500 Spring Creek near Spring, Tex. (hereinafter the Spring Creek site), and 08070200 East Fork San Jacinto River near New Caney, Tex. (hereinafter the East Fork San Jacinto River site). The regression models for each constituent are presented for each site. Lastly, examples of detailed computational analysis are provided to give the reader a step-by-step algorithmic or procedural guideline to independently use the regression models herein to calculate constituent concentrations, 90-percent prediction intervals, and instantaneous loads.

Results from these regression models can be used to better understand fluctuations of concentration and loads during changing seasons and flow conditions and to assess water-quality conditions relative to total maximum daily load goals and water-quality standards. The information also is useful for evaluating loading characteristics, such as range and variability, and for determining effectiveness of best management practices (Rasmussen and others, 2008).

While this report serves primarily as an update to Oden and others (2009), the 2 years of additional data used to update the models also extends the scientific and statistical understanding between continuously measured streamflow and water-quality properties and constituent concentrations measured in the laboratory. The regression models presented in this report are therefore considered to represent actual conditions more accurately than the earlier report and should serve as a replacement to previous models and the primary source for constituent concentration and load calculations.

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 less developed than the western subbasin. The western subbasin encompasses three tributary subwatersheds, and the eastern subbasin encompasses four tributary subwatersheds (table 1). The study area of this report consists of subwatersheds from the western

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.

and eastern subbasins— 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 account for about 41 percent of the 453 square miles of the subwatershed and the predominant land-use classification is forest (31 percent) (Multi-Resolution Land Characteristics Consortium, 2003).

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 about 18 percent of the 404 square miles of the subwatershed and, as in the Spring Creek subwatershed, the predominant land-use classification in the subwatershed is forest (47 percent) (Multi-Resolution Land Characteristics Consortium, 2003).

The climate in the study area is classified as humid subtropical (Texas State Climatologist, 2011), characterized by cool, temperate winters and long, hot summers with high humidity. During 2005–9, annual rainfall ranged from 41.2 to 65.5 inches at George Bush Intercontinental Airport, Houston, Tex. (National Oceanic and Atmospheric Administration, 2011).

Methods

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) and Turnipseed and Sauer (2010). Streamflow measurements during the course of the study (2005–9) were made about five times per year at the Spring Creek site and about five times per year at the East Fork San Jacinto River site. Stage, or gage height, was measured every 15 minutes by using submersible pressure transducers (or other conventional stage-measurement technology as needed) to the nearest 0.01 foot at the Spring Creek and East Fork San Jacinto River 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. Discharge measurements in this study were made to verify and modify a stage-discharge relation 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, 2011).

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 by using a multiparameter monitor. In November 2005, a multiparameter monitor 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 River site in November 2005. Each of the five sensors on the multiparameter monitors were calibrated as described in the USGS “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 River sites feature a swinging well design for monitoring real-time water-quality properties. Swinging wells respond to and swing in the direction of the force of the flowing water. The wells are constructed of schedule 80 polyvinyl-chloride pipe with holes in the lower 3 feet, allowing water to pass through wherever a multiparameter monitor is located. Each multiparameter monitor is positioned near the centroid of base flow in each stream in a swinging well. The data from each multiparameter monitor were electronically recorded and transmitted by satellite to a downlink site and then to the USGS Texas Water Science Center in Austin, Tex. 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 River sites at the present (2012) remain operational for these five physical water-quality properties.

Discrete Water-Quality Sample Collection, Analysis, and Results

Discrete water-quality samples were manually collected at each sampling site. For this analysis, 58 samples were collected at the Spring Creek site, and 51 samples were collected at the East Fork San Jacinto River site. Samples were analyzed for nutrients, total organic carbon, bacteria, atrazine, and suspended sediment. The actual number of results reported varied by constituent because of a variety of reasons, for example, broken sample bottles, lost samples, or samples were determined to fail internal quality-assurance checks and therefore reviewed and rejected.

Sample Design and Collection

Hydrologic conditions in the Spring Creek and East Fork San Jacinto River sites 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 facilitate detection of 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. During the second and third year of the study (December 2006–December 2008) discrete water-quality samples were collected approximately once a month at both the Spring Creek and East Fork San Jacinto River sites. During the fourth year of the study (December 2008–December 2009), an approximate monthly sampling schedule was maintained for the Spring Creek site, whereas samples collected at East Fork San Jacinto River site were reduced to a quarterly schedule. Instead of fixed frequency sampling during the second through fourth years of the study, sampling focused on stormwater-runoff whenever possible. Changes to the sampling design during the course of the study (in terms of timing or responding to specific hydrologic events) added considerable complexity to the effort of assessing temporal trends in the selected water-quality constituents.

Discrete water-quality samples were collected either by wading, when flow conditions permitted, or by sampling from bridges during higher flows. All samples were collected and processed as described in the USGS "National Field Manual for the Collection of Water-Quality Data" (U.S. Geological Survey, variously dated). Depth-integrated samples were collected, by 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 approximate 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., by 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 in Baton Rouge, La. (December 2005 – September 2007), or Louisville, Ky. (October 2007–9), by using procedures described in Guy (1969) and Mathes and others (1992). Atrazine samples were analyzed by the USGS Organic Geochemistry Research Laboratory, Lawrence, Kansas, by 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 Lab at Shenandoah, Tex., within the Gulf Coast Program Office of the USGS Texas Water Science Center by 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 River sites are stored in the USGS NWIS database and can be publicly accessed online (U.S. Geological Survey, 2011).

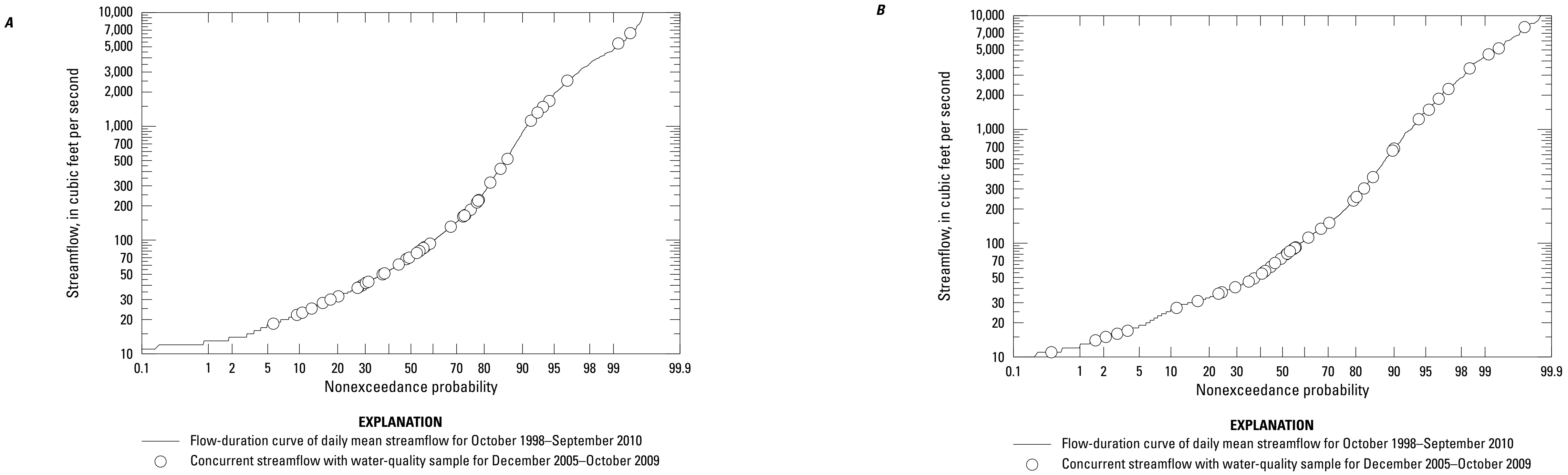


Figure 2. Flow duration curve and corresponding discrete water-quality samples, (*A*) Spring Creek near Spring, Texas, and (*B*) East Fork San Jacinto River near New Caney, Tex.

Table 2. Summary statistics for samples collected from two tributaries to Lake Houston near Houston, Texas, 2005–9.

[n, number of samples; <, less than1; E, estimated[1]; bold values indicate change in summary statistic from Oden and others (2009) report]

U.S. Geological Survey streamflow-gaging station name	U.S. Geological Survey stream-flow-gaging 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)	Ortho-phosphate, water, filtered (milligrams per liter as phosphorus)	Phosphorus, water, filtered (milligrams per liter)	Phosphorus, 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)	Organic carbon, water, unfiltered (milligrams per liter)	Suspended sediment (milligrams per liter)
Spring Creek near Spring, Tex.	08068500	Minimum	0.48	0.64	<0.01	0.11	0.005	0.06	0.08	0.15	20.00	2,420	<0.1	7.51	15
		Maximum	1.25	2.28	0.41	8.08	0.24	2.25	2.12	2.29	41,000	1,300,000	14	21.00	987
		Median	0.82	1.21	0.06	2.12	0.037	0.67	0.67	0.86	300	38,800	0.80	11.53	44.5
		Number of Samples	57	57	57	57	57	57	57	57	56	56	57	55	58
East Fork San Jacinto River near New Caney, Tex.	08070200	Minimum	0.14	0.23	<0.01	<0.026	<0.00107	<0.00432	0.013	0.054	13.20	1,000	<0.1	3.97	6
		Maximum	0.74	1.29	0.05	0.51	0.03	0.059	0.075	0.25	43,500	260,250	0.53	32.64	362
		Median	0.37	0.43	E0.02	0.10	0.003	0.02	0.034	0.098	108	9,200	<0.1	7.77	25
		Number of Samples	49	51	51	51	51	51	51	51	49	49	51	50	49

[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 (Childress and others, 1999).

Quality Control

Quality-control (QC) samples were collected as described in the USGS “National Field Manual for the Collection of Water-Quality Data” (variously dated) and analyzed by the same laboratories and methods as the environmental samples. QC samples included equipment blanks, field blanks, and split replicate samples. 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. Equipment blanks were collected about 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. Field blanks were collected and processed at sampling sites prior to the collection of environmental samples Split replicate samples (referred to as replicate samples in this report) were collected and 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 calculating the relative percent difference (RPD) for each constituent. RPD was computed by using the equation

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

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 greater compared to the laboratory reporting level. The RPD exceeded 10 percent for 8 of the 80 sample pairs of nutrients, 6 of 10 for total organic carbon, 3 of 11 for atrazine, and 6 of 11 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. Cases for which the analyte in both of the samples either was not detected or was detected at a concentration less than the LRL were defined as in agreement. Cases for which the analyte was detected at a concentration equal to or greater than the LRL in one of the samples and not detected or detected at a concentration less than the LRL were defined as being in nonagreement. This distinction was necessary because of a few instances in which an analyte was not detected or was detected at a concentration less than the LRL.

The 19 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 8 of 38 bacteriological sample pairs. Mean and median RPDs for all bacteriological samples, however, were 19 percent and 15 percent, respectively. The confidence intervals for the eight replicate samples with RPDs exceeding 30 percent overlapped, indicating there were no statistically significant differences between replicate samples. Additionally, seven of the eight replicate samples with RPDs that exceeded 30 percent originated from the Spring Creek site. Environmental sample and associated QC sample pair results are listed in appendix 1.

Regression Model Development

The R environment for statistical computing (R Development Core Team, 2010) was used to develop algebraically representable, multiple-linear regression equations (Faraway, 2005; Helsel and Hirsch, 2002; Helsel, 2005; Maindonald and Braun, 2003) to estimate concentrations for selected water-quality constituents and estimate prediction intervals or quantification of uncertainty. The authors used an open-source computational environment (the R environment for statistical computing) because the code syntax in this environment accommodates the syntax required for mathematical operations, and because the R environment is available free of charge to readers and is available for different computer platforms.

The regression models reported here are based on selected statistical relations between the constituent concentration acquired during discrete sampling and contemporaneous values of predictor variables that normally are measured continuously at a particular site (East Fork San Jacinto River or Spring Creek). Each equation provides estimates of the concentration of a single constituent as opposed to multivariate equations/methods that can provide for simultaneous estimation and inference of an ensemble of constituents; multivariate techniques are outside of the scope of this report. The relation between specific measurements and associated regression equations are site specific, and each equation is uniquely applicable for a particular site, but the methods could be applied to most perennial streams for the purpose of estimating real-time water quality data.

The potential predictor variables included streamflow, specific conductance, pH, water temperature, turbidity, dissolved oxygen, and various trigonometric functions of time (days into the year). Time was used when necessary to accommodate systematic seasonal variations of some constituents. The predictor variables are measured at the comparatively frequent sampling rate of every 15 minutes compared to the discrete sampling (intervals of every 2 weeks or longer) for the constituents. Lastly, the instantaneous daily load of a constituent can be estimated by multiplying the estimated constituent concentration (mass per volume) by contemporaneous streamflow (volume per time), accommodation of requisite unit-conversion factors, and applying an appropriate bias correction factor.

The application of multiple-linear regression equations to estimate constituent concentration and subsequent load estimation is well documented; for example, see Christensen and others (2000), Ryberg (2006), and Rasmussen and others (2009). Normally distributed response and explanatory variables with linear relations and constant variance are required for highly reliable equations. Logarithmic transformations on the response and explanatory variables are commonly used to improve linearity and to mitigate for nonnormality and heteroscedasticity (nonconstant variance about the regression line) in model residuals (Helsel and Hirsch, 2002). Consideration of nonlogarithmic transformations have been previously made (Oden and others, 2009) with a subset of the data considered in this report; however, additional analysis and implementation considerations resulted in a decision by the authors to exclusively use logarithmic transformation (base-10) trigonometric operations.

The coefficient of determination, R-squared, 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-squared statistic 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). For the current (2012) investigation, reader attention is drawn to adjusted R-squared values provided with each regression figure (figs. 3-14) because the adjusted R-squared can provide a more realistic evaluation of the model fit or ability of the model to characterize uncertainty in this study (Helsel and Hirsch, 2002).

The residual standard errors (RSE) of the regression equations reported here are exclusively reported in logarithmic units, which is consistent with the application of the logarithmic transformation of the response (water-quality constituent concentration). The authors use base-10 logarithms because these logarithms are conventionally most accessible to water-resources managers and the supporting engineering community.

After the transformation or transformations for the response and explanatory variables are selected, the analysis continues with the selection of explanatory variables that produce reliable regression models. The preferred regression model contains the fewest explanatory variables for which model diagnostics (including adjusted R-squared, the residual standard error, residual plots, and p-values) are acceptable. The p-value represents the probability (ranging from zero to one), that the statistical test result could have occurred if the null hypothesis (the hypothesis representing no change or no difference) was true (Helsel and Hirsch, 2002). When the p-value is less than a specified significance level (.05 in our application), the null hypothesis is rejected. Preferable models are those judged to have an acceptable balance between model fit and the number of variables in the model. Variables with small statistical significance and (or) substantial variance inflation potential are excluded (Helsel and Hirsch, 2002).

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 and requisite expansion of prediction intervals. A VIF represents the increase in variance because of correlation between predictive variables, whereas a minimum value of 1 occurs when no correlation is present. Typically, VIFs numerically 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. Assessment of VIFs was made and no values greater than 10 were present for the variables shown in the regression equations reported here.

Graphical analysis is a vital component of regression analysis and subsequent interpretation; it facilitates visual inspection and verification of data patterns such as linearity and constant variance underlying linear regression equation theory. The patterns seen in residual plots facilitate judgments in model reliability and are used to check if regression equations fit the observed data.

Retransformation Bias Correction

When a water-quality constituent is transformed (that is, into logarithmic units) as part of the building of a regression equation, the constituent must be retransformed to obtain an estimate in the original units. Estimates of constituent concentration that are unbiased in the transformed scale will be biased upon retransformation to the original scale. Retransformation bias corrections are made to mitigate or remove bias; the form of the bias correction factor will depend on the transformation used in the regression analysis (Duan,1983; Helsel and Hirsch, 2002).

Because logarithmic transformation was used for the current (2012) investigation, a bias correction is necessary because retransformation yields a median estimate of a constituent; median estimates tend to underestimate the actual arithmetic mean for the data considered here. Simply inverting a log-transformed response will return a biased low and therefore inconsistent estimate of the arithmetic mean. This bias greatly influences how loadings (say in units of tons per unit time) are computed. Extensive research has been done by others to find estimators that return the expected value (mean) of the streamflow load of a constituent if the response was log-transformed, for example Duan (1983), Crawford (1991), and Cohn (2005).

A minimum variance unbiased estimator (MVUE) was derived by Finney (1941). This estimator adjusts for bias and returns an efficient estimate of the mean. The Finney estimator is a commonly used and a reliable choice when the log-normal model is correct and the residual errors are normally distributed. However, the requirement that the residuals are normally distributed is an assumption that is difficult to achieve, assess, or interpret for the water-quality data considered in this report. Nonparametric estimators can provide a useful alternative to the retransformation methods. Duan (1983) derived a "smearing" estimator that requires only the residuals to be independent and homoscedastic (constant variance about the regression line). In the case of a log-transformation, the correction factor involves re-expressing the residuals in the original units and computing their mean. This factor for a given regression equation is to be multiplied to the regression equation in circumstances involving the computation of loads (concentration multiplied by streamflow along with necessary unit conversions).

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 (Childress and others, 1999). This kind of reporting results in what is referred to as left-censored observations (Helsel, 2005). Another complicating feature of water-quality data is that the censored values can vary depending on the laboratory reporting level (LRL) at the time the analyses were done. Censored values are those less than the laboratory reporting level applicable at the time the analyses were done. The mathematical theory is thoroughly described by Helsel (2005).

For this report, the foundational computational scripts were constructed by the authors to auto-adapt to the presence of left-censored data values, identify these values, and subsequently apply a maximum-likelihood estimator (MLE) for regression in lieu of conventional ordinary least-squares (OLS) regression. Both regression estimation techniques are provided by the R environment for statistical computing and standard packages of R; the two specific functions for OLS and MLE are *lm( )* (a linear regression modeling function) and *survreg( )* (a survival regression modeling function), respectively (R Development Core Team, 2010). The end result for the current (2012) investigation is that the basic algebraic implementation of the equations, whether produced by OLS or MLE regression, will be familiar to water-resources managers and the engineering community.

To conclude the broader discussion of censored data, an additional remark concerning judgment exercised by the authors is needed. Along with censored data, the laboratory might report estimated values; these are identified in the National Water Information System database with an "E" remark code. A constituent concentration is considered estimated by the laboratory when results are greater than the long-term method detection level (LT–MDL) and less than the LRL; that is, a detection is considered likely, but numerical quantification is considered questionable. For this investigation, all occurrences of "E" were dropped and the remaining numerical values used in the regression analysis. Additionally, individual samples collected prior to 2008 that were deemed contaminated or determined to be substantial outliers by Oden and others (2009) also were not used in the development of the models reported herein. To further clarify, for the current (2012) investigation, the data files used by Oden and others (2009) were extended by using data collected after 2008.

Censored Data at Spring Creek and East Fork San Jacinto River Site

The data for the Spring Creek and East Fork San Jacinto River sites used in this study are stored in the USGS NWIS database and can be publicly accessed online (U.S. Geological Survey, 2011). Only one of the atrazine concentrations measured in environmental samples collected at the Spring Creek site was less than the LRL of 0.10 mg/L. In contrast, most of the atrazine concentrations measured from East Fork San Jacinto River site were less than the LRL. Because of the large amount of censored atrazine data in discrete environmental samples collected at the East Fork San Jacinto River site, a defensible multilinear regression equation to estimate atrazine concentrations could not be developed. None of the nitrite plus nitrate concentrations measured in environmental samples collected at the Spring Creek site was less than the LRL of 0.06 mg/L, whereas four of the nitrite plus nitrate concentration measured in samples collected from the East Fork San Jacinto River site were less than the LRL.

Regression Analysis Summaries and Presentation of Equations

Summaries of the developed regression equations are provided in figures 3–14. Figure 3 shows abbreviations of terms used in figures 4-14. Each figure encapsulates an individual constituent per location and provides the significant explanatory variables used in the model, diagnostics factors used as indicators of general model reliability (including adjusted R-squared, the residual standard error, residual plots, and p-values), summary statistics for the explanatory variables and calculated constituent, the algebraic representation of the resultant model, and additional statistical parameters required for calculating prediction intervals, which were computed to display the uncertainty associated with the estimate (Helsel and Hirsch, 2002).

Regression equations for Spring Creek were developed for all constituents (table 2) analyzed for the study. A preliminary assessment (not reported here) of total ammonia plus organic nitrogen, dissolved ammonia plus organic nitrogen, ammonia nitrogen, nitrite nitrogen, orthophosphate phosphorus, dissolved phosphorus, and total coliform bacteria did not result in reliable equations because of large residual standard errors, and other unsatisfactory results from regression equation diagnostics. The regression equations developed for nitrite plus nitrate nitrogen, total phosphorus, total organic carbon, *E. coli* bacteria, atrazine, and suspended sediment are described in figures 4–9. Adjusted R-squared values for the Spring Creek models ranged from .582–.922 (dimensionless). The residual standard errors ranged from .073–.447 (base-10 logarithm).

Regression equations for East Fork San Jacinto River were developed for all constituents analyzed for the study (table 2), except atrazine. A preliminary assessment (not reported here) of total ammonia plus organic nitrogen, dissolved ammonia plus organic nitrogen, ammonia nitrogen, nitrite nitrogen, orthophosphate phosphorus, dissolved phosphorus, and total coliform bacteria did not result in reliable equations because of large residual standard errors, and other unsatisfactory results from regression equation diagnostics. Furthermore, an atrazine regression equation was not developed for the study because more than 50 percent of the data were below the LRL. The regression equations developed for nitrite plus nitrate nitrogen, total phosphorus, organic carbon, *E. coli* bacteria, and suspended sediment are described in figures 10–14. Adjusted R-squared values for the East Fork San Jacinto River models ranged from .253–.853(dimensionless). The residual standard errors ranged from .076–.388 (base-10 logarithm).

```
ABBREVIATIONS OF MATHEMATICAL FUNCTIONS AND STATISTICAL TERMS SHOWN IN
    FIGURES 4 THROUGH 14

Summary Statistics and Miscellaneous
Min.          Minimum
1st Qu.       First quartile
3rd Qu.       Third quartile
Max.          Maximum
log10()       Base-10 logarithm
cos2piD       The cosine of the cosine of the products 2 * pi * (Date)
sin2piD       The   sine of the   sine of the products 2 * pi * (Date)
cos4piD       The cosine of the cosine of the products 4 * pi * (Date)
sin4piD       The   sine of the   sine of the products 4 * pi * (Date)
pi            A mathematical constant approximately equal to 3.1415.
Date          Julian days into year divided by 365.25

Regression Model, linear model (ordinary least squares for uncensored data)
lm()          Linear regression modeling function
Std.Error     Standard error
t-value       T-statistic for the t-test
Pr(>|t|)      Probability of absolute value of t-value
Signif.codes  Default textual flags related to parameter significance
R-squared     Coefficient of determination
F-statistic   A statistic for the F-test
DF            Degrees of freedom
p-value       p-value (a standard computed probability)

Regression Model, survival model (for censored data)
survreg()     Survival regression modeling function
dist = "guassian"  The normal distribution is used for model estimation
Std.Error     Standard error
z             Standard normal z-statistic for survreg()
p-value       p-value (a standard computed probability)
Scale         Equivalent to residual standard error in transformed units
Loglik()      Log-likelihood (a statistic for method of maximum likelihood)
Chisq         A statistic for the Chi-squared distribution
```

Figure 3. Abbreviations of mathematical functions and statistical terms related to regression analysis of water-quality constituents shown in figures 4-14.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)        log10(SC)
 Min.   :1.230   Min.   :1.806
 1st Qu.:1.580   1st Qu.:2.413
 Median :1.914   Median :2.562
 Mean   :2.112   Mean   :2.521
 3rd Qu.:2.346   3rd Qu.:2.680
 Max.   :3.867   Max.   :2.847
```

Summary of Regression Analysis for the Constituent of: nitrite plus nitrate (NO_2 NO_3)

```
SUMMARY STATISTICS FOR NO2NO3, IN LOG10() milligrams per liter
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.9586  0.1239  0.3263  0.2678  0.6031  0.9074

REGRESSION EQUATION
lm(formula = log10(NO2NO3) ~ log10(Q) + log10(SC), data = the.data)

Residuals:
     Min    1st Qu.   Median   3rd Qu.      Max
-0.42257 -0.04910  0.02540  0.06092  0.32012

Coefficients:
            Estimate  Std.Error t-value Pr(>|t|)
(Intercept) -3.38330    0.59894  -5.649 6.20e-07 ***
log10(Q)    -0.10225    0.06142  -1.665    0.102
log10(SC)    1.53375    0.18900   8.115 6.44e-11 ***
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.1241 on 54 degrees of freedom
Multiple R-squared: 0.9245,     Adjusted R-squared: 0.9217
F-statistic: 330.7 on 2 and 54 DF,  p-value: < 2.2e-16

Variance inflation factors:
 log10(Q) log10(SC)
   7.0611    7.0611

Duan (1983) smearing factor: 1.037  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

Nomenclature (all potential variables)

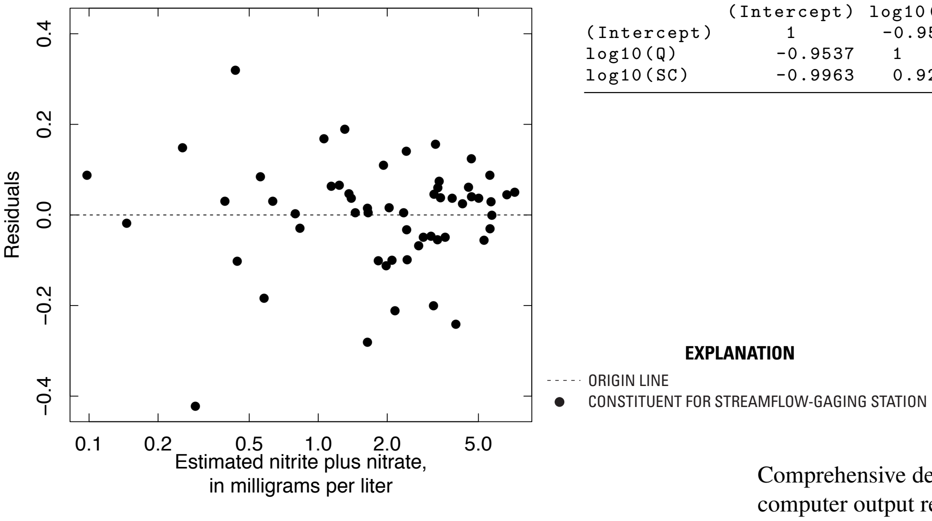
- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- SC is specific conductance, in microsiemens per centimeter at 25° Celsius;
- $Turb$ is turbidity in Formazine Nephelometric Units;
- $Temp$ is water temperature, in ° Celsius;
- $Date$ is Julian day DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\begin{aligned}\log_{10}(NO_2NO_3) = &-3.3833\\ &-0.1023\log_{10}(Q)\\ &+1.5338\log_{10}(SC)\end{aligned}$$

Residual Plot for Regression

```
CORRELATION OF COEFFICIENTS NO2NO3
             (Intercept) log10(Q) log10(SC)
(Intercept)      1        -0.9537   -0.9963
log10(Q)       -0.9537    1          0.9265
log10(SC)      -0.9963    0.9265    1
```



Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 4. Summary of regression analysis for nitrite plus nitrate nitrogen for U.S. Geological Survey streamflow-gaging station 08068500 Spring Creek near Spring, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
   log10(SC)       log10(Turb)        sin2piD             cos2piD
 Min.   :1.806   Min.   :0.9956   Min.   :-0.999761   Min.   :-0.9976
 1st Qu.:2.411   1st Qu.:1.2729   1st Qu.:-0.736086   1st Qu.:-0.6996
 Median :2.565   Median :1.5796   Median :-0.175941   Median :-0.2322
 Mean   :2.522   Mean   :1.6364   Mean   : 0.005503   Mean   :-0.1095
 3rd Qu.:2.684   3rd Qu.:1.9303   3rd Qu.: 0.793275   3rd Qu.: 0.4928
 Max.   :2.847   Max.   :2.4771   Max.   : 0.999091   Max.   : 0.9839
```

Summary of Regression Analysis for the Constituent of: total phosphorus (Phos)

```
SUMMARY STATISTICS FOR PHOS, IN LOG10() milligrams per liter
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.82390 -0.30550 -0.06299 -0.12350  0.05011  0.35980

REGRESSION EQUATION
lm(formula = log10(PHOS) ~ log10(SC) + log10(Turb) + sin2piD +
    cos2piD, data = the.data)

Residuals:
     Min   1st Qu.  Median   3rd Qu.     Max
-0.24310 -0.05750 -0.00554  0.05387  0.22659

Coefficients:
            Estimate  Std.Error t-value Pr(>|t|)
(Intercept) -3.58078    0.26045 -13.749  < 2e-16 ***
log10(SC)    1.25921    0.07907  15.926  < 2e-16 ***
log10(Turb)  0.16780    0.04699   3.571 0.000786 ***
sin2piD     -0.02472    0.01722  -1.436 0.157185
cos2piD     -0.06272    0.01973  -3.178 0.002519 **
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.09536 on 51 degrees of freedom
Multiple R-squared: 0.8838,     Adjusted R-squared: 0.8747
F-statistic: 96.96 on 4 and 51 DF,  p-value: < 2.2e-16

Variance inflation factors:
  log10(SC) log10(Turb)     sin2piD     cos2piD
     2.0915      2.1556      1.0197      1.0297

Duan (1983) smearing factor: 1.0222  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

Nomenclature (all potential variables)

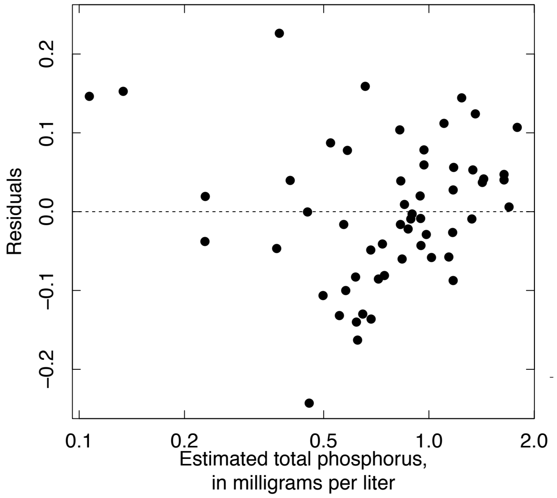
- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- SC is specific conductance, in microsiemens per centimeter at 25° Celsius;
- $Turb$ is turbidity in Formazine Nephelometric Units;
- $Temp$ is water temperature, in ° Celsius;
- $Date$ is Julian day DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\begin{aligned}\log_{10}(Phos) = &-3.5808\\ &+1.2592\log_{10}(SC)\\ &+0.1678\log_{10}(Turb)\\ &-0.0247\sin[2\pi(Date)]\\ &-0.0627\cos[2\pi(Date)]\end{aligned}$$

Residual Plot for Regression

```
CORRELATION OF COEFFICIENTS PHOS
            (Intercept) log10(SC) log10(Turb) sin2piD cos2piD
(Intercept)      1        -0.9777     -0.8447 -0.0966 -0.0567
log10(SC)       -0.9777    1           0.7190  0.0750  0.0324
log10(Turb)     -0.8447    0.7190      1       0.1310  0.1361
sin2piD         -0.0966    0.0750      0.1310  1      -0.0164
cos2piD         -0.0567    0.0324      0.1361 -0.0164  1
```



EXPLANATION

- - - - ORIGIN LINE

● CONSTITUENT FOR STREAMFLOW-GAGING STATION

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 5. Summary of regression analysis for total phosphorus for U.S. Geological Survey streamflow-gaging station 08068500 Spring Creek near Spring, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)       log10(Turb)
 Min.   :1.230   Min.   :0.9956
 1st Qu.:1.580   1st Qu.:1.2899
 Median :1.914   Median :1.5911
 Mean   :2.115   Mean   :1.6457
 3rd Qu.:2.349   3rd Qu.:1.9414
 Max.   :3.867   Max.   :2.4771
```

Summary of Regression Analysis for the Constituent of:
total organic carbon (OrgC)

```
SUMMARY STATISTICS FOR ORGC, IN LOG10() milligrams per liter
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8751  0.9800  1.0610  1.0640  1.1480  1.3220

REGRESSION EQUATION
lm(formula = log10(ORGC) ~ log10(Q) + log10(Turb), data = the.data)

Residuals:
     Min   1st Qu.  Median   3rd Qu.     Max
-0.11713 -0.05173 -0.01510  0.03307  0.18542

Coefficients:
            Estimate  Std.Error t-value Pr(>|t|)
(Intercept)  0.71382    0.04244  16.819  < 2e-16 ***
log10(Q)     0.02522    0.02233   1.130    0.264
log10(Turb)  0.18017    0.04054   4.444 4.65e-05 ***
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.07333 on 52 degrees of freedom
Multiple R-squared: 0.597,      Adjusted R-squared: 0.5815
F-statistic: 38.51 on 2 and 52 DF,  p-value: 5.479e-11

Variance inflation factors:
   log10(Q) log10(Turb)
     2.6598      2.6598

Duan (1983) smearing factor: 1.0142  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

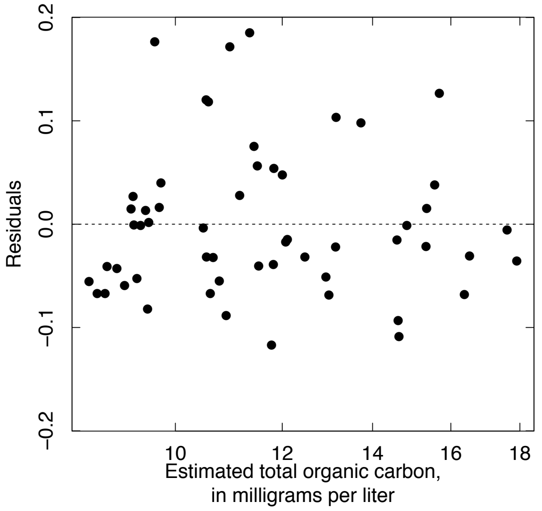
Nomenclature (all potential variables)

- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- 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 DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\begin{aligned}\log_{10}(OrgC) = {} & 0.7138 \\ & + 0.0252 \log_{10}(Q) \\ & + 0.1802 \log_{10}(Turb)\end{aligned}$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS ORGC
             (Intercept) log10(Q) log10(Turb)
(Intercept)       1        0.1291      -0.693
log10(Q)        0.1291   1             -0.790
log10(Turb)     -0.6930  -0.7900        1.000
```

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 6. Summary of regression analysis for total organic carbon for U.S. Geological Survey streamflow-gaging station 08068500 Spring Creek near Spring, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)        log10(Turb)
 Min.   :1.230   Min.   :0.9956
 1st Qu.:1.564   1st Qu.:1.2553
 Median :1.860   Median :1.5623
 Mean   :2.079   Mean   :1.6230
 3rd Qu.:2.283   3rd Qu.:1.9303
 Max.   :3.867   Max.   :2.4771
```

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

```
SUMMARY STATISTICS FOR ECB, IN LOG10() most probable number per 100
    milliliters
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.301   1.995   2.477   2.701   3.519   4.613

REGRESSION EQUATION
lm(formula = log10(ECB) ~ log10(Q) + log10(Turb), data = the.data)

Residuals:
     Min   1st Qu.  Median   3rd Qu.     Max
-0.73223 -0.33382 -0.06325  0.22911  1.21151

Coefficients:
            Estimate  Std.Error t-value Pr(>|t|)
(Intercept)  -0.1913     0.2512  -0.762 0.449533
log10(Q)      0.4969     0.1394   3.564 0.000782 ***
log10(Turb)   1.1458     0.2468   4.642 2.31e-05 ***
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.4468 on 53 degrees of freedom
Multiple R-squared: 0.7578,     Adjusted R-squared: 0.7487
F-statistic: 82.92 on 2 and 53 DF,  p-value: < 2.2e-16

Variance inflation factors:
   log10(Q) log10(Turb)
     2.7361      2.7361

Duan (1983) smearing factor: 1.8071  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

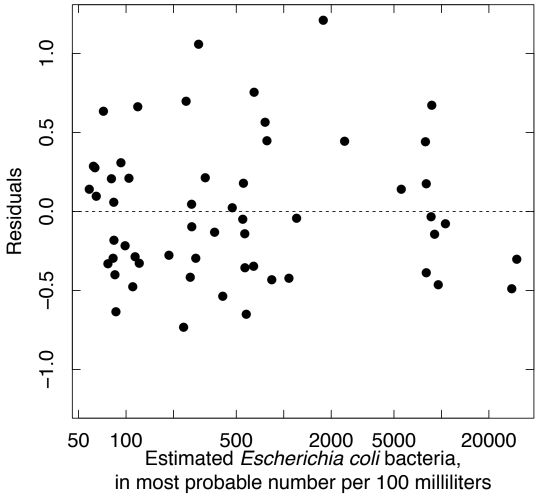
Nomenclature (all potential variables)

- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- 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 DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\begin{aligned}\log_{10}(ECB) = &-0.1913\\ &+0.4969\log_{10}(Q)\\ &+1.1458\log_{10}(Turb)\end{aligned}$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS ECB
             (Intercept) log10(Q) log10(Turb)
(Intercept)       1         0.1167     -0.6759
log10(Q)          0.1167    1          -0.7966
log10(Turb)      -0.6759   -0.7966      1
```

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 7. Summary of regression analysis for *Escherichia coli* for U.S. Geological Survey streamflow-gaging station 08068500 Spring Creek near Spring, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)        log10(Turb)        sin2piD            cos2piD           sin4piD            cos4piD
 Min.   :1.230   Min.   :0.9956   Min.   :-0.99976   Min.   :-0.9976   Min.   :-0.99901   Min.   :-0.99904
 1st Qu.:1.580   1st Qu.:1.2611   1st Qu.:-0.71865   1st Qu.:-0.7164   1st Qu.:-0.61393   1st Qu.:-0.71440
 Median :1.894   Median :1.5796   Median :-0.05209   Median :-0.2322   Median :-0.02074   Median :-0.14635
 Mean   :2.101   Mean   :1.6314   Mean   : 0.01822   Mean   :-0.1065   Mean   : 0.03735   Mean   :-0.08866
 3rd Qu.:2.342   3rd Qu.:1.9191   3rd Qu.: 0.77246   3rd Qu.: 0.5273   3rd Qu.: 0.74837   3rd Qu.: 0.57287
 Max.   :3.867   Max.   :2.4771   Max.   : 0.99909   Max.   : 0.9839   Max.   : 0.99948   Max.   : 0.99045
```

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

```
SUMMARY STATISTICS FOR ATZ, IN LOG10() milligrams per liter
The variable as "survival time"
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.00000 -0.28830 -0.09971 -0.03906  0.24180  1.14600

The variable as "survival status"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  1.0000  1.0000  0.9828  1.0000  1.0000

REGRESSION EQUATION
survreg(formula = log10(ATZ) ~ log10(Q) + log10(Turb) + sin2piD +
    cos2piD + sin4piD + cos4piD, data = the.data, dist = "gaussian")
              Value  Std.Error       z  p-value
(Intercept) -0.8096     0.1537  -5.266 1.40e-07
log10(Q)    -0.2253     0.0760  -2.966 3.02e-03
log10(Turb)  0.7473     0.1359   5.498 3.84e-08
sin2piD      0.4027     0.0451   8.932 4.20e-19
cos2piD     -0.0718     0.0489  -1.469 1.42e-01
sin4piD     -0.0105     0.0497  -0.212 8.32e-01
cos4piD     -0.2353     0.0486  -4.847 1.25e-06
Log(scale)  -1.4200     0.0935 -15.181 4.70e-52

Scale= 0.242

Gaussian distribution
Loglik(model)= 0   Loglik(intercept only)= -35.6
        Chisq= 71.35 on 6 degrees of freedom, p-value:  2.2e-13
Number of Newton-Raphson Iterations: 5
Sample size:  58

Variance inflation factors are not computable for a survival
    regression using the DAAG:::vif() function..
McFadden (1974) R-squared:  1.001 and adjusted R-squared:  0.8046
Bias correction factor: 1.0696  = 10^[(Scale * Scale)/2]
 (Detransformed estimates are to be multiplied this factor.)
```

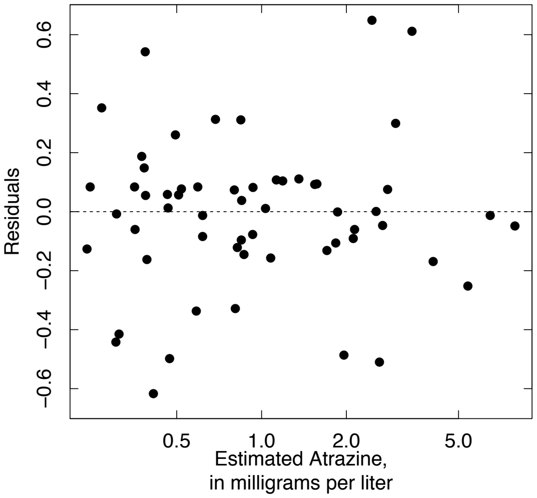
Nomenclature (all potential variables)

- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *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 *DOY* (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of *x*.

Algebraic Equation

$$\begin{aligned}\log_{10}(Atz) = &-0.8096 \\ &-0.2253\log_{10}(Q) \\ &+0.7473\log_{10}(Turb) \\ &+0.4027\sin[2\pi(Date)] \\ &-0.0718\cos[2\pi(Date)] \\ &-0.0105\sin[4\pi(Date)] \\ &-0.2353\cos[4\pi(Date)]\end{aligned}$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS ATZ
            (Intercept) log10(Q) log10(Turb) sin2piD cos2piD sin4piD cos4piD
(Intercept)      1        0.0223     -0.6761 -0.2079 -0.1187 -0.4198 -0.0427
log10(Q)        0.0223    1          -0.7376  0.0807 -0.0864  0.1763  0.1573
log10(Turb)    -0.6761   -0.7376      1       0.0791  0.1681  0.1588 -0.0654
sin2piD        -0.2079    0.0807      0.0791  1       0.0031  0.1904  0.0994
cos2piD        -0.1187   -0.0864      0.1681  0.0031  1       0.0582  0.1181
sin4piD        -0.4198    0.1763      0.1588  0.1904  0.0582  1       0.1052
cos4piD        -0.0427    0.1573     -0.0654  0.0994  0.1181  0.1052  1
```

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 8. Summary of regression analysis for atrazine for U.S. Geological Survey streamflow-gaging station 08068500 Spring Creek near Spring, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)       log10(Turb)
 Min.   :1.230   Min.   :0.9956
 1st Qu.:1.580   1st Qu.:1.2843
 Median :1.919   Median :1.5966
 Mean   :2.135   Mean   :1.6441
 3rd Qu.:2.351   3rd Qu.:1.9191
 Max.   :3.867   Max.   :2.4771
```

Summary of Regression Analysis for the Constituent of: suspended sediment (SS)

```
SUMMARY STATISTICS FOR SS, IN LOG10() milligrams per liter
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.176   1.535   1.648   1.867   2.128   2.994

REGRESSION EQUATION
lm(formula = log10(SS) ~ log10(Q) + log10(Turb), data = the.data)

Residuals:
     Min    1st Qu.  Median   3rd Qu.     Max
-0.41044  -0.14107 -0.00053  0.15352  0.29726

Coefficients:
            Estimate  Std.Error t-value Pr(>|t|)
(Intercept)  0.23531    0.10339   2.276   0.0268 *
log10(Q)     0.42817    0.05009   8.548 1.13e-11 ***
log10(Turb)  0.43664    0.09587   4.554 2.96e-05 ***
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.179 on 55 degrees of freedom
Multiple R-squared: 0.8755,     Adjusted R-squared: 0.871
F-statistic: 193.4 on 2 and 55 DF,  p-value: < 2.2e-16

Variance inflation factors:
   log10(Q) log10(Turb)
     2.5089      2.5089

Duan (1983) smearing factor: 1.0801  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

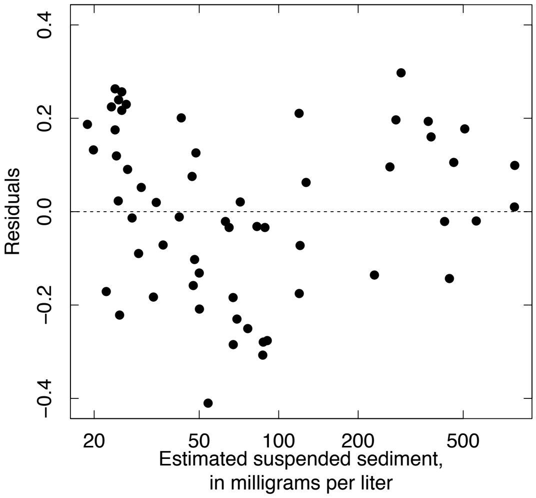
Nomenclature (all potential variables)

- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- 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 DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\log_{10}(SS) = 0.2353 + 0.4282\log_{10}(Q) + 0.4366\log_{10}(Turb)$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS SS
            (Intercept) log10(Q) log10(Turb)
(Intercept)      1        0.1478    -0.7223
log10(Q)         0.1478   1         -0.7755
log10(Turb)     -0.7223  -0.7755     1
```

EXPLANATION

- - - - ORIGIN LINE

● CONSTITUENT FOR STREAMFLOW-GAGING STATION

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 9. Summary of regression analysis for suspended sediment for U.S. Geological Survey streamflow-gaging station 08068500 Spring Creek near Spring, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)         sin2piD            cos2piD
 Min.   :1.041   Min.   :-0.99894   Min.   :-0.99777
 1st Qu.:1.568   1st Qu.:-0.71442   1st Qu.:-0.64457
 Median :1.903   Median : 0.11502   Median : 0.04604
 Mean   :2.086   Mean   : 0.03078   Mean   : 0.03567
 3rd Qu.:2.474   3rd Qu.: 0.77167   3rd Qu.: 0.65656
 Max.   :3.899   Max.   : 0.99998   Max.   : 0.99336
```

Summary of Regression Analysis for the Constituent of: nitrite plus nitrate (NO_2 NO_3)

```
SUMMARY STATISTICS FOR NO2NO3, IN LOG10() milligrams per liter
The variable as "survival time"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.5230 -1.2220 -1.0000 -0.9825 -0.7959 -0.2924

The variable as "survival status"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  1.0000  1.0000  0.9216  1.0000  1.0000

REGRESSION EQUATION
survreg(formula = log10(NO2NO3) ~ log10(Q) + sin2piD + cos2piD,
    data = the.data, dist = "gaussian")
              Value  Std.Error       z  p-value
(Intercept) -0.8938     0.1251  -7.144 9.05e-13
log10(Q)    -0.0288     0.0596  -0.483 6.29e-01
sin2piD      0.0289     0.0504   0.575 5.66e-01
cos2piD     -0.1911     0.0590  -3.241 1.19e-03
Log(scale)  -1.3788     0.1030 -13.390 6.94e-41

Scale= 0.252

Gaussian distribution
Loglik(model)= -3.7   Loglik(intercept only)= -10.3
        Chisq= 13.2 on 3 degrees of freedom, p-value:  0.0042
Number of Newton-Raphson Iterations: 4
Sample size:  51

Variance inflation factors are not computable for a survival
    regression using the DAAG:::vif() function..
McFadden (1974) R-squared:  0.6412 and adjusted R-squared:  0.2527
Bias correction factor: 1.0758  = 10^[(Scale * Scale)/2]
 (Detransformed estimates are to be multiplied this factor.)
```

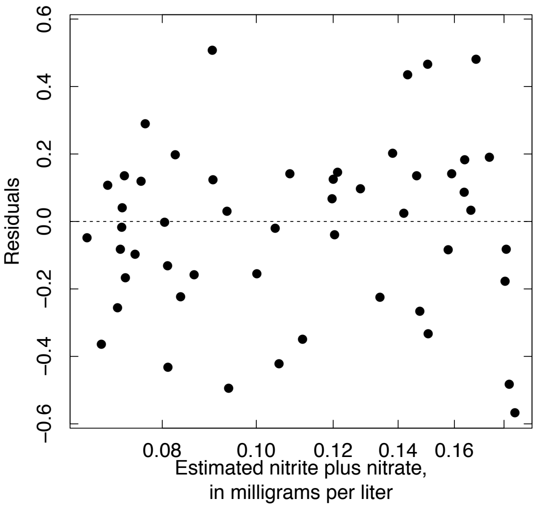
Nomenclature (all potential variables)

- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- 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 DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\begin{aligned}\log_{10}(NO_2NO_3) = &-0.8938\\ &-0.0288\log_{10}(Q)\\ &+0.0289\sin[2\pi(Date)]\\ &-0.1911\cos[2\pi(Date)]\end{aligned}$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS NO2NO3
            (Intercept) log10(Q) sin2piD cos2piD
(Intercept)      1       -0.9572 -0.0923  0.4401
log10(Q)        -0.9572   1       0.0703 -0.4678
sin2piD         -0.0923   0.0703  1      -0.0730
cos2piD          0.4401  -0.4678 -0.0730  1
```

EXPLANATION

----- ORIGIN LINE

● CONSTITUENT FOR STREAMFLOW-GAGING STATION

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 10. Summary of regression analysis for nitrite plus nitrate nitrogen for U.S. Geological Survey streamflow-gaging station 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)        log10(Turb)        sin2piD            cos2piD            sin4piD            cos4piD
 Min.   :1.041   Min.   :0.8633   Min.   :-0.99894   Min.   :-0.99777   Min.   :-0.99919   Min.   :-0.99993
 1st Qu.:1.568   1st Qu.:1.0792   1st Qu.:-0.71442   1st Qu.:-0.64457   1st Qu.:-0.56101   1st Qu.:-0.74933
 Median :1.903   Median :1.2041   Median : 0.11502   Median : 0.04604   Median :-0.08091   Median :-0.12020
 Mean   :2.086   Mean   :1.3546   Mean   : 0.03078   Mean   : 0.03567   Mean   : 0.04246   Mean   :-0.05625
 3rd Qu.:2.474   3rd Qu.:1.6127   3rd Qu.: 0.77167   3rd Qu.: 0.65656   3rd Qu.: 0.73417   3rd Qu.: 0.69467
 Max.   :3.899   Max.   :2.2304   Max.   : 0.99998   Max.   : 0.99336   Max.   : 0.99926   Max.   : 0.99108
```

Summary of Regression Analysis for the Constituent of: total phosphorus (Phos)

```
SUMMARY STATISTICS FOR PHOS, IN LOG10() milligrams per liter
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.2680 -1.1110 -1.0000 -1.0070 -0.9393 -0.5986

REGRESSION EQUATION
lm(formula = log10(PHOS) ~ log10(Q) + log10(Turb) + sin2piD +
    cos2piD + sin4piD + cos4piD, data = the.data)

Residuals:
      Min    1st Qu.   Median    3rd Qu.      Max
-0.207627 -0.060346 -0.001328  0.055656  0.236860

Coefficients:
             Estimate  Std.Error t-value Pr(>|t|)
(Intercept) -1.458421   0.050022 -29.156  < 2e-16 ***
log10(Q)    -0.081051   0.029817  -2.718  0.00935 **
log10(Turb)  0.457366   0.056263   8.129 2.63e-10 ***
sin2piD     -0.044349   0.016831  -2.635  0.01158 *
cos2piD     -0.019996   0.020621  -0.970  0.33750
sin4piD     -0.003777   0.017487  -0.216  0.82998
cos4piD     -0.052928   0.017576  -3.011  0.00430 **
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.08579 on 44 degrees of freedom
Multiple R-squared: 0.7314,     Adjusted R-squared: 0.6948
F-statistic: 19.97 on 6 and 44 DF,  p-value: 4.201e-11

Variance inflation factors:
   log10(Q) log10(Turb)     sin2piD     cos2piD     sin4piD
       cos4piD
     3.4164      2.6701      1.0348      1.3866      1.0624
         1.0567

Duan (1983) smearing factor: 1.0172  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

Nomenclature (all potential variables)

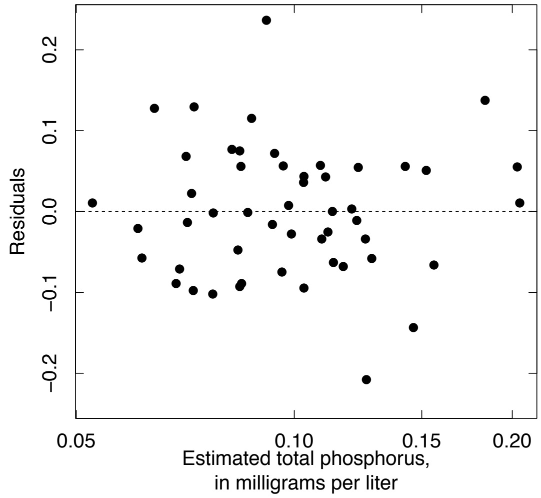
- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- SC is specific conductance, in microsiemens per centimeter at 25° Celsius;
- $Turb$ is turbidity in Formazine Nephelometric Units;
- $Temp$ is water temperature, in ° Celsius;
- $Date$ is Julian day DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\begin{aligned}\log_{10}(Phos) = &-1.4584\\ &-0.0811\log_{10}(Q)+\\ &+0.4574\log_{10}(Turb)\\ &-0.0443\sin[2\pi(Date)]\\ &-0.0200\cos[2\pi(Date)]\\ &-0.0038\sin[4\pi(Date)]\\ &-0.0529\cos[4\pi(Date)]\end{aligned}$$

Residual Plot for Regression

```
CORRELATION OF COEFFICIENTS PHOS
            (Intercept) log10(Q) log10(Turb) sin2piD cos2piD sin4piD cos4piD
(Intercept)       1      -0.0607      -0.5696 -0.1556  0.2165 -0.0813 -0.0310
log10(Q)        -0.0607   1           -0.7713  0.0769 -0.4853  0.2306  0.2231
log10(Turb)     -0.5696  -0.7713       1       0.0346  0.2444 -0.1423 -0.1484
sin2piD         -0.1556   0.0769       0.0346  1      -0.1018 -0.0213  0.0660
cos2piD          0.2165  -0.4853       0.2444 -0.1018  1      -0.1299 -0.1356
sin4piD         -0.0813   0.2306      -0.1423 -0.0213 -0.1299  1       0.0620
cos4piD         -0.0310   0.2231      -0.1484  0.0660 -0.1356  0.0620  1
```



Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 11. Summary of regression analysis for total phosphorus for U.S. Geological Survey streamflow-gaging station 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
    log10(Q)        log10(SC)       log10(Turb)
 Min.   :1.041   Min.   :1.690   Min.   :0.8633
 1st Qu.:1.568   1st Qu.:2.161   1st Qu.:1.0792
 Median :1.908   Median :2.248   Median :1.2553
 Mean   :2.111   Mean   :2.185   Mean   :1.3653
 3rd Qu.:2.542   3rd Qu.:2.328   3rd Qu.:1.6232
 Max.   :3.899   Max.   :2.442   Max.   :2.2304
```

Summary of Regression Analysis for the Constituent of: total organic carbon (OrgC)

```
SUMMARY STATISTICS FOR ORGC, IN LOG10() milligrams per liter
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6021  0.7853  0.8865  0.9247  1.1140  1.2600

REGRESSION EQUATION
lm(formula = log10(ORGC) ~ log10(Q) + log10(SC) + log10(Turb),
    data = the.data)

Residuals:
     Min   1st Qu.  Median   3rd Qu.     Max
-0.16068 -0.03881 -0.01109  0.04640  0.20723

Coefficients:
            Estimate  Std.Error t-value Pr(>|t|)
(Intercept) -0.53685    0.25485  -2.107 0.040767 *
log10(Q)     0.24611    0.02745   8.967 1.41e-11 ***
log10(SC)    0.32656    0.09239   3.535 0.000958 ***
log10(Turb)  0.16730    0.04918   3.402 0.001413 **
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.07622 on 45 degrees of freedom
Multiple R-squared: 0.8619,     Adjusted R-squared: 0.8527
F-statistic: 93.64 on 3 and 45 DF,  p-value: < 2.2e-16

Variance inflation factors:
   log10(Q)   log10(SC) log10(Turb)
     3.5481      2.9318      2.5231

Duan (1983) smearing factor: 1.0144  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

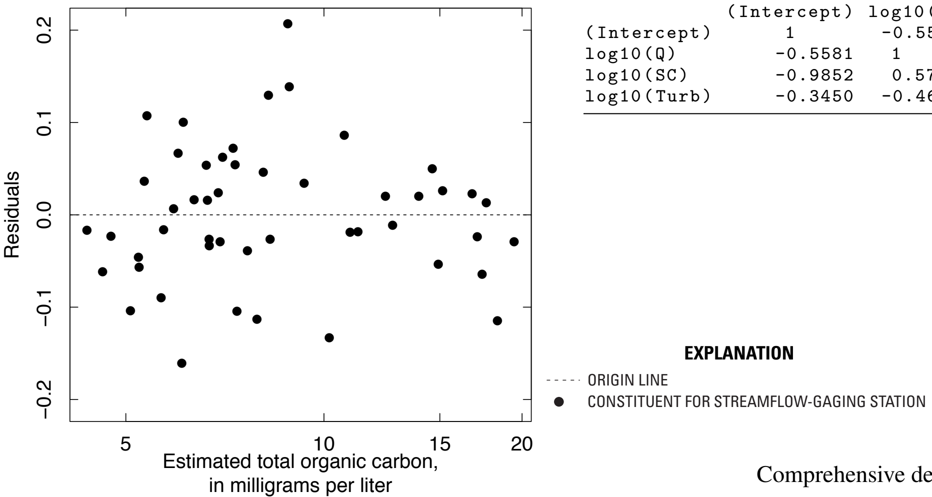
Nomenclature (all potential variables)

- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- 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 DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\begin{aligned}\log_{10}(OrgC) = &-0.5369\\ &+0.2461\log_{10}(Q)\\ &+0.3266\log_{10}(SC)\\ &+0.1673\log_{10}(Turb)\end{aligned}$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS ORGC
            (Intercept) log10(Q) log10(SC) log10(Turb)
(Intercept)      1       -0.5581   -0.9852     -0.3450
log10(Q)       -0.5581   1          0.5737     -0.4695
log10(SC)      -0.9852   0.5737    1            0.2378
log10(Turb)    -0.3450  -0.4695    0.2378      1
```

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 12. Summary of regression analysis for total organic carbon for U.S. Geological Survey streamflow-gaging station 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
  log10(Turb)      I(log10(Turb)^2)    sin2piD           cos2piD
 Min.   :0.8633   Min.   :0.7453   Min.   :-0.99894   Min.   :-0.99777
 1st Qu.:1.0792   1st Qu.:1.1646   1st Qu.:-0.59438   1st Qu.:-0.66325
 Median :1.2041   Median :1.4499   Median : 0.13659   Median : 0.19970
 Mean   :1.3642   Mean   :1.9853   Mean   : 0.08057   Mean   : 0.05389
 3rd Qu.:1.6232   3rd Qu.:2.6349   3rd Qu.: 0.77665   3rd Qu.: 0.67111
 Max.   :2.2304   Max.   :4.9749   Max.   : 0.99998   Max.   : 0.99336
```

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

```
SUMMARY STATISTICS FOR ECB, IN LOG10() most probable number per 100
    milliliters
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.114   1.740   2.041   2.277   2.716   4.643

REGRESSION EQUATION
lm(formula = log10(ECB) ~ log10(Turb) + log10(Turb)^2    + sin2piD +
    cos2piD, data = the.data)

Residuals:
    Min   1st Qu. Median  3rd Qu.    Max
-0.7373 -0.1958 -0.0143  0.2422  0.9307

Coefficients:
                 Estimate  Std.Error t-value Pr(>|t|)
(Intercept)       4.21852    1.02533   4.114 0.000168 ***
log10(Turb)      -4.43709    1.45936  -3.040 0.003970 **
log10(Turb)^2     2.06978    0.49393   4.190 0.000132 ***
sin2piD          -0.12086    0.07790  -1.552 0.127932
cos2piD           0.22499    0.08888   2.531 0.015009 *
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.388 on 44 degrees of freedom
Multiple R-squared: 0.7968,     Adjusted R-squared: 0.7783
F-statistic: 43.12 on 4 and 44 DF,  p-value: 1.106e-14

Variance inflation factors:
     log10(Turb) log10(Turb)^2            sin2piD           cos2piD
         86.0980          87.7590            1.0155            1.2253

Duan (1983) smearing factor: 1.4893  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

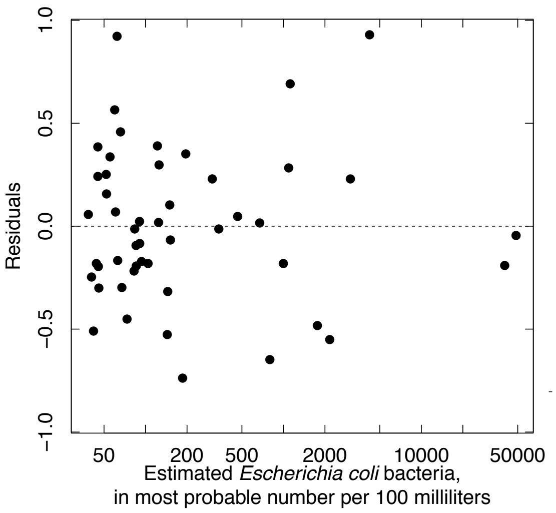
Nomenclature (all potential variables)

- *Q* is streamflow, in cubic feet per second;
- *pH* is pH, in standard units;
- *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 *DOY* (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of *x*.

Algebraic Equation

$$
\begin{aligned}
\log_{10}(ECB) = {} & 4.2185 \\
& - 4.4371 \log_{10}(Turb) \\
& + 2.0700 \log_{10}(Turb)^2 \\
& - 0.1209 \sin[2\pi(Date)] \\
& + 0.2250 \cos[2\pi(Date)]
\end{aligned}
$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS ECB
               (Intercept) log10(Turb) log10(Turb)^2    sin2piD cos2piD
(Intercept)        1         -0.9929            0.9752 -0.0179 -0.3501
log10(Turb)       -0.9929     1                -0.9939 -0.0012  0.3715
log10(Turb)^2      0.9752    -0.9939            1        0.0149 -0.3929
sin2piD           -0.0179    -0.0012            0.0149  1       -0.0248
cos2piD           -0.3501     0.3715           -0.3929 -0.0248  1
```

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 13. Summary of regression analysis for *Escherichia coli* for U.S. Geological Survey streamflow-gaging station 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–9.

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

```
EXPLANATORY VARIABLE SUMMARY STATISTICS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8633  1.0790  1.2040  1.3610  1.6130  2.2300
```

Summary of Regression Analysis for the Constituent of: suspended sediment (SS)

```
SUMMARY STATISTICS FOR SS, IN LOG10() milligrams per liter
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7782  1.1900  1.4310  1.5460  1.7310  3.0370

REGRESSION EQUATION
lm(formula = log10(SS) ~ log10(Turb), data = the.data)

Residuals:
     Min   1st Qu.  Median   3rd Qu.      Max
-0.59759 -0.18961 -0.05024  0.08017  1.03328

Coefficients:
            Estimate  Std.Error t-value Pr(>|t|)
(Intercept)  0.06566    0.19021   0.345    0.731
log10(Turb)  1.08800    0.13538   8.037 1.68e-10 ***
---
Signif.codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

Residual standard error: 0.3382 on 49 degrees of freedom
Multiple R-squared: 0.5686,     Adjusted R-squared: 0.5598
F-statistic: 64.59 on 1 and 49 DF,  p-value: 1.680e-10

Variance inflation factors:
log10(Turb)
          1

Duan (1983) smearing factor: 1.4908  = mean(10^[residuals(model)])
 (Detransformed estimates are to be multiplied this factor.)
```

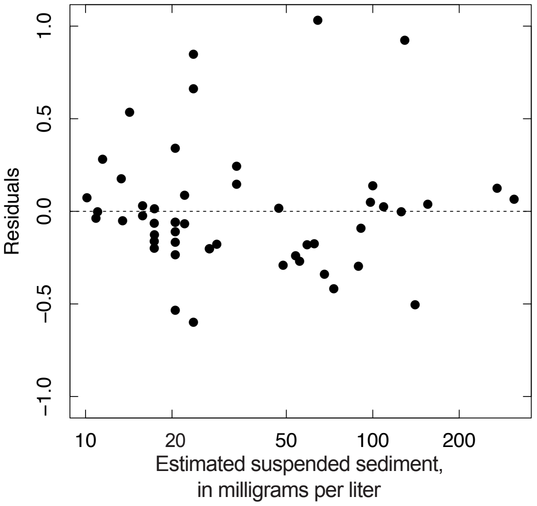
Nomenclature (all potential variables)

- Q is streamflow, in cubic feet per second;
- pH is pH, in standard units;
- 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 DOY (days into year) divided by 365.25; and
- $\log_{10}(x)$ is base-10 logarithm of x.

Algebraic Equation

$$\log_{10}(SS) = 0.0657 + 1.0880\log_{10}(Turb)$$

Residual Plot for Regression



```
CORRELATION OF COEFFICIENTS SS
              (Intercept) log10(Turb)
(Intercept)        1          -0.9685
log10(Turb)      -0.9685       1
```

Comprehensive descriptions of abbreviations for the computer output related to the regression analysis documented in this figure are shown in figure 3.

Figure 14. Summary of regression analysis for suspended sediment for U.S. Geological Survey streamflow-gaging station 08070200 East Fork San Jacinto River near New Caney, Texas, 2005–9.

Computational Procedures of Concentration and Loads

Implementation of the 11 equations shown in figures 4-14 for estimation of constituent concentration is done in this report in the context of concurrent reporting of 90-percent prediction intervals. The prediction interval computation is important because the equations documented in this report are readily used for real-time estimation of the constituents. The regression equations documented in this report provide a computed constituent record available through the online USGS National Real-Time Water-Quality website (NRTWQ) (http://nrtwq.usgs.gov/ks).

A thorough description and presentation of numerical results for a selected equation for selected stream conditions has been provided so that interested parties can verify the basic algorithmic framework needed to implement the equations in this report. Example algorithmic frameworks are described in succession in appendix 2 (for constituents with noncensored data) and appendix 3 (for constituents containing censored data). By way of example, these appendixes describe the application of the equations described in figures 4 and 14 for computation of an instantaneous 15-minute load (15 minutes is the smallest time interval for computing continuously measured water-quality values). The implementation does not involve a particularly high level of mathematics, but basic knowledge of logarithms, statistics associated with regression, and matrix algebra is required.

Summary

In December 2005, the U.S. Geological Survey (USGS), in cooperation with the City of Houston, Texas, began collecting discrete water-quality samples for nutrients, total organic carbon, bacteria (*Escherichia coli* and total coliform), atrazine, and suspended sediment at two USGS streamflow-gaging stations that represent watersheds contributing to Lake Houston (08068500 Spring Creek near Spring, Tex., and 08070200 East Fork San Jacinto River near New Caney, Tex.). These sampling sites are respectively referred to as the Spring Creek site and East Fork San Jacinto River site. Data from discrete water-quality samples collected during 2005–9, in conjunction with continuously monitored real-time data that included streamflow and other physical water-quality properties (specific conductance, pH, water temperature, turbidity, and dissolved oxygen), were used to develop regression models for estimation of concentrations of water-quality constituents of substantial source watersheds to Lake Houston.

The regression models 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 most perennial streams for the purpose of estimating real-time water quality data.

The continuously monitored streamflow and other physical water-quality properties, in conjunction with regression models that use 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 lacking continuous record.

Streamflow, physical water-quality properties, and selected constituent concentrations were collected at the two sites. During 2005–9, discrete samples were collected at the Spring Creek site (58 samples) and at the East Fork San Jacinto River site (51 samples). Hydrologic conditions within the Spring Creek and East Fork San Jacinto River sites 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 facilitate detection of 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 streamflow. During storms or periods of high flow, unscheduled samples were also periodically collected during the first year of the study. During the second and third year of the study (December 2006–December 2008) discrete water-quality samples were collected approximately once a month at both the Spring Creek and East Fork San Jacinto River sites. During the fourth year of the study (December 2008–December 2009), an approximate monthly sampling schedule was maintained for the Spring Creek site, whereas samples collected at East Fork San Jacinto River site were reduced to a quarterly schedule. Instead of sampling on a fixed frequency during the second through fourth years of the study, stormwater-runoff samples were collected whenever possible.

Regression analyses were done by using streamflow, continuous water-quality, and discrete water-quality data collected during 2005–9 at the Spring Creek and East Fork San Jacinto River sites. The R environment for statistical computing was used to develop algebraically representable, multiple-linear regression equations to (1) estimate concentrations for selected water-quality constituents and (2) estimate prediction intervals or quantification of uncertainty. The potential explanatory or predictive variables included streamflow, specific conductance, pH, water temperature, turbidity, dissolved oxygen, 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. Logarithmic transformations (base-10) on the response and explanatory variables were used to improve linearity and to mitigate for nonnormality and heteroscedasticity in model residuals. 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. Factors used as indicators of general model reliability include the adjusted R-squared, the residual standard error, residual plots, and p-values.

Regression equations for the Spring Creek site were developed for nitrite plus nitrate nitrogen, total phosphorus, organic carbon, *E. coli* bacteria, atrazine, and suspended sediment. Adjusted R-squared values for the Spring Creek models ranged from .582–.922 (dimensionless). The residual standard errors ranged from .073–.447 (base-10 logarithm).

Regression equations for the East Fork San Jacinto River site were developed for nitrite plus nitrate nitrogen, total phosphorus, organic carbon, *E. coli* bacteria, and suspended sediment. Adjusted R-squared values for the East Fork San Jacinto River models ranged from .253–.853 (dimensionless). The residual standard errors ranged from .076–.388 (base-10 logarithm).

In conjunction with estimated concentrations, constituent loads can be estimated by multiplying the estimated concentration by the corresponding streamflow and by applying the appropriate conversion factor. By calculating loads from estimated constituent concentrations, a continuous record of estimated loads can be produced.

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.

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

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.

City of Houston, 2011, Public Works and Engineering Drinking Water Operations: accessed March 2011, at http://www.publicworks.houstontx.gov/utilities/drinkingwater.html.

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, 13 p.

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

Duan, Naihua, 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 Hall, 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.

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/.

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.

Kaplan, E.L., and Meier, Paul, 1958, Nonparametric estimation of incomplete observations: Journal of the American Statistical Association, v. 53, p. 457–481.

Kasmarek, M.C., Johnson, M.R., and Ramage, J.K., 2010, Water-level altitudes 2010 and water-level changes in the Chicot, Evangeline, and Jasper aquifers and compaction 1973–2009 in the Chicot and Evangeline aquifers, Houston-Galveston region, Texas: U.S. Geological Survey Scientific Investigations Map 3138, 17 p., 16 sheets, 1 appendix.

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 Investigation, 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 Investigation, book 3, chapter A10, 59 p., accessed May 3, 2008, at http://pubs.usgs.gov/twri/twri3-a10/.

Lee, Lopaka, 2009, NADA—Nondetects and data analysis for environmental data: R package version 1.5-3, dated December 22, 2010, initial package release June 24, 2004, http://www.cran.r-project.org/package=NADA.

Maindonald, J.H., and Braun, John, 2003, Data analysis and graphics using R—An example-based approach: Cambridge, Cambridge University Press, 362 p.

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.

McFadden, Daniel, 1974, The measurement of urban travel demand: Journal of Public Economics, v. 3, 303–328.

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/.

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, 2011, Southeast Texas climate data: National Weather Service Forecast Office, Houston/Galveston, Tex., accessed March 13, 2011, at http://www.srh.noaa.gov/hgx/climate.htm.

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.

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, 2010, R—A language and environment for statistical computing (version 2.12.2): R Foundation for Statistical Computing, Vienna, Austria. (Also available at http://www.R-project.org.)

Rasmussen, P.P., Gray, J.R., Glysson, G.D., and Ziegler, A.C., 2009, Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity-sensor and streamflow data: U.S. Geological Survey Techniques and Methods, book 3, chap. C4, 52 p.

Rasmussen, T.J., Lee, C.J., and Ziegler, A.C., 2008, Estimation of constituent concentrations, loads, and yields in streams of Johnson County, northeast Kansas, using continuous water-quality monitoring and regression models, October 2002 through December 2006: U.S. Geological Survey Scientific Investigations Report 2008–5014, 103 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.

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.

Texas Commission on Environmental Quality, 2011, Texas water quality inventory and 303(d) list: accessed August 10, 2011, at http://www.tceq.texas.gov/assets/public/compliance/monops/water/08twqi/2008_303d.pdf.

Texas State Climatologist, 2011, Texas climatic bulletin: Office of the Texas State Climatologist, College of Geosciences, Department of Atmospheric Sciences, Texas A&M University, accessed March 16, 2011, at http://www.met.tamu.edu/osc/TXclimat.html.

Texas State Data Center, 2011, 2009 total population estimates for Texas metropolitan statistical areas: Office of the State Demographer, Texas population estimates and projection program, accessed March 16, 2011, at http://txsdc.utsa.edu/tpepp/2009_txpopest_msa.php.

Turnipseed, D.P., and Sauer, V.B., 2010, Discharge measurements at gaging stations: U.S. Geological Survey Techniques and Methods, book 3, chap. A8, 87 p.

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. (Also available at http://pubs.water.usgs.gov/twri9A.)

U.S. Geological Survey, 2011, National Water Information System: accessed March 16, 2011, at http://waterdata.usgs.gov/tx/nwis/.

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/.

Wang, Dongliang, Hutson, A.D., and Miecznikowski, J.C., 2010, L-moment estimation for parametric survival models given censored data: Statistical Methodology, v. 7, no. 6, p. 655–667.

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.

Appendix 1

Appendix 1. Results from environmental and quality-control sample pairs and equipment blanks collected for two tributaries (Spring Creek and East Fork San Jacinto River) to Lake Houston near Houston, Texas, 2005–9.

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

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)	Phosphorus, water, filtered (milligrams per liter)	Phosphorus, 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)	Organic carbon, water, unfiltered (milligrams per liter)	Suspended sediment (milligrams per liter)
Spring Creek near Spring, Tex. 08068500															
12/1/2005	1330	Environ.	0.85	1.3	0.08	6.87	0.028	1.56	1.61	1.73	130	>2400	0.3	9.6	36
12/1/2005	1331	Replicate	—	—	—	—	—	—	—	—	90	>2400	—	—	—
2/7/2006	1400	Environ.	0.75	1.1	0.06	4.52	0.02	0.95	0.97	1.11	60	3400	0.95	9.8	33
2/7/2006	1401	Replicate	—	—	—	—	—	—	—	—	55	3300	—	—	—
2/21/2006	1400	Environ.	1.1	1.4	0.39	4.01	0.054	0.91	0.87	1.05	37	6100	1.51	9.6	24
2/21/2006	1401	Replicate	1.2	1.4	0.38	4.04	0.054	0.91	0.86	1.02	27	6500	1.34	9.3	23
5/16/2006	1115	Environ.	0.78	1.8	<.04	1.33	0.069	0.63	0.66	0.87	490	110000	2.56	13.4	60
5/16/2006	1116	Replicate	0.78	1.9	<.04	1.32	0.07	0.64	0.68	0.85	690	98000	2.68	15.3	61
6/20/2006	1000	Environ.	0.86	1.5	0.035	0.35	0.034	0.197	0.24	0.44	8000	410000	1.45	17	547
6/20/2006	1001	Replicate	—	—	—	—	—	—	—	—	7100	460000	—	—	—
6/28/2006	845	Blank	0.12	<.10	0.015	<.06	<.002	<.006	<.02	<.02	—	—	<.10	<.4	<1
6/28/2006	930	Environ.	1	1.2	0.088	4.66	0.028	1.01	1	1.25	43	>2400	0.62	8.9	31
7/12/2006	1300	Environ.	0.83	1.1	0.058	3.36	0.057	0.719	0.7	0.89	120	41000	0.68	9	38
7/12/2006	1346	Replicate	—	—	—	—	—	—	—	—	190	69000	—	—	—
8/23/2006	1320	Blank	E.07	<.10	<.010	<.06	<.002	<.006	<.02	<.02	—	—	<.10	<.4	<1
8/23/2006	1345	Environ.	0.77	1.8	<.010	2.12	0.021	0.738	0.78	1	6800	440000	1.75	14.1	147
9/20/2006	1230	Environ.	1.1	1.5	0.166	2.01	0.101	0.891	0.94	1.1	270	82000	1.73	13.1	56
9/20/2006	1231	Replicate	1.1	1.5	0.17	2.02	0.101	0.901	0.95	1.09	360	49000	1.77	10.3	46
8/15/2007	1420	Environ.	0.64	0.73	0.055	3.56	0.053	0.735	0.75	0.92	310	17000	0.18	8.1	42
8/15/2007	1421	Replicate	—	—	—	—	—	—	—	—	190	17000	—	—	—
2/14/2008	1110	Environ.	0.7	1.1	0.011	1.47	0.022	0.252	0.27	0.41	2200	33000	1.95	11.6	43
2/14/2008	1110	Replicate	—	—	—	—	—	—	—	—	2100	37000	—	—	—
11/13/2008	1315	Environ.	0.5	0.9	<.020	0.14	0.006	0.069	0.09	0.19	3300	57000	0.43	11.4	587
11/13/2008	1316	Replicate	0.51	0.91	<.020	0.15	0.006	0.068	0.09	0.19	—	—	0.45	11.3	755
3/10/2009	1027	Environ.	1	1.3	0.189	4.66	0.066	1.24	1.22	1.34	34	37000	0.64	8.3	32
3/10/2009	1028	Replicate	1	1.3	0.189	4.65	0.066	1.25	1.21	1.32	—	—	1	8.3	37
7/29/2009	1120	Environ.	0.94	1.3	0.085	5.72	0.101	1.65	1.62	1.83	20	18000	0.3	9.9	46
7/29/2009	1121	Replicate	0.96	1.3	0.078	5.8	0.102	1.66	1.61	1.77	—	—	0.39	7.8	139
East Fork San Jacinto River near New Caney, Tex. 08070200															
12/1/2005	1020	Environ.	0.14	0.23	<.04	0.1	<.008	0.03	0.039	0.086	43	1700	<.10	4	11
12/1/2005	1021	Replicate	—	—	—	—	—	—	—	—	34	1700	—	—	—
12/21/2005	930	Blank	E.06	<.10	<.04	<.06	<.008	<.02	E.002	<.004	<1	<1	<.10	<.4	1
12/21/2005	1030	Environ.	0.45	0.38	<.04	0.07	<.008	<.02	0.025	0.058	130	2000	0.1	7.7	16
3/7/2006	1130	Environ.	0.46	0.44	<.04	0.07	<.008	E.01	0.025	0.076	34	1600	<.10	7.3	17
3/7/2006	1131	Replicate	—	—	—	—	—	—	—	—	34	1700	—	—	—
4/4/2006	1100	Environ.	0.74	0.8	0.05	0.17	E.006	0.02	0.037	0.09	43	5200	0.13	14.2	19

Appendix 1. Results from environmental and quality-control sample pairs and equipment blanks collected for two tributaries (Spring Creek and East Fork San Jacinto River) to Lake Houston near Houston, Texas, 2005–9.

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

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)	Phosphorus, water, filtered (milligrams per liter)	Phosphorus, 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)	Organic carbon, water, unfiltered (milligrams per liter)	Suspended sediment (milligrams per liter)
East Fork San Jacinto River near New Caney, Tex. 08070200—Continued											East Fork San Jacinto River near New Caney, Tex. 08070200—Continued				
4/4/2006	1101	Replicate	0.65	0.8	0.05	0.17	E.007	0.02	0.041	0.101	50	10000	<.10	12.7	14
7/25/2006	945	Environ.	0.35	0.47	0.027	0.51	0.007	0.059	0.071	0.158	93	9200	<.10	7	6
7/25/2006	946	Replicate	—	—	—	—	—	—	—	—	84	12000	—	—	—
8/8/2006	1000	Environ.	0.38	0.39	0.023	0.22	0.002	0.034	0.043	0.111	120	8100	<.10	5.5	12
8/8/2006	1001	Replicate	0.42	0.39	0.022	0.23	0.003	0.032	0.043	0.111	150	6900	<.10	6.1	13
9/6/2006	1100	Environ.	0.36	0.27	0.025	0.08	E.001	0.024	0.034	0.113	26	13000	<.10	4	6
9/6/2006	1130	Blank	0.14	<.10	E.009	<.06	<.002	<.006	<.004	<.004	—	—	<.10	<.4	<1
10/4/2006	945	Environ.	0.22	0.3	E.016	0.15	<.002	0.047	0.056	0.124	63	4900	<.10	4.2	15
10/4/2006	946	Replicate	—	—	—	—	—	—	—	—	52	3900	—	—	—
10/18/2006	1330	Environ.	0.65	0.98	<.020	E.05	0.003	0.013	0.036	0.134	610	44000	0.12	18.2	125
10/18/2006	1331	Replicate	—	—	—	—	—	—	—	—	690	39000	—	—	—
1/15/2007	1340	Environ.	0.5	0.76	0.038	0.09	0.003	0.021	0.029	0.122	2100	18000	<.10	16.1	110
1/15/2007	1341	Replicate	0.52	0.78	0.035	0.09	0.003	0.019	0.028	0.123	2000	17000	<.10	13.3	81
5/24/2007	1038	Environ.	—	—	—	—	—	—	—	—	250	10000	<.10	—	39
5/24/2007	1039	Replicate	—	—	—	—	—	—	—	—	240	9900	<.10	—	39
8/15/2007	1059	Blank	0.35	<.10	0.031	<.06	E.001	<.006	<.006	<.008	—	—	—	—	1
8/15/2007	1132	Environ.	—	0.33	<.020	0.07	0.004	0.015	0.017	0.081	71	15000	<.10	6.5	27
4/23/2008	1030	Environ.	0.2	0.32	<0.02	0.15	0.005	0.021	0.031	0.068	100	5200	<.10	5.77	45
4/23/2008	1030	Replicate	—	—	—	—	—	—	—	—	108	4600	—	—	—
1/7/2009	1050	Environ.	0.27	0.37	E.012	0.08	0.005	0.015	0.02	0.069	300	7600	<.10	7.8	168
1/7/2009	1051	Replicate	0.3	0.36	E.011	0.08	0.005	0.014	0.018	0.065	—	—	<.10	6.5	174
Houston Lab at Shenandoah, Tex. 301056095265000											Houston Lab at Shenandoah, Tex. 301056095265000				
8/22/2006	1330	Equip. Blank	<.10	<.10	<.010	<.06	<.002	E.003	<.02	<.02	—	—	<.10	<.4	—
11/28/2007	1358	Equip. Blank	<.14	<.14	<.020	<.04	<.002	<.006	<.006	E.006	—	—	—	—	1
11/28/2007	1359	Equip. Blank	—	—	—	—	—	—	—	—	—	—	—	0.9	—
12/3/2008	1035	Equip. Blank	E.09	<.10	<.020	<.04	<.002	<.008	<.02	<.02	—	—	—	<.6	1
12/3/2008	1555	Equip. Blank	—	<.10	<.020	<.016	—	<.008	—	<.008	—	—	—	E.4	—
12/3/2008	1615	Equip. Blank	—	E.07	<.020	<.016	—	<.008	—	<.008	—	—	—	E.5	—

Appendix 2—Example computations

(definition of terms and quantities, concentration estimates, 90-percent prediction intervals, and loads) for nitrite plus nitrate (N02 N03) at U.S. Geological Survey streamflow-gaging station 08070200 East Fork San Jacinto River near New Caney, Texas.

Example computations (definition of terms and quantities) for nitrite plus nitrate (NO_2 NO_3) for 08070200 East Fork San Jacinto River near New Caney, Tex.

Given values for example problem

- $Q = 7{,}930$ cubic feet per second;
- $DOY = 292.5243$ day of year (October 19, 2006 at 12:35 hours); and
- $C_o = <0.06$ observed NO_2NO_3 concentration in milligrams per liter for DOY.

Nomenclature (select variables and symbols)

- NO_2NO_3 is nitrite plus nitrate in milligrams per liter;
- Q is streamflow in cubic feet per second;
- S is residual scale in base-10 logarithms of milligrams per liter. The survival regression provides the natural logarithm of the residual scale as part of standard (conventional) output. A detransformation by exp() is seen in these examples;
- $Date$ is Julian day DOY (days into year) divided by 365.25;
- $\log_{10}(x)$ is base-10 logarithm of x; and
- π or `pi` is the quantity “pi” of approximately 3.14159.

Note, text set in a `monospaced` font (often within brackets) reflects the precise nomenclature in computer output seen in figure 10. This specific font change is used to promote parallelism between the regression figure and these example computations.

Nitrite plus nitrate equation and diagnostics for 08070200 East Fork San Jacinto River near New Caney, Tex.

$$\log_{10}(NO_2NO_3) = -0.8938 - 0.0288\log_{10}(Q) + 0.0289\sin[2\pi(Date)] - 0.1911\cos[2\pi(Date)]$$

$$NO_2NO_3 = \aleph \times 10^{\log_{10}(NO_2NO_3)} \quad \text{in milligrams per liter}$$

$S = 0.2519 = (\exp(-1.3788))$	[`Log(scale)`]	base-10 logarithms of milligrams per liter
$\mathrm{DF} = 51\,(\text{samples}) - 4\,(\text{parameters}) = 47$	[`DF`]	degrees of freedom
$\mathrm{SE}_B = 0.1251$	[`Std.Error`]	Standard error of coefficient on intercept term
$\mathrm{SE}_Q = 0.0596$	[`Std.Error`]	Standard error of coefficient on streamflow term
$\mathrm{SE}_{sin} = 0.0504$	[`Std.Error`]	Standard error of coefficient on 2π sine term
$\mathrm{SE}_{cos} = 0.0590$	[`Std.Error`]	Standard error of coefficient on 2π cosine term
$\aleph = 1.0758$	[`ALEPH`]	Duan (1983) smearing factor (dimensionless)

Correlation of coefficients matrix for given constituent equation

$$\begin{bmatrix} 1 & -0.9572 & -0.0923 & 0.4401 \\ -0.9572 & 1 & 0.0703 & -0.4678 \\ -0.0923 & 0.0703 & 1 & -0.0730 \\ 0.4401 & -0.4678 & -0.0730 & 1 \end{bmatrix}$$

The matrix on the left is used for prediction-limit computation. The matrix will require additional operations to convert it to the inverted X-prime X-transverse matrix. This second matrix is critical for computation of leverage for a given prediction. Once the leverage is known, the subsequent computations to acquire the prediction limits are relatively straightforward.

Example computations (computations of estimates and 90-percent prediction limits) for nitrite plus nitrate (NO_2 NO_3) for 08070200 East Fork San Jacinto River near New Caney, Tex.

Computation of an estimate and 90-percent prediction limits

The estimate of nitrite plus nitrate for the example problem is readily computed by

$$NO_2NO_3 = 1.0758 \times 10^{-0.8938-0.0288\log_{10}(Q)+0.0289\sin[2\pi(292.5243/365.25)]-0.1911\cos[2\pi(292.5243/365.25)]}$$

$$NO_2NO_3 = 0.0867 \text{ milligrams per liter}$$

The lower (↓) and upper (↑) 90-percent prediction limits of NO_2NO_3 are respectively computed by

$${}^{\downarrow}NO_2NO_3^{[\alpha/2]} = 10^{\log_{10}(NO_2NO_3)-|\mathrm{t}_{[\alpha/2,\mathrm{DF}]}|\,S\sqrt{1+h_o}} \quad \text{and} \quad {}^{\uparrow}NO_2NO_3^{[\alpha/2]} = 10^{\log_{10}(NO_2NO_3)+|\mathrm{t}_{[\alpha/2,\mathrm{DF}]}|\,S\sqrt{1+h_o}},$$

where NO_2NO_3 is the estimate, $\mathrm{t}_{[\alpha/2,\mathrm{DF}]}$ is the lower tail of the t-distribution for $\mathrm{DF}^{[NO_2NO_3]}$ degrees of freedom at the α significance level ($\alpha = [100-90 \text{ percent}]/100$), S is the residual scale of the maximum likelihood regression, and h_o is the leverage of the estimate. The lower and upper prediction limits of the estimated $NO_2NO_3 = 0.0867$ milligrams per liter are respectively computed by

$${}^{\downarrow}NO_2NO_3 = 10^{\log_{10}(0.0867)-1.6779\times 0.2519\sqrt{1+0.2256}} = 0.0295 \text{ milligrams per liter and}$$

$${}^{\uparrow}NO_2NO_3 = 10^{\log_{10}(0.0867)+1.6779\times 0.2519\sqrt{1+0.2256}} = 0.2546 \text{ milligrams per liter.}$$

Thus, the 90-percent prediction interval is $\boxed{0.030 \le NO_2NO_3 = 0.087 \le 0.255}$ milligrams per liter.

Computation of an estimate and 90-percent prediction limits from maximum likelihood regression by using the R environment for statistical computing (R Development Core Team, 2010)

```
Q    <- 7930      # streamflow in cubic feet per second
DOY  <- 292.5243  # Days into year (Day of Year)

Date <- DOY/365.25; DF <- 47; ALEPH <- 1.0758; PERCENT <- 90
X <-c(1, log10(Q), sin(2*pi*Date), cos(2*pi*Date)) # array of predictor variables

CORMAT <- # Correlation of Coefficients Matrix
matrix(c( 1      , -0.9572, -0.0923,  0.4401,
          -0.9572,  1      ,  0.0703, -0.4678,
          -0.0923,  0.0703,  1      , -0.0730,
           0.4401, -0.4678, -0.0730,  1      ), ncol=4) # a 4x4 matrix
S <- exp(-1.3788) # Note the use of exp() for the "Log(scale)" of the regression
print(S) # in order to see at least four decimal points, compare to Scale= 0.252
[1] 0.2518806  # confirming, this value is in base-10 logarithmic units
SIGMAS <- c(0.1251, 0.0596, 0.0504, 0.0590) # Std. Error of the model coefficients

# The next two lines of code represent substantially complex mathematical operations for
# which the the use of a computing environment (not calculator or spreadsheet) such as R
# for this report is justified.
XPXI <- CORMAT * outer(SIGMAS, SIGMAS) / S^2 # inverted X-prime X-transverse matrix
ho <- t(X) %*% XPXI %*% X  # leverage of the prediction
print(ho)    # show the leverage of the prediction (ho used in hand computations shown above)
[1] 0.2256367

alpha <- 1 - PERCENT/100   # for the prediction interval
QT <- abs(qt(alpha/2, df= DF)) # t-distribution multiplier (sign change handled later)
print(QT)    # show the lower-tail of t-distribution (QT used in hand computations shown above)
[1] 1.677927

Xbar <- -0.8938*X[1] - 0.0288*X[2] + 0.0289*X[3] - 0.1911*X[4]  # the prediction
Xlo <- Xbar - QT*S*sqrt(1+ho); Xhi <- Xbar + QT*S*sqrt(1+ho) # lower/upper predict. limits

C <- ALEPH * 10^c(Xlo, Xbar, Xhi); names(C) <- c("lwr", "fit", "upr")
print(C) # show the estimated limits and value using numerical values of this report
      lwr        fit        upr
0.02952812 0.08672235 0.25469844   # Lower limit, the prediction, and upper limit

# The predict.survreg() function does not provide direct capacity for prediction limits.
# If one has the model object (the.lm) that contains the NO2NO3 regression, then prediction is
print(ALEPH * 10^(predict(the.lm))[24]) # the 24th value in the underlying data
[1] 0.08719011   # numerical slight numerical difference because of rounding
```

Example computations (computation of load) for nitrite plus nitrate (NO_2 NO_3) for 08070200 East Fork San Jacinto River near New Caney, Tex.

Unit conversion to 15-minute load from streamflow and constituent concentration

The relation between a 15-minute load L, streamflow Q, and NO_2NO_3 concentration C is determined as follows

$$L\left[\frac{\text{kilogram}}{\text{15 minutes}}\right] = Q\left[\frac{\text{cubic feet}}{\text{second}}\right] \times C\left[\frac{\text{milligram}}{\text{liter}}\right] \times \underbrace{\left[\frac{\text{900 seconds}}{\text{15 minutes}}\right] \times \left[\frac{\text{28.317 liter}}{\text{cubic feet}}\right] \times \left[\frac{\text{kilogram}}{1 \times 10^6 \text{ milligram}}\right]}_{K = 0.02549}$$

Computation of 15-minute load of nitrite plus nitrate

The 15-minute load L in kilograms per 15 minutes for streamflow $Q = 7{,}930$ cubic feet per second of nitrite plus nitrate with a concentration of $C = 0.0867$ milligrams per liter is computed by

$$L = Q \times C \times K$$
$$L = 7{,}930 \times 0.0867 \times 0.02549$$
$$L = 17.53 \text{ kilograms per 15 minutes}$$

Computation of 15-minute load of nitrite plus nitrate by using R environment for statistical computing (R Development Core Team, 2010)

```
# Q is streamflow from code shown on previous page (7,930 cubic feet per second)
# C is the lower, estimate, and upper concentrations shown on previous page
#               (0.02952812, 0.08672235, 0.25469844)
K <- 0.02549 # unit conversion factor shown in previous hand derivation shown
    above
L <- Q * C * K # compute the 15-minute load prediction and prediction limits
print(L) # show the 15-minute load along with prediction limits
      lwr       fit       upr
[1]  5.968687 17.529683 51.483647
```

Thus, the 15-minute load of nitrite plus nitrate with 90-percent prediction limits is $\boxed{5.97 \le L = 17.53 \le 51.48}$ kilograms per 15 minutes.

Appendix 3—Example computations (definition of terms and quantities, concentration estimates, 90-percent prediction intervals, and loads) for total phosphorus (Phos) at U.S. Geological Survey streamflow-gaging station 08068500 Spring Creek near Spring, Texas.

Example computations (definition of terms and quantities) for total phosphorus (Phos) for 08068500 Spring Creek near Spring, Tex.

Given values for example problem

- $Q = 40$ cubic feet per second;
- $SC = 454$ microsiemens per centimeter at 25° Celsius;
- $Turb = 18$ Formazine Nephelometric units;
- $DOY = 52.5833$ day of year (February 21, 2006 at 14:00 hours); and
- $C_o = 1.05$ observed concentration, in milligrams per liter for DOY.

Nomenclature (select variables and symbols)

- $Phos$ is total phosphorus in milligrams per liter;
- Q is streamflow in cubic feet per second;
- RSE is residual standard error σ in base-10 logarithms of milligrams per liter;
- SC is specific conductance, in microsiemens per centimeter at 25° Celsius;
- $Turb$ is turbidity in Formazine Nephelometric units;
- $Date$ is Julian day DOY (days into year) divided by 365.25;
- $\log_{10}(x)$ is base-10 logarithm of x; and
- π or `pi` is the quantity “pi” of approximately 3.14159.

Note, text set in a `monospaced` font (often within brackets) reflects the precise nomenclature in computer output seen in figure 5. This specific font change is used to promote parallelism between the regression figure and these example computations.

Total phosphorus equation and diagnostics for 08068500 Spring Creek near Spring, Tex.

$$\log_{10}(Phos) = -3.5808 - 1.2592\log_{10}(SC) + 0.1678\log_{10}(Turb) - 0.0247\sin[2\pi(Date)] - 0.0627\cos[2\pi(Date)]$$

$$Phos = \aleph \times 10^{\log_{10}(Phos)} \quad \text{in milligrams per liter}$$

$\mathrm{RSE} = 0.0954$	`[Residual standard error]`	base-10 logarithms of milligrams per liter
$\mathrm{DF} = 56\,(\text{samples}) - 5\,(\text{parameters}) = 51$	`[DF]`	degrees of freedom
$\mathrm{SE}_B = 0.2605$	`[Std.Error]`	Standard error of coefficient on intercept term
$\mathrm{SE}_{SC} = 0.0791$	`[Std.Error]`	Standard error of coefficient on specific conductance term
$\mathrm{SE}_{Turb} = 0.0470$	`[Std.Error]`	Standard error of coefficient on turbidity term
$\mathrm{SE}_{sin} = 0.0172$	`[Std.Error]`	Standard error of coefficient on 2π sine term
$\mathrm{SE}_{cos} = 0.0197$	`[Std.Error]`	Standard error of coefficient on 2π cosine term
$\aleph = 1.0222$	`[ALEPH]`	Duan (1983) smearing factor (dimensionless)

Correlation of coefficients matrix for given constituent equation

$$\begin{bmatrix} 1 & -0.9777 & -0.8447 & -0.0966 & -0.0567 \\ -0.9777 & 1 & 0.7190 & 0.0750 & 0.0324 \\ -0.8447 & 0.7190 & 1 & 0.1310 & 0.1361 \\ -0.0966 & 0.0750 & 0.1310 & 1 & -0.0164 \\ -0.0567 & 0.0324 & 0.1361 & -0.0164 & 1 \end{bmatrix}$$

The matrix on the left is used for prediction-limit computation. The matrix will require additional operations to convert it to the inverted X-prime X-transverse matrix. This second matrix is critical for computation of leverage for a given prediction. Once the leverage is known, the subsequent computations to acquire the prediction limits are relatively straightforward.

Example computations (computations of estimates and 90-percent prediction limits) for total phosphorus (Phos) for 08068500 Spring Creek near Spring, Tex.

Computation of an estimate and 90-percent prediction limits

The estimate of total phosphorus for the example problem is readily computed by

$$Phos = 1.0222 \times 10^{-3.5808+1.2592\log_{10}(SC)+0.1678\log_{10}(Turb)-0.0247\sin[2\pi(52.8333/365.25)]-0.0627\cos[2\pi(52.8333/365.25)]}$$

$$Phos = 0.8452 \text{ milligrams per liter}$$

The lower (↓) and upper (↑) 90-percent prediction limits of *Phos* are respectively computed by

$$^{\downarrow}Phos^{[\alpha/2]} = 10^{\log_{10}(Phos)-|\mathrm{t}_{[\alpha/2,\mathrm{DF}]}|\,RSE\sqrt{1+h_o}} \quad \text{and} \quad ^{\uparrow}Phos^{[\alpha/2]} = 10^{\log_{10}(Phos)+|\mathrm{t}_{[\alpha/2,\mathrm{DF}]}|\,RSE\sqrt{1+h_o}},$$

where *Phos* is the estimate, $\mathrm{t}_{[\alpha/2,\mathrm{DF}]}$ is the lower tail of the t-distribution for $\mathrm{DF}^{[Phos]}$ degrees of freedom at the α significance level ($\alpha = [100 - 90 \text{ percent}]/100$), *RSE* is the residual standard error, and h_o is the leverage of the estimate.
The lower and upper prediction limits of the estimated $Phos = 0.8452$ milligrams per liter are respectively computed by

$$^{\downarrow}Phos = 10^{\log_{10}(0.8452)-1.6753\times 0.0954\sqrt{1+0.0664}} = 0.5780 \text{ milligrams per liter and}$$

$$^{\uparrow}Phos = 10^{\log_{10}(0.8452)+1.6753\times 0.0954\sqrt{1+0.0664}} = 1.2360 \text{ milligrams per liter.}$$

Thus, the 90-percent prediction interval is $\boxed{0.578 \leq Phos = 0.845 \leq 1.24}$ milligrams per liter.

Computation of an estimate and 90-percent prediction limits from least-squares regression by using the R environment for statistical computing (R Development Core Team, 2010)

```
Q    <- 40       # streamflow in cubic feet per second
SC   <- 454      # specific conductance in microsiemens per centimeter at 25 degree Celsius
Turb <- 18       # turbidity in Formazine Nephelometric units
DOY  <- 52.5833  # Days into year (Day of Year)

Date <- DOY/365.25; DF <- 51; ALEPH <- 1.0222; PERCENT <- 90
X <-c(1, log10(SC), log10(Turb), sin(2*pi*Date), cos(2*pi*Date)) # array of predictor variables

CORMAT <- # Correlation of Coefficients Matrix
matrix(c( 1     , -0.9777, -0.8447, -0.0966, -0.0567,
         -0.9777,  1     ,  0.7190,  0.0750,  0.0324,
         -0.8447,  0.7190,  1     ,  0.1310,  0.1361,
         -0.0966,  0.0750,  0.1310,  1     , -0.0164,
         -0.0567,  0.0324,  0.1361, -0.0164,  1     ), ncol=5) # a 5x5 matrix
RSE <- 0.0954    # residual standard error
SIGMAS <- c(0.2605, 0.0791, 0.0470, 0.0172, 0.0197) # Std. Error of the model coefficients

# The next two lines of code represent substantially complex mathematical operations for
# which the the use of a computing environment (not calculator or spreadsheet) such as R
# for this report is justified.
XPXI <- CORMAT * outer(SIGMAS, SIGMAS) / RSE^2 # inverted X-prime X-transverse matrix
ho <- t(X) %*% XPXI %*% X  # leverage of the prediction
print(ho)    # show the leverage of the prediction (ho used in hand computations shown above)
[1] 0.06637472

alpha <- 1 - PERCENT/100   # for the prediction interval
QT <- abs(qt(alpha/2, df= DF)) # t-distribution multiplier (sign change handled later)
print(QT)    # show the lower-tail of t-distribution (QT used in hand computations shown above)
[1] 1.675285

Xbar <- -3.5808*X[1] + 1.2592*X[2] + 0.1678*X[3] - 0.0247*X[4] - 0.0627*X[5]  # the prediction
Xlo <- Xbar - QT*RSE*sqrt(1+ho); Xhi <- Xbar + QT*RSE*sqrt(1+ho) # lower/upper predict. limits

C <- ALEPH * 10^c(Xlo, Xbar, Xhi); names(C) <- c("lwr", "fit", "upr")
print(C) # show the estimated limits and value using numerical values of this report
      lwr       fit       upr
0.5779972 0.8452142 1.2359697   # Lower limit, the prediction, and upper limit
# If one has the model object (the.lm) that contains the PHOS regression, then
# the prediction limits are extracted using more significant figures than in this report.
print(sort(ALEPH * 10^(predict.lm(the.lm, interval="prediction", level=0.90))[4,]))
      lwr       fit       upr
0.5775619 0.8444275 1.2345999   # Values acceptably similar to previous, reliability is shown.
```

Example computations (computation of load) for total phosphorus (Phos) for 08068500 Spring Creek near Spring, Tex.

Unit conversion to 15-minute load from streamflow and constituent concentration

The relation between an 15-minute load L, streamflow Q, and *Phos* concentration C is determined as follows

$$L\left[\frac{\text{kilogram}}{\text{day}}\right] = Q\left[\frac{\text{cubic feet}}{\text{second}}\right] \times C\left[\frac{\text{milligram}}{\text{liter}}\right] \times \underbrace{\left[\frac{\text{900 seconds}}{\text{15 minutes}}\right] \times \left[\frac{\text{28.317 liter}}{\text{cubic feet}}\right] \times \left[\frac{\text{kilogram}}{1 \times 10^6\ \text{milligram}}\right]}_{K\,=\,0.02549}$$

Computation of 15-minute load of total phosphorus

The 15-minute load L in kilograms per 15 minutes for streamflow $Q = 40$ cubic feet per second of total phosphorus with a concentration of $C = 0.8452$ milligrams per liter is computed by

$$L = Q \times C \times K$$
$$L = 40 \times 0.8452 \times 0.02549$$
$$L = 0.862 \text{ kilograms per 15 minutes}$$

Computation of 15-minute load of total phosphorus by using the R environment for statistical computing (R Development Core Team, 2010)

```
# Q is streamflow from code shown on previous page (40 cubic feet per second)
# C is the lower, estimate, and upper concentrations shown on previous page
#                (0.5779972, 0.8452142, 1.2359697)
K <- 0.02549 # unit conversion factor shown in previous hand derivation shown
    above
L <- Q * C * K # compute the 15-minute prediction and prediction limits
print(L) # show the 15-minute load along with prediction limits
      lwr       fit       upr
[1] 0.5893259 0.8617804 1.2601947
```

Thus, the 15-minute load of total phosphorus with 90-percent prediction limits is
$\boxed{0.589 \le L = 0.862 \le 1.26}$ kilograms per 15 minutes.

Lee and others—**Regression Model Development and Computational Procedures in Two Tributaries to Lake Houston, 2005–9**—Scientific Investigations Report 2012–5006

ISBN 978-1 4113 3368-0

Printed on recycled paper