USGS
science for a changing world

Prepared in cooperation with the Kansas Water Office, the City of Lawrence, the City of Topeka, the City of Olathe, and Johnson County Water One

Logistic and Linear Regression Model Documentation for Statistical Relations Between Continuous Real-Time and Discrete Water-Quality Constituents in the Kansas River, Kansas, July 2012 through June 2015

Open-File Report 2016–1040

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

Logistic and Linear Regression Model Documentation for Statistical Relations Between Continuous Real-Time and Discrete Water-Quality Constituents in the Kansas River, Kansas, July 2012 through June 2015

By Guy M. Foster and Jennifer L. Graham

Prepared in cooperation with the Kansas Water Office, the City of Lawrence, the City of Topeka, the City of Olathe, and Johnson County Water One

Open-File Report 2016–1040

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

U.S. Department of the Interior
SALLY JEWELL, Secretary

U.S. Geological Survey
Suzette M. Kimball, Director

U.S. Geological Survey, Reston, Virginia: 2016

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

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

Suggested citation:
Foster, G.M., and Graham, J.L., 2016, Logistic and linear regression model documentation for statistical relations between continuous real-time and discrete water-quality constituents in the Kansas River, Kansas, July 2012 through June 2015: U.S. Geological Survey Open-File Report 2016–1040, 27 p., http://dx.doi.org/10.3133/ofr20161040.

ISSN 2331-1258 (online)

Contents

Figures

Tables

Conversion Factors

[U.S. customary units to International System of Units]

Multiply	By	To obtain
	Volume	
liter (L)	33.82	ounce, fluid (fl. oz)
milliliter (mL)	0.0338	ounce, fluid (fl. oz)
	Flow rate	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
	Mass	
kilogram (kg)	2.204	pounds (lb)
microgram (µg)	0.000001	gram (g)
milligram (mg)	0.001	gram (g)
nanogram (ng)	0.000000001	gram (g)
	Length	
centimeter (cm)	0.03281	foot (ft)
meter (m)	3.281	foot (ft)
mile (mi)	1.609	kilometer (km)
micrometer (µm)	0.001	millimeter (mm)
nanometer (nm)	0.000001	millimeter (mm)
	Area	
square mile (mi^2)	2.590	square kilometer (km^2)
square meter (m^2)	10.76	square feet (ft^2)

Datum

Horizontal coordinate information is referenced to the North American Datum of 1927.

Supplemental Information

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), micrograms per liter (µg/L), or nanograms per liter (ng/L).

Logistic and Linear Regression Model Documentation for Statistical Relations Between Continuous Real-Time and Discrete Water-Quality Constituents in the Kansas River, Kansas, July 2012 through June 2015

By Guy M. Foster and Jennifer L. Graham

Abstract

The Kansas River is a primary source of drinking water for about 800,000 people in northeastern Kansas. Source-water supplies are treated by a combination of chemical and physical processes to remove contaminants before distribution. Advanced notification of changing water-quality conditions and cyanobacteria and associated toxin and taste-and-odor compounds provides drinking-water treatment facilities time to develop and implement adequate treatment strategies. The U.S. Geological Survey (USGS), in cooperation with the Kansas Water Office (funded in part through the Kansas State Water Plan Fund), and the City of Lawrence, the City of Topeka, the City of Olathe, and Johnson County Water One, began a study in July 2012 to develop statistical models at two Kansas River sites located upstream from drinking-water intakes. Continuous water-quality monitors have been operated and discrete-water quality samples have been collected on the Kansas River at Wamego (USGS site number 06887500) and De Soto (USGS site number 06892350) since July 2012. Continuous and discrete water-quality data collected during July 2012 through June 2015 were used to develop statistical models for constituents of interest at the Wamego and De Soto sites. Logistic models to continuously estimate the probability of occurrence above selected thresholds were developed for cyanobacteria, microcystin, and geosmin. Linear regression models to continuously estimate constituent concentrations were developed for major ions, dissolved solids, alkalinity, nutrients (nitrogen and phosphorus species), suspended sediment, indicator bacteria (*Escherichia coli*, fecal coliform, and enterococci), and actinomycetes bacteria. These models will be used to provide real-time estimates of the probability that cyanobacteria and associated compounds exceed thresholds and of the concentrations of other water-quality constituents in the Kansas River. The models documented in this report are useful for characterizing changes in water-quality conditions through time, characterizing potentially harmful cyanobacterial events, and indicating changes in water-quality conditions that may affect drinking-water treatment processes.

Introduction

Cyanobacteria (also called blue-green algae) cause a multitude of water-quality concerns, including the potential to produce toxins and taste-and-odor compounds. Toxins and taste-and-odor compounds may cause substantial economic and public health concerns and are of particular interest in lakes, reservoirs, and rivers that are used for drinking-water supply (Graham, 2006). Cyanobacterial toxins (cyanotoxins) have been implicated in human and animal illness and death in at least 36 states in the United States, including Kansas (Graham and others, 2009; Trevino-Garrison and others, 2015). Several countries have set national standards or guidelines for cyanotoxins in drinking water (Hudnell, 2008). The U.S. Environmental Protection Agency (EPA) recently (2015) released health advisory values for the cyanotoxins microcystin and cylindrospermopsin in finished drinking water. The 10-day health advisory values for microcystin in finished drinking water are 0.3 micrograms per liter (µg/L) for young children and 1.6 µg/L for all other ages; the health advisory values for cylindrospermopsin are 0.7 µg/L for young children and 3.0 µg/L for all other ages (U.S. Environmental Protection Agency, 2015). Unlike cyanotoxins, taste-and-odor compounds have no known effects on human health and there are no regulations or advisory values for these compounds. Aesthetic issues associated with taste and odor occur at low concentrations (5 to 10 nanograms per liter [ng/L]) and remedial actions commonly are implemented as soon as taste or odor is detected in a drinking-water supply (Taylor and others, 2005).

The Kansas River is a primary source of drinking water for about 800,000 people in northeastern Kansas (Graham and others, 2012). Cyanobacterial blooms typically do not develop in the Kansas River; however, reservoirs in or near the lower

Kansas River Basin (fig. 1) do occasionally develop blooms (Graham and others, 2012; Trevino-Garrison and others, 2015). The Kansas River has periodic taste-and-odor episodes that may be caused by either cyanobacterial production of taste-and-odor compounds in upstream reservoirs or actinomycetes bacteria production and transport during runoff events. Downstream transport of cyanobacteria and associated toxins and taste-and-odor compounds from reservoirs in the lower Kansas River Basin was documented during water releases from Milford, Tuttle Creek, and Perry reservoirs during September and October, 2011 (Graham and others, 2012).

Source-water supplies are treated by a combination of chemical and physical processes to remove contaminants before distribution. Water-quality conditions, such as turbidity, specific conductance, and pH, may require alteration of treatment processes to ensure effective removal of contaminants. An advanced notification system of changing water-quality conditions and cyanotoxin and taste-and-odor occurrences provides drinking-water treatment facilities time to develop and implement adequate treatment strategies.

Statistical models using discretely sampled and continuously measured physicochemical properties to estimate concentrations of constituents of concern may be used to provide advanced notification of changing water-quality conditions for constituents of interest (for example, Stone and others, 2013). The U.S. Geological Survey (USGS), in cooperation with the Kansas Water Office (funded in part through the Kansas Water Plan), the City of Lawrence, the City of Topeka, the City of Olathe, and Johnson County Water One began a study in July 2012 to develop models at two Kansas River sites located upstream from drinking-water intakes. Continuous water-quality monitors were operated and discrete-water quality samples collected on the Kansas River at the Wamego and De Soto sites (fig. 1) since July 2012. The study involved development of logistic regression models that continuously estimate the probability of cyanobacteria and associated compounds exceeding selected thresholds and linear regression models that continuously estimate concentrations of nutrients, sediment, and additional water-quality constituents of interest in the Kansas River.

Purpose and Scope

The purpose of this report is to document regression models that establish relations between continuous and discrete water-quality data collected from the Kansas River during July 2012 through June 2015 at the Wamego and De Soto sites (fig. 1). Logistic regression models that estimate the probability of occurrence above selected thresholds were developed for cyanobacteria, microcystin, and geosmin because the necessary assumptions for using linear approaches were not met for these constituents. Linear regression models were developed for major ions, nutrients, sediment, and fecal indicator bacteria. Logistic models provide real-time estimates of the probability that cyanobacteria and associated compounds occur above selected thresholds upstream from drinking-water supply intakes on the Kansas River. Linear models characterize changes in concentrations of other water-quality constituents of interest as well as water-quality conditions through time, potentially harmful cyanobacterial events, and changes in water-quality conditions that may affect drinking-water treatment processes.

Linear regression models for water-quality constituents for the Kansas River at the Wamego and De Soto sites, including major ions, nutrients, sediment, and indicator bacteria, originally were published by Rasmussen and others (2005) using data collected during 2000 through 2003. Because of the elapsed time between these previously published models and data in this report; updated analytical methods and sensor technology; and potential changes in watershed practices, water-quality conditions, and riverine processes through time; new models for the Wamego and De Soto sites were developed using data collected only during July 2012 through June 2015. Evaluating causes for changes in model form and potential causes for change are beyond the scope of this report.

Description of Study Area

The Kansas River Basin has an area of 60,097 square miles (mi^2) and includes most of the northern one-half of Kansas and parts of Nebraska and Colorado. The Kansas River is formed by the confluence of the Smoky Hill and Republican Rivers, near Junction City, Kansas, and flows about 173 miles east into the Missouri River (not shown), at Kansas City, Kansas (fig. 1). The area downstream from the Smoky Hill and Republican Rivers commonly is called the lower Kansas River Basin, which has an area of 5,448 mi^2 and forms the study area. Eighteen Federal reservoir projects impound water on all major tributaries of the Kansas River and control streamflow in 85 percent of the drainage area (Perry, 1994). The main stem of the Kansas River, however, has only minor control structures that include Bowersock Dam (in Lawrence, Kansas; not shown), a low-head hydroelectric dam, and several diversion weirs for water supply (Kansas Riverkeeper, 2012).

About 77 percent of land use in the study area is agricultural (cropland and grassland) and includes some urban areas (about 9 percent of land use; Fry and others, 2011). Although urban development represents a small part of total land use, major urban and industrial areas are located along the Kansas River at Manhattan, Topeka, Lawrence, and Kansas City, Kansas (fig. 1). All of these cities, in addition to many smaller communities, withdraw water from the Kansas River and associated alluvial aquifer for municipal water supply.

To characterize cyanotoxin and water-quality concentrations on the Kansas River, two sites were selected on the main stem (fig. 1). The Kansas River at Wamego (USGS site number 06887500) has a drainage area of 55,280 mi^2 and is 128 river miles upstream from the confluence of the

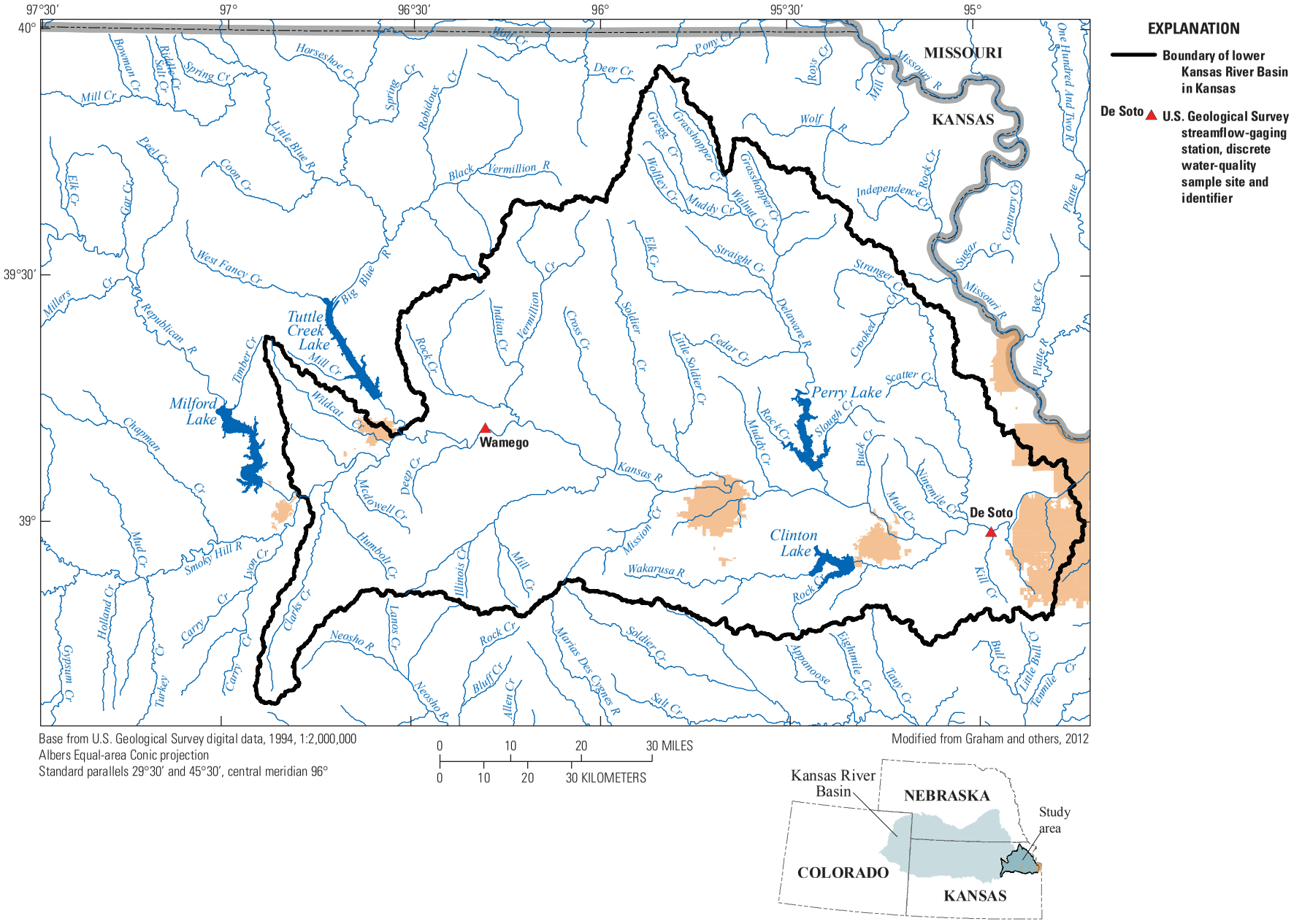

Figure 1. Location of reservoirs and the Wamego and De Soto streamflow-gaging stations and discrete water-quality sampling sites in the lower Kansas River Basin.

Kansas River with the Missouri River. Wamego is upstream from the major urban water suppliers in the cities of Topeka, Lawrence, Olathe, and Johnson County, and is downstream from the large federal reservoirs Milford Lake and Tuttle Creek Lake (fig. 1).The Kansas River at De Soto (USGS site number 06892350) site has a drainage area of 59,756 mi^2 and is 31 river miles upstream from the confluence of the Kansas River with the Missouri River and is upstream from water withdrawals in Johnson County. DeSoto is downstream from Topeka and Lawrence, as well as the large federal reservoirs Perry Lake and Clinton Lake (fig. 1).

Methods

Continuous and discrete water-quality data were collected at the Wamego and De Soto sites on the Kansas River. Water quality has been measured continuously and discrete water-quality samples have been routinely collected since July 2012. Continuous and discrete water-quality data collected by the USGS at the Wamego and De Soto sites over a range of streamflows during July 2012 through June 2015 (fig. 2) were used to develop site-specific regression models.

Continuous Water-Quality Monitoring

Continuous streamflow and water-quality data were collected from the Wamego and De Soto sites. Streamflow was measured using standard USGS methods (Sauer and Turnipseed, 2010; Turnipseed and Sauer, 2010). Both sites were equipped with YSI 6600 V2 water-quality monitors from July 2012 through June 2014, and then replaced by Xylem YSI EXO2 water-quality monitors from June 2014 through June 2015. Water-quality monitors measured water temperature, specific conductance, pH, dissolved oxygen, turbidity, and chlorophyll fluorescence. Data from these different monitors were linearly comparable, and there were no notable shifts in data quality between the two monitors; however, the YSI EXO2 turbidity sensor had a greater upper detection limit (4,000 formazin nephelometric units[FNU]) than the YSI 6600 V2 (1,000 FNU) (YSI Incorporated, 2015). Monitors were housed in 4-inch diameter polyvinyl chloride (PVC) or steel pipes with holes drilled to facilitate flow across sensors and were suspended from bridges approximately 1 to 2 feet (ft) below the water surface in the main flow zone of the river. Data were recorded every 15 minutes and are available from the USGS National Water Information System (NWIS) at http://dx.doi.org/10.5066/F7P55KJN.

Monitor maintenance and data reporting generally followed procedures described in Wagner and others (2006) with the exception of increased length between calibration checks (approximately 2–3 months) because of minimal sensor calibration drift. Monitors were cleaned approximately every 6 weeks or as needed. Continuous data during the study period required corrections of less than 10 percent, which classifies the data quality rating as good according to established guidelines (Wagner and others, 2006).

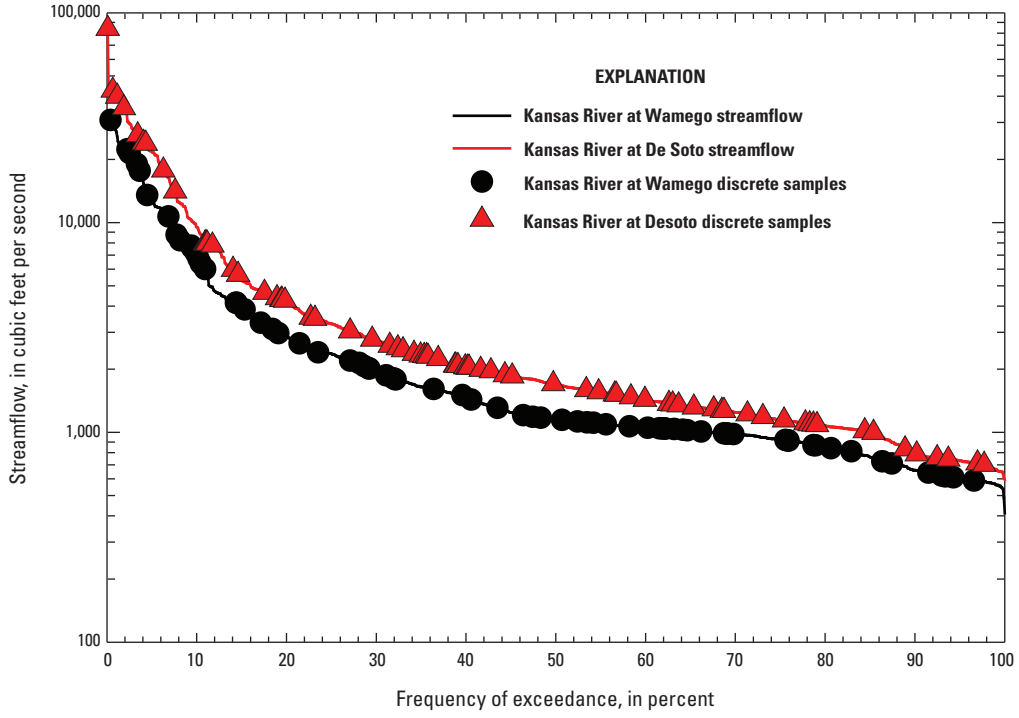

Figure 2. Streamflow and timing of discrete-sample collection for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.

Discrete Water-Quality Sampling

During July 2012 through June 2015, discrete water-quality samples were collected bi-weekly during May through October and monthly during November through April. Discrete water-quality samples were collected over a range of streamflow conditions using this fixed-schedule sampling regime (fig. 2). Most discrete samples were collected using depth- and width-integrating sample-collection techniques (Wilde, 2008; U.S. Geological Survey, 2006). Samples collected using this approach are representative of the average chemical and biological composition of the stream cross-sectional area. Bacteria samples were collected at the centroid of flow using a weighted basket in accordance with USGS protocols (U.S. Geological Survey, 2006). A weighted basket also was used to collect samples when ice or extreme cold made it impossible to collect samples using depth- and width-integrating techniques. All water samples were analyzed for cyanobacterial abundance, the cyanotoxin microcystin, the taste-and-odor compounds geosmin and 2-methylisoborneol (MIB), major ions (including dissolved solids), alkalinity, nutrients (nitrogen and phosphorus species), suspended sediment, indicator bacteria, and actinomycetes bacteria. Discrete water-quality data are available through the USGS National Water Information System (NWIS) at http://dx.doi.org/10.5066/F7P55KJN.

Phytoplankton samples (preserved with a 9:1 Lugol's iodine:acetic acid solution) to determine cyanobacterial abundance were analyzed for taxonomic identification and enumeration by BSA Environmental Services, Inc., Beachwood, Ohio, as described in Stone and others (2013). Geosmin and MIB were analyzed by Engineering Performance Solutions (EPS), LLC, Jacksonville, Florida, using solid phase microextraction gas chromatography/mass spectrometry (Zimmerman and others, 2002). Total microcystin was analyzed by the USGS Organic Geochemistry Research Laboratory, Lawrence, Kansas. All samples were lysed by three sequential freeze-thaw cycles and filtered using 0.7-micrometer (µm) glass-fiber filters before analysis for microcystin (Loftin and others, 2008). Abraxis® enzyme-linked immunosorbent assays (ELISA) were used to measure microcystin (congener independent).

Major ions (including dissolved solids), alkalinity, and nutrients were analyzed by the USGS National Water Quality Laboratory in Lakewood, Colorado, using methods described by Fishman and Friedman (1989). *Escherichia coli* (*E. coli*), fecal coliform, and enterococci bacteria were analyzed at the USGS Kansas Water Science Center using methods described by Wilde (2008). Suspended-sediment concentration was analyzed at the USGS Iowa Sediment Laboratory in Iowa City, Iowa, according to methods described in Guy (1969). Actinomycetes bacteria were analyzed by the USGS Ohio Water Science Center Microbiology Laboratory in Columbus, Ohio, using standard plate counts (American Public Health Association, American Water Works Association, and Water Environment Federation, 2010).

Quality-assurance and quality-control (QA/QC) samples were collected to evaluate variability in sample collection and processing techniques and among-laboratory variability in analytical techniques. About 10 percent of discrete water-quality samples were QA/QC samples. Eighteen concurrent or replicate constituent pairs were collected during July 2012 through June 2015. Relative percentage difference (RPD) was used to evaluate differences in analyte concentrations detected in replicate water samples. The RPD was calculated by dividing the difference between replicate pairs by the mean and multiplying that value by 100, thereby creating a value that represents the percent difference between replicate samples (Zar, 1999). Replicate pairs with an RPD within 10 percent were considered acceptable for inorganic constituents (Rasmussen and others, 2014) and replicate pairs with an RPD within 20 percent were considered acceptable for nutrient and organic constituents. All inorganic and nutrient constituent replicate pairs had median RPDs that were less than or equal to 2 percent, except for suspended sediment (7 percent). Microcystin, geosmin, and MIB replicate median RPDs were 10 percent or less. Replicate median RPD of bacteria samples ranged from 12 to 51 percent. Eleven concurrent replicates were collected for total cyanobacteria analysis. Median RPD for total cyanobacteria was 132 percent; large RPDs were generally caused by low cyanobacterial abundance relative to the overall community (less than 25 percent of the overall abundance). Large RPDs not affected by low cyanobacterial abundance likely were caused by the extreme spatial variability that may be present in cyanobacterial communities that create challenges when collecting and processing replicate samples (Graham and others, 2012). Quality-assurance data are available through the USGS National Water Information System (NWIS) at http://dx.doi.org/10.5066/F7P55KJN.

Equipment and field blank samples were collected to analyze for potential contaminants introduced by ambient conditions at the site or through the equipment used for sampling (U.S. Geological Survey, 2006). A total of 10 field and equipment blanks were collected from July 2012 to June 2015. All constituents of interest were determined to be below laboratory minimum reporting levels (microsystin, 0.1 ng/L; geosomin, 1.0 ng/L; MIB, 1.0 ng/L). Quality-assurance data, including laboratory minimum reporting levels, are available through the USGS National Water Information System (NWIS) at http://dx.doi.org/10.5066/F7P55KJN.

Development of Logistic Regression Models

Cyanobacteria, the cyanotoxin microcystin, and the taste-and-odor compounds geosmin and MIB were commonly detected at the Wamego and De Soto sites during July 2012 through June 2015; however, between 10 and 70 percent of samples were below laboratory minimum reporting levels depending on constituent and site. Because of the high percentage of censored values, ordinary least squares regression is not an appropriate modeling technique for these constituents

(Helsel and Hirsch, 2002). Multiple logistic regression was used to develop models to identify factors that best explained the probability of cyanobacteria, microcystin, geosmin, and MIB concentrations exceeding selected thresholds.

Logistic regression models the probability of the response variable being in one of two categorical response groups (for example, 0 equals a reference or negative response and 1 equals a positive response). Logistic regression transforms estimated probabilities into a continuous response variable, with possible values ranging from negative to positive infinity. The transformed response variable can then be modeled as a linear function of one or more explanatory variables in a logistic regression (Helsel and Hirsch, 2002). The logistic regression equation can be expressed as follows:

$$\ln\left(\frac{p}{1-p}\right) = b0 + bX \quad (1)$$

where

ln is the natural logarithm,

$\left(\frac{p}{1-p}\right)$ is the odds ratio, with p equal to the probability of a 1 (positive) response,

b0 is the intercept,

X is a vector of k explanatory variables, and

bX includes the slope coefficients for each explanatory variable so that $bX=b1X1+b2X2\ldots bkXk$.

In this form, the logistic regression models the probability of obtaining a 1 (positive) response (Helsel and Hirsch, 2002; Systat Software, Inc., 2008). Model output is the natural logarithm of the odds ratio. The natural logarithm of the odds ratio can be converted into a probability using the following equation:

$$p = e^{b0+bX} / \left[1 + e^{b0+bX}\right] \quad (2)$$

where

p is the probability of a response of 1,

e is the base of the natural logarithm (approximately equal to 2.71828),

$b0$ is the intercept,

X is a vector of k explanatory variables, and

bX includes the slope coefficients for each explanatory variable so that $bX=b1X1+b2X2\ldots bkXk$.

Cyanobacteria frequently are present without the occurrence of microcystin, geosmin or MIB. Therefore, to select thresholds for cyanobacteria, bivariate plots of cyanobacteria and microcystin, geosmin, and MIB were inspected for each site to determine if there were cyanobacterial abundances above which relatively high concentrations of these compounds occurred. The only compound with clear breakpoints in the data was microcystin; microcystin concentrations only exceeded the 10-day health advisory value of 0.3 µg/L for finished drinking water (U.S. Environmental Protection Agency, 2015) when cyanobacterial abundances were greater than 2,000 and 10,000 cells/mL at the Wamego and De Soto sites, respectively. Therefore, at the Wamego site, a categorical response value of 1 (positive response) was assigned to samples with cyanobacterial abundances greater than 2,000 cells/mL and 0 (negative response) was assigned to samples in which abundance was less than or equal to 2,000 cells/mL; at the De Soto site a cyanobacterial abundance of 10,000 cells/mL was used as the threshold for categorical response.

The EPA 10-day health advisory values for microcystin in finished drinking water are 0.3 µg/L for young children and 1.6 µg/L for all other ages (U.S. Environmental Protection Agency, 2015). At the Wamego (n=59) and De Soto (n=60) sites, 8 percent of samples exceeded 0.3 µg/L and 2 percent of samples exceeded 1.6 µg/L. Because microcystin concentrations rarely exceeded the advisory values, they could not be used as meaningful thresholds for logistic model development. Therefore, for microcystin model development, a categorical response value of 1 was assigned to samples with concentrations greater than or equal to the laboratory minimum reporting level (0.1 µg/L) and 0 was assigned to samples in which microcystin was not detected. Using the laboratory minimum reporting level for model development will provide an indicator of when microcystin may be detected before concentrations reach the health advisory levels for finished drinking water.

Geosmin and MIB concentrations greater than the human detection threshold of 5.0 ng/L (Taylor and others, 2005) did not occur frequently enough to develop meaningful models using this as a categorical response. Geosmin concentrations exceeded 5 ng/L in about 11 percent of samples at Wamego (n=59) and 18 percent of samples at the De Soto (n=61) sites. Concentrations of MIB exceeded 5 ng/L in about 7 and 5 percent of samples collected at the Wamego and De Soto sites, respectively. Geosmin and MIB model development assigned a categorical response value of 1 to samples with concentrations greater than to 2.0 ng/L and 0 was assigned to samples in which concentrations were equal to or less than 2.0 ng/L. The 2.0 ng/L threshold was selected because it was greater than the laboratory minimum reporting level for both, (1.0 ng/L) and created categories that allowed model development. In addition, using the 2.0 ng/L threshold for model development will provide an indicator of when geosmin and MIB may be detected before concentrations reach the human detection threshold.

Explanatory variables available as inputs to the multiple logistic regression analyses for the entire study period were specific conductance, pH, water temperature, dissolved oxygen, turbidity, chlorophyll fluorescence, and streamflow. Seasonal components (sine and cosine variables) were used as explanatory variables to determine if seasonal changes affected the model. All combinations of physicochemical properties and a seasonal component were evaluated to determine which combinations produced the best models.

Logistic model equations were developed using the multiple logistic regression routine in SigmaPlot® version 11.0 (Systat Software, Inc., 2008). Explanatory variables were evaluated individually and in selected combinations. Model combinations and the final best model were selected based on the statistical tests described in Stone and others (2013) in the following order: Pearson Chi-Square Statistic, Likelihood Ratio Test statistic, Hosmer-Lemeshow Statistic (Hosmer and others, 2013), and the -2 log likelihood ratio (see the model archive summaries, appendixes 1-6). Variance inflation factors and Wald Statistic *p*-values (Menard, 2002) were used to evaluate the redundancy of multiple explanatory variables included in the models and the association between explanatory and dependent variables. Model simplicity also was considered for model selection because as more variables are included the greater likelihood that the variability of the system is not described by the sampling dataset increases. A model classification table with a threshold probability for positive classification (TPPC) of 0.5 was also used in final model selection. A model classification table places dependent variable data into one of four categories: positive response predicted as positive (true positive; model sensitivity), reference response predicted as reference (true negative; model specificity), positive response predicted as reference (false negative), and reference response predicted as positive (false positive) (Systat Software, Inc., 2008). A model was considered suitable for use for constituent probability computations if the model properly classified 65 percent or more of the sample data as positive or reference, and the positively classified data included the highest measured concentrations. After the best model was selected, the TPPC for the model was adjusted to maximize the number of samples classified as positive to make the model more conservative (more likely to give a false positive than a false negative) by guarding more strongly against false negatives. The regression then used the newly adjusted thresholds, which changed the number of sample data classified as positive and reference, but the model constants and other statistical outputs remained the same.

Development of Linear Regression Models

Models were developed using simple linear (ordinary least squares) regression analyses (Helsel and Hirsch, 2002) to relate discrete sample concentrations or densities of water-quality constituents to continuously measured water-quality physicochemical properties (Rasmussen and others, 2009). Linear regression models were not developed if more than 10 percent of the data for a given constituent was censored (that is, ammonia and total suspended solids at the Wamego and De Soto sites, nitrate plus nitrite at the De Soto site). Concomitant in-situ continuous measurements were used to correspond with discrete measurements as described in Rasmussen and others (2009). Comparisons of cross-sectional averages and continuously measured data were within 8 percent or less for all final explanatory variables except chlorophyll fluorescence, which was 20 percent. Cross-sectional averages from the field water-quality monitors were used when concomitant continuously measured data were not available (for example, were deleted due to excessive fouling or equipment malfunction). Methods used for model development, quantifying uncertainty, and identifying and removing outliers are described in detail in Rasmussen and others (2009). Outliers that were removed from analysis are identified in the model archive summary for each model (appendixes 7–31).

Linear regression models were developed using R 3.2.0 (R Core Team, 2015). Explanatory variables available as inputs to linear regression for the period of record were those physicochemical properties that were used in the logistic models: specific conductance, pH, water temperature, dissolved oxygen, turbidity, chlorophyll fluorescence, and streamflow. Seasonal components (sine and cosine variables) were used as explanatory variables to determine if seasonal changes affected the model. All combinations of physicochemical properties and a seasonal component were evaluated to determine which combinations produced the best models.

Linear regression models were evaluated based on diagnostic statistics (R^2, coefficient of determination; Mallow's *C*p; *RMSE*, root mean square error; *PRESS*, prediction error sum of squares), patterns in residual plots, and the range and distribution of discrete and continuous data (Helsel and Hirsch, 2002). The best model for each constituent was selected to maximize the amount of variance in the response variable explained by the model (multiple R^2 for models with one explanatory variable and adjusted R^2 for models with more than one explanatory variable), fit the data (Mallow's *C*p), and minimize heteroscedasticity (irregular scatter) in the residual plots and uncertainty associated with computed values (*RMSE* and *PRESS*). Variance inflation factor was used to measure the exact or approximate linear relation between variables (collinearity; Marquardt, 1970). Model simplicity also was considered for model selection because as more variables are included the likelihood that the variability of the system is not described by the sampling dataset increases. Significant (*p*-value less than 0.05) additional explanatory variables were included in final models if retaining them increased the amount of variance explained by the model by 10 percent or more, decreased Mallow's *C*p, and minimized heteroscedasticity in residual plots. Models were considered suitable to use for constituent concentration computations if the amount of variance explained by the models (R^2) was 0.55 or greater.

Mean square error (*MSE*) and *RMSE* were calculated for each model to assess the variance between computed and observed values (Helsel and Hirsch, 2002). The model standard percentage error (MSPE) was calculated as a percentage of the RMSE (Hardison, 1969). Because transformation of estimates back into original units results in a low biased estimate (Helsel and Hirsch, 2002), a bias correction factor (BCF) was calculated for models with logarithmically transformed response variables (Duan, 1983).

Results of Logistic Regression Analysis for Cyanobacteria and Associated Compounds

Logistic regression models that estimate probability of occurrence above selected thresholds were successfully developed for cyanobacteria, microcystin, and geosmin at the Wamego and De Soto sites. Final models are presented in table 1 (at the back of this report). Statistical model output and model datasets are presented in appendixes 1–6.

Cyanobacteria

Cyanobacteria were present in 66 percent (*n*=56) of samples collected at the Wamego site during July 2012 through June 2015 and concentrations ranged from 0 to 69,000 cells/mL (median: 280 cells/mL). Cyanobacteria were measured above the 2,000 cells/mL threshold used for logistic model development in 28 percent of samples. The best fit model at the Wamego site for probability of cyanobacteria occurrence at concentrations more than the 2,000 cells/mL threshold included a seasonal component and turbidity, likely because cyanobacteria occurrence at the Wamego site has seasonal patterns and are also affected by light. The threshold of the cyanobacteria model for the Wamego site was reset from 0.50 to 0.31. The final logistic model correctly estimated the likelihood of cyanobacteria abundance exceeding 2,000 cells/mL 75 percent of the time and not exceeding the detection threshold 82 percent of the time, resulting in an overall sensitivity of 80 percent (table 1, appendix 1).

Cyanobacteria were present in 67 percent (*n*=57) of samples collected at the De Soto site during July 2012 through June 2015 and concentrations ranged from 0 to 295,000 cells/mL (median: 340 cells/mL). Cyanobacteria occurred at more than the 10,000 cells/mL threshold used for logistic model development in 23 percent of samples. The best fit model at the De Soto site for cyanobacteria occurrence at concentrations more than the 10,000 cells/mL threshold included a seasonal component and turbidity, likely because cyanobacteria occurrence at the De Soto site has seasonal patterns and are also affected by light. The threshold of the cyanobacteria model for the De Soto site was reset from 0.50 to 0.48. The final logistic model correctly estimated the likelihood of cyanobacteria abundance exceeding 10,000 cells/mL 85 percent of the time and not exceeding the detection threshold 93 percent of the time, resulting in an overall sensitivity of 91 percent (table 1, appendix 2).

Microcystin

Microcystin was detected in 22 percent (*n*=59) of samples collected at the Wamego site during July 2012 through June 2015 and concentrations ranged from <0.1 to 1.7 µg/L (median: <0.1 µg/L). The best fit model at the Wamego site for microcystin occurrence at concentrations more than the 0.1 µg/L detection threshold included a seasonal component, streamflow, and turbidity as explanatory variables, likely because microcystin occurrences at the Wamego site have seasonal patterns mediated by streamflow as affected by reservoir releases. The threshold of the microcystin model for the Wamego site was reset from 0.50 to 0.40. The final logistic model correctly estimated the likelihood of microcystin concentrations exceeding the detection threshold 77 percent of the time and not exceeding the detection threshold 87 percent of the time, resulting in an overall sensitivity of 85 percent (table 1, appendix 3).

Microcystin was detected in 27 percent (*n*=60) of samples collected at the De Soto site during July 2012 through June 2015 and concentrations ranged from <0.1 to 2.4 µg/L (median: <0.1 µg/L). The best fit model at the De Soto site for microcystin occurrence at concentrations more than 0.1 µg/L included a seasonal component and streamflow as explanatory variables, likely because microcystin occurrences at the De Soto site have seasonal patterns mediated by streamflow as affected by reservoir releases. The threshold of the microcystin model for the De Soto site was reset from 0.50 to 0.36. The final logistic model correctly estimated the likelihood of microcystin concentrations exceeding the detection threshold 75 percent of the time and not exceeding the detection threshold 80 percent of the time, resulting in an overall sensitivity of 78 percent (table 1, appendix 4).

Geosmin

Geosmin was detected (>1.0 ng/L) in 78 percent (*n*=59) of samples collected at the Wamego site during July 2012 through June 2015 and concentrations ranged from <1.0 to 16 ng/L (median: 2.1 ng/L). Geosmin occurred above the 2.0 ng/L threshold used for logistic model development in 52 percent of samples. The best fit model at the Wamego site for geosmin occurrence at concentrations more than 2.0 ng/L included a seasonal component and chlorophyll fluorescence as explanatory variables, likely because geosmin occurrences at the Wamego site have seasonal patterns mediated by algal community composition and abundance. The threshold of the geosmin model for the Wamego site was reset from 0.50 to 0.53. The final logistic model correctly estimated the likelihood of geosmin concentrations exceeding the 2 ng/L threshold 71 percent of the time and not exceeding the threshold 75 percent of the time, resulting in an overall sensitivity of 73 percent (table 1, appendix 5).

Geosmin was detected (>1.0 ng/L) in 88 percent (*n*=61) of samples collected at the De Soto site during July 2012 through June 2015 and concentrations ranged from <1.0 to 42 ng/L (median: 2.3 ng/L). Geosmin occurred above the 2.0 ng/L threshold used for logistic model development in 56 percent of samples. The best fit model at the De Soto site for geosmin occurrence at concentrations above 2.0 ng/L

included a seasonal component and streamflow as explanatory variables, likely because geosmin occurrences at the De Soto site have seasonal patterns mediated by streamflow as affected by reservoir releases. The threshold of the geosmin model for the De Soto site did not need to be changed from 0.50. The final logistic model correctly estimated the likelihood of geosmin concentrations exceeding the 2.0 ng/L threshold 76 percent of the time and not exceeding the threshold 74 percent of the time, resulting in an overall sensitivity of 75 percent (table 1, appendix 6).

MIB

MIB was detected (>1.0 ng/L) in 47 percent (n=59) of samples collected at the Wamego site during July 2012 through June 2015 and concentrations ranged from <1.0 to 11 ng/L (median: <1.0 ng/L). At the De Soto site, MIB was detected in 33 percent (n=61) of samples and concentrations ranged from <1.0 to 26 ng/L (median: <1.0 ng/L). MIB occurred above the 2.0 ng/L threshold used for logistic model development in 34 percent of the Wamego site samples and 15 percent of the De Soto site samples. Logistic models for MIB that met model selection criteria were not successfully developed. At the Wamego site, there were no models that had statistically significant explanatory variables (Wald Statistic) and the highest MIB values were incorrectly classified. At the De Soto site, models properly classified less than 65 percent of sample data at a TPPC of 0.50. In addition, all explanatory variables show strong diurnal patterns, which may lead to artificial variability in the probability of occurrence throughout a 24-hour period.

Results of Linear Regression Analysis for Selected Constituents

Linear regression models for 15 constituents were developed for constituent concentration computations at the Wamego and De Soto sites. Modeled constituents included major ions and dissolved solids, nutrients (nitrogen and phosphorus species), suspended sediment, indicator bacteria (*E. coli*, fecal coliform, and enterococci), and actinomycetes bacteria. Final models are presented in table 2 (at the back of this report). Statistical model output and model datasets are presented in appendixes 7–31.

Specific conductance was the single explanatory variable for dissolved solids, major ions, and alkalinity models at both study sites (table 2; appendixes 7–20). All constituents were positively related to specific conductance and models explained between 65 and 98 percent of the variance in constituent concentrations. Explanatory variables for nutrient models varied by site and constituent, and included specific conductance, chlorophyll fluorescence, turbidity, temperature, and a seasonal component (table 2; appendixes 21–24). Nutrient models explained between 57 and 85 percent of the variance in concentrations. Neither a total phosphorus or nitrate+nitrite model that met selection criteria could be developed for the De Soto site. The variability in nutrient models by site and constituent likely reflects the complex range of hydrological and biogeochemical processes than influence nutrient concentrations in the Kansas River. Turbidity was the single explanatory variable for suspended-sediment concentrations at both study sites (table 2; appendixes 25–26). Suspended sediment was positively related to turbidity and models explained between 78 and 84 percent of the variance in concentrations. Turbidity and seasonality were explanatory variables for bacteria concentrations at both study sites (table 2; appendixes 27–31). Bacteria models explained between 64 and 84 percent of the variance in concentrations. Models for *E. coli*, fecal coliform, and enterococci bacteria that met model selection criteria could not be developed for the Wamego site, likely because of the relatively narrow range of turbidity conditions sampled as part of the fixed-schedule sampling regime.

Summary

Cyanobacteria cause a multitude of water-quality concerns, including the potential to produce toxins and taste-and-odor compounds. Toxins and taste-and-odor compounds are of particular interest in lakes, reservoirs, and rivers that are used for drinking-water supply. The Kansas River is a primary source of drinking water for about 800,000 people in northeastern Kansas. Before distribution, source-water supplies are treated by a combination of chemical and physical processes to remove contaminants. An advanced notification system of changing water-quality conditions and cyanotoxin and taste-and-odor occurrences would allow drinking-water treatment facilities time to develop and implement adequate treatment strategies. Surrogate models using discretely sampled and continuously measured physicochemical properties to estimate concentrations of constituents of concern that are not easily measured in real time may be used to provide advanced notification. In July 2012 the U.S. Geological Survey (USGS), in cooperation with the Kansas Water Office (funded in part through the Kansas State Water Plan Fund), the City of Lawrence, the City of Topeka, the City of Olathe, and Johnson County Water One began a study to develop models at two Kansas River sites located upstream from drinking-water intakes.

Regression models were developed that establish relations between continuous and discrete water-quality data collected from the Kansas River during July 2012 through June 2015 at the Wamego and De Soto sites, Kansas. Logistic regression models that estimate the probability of occurrence above selected thresholds were developed for cyanobacteria, microcystin, and geosmin because the necessary assumptions for using linear approaches were not met for these constituents. Linear regression models were developed for major ions,

nutrients, sediment, and bacteria. These models will be used to provide real-time estimates of the probability of occurrence of cyanobacteria and associated compounds and concentrations of other water-quality constituents upstream from drinking-water supply intakes on the Kansas River. These models are useful for characterizing changes in water-quality conditions through time, characterizing potentially harmful cyanobacterial events, and indicating changes in water-quality conditions that may affect drinking-water treatment processes.

References Cited

American Public Health Association, American Water Works Association , and Water Environment Federation, 2010, Standard methods for examination of water and wastewater: American Public Health Association, part 9250, Detection of Actinomycetes.

Duan, N., 1983, Smearing estimate a nonparametric retransformation method: Journal of the American Statistical Association, v. 78, no. 383, p. 605–610.

Fishman, M.J., and Friedman, L.C., 1989, Methods for determination of inorganic substances in water and fluvial sediments: U.S. Geological Survey Techniques of Water-Resources Investigations, book 5, chap. A1, 545 p.

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., and Wickham, J., 2011, Completion of the 2006 national land cover database for the conterminous United States: Photogrammetric Engineering & Remote Sensing, v. 77, p. 858–864.

Graham, J.L., 2006, Harmful algal blooms: U. S. Geological Survey Fact Sheet 2006–3147, 2 p.

Graham, J.L., Loftin, K.A., and Kamman, N., 2009, Monitoring recreational freshwaters: LakeLine, v. 29, p. 16–22.

Graham, J.L., Ziegler, A.C., Loving, B.L., and Loftin K.A., 2012, Fate and transport of cyanobacteria and associated toxins and taste-and-odor compounds from upstream reservoir releases in the Kansas River, Kansas, September and October 2011: U.S. Geological Survey Scientific Investigations Report 2012–5120, 65 p.

Guy, H.P., 1969, Laboratory theory and methods for sediment analysis: U.S. Geological Survey Techniques of Water-Resources Investigations, book 5, chap. C1, 58 p.

Hardison, C.H., 1969, Accuracy of streamflow characteristics, *in* Geological Survey Research, 1969: U.S. Geological Survey Professional Paper 650–D, p. D210–D214.

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources—Hydrologic analysis and interpretation: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 510 p.

Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X., 2013, Applied logistic regression (3rd ed.): John Wiley and Sons, Inc., Hoboken, New Jersey, 528 p.

Hudnell, H.K., ed., 2008, Cyanobacterial harmful algal blooms—State of the Science and Research Needs: Advances in Experimental Medicine and Biology, v. 619, 950 p.

Kansas Riverkeeper, 2012, Kansas River Inventory: accessed June 18, 2012, at http://www.kansasriver.org/.

Loftin, K.A., Meyer, M.T., Rubio, F., Kamp, L., Humphries, E., and Whereat, E., 2008, Comparison of two cell lysis procedures for recovery of microcystins in water samples from Silver Lake in Dover, Delaware, with microcystin producing cyanobacterial accumulations: U.S. Geological Survey Open-File Report 2008–1341, 9 p.

Marquardt, D.W., 1970, Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation: Technometrics, v. 12, p. 591–612.

Menard, S.W., 2002, Applied logistic regression analysis (2nd ed.): Thousand Oaks, California, Sage Publications, Inc., 119 p.

Perry, C.A., 1994, Effects of reservoirs on flood discharges in Kansas and the Missouri River Basins, 1993: U.S. Geological Survey Circular 1120–E, 20 p.

R Core Team, 2015, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing. [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, 53 p.

Rasmussen, T.J., Bennett, Trudy, Stone, M.L., Foster, G.M., Graham, J.L., Putnam, J.E., 2014, Quality-assurance and data-management plan for water-quality activities in the Kansas Water Science Center, 2014: U.S. Geological Survey Open-File Report 2014–1233, 41 p.

Rasmussen, T.J., Ziegler, A.C., and Rasmussen, P.P., 2005, Estimation of constituent concentrations, densities, loads, and yields in lower Kansas River, northeast Kansas, using regression models and continuous water-quality monitoring, January 2000 through December 2003: U.S. Geological Survey Scientific Investigations Report 2005–5165, 117 p.

Sauer, V.B., and Turnipseed, D.P., 2010, Stage measurement at gaging stations: U.S. Geological Survey Techniques and Methods, book 3, chap. A7, 45 p. [Also available at http://pubs.usgs.gov/tm/tm3-a7/.)

Stone, M.L., Graham, J.L., and Gatotho, J.W., 2013, Model documentation for relations between continuous real-time and discrete water-quality constituents in Cheney Reservoir near Cheney, Kansas, 2001–2009: U.S. Geological Survey Open-File Report 2013–1123, 100 p. [Also avaialble at http://pubs.usgs.gov/of/2013/1123/.]

Systat Software, Inc., 2008, SigmaPlot® 11.0 Statistics User's Guide, 564 p.

Taylor, W.D., Losee, R.F., Torobin, M., Izaguirre, G., Sass, D., Khiari, D., and Atasi, K., 2005, Early warning and management of surface water taste-and-odor events: American Water Works Association Research Foundation, 373 p.

Trevino-Garrison, I., DeMent, J., Ahmed, F.S., Haines-Lieber, P., Langer, T., Menager, H., Neff, J., van der Merwe, D., Carney, E., 2015, Human illnesses and animal deaths associated with freshwater harmful algal blooms–Kansas: Toxins, v. 7, p. 353–366.

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. [Also available at http://pubs.usgs.gov/tm/tm3-a8/.)

U.S. Environmental Protection Agency, 2015, Drinking water health advisory for the cyanobacterial microcystin toxins: Washington D.C., U.S. Environmental Protection Agency, Office of Water, EPA 820R15100, 67 p., accessed December 2015 at http://www.epa.gov/sites/production/files/2015-06/documents/microcystins-report-2015.pdf.

U.S. Geological Survey, 2006, Collection of water samples (ver. 2.0): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A4, 166 p. [Also available at http://pubs.water.usgs.gov/twri9A4/].

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, book 1, chap. D3, 51 p., plus 8 attachments, accessed April 2012, at http://pubs.water.usgs.gov/tm1d3.

Wilde, F.D., ed., 2008, Field measurements, *in* National field manual for the collection of water-quality data: U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A6, p. 3–20.

YSI Incorporated, 2015, YSI 6136 turbidity sensor documentation: YSI product documentation, accessed November 20, 2015, at https://www.ysi.com/File%20Library/Documents/Specification%20Sheets/E56-6136-Turbidity-Sensor.pdf.

Zar, J.H., 1999, Biostatistical analysis (4th ed): New Jersey, Prentice-Hall Inc., 663 p.

Zimmerman, L.R., Ziegler, A.C., and Thurman, E.M., 2002, Method of analysis and quality-assurance practices by U.S. Geological Survey Organic Geochemistry Research Group—Determination of geosmin and methylisoborneol in water using solid-phase microextraction and gas chromatograhphy/mass spectrometery: U.S. Geological Survey Open-File Report 02–337, 12 p.

Tables

Table 1. Best fit multiple logistic regression models and summary statistics for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.

[Logit *P*, logistic probability of presence; sin, sine; *D*, day of year; cos, cosine; *Turb*, turbidity in formazin nephelometric units; <, less than; *Chl*, fluorescence at wavelength of 650 to 700 nanometers in micrograms per liter as chlorophyll; *Q*, streamflow in cubic feet per second]

Site	Multiple logistic regression equation	Model archival summary	Threshold probability for positive classification	Response information					
				Classification table	Predicted reference responses	Predicted positive responses	Diagnostic	Total actual responses	Percent correctly classified responses
Cyanobacteria (Cyano)									
Wamego	Logit $P = 0.0887 - 0.875\sin(2\pi D/365) - 1.914\cos(2\pi D/365) - 0.0319(Turb)$	Appendix 1	0.31	Actual reference responses	33	7	Specificity	40	82
				Actual positive responses	4	12	Sensitivity	16	75
				Totals	37	19	Overall	56	80
De Soto	Logit $P = 0.724 - 1.254\sin(2\pi D/365) - 3.725\cos(2\pi D/365) - 0.0928(Turb)$	Appendix 2	0.48	Actual reference responses	41	3	Specificity	44	93
				Actual positive responses	2	11	Sensitivity	13	85
				Totals	43	14	Overall	57	91
Microcystin (MC)									
Wamego	Logit $P = -0.872 - 1.716\sin(2\pi D/365) - 1.313\cos(2\pi D/365) + 0.000349(Q) - 0.0490(Turb)$	Appendix 3	0.40	Actual reference responses	40	6	Specificity	46	87
				Actual positive responses	3	10	Sensitivity	13	77
				Totals	43	16	Overall	59	85
De Soto	Logit $P = -1.021 - 1.141\sin(2\pi D/365) - 0.824\cos(2\pi D/365) - 0.000115(Q)$	Appendix 4	0.36	Actual reference responses	35	9	Specificity	44	80
				Actual positive responses	4	12	Sensitivity	16	75
				Totals	39	21	Overall	60	78
Geosmin (GEO)									
Wamego	Logit $P = 1.325 + 0.830\sin(2\pi D/365) + 1.219\cos(2\pi D/365) - 0.0527(Chl)$	Appendix 5	0.53	Actual reference responses	21	7	Specificity	28	75
				Actual positive responses	9	22	Sensitivity	31	71
				Totals	30	29	Overall	59	73
De Soto	Logit $P = 0.236 + 0.585\sin(2\pi D/365) + 1.084\cos(2\pi D/365) + 0.0000473(Q)$	Appendix 6	0.5	Actual reference responses	20	7	Specificity	27	74
				Actual positive responses	8	26	Sensitivity	34	76
				Totals	28	33	Overall	61	75

Table 1. Best fit multiple logistic regression models and summary statistics for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[Logit *P*, logistic probability of presence; sin, sine; *D*, day of year; cos, cosine; *Turb*, turbidity in formazin nephelometric units; <, less than; Chl, fluorescence at wavelength of 650 to 700 nanometers in micrograms per liter as chlorophyll; *Q*, streamflow in cubic feet per second]

Site	Multiple logistic regression equation	Model archival summary	Pearson chi-square statistic (*p*-value)	Likelihood ratio test statistic (*p*-value)	-2 log likelihood	Hosmer-Lemshow statistic (*p*-value)
		Cyanobacteria (Cyano)				
Wamego	Logit $P = 0.0887 - 0.875\sin(2\pi D/365) - 1.914\cos(2\pi D/365) - 0.0319(Turb)$	Appendix 1	57.812	19.962	47.044	8.164
			(0.238)	(<0.001)		(0.418)
De Soto	Logit $P = 0.724 - 1.254\sin(2\pi D/365) - 3.725\cos(2\pi D/365) - 0.0928(Turb)$	Appendix 2	37.949	33.185	28.025	2.415
			(0.928)	(<0.001)		(0.966)
		Microcystin (MC)				
Wamego	Logit $P = -0.872 - 1.716\sin(2\pi D/365) - 1.313\cos(2\pi D/365) + 0.000349(Q) - 0.0490(Turb)$	Appendix 3	46.475	20.411	41.814	6.534
			(0.725)	(<0.001)		(0.588)
De Soto	Logit $P = -1.021 - 1.141\sin(2\pi D/365) - 0.824\cos(2\pi D/365) - 0.000115(Q)$	Appendix 4	61.82	11.186	58.404	10.619
			(0.246)	(0.011)		(0.224)
		Geosmin (GEO)				
Wamego	Logit $P = 1.325 + 0.830\sin(2\pi D/365) + 1.219\cos(2\pi D/365) - 0.0527(Chl)$	Appendix 5	56.577	19.603	62.036	9.979
			(0.379)	(<0.001)		(0.266)
De Soto	Logit $P = 0.236 + 0.585\sin(2\pi D/365) + 1.084\cos(2\pi D/365) + 0.0000473(Q)$	Appendix 6	60.455	10.502	73.257	12.503
			(0.318)	(0.015)		(0.130)

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Percent censored data	Multiple R^2	Adjusted R^2	*MSE*	*RMSE*	*MSPE* (upper)	*MSPE* (lower)	Bias correction factor (Duan, 1983)
Total dissolved solids (TDS), mg/L										
Wamego	$\log(TDS) = 0.911 \times \log(SC) + 0.0548$	Appendix 7	0	0.977	0.976	0.00065	0.0255	6.04	5.7	1
DeSoto	$\log(TDS) = 0.938 \times \log(SC) - 0.0345$	Appendix 8	0	0.960	0.959	0.00106	0.0325	7.77	7.21	1
Calcium (Ca), dissolved, mg/L										
Wamego	$\log(Ca) = 0.646 \times \log(SC) - 0.0258$	Appendix 9	0	0.925	0.923	0.00112	0.0334	7.98	7.39	1
DeSoto	$\log(Ca) = 0.645 \times \log(SC) - 0.03$	Appendix 10	0	0.872	0.870	0.00174	0.0417	10.1	9.16	1
Magnesium (Mg), dissolved, mg/L										
Wamego	$\log(Mg) = 0.768 \times \log(SC) - 0.996$	Appendix 11	0	0.920	0.919	0.00166	0.0408	9.86	8.98	1
DeSoto	$\log(Mg) = 0.845 \times \log(SC) - 1.22$	Appendix 12	0	0.952	0.951	0.00102	0.0319	7.62	7.08	1
Sodium (Na), dissolved, mg/L										
Wamego	$\log(Na) = 1.52 \times \log(SC) - 2.55$	Appendix 13	0	0.965	0.965	0.00271	0.0521	12.7	11.3	1.01
DeSoto	$\log(Na) = 1.72 \times \log(SC) - 3.13$	Appendix 14	0	0.961	0.961	0.00338	0.0581	14.3	12.5	1.01

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Discrete data			
			n	Range of values in variable measurements	Mean	Median
Total dissolved solids (TDS), mg/L						
Wamego	$\log(TDS) = 0.911 \times \log(SC) + 0.0548$	Appendix 7	55	DS: 156–784	494	522
				SC: 260–1,280	793	829
DeSoto	$\log(TDS) = 0.938 \times \log(SC) - 0.0345$	Appendix 8	59	DS: 138–645	446	470
				SC: 197–1,110	726	770
Calcium (Ca), dissolved, mg/L						
Wamego	$\log(Ca) = 0.646 \times \log(SC) - 0.0258$	Appendix 9	55	Ca: 32.1–99.0	69.7	72.6
				SC: 260–1,280	793	829
DeSoto	$\log(Ca) = 0.645 \times \log(SC) - 0.03$	Appendix 10	59	Ca: 29.2–95.3	65	67.9
				SC: 197–1,110	726	770
Magnesium (Mg), dissolved, mg/L						
Wamego	$\log(Mg) = 0.768 \times \log(SC) - 0.996$	Appendix 11	55	Mg: 5.92–26.5	16.8	17.9
				SC: 260–1,280	793	829
DeSoto	$\log(Mg) = 0.845 \times \log(SC) - 1.22$	Appendix 12	59	Mg: 5.14–23.1	15.8	17
				SC: 197–1,110	726	770
Sodium (Na), dissolved, mg/L						
Wamego	$\log(Na) = 1.52 \times \log(SC) - 2.55$	Appendix 13	55	Na: 9.79–134	74.7	80.6
				SC: 260–1,280	793	829
DeSoto	$\log(Na) = 1.72 \times \log(SC) - 3.13$	Appendix 14	59	Na: 5.29–118	63.6	68.3
				SC: 197–1,110	726	770

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Percent censored data	Multiple R^2	Adjusted R^2	*MSE*	*RMSE*	*MSPE* (upper)	*MSPE* (lower)	Bias correction factor (Duan, 1983)
Sulfate (SO_4), dissolved, mg/L										
Wamego	$\log(Sulf) = 1.05 \times \log(SC) - 1.06$	Appendix 15	0	0.938	0.937	0.0024	0.049	12	10.7	1.01
DeSoto	$\log(Sulf) = 1.29 \times \log(SC) - 1.73$	Appendix 16	0	0.935	0.934	0.00327	0.0572	14.1	12.3	1.01
Chloride (Cl), dissolved, mgL										
Wamego	$\log(Cl) = 1.82 \times \log(SC) - 3.33$	Appendix 17	0	0.963	0.962	0.00417	0.0646	16	13.8	1.01
DeSoto	$\log(Cl) = 1.97 \times \log(SC) - 3.78$	Appendix 18	0	0.961	0.960	0.00445	0.0667	16.6	14.2	1.01
Alkalinity (ALK), mg/L										
Wamego	$\log(Alk) = 0.526 \times \log(SC) + 0.733$	Appendix 19	0	0.900	0.898	0.0011	0.0332	7.95	7.37	1
DeSoto	$\log(Alk) = 0.567 \times \log(SC) + 0.616$	Appendix 20	0	0.667	0.654	0.00188	0.0434	10.5	9.52	1
Nitrate + Nitrite (NO_x), dissolved, mg/L										
Wamego	$NO_x = -0.00102(SC) - 0.0176 \times (Chl) + 1.85$	Appendix 21	3.7	0.657	0.644	0.0841	0.294	38.8	38.8	--

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Discrete data			
			n	Range of values in variable measurements	Mean	Median
Sulfate (SO_4), dissolved, mg/L						
Wamego	$\log(Sulf) = 1.05 \times \log(SC) - 1.06$	Appendix 15	55	Sulf: 31.4–170	99.7	98.3
				SC: 260–1,280	793	829
DeSoto	$\log(Sulf) = 1.29 \times \log(SC) - 1.73$	Appendix 16	59	Sulf: 12.7–149	92.5	100
				SC: 197–1,110	726	770
Chloride (Cl), dissolved, mgL						
Wamego	$\log(Cl) = 1.82 \times \log(SC) - 3.33$	Appendix 17	55	Cl: 10.5–175	93.4	96.3
				SC: 260–1,280	793	829
DeSoto	$\log(Cl) = 1.97 \times \log(SC) - 3.78$	Appendix 18	59	Cl: 4.93–149	76.3	82.3
				SC: 197–1,110	726	770
Alkalinity (ALK), mg/L						
Wamego	$\log(Alk) = 0.526 \times \log(SC) + 0.733$	Appendix 19	49	Alk: 85–255	177	180
				SC: 260–1,280	784	818
DeSoto	$\log(Alk) = 0.567 \times \log(SC) + 0.616$	Appendix 20	54	Alk: 80.4–234	170	175
				SC: 197–1,110	720	769
Nitrate + Nitrite (NO_x), dissolved, mg/L						
Wamego	$NO_x = -0.00102(SC) - 0.0176(Chl) + 1.85$	Appendix 21	54	NO_x: 0.005–2.15	0.748	0.671
				SC: 260–1,280	800	833
				Chl: 1.7–51.2	16.4	11.6

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Percent censored data	Multiple R^2	Adjusted R^2	*MSE*	*RMSE*	*MSPE* (upper)	*MSPE* (lower)	Bias correction factor (Duan, 1983)
Nitrogen, total (TN), mg/L; total particulate nitrogen plus dissolved nitrogen (TPN + DN)										
Wamego	$\log(TN) = 0.314 \times \log(Turb) - 0.149 \times \log(Chl) - 0.131$	Appendix 22	0	0.829	0.819	0.00668	0.0817	20.7	17.1	1.02
DeSoto	$TN = -0.0362(Temp) + 1.93 \times \log(Turb) - 0.735$	Appendix 23	0	0.854	0.846	0.148	0.385	19.5	19.5	--
Phosphorus, total (TP), mg/L										
Wamego	$\log(TP) = 0.27 \times \log(Turb) - 0.0763 \times \sin(2\pi DY/365) + 0.00703 \times \cos(2\pi DY/365) - 0.893$	Appendix 24	0	0.600	0.574	0.146	0.121	32.1	24.3	1.04
Suspended-sediment concentration (SSC), mg/L										
Wamego	$\log(SSC) = 0.969 \times \log(Turb) + 0.461$	Appendix 25	0	0.781	0.776	0.0615	0.248	77.2	43.6	1.21
DeSoto	$\log(SSC) = 0.942 \times \log(Turb) + 0.441$	Appendix 26	0	0.847	0.844	0.0408	0.202	59.2	37.2	1.12
Escherichia coli bacteria (*Ecoli*), colonies/100 mL										
DeSoto	$\log(Ecoli) = 1.54 \times \log(Turb) - 0.803$	Appendix 27	4	0.731	0.725	0.252	0.502	217	38.5	1.92
Fecal coliform bacteria (Fcoli), colonies/100 mL										
DeSoto	$\log(Fcoli) = 0.877 \times \log(Turb) - 0.228 \times \sin(2\pi DY/365) - 0.59 \times \cos(2\pi DY/365) - 0.481$	Appendix 28	2.08	0.688	0.666	0.328	0.573	274	73.2	1.86

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Discrete data			
			n	Range of values in variable measurements	Mean	Median
Nitrogen, total (TN), mg/L; total particulate nitrogen plus dissolved nitrogen (TPN + DN)						
Wamego	$\log(TN) = 0.314 \times \log(Turb) - 0.149 \times \log(Chl) - 0.131$	Appendix 22	36	TN: 0.734–3.62	1.82	1.54
				Chl: 2.37–51.2	18.2	10.9
				Turb: 7.42–299	76.1	41.2
DeSoto	$TN = -0.0362(Temp) + 1.93 \times \log(Turb) - 0.735$	Appendix 23	40	TN: 0.758–4.65	1.98	1.64
				Temp: 0.04–30.6	18.7	20.6
				Turb: 7.4–982	129	40.4
Phosphorus, total (TP), mg/L						
Wamego	$\log(TP) = 0.27 \times \log(Turb) - 0.0763 \times \sin(2\pi DY/365) + 0.00703 \times \cos(2\pi DY/365) - 0.893$	Appendix 24	49	TP: 0.117–0.917	0.389	0.346
				Turb: 6.2–299	74.9	39
Suspended-sediment concentration (SSC), mg/L						
Wamego	$\log(SSC) = 0.969 \times \log(Turb) + 0.461$	Appendix 25	48	SSC: 14–1,800	226	104
				Turb: 6.2–299	76.1	39.2
DeSoto	$\log(SSC) = 0.942 \times \log(Turb) + 0.441$	Appendix 26	51	SSC: 11–2,190	243	87
				Turb: 7–1,529	108	35
Escherichia coli bacteria (*Ecoli*), colonies/100 mL						
DeSoto	$\log(Ecoli) = 1.54 \times \log(Turb) - 0.803$	Appendix 27	50	*Ecoli*: 0.5–14,000	618	46.5
				Turb: 7–982	98.4	31
Fecal coliform bacteria (Fcoli), colonies/100 mL						
DeSoto	$\log(Fcoli) = 0.877 \times \log(Turb) - 0.228 \times \sin(2\pi DY/365) - 0.59 \times \cos(2\pi DY/365) - 0.481$	Appendix 28	48	Fcoli: 0.5–7,670	569	172
				Turb: 7–674	90.1	31

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Percent censored data	Multiple R^2	Adjusted R^2	*MSE*	*RMSE*	*MSPE* (upper)	*MSPE* (lower)	Bias correction factor (Duan, 1983)
Enterococci bacteria (Ent), colonies/100 mL										
DeSoto	$\log(Ent) = 1.39 \times \log(Turb) + 0.211 \times \sin(2\pi DY/365) + 0.214 \times \cos(2\pi DY/365) - 0.292$	Appendix 29	0	0.661	0.639	0.22	0.469	194	66	2.23
Actinomycetes bacteria (Act), colonies/100 mL										
Wamego	$\log(Act) = 1.35 \times \log(Turb) + 0.198$	Appendix 30	0	0.739	0.734	0.154	0.393	147	59.5	1.58
DeSoto	$\log(Act) = 1.54 \times \log(Turb) + 0.246 \times \sin(2\pi DY/365) + 0.194 \times \cos(2\pi DY/365) - 0.171$	Appendix 31	0	0.850	0.840	0.0936	0.306	102	50.6	1.24

Table 2. Regression models and summary statistics for continuous concentration computations for Kansas River at Wamego (06887500) and De Soto (06892350), Kansas, July 2012 through June 2015.—Continued

[R^2, coefficient of determination; *MSE*, mean square error; *RMSE*, root mean square error; *MSPE*, model standard percentage error; ±, plus or minus; *n*, number of discrete samples; mg/L, milligrams per liter; log, $\log_{10}$; SC, specific conductance in microsiemens per centimeter at 25 degrees Celsius; DS; dissolved solid; sulf, sulfate; --, no data; *Turb*, turbidity in formazin nephelometric units; temp, temperature; sin, sine; *DY*, day of year; cos, cosine; colonies/100 mL, colonies per 100 milliliters]

Site	Regression model	Model archival summary	Discrete data			
			n	Range of values in variable measurements	Mean	Median
Enterococci bacteria (Ent), colonies/100 mL						
DeSoto	$\log(Ent) = 1.39 \times \log(Turb) + 0.211 \times \sin(2\pi DY/365) + 0.214 \times \cos(2\pi DY/365) - 0.292$	Appendix 29	50	Ent: 4–16,700	601	62
				Turb: 7–674	87.8	31
Actinomycetes bacteria (Act), colonies/100 mL						
Wamego	$\log(Act) = 1.35 \times \log(Turb) + 0.198$	Appendix 30	53	Act: 10–9,000	847	160
				Turb: 5.2–299	67.1	32
DeSoto	$\log(Act) = 1.54 \times \log(Turb) + 0.246 \times \sin(2\pi DY/365) + 0.194 \times \cos(2\pi DY/365) - 0.171$	Appendix 31	50	Act: 8–8,500	789	155
				Turb: 7–674	86	32.4

Appendixes 1–31

Appendixes 1–31 contain the model archive summaries for the logistic and linear models. Logistic models have not been previously developed for the Wamego and De Soto sites on the Kansas River. Linear models for some constituents were previously developed for both sites and are indicated in the model summaries. Previously developed linear regression models are documented in Rasmussen and others (2005).

The following appendixes are availabe at http://dx.doi.org/10.3133/ofr20161040.

Appendix 17. Model Archival Summary for Chloride Concentration at Station 06887500; Kansas River at Wamego, Kansas

Appendix 18. Model Archival Summary for Chloride Concentration at Station 06892350; Kansas River at De Soto, Kansas

Appendix 19. Model Archival Summary for Alkalinity Concentration at Station 06887500; Kansas River at Wamego, Kansas

Appendix 20. Model Archival Summary for Alkalinity Concentration at Station 06892350; Kansas River at De Soto, Kansas

Appendix 21. Model Archival Summary for Nitrate + Nitrite Concentration at Station 06887500; Kansas River at Wamego, Kansas

Appendix 22. Model Archival Summary for Total Nitrogen Concentration at Station 06887500; Kansas River at Wamego, Kansas

Appendix 23. Model Archival Summary for Total Nitrogen Concentration at Station 06892350; Kansas River at De Soto, Kansas

Appendix 24. Model Archival Summary for Total Phosphorous Concentration at Station 06887500; Kansas River at Wamego, Kansas

Appendix 25. Model Archival Summary for Suspended-Sediment Concentration at Station 06887500; Kansas River at Wamego, Kansas

Appendix 26. Model Archival Summary for Suspended-Sediment Concentration at Station 06892350; Kansas River at De Soto, Kansas

Appendix 27. Model Archival Summary for *Escherichia coli* Concentration at Station 06892350; Kansas River at De Soto, Kansas

Appendix 28. Model Archival Summary for Fecal Coliform Concentration at Station 06892350; Kansas River at De Soto, Kansas

Appendix 29. Model Archival Summary for Enterococci Concentration at Station 06892350; Kansas River at De Soto, Kansas

Appendix 30. Model Archival Summary for Actinomycetes Concentration at Station 06887500; Kansas River at Wamego, Kansas

Appendix 31. Model Archival Summary for Actinomycetes Concentration at Station 06892350; Kansas River at De Soto, Kansas

Publishing support provided by:
Rolla Publishing Service Center

For additional information concerning this publication, contact:
Director, USGS Kansas Water Science Center
4821 Quail Crest Place
Lawrence, KS 66049
(785) 842–9909

Or visit the Kansas Water Science Center Web site at:
http://ks.water.usgs.gov

ISSN 2331-1258 (online)
http://dx.doi.org/10.3133/ofr20161040