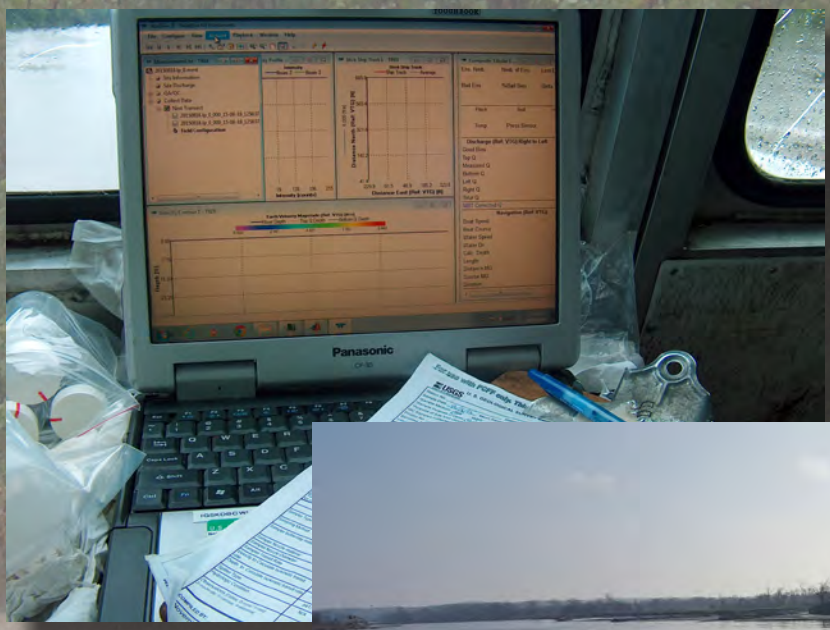
USGS
science for a changing world

Prepared in cooperation with the city of Omaha, Nebraska

Modeling *Escherichia coli* in the Missouri River near Omaha, Nebraska, 2012–16



Scientific Investigations Report 2020–5045

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

Cover: **Background**: A perfect day for sampling runoff on the Missouri River near Omaha, Nebraska, August 18, 2015.
Top Left: Computer collecting discharge information and notes sheet used during a Missouri River water quality sample near Omaha, Nebraska, August 18, 2015.
Bottom Right: The U.S. Geological Survey collects a discharge measurement on the Missouri River at Freedom Park, February 17, 2016.

Modeling *Escherichia coli* in the Missouri River near Omaha, Nebraska, 2012–16

By Brenda K. Densmore, Brent M. Hall, and Matthew T. Moser

Prepared in cooperation with the city of Omaha, Nebraska

Scientific Investigations Report 2020–5045

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

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2020

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 https://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov.

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

Although this information product, for the most part, is in the public domain, it also may contain copyrighted materials as noted in the text. Permission to reproduce copyrighted items must be secured from the copyright owner.

Suggested citation:
Densmore, B.K., Hall, B.M., and Moser, M.T., 2020, Modeling *Escherichia coli* in the Missouri River near Omaha, Nebraska, 2012–16: U.S. Geological Survey Scientific Investigations Report 2020–5045, 24 p., https://doi.org/10.3133/sir20205045.

Associated data for this publication:
Densmore, B.K., and Hall, B.M., 2020, Modeling *Escherichia coli* in the Missouri River near Omaha, Nebraska, 2012–16: Model Inputs and Outputs: U.S. Geological Survey data release, https://doi.org/10.5066/P97S6WSV.

ISSN 2328-0328 (online)

Acknowledgments

The U.S. Geological Survey Nebraska Water Science Center would like to acknowledge the city of Omaha for providing funding and the opportunity to cooperate with them on this long-term (2012–20) Missouri River water-quality sampling effort. This preliminary analysis of *Escherichia coli* has been greatly improved by the city's support, scientific discussions, and supply of several needed datasets on wastewater treatment and combined sewer overflow operations.

Contents

Figures

Tables

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
	Length	
inch (in.)	2.54	centimeter (cm)
inch (in.)	25.4	millimeter (mm)
foot (ft)	0.3048	meter (m)
mile (mi)	1.609	kilometer (km)
	Flow rate	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)

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

$$°C = (°F - 32) / 1.8.$$

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) or micrograms per liter (µg/L).

Bacteria concentrations are given in most probable number of bacteria per 100 milliliters (MPN/100 mL).

A water year is the 12-month period, October 1 through September 30, and is designated by the calendar year in which it ends.

Abbreviations

ADCP	acoustic Doppler current profiler
AMLE	adjusted maximum likelihood estimation
API	antecedent precipitation index
CSO	combined sewer overflow
CSS	combined sewer system
E. coli	*Escherichia coli*
LOADEST	Load Estimator
LTCP	Long Term Control Plan
MLE	maximum likelihood estimation
NDEQ	Nebraska Department of Environmental Quality
NEWSC	Nebraska Water Science Center
PPCC	probability plot correlation coefficient
R^2	coefficient of determination
USGS	U.S. Geological Survey
VIF	variance inflation factor

Modeling *Escherichia coli* in the Missouri River near Omaha, Nebraska, 2012–16

By Brenda K. Densmore, Brent M. Hall, and Matthew T. Moser

Abstract

The city of Omaha, Nebraska, has a combined sewer system in some areas of the city. In Omaha, Nebr., a moderate amount of rainfall will lead to the combination of stormwater and untreated sewage or wastewater being discharged directly into the Missouri River and Papillion Creek and is called a combined sewer overflow (CSO) event. In 2009, the city of Omaha began the implementation of their Long Term Control Plan (LTCP) to mitigate the effects of CSOs on the Missouri River and Papillion Creek. As part of the LTCP, the city partnered with the U.S. Geological Survey (USGS) in 2012 to begin monitoring in the Missouri River. Since 2012, monthly discrete water-quality samples for many constituents have been collected from the Missouri River at four sites. At 3 of the 4 sites, water quality has been monitored continuously for selected constituents and physical properties. These discrete water-quality samples and continuous water-quality monitoring data (from July 2012 to 2020) have been collected to better understand the water quality of the Missouri River, how it is changing with time, how it changes upstream from the city of Omaha to downstream, and how it varies during base-flow conditions and during periods of runoff.

The purpose of this report is to document the development of *Escherichia coli* (*E. coli*) concentration models for these four Missouri River sites. Analysis was completed using the first 5 years of data (through 2016) to determine if the current approach is sufficient to meet future analysis goals and to understand if proposed models such as Load Estimator (LOADEST) models will be able to represent water-quality changes in the Missouri River.

Multiple linear regression models were developed to estimate *E. coli* concentration using LOADEST as implemented in the rloadest package in the R statistical software program. A set of explanatory variables, including streamflow and streamflow anomalies, precipitation, information about CSOs, and continuous water quality, were evaluated for potential inclusion in regression models. The best model at Missouri River at NP Dodge Park at Omaha, Nebr. (USGS station 412126095565201; hereafter "NP Dodge") included basin explanatory variables of upstream antecedent precipitation index measured at Tekamah, Nebr.; decimal time; season; and turbidity. The best model at Missouri River at Freedom Park Omaha, Nebr. (USGS station 411636095535401; hereafter "Freedom Park") included the same explanatory variables as the NP Dodge model with the addition of turbidity anomalies and flow anomalies. The best models at the two downstream sites (Missouri River near Council Bluffs, Iowa, USGS station 06610505 and Missouri River near La Platte, Nebr., USGS station 410333095530101) included the same explanatory variables as the Freedom Park model with the addition of local antecedent precipitation index as measured at Eppley Airport in Omaha, Nebr., and additional turbidity and flow anomalies. The final selected models were the best models given our modeling design constraint in which explanatory variables included in the model for the upstream site were included in the downstream models.

Explanatory variables currently (2020) being collected and included in the selected models through 2016 explained 64–75 percent of the variability of *E. coli* concentration in the Missouri River. Explaining 64–75 percent of the variability might be considered low when working with physical constituents (total nitrogen or sediment), but with the natural variability of biological constituents such as *E. coli*, the uncertainty of *E. coli* laboratory measurements, and the added complexity of modeling in a large drainage basin with multiple sources, these results are adequate and indicate that the explanatory variables being collected and models such as LOADEST can represent water-quality changes in the Missouri River for *E. coli* concentration from 2012 to 2016.

Introduction

The city of Omaha, Nebraska, has a combined sewer system (CSS) in some areas of the city. A CSS collects wastewater from multiple sources, including stormwater runoff, domestic sewage, and industrial wastewater, into one pipe. Typically, a CSS transports all collected wastewater to a wastewater treatment plant for treatment, then discharges treated water to a water body; however, during some rainfall events, the volume of stormwater runoff can cause the total volume to exceed the capacity of the CSS or wastewater treatment plant. When capacity is exceeded, untreated wastewater, including stormwater runoff, domestic sewage, and industrial wastewater, discharges directly to nearby streams, rivers, and other water bodies (U.S. Environmental Protection Agency, 2017). In Omaha, Nebr., a moderate amount of rainfall will lead to untreated wastewater being discharged directly into the Missouri River and Papillion Creek; this is called a combined sewer overflow (CSO) event. There are nearly 860 municipalities throughout the United States that have CSOs as a priority water pollution concern (U.S. Environmental Protection Agency, 2017). In 2009, the city of Omaha began implementation of their Long Term Control Plan (LTCP; City of Omaha, 2014; Clean Solutions for Omaha, 2017) to mitigate the effects of CSOs on the Missouri River and Papillion Creek. The CSOs and stormwater discharges are affecting the water quality of the streams in the Omaha area, often resulting in *Escherichia coli* (*E. coli)* densities greater than 126 units per 100 milliliters and in concentrations greater than their respective health-based screening levels for other constituents (Vogel and others, 2009).

The city of Omaha's LTCP includes several improvements to the sewer system that will eliminate some CSO outfalls and will reduce the volume of raw sewage discharged at other CSO outfalls. Some of the improvements included in the LTCP are stormwater and sewer line separations, stormwater retention and green infrastructure projects, and increased treatment capacity at wastewater treatment plants (City of Omaha, 2014). The city of Omaha plans to complete all the proposed improvements to the sewer system that are described in the LTCP by 2037. As part of the LTCP implementation, in 2012, the city partnered with the U.S. Geological Survey (USGS) to begin monitoring water quality in the Missouri River, one of the streams that receives CSO discharges. Discrete water-quality samples and continuous water-quality monitoring data (from July 2012 to 2020) have been collected to better understand the water quality of the Missouri River, how it is changing with time, how it changes upstream from the city of Omaha to downstream, and how it varies during base-flow conditions and during periods of runoff.

One constituent that is of interest to the city of Omaha is *E. coli*. *E. coli* can come from many sources, which may be natural or anthropogenic. *E. coli* is commonly detected in the gastrointestinal tract of warm-blooded animals and as such is a good indicator of fecal contamination in recreational waters (Ishii and Sadowsky, 2008). Ishii and Sadowsky (2008) reported that *E. coli* can grow in the environment under aerobic and anaerobic conditions, using a variety of energy sources, and in temperatures ranging from 7.5 to 49 degrees Celsius (°C) (with long-term survival even under freezing conditions). However, Ishii and Sadowsky (2008) also reported that *E. coli* replicate best in the environment under conditions of high nutrients and temperature that are most commonly found in tropical or subtropical climates. Since July 2012, the USGS has been collecting *E. coli* samples to better quantify the concentration of *E. coli* in the river over time. The Missouri River is designated for recreational use by the Nebraska Department of Environmental Quality (NDEQ; Nebraska Department of Environmental Quality, 2014). The *E. coli* standard most often used by the NDEQ is *E. coli* bacteria shall not exceed a geometric mean of 126 colonies in 100 milliliters.

Purpose and Scope

The purpose of this report is to document the development of *E. coli* concentration models for the four Missouri River sites near Omaha, Nebr. This initial data analysis used the first 5 years of data (July 2012–September 2016) and was completed to determine if the current sampling and analysis approach is sufficient to meet planned analysis goals and to understand if proposed models such as Load Estimator (LOADEST) (Runkel and others, 2004) are able to represent *E. coli* changes in the Missouri River near Omaha, Nebr., from 2012 to 2016. The intent of the initial analysis and the model development is not to document change over time or difference between sites with this limited dataset. Future analysis is planned to focus on understanding the water quality of the Missouri River (nutrients, biological oxygen demand, suspended solids, and *E. coli*), how it is changing with time, and how it changes upstream from the city of Omaha to downstream. This report also includes *E. coli* sample collection and processing methods.

Study Area Description

The Missouri River is the longest river in the United States. The river travels more than 2,300 miles (mi) starting at the confluence of the Jefferson, Madison, and Gallatin Rivers at Missouri River Headwaters Park in Montana to St. Louis, Missouri, where it joins the Mississippi River (not shown in fig. 1). The river is controlled by six main-stem dams and is managed for authorized purposes including fish and wildlife, flood control, hydropower, irrigation, navigation, recreation, water supply, and water quality (U.S. Army Corps of Engineers, 2013). The U.S. Army Corps of Engineers controls the volume of water being released from these dams. Gavin's Point Dam near Yankton, South Dakota (not shown in fig. 1), is the lowest dam on the Missouri River and controls most flow year round. During navigation season (about March to November), higher flow regimes allow barge traffic on the river. In the nonnavigation season (about December to February), the U.S. Army Corps of Engineers lowers flow to prevent ice jams on the Missouri River and because there is no need for river navigation. Other streams can contribute to the flow in the Missouri River at Omaha. These streams include the Big Sioux River, Floyd River, Little Sioux River, and Boyer River (not shown in fig. 1). Seasonal runoff from these streams during large rain events can contribute large amounts of sediment and nutrients to the Missouri River. The mean daily streamflow at the Missouri River at Omaha, Nebr., streamflow-gaging station (USGS station 06610000) is 36,000 cubic feet per second (ft^3/s) based on 20 years of record (water year 1997–2017; U.S. Geological Survey, 2018). A water year is the 12-month period, October 1 through September 30, and is designated by the calendar year in which it ends.

The Missouri River runs along the east side of the city of Omaha (fig. 1) between river miles 596 and 629 and flows from north to south. The river is channelized as it flows through this section with wing dikes and rip rapped banks. As the Missouri River flows past Omaha, inflows come from direct surface runoff, wastewater treatment plants, tributaries, and CSO outfalls (fig. 1). The city of Council Bluffs, Iowa, has one major wastewater treatment plant and the city of Omaha has two major wastewater treatment plants—the Missouri River Water Resource Recovery Facility, which is just downstream from the US–275 Missouri River Bridge, and the Papillion Creek Water Resource Recovery Facility, which is north of the confluence of Papillion Creek and the Missouri River (fig. 1). The only major tributary entering the Missouri River within the boundary of the city of Omaha is Papillion Creek (fig. 1). Tributaries from Iowa in the Omaha section of river include Pigeon Creek, Indian Creek, and Mosquito Creek (fig. 1). The Boyer River (not shown in fig. 1) enters the Missouri River from Iowa approximately 8 mi upstream from Omaha. Historically, the maximum number of CSO outfalls into receiving streams in Omaha was 32. As of 2009, there were 29 operational CSO outfalls. Since 2009, three CSO outfalls have been deactivated: one in December 2011, one in September 2012, and one in August 2014. As of 2018, the city of Omaha had 26 permitted CSO outfalls: 9 to Papillion Creek and its tributaries and 17 to the Missouri River (fig. 1; City of Omaha, 2017).

The total volume of discharge into the Missouri River from Omaha CSOs and tributaries is a small part of the overall Missouri River streamflow. Although no discharge records are available for CSOs, their contribution, even during large local runoff events, is likely much less than 1 percent of the overall Missouri River streamflow. The mean daily streamflow during ice-free conditions at the Papillion Creek at Fort Crook, Nebr., streamflow-gaging station (USGS station 06610795) is 471 ft^3/s based on 6 years of record (U.S. Geological Survey, 2018). This indicates that during mean streamflow conditions, Papillion Creek streamflow is just barely more than 1 percent of the overall Missouri River streamflow, and it is estimated that CSO contributions are less than Papillion Creek streamflow.

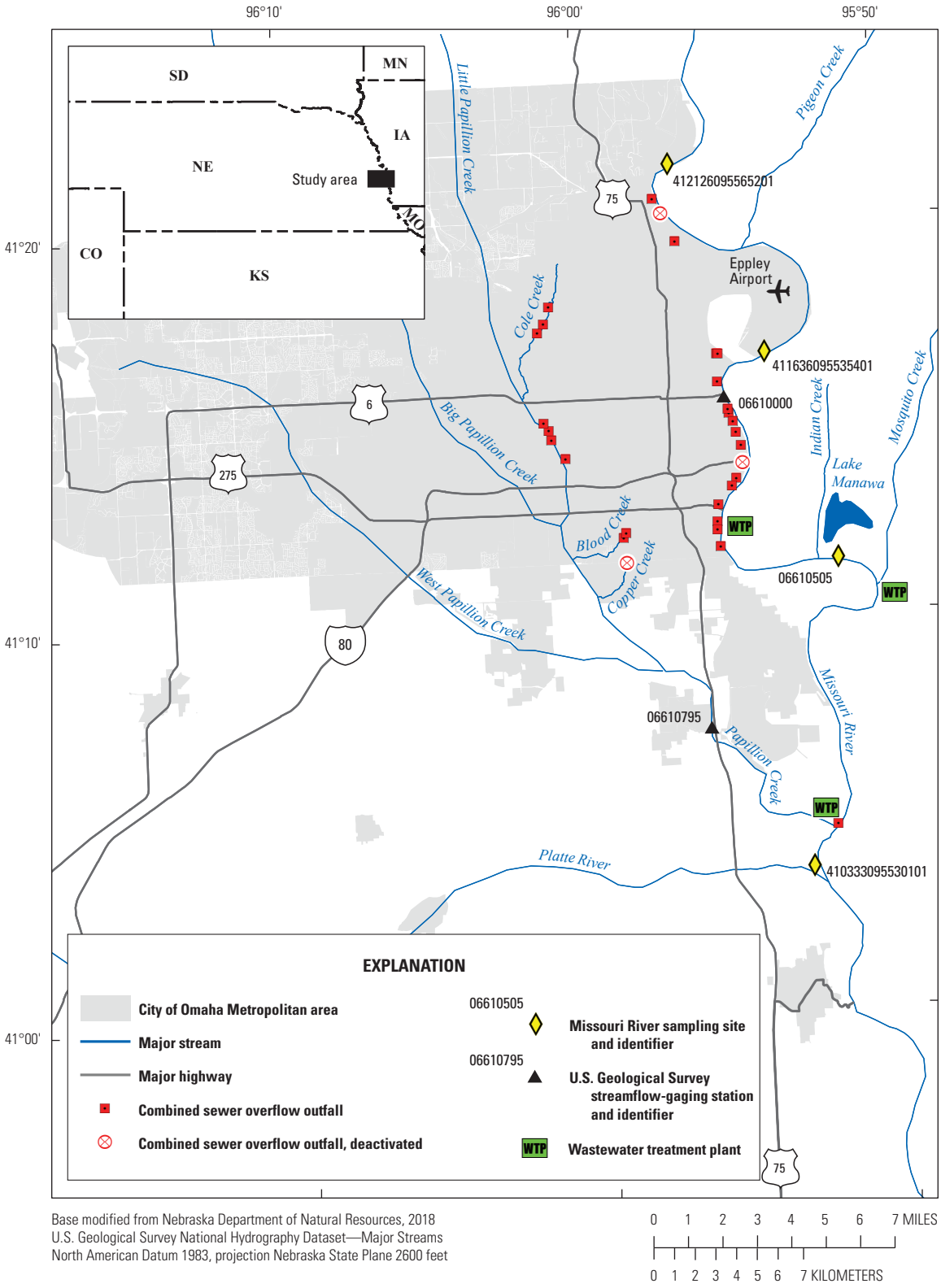


Figure 1. Missouri River near Omaha, Nebraska, including U.S. Geological Survey (USGS) sampling sites with station identifiers, USGS streamflow-gaging stations with station identifiers, combined sewer overflow outfalls, wastewater treatment plants, and tributaries.

Methods of Study

The following section describes the sampling design, methods used for the collection of water-quality samples, and operation of continuous water-quality monitors. This section also describes the methods used to develop concentration models of *E. coli*.

Site Selection and Sampling Frequency

A total of four sampling sites were selected on the Missouri River near the city of Omaha. The four sites were chosen based on the location of the site, CSO outfalls, wastewater treatment plants, and tributary location. The 4 sites include 1 site upstream from the city of Omaha, 2 sites within the city, and 1 site downstream from the city. The four sampling sites were as follows (fig. 1):

1. Missouri River at NP Dodge Park at Omaha, Nebr. (USGS station 412126095565201, MR-5; Vogel and others, 2009; hereinafter referred to as "NP Dodge");

2. Missouri River at Freedom Park Omaha, Nebr. (USGS station 411636095535401, MR-4; Vogel and others, 2009; hereinafter referred to as "Freedom Park");

3. Missouri River near Council Bluffs, Iowa, about 4.5 mi downstream from the Missouri River Water Resources Recovery Plant (USGS station 06610505, hereinafter referred to as "Council Bluffs"); and

4. Missouri River near La Platte, Nebr., between the Papillion Creek confluence and the Platte River confluence (USGS station 410333095530101, hereinafter referred to as "La Platte").

All four sites were sampled from a boat or at the bank during ice conditions once per month, beginning in July 2012 (fig. 2, showing one point on each date that all four sites were sampled). Wet weather sampling was always targeted. Wet weather was defined as a precipitation event with at least 0.1 inch (in.) of precipitation. For each month a specific week was targeted ahead of time—typically the third week of the month. The exact day of sampling during the sampling week was chosen based on precipitation forecasts. If a sampling week did not have any precipitation events forecasted, a sample was still collected and categorized as nonwet weather. To collect additional information on Missouri River water quality during times of local Omaha runoff and possible CSO events, two additional wet weather samples were collected per year beginning in 2015 (fig. 2). These two wet weather samples were collected during the city's disinfection season, which coincides with the recreation season on the Missouri River (May 1 to September 30). These samples were collected any time during disinfection months when a rain event with more than 0.1 in. of precipitation occurred during the late night or early morning before the sample.

Monitoring Data Collection

Data collection efforts include discrete water-quality sample collection for laboratory analysis and continuous monitoring of several constituents and physical properties of the river. Procedures for the collection of these data are outlined in the following sections.

Discrete Water-Quality Sample Collection

During nonice conditions, all samples were collected from a boat. Most samples were collected using isokinetic, depth-integrated sampling procedures that have been designed to obtain samples that represent a composite of the cross section (U.S. Geological Survey, variously dated). Streamflow was measured using an acoustic Doppler current profiler (ADCP) at the sampling cross section prior to sample collection during nonice conditions. A USGS computer program, Equal Discharge Increment Version 3.32, used the streamflow information collected by the ADCP to calculate five sampling points that represent equal discharge on the cross section, including the midpoint of streamflow. A list of all constituents and physical properties measured during a discrete sample is given in table 1.

During winter months when ice was on the river and a boat could not be launched safely, bank samples were collected at all four sites. These types of conditions typically occurred at least once per year. During a bank sample, shore ice was broken, and a single vertical sample was collected in 2 to 3 feet of water near the edge of water. The sample was collected in flowing water that was free of any disturbance from the ice breakup.

Water samples for *E. coli* determinations were collected during every sampling trip. Unlike the other discrete samples, *E. coli* was collected at only one point in the cross section or vertical sampling point. Although the rest of the samples are a complete composite of the channel, *E. coli* samples were collected as grab samples so that the sample only came into contact with one bottle or surface during collection, to maintain consistency with wastewater sampling protocols associated with bacteria (Title 40 Code of Federal Regulations part 136). The water samples for *E. coli* were collected by hand dipping a sample bottle at the midpoint of streamflow in the channel or at the location of the single-vertical sampling point during ice conditions. The hand dip sample was collected by taking the bottle and opening it up underneath the water surface; this was done to avoid any floating debris on the surface of the water. *E. coli* samples were collected at the surface at the midpoint of streamflow in the channel, not at the midpoint of wetted width. In addition to midpoint of streamflow, *E. coli* samples also were collected during most sampling years at the bank as a grab sample to help understand how *E. coli* varied by location in the cross section; however, these samples are not included in this analysis and modeling effort.

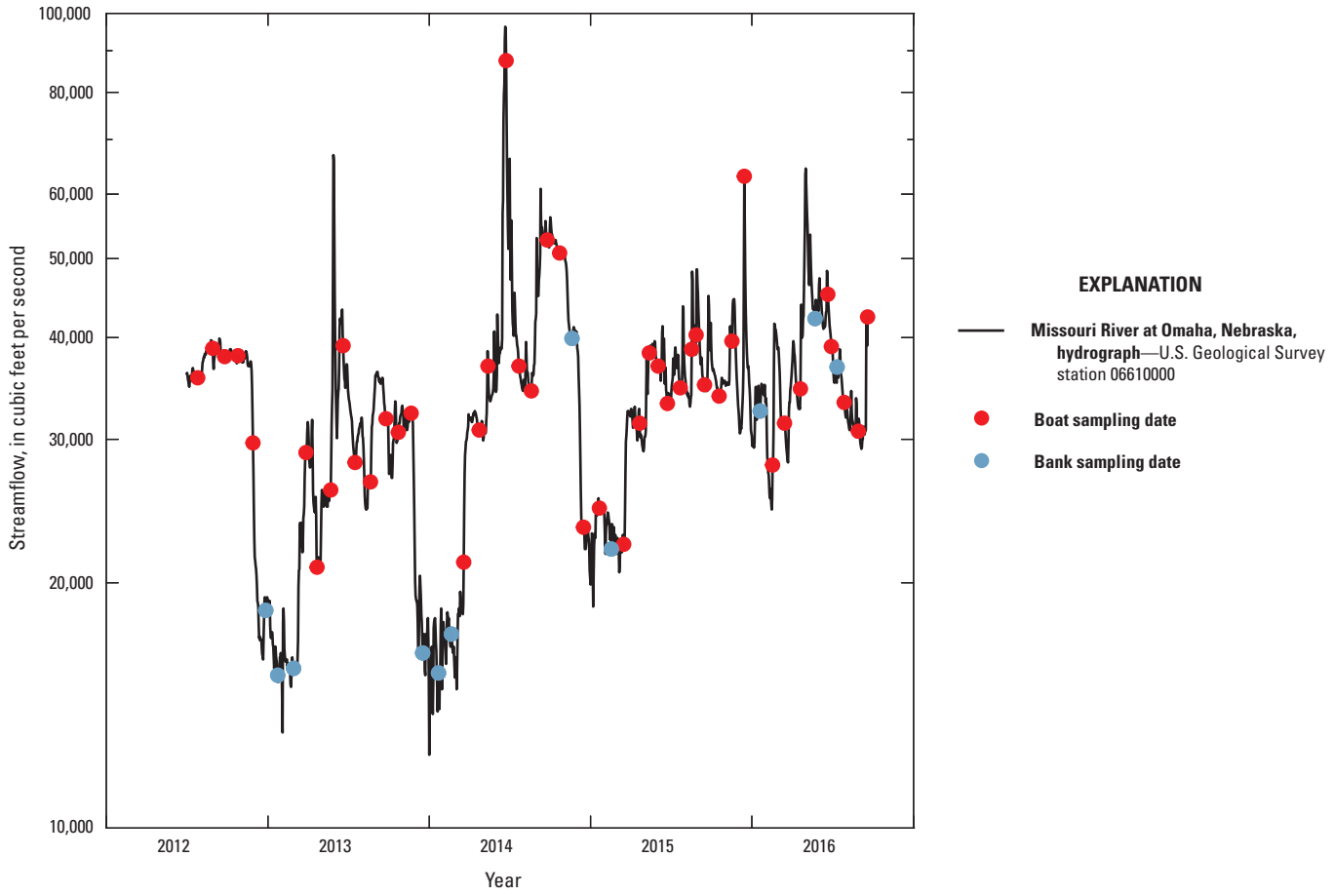


Figure 2. Missouri River streamflow at U.S. Geological Survey streamflow-gaging station with boat and bank sampling events.

Table 1. Water-quality constituents and physical properties analyzed or measured during discrete sampling during both ice and nonice conditions, 2012–16.

[BOD, biochemical oxygen demand, 5-day; TP, total phosphorus; TKN, total Kjeldahl nitrogen; NO_3, nitrate; NH_3, ammonia; TSS, total suspended solids; SC, specific conductance; DO, dissolved oxygen]

Sample location	Analysis method	Water-quality constituent or property
Composited sample mixed in a churn	Constituent analyzed at laboratory	BOD, TP, TKN, NO_3, NH_3, TSS, and chloride.
	Property measured in field	SC, pH, turbidity.
In-situ or grab sample at midpoint of discharge	Constituent analyzed at laboratory	*Escherichia coli*, total coliforms.
	Property measured in field	DO, temperature.

During discrete sampling, field notes were collected that described what the hydrologist observed at each site. This included notes about the amount of debris coming down the river and other indications of river condition including suds, trash, and dead fish. The hydrologist completing the sampling also noted if the Missouri River flow was elevated or if there was local runoff (potential CSO discharge). The hydrologist typically classified the Missouri River as elevated if there was increased debris at NP Dodge, increased turbidity at all sites, and if the hydrograph indicated a recent (in the past day) increase in streamflow. If local runoff was noted for a sample, the hydrologist verified the onsite observations by checking precipitation observations at Eppley Airport. Because this study is focused on understanding the water quality of the Missouri River in respect to different river runoff conditions, it was important to document this information while collecting the sample.

An IDEXX Quantitray 2000 system (IDEXX Laboratories, Inc., 2018) was used for determination of *E. coli* concentrations. This system utilizes IDEXX Quantitray 2000 sealer, medium, and trays. Water samples were transported to the USGS Nebraska Water Science Center (NEWSC) or to Midwest Laboratories (https://midwestlabs.com/about-mwl/, who analyzed samples from July 2012 to March 2013) to be analyzed for *E. coli*. The samples were diluted at the laboratory if necessary based on environmental conditions, including turbidity and runoff. Dilutions were either 1:1, 1:10, or 1:100. If turbidity concentration was greater than 50 formazin nephelometric units, a dilution of 1:10 was often used. If turbidity concentration was greater than 100 formazin nephelometric units, 1:100 dilutions were often used. Runoff was also a factor in determining dilutions. Dilutions were increased if upstream runoff (runoff upstream from Omaha, Nebr.) or local runoff was present. During local runoff events, there is a chance for high *E. coli* concentrations without high turbidity concentrations. The dilutions used and the determining factors for the dilutions are based on hydrologist knowledge from past *E. coli* results at these Missouri River sites and at other sites in Nebraska (Vogel and others, 2009). All samples were processed following USGS standards (U.S. Geological Survey, variously dated).

Continuous Monitoring

Between 2012 and 2013, three multiparameter water-quality monitors were deployed in the Missouri River to collect continuous water-quality data. Monitors were deployed at the NP Dodge and Council Bluffs sites in July 2012 and at the La Platte site in April 2013. Data collection has been ongoing since those dates. These monitors collect specific conductance, dissolved oxygen, turbidity, pH, and temperature data every 15 minutes throughout the year and transmit the data to the USGS National Water Information System (U.S. Geological Survey, 2018) in near real time. Monitors and the resulting data records were operated and maintained in accordance with standard procedures described in Wagner and others (2006).

Quality Assurance

Because the natural variability of *E. coli* concentration is large, replicates were collected for nearly every sample since June 2014. These replicates had similar dilutions to the environmental sample and were processed in the same fashion. The USGS NEWSC has collected and analyzed 125 replicate *E. coli* samples using the IDEXX Quantitray 2000 method (IDEXX Laboratories, Inc., 2018) since 2012; this includes Missouri River samples collected as part of this project as well as samples collected on other streams throughout Nebraska for other USGS projects. These replicate samples are collected immediately after the primary *E. coli* sample. The standard deviation of each replicate was calculated by taking the log base 10 of the replicate, subtracting it from the log base 10 of the primary (environmental) sample, squaring this result, dividing by two, and calculating the square root. This calculation was completed for each of the 125 samples and the mean of these standard deviations was calculated. Conversion to percentage was calculated as observed divided by lower confidence limit. The mean standard deviation is 0.10 log base 10 units which is plus or minus (±) 26 percent. Many of the replicate samples differ from the environmental sample by more than ±0.10 log base 10 units and 101 of these replicates fall within ±0.3 log base 10 units (± 97 percent). The mean standard deviation of replicate samples can provide information on the performance of the IDEXX Quantitray 2000 method. The performance of the method being used provides some information about the minimum amount of change that is detectable with this method.

Many quality assurance procedures and checks are required for proper operation of continuous water-quality monitors. The monitors and the resulting data records were quality controlled following the standard procedures described in Wagner and others (2006).

Statistical Analysis

Linear regression analysis methods were used to develop daily *E. coli* concentration models for the four Missouri River sampling sites because sampling and analysis cost constraints prohibit daily measurements of *E. coli*. For the methods being used in this analysis, daily *E. coli* concentrations are needed to compare differences between sites and between years. The models were developed from sampling data and daily measurements of explanatory variables; these models are used to estimate daily, monthly, or yearly *E. coli* concentrations. These estimated *E. coli* concentrations can then be compared to understand differences among sites and years.

Helsel and Hirsch (2002) describe linear regression as an important tool for the statistical analysis of water resources data. Linear regression is used to describe the covariation between some variable of interest and one or more other variables. In this analysis, regression was used to estimate or predict values of one variable based on knowledge of another variable, for which more data are available. Multiple variables were used to estimate or predict the concentration of *E. coli*. Because multiple explanatory variables are needed to explain the variation observed in *E. coli* concentration, multiple linear regression models were developed. The general form of a multiple linear regression model is shown in equation 1 (Helsel and Hirsch, 2002).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \quad (1)$$

where

y is the response variable,
β_0 is the intercept,
β_1 is the slope coefficient for the first explanatory variable,
β_2 is the slope coefficient for the second explanatory variable,
β_K is the slope coefficient for the Kth explanatory variable, and
ε is the remaining unexplained noise in the data (the error).

Multiple linear regression models were developed to estimate *E. coli* concentration in the Missouri River using LOADEST (Runkel and others, 2004) as implemented in the R statistical software program (rloadest package, Lorenz and others, 2013). Model coefficients were estimated using maximum likelihood estimation (MLE), also called Tobit estimation (Cohen, 1950), or adjusted MLE (AMLE; Cohen, 1976; Cohn, 1988; Runkel and others, 2004). The AMLE provides maximum likelihood estimates of regression model coefficients, corrects for bias in the model coefficients and model estimates, and can be useful when data are censored (or contain less than values, Runkel and others, 2004). When using MLE to develop linear regression models, transformations of y and x variables are frequently required to make the data more nearly normal and improve the fit of the MLE regression because the MLE method assumes a linear model with normally distributed errors (Helsel and Hirsch, 2002). Failure of the data to conform to these assumptions will tend to lower the statistical power of the test and give unreliable estimates of the model parameters (Helsel and Hirsch, 2002). In addition, LOADEST centers some explanatory variables including streamflow and decimal time (Runkel and others, 2004), which removes the effects of multicollinearity. Mayo and Leib (2012) provide the equations and additional descriptions of centering and calculating decimal time that can be implemented in the R statistical software program when rloadest does not do it automatically.

The rloadest package can be used in two ways—automatic variable selection or manual variable selection for models. The automatic variable selection for models determines the best load model from nine available models using various combinations of daily streamflow, daily streamflow squared, time, time squared, and season. Because this analysis is focused on developing concentration models for *E. coli* and because other explanatory variables are needed to explain the variation observed in *E. coli* concentration, rloadest manual variable selection was used to develop custom models. When using manual variable selection and developing custom models, the user must calculate decimal time and complete the centering steps. Variables other than streamflow and time can be centered if multicollinearity is a concern.

Collection of Ancillary Data

A set of explanatory variables, including streamflow and streamflow anomalies, precipitation, information about CSOs, and continuous water quality, were obtained or developed for potential inclusion in regression models (table 2). Explanatory variables of precipitation, LTCP progress, and CSO overflow are available as a USGS data release (Densmore and Hall, 2020). Missouri River streamflow from Missouri River at Omaha, Nebr., streamflow-gaging station (USGS station 06610000) was log transformed and automatically centered by the rloadest package. Streamflow data from local Missouri River tributaries (Papillion Creek at Fort Crook, Nebr. [USGS station 06610795], and Boyer River at Logan, Iowa [USGS station 06609500]; not shown in figures) were log transformed and centered. Information about construction progress was compiled from the City of Omaha project website (City of Omaha, 2017). City of Omaha personnel provided dates for disinfection of wastewater and a record of combined sewer overflow inspections that recorded when there was discharge from each overflow point (Evan Wickham, City of Omaha, written commun., 2017). Hourly precipitation data were obtained from the National Weather Service stations at the airports in Omaha (National Center for Environmental Information, 2017) and Tekamah, Nebr. (National Center for Environmental Information, 2018). These hourly data were then totaled to get daily values with a day consisting of 24 hours and ending at noon local time. Because of the distance from Tekamah to Omaha (about 40 mi), and based on an analysis of Missouri River hydrographs, a time lag of 1 day was added to the precipitation data from Tekamah. The antecedent precipitation index (API; Heggen, 2001) was calculated for both sites using the following equation:

$$API = \sum_{i=0}^{-29} 0.75^{-i} P_i \quad (2)$$

where

API is the antecedent precipitation index, in inches;

i is the day, with i=0 being the present day and i=−1 being the previous day; and

P_i is the precipitation on day i, in inches.

Turbidity data were downloaded from the USGS National Water Information System (U.S. Geological Survey, 2018) as daily mean values. Turbidity and precipitation data were transformed so that the distribution was as close as possible to a normal distribution. Turbidity data were transformed by taking the negative of the inverse of the square root of the recorded value because this transformation produced a dataset that was close to a normal distribution at all sites. The API data from Eppley Airport were transformed by taking the fifth root of the daily value, and lagged API data from the Tekamah airport were transformed by taking the fourth root of the daily value (table 2).

LOADEST requires complete daily datasets of the explanatory variables to calculate daily estimates of loads and concentrations. Several steps were taken to fill gaps in the turbidity records from the continuous water-quality monitors. At each site, the turbidity record from the monitor at that site was used as the primary data source. If there were gaps in the daily mean record, those gaps were first filled using a linear regression between that site and data from another continuous site. The data gaps at NP Dodge and La Platte were filled using estimates based on regression relations with the Council Bluffs monitor. The Council Bluffs data were filled with regression equation estimates from a secondary turbidity sensor deployed less than (<) 16 feet from the continuous monitor. One data gap that occurred during November and December 2013 at the Council Bluffs monitor was filled with data from the NP Dodge monitor. If there were gaps in the continuous water-quality turbidity record on days when discrete water quality was sampled, the record was filled at NP Dodge and Council Bluffs by substituting the turbidity value measured from the sample churn. Finally, any remaining gaps at all three sites were filled using the fillMissing command from the USGS statistical package for R (Lorenz, 2015). The fillMissing command uses simple interpolation with data from the adjacent five days of the gap. These steps were used to fill approximately 12 percent of the record at NP Dodge, 10 percent of the record at Council Bluffs, and 3 percent of the record at La Platte. Continuous turbidity data at Freedom Park were estimated using the continuous turbidity record from Council Bluffs, because turbidity was not monitored at Freedom Park. The Council Bluffs turbidity record was selected because comparisons between continuous data from NP Dodge and Council Bluffs with turbidity values from discrete samples collected at Freedom Park indicated that Freedom Park turbidity was most similar to turbidity at Council Bluffs.

Table 2. Daily datasets explored as potential explanatory variables to model Missouri River *Escherichia coli* concentrations, 2012–16.

[Nebr., Nebraska; API, antecedent precipitation index; CSO, combined sewer overflow; LTCP, Long Term Control Plan]

Variables
Basin
Missouri River at Omaha, Nebr. (06610000) daily streamflow and daily streamflow squared.
Missouri River at Omaha, Nebr. (06610000) daily streamflow short and medium term anomalies.
Decimal time and decimal time squared.
Season.
Transformed API from Tekamah, Nebr., lagged by 1 day.
Log of Boyer River at Logan, Iowa (06609500) daily streamflow.
Local
Transformed daily mean turbidity.
Turbidity short and medium term anomalies.
Transformed API from Eppley Airport in Omaha, Nebr.
Specific conductance.
Chlorination season.
Log of Papillion Creek at Fort Crook, Nebr. (06610795) daily streamflow.
Number of CSOs discharging upstream from sampling point.
Number of LTCP projects completed.

Anomaly values, or values representing how a daily value differs from the mean of past daily values, were calculated for streamflow and turbidity using the waterData package for the R statistical software program (Ryberg and Vecchia, 2012). The anomalies were calculated by subtracting the mean value over a set number of prior days from the observed value for a given day. Two anomaly periods were selected for inclusion in the analysis: short and medium. An example of the calculation for the short-term streamflow anomaly is the log of the streamflow on day *t* minus the 10-day mean of the log of streamflow (this 10-day period starts on day *t* and includes the 9 previous days). An example of the calculation for the medium-term streamflow anomaly for day *t* is the 10-day mean of the log of streamflow minus the 100-day mean of the log of streamflow (again both periods start on day *t* and include either the 9 previous days or the 99 previous days).

Loadest *Escherichia Coli* Concentration Model Development

E. coli models were developed for all four sites. The first model developed was for the site at NP Dodge, and this model served as a baseline model for sites downstream from NP Dodge. As the most upstream model, these estimates represent *E. coli* concentrations in the Missouri River coming into the Omaha area. This model was intended to include many of the basin explanatory variables (table 2) because those variables most likely would represent the *E. coli* concentrations entering the Omaha reach, whereas local variables such as Omaha rainfall and CSOs would not be appropriate. Once the NP Dodge model was completed, models were developed for the other sites in downstream order. These models included the explanatory variables used in the NP Dodge model as well as additional local explanatory variables. Only *E. coli* grab samples from the midpoint of streamflow were included in the analysis, which means that a few November, December, January, and February samples were not used if they were bank samples during ice conditions. Models were developed with all *E. coli* grab samples from the midpoint of streamflow; however, if there were less than three of these samples collected per month per site, then the model was not used to predict *E. coli* concentration for that month. For example, from July 2012 through September 2016, boat samples were not able to be collected at NP Dodge in January 2013, 2014, or 2016; therefore, there were less than three samples collected in January at NP Dodge so the NP Dodge model was not used to predict *E. coli* concentrations for the month of January. The *E. coli* grab samples from the bank during ice conditions were not used because several years of collecting both midpoint of streamflow and bank samples have shown that these concentrations often are different. At the time this data analysis was completed (2018), there were not enough data collected to develop a good relation that would allow *E. coli* concentration at the midpoint of streamflow to be estimated from bank *E. coli* samples.

LOADEST manual model development was used to create concentration models with different combinations of explanatory variables (table 2) to determine which had the strongest relation to *E. coli* concentration. At NP Dodge, a systematic approach was used to add and remove explanatory variables into a model to first determine which explanatory variables were the most significant. Once the most significant explanatory variables were determined, then additional explanatory variables were added one at a time to determine if any of these could substantially improve the model without correlating with the very significant explanatory variables. All explanatory variables were tried. This same type of approach was used at each of the downstream sites with the exception that model development started with all the explanatory variables that had been included in the upstream models. At each site, the best working model was selected based on model diagnostics, residual plots, explanatory variable correlation, and bias statistics comparing the observed and estimated loads (Lee and others, 2017).

Mean *E. coli* concentrations for selected periods can be obtained from LOADEST using load prediction functions in the rloadest package by using a synthetic flow value that converts the output loads to concentrations (Runkel and others, 2004; Lorenz, 2015, 2017b). The synthetic flow is set as the inverse of the concentration-to-load unit conversion factor so that the flow equals 1 after conversion and the output value for load is actually the mean concentration. This gives a mean monthly or annual time weighted concentration, which is the mean of all the daily concentration values for that month or year. LOADEST also calculates 95-percent confidence intervals for each mean concentration.

Model Evaluation

Several metrics were used to evaluate model quality. The coefficient of determination (R^2) value indicates the variation in the water-quality constituent that is explained by the explanatory variables in the model. All models were evaluated for residual normality using the probability plot correlation coefficient (PPCC), which is the *r*-value with the *p*-value statistic. Models were evaluated to ensure the PPCC *r*-value was near 1 and the *p*-value was greater than 0.05. The serial correlation of residuals also is calculated by rloadest (Lorenz and others, 2013), and all models were evaluated to ensure this value was low (<0.2). The variance inflation factor (VIF) calculated by rloadest aids in identifying multicollinearity between explanatory variables included in the model (Helsel and Hirsch, 2002). Models that contained explanatory variables with VIFs <10 were selected to ensure mostly independent explanatory variables. The bias of a model is evaluated using model bias diagnostics, which are based on the comparison of the sampled data to the model predicted value. The load or concentration bias is given in percent, with positive bias indicating overestimation and negative bias indicating underestimation (Lorenz, 2017a). Models were evaluated to ensure concentration bias was < ±25 percent (except at La Platte where the model included many variables from upstream models). The partial concentration ratio uses only estimates that have an observed value and is the sum of estimated values divided by the sum of observed values, so ratios greater than 1 indicate overestimation (Lorenz, 2017a). Models were evaluated to ensure partial concentration ratios were between 0.75 and 1.2 (except at La Platte where the model included many variables from upstream models). In addition to the quality indicators, diagnostic plots were used to evaluate the best models including predicted versus sampled values plot, residuals versus predicted values, partial residual plots for each explanatory variable, residuals versus time, residuals versus streamflow, and normal quantile plot of the residuals (Helsel and Hirsch, 2002).

Missouri River *Escherichia Coli* Concentration Model Results

E. coli samples were collected at all four Missouri River sampling sites from 2012 to 2016 and sample collection continues to present (2020). At each site, a total of 47 *E. coli* samples collected between July 2012 and September 2016 were included in the regression model analysis; bank samples collected during ice conditions on the river and at other times were not included. Sampling occurred over a range of flow conditions (fig. 2). Approximately 15 of the 47 samples collected at each site were collected during local runoff (fig. 3) with a range in precipitation from 0.03 to 3.79 in.

Selected Models

The best model at NP Dodge included basin explanatory variables of upstream precipitation measured at Tekamah, Nebr.; decimal time; season; and turbidity. The best model at Freedom Park included the same explanatory variables as the NP Dodge model with the addition of turbidity anomalies and flow anomalies. The best models at the two downstream sites included the same explanatory variables as the Freedom Park model with the addition of local antecedent precipitation index as measured at Eppley Airport in Omaha, Nebr. (fig. 1) and additional turbidity and flow anomalies. The form of the selected regression equation for the models is as follows:

$$\begin{aligned} Ln(C) = {} & a_0 + a_1(dtime) + a_2(dtime^2) + a_3(TU) + \\ & a_4(API\_T) + a_5\sin(2\pi dtime) + a_6\cos(2\pi dtime) + \\ & a_7(TU\ short) + a_8(Q\ medium) + a_9(API) + \\ & a_{10}(TU\ medium) + a_{11}(Q\ short) \end{aligned} \quad (3)$$

where

C	is *E. coli* concentration, in most probable number of bacteria per 100 milliliters;
a_n	are model coefficients;
dtime	is decimal time centered;
TU	is transformed turbidity;
API_T	is transformed and 1-day lagged antecedent precipitation index from Tekamah, Nebr.;
TU short	is turbidity short-term anomaly;
Q medium	is Missouri River at Omaha, Nebr. (06610000) daily streamflow medium-term anomaly;
API	is transformed antecedent precipitation index from Eppley Airport Omaha, Nebr.;
TU medium	is turbidity medium-term anomaly; and
Q short	is Missouri River at Omaha, Nebr. (06610000) daily streamflow short-term anomaly.

The selected models at each site are shown in table 3 and coefficients and some quality indicators are included.

Many potential explanatory variables were not included in the selected models for several reasons. Some explanatory variables were excluded because of multicollinearity with variables already included in the upstream models. The explanatory variable of LTCP projects completed could not be used in the same model as time because of multicollinearity.

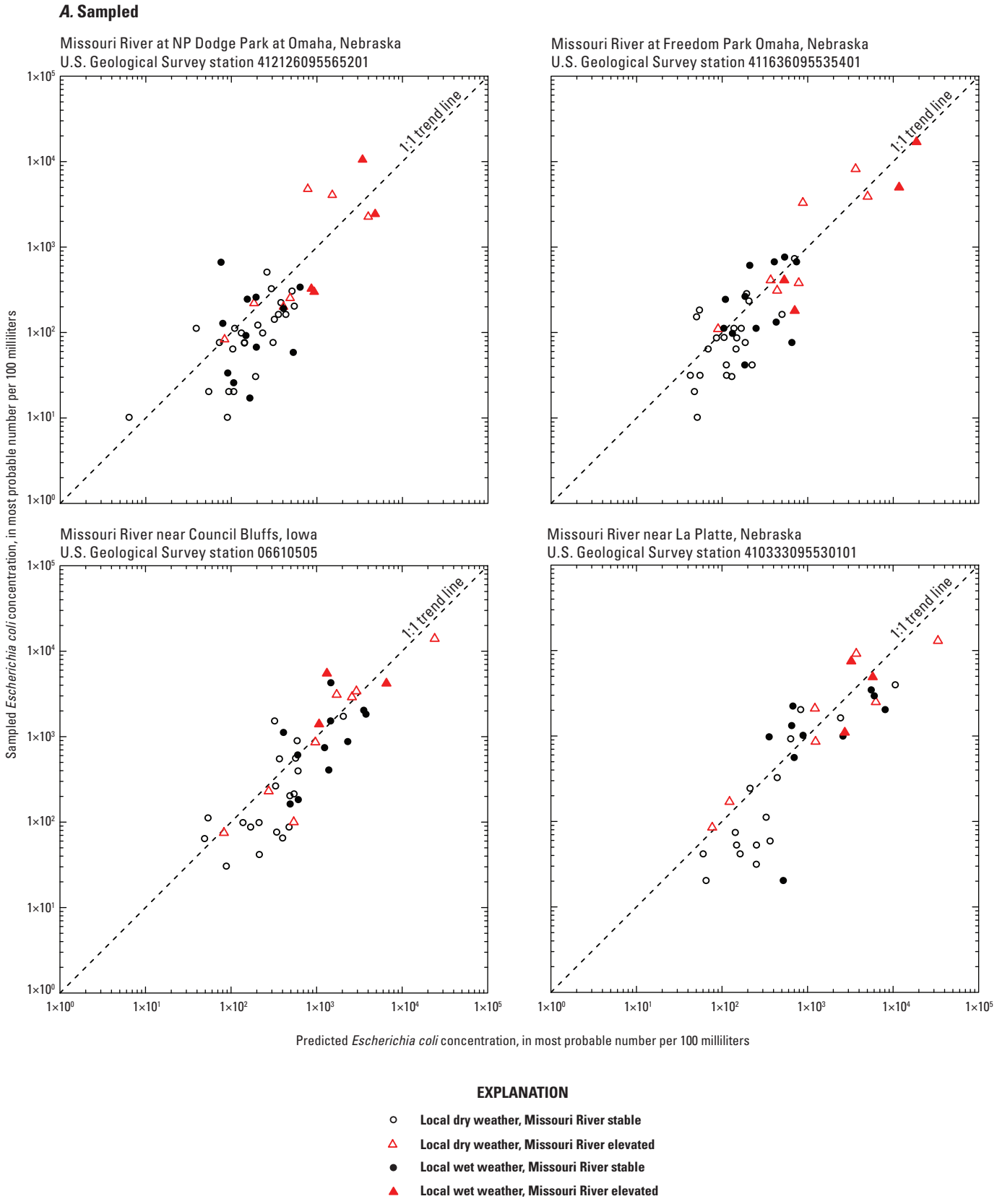


Figure 3. Model diagnostic plots at sampling sites near Omaha, Nebraska. *A*, predicted *Escherichia coli* concentration versus sampled *Escherichia coli* concentration; and *B*, predicted *Escherichia coli* concentration versus residuals.

B. **Residual**

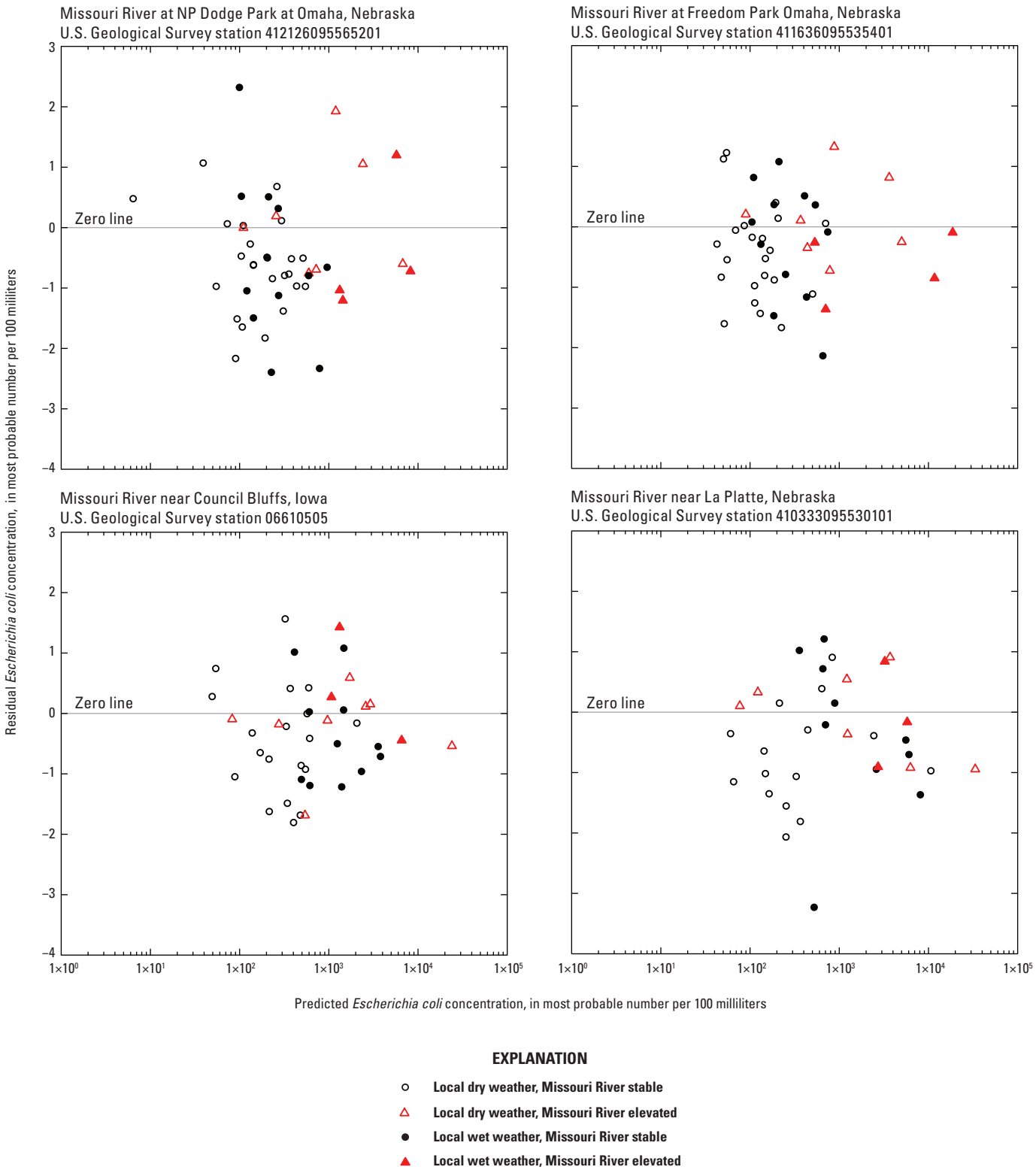


Figure 3. —Continued

Table 3. Selected *Escherichia coli* concentration models with quality indicators, 2012–16.

[*n*, number of samples included in model development; a_n, model coefficients; API, antecedent precipitation index; sin, sine; cos, cosine; --, no data; R^2, coefficient of determination that indicates the percent of *Escherichia coli* variability explained by the model; VIF, variance inflation factor (highest value from all the included explanatory variable); PPCC, probability plot correlation coefficient; <, less than]

Site	*n*	Coefficients															
		Intercept a_0	Intercept *p*-value	Decimal time a_1	Decimal time *p*-value	Decimal time^2 a_2	Decimal time^2 *p*-value	Turbidity a_3	Turbidity *p*-value	API Tekamah with 1-day lag a_4	API Tekamah with 1-day lag *p*-value	sin(date) a_5	sin(date) *p*-value	cos(date) a_6	cos(date) *p*-value	Turbidity short-term anomaly a_7	Turbidity short-term anomaly *p*-value
NP Dodge	47	6.1256	<0.0001	0.1694	0.1745	−0.3483	0.0027	12.0724	0.0007	3.0577	<0.0001	0.1494	0.4733	0.9498	0.0004	--	--
Freedom Park	45	6.5372	<0.0001	0.1650	0.1274	−0.3054	0.0043	11.9928	0.0006	1.8188	0.0040	0.3407	0.0678	0.8621	0.0035	1.8870	0.0175
Council Bluffs	41	4.5969	0.0033	0.2327	0.1481	−0.0275	0.8279	2.2950	0.7319	−0.1276	0.9001	−0.1214	0.6801	1.0491	0.0058	2.2078	0.0486
La Platte	35	8.0082	0.0001	0.2601	0.3950	−0.2830	0.2214	14.2702	0.0944	−0.8979	0.4571	−0.0040	0.9915	1.3898	0.0037	0.2529	0.8703

Site	*n*	Coefficients							
		Missouri River at Omaha, Nebraska (06610000) daily streamflow medium-term anomaly a_8	Missouri River at Omaha, Nebraska (06610000) daily streamflow medium-term anomaly *p*-value	API Omaha a_9	API Omaha *p*-value	Turbidity medium-term anomaly a_{10}	Turbidity medium-term anomaly *p*-value	Missouri River at Omaha, Nebraska (06610000) daily streamflow short-term anomaly a_{11}	Missouri River at Omaha, Nebraska (06610000) daily streamflow short-term anomaly *p*-value
NP Dodge	47	--	--	--	--	--	--	--	--
Freedom Park	45	3.4639	0.0438	--	--	--	--	--	--
Council Bluffs	41	2.4103	0.3189	2.6850	0.0036	2.1436	0.0656	−7.7625	0.0755
La Platte	35	−3.5947	0.2243	2.1336	0.0913	3.3919	0.0160	−3.9432	0.5467

Site	*n*	Quality indicators						
		R^2	Highest VIF from all included coefficients	PPCC *r*-value	PPCC *p*-value	Partial concentration ratio	Adjusted R^2	*p*-value of overall model
NP Dodge	47	64	1.80	0.9803	0.1068	0.758	0.59	<0.0001
Freedom Park	45	75	2.35	0.9956	0.9465	1.098	0.69	<0.0001
Council Bluffs	45	72	6.89	0.9927	0.7868	1.191	0.61	<0.0001
La Platte	39	74	8.50	0.9852	0.3897	1.515	0.61	<0.0001

This was the same for the explanatory variable disinfection season, which could not be used with the explanatory variable season. Other explanatory variables had low predictive power and did not substantially improve the model, such as specific conductance. The final models selected were the best models given the constraint of our modeling design in which explanatory variables included in the models for upstream sites were also included in the downstream models. Some of the explanatory variables used in the NP Dodge model were not significant and had very small coefficients in the downstream models. For example, the API from Tekamah, lagged by 1 day, was very significant (*p*-value less than 0.0001) in the NP Dodge model, significant (*p*-value less than 0.05) in the Freedom Park model, and very insignificant (*p*-value greater than 0.45) with a slightly negative coefficient in the Council Bluffs and La Platte models (table 3). In addition, the turbidity short-term anomaly was significant at Freedom Park and Council Bluffs but not at La Platte. The turbidity medium-term anomaly was significant in the La Platte model and also included in the Council Bluffs model because it was nearly significant and improved other model quality indicators. Similarly, the Missouri River at Omaha, Nebr., daily streamflow short-term anomaly was included in the Council Bluffs model because it also improved other model quality indicators. The insignificant explanatory variables in each model have small coefficients except with the Missouri River at Omaha, Nebr., daily streamflow short- and medium-term anomalies; however, the slightly larger coefficients still have only a small effect on the *E. coli* concentrations predicted because the magnitude of the anomaly values is small.

The models accounted for 64–75 percent (R^2 value, table 3) of the variability in sampled *E. coli* concentrations. The total number of samples included in the development of the models slightly varied (table 3) depending on the exact start date of continuous water-quality monitoring and the explanatory variable used in the model. The highest VIF for the explanatory variables included in the selected models indicates that multicollinearity is not present. The PPCC *p*-values were greater than 0.05, which indicates residuals from each model were normally distributed (table 3). Residuals were also evaluated from plots of residuals on a normal probability plot (not shown), residuals versus time (not shown), and residuals versus predicted *E. coli* (fig. 3*B*). Diagnostic plots of predicted *E. coli* concentration versus sampled *E. coli* concentration show that the selected models at each site are predicting *E. coli* concentration adequately (fig. 3*A*). The highest concentrations (near or greater than $1x10^4$ most probable number per 100 milliliters) at the two upstream sites only occur when Missouri River flow is elevated. However, at La Platte, high *E. coli* concentrations were sampled and predicted during times of stable Missouri River flow, especially during local wet weather. The residuals at all sites ranged from −4 to 3 and showed no pattern to indicate that the models are biased (fig. 3*B*).

Estimation of Daily, Annual, and Recreation Season *Escherichia Coli* Concentrations

Daily *E. coli* concentrations were estimated (predicted) for all four sites (except in the winter months of December, January, and February) using the selected models (fig. 4). The models slightly overpredict *E. coli* concentrations at values below approximately 100 most probable number of bacteria per 100 milliliters (figs. 3 and 4).

Annual mean *E. coli* concentrations were calculated at all four sites using the selected models (fig. 5). Annual mean *E. coli* concentrations represent the mean of all daily estimated *E. coli* concentrations for that year and are useful for seeing how total *E. coli* concentrations change from year to year. Annual mean *E. coli* concentrations are calculated by water year (October 1 through September 30) and only for years with complete continuous monitoring data. Annual mean *E. coli* concentrations at NP Dodge and Freedom Park in 2015 and 2016 were slightly less than 2014 but the annual mean *E. coli* concentrations at Council Bluffs and La Platte remained about the same during these 3 years. Although it appears that annual mean *E. coli* concentrations at NP Dodge and Freedom Park were lower than Council Bluffs and La Platte in all years, the overlap of 95-percent confidence intervals indicates that a statistical difference between the sites cannot be determined with current models (fig. 5). The wide 95-percent confidence intervals at La Platte, and in some years Council Bluffs, are because of the variability in the sampled and predicted concentrations through the water year as well as the number of variables included in the models. In addition, differences between sites from year to year are not a focus of this analysis with this limited dataset; looking for trends from year to year requires many years of data because of yearly variability especially in such a large river system.

Mean *E. coli* concentrations during the recreation season are slightly different than annual mean *E. coli* concentrations. Recreation season mean *E. coli* concentrations were calculated at all four sites using the selected models (fig. 6). Recreation season mean *E. coli* concentrations represent the mean of daily estimated *E. coli* concentrations for the recreation season each year, May 1 to September 30. The recreation season mean *E. coli* concentrations in general are slightly higher than the annual mean concentrations, but not always. The relation between sites each year is similar between the annual means and the recreation season means with a few slight differences most noticeably at La Platte in 2016. Predicted daily, annual mean, and recreation season mean *E. coli* concentrations are available as a USGS data release (Densmore and Hall, 2020).

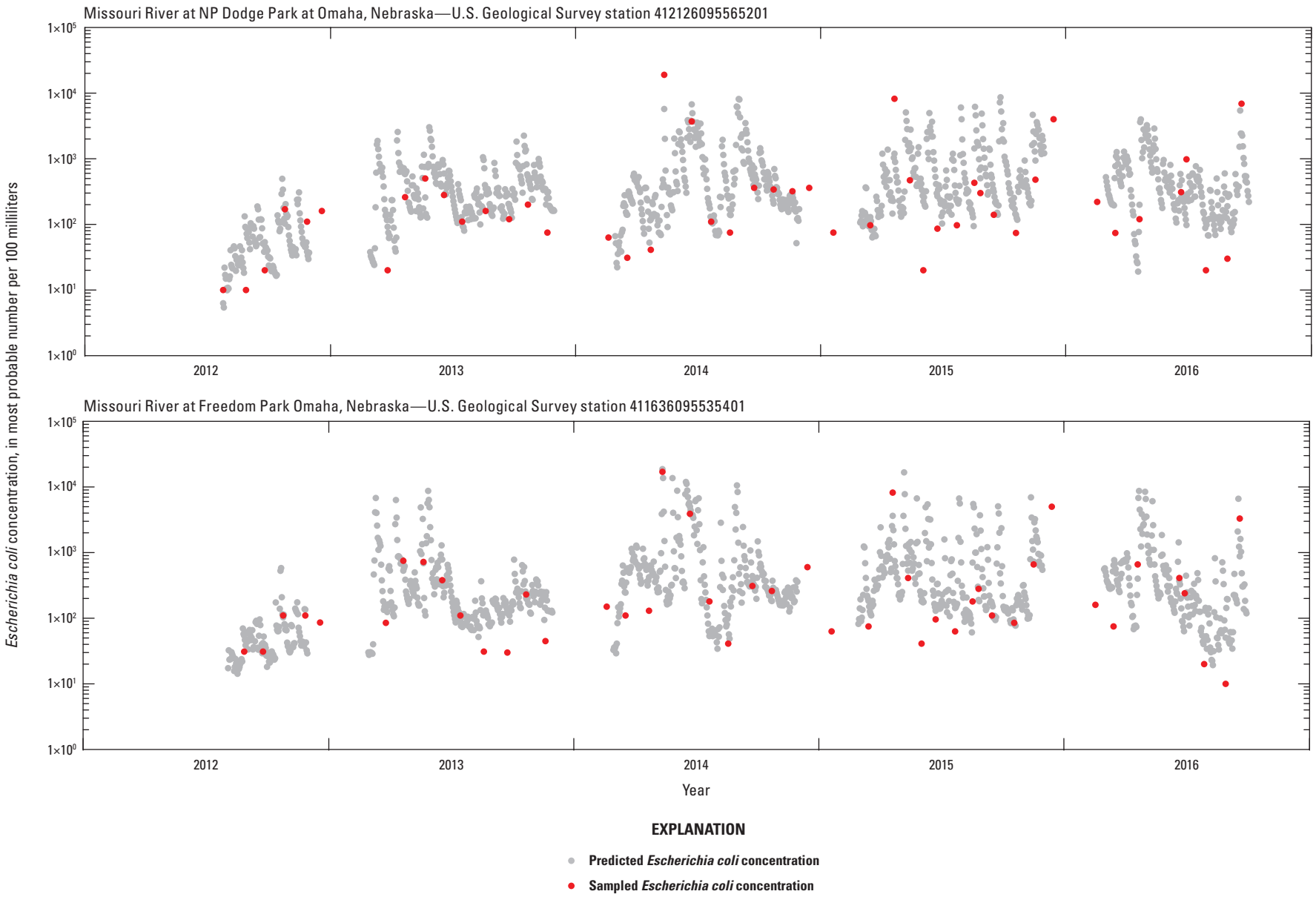


Figure 4. Daily *Escherichia coli* concentrations predicted from selected models and sampled *Escherichia coli* concentrations at Nebraska sampling sites.

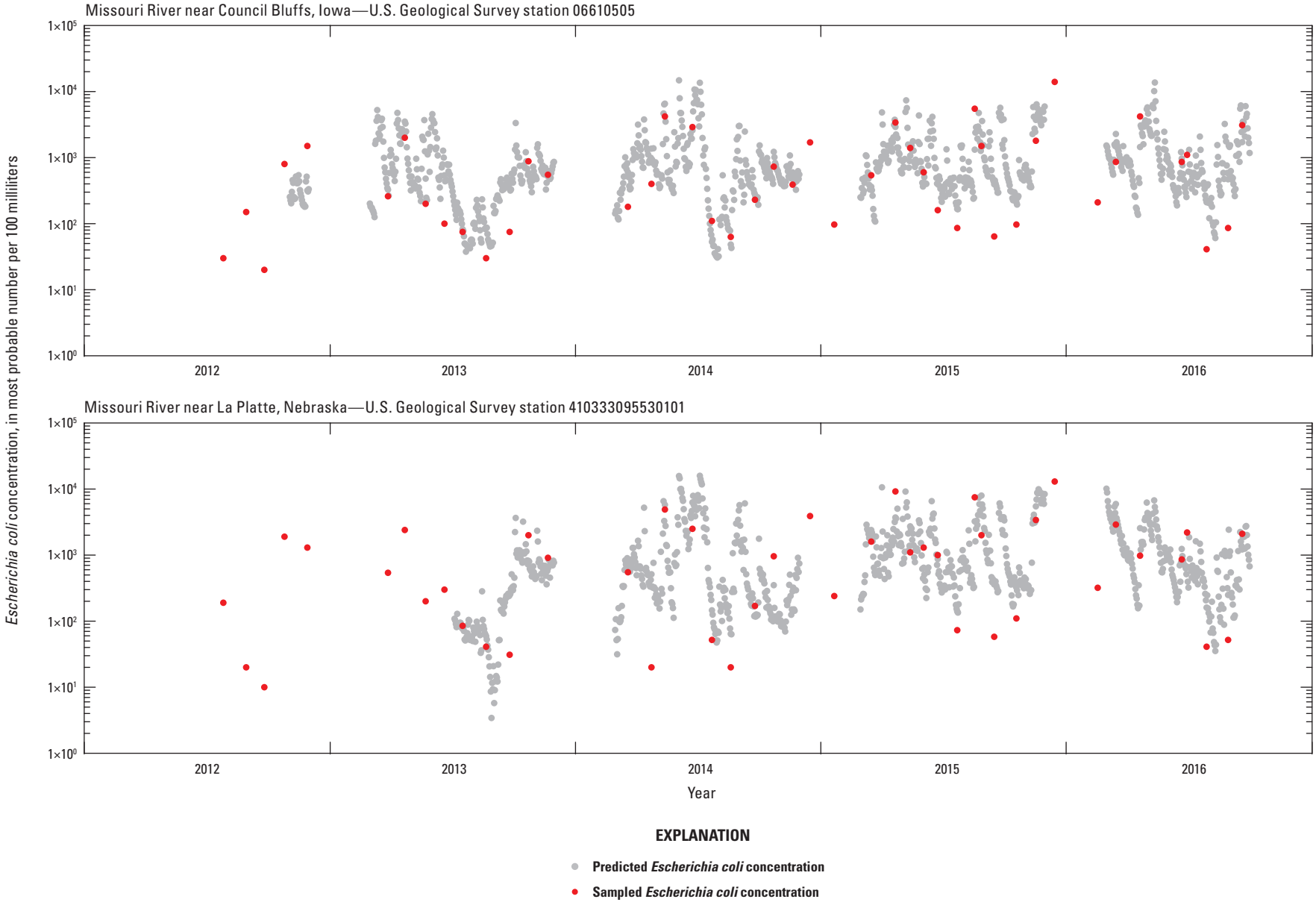


Figure 4. —Continued

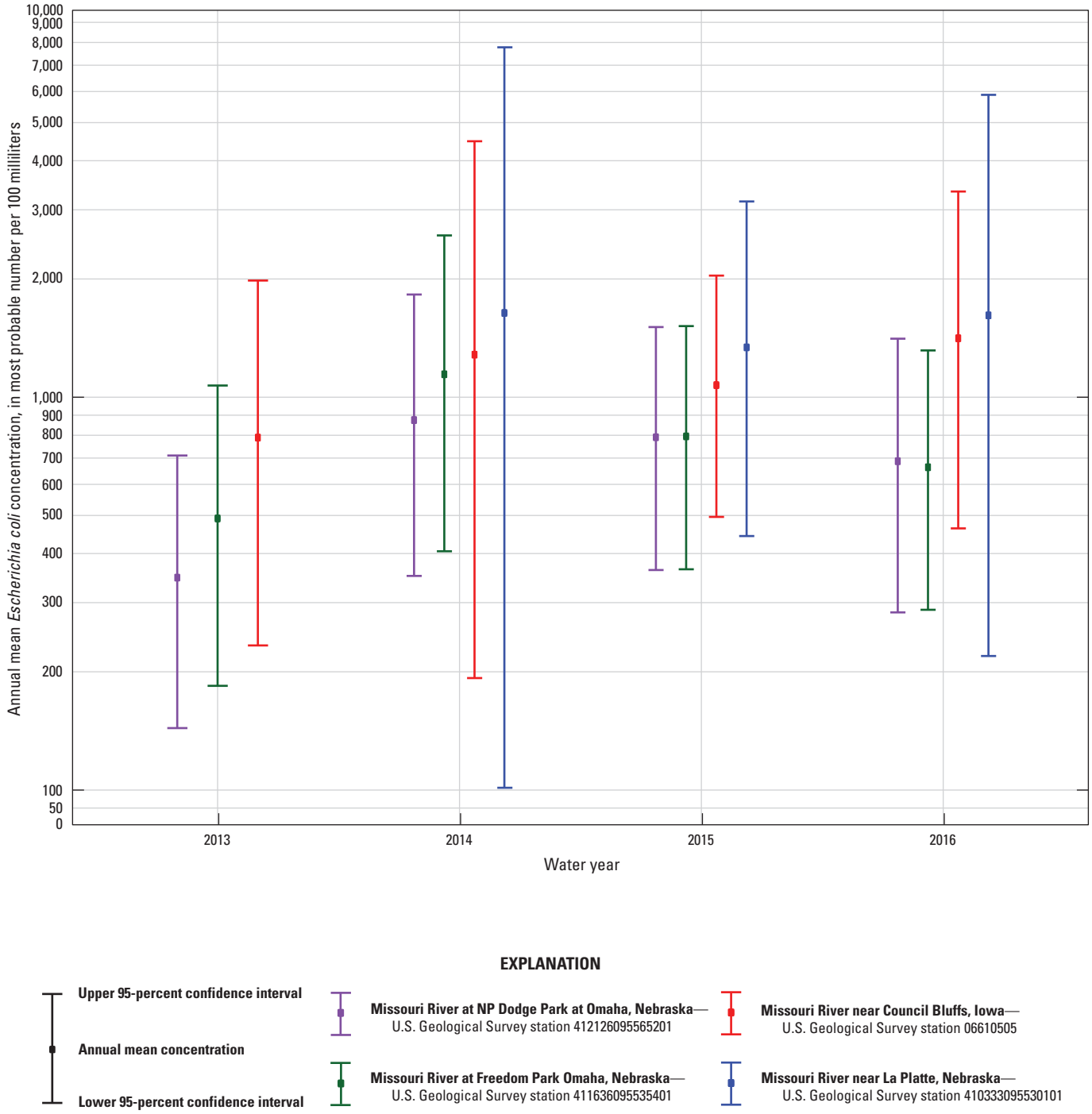


Figure 5. Annual mean *Escherichia coli* concentrations predicted from selected models at Nebraska sampling sites.

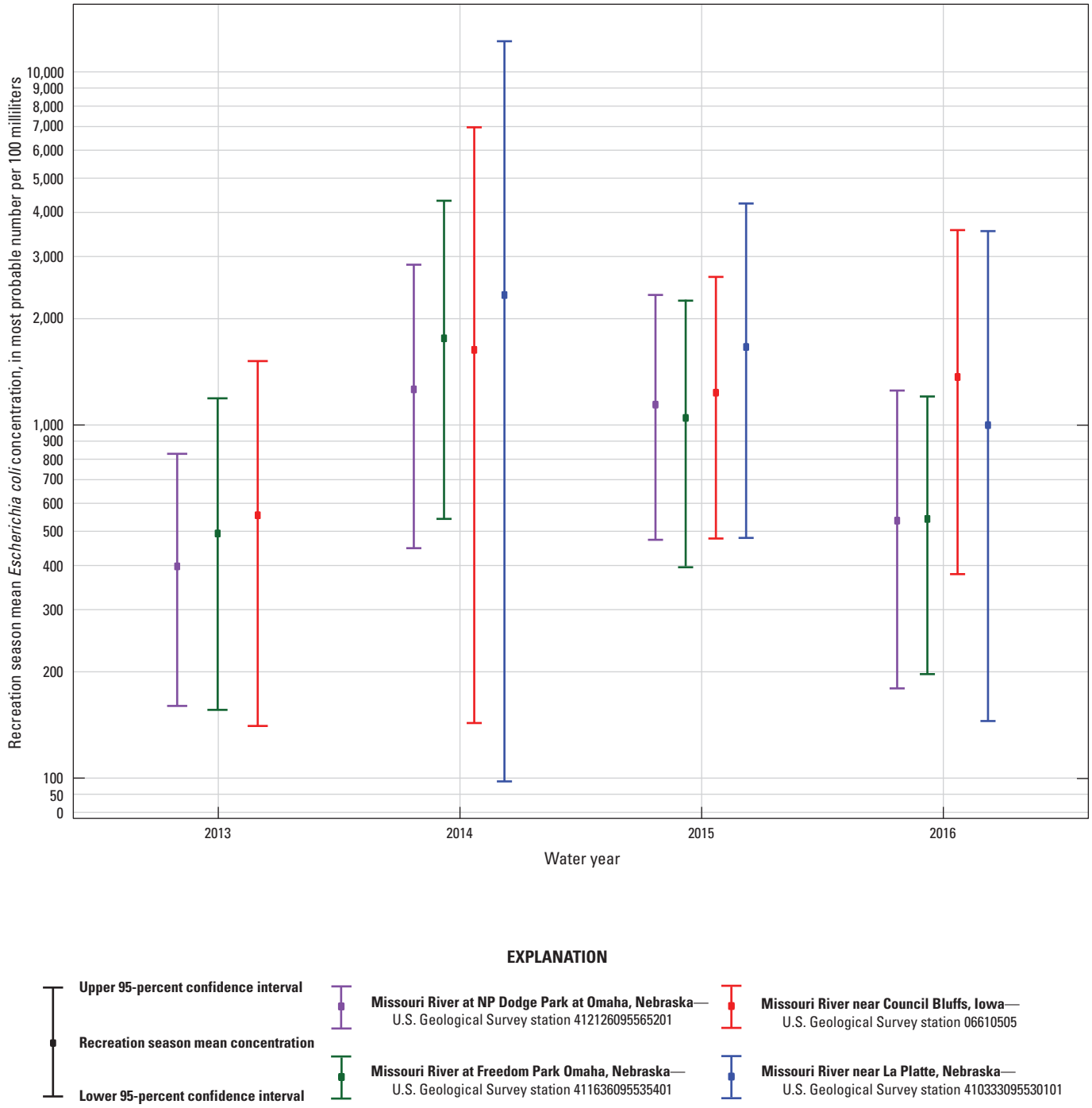


Figure 6. Recreation season mean *Escherichia coli* concentrations predicted from selected models at Nebraska sampling sites.

Model Capabilities and Limitations

The goal of this initial model regression analysis was to determine if the datasets currently being collected for this study are sufficient to meet future analysis goals and to understand if proposed models such as LOADEST can adequately represent water-quality changes in the Missouri River. Explanatory variables currently being collected and included through 2016 in the selected models explained 64–75 percent of the variability of *E. coli* concentration in the Missouri River. Explaining 64–75 percent of the variability might be considered low when working with physical constituents (total nitrogen or sediment), but with the natural variability of biological constituents such as *E. coli*, the uncertainty of *E. coli* laboratory measurements, and the added complexity of modeling in a large drainage basin with multiple sources, these results indicate that the explanatory variables being collected and models such as LOADEST can adequately represent water-quality changes in the Missouri River for *E. coli* concentration. Because of the factors mentioned above, one challenge with using LOADEST models to estimate monthly or annual mean *E. coli* concentrations will be the potentially large uncertainty around these estimates. LOADEST models might be expected to model physical constituents with better performance, but the complexities of a large upstream drainage basin are important to consider. The complexity of a large upstream drainage basin means that continued investigation into basin explanatory variables is needed with special consideration for each specific water-quality constituent of interest. Additional basin explanatory variables, such as upstream turbidity or other continuously monitored water-quality constituents and physical properties on the Missouri River or large tributaries, information on water release from Gavins Point Dam (not shown in figures), or land use information, might help create a better model of the water-quality constituents from the most upstream site, NP Dodge, which might also improve all subsequent downstream models. Data collected for explanatory variables not used in the *E. coli* concentration models described in this report, may still be necessary to model other water-quality constituents. These include LTCP progress, CSO overflows, disinfection season, and other continuously monitored constituents and physical properties (such as specific conductance, temperature, and pH). Additional continuous water-quality explanatory variables, such as differences in continuous water-quality constituents and physical properties from upstream sites to downstream sites, might also be considered in future analyses.

LOADEST regression models were able to model *E. coli* concentration adequately with the datasets for 2012–16, and models likely would improve with a larger (longer term) dataset. With larger datasets, future analyses could consider LOADEST regression models for various river runoff conditions: local dry weather, Missouri River stable; local dry weather, Missouri River elevated; local wet weather, Missouri River stable; and local wet weather, Missouri River elevated. These different conditions likely affect how the explanatory variables relate to *E. coli* concentrations. Future analysis could also consider using LOADEST differently, such as creating one model with data from all four sites and including an explanatory variable for each site. This approach might produce a stronger model with better estimation ability and would be similar to a fixed-effect model. In addition, other modeling software options are available for larger datasets (greater than 10 years) that could be considered once the LTCP has been fully implemented and more years of data become available. One potential model application is Weighted Regressions on Time, Discharge, and Season (Hirsch and others, 2015), which might be applicable for constituents that do not require explanatory variables other than time, discharge (streamflow), and season and that have 10 or more years of data. An additional R-package, seawaveQ (Ryberg and Vecchia, 2013), has enhanced options for modeling seasonality, which is typically used for pesticides. This R-package does not necessarily require a large dataset but could be considered for future data analysis of other water-quality constituents being collected. Finally, once a larger dataset is acquired, more direct comparisons of samples between sites can be made in addition to modeling results. Direct comparisons of samples likely will have more power to detect changes between sites and over time, especially for *E. coli* because modeling estimates have such large uncertainty associated with them. These direct comparisons can be made in part because all four sites are typically being sampled on the same day. One example of analysis that could be used to directly compare the samples would be a seasonal Mann-Kendall test to detect changes over time. These direct comparison methods might be used in addition to modeling because modeling helps us understand what factors affect the changes detected. The modeling application chosen for future analysis will likely depend on the water-quality constituent being modeled, the number of censored values in the dataset, and the length of the dataset.

During the development of *E. coli* concentration models, it was recognized that targeting local runoff might not result in samples being collected during all extreme conditions including high turbidity or high streamflow; therefore, future sampling efforts might consider these variables in addition to precipitation when planning sampling events. High turbidity and high streamflow might occur during runoff events from upstream tributaries that are located downstream from Gavins Point Dam (not shown in figures).

Although we were able to produce adequate models while constraining model development by first developing the best upstream model and using the explanatory variables from that model as the basis for all subsequent downstream models, some of these explanatory variables were insignificant in the downstream models. In some cases, including them resulted in multicollinearity with other explanatory variables that might have improved the model. Further investigation may be necessary to determine the best modeling approach.

The use of similar models for all sites could be evaluated against developing individual highly precise models at each site and making comparisons between sites with potentially very different models. The development of models in downstream order could be evaluated against developing the best model for each site and then basing the final models on the explanatory variables common between all best fit models. Another topic for consideration for future model development is the explanatory variable selection process. This initial analysis used strictly empirical methods to select explanatory variables, meaning that explanatory variables were only included in the models if they improved the metrics that were used to evaluate model quality and not because they made logical sense to include. However, one focus of the modeling effort is to better understand the water-quality changes in the Missouri River in relation to the implementation of the city of Omaha's LTCP. The best models selected did not include explanatory variables that measured the LTCP progress (number of CSOs discharging upstream from sampling point or number of LTCP projects completed) because these did not improve model diagnostics. These explanatory variables did not improve the models possibly because of the small dataset, the variability in *E. coli* concentration in the system, and the multicollinearity between other explanatory variables. Future modeling efforts might consider using some nonsystematic subjectivity during the explanatory variable selection process and include measurements of the LTCP progress even if these explanatory variables do not significantly improve the models. This approach might be more useful taking into consideration the intent of future models to understand the water quality of the Missouri River (nutrients, biological oxygen demand, suspended solids, and *E. coli*), how it is changing with time, and how it changes upstream from the city of Omaha to downstream. This initial analysis used strictly empirical methods for model selection because of the small dataset and the poor performance of models that used more logical explanatory variables. However, future analysis completed on a larger dataset could re-evaluate a better method for selecting explanatory variables for the models.

Summary

The city of Omaha, Nebraska, has a combined sewer system in some areas of the city. In Omaha, Nebr., a moderate amount of rainfall will lead to the combination of stormwater and untreated sewage or wastewater being discharged directly into the Missouri River and Papillion Creek and is called a combined sewer overflow (CSO) event. In 2009, the city of Omaha began the implementation of their Long Term Control Plan (LTCP) to mitigate the effects of CSO events on the Missouri River and Papillion Creek. As part of the LTCP implementation, in 2012, the city partnered with the U.S. Geological Survey to begin water-quality monitoring in the Missouri River. Since 2012, monthly discrete water-quality samples have been collected from the Missouri River at four sites. The four sites were chosen based on location of the site, CSO outfalls, wastewater treatment plants, and tributary locations. The 4 sites include 1 site upstream from the city of Omaha, 2 sites within the city, and 1 site downstream from the city. At 3 of the 4 sites, selected water-quality constituents and physical properties have been monitored continuously. These discrete water-quality samples and continuous water-quality monitoring data (from July 2012 to 2020) are being collected to better understand the water quality of the Missouri River, how it is changing with time, how it changes upstream from the city of Omaha to downstream, and how it varies during base flow conditions and during periods of runoff.

The purpose of this report is to document the development of *Escherichia coli* (*E. coli*) concentration models for the four Missouri River sites. This report describes the initial data analysis and a modeling approach. Analysis was completed using the first 5 years of data (July 2012 through September 2016) to determine if the current sampling and analysis approach is sufficient to meet future analysis goals and to understand if proposed models such as Load Estimator (LOADEST) models will be able to represent water-quality changes in the Missouri River.

During nonice conditions, *E. coli* samples were collected by hand dipping the sample bottle at the midpoint of streamflow in the channel. Wet weather sampling was always targeted. An IDEXX Quantitray 2000 system was used for determination of *E. coli* concentrations. Included in this analysis are 47 *E. coli* samples per site, collected between July 2012 and September 2016.

Multiple linear regression models were developed to estimate *E. coli* concentrations in the Missouri River using LOADEST as implemented in the R statistical software package rloadest. A set of explanatory variables, including streamflow and streamflow anomalies, precipitation, information about CSOs, and continuous water quality, were evaluated for potential inclusion in regression models. Hourly precipitation data were totaled to get daily values and the antecedent precipitation index (API) was calculated. Turbidity and precipitation data were transformed so that the distribution was as close as possible to a normal distribution.

The model for the Missouri River at NP Dodge Park at Omaha, Nebr. (USGS station 412126095565201; hereafter referred to as "NP Dodge")—the most upstream site—was developed first with the intention of developing the best model to predict *E. coli* concentration coming into the Omaha reach. This model was intended to include basin explanatory variables. Model development for the downstream sites included the explanatory variables used in the NP Dodge model as well as local explanatory variables. The best model at NP Dodge included basin explanatory variables of upstream API measured at Tekamah, Nebr.; decimal time; season; and turbidity. The best model at Missouri River at Freedom Park Omaha, Nebr. (USGS station 411636095535401; hereafter "Freedom Park") included the same explanatory variables as the NP Dodge model with the addition of turbidity anomalies and flow anomalies. The best models at the two downstream sites (Missouri River near Council Bluffs, Iowa, USGS station 06610505 and Missouri River near La Platte, Nebr., USGS station 410333095530101) included the same explanatory variables as the Freedom Park model with the addition of local antecedent precipitation index as measured at Eppley Airport in Omaha, Nebr., and additional turbidity and flow anomalies. Many potential explanatory variables were not included in the selected models for several reasons. Some explanatory variables were excluded because of multicollinearity with variables already included in the upstream models. The explanatory variable of LTCP projects completed could not be used in the same model as time because of multicollinearity. For the same reason, the explanatory variable disinfection season could not be used in the same model as the explanatory variable season. Other explanatory variables had low predictive power and did not substantially improve the model, such as specific conductance. The final selected models were the best models given our modeling design constraint in which explanatory variables included in the model for the upstream site were included in the downstream models.

Explanatory variables included in the selected models were able to explain 64–75 percent of the variability of *E. coli* concentration in the Missouri River for 2012–16. Explaining 64–75 percent of the variability might be considered low when working with physical constituents (total nitrogen or sediment), but with the natural variability of biological constituents such as *E. coli*, the uncertainty of *E. coli* laboratory measurements, and the added complexity of modeling in such a large drainage basin with multiple sources, these results indicate that the explanatory variables being collected and models such as LOADEST were able to adequately represent water-quality changes in the Missouri River for *E. coli* concentration from 2012 to 2016.

References Cited

City of Omaha, 2014, Update to the long-term control plan for the Omaha Combined Sewer Overflow Control Program: City of Omaha, 550 p. [Also available at http://omahacso.com/files/6814/1450/8302/Final_Omaha_LTCPUpdate-Appendices_Oct2014.pdf.]

City of Omaha, 2017, City of Omaha combined sewer overflow annual report, NPDES permit no. NE0133680 October 1, 2016, through September 30, 2017: City of Omaha, 275 p. [Also available at http://omahacso.com/files/5815/1570/8558/2017_CSO_Annual_Report_FINAL_Web.pdf.]

Clean Solutions for Omaha, 2017, Clean Solutions for Omaha: City of Omaha web page, accessed November 2017 at http://omahacso.com/.

Cohen, A.C., 1950, Estimating the mean and variance of normal populations for singly truncated and doubly truncated samples: Annals of Mathematical Statistics, v. 21, no. 4, p. 557–569. [Also available at https://doi.org/10.1214/aoms/1177729751.]

Cohen, A.C., 1976, Progressively censored sampling in the three parameter log-normal distribution: Technometrics, v. 18, no. 1, p. 99–103. [Also available at https://doi.org/10.2307/1267922.]

Cohn, T.A., 1988, Adjusted maximum likelihood estimation of the moments of lognormal populations from type I censored samples: U.S. Geological Survey Open-File Report 88–350, 34 p. [Also available at https://doi.org/10.3133/ofr88350.]

Densmore, B.K., and Hall, B.M., 2020, Modeling *Escherichia coli* in the Missouri River near Omaha, Nebraska, 2012–16—Model inputs and outputs: U.S. Geological Survey data release, https://doi.org/10.5066/P97S6WSV.

Heggen, R.J., 2001, Normalized antecedent precipitation index: Journal of Hydrologic Engineering, v. 6, no. 5, p. 377–381. [Also available at https://doi.org/10.1061/(ASCE)1084-0699(2001)6:5(377).]

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p. [Also available at https://doi.org/10.3133/twri04A3.]

Hirsch, R.M., Archfield, S.A., and De Cicco, L.A., 2015, A bootstrap method for estimating uncertainty of water quality trends: Environmental Modelling & Software, v. 73, p. 148–166. [Also available at https://doi.org/10.1016/j.envsoft.2015.07.017.]

IDEXX Laboratories, Inc., 2018, Quanti-Tray System—Take the guess work out of bacterial counts: IDEXX Laboratories, Inc., web page, accessed December 2018 at https://www.idexx.com/en/water/water-products-services/quanti-tray-system/.

Ishii, S., and Sadowsky, M.J., 2008, *Escherichia coli* in the environment—Implications for water quality and human health: Microbes and Environments, v. 23, no. 2, p. 101–108. [Also available at https://doi.org/10.1264/jsme2.23.101.]

Lee, C.J., Murphy, J.C., Crawford, C.G., and Deacon, J.R., 2017, Methods for computing water-quality loads at sites in the U.S. Geological Survey National Water Quality Network: U.S. Geological Survey Open-File Report 2017–1120, 20 p., accessed January 2018 at https://doi.org/10.3133/ofr20171120.

Lorenz, D.L., 2015, smwrBase—An R package for managing hydrologic data, version 1.1.1: U.S. Geological Survey Open-File Report 2015–1202, 7 p., accessed March 2018 at https://doi.org/10.3133/ofr20151202.

Lorenz, D.L., 2017a, Application 1—Analysis of an uncensored constituent using a predefined model: U.S. Geological Survey rloadest index, accessed September 2017 at https://rdrr.io/github/USGS-R/rloadest/f/inst/doc/app1.pdf.

Lorenz, D.L., 2017b, Application 6—Regression model for concentration: U.S. Geological Survey rloadest index, accessed September 2017 at https://rdrr.io/github/USGS-R/rloadest/f/inst/doc/app6.pdf.

Lorenz, D.L., Runkel, R.L., and De Cicco, L., 2013, rloadest—USGS water science R functions for LOAD ESTimation of constituents in rivers and streams, v 0.4.1: GitHub, Inc., web page, accessed September 2017 at https://github.com/USGS-R/rloadest.

Mayo, J.W., and Leib, K.J., 2012, Flow-adjusted trends in dissolved selenium load and concentration in the Gunnison and Colorado Rivers near Grand Junction, Colorado, water years 1986–2008: U.S. Geological Survey Scientific Investigations Report 2012–5088, 33 p. [Also available at https://doi.org/10.3133/sir20125088.]

National Center for Environmental Information, 2017, Climate data online—Asheville, N. Car.: National Oceanic and Atmospheric Administration, digital data, accessed August 22, 2017, at https://www.ncdc.noaa.gov/cdo-web/search.

National Center for Environmental Information, 2018, Climate data online—Asheville, N. Car.: National Oceanic and Atmospheric Administration, digital data, accessed January 31, 2018, at https://www.ncdc.noaa.gov/cdo-web/search.

Nebraska Department of Environmental Quality, 2014, Title 117—Nebraska Surface Water Quality Standards: State of Nebraska, 261 p., accessed May 14, 2018, at http://deq.ne.gov/RuleAndR.nsf/pages/PDF/%24FILE/Title117_2014.pdf.

Runkel, R.L., Crawford, C.G., and Cohn, T.A., 2004, Load estimator (LOADEST)—A FORTRAN program for estimating constituent loads in streams and rivers: U.S. Geological Survey Techniques and Methods, book 4, chap. A5, 69 p. [Also available at https://doi.org/10.3133/tm4A5.]

Ryberg, K.R., and Vecchia, A.V., 2012, waterData—An R package for retrieval, analysis, and anomaly calculation of daily hydrologic time series data, version 1.0: U.S. Geological Survey Open-File Report 2012–1168, 8 p., accessed January 2018 at https://doi.org/10.3133/ofr20121168.

Ryberg, K.R., and Vecchia, A.V., 2013, seawaveQ—An R package providing a model and utilities for analyzing trends in chemical concentrations in streams with a seasonal wave (seawave) and adjustment for streamflow (Q) and other ancillary variables: U.S. Geological Survey Open-File Report 2013–1255, 13 p., with 3 appendixes, accessed January 2016 at https://doi.org/10.3133/ofr20131255.

U.S. Army Corps of Engineers, 2013, Missouri River Recovery Program: U.S. Army Corps of Engineers web page, accessed May 7, 2018, at https://www.nwo.usace.army.mil/MRRP/.

U.S. Environmental Protection Agency, 2017, National Pollutant Discharge Elimination System—Combined sewer overflows (CSOs): U.S. Environmental Protection Agency web page, accessed May 7, 2018, at https://www.epa.gov/npdes/combined-sewer-overflows-csos.

U.S. Geological Survey, 2018, USGS water data for the Nation: U.S. Geological Survey National Water Information System database: accessed May 8, 2018, at https://doi.org/10.5066/F7P55KJN.

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, chaps. A1–A10, accessed March 2018 at https://pubs.water.usgs.gov/twri9A.

Vogel, J.R., Frankforter, J.D., Rus, D.L., Hobza, C.M., and Moser, M.T., 2009, Water quality of combined sewer overflows, stormwater, and streams, Omaha, Nebraska, 2006–07: U.S. Geological Survey Scientific Investigations Report 2009–5175, 152 p. plus appendixes. [Also available at https://doi.org/10.3133/sir20095175.]

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 May 18, 2018, at https://doi.org/10.3133/tm1D3.

For more information about this publication, contact:
Director, USGS Nebraska Water Science Center
5231 South 19th Street Lincoln, NE 68512
402–328–4100

For additional information, visit:
https://www.usgs.gov/centers/ne-water

Publishing support provided by the Rolla and Sacramento Publishing Service Centers

USGS

Densmore and others—**Modeling *Escherichia coli* in the Missouri River near Omaha, Nebraska, 2012–16**—Scientific Investigations Report 2020–5045

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20205045