Appendix 9. Model Archive Summary for *Escherichia coli* Density at U.S. Geological Survey Station 07144780, North Fork Ninnescah River above Cheney Reservoir, Kansas, during November 14, 2015, through September 30, 2021

This model archive summary summarizes the *Escherichia coli* (*E. coli*) model developed to compute 15-minute, hourly, or daily *E. coli* densities during November 14, 2015, onward. This model supersedes all prior models used during this period. The methods follow U.S. Geological Survey (USGS) guidance as referenced in relevant Office of Surface Water/Office of Water Quality Technical Memoranda and USGS Techniques and Methods, book 3, chapter C4 (Rasmussen and others, 2009; U.S. Geological Survey, 2016).

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

Site and Model Information

Site number: 07144780

Site name: North Fork Ninnescah River above Cheney Reservoir, Kansas

Location: Lat 37°51'45", long 98°00'49" referenced to North American Datum of 1927, in NE 1/4 SE 1/4 NE 1/4 sec.19, T.25 S., R.6 W., Reno County, Kans., hydrologic unit 11030014, on right bank at upstream side of county highway bridge, 10 miles south of Hutchinson, 18.1 miles upstream from Cheney Dam.

Equipment: A YSI, Inc., EXO water-quality monitor (YSI, Inc., 2017) equipped with sensors for water temperature, specific conductance, dissolved oxygen, pH, and turbidity was installed November 14, 2015. The EXO monitor was installed in a 4-inch-diameter metal or polyvinyl chloride (or PVC) pipe suspended from the downstream side of the bridge in the deepest, fastest flowing water. Measurements from the EXO were recorded every 15 minutes to hourly and transmitted hourly via satellite. Real-time stage was measured using a Design Analysis Water Log H–350/355 nonsubmersible pressure transducer.

Date model was created: August 9, 2022

Model calibration data period: April 19, 2016, through August 12, 2021 (dataset consisted of 25 discrete water-quality samples).

Model application date: November 14, 2015, onward (date of EXO continuous water-quality monitor installation).

Model developed by: Ariele Kramer, USGS, Lawrence, Kans. (akramer@usgs.gov)

Model Calibration Dataset

All data were collected using USGS protocols (U.S. Geological Survey, 2006; Wagner and others, 2006; Bennett and others, 2014) and are stored in the USGS National Water Information System database (https://doi.org/10.5066/F7P55KJN; U.S. Geological Survey, 2022). Potential explanatory variables evaluated individually and in combination were water temperature, specific conductance, pH, dissolved oxygen, turbidity, seasonality (sine and cosine variables), and streamflow.

The regression model is based on 25 concomitant values of discretely collected *E. coli* and continuously measured turbidity and streamflow during April 19, 2016, through August 12, 2021. Discrete samples were collected throughout the range of continuously observed hydrologic conditions. One sample had an *E. coli* density that was reported as a right-censored (greater than reported) value due to analytical dilution methods. All potential explanatory variables were time interpolated within the 15-minute to hourly continuous record based on the discrete sample time. The maximum time span between two continuous data points used for interpolation was 4 hours (to preserve the sample dataset, field monitor averages obtained during sample collection were used for model development data if no continuous data were available or if gaps larger than 4 hours in the continuous data record resulted in missing interpolated data). Summary statistics and the complete model-calibration dataset are provided below. Potential outliers were identified using the methods described in Rasmussen and others (2009) and Helsel and others (2020). All potential outliers were investigated by reviewing sample collection information sheets and laboratory reports; if there were no clear issues, explanations, or conditions that would cause a result to be invalid for model calibration, the sample was retained in the dataset. One right-censored sample was examined more closely due to exceeding laboratory detection limits and was ultimately removed from the data set due to laboratory processing irregularities.

E. coli Sampling Details

Discrete water-quality samples were collected over a range of hydrologic conditions primarily using a grab sample collection technique (U.S. Geological Survey, 2006). Grab samples were collected either instream as a wading sample within 300 feet of the bridge or from the downstream side of the bridge using an autoclaved wide mouth 1-liter polytetrafluoroethylene bottle. Discrete samples were collected on a semifixed to event-based schedule one to six times per year. Samples were analyzed for *E. coli* by the Wichita Municipal Water and Wastewater Laboratory in Wichita, Kans., according to standard methods (Eaton and others, 1995).

Continuous Water-Quality Data

Turbidity was continuously measured (15 minutes to hourly) using a YSI, Inc., EXO multiparameter sonde (YSI, Inc., 2017). The water-quality monitor was operated and maintained according to standard USGS methods (Wagner and others, 2006; Bennett and others, 2014). Discharge was computed using a nonsubmersible pressure transducer following standard USGS methods (Turnipseed and Sauer, 2010; Painter and Loving, 2015). All continuous water-quality data at the North Fork Ninnescah River above Cheney Reservoir are available in near-real time (updated hourly) from the USGS National Water Information System database

(https://doi.org/10.5066/F7P55KJN; U.S. Geological Survey, 2022) using the site number 07144780.

Model Development

Ordinary least squares linear regression was used to develop surrogate regression models that relate continuous water-quality conditions to discretely sampled constituent densities. All regressions were computed using the R software environment (R Core Team, 2020). The data and subsequent regression equation must meet the five assumptions necessary to apply ordinary least squares regression: the dependent variable is linearly related to the explanatory variables, data used to fit the model are representative of the data of interest, the variance of the residuals is constant (homoscedastic), the residuals are independent of the explanatory variables, and the residuals are normally distributed (Helsel and others, 2020). Previously published explanatory variables also were considered for continuity.

Turbidity and streamflow were selected as a good surrogate for *E. coli* based on residual plots, coefficient of determination (R^2), and model standard percentage error (MSPE). Values for the aforementioned statistics were computed and are included below along with all relevant sample data and additional statistical information.

Model Summary

Summary of final *E. coli* (ecoli) regression analysis at USGS site 07144780:

E. coli density-based model:

$$\log_{10}(ecoli) = (1.09 \times \log_{10}(TBY)) + (0.619 \times \log_{10}(Q)) - 0.348,$$

where,

ecoli = *E. coli*, most probable number per 100 milliliters (MPN/100 mL) (USGS parameter code 50468);

TBY = turbidity, monochrome near infra-red light-emitting diode light, 780-900 nanometers, detection angle 90 +-2.5 degrees, formazin nephelometric units (FNU) (USGS parameter code 63680);

Q = streamflow, instantaneous, cubic feet per second (ft^3/s) (USGS parameter code 00060); and

$\log_{10}$ = decimal logarithm.

The $\log_{10}$-transformed model may be retransformed to the original units so that *E. coli* can be calculated directly. The retransformation introduces a negative bias in the retransformed calculated constituent (Helsel and others, 2020). This bias may be corrected using Duan's bias correction factor (BCF; Duan, 1983; Helsel and others, 2020). For this model, the calculated BCF was 1.63. The retransformed model, accounting for BCF, is as follows:

$$ecoli = (TBY^{1.09} \times Q^{0.619} \times 10^{-0.348}) \times 1.63.$$

Fecal indicator bacteria, such as *E. coli*, can sorb to suspended particles; therefore, turbidity can be a good indicator of fecal indicator bacteria. Including streamflow as an explanatory variable made sense statistically and intuitively since sources of *E. coli* likely come from agricultural non-point sources in areas surrounding the North Fork Ninnescah River.

Extrapolation, defined as computation beyond the range of the model calibration dataset, may be used to extrapolate no more than 10 percent outside the range of the calibration data used to fit the model and is therefore limited. The extrapolation limit for *E. coli* using this model is 50,600 most probable number per 100 milliliters. Computed estimates outside that limit are not supported by the current model calibration dataset.

Model Statistics, Data, and Plots

Definitions

Variable	Explanation
BCF	Bias Correction Factor, used to correct logarithmic bias (Duan 1983)
Cook's D	Cook's distance, a measure of influence (Helsel and others, 2020)
DFFITS	Difference in fits, a measure of influence (Helsel and others, 2020)
ecoli	*E. coli*, most probable number per 100 milliliters (MPN/100 mL) (USGS parameter code 50468)
E.vars	Explanatory variables
Leverage	An outlier's measure in the x direction (Helsel and others, 2020)
LOESS	Local polynomial regression fitting (Helsel and others, 2020)
logecoli	*E. coli*, most probable number per 100 milliliters (MPN/100 mL), $\log_{10}$ transformed
logTBY	Turbidity, monochrome near infra-red LED light, 780-900 nm, formazin nephelometric units (FNU) (USGS parameter code 63680), $\log_{10}$ transformed
logQ	Streamflow, instantaneous, cubic feet per second (ft^3/s) (USGS parameter code 00060), $\log_{10}$ transformed
MSE	Model standard error (Helsel and others, 2020)
MSPE	Model standard percentage error (Helsel and others, 2020)
Pr(>|*t*|)	The probability that the independent variable has no effect on the dependent variable (Helsel and others, 2020)
RMSE	Root mean square error (Helsel and others, 2020)
t value	Student's *t* value; the coefficient divided by its associated standard error (Helsel and others, 2020)

Model

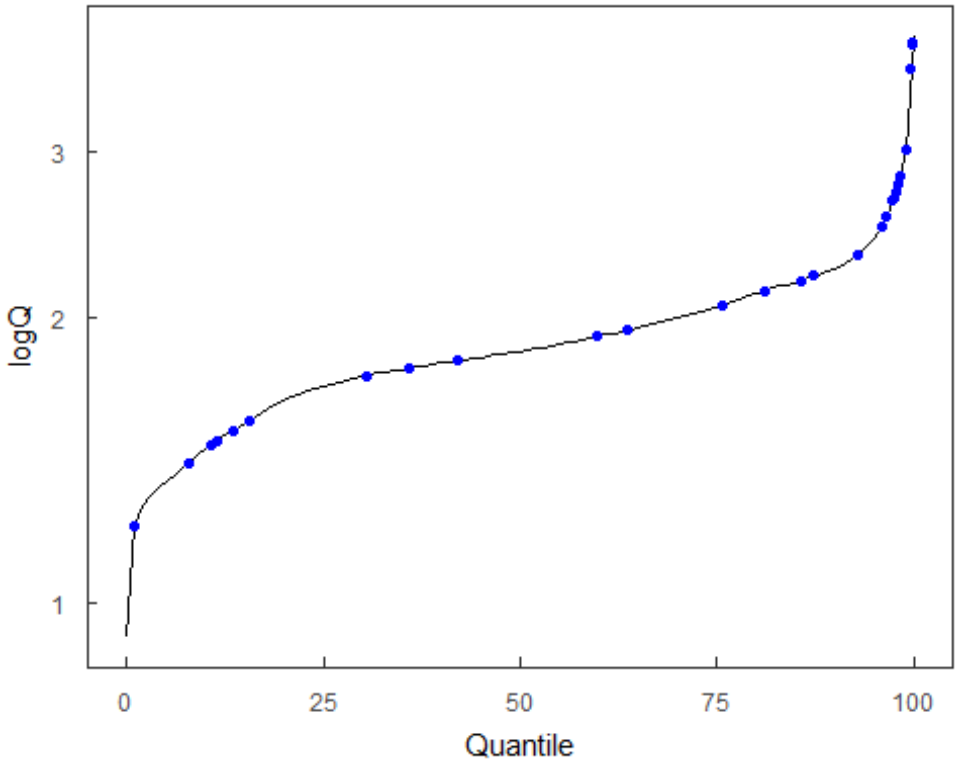
$$\log_{10}(ecoli) = (1.09 \times \log_{10}(TBY)) + (0.619 \times \log_{10}(Q)) - 0.348$$

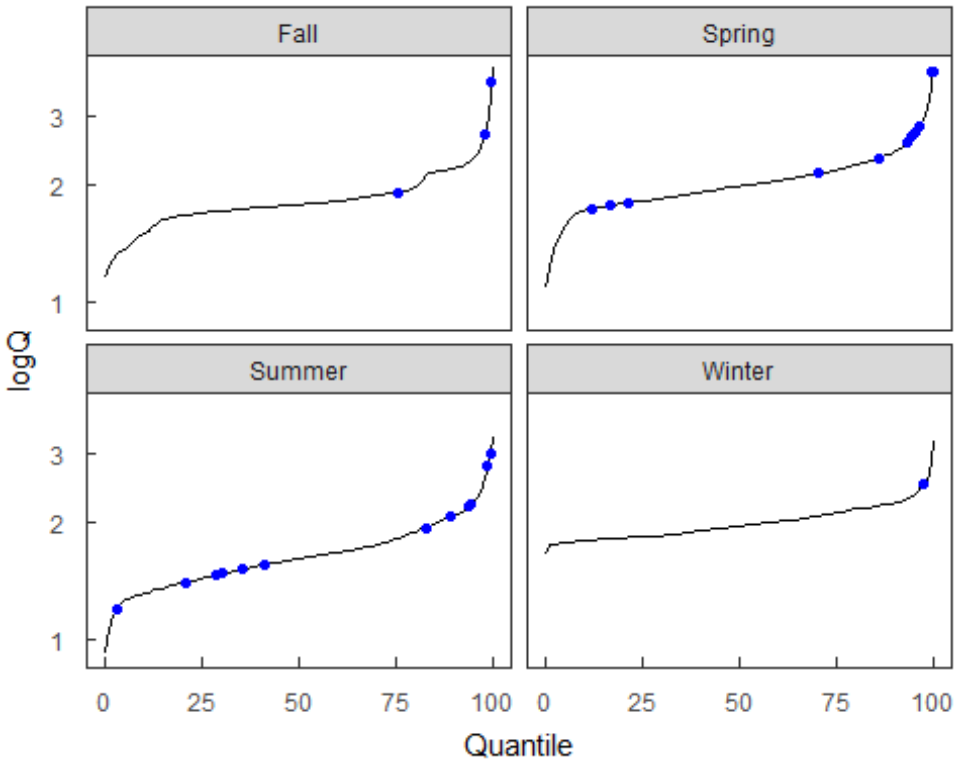
Variable summary statistics

Variable	Minimum	Q1	Median	Mean	Q3	Maximum
ecoli	12	140	440	5,030	6,900	46,000
logecoli	1.08	2.15	2.64	2.9	3.84	4.66

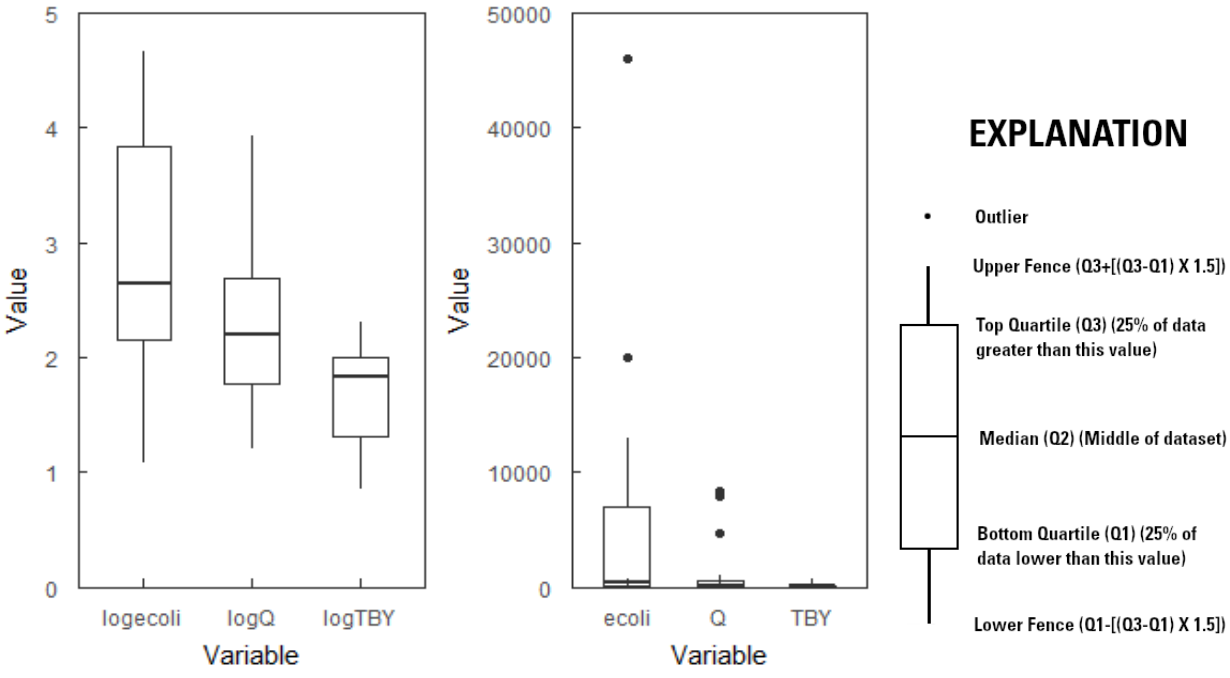
Variable	Minimum	Q1	Median	Mean	Q3	Maximum
logQ	1.21	1.77	2.19	2.31	2.69	3.92
logTBY	0.847	1.3	1.84	1.66	2	2.3
Q	16	59.2	155	1,040	485	8,290
TBY	7.03	20.1	68.7	66.4	99.3	201

Duration plots

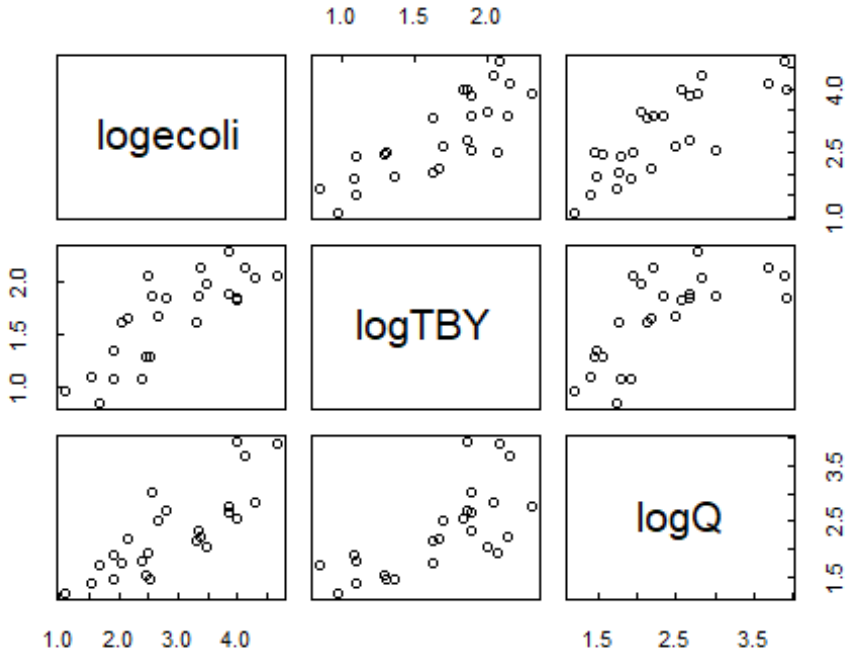



Fall
Spring
Summer
Winter
logQ
Quantile
0
25
50
75
100
1
2
3

Box plots


Value
Variable
logecoli
logQ
logTBY
ecoli
Q
TBY
EXPLANATION
Outlier
Upper Fence (Q3+[(Q3-Q1) X 1.5])
Top Quartile (Q3) (25% of data greater than this value)
Median (Q2) (Middle of dataset)
Bottom Quartile (Q1) (25% of data lower than this value)
Lower Fence (Q1-[(Q3-Q1) X 1.5])

Scatter plots



The x- and y-axis labels for a given bivariate plot are defined by the intersecting row and column labels.

Basic model statistics

Statistic	Value
Observations	25
R^2	0.767
Adjusted R^2	0.746
RMSE	0.49
Upper MSPE (90%)	209
Lower MSPE (90%)	67.7
BCF	1.63

Model coefficients

	Estimate	Standard Error	t value	Pr(>|t|)
(Intercept)	-0.3482747	0.4117628	-0.8458139	0.4067624
logTBY	1.0895143	0.3381046	3.2224181	0.0039193
logQ	0.6191306	0.1881118	3.2912910	0.0033301

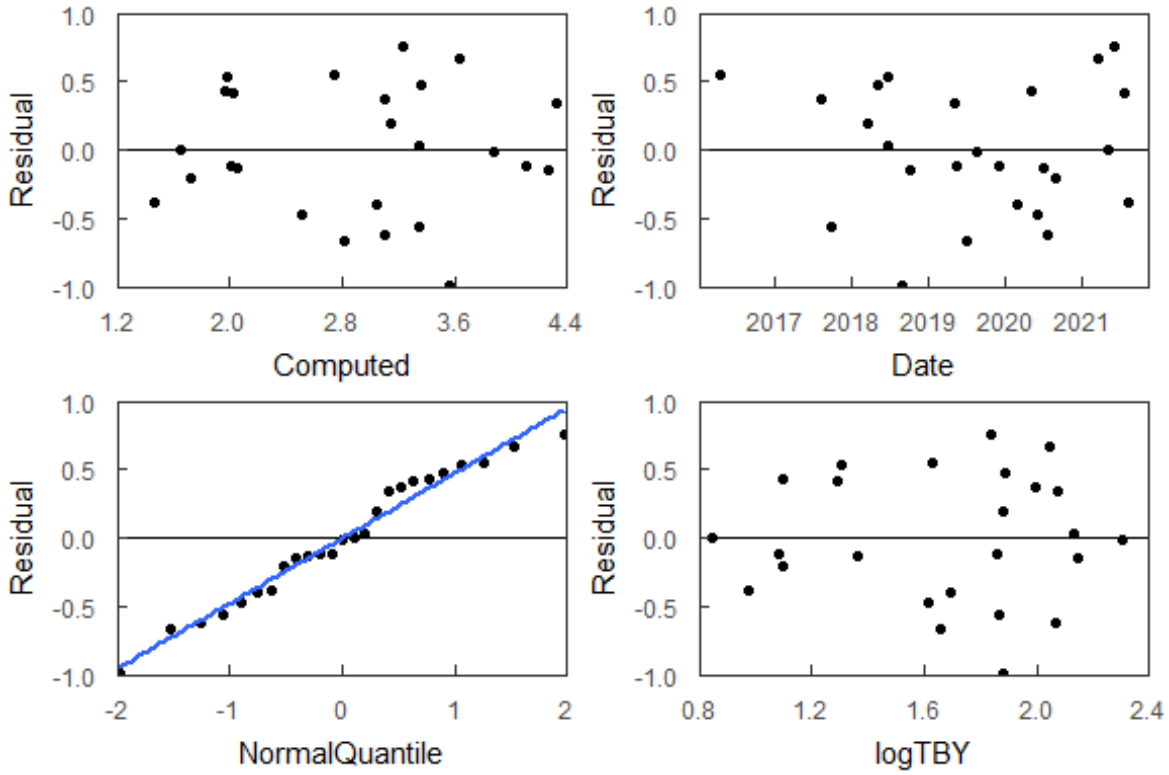
Correlation matrix

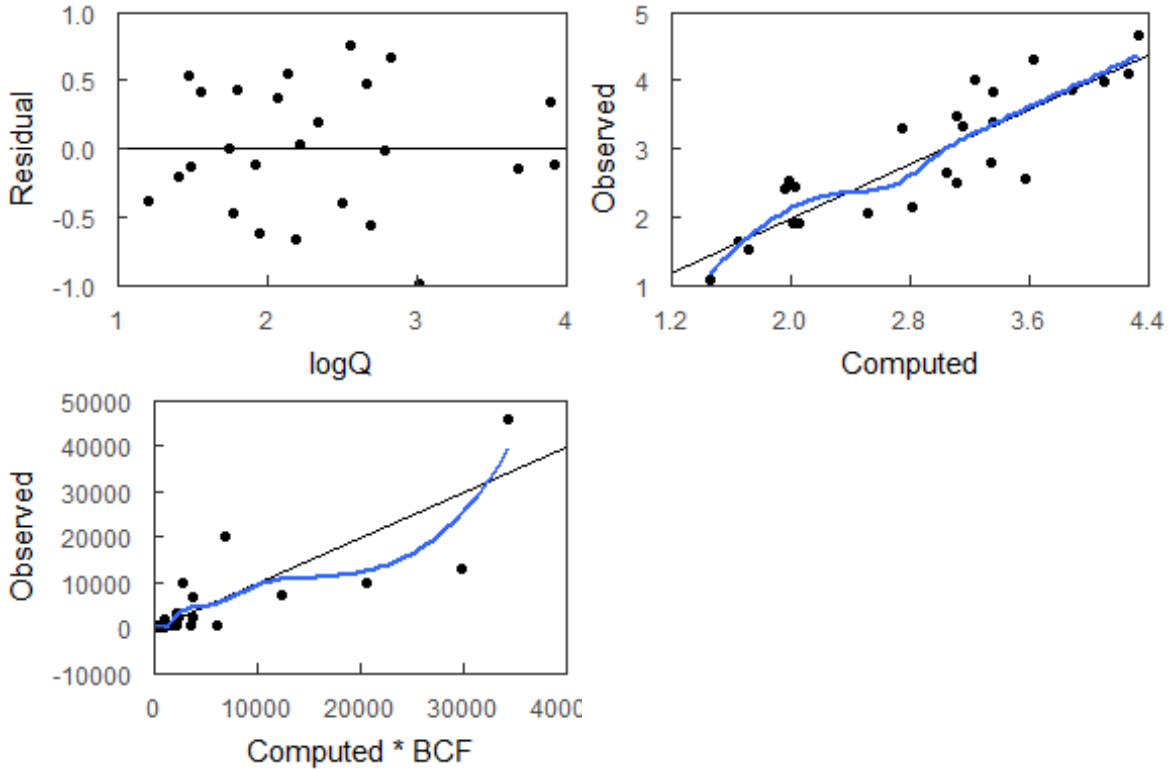
	logecoli	logTBY	logQ
logecoli	1.0000000	0.8073778	0.8103184
logTBY	0.8073778	1.0000000	0.7066090
logQ	0.8103184	0.7066090	1.0000000

Outlier test criteria

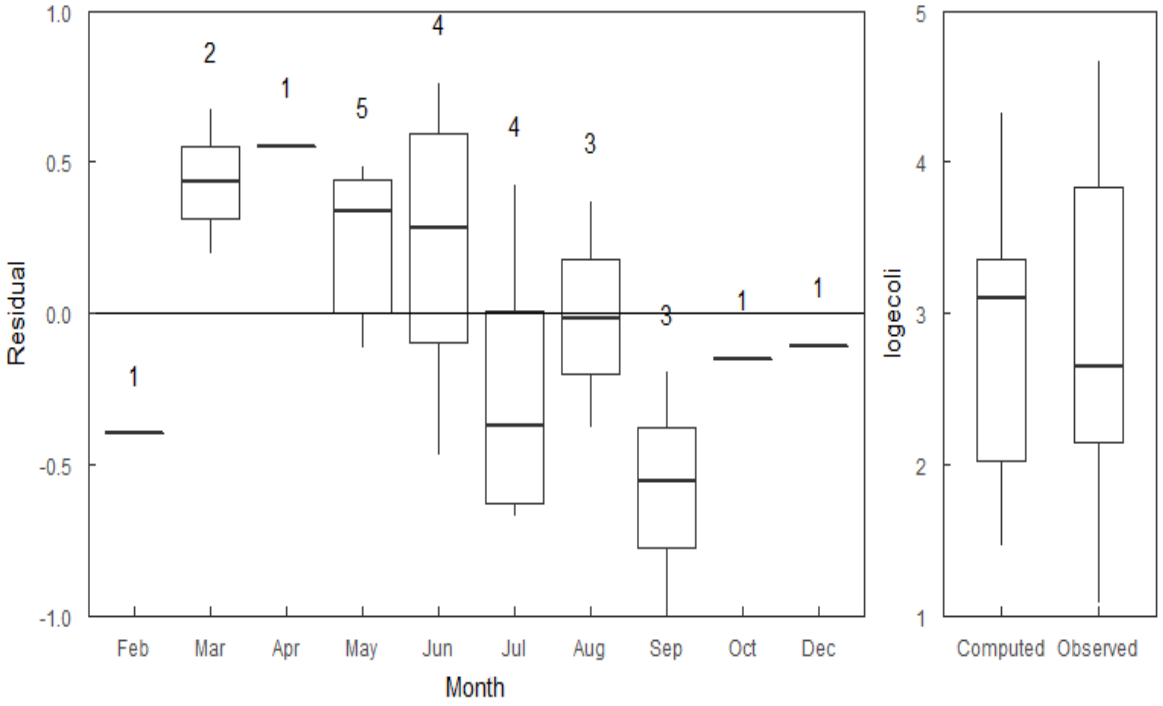
Leverage	DFFITS	CooksD
0.36	0.6928	0.2606

Statistical plots





The blue line shows the locally estimated scatterplot smoothing (LOESS). The black dots correspond to observed values. The black line represents the 1:1 line.



EXPLANATION

1 **Number of values**

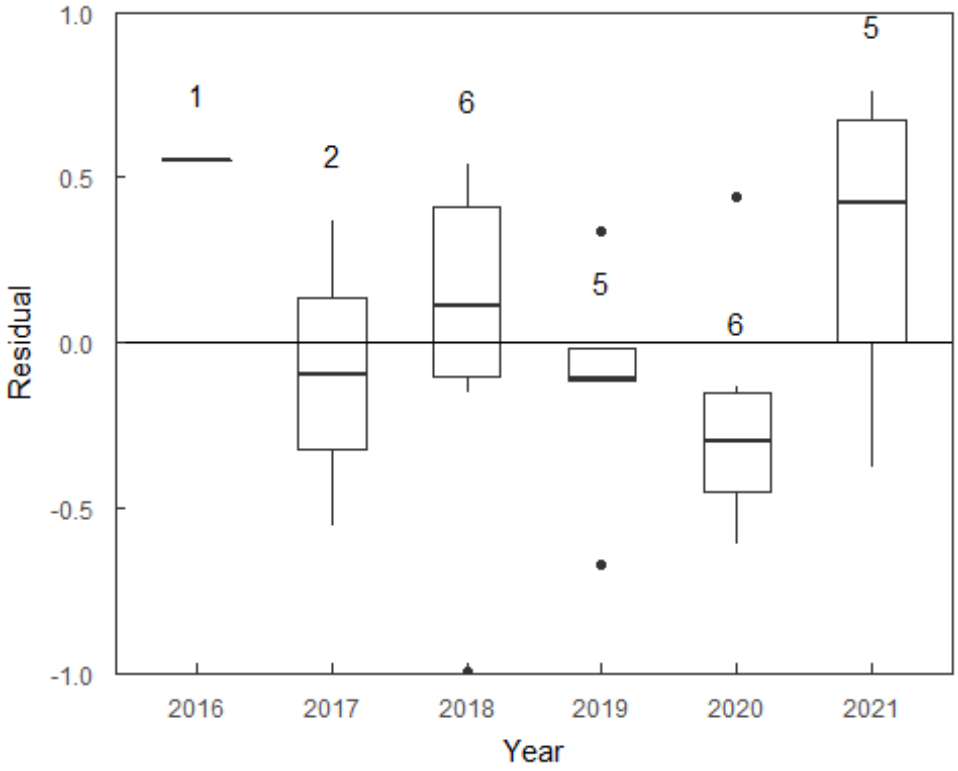
Upper Fence (Q3+[(Q3-Q1) X 1.5])

Top Quartile (Q3) (25% of data greater than this value)

Median (Q2) (Middle of dataset)

Bottom Quartile (Q1) (25% of data lower than this value)

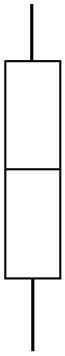
Lower Fence (Q1-[(Q3-Q1) X 1.5])



EXPLANATION

1 Number of values

• Outlier



Upper Fence (Q3+[(Q3-Q1) X 1.5])

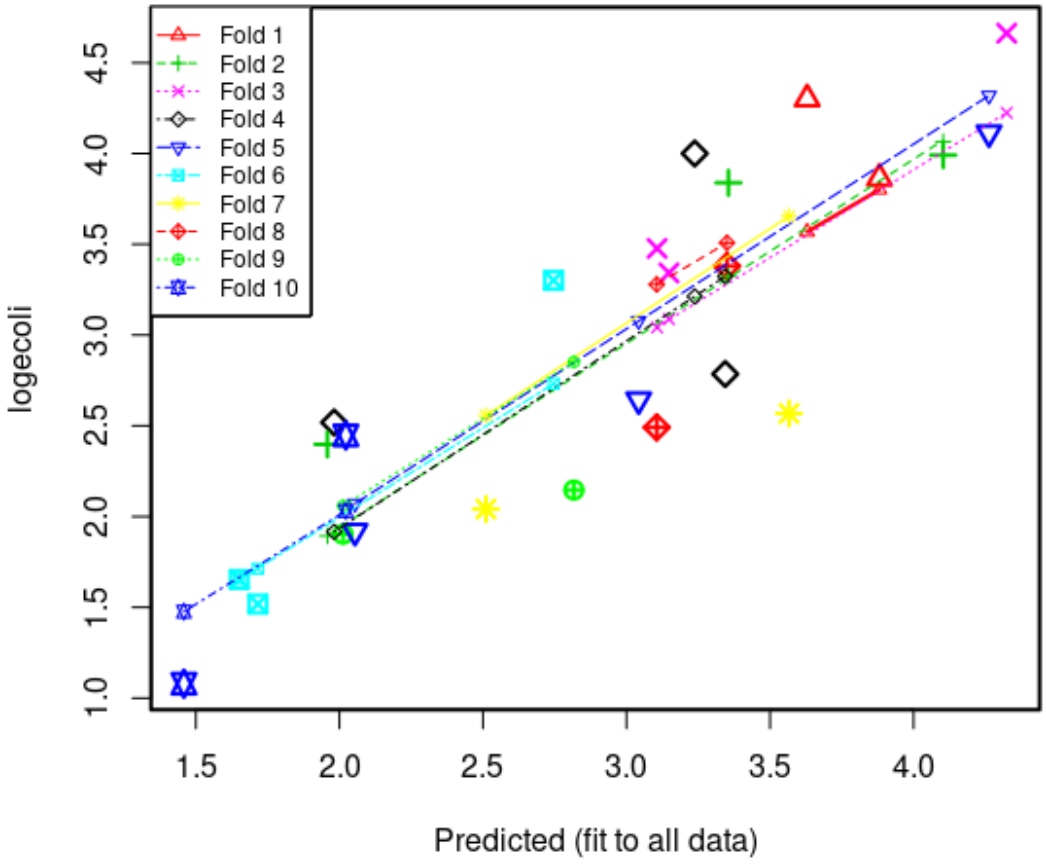
Top Quartile (Q3) (25% of data greater than this value)

Median (Q2) (Middle of dataset)

Bottom Quartile (Q1) (25% of data lower than this value)

Lower Fence (Q1-[(Q3-Q1) X 1.5])

Cross Validation



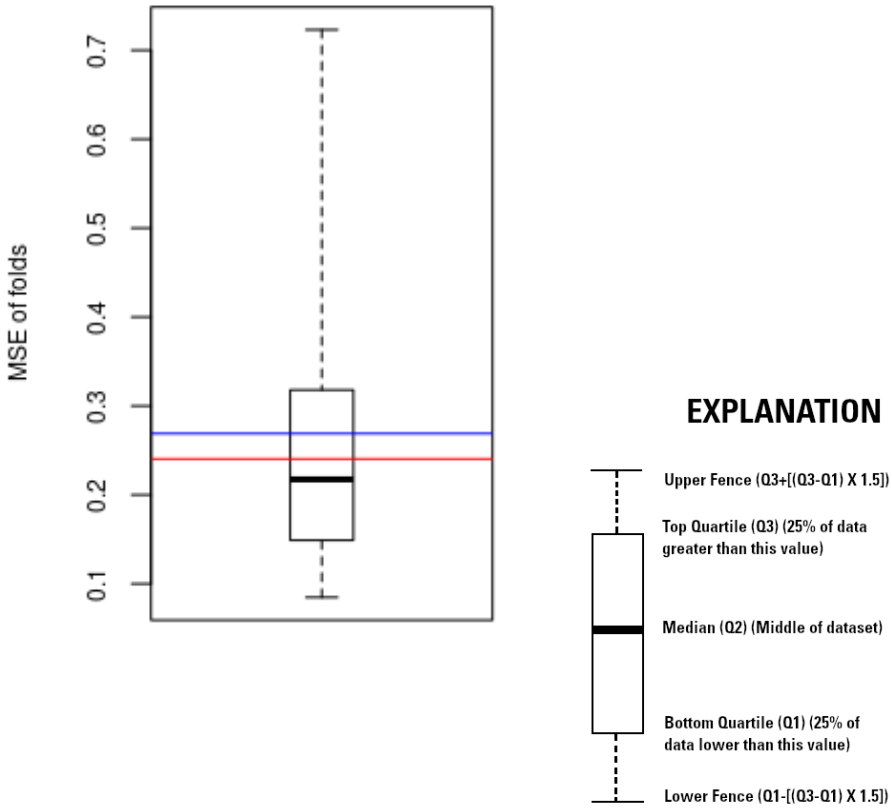
Fold - equal partition of the data (10 percent of the data).

Large symbols – observed value of a data point removed in a fold.

Small symbols – recomputed value of a data point removed in a fold.

Recomputed regression lines – adjusted regression line with one fold removed.

Statistic	Value
Minimum MSE of folds	0.0847
Median MSE of folds	0.2180
Mean MSE of folds	0.2690
Maximum MSE of folds	0.7230
(Mean MSE of folds) / (Model MSE)	1.1200




```
Red line - Model MSE

Blue line - Mean MSE of folds
```

Model calibration dataset

datetime	logecoli	logTBY	logQ	ecoli	Computed	Retransformed
2016-04-19 10:25:00	3.3	1.62	2.14	2,000	2.75	908
2017-08-11 11:00:00	3.48	2	2.07	3,000	3.11	2,080
2017-09-28 10:30:00	2.79	1.86	2.69	610	3.34	3,600
2018-03-20 10:30:00	3.34	1.88	2.34	2,200	3.15	2,290
2018-05-04 10:00:00	3.84	1.89	2.66	6,900	3.36	3,690
2018-06-21 10:10:00	2.52	1.3	1.47	330	1.98	156
2018-06-26 13:20:00	3.38	2.13	2.22	2,400	3.35	3,650
2018-09-05 09:55:00	2.57	1.88	3.01	370	3.57	5,990
2018-10-09 10:10:00	4.11	2.15	3.67	13,000	4.26	29,800

datetime	logecoli	logTBY	logQ	ecoli	Computed	Retransformed
2019-05-08 12:00:00	4.66	2.07	3.9	46,000	4.32	34,400
2019-05-21 12:30:00	3.99	1.86	3.92	9,800	4.1	20,700
2019-07-08 11:30:00	2.15	1.66	2.19	140	2.82	1070
2019-08-26 11:30:00	3.86	2.3	2.78	7,300	3.88	12,400
2019-12-03 10:20:00	1.9	1.08	1.91	80	2.01	168
2020-02-26 10:30:00	2.64	1.69	2.5	440	3.04	1,800
2020-05-07 10:30:00	2.4	1.09	1.8	250	1.96	148
2020-06-04 10:20:00	2.04	1.62	1.77	110	2.51	528
2020-07-08 11:00:00	1.92	1.36	1.48	83	2.05	185
2020-07-21 10:10:00	2.49	2.07	1.94	310	3.11	2,080
2020-09-03 10:20:00	1.52	1.1	1.4	33	1.72	84.7
2021-03-23 11:40:00	4.3	2.04	2.83	20,000	3.63	6,940
2021-05-10 10:50:00	1.65	0.847	1.74	45	1.65	72.9
2021-06-01 10:40:00	4	1.84	2.56	10,000	3.24	2,810
2021-07-22 10:40:00	2.45	1.29	1.56	280	2.02	171
2021-08-12 11:00:00	1.08	0.973	1.21	12	1.46	46.8

References Cited

Bennett, T.J., Graham, J.L., Foster, G.M., Stone, M.L., Juracek, K.E., Rasmussen, T.J., and Putnam, J.E., 2014, U.S. Geological Survey quality-assurance plan for continuous water-quality monitoring in Kansas, 2014: U.S. Geological Survey Open-File Report 2014–1151, 34 p. plus appendixes, accessed September 7, 2022, at https://doi.org/10.3133/ofr20141151.

Eaton, A.D., Clesceri, L.S., and Greenberg, A.E., eds., 1995, Standard methods for the examination of water and wastewater (19th ed.): New York, American Public Health Association, 905 p.

Duan, N., 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, n. 383 p. 605–610.

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, Statistical methods in water resources: U.S. Geological Survey Techniques and Methods, book 4, chap. A3, 458 p. [Also available at https://doi.org/10.3133/tm4A3.] [Supersedes USGS Techniques of Water-Resources Investigations, book 4, chap. A3, version 1.1.]

Painter, C.C., and Loving, B.L., 2015, U.S. Geological Survey quality-assurance plan for surface-water activities in Kansas, 2015: U.S. Geological Survey Open-File Report 2015-1074, 33 p., https://doi.org/10.3133/ofr20151074.

R Core Team, 2020, R—A language and environment for statistical computing: R Foundation for Statistical Computing software release (version 4.0.2), accessed September 7, 2022, at https://www.R-project.org/.

Rasmussen, P.P., Gray, J.R., Glysson, G.D., and Ziegler, A.C., 2009, Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity-sensor and streamflow data: U.S. Geological Survey Techniques and Methods, book 3, chap. C4, 52 p. [Also available at https://doi.org/10.3133/tm3C4.]

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., accessed July 13, 2022, at https://doi.org/10.3133/tm3A8.

U.S. Geological Survey, 2006, Collection of water samples (ver. 2.0, September 2006): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A4 [variously paged]. [Also available at https://doi.org/10.3133/twri09A4.]

U.S. Geological Survey, 2016, Policy and guidance for approval of surrogate regression models for computation of time series suspended-sediment concentration and loads: U.S. Geological Survey Office of Surface Water Technical Memorandum 2016.07, Office of Water Quality Technical Memorandum 2016.10, 40 p., accessed September 7, 2022, at https://water.usgs.gov/water-resources/memos/memo.php?id=467.

U.S. Geological Survey, 2022, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed September 7, 2022, at https://doi.org/10.5066/F7P55KJN.

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. [Also available at https://doi.org/10.3133/tm1D3.]

YSI, Inc., 2017, EXO user manual—Advanced water quality monitoring platform (rev. G): Yellow Springs, Ohio, YSI, Inc., 154 p., accessed September 7, 2022, at https://www.ysi.com/file%20library/documents/manuals/exo-user-manual-web.pdf.