Appendix 1.23. Model Archive Summary for *Escherichia coli* Bacteria Density at U.S. Geological Survey site 07143672; Little Arkansas River at Highway 50 near Halstead, Kansas, during March 2017 through December 2019

This model archive summary summarizes the *Escherichia coli* bacteria model developed to compute hourly or daily *Escherichia coli* bacteria. Model development methods follow U.S. Geological Survey (USGS) guidance from Office of Surface Water/Office of Water Quality Technical Memoranda and USGS Techniques and Methods, book 3, chap. C4 (Rasmussen and others, 2009).

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: 07143672

Site Name: Little Arkansas River at Highway 50 near Halstead, Kansas

Location: Latitude 38°01'43", longitude 97°32'25" referenced to North American Datum of 1927, in NW 1/4 NE 1/4 NE 1/4 sec.28, T.23 S., R.2 W., Harvey County, Kansas, hydrologic unit 11030012.

Equipment: A Sutron Satlink II High Data Rate Collection Platform and a Design Analysis Water Log H350/355 nonsubmersible pressure transducer transfers real-time stage and water-quality data via satellite. The primary reference gage is a Type-A wire-weight gage located on the downstream bridge guardrail. Check-bar elevation is 33.396 feet. The orifice tube is enclosed in 1.25-inch steel conduit trenched into the ground down to the edge of water, where the orifice emerges from the bank and culminates in a 2-inch open-end orifice tethered to a steel fencepost near the left edge of water. Gage height was measured during May 1998 through December 2019. A YSI 6600 water-quality monitor equipped with water temperature, specific conductance, pH, dissolved oxygen, and turbidity (a YSI Model 6026 [December 1998 through December 2006] and YSI Model 6136 [July 2004 through December 2017]) sensors collected data during May 1998 through December 2017. A YSI EXO2 water-quality monitor equipped with water temperature, specific conductance, pH, dissolved oxygen, turbidity, and fluorescent dissolved organic matter sensors collected data during January 2017 through December 2019. A Hach Nitratax monitor collected nitrate data during February 2017 through December 2019.

Date model was developed: June 1, 2020

Model calibration data period: May 3, 2017 through December 10, 2019

Model Data

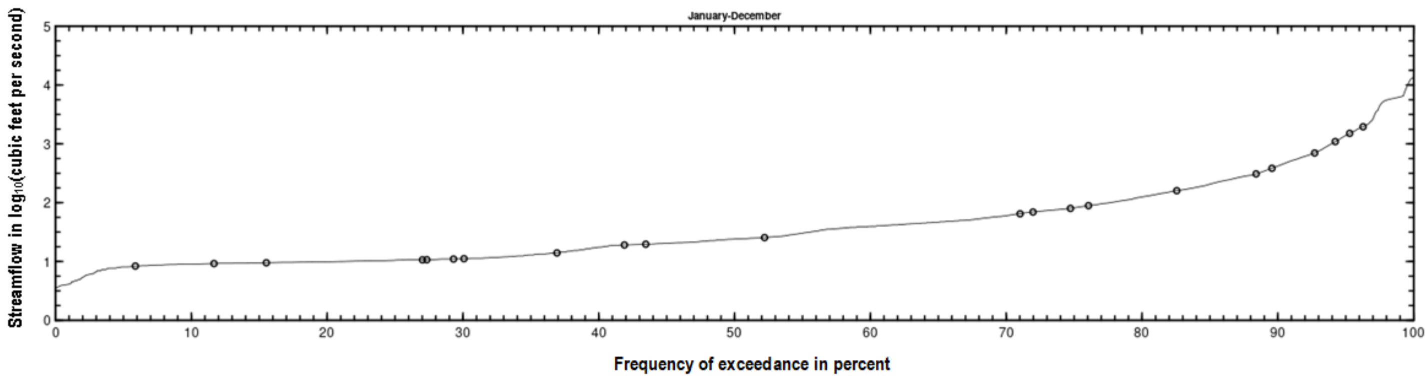
All data were collected using USGS protocols (U.S. Geological Survey, variously dated; Wagner and others, 2006; Sauer and Turnipseed, 2010; Turnipseed and Sauer, 2010) and are stored in the National Water Information System (NWIS) database (U.S. Geological Survey, 2021). Explanatory variables were evaluated individually and in combination. Potential explanatory variables included streamflow, water temperature, specific conductance, pH, dissolved oxygen, YSI EXO2 turbidity, nitrate, and fluorescent dissolved organic matter. Seasonal components (sine and cosine variables) also were evaluated as explanatory variables.

The regression model is based on 22 concomitant values of discretely collected *Escherichia coli* bacteria and continuously measured turbidity during March 2017 through December 2019. Discrete samples were collected over a range of streamflow and turbidity conditions. No samples had densities that were below laboratory detection limits. Summary statistics and the complete model-calibration dataset are provided below. Outliers and influential points were identified using studentized residuals, DFITS, Cook's D (Cook, 1977), and leverage. Outliers in previously published versions of this model (Christensen and others, 2003; Rasmussen and others, 2016) were examined and retained in the dataset if there were no clear issues, explanations, or conditions that would cause a result to be invalid for model calibration. Two samples (collection dates March 30, 2017 and March 14, 2019) were not representative of the dataset and exceeded Cook's D and DFITS outlier criteria and were removed from the model dataset to avoid erroneous inflation of model-computed values at the upper range of surrogate relations. Removing data points based only on outlier criteria may only overestimate the certainty of the model.

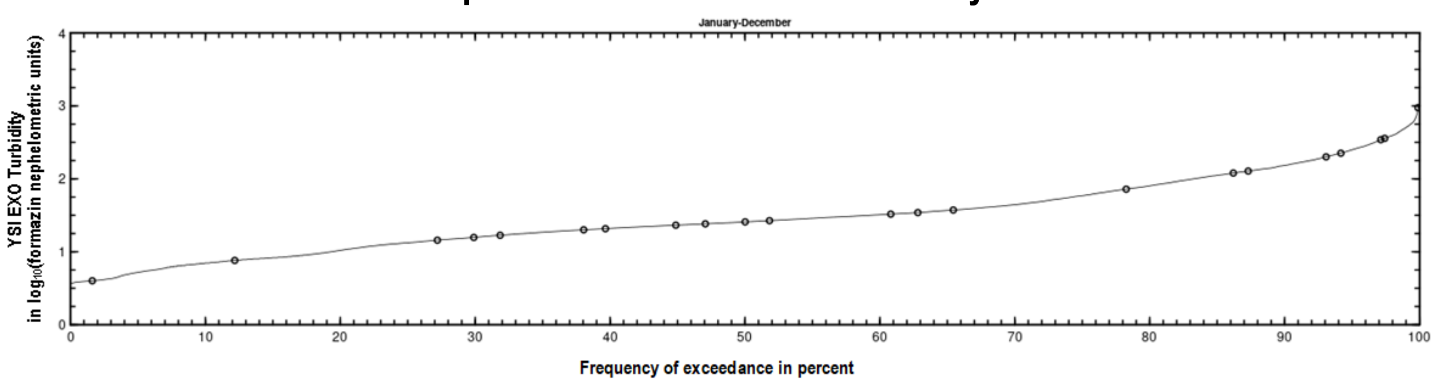
Escherichia coli Bacteria

Discrete samples were collected from the downstream side of the bridge or instream within 50 feet of the bridge using equal-width-increment, multi-vertical, single vertical or grab-dip methods following U.S. Geological Survey (variously dated) and Rasmussen and others (2014). Discrete samples were collected on a semifixed to event-based schedule ranging from 6 to 8 samples per year with a FISP US DH–95 or D–95 with a Teflon bottle, cap, and nozzle depth-integrating sampler, a DH–81 with a Teflon bottle, cap, and nozzle hand sampler or a grab sample with a Teflon bottle depending on sample location. Samples were analyzed for *Escherichia coli* bacteria by the U.S. Geological Survey Kansas Water Science Center.

Escherichia coli Bacteria Samples Plotted on Streamflow Duration Curve



Escherichia coli Bacteria Samples Plotted on YSI EXO Turbidity Duration Curve



Continuous Data

Concomitant turbidity values were time interpolated. If no concomitant continuous data were available within 2 hours of sample collection, the sample was not included in the dataset.

Model Development

Ordinary least squares regression analysis was done using R (version 4.0.0) programming language (R Core Team, 2020) to relate discretely collected *Escherichia coli* bacteria to turbidity and other continuously measured data. The distribution of residuals was examined for normality and plots of residuals (the difference between the measured and model-calculated values) compared to model-computed *Escherichia coli* bacteria were examined for homoscedasticity (departures from zero did not change substantially over the range of model-calculated values). Previously published explanatory variables were also strongly considered for continuity; however, the best explanatory variable(s) were ultimately selected.

Turbidity was selected as the best predictor of *Escherichia coli* bacteria based on residual plots, high coefficient of determination (R^2), and low model standard percentage error (MSPE). Turbidity was positively correlated with *Escherichia coli* bacteria because turbidity measures light scattered by particulates in water.

Model Summary

Summary of final *Escherichia coli* bacteria regression analysis at USGS site number 07143672:

Escherichia coli bacteria-based model:

$$\log_{10}(EC) = 0.964 \times \log_{10}(TBY) + 1.08$$

where,

$\log_{10}$ = logarithm base 10;
EC = *Escherichia coli* bacteria, in colony forming units per 100 milliliters (cfu/100 mL); and
TBY = turbidity, in formazin nephelometric units (FNU)

The log-transformed model may be retransformed to original units so that EC can be calculated directly. The retransformation introduces a bias in the calculated constituent. This bias may be corrected using Duan's bias correction factor (BCF; Duan, 1983). For this model, the calculated BCF is 1.41. The retransformed model, accounting for BCF is:

$$EC = 16.95 \times TBY^{0.964}$$

Model Statistics, Data, and Plots

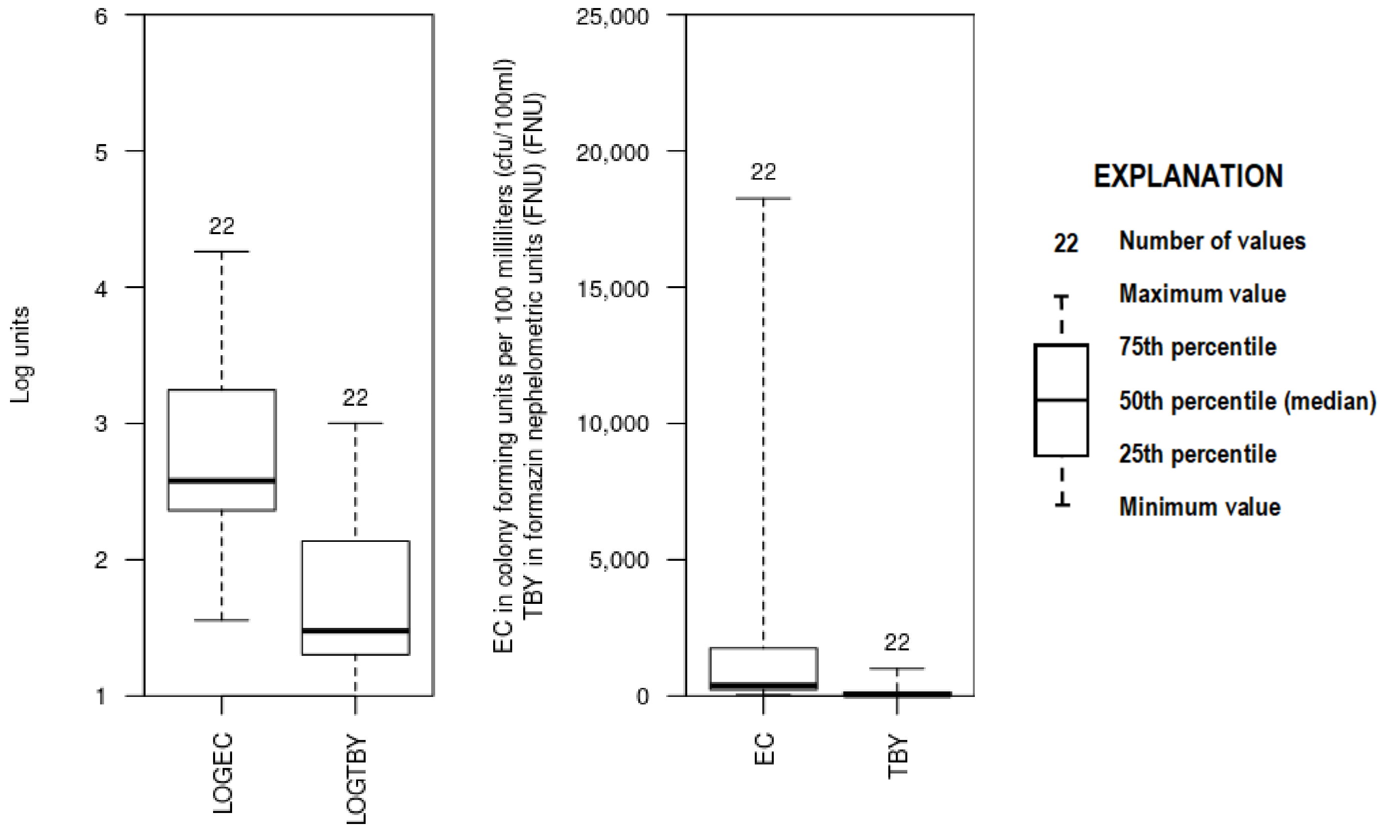
Model

LOGEC = + 0.964 * LOGTBY + 1.08

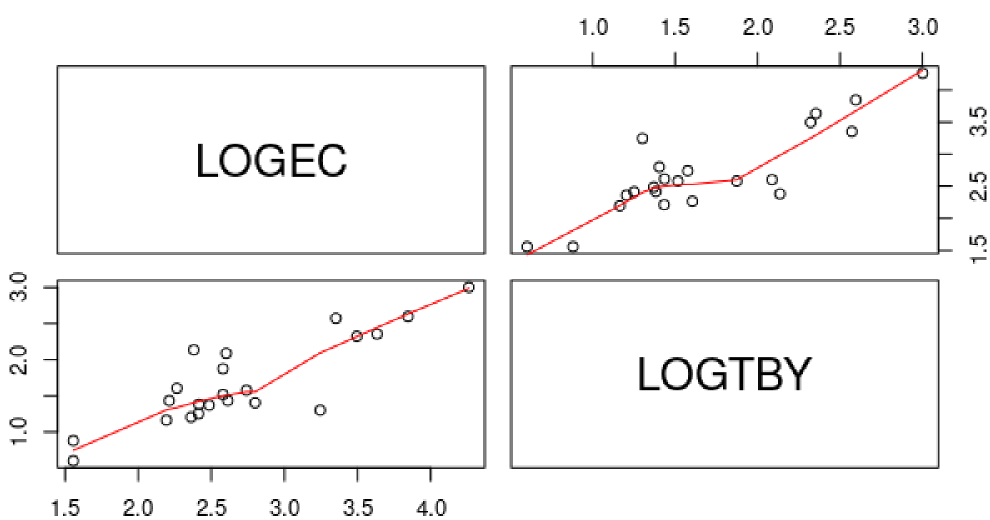
Variable Summary Statistics

	LOGEC	EC	LOGTBY	TBY
Minimum	1.56	36	0.602	4.0
1st Quartile	2.36	230	1.300	20.0
Median	2.58	380	1.480	30.1
Mean	2.71	1880	1.690	130.0
3rd Quartile	3.24	1760	2.140	137.0
Maximum	4.26	18300	3.000	1000.0

Box Plots



Exploratory Plots



Basic Model Statistics

Number of Observations	22
Standard error (RMSE)	0.361
Average Model standard percentage error (MSPE)	92.9
Coefficient of determination (R^2)	0.733
Adjusted Coefficient of Determination (Adj. R^2)	0.72
Bias Correction Factor (BCF)	1.41

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	1.080	0.232	4.67	1.47e-04
LOGTBY	0.964	0.130	7.42	3.68e-07

Correlation Matrix

	Intercept	E.vars
Intercept	1.000	-0.944
E.vars	-0.944	1.000

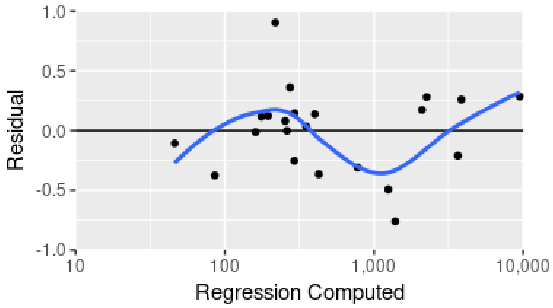
Outlier Test Criteria

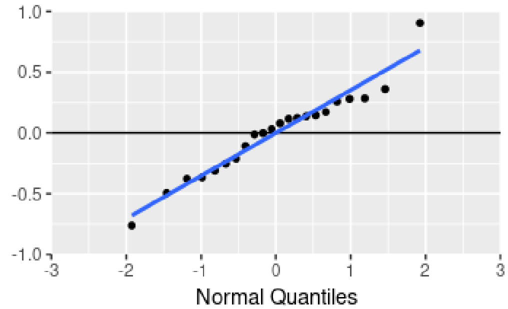
Leverage	Cook's D	DFFITS
0.273	0.193	0.603

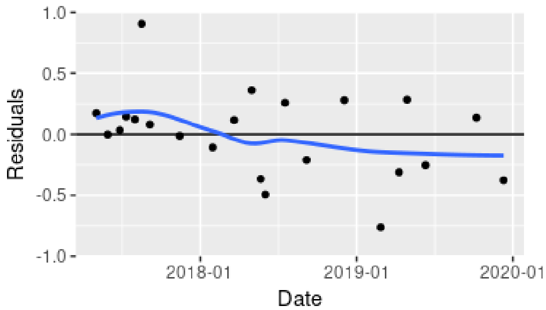
Flagged Observations

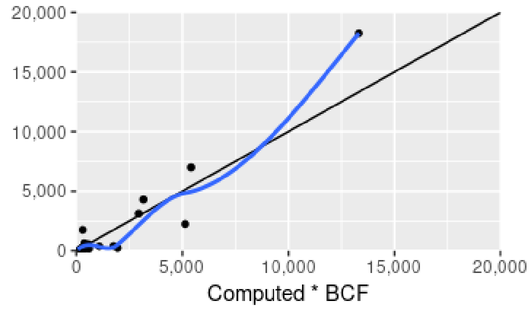
datetime	LOGEC	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
8/17/2017 10:05	3.24	2.34	0.906	2.6	3.11	0.0645	0.232	0.816
2/26/2019 11:40	2.38	3.14	-0.763	-2.19	-2.45	0.0718	0.186	-0.683

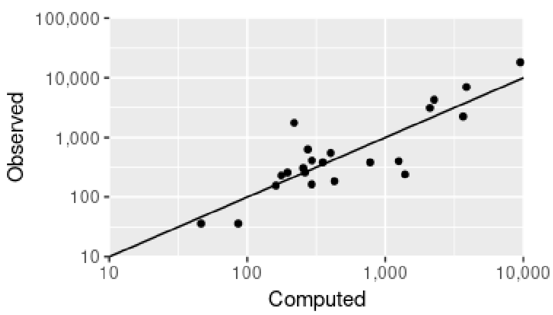
Statistical Plots

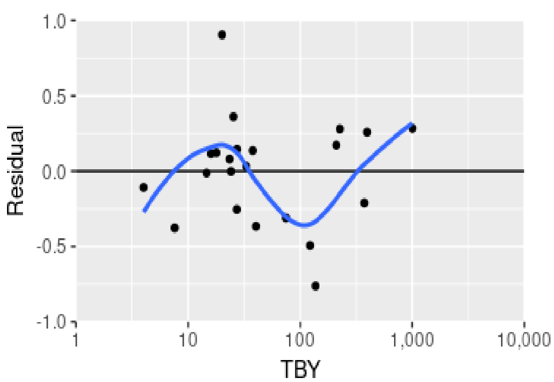


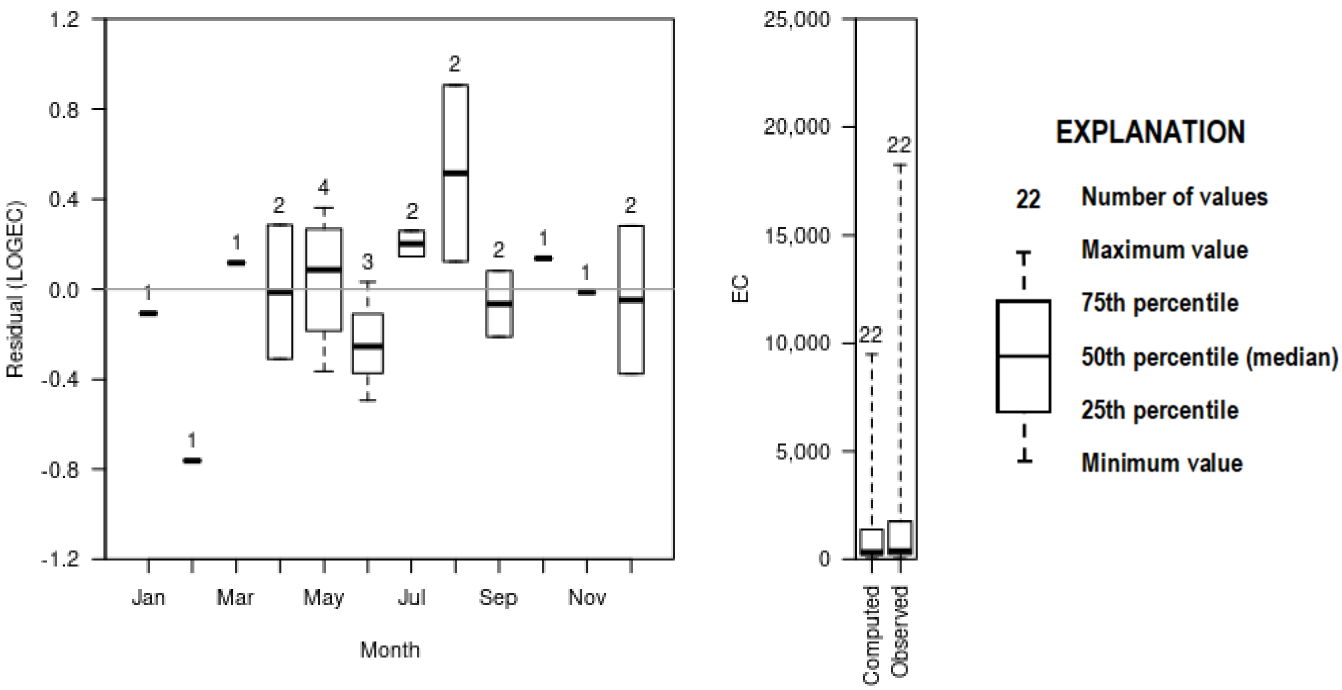




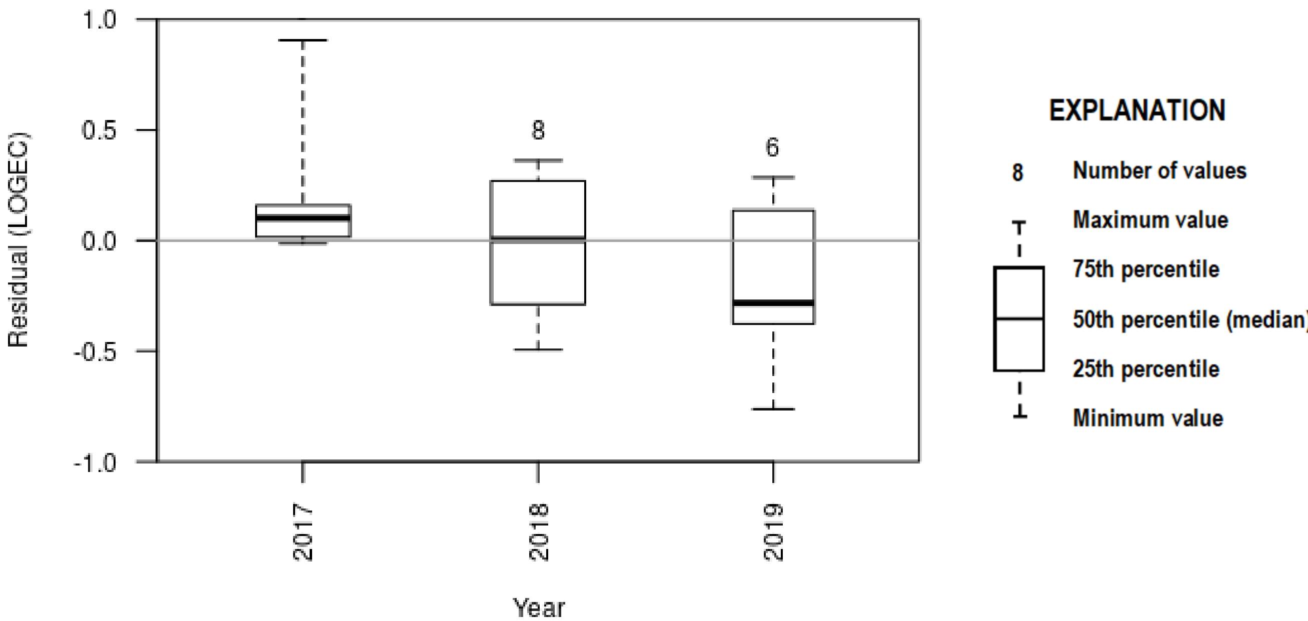






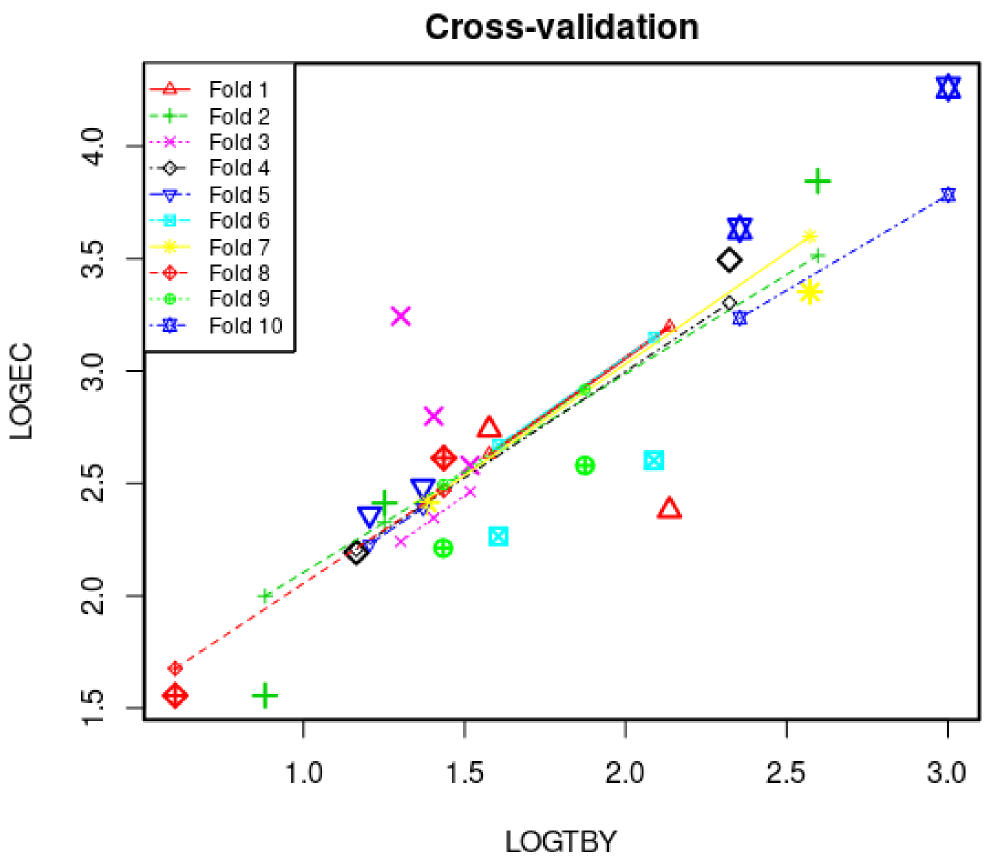


1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
Residual (LOGEC)
1
1
1
2
4
3
2
2
2
1
1
2
Jan
Mar
May
Jul
Sep
Nov
Month
25,000
20,000
15,000
10,000
5,000
0
EC
22
22
Computed
Observed
EXPLANATION
22
Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value

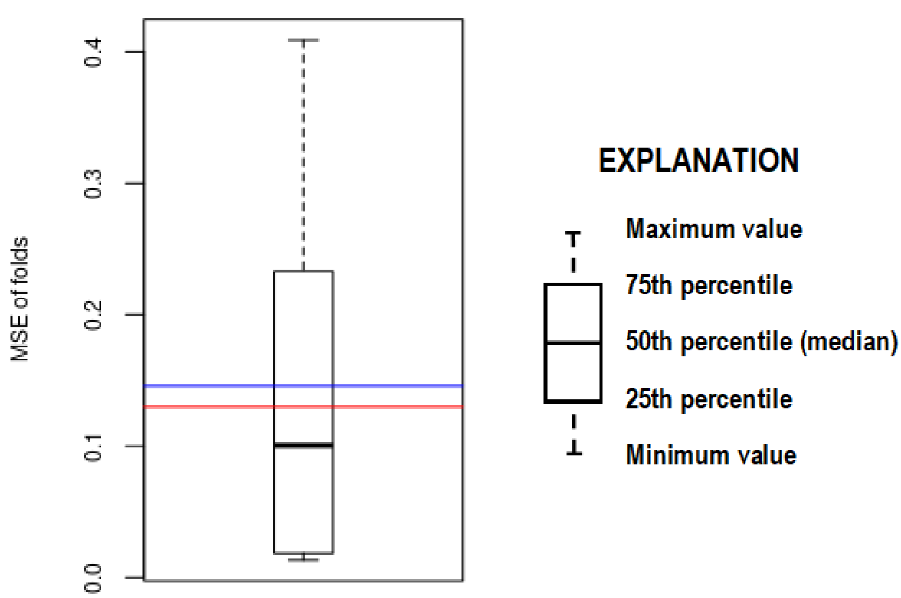


1.0
0.5
0.0
-0.5
-1.0
Residual (LOGEC)
8
6
2017
2018
2019
Year
EXPLANATION
8
Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value

Cross Validation




```
               Minimum MSE of folds:  0.0132
                  Mean MSE of folds:  0.1460
                Median MSE of folds:  0.1000
               Maximum MSE of folds:  0.4090
 (Mean MSE of folds) / (Model MSE):  1.1200
```




```
Red line - Model MSE

Blue line - Mean MSE of folds
```

Model-Calibration Dataset

	Date	LOGEC	LOGTBY	EC	TBY	Computed LOGEC	Computed EC	Residual	Normal Quantiles
1	5/3/2017	3.49	2.32	3130	210	3.32	2950	0.173	0.667
2	5/30/2017	2.41	1.38	260	24.2	2.42	367	-0.00214	-0.17
3	6/27/2017	2.58	1.52	380	32.9	2.55	494	0.0334	-0.0565
4	7/12/2017	2.61	1.44	410	27.3	2.47	412	0.145	0.532
5	8/1/2017	2.41	1.25	259	17.8	2.29	274	0.122	0.286
6	8/17/2017	3.24	1.3	1760	20	2.34	307	0.906	1.93
7	9/5/2017	2.49	1.37	306	23.5	2.41	357	0.0802	0.0565
8	11/14/2017	2.19	1.16	156	14.6	2.21	226	-0.0131	-0.286
9	1/30/2018	1.56	0.602	36	4	1.66	64.9	-0.108	-0.406
10	3/21/2018	2.36	1.21	230	16	2.25	247	0.116	0.17
11	5/1/2018	2.8	1.4	630	25.4	2.44	385	0.362	1.46
12	5/22/2018	2.26	1.6	184	40.3	2.63	601	-0.366	-0.986
13	6/2/2018	2.6	2.09	400	122	3.1	1750	-0.494	-1.46
14	7/18/2018	3.85	2.6	7000	395	3.59	5420	0.259	0.816
15	9/6/2018	3.35	2.57	2250	373	3.56	5140	-0.211	-0.532
16	12/3/2018	3.63	2.35	4300	226	3.35	3170	0.28	0.986
17	2/26/2019	2.38	2.14	240	137	3.14	1950	-0.763	-1.93
18	4/10/2019	2.58	1.87	380	74.8	2.89	1090	-0.31	-0.816
19	4/29/2019	4.26	3	18300	1000	3.98	13300	0.284	1.19
20	6/11/2019	2.21	1.43	163	27.2	2.47	411	-0.254	-0.667
21	10/8/2019	2.74	1.58	550	37.7	2.6	564	0.137	0.406
22	12/10/2019	1.56	0.881	36	7.6	1.93	120	-0.377	-1.19

Definitions

```
EC: Escherichia coli in cfu/100mL (90902)
TBY: Turbidity in FNU (63680)
```

References Cited

American Public Health Association, American Water Works Association, and Water Environment Federation, 1995, Standard methods for the examination of water and wastewater (19th ed.): Washington D.C., American Public Health Association, 905 p.

Christensen, V.G., Ziegler, A.C., Rasmussen P.P., and Jian X., 2003, Continuous real-time water-quality monitoring of Kansas streams, *in* Proceedings of 2003 Spring Specialty Conference on Agricultural Hydrology and Water Quality, Kansas City, Mo., May 12–14, 2003: Middleburg, Va., American Water Resources Association Technical Publication Series No. TPS–03–1, compact disc. [Also available at https://nrtwq.usgs.gov/ks/methods/christensen2003.]

Cook, D.R., 1977, Detection of influential observation in linear regression: Technometrics, v. 19, no. 1, p. 15–18. [Also available at https://www.jstor.org/stable/1268249?seq=4#metadata_info_tab_contents.]

Duan, N., 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, no. 383, p. 605–610. [Also available at https://doi.org/10.1080/01621459.1983.10478017.]

R Core Team, 2020, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, version 4.0.0. [Also available at https://www.r-project.org.]

Rasmussen, P.P., Eslick, P.J., and Ziegler, A.C., 2016, Relations between continuous real-time physical properties and discrete water-quality constituents in the Little Arkansas River, south-central Kansas, 1998–2014: U.S. Geological Survey Open-File Report 2016–1057, 16 p. [Also available at https://doi.org/10.3133/ofr20161057.]

Rasmussen, P.P., Gray, J.R., Glysson, G.D., and Ziegler, A.C, 2009, Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity sensor and streamflow data: U.S. Geological Survey Techniques and Methods, book 3, chap. C4, 53 p. [Also available at https://doi.org/10.3133/tm3C4.]

Rasmussen, T.J., Bennett, T.J., Stone, M.L., Foster, G.M., Graham, J.L., and Putnam, J.E., 2014, Quality-assurance and data-management plan for water-quality activities in the Kansas Water Science Center, 2014: U.S. Geological Survey Open-File Report 2014–1233, 41 p. [Also available at https://doi.org/10.3133/ofr20141233.]

Sauer, V.B., and Turnipseed, D.P., 2010, Stage measurement at gaging stations: U.S. Geological Survey Techniques and Methods, book 3, chap. A7, 45 p. [Also available at https://doi.org/10.3133/tm3A7.]

Turnipseed, D.P., and Sauer, V.B., 2010, Discharge measurements at gaging stations: U.S. Geological Survey Techniques and Methods, book 3, chap. A8, 87 p. [Also available at https://doi.org/10.3133/tm3A8.]

U.S. Geological Survey, 2021, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed December 8, 2021, 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–A9 [variously paged]. [Also available at https://water.usgs.gov/owq/FieldManual/.]

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