Appendix 28. Model Archival Summary for *Escherichia coli* Bacteria Density at U.S. Geological Survey Site 06892350, Kansas River at De Soto, Kansas, during September 2013 through September 2019

This model archival summary summarizes the *Escherichia coli* bacteria (ECB; U.S. Geological Survey [USGS] parameter code 90902) density model developed to compute 15-minute ECB densities from September 2013 onward. This model supersedes all previous models.

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: 06892350
Site name: Kansas River at De Soto, Kansas
Location: Lat 38°59'00", long 94°57'52" referenced to North American Datum of 1927, in NE 1/4 SE 1/4 SE 1/4 sec.28, T.12 S., R.22 E., Leavenworth County, Kans., hydrologic unit 10270104.

Equipment: A YSI 6600 water-quality monitor equipped with sensors for water temperature, specific conductance, dissolved oxygen, pH, and turbidity (TBY) was installed from August 2012 through June 2014. A Xylem YSI EXO2 water-quality monitor equipped with sensors for water temperature, specific conductance, dissolved oxygen, pH, TBY, and chlorophyll and phycocyanin fluorescence was installed during June 2014 through September 2019. A Hach Nitratax plus sc sensor (5-millimeter path length) that monitors ultraviolet (UV) nitrate concentrations was installed from June 2013 through September 2019. The monitors were housed in side-by-side 4-inch-diameter galvanized steel pipes. Readings from the water-quality and nitrate plus nitrite monitors were recorded every 15 minutes and transmitted by way of satellite, hourly.

Date model was created: April 13, 2020

Model calibration data period: September 23, 2013, through September 24, 2019

Model application date: September 23, 2013, onward

Model-Calibration Dataset

All data were collected using USGS protocols (Wagner and others, 2006; U.S. Geological Survey, variously dated) and are stored in the National Water Information System (U.S. Geological Survey, 2020) database and available to the public. Ordinary least squares analysis was used to develop regression models using R programming language (R Core Team, 2020). Potential explanatory variables that were evaluated individually and in combination included streamflow, water temperature, specific conductance, dissolved oxygen, pH, TBY, chlorophyll and phycocyanin fluorescence, and UV nitrate sensor data. The maximum time span between two continuous data points used for interpolation was 2 hours (in order 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 1 hour in the continuous data record resulted in missing interpolated data). Seasonal components (sine and cosine variables) were also evaluated as potential explanatory variables.

The final selected regression model was based on 77 concurrent measurements of ECB density and sensor-measured TBY during September 23, 2013, through September 24, 2019. Samples were collected throughout the range of continuously observed hydrologic conditions. One sample had a density below the laboratory detection limit; therefore, a Tobit regression model was developed to compute estimates of linear regression model parameters using the absolute maximum likelihood estimation approach (Cohen, 1950; Hald, 1949; Helsel and others, 2020; Tobin, 1958). Thirty-seven sample densities were qualified as “estimated.” Summary statistics and the complete model-calibration dataset are provided below. Potential outliers were identified using the methods described in Rasmussen and others (2009). Additionally, outlier test criteria, including leverage and Cook’s distance (Cook’s D; Cook, 1977), were used to estimate potential outlier influence on the final Tobit regression model. All potential outliers were not found to have errors associated with collection, processing, or analysis and were therefore considered valid.

This model is specific to the Kansas River at De Soto, Kans., during this study period and cannot be applied to data collected from other sites on the Kansas River or data collected from other waterbodies.

Escherichia coli Bacteria Sampling Details

Indicator bacteria samples typically were collected either from the downstream side of the bridge or instream within 100 feet of the bridge. The grab sample collection method with weighted basket was used for all indicator bacteria samples (contrary to the equal-width-increment collection method used for all other analytes; U.S. Geological Survey, variously dated). During July 2012 through

June 2017, grab samples were collected every 2 weeks during March through October, once a month during November through February, and during selected reservoir release and runoff events. During July 2017 through September 2019, grab samples were collected on a monthly to bimonthly basis, depending on flow conditions. An open-mouth bottle with weighted-basket sampler was used. Additional detail on sample collection is available in Foster and Graham (2016) and Graham and others (2018). Samples were analyzed for ECB density at the USGS Kansas Water Science Center in Lawrence, Kans.

Model Development

Discretely collected ECB was related to sensor-measured TBY and other continuous sensor-measured data using stepwise regression analysis in R programming language (R Core Team, 2020). The distribution of residuals was examined for normality, and the plots of residuals (the difference between the measured and computed values) were examined for homoscedasticity (departures from zero did not change substantially over the range of computed values). Previously published explanatory variables were also strongly considered for continuity.

1.3 percent of the model-calibration dataset were censored results (less than the minimum reporting level). Tobit regression models were developed using absolute maximum likelihood estimation methods to relate discretely collected ECB density to sensor-measured TBY. Tobit model parameter estimates were calculated using the *smwrQW* (*v0.7.9*) package in R programming language (R Core Team, 2020).

TBY was selected as a good surrogate for ECB based on residual plots, pseudocoefficient of determination (pseudo-R^2), and estimated residual standard error. Values for all the aforementioned statistics were computed and are included below along with all relevant sample data and additional statistical information.

Model Summary

The following is a summary of final Tobit regression analysis for ECB density at USGS site 06892350:

ECB density-based model:

$$\log ECB = 1.68 \times \log TBY - 1.02$$

where

log = logarithm base 10;

ECB = *Escherichia coli* bacteria density, in colonies per 100 milliliters; and

TBY = turbidity, in formazin nephelometric units.

TBY makes physical and statistical sense as an explanatory variable for ECB because of its positive correlation with suspended material to which fecal indicator bacteria can physically bind.

The logarithmically (log) transformed model may be retransformed to the original units so that ECB 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 1.60. The retransformed model, accounting for BCF is as follows:

$$ECB = 1.6 \times (TBY^{1.68} \times 10^{-1.02})$$

Previous Models

Start Year	End Year	Model Equation	Reference
2012	2019	logECB = 1.54logTBY - 0.803	Foster and Graham (2016)
1999	2003	logECB = 1.55logTBY - 1.16	Rasmussen and others (2005)

Model Statistics, Data, and Plots

Model

logECB = + 1.68 * logTBY - 1.02

Computation method: Absolute Maximum Likelihood Estimation (AMLE)

Variable Summary Statistics

	ECB	TBY
Minimum	<1.0	6.1
1st Quartile	17.0	26.0
Median	77.0	58.0
Mean	1588.0	161.5
3rd Quartile	340.0	150.0
Maximum	22000.0	1530.0

Basic Model Statistics

Estimated residual standard error (unbiased)	0.4392	
Number of observations	77	
Number censored	1	(1.3 percent)
Log-likelihood (model)	-45.48	
Log-likelihood (intercept only)	-142.7	
Chi-square	194.4	
Degrees of freedom	1	
p-value	<0.0001	
Pseudo-R-squared	0.8216	
Akaike Information Criterion	96.96	
Bayesian Information Criterion	104	
Bias Correction Factor	1.600882	

Explanatory Variables

Coefficients:

	Estimate	Std. Error	z-score	p-value
(Intercept)	-1.022	0.17238	-5.927	0
logTBY	1.676	0.09045	18.532	0

Outlier Test Criteria

Leverage	Cook's D
0.03896	0.69951

Flagged Observations

Observations exceeding at least one test criterion

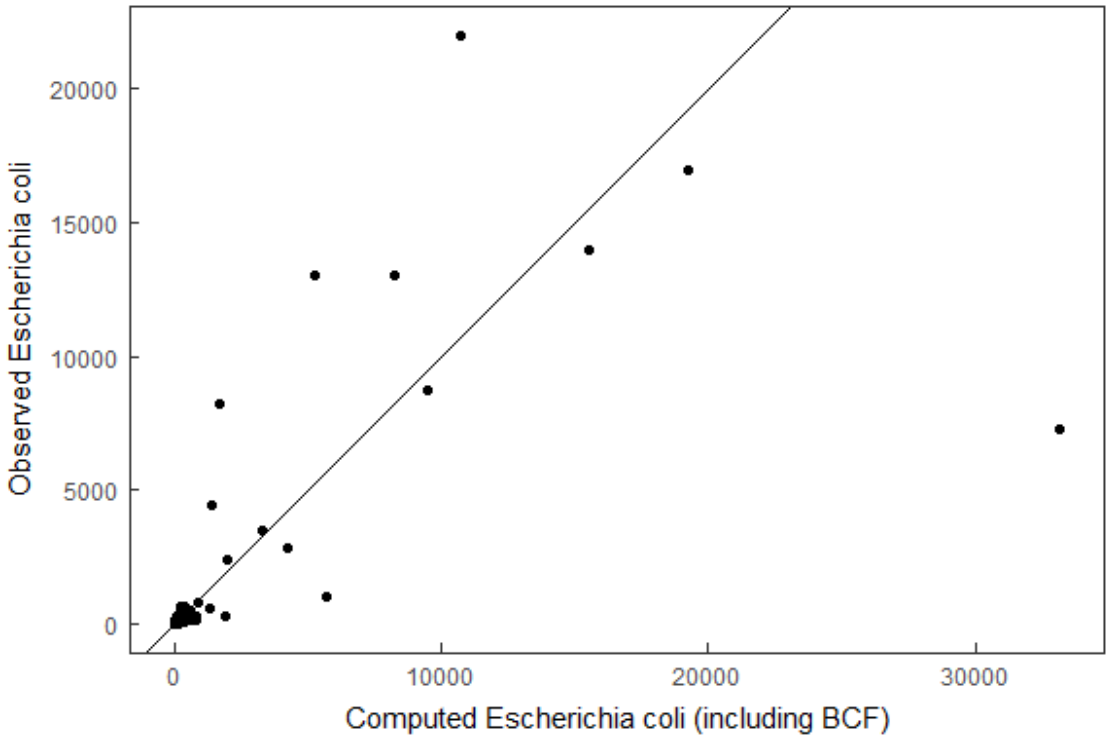
	logECB	ycen	yhat	resids	leverage	cooksD
5	1.2041	FALSE	0.4452	0.75896	0.05075	8.411e-02
6	0.9542	FALSE	0.2947	0.65950	0.05825	7.406e-02
11	3.8633	FALSE	4.3168	-0.45343	0.09132	5.895e-02
22	0.0000	TRUE	0.5251	-0.73952	0.04705	7.345e-02
24	0.3010	FALSE	0.4288	-0.12777	0.05154	2.425e-03
27	3.9395	FALSE	3.7721	0.16738	0.05843	4.787e-03
28	4.1461	FALSE	3.9866	0.15948	0.07032	5.365e-03
45	4.1139	FALSE	3.5170	0.59698	0.04608	4.679e-02
54	4.3424	FALSE	3.8263	0.51615	0.06130	4.805e-02
64	3.4472	FALSE	3.4204	0.02673	0.04192	8.457e-05

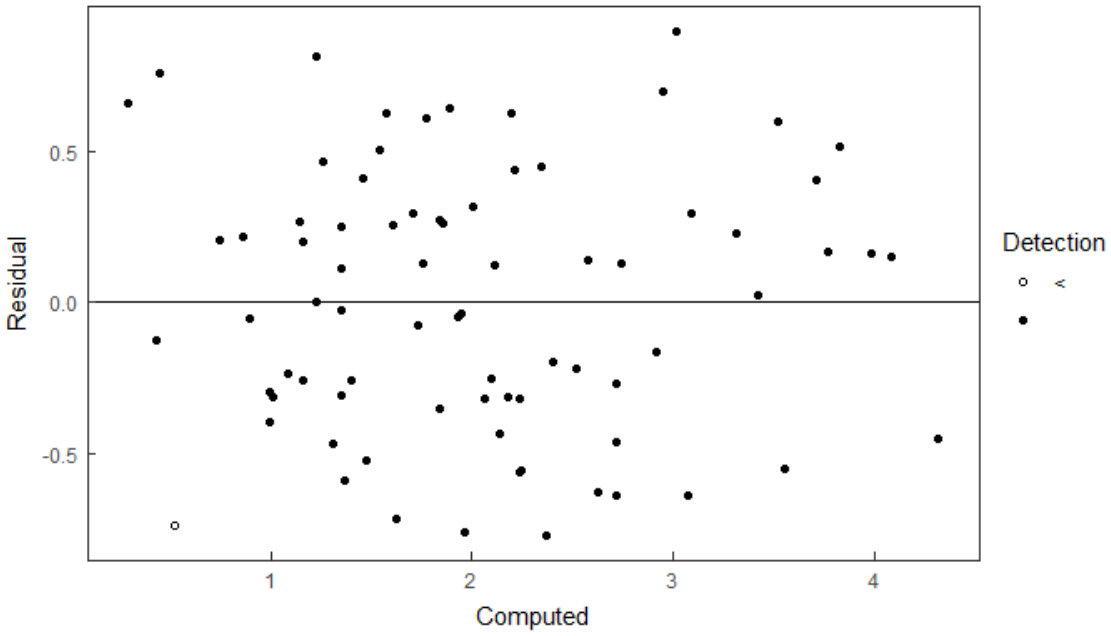
```
67 4.1139 FALSE 3.7110  0.40291  0.05530 2.607e-02
71 3.0000 FALSE 3.5490 -0.54898  0.04752 4.093e-02
73 4.2304 FALSE 4.0803  0.15020  0.07595 5.201e-03
```

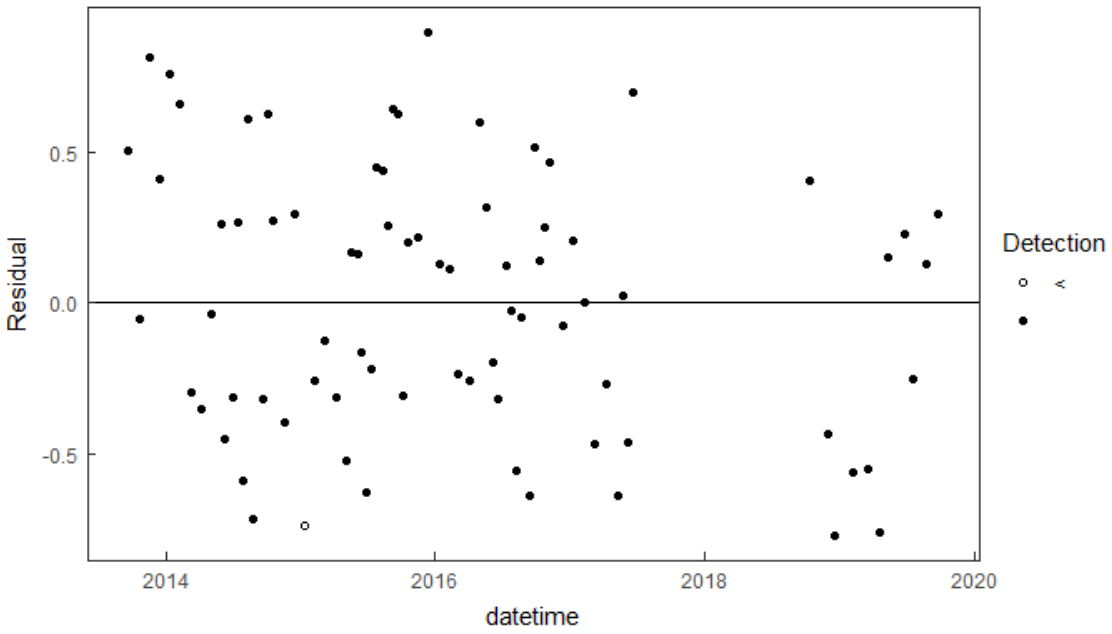
95 Percent Confidence Interval

```
                2.5 %     97.5 %
(Intercept) -1.359554 -0.6838255
logTBY       1.499000  1.8535640
```

Plots







Model-Calibration Dataset

	datetime	logECB	logTBY	ECB	TBY	Computed_logECB	Computed ECB
1	2013-09-23 09:40:00	2.04	1.528	110	33.70	1.538	55.29
2	2013-10-21 10:30:00	0.845	1.146	7	14.00	0.899	12.68
3	2013-11-18 14:00:00	2.04	1.342	110	22.00	1.228	27.06
4	2013-12-16 08:00:00	1.86	1.477	73	30.00	1.454	45.50
5	2014-01-13 08:00:00	1.2	0.875	16	7.50	0.445	4.46
6	2014-02-10 07:40:00	0.954	0.785	9	6.10	0.294	3.15
7	2014-03-10 13:40:00	0.699	1.204	5	16.00	0.996	15.87
8	2014-04-07 11:40:00	1.49	1.708	31	51.00	1.840	110.73
9	2014-05-05 14:00:00	1.91	1.771	81	59.00	1.946	141.35
10	2014-06-02 13:00:00	2.11	1.716	130	52.00	1.854	114.39
11	2014-06-11 16:00:00	3.86	3.185	7300	1530.00	4.316	33105.61
12	2014-06-30 12:20:00	1.86	1.908	73	81.00	2.177	240.43
13	2014-07-14 12:50:00	1.41	1.294	26	19.67	1.146	22.42
14	2014-07-28 13:50:00	0.778	1.426	6	26.67	1.368	37.35
15	2014-08-11 15:10:00	2.38	1.666	240	46.33	1.770	94.27
16	2014-08-25 14:20:00	0.903	1.576	8	37.67	1.619	66.63
17	2014-09-22 10:00:00	1.74	1.839	55	69.00	2.060	183.77
18	2014-10-06 13:30:00	2.82	1.918	660	82.80	2.193	249.45
19	2014-10-20 14:00:00	2.11	1.708	130	51.00	1.840	110.73
20	2014-11-17 12:00:00	0.602	1.204	4	16.00	0.996	15.87
21	2014-12-15 10:00:00	2	1.625	100	42.20	1.702	80.61
22	2015-01-12 11:30:00	<0	0.923	<1	8.37	0.524	5.36
23	2015-02-09 13:50:00	1.15	1.447	14	28.00	1.403	40.53
24	2015-03-09 12:20:00	0.301	0.865	2	7.33	0.428	4.29
25	2015-04-06 14:50:00	0.699	1.213	5	16.33	1.011	16.42
26	2015-05-04 13:30:00	0.954	1.490	9	30.88	1.475	47.76
27	2015-05-18 15:30:00	3.94	2.860	8700	724.10	3.771	9448.90
28	2015-06-06 19:50:00	4.15	2.988	14000	972.22	3.985	15482.95
29	2015-06-15 14:50:00	2.75	2.348	560	222.96	2.914	1312.16
30	2015-06-29 12:30:00	2	2.176	100	150.00	2.625	675.29
31	2015-07-13 14:30:00	2.3	2.114	200	130.00	2.521	531.29
32	2015-07-27 14:20:00	2.79	2.009	620	102.00	2.344	353.81
33	2015-08-10 13:50:00	2.65	1.931	450	85.33	2.215	262.37

34	2015-08-24 09:50:00	1.86	1.568	73	37.00	1.606	64.66
35	2015-09-08 09:50:00	2.53	1.735	340	54.33	1.886	123.12
36	2015-09-21 09:30:00	2.2	1.550	160	35.50	1.576	60.33
37	2015-10-05 09:20:00	1.04	1.415	11	26.00	1.349	35.80
38	2015-10-19 09:30:00	1.36	1.301	23	20.00	1.159	23.06
39	2015-11-16 09:10:00	1.08	1.125	12	13.33	0.863	11.69
40	2015-12-14 15:10:00	3.91	2.409	8200	256.67	3.016	1661.35
41	2016-01-11 10:10:00	1.89	1.656	77	45.33	1.754	90.89
42	2016-02-08 08:30:00	1.46	1.415	29	26.00	1.349	35.80
43	2016-03-03 09:00:00	0.845	1.255	7	18.00	1.082	19.33
44	2016-04-04 09:20:00	0.903	1.301	8	20.00	1.159	23.06
45	2016-05-02 09:30:00	4.11	2.708	13000	510.00	3.516	5251.05
46	2016-05-16 09:10:00	2.32	1.806	210	64.00	2.005	162.00
47	2016-06-06 11:30:00	2.2	2.041	160	110.00	2.399	401.54
48	2016-06-20 08:20:00	1.92	1.944	83	88.00	2.237	276.26
49	2016-07-11 08:40:00	2.23	1.867	170	73.67	2.108	205.07
50	2016-07-25 08:30:00	1.32	1.415	21	26.00	1.349	35.80
51	2016-08-08 10:10:00	1.69	1.949	49	89.00	2.245	281.54
52	2016-08-22 09:00:00	1.89	1.763	77	58.00	1.934	137.36
53	2016-09-12 08:50:00	2.43	2.442	270	276.67	3.071	1883.99
54	2016-09-26 10:30:00	4.34	2.892	22000	780.00	3.825	10703.09
55	2016-10-11 09:20:00	2.72	2.146	520	140.00	2.575	601.55
56	2016-10-24 09:30:00	1.6	1.415	40	26.00	1.349	35.80
57	2016-11-07 08:40:00	1.72	1.362	53	23.00	1.260	29.15
58	2016-12-12 09:50:00	1.65	1.640	45	43.67	1.727	85.36
59	2017-01-09 09:50:00	0.954	1.054	9	11.33	0.745	8.90
60	2017-02-06 10:00:00	1.23	1.342	17	22.00	1.228	27.06
61	2017-03-06 10:40:00	0.845	1.392	7	24.67	1.311	32.77
62	2017-04-10 09:20:00	2.45	2.230	280	170.00	2.716	832.90
63	2017-05-08 09:20:00	2.08	2.230	120	170.00	2.716	832.90
64	2017-05-22 09:20:00	3.45	2.650	2800	446.67	3.419	4204.66
65	2017-06-05 08:50:00	2.26	2.230	180	170.00	2.716	832.90
66	2017-06-19 09:10:00	3.64	2.368	4400	233.33	2.947	1416.08
67	2018-10-11 13:50:00	4.11	2.823	13000	665.80	3.710	8208.82
68	2018-11-29 12:20:00	1.7	1.883	50	76.37	2.134	217.82
69	2018-12-18 10:40:00	1.6	2.024	40	105.77	2.371	376.01
70	2019-02-06 10:20:00	1.68	1.946	48	88.33	2.240	278.01
71	2019-03-19 09:20:00	3	2.727	1000	532.93	3.548	5652.79
72	2019-04-16 09:30:00	1.2	1.782	16	60.60	1.965	147.84
73	2019-05-09 11:10:00	4.23	3.044	17000	1105.62	4.079	19206.24
74	2019-06-26 12:30:00	3.54	2.586	3500	385.50	3.312	3285.00
75	2019-07-16 10:10:00	1.85	1.861	70	72.60	2.097	200.12
76	2019-08-20 09:20:00	2.88	2.247	750	176.60	2.744	887.80
77	2019-09-24 09:50:00	3.38	2.451	2400	282.23	3.085	1947.95

Definitions

Cook's D: Cook's distance (Helsel and others, 2020).

ECB: *Escherichia coli*, in colonies per 100 milliliters (90902).

Leverage: An outlier's measure in the x direction (Helsel and others, 2020).

p-value: The probability that the independent variable has no effect on the dependent variable (Helsel and others, 2020).

Pseudo-R-squared: Pseudocoefficient of determination. An estimation of the proportion of variance in the response variable explained by the model (McKelvey and Zavoina, 1975).

TBY: Turbidity, in formazin nephelometric units (63680).

z-score: The estimated coefficient divided by its associated standard error (Helsel and others, 2020).

References Cited

Cohen, A.C., Jr., 1950, Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples: Annals of Mathematical Statistics, v. 21, no. 4, p. 557–569, accessed October 2019 at https://doi.org/10.1214/aoms/1177729751.

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

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.]

Foster, G.M., and Graham, J.L., 2016, Logistic and linear regression model documentation for statistical relations between continuous real-time and discrete water-quality constituents in the Kansas River, Kansas, July 2012 through June 2015: U.S. Geological Survey Open-File Report 2016–1040, 27 p., accessed July 2020 at https://doi.org/10.3133/ofr20161040.

Graham, J.L., Foster, G.M., Williams, T.J., Mahoney, M.D., May, M.R., and Loftin, K.A., 2018, Water-quality conditions with an emphasis on cyanobacteria and associated toxins and taste-and-odor compounds in the Kansas River, Kansas, July 2012 through September 2016: U.S. Geological Survey Scientific Investigations Report 2018–5089, 55 p. [Also available at https://doi.org/10.3133/sir20185089.]

Hald, A., 1949, Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point: Scandinavian Actuarial Journal, v. 1949, no. 1, p. 119–134. [Also available at https://doi.org/10.1080/03461238.1949.10419767.]

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, ver. 1.1.]

McKelvey, R.D., and Zavoina, W., 1975, A statistical model for the analysis of ordinal level dependent variables: The Journal of Mathematical Sociology, v. 4, no. 1, p. 103–120. [Also available at https://doi.org/10.1080/0022250X.1975.9989847.]

R Core Team, 2020, R—A language and environment for statistical computing, version 4.0.3: Vienna, Austria, R Foundation for Statistical Computing, accessed December 2020 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, 53 p. [Also available at https://doi.org/10.3133/tm3C4.]

Rasmussen, T.J., Ziegler, A.C., and Rasmussen, P.P., 2005, Estimation of constituent concentrations, densities, loads, and yields in lower Kansas River, northeast Kansas, using regression models and continuous water-quality monitoring, January 2000 through December 2003: U.S. Geological Survey Scientific Investigations Report 2005–5165, 117 p. [Also available at https://doi.org/10.3133/sir20055165.]

Tobin, J., 1958, Estimation of relationships for limited dependent variables: Econometrica, v. 26, no. 1, p. 24–36. [Also available at https://doi.org/10.2307/1907382.]

U.S. Geological Survey, 2020, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed April 2020 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], accessed July 2020 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, 51 p. plus 8 attachments. [Also available at https://doi.org/10.3133/tm1D3.] [Supersedes USGS Water-Resources Investigations Report 2000–4252.]