Appendix 31. Model Archival Summary for Enterococci Bacteria Density at U.S. Geological Survey Site 06887500, Kansas River at Wamego, Kansas, during September 2013 through September 2019

This model archival summary summarizes the enterococci bacteria (ENT; U.S. Geological Survey [USGS] parameter code 90909) density model developed to compute 15-minute ENT 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: 06887500
Site name: Kansas River at Wamego, Kansas
Location: Lat 39°11'54", long 96°18'19" referenced to North American Datum of 1927, in SW 1/4 NW 1/4 SE 1/4 sec.9, T.10 S., R.10 E., Pottawatomie County, Kans., hydrologic unit 10270102.

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. The monitor was housed in a 4-inch-diameter galvanized steel pipe. Readings from the water-quality monitor were recorded every 15 minutes and transmitted by way of satellite, hourly.

Date model was created: May 5, 2020

Model calibration data period: September 23, 2013, through September 23, 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, and chlorophyll and phycocyanin fluorescence. These potential explanatory variables were interpolated within the 15-minute continuous record based on sample time. 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 ENT density and sensor-measured TBY during September 23, 2013, through September 23, 2019. Samples were collected throughout the range of continuously observed hydrologic conditions. No samples had densities below laboratory detection limits. Thirteen sample densities were qualified as "estimated." One sample, from December 14, 2015, exceeded the laboratory detection limit (greater than 20,000 colony forming units) and was removed from the model calibration dataset due to a high level of uncertainty and high leverage it would have on the final model. 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, studentized residuals from the final model were inspected for values greater than three or less than negative three. Values outside of that range were considered potential outliers and were investigated. 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 Wamego, 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.

Enterococci 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 ENT density at the USGS Kansas Water Science Center in Lawrence, Kans.

Model Development

Ordinary least squares regression analysis was done using R programming language (R Core Team, 2020) to relate discretely collected ENT density to sensor-measured TBY. 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.

TBY was selected as a good surrogate for ENT based on residual plots, coefficient of determination (R^2), and model standard percentage 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 regression analysis for ENT density at USGS site 06887500:

ENT density-based model:

$$\log ENT = 1.03 \times \log TBY + 0.578$$

where

log = logarithm base 10;

ENT = enterococci 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 ENT 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 ENT 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 1.73. The retransformed model, accounting for BCF is as follows:

$$ENT = 1.73 \times (TBY^{1.03} \times 10^{0.578})$$

Previous Models

Start Year	End Year	Model Equation	Reference
1999	2003	logENT = 1.35logTBY - 0.048	Rasmussen and others (2005)

Model Statistics, Data, and Plots

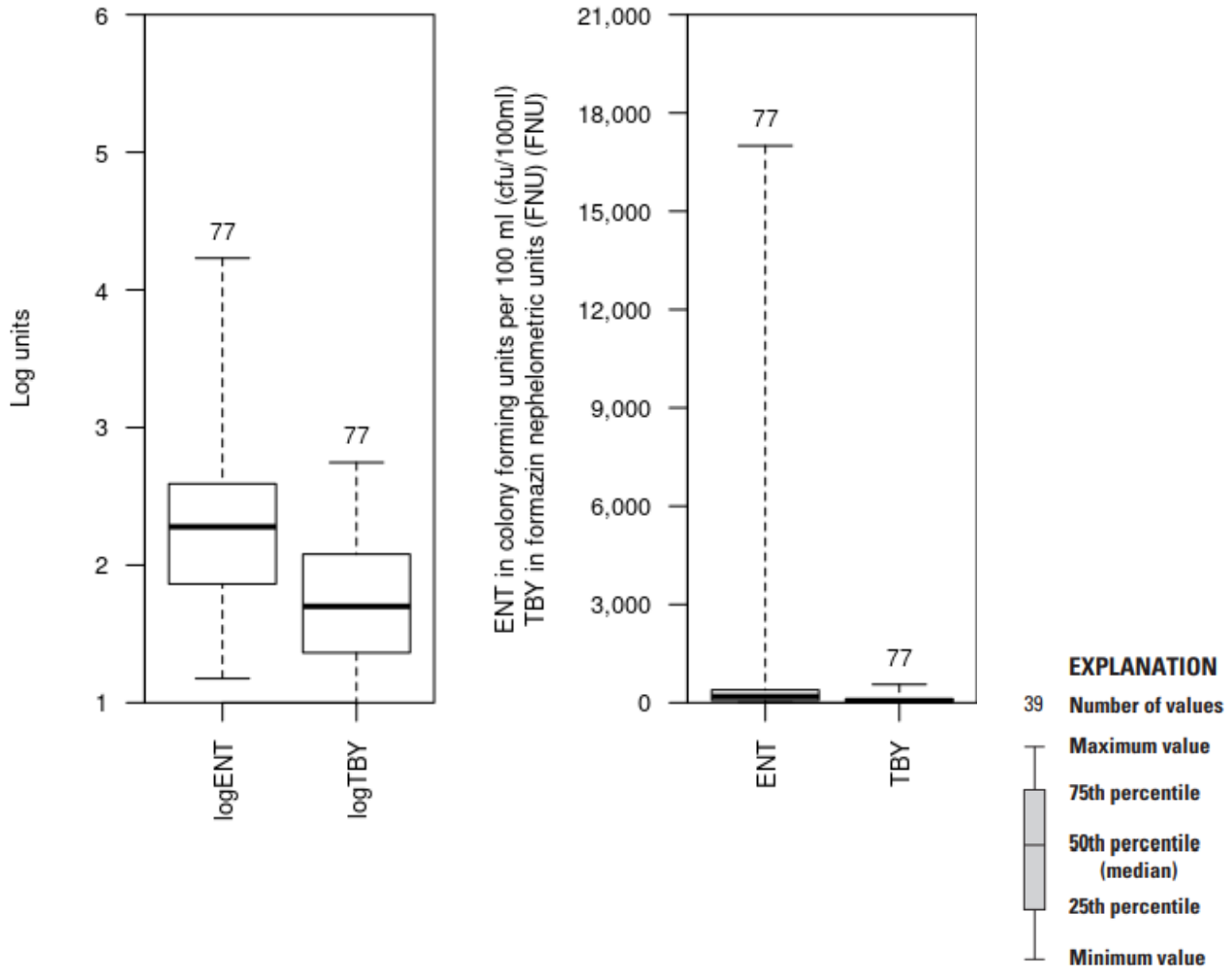
Model

logENT = + 1.03 * logTBY + 0.578

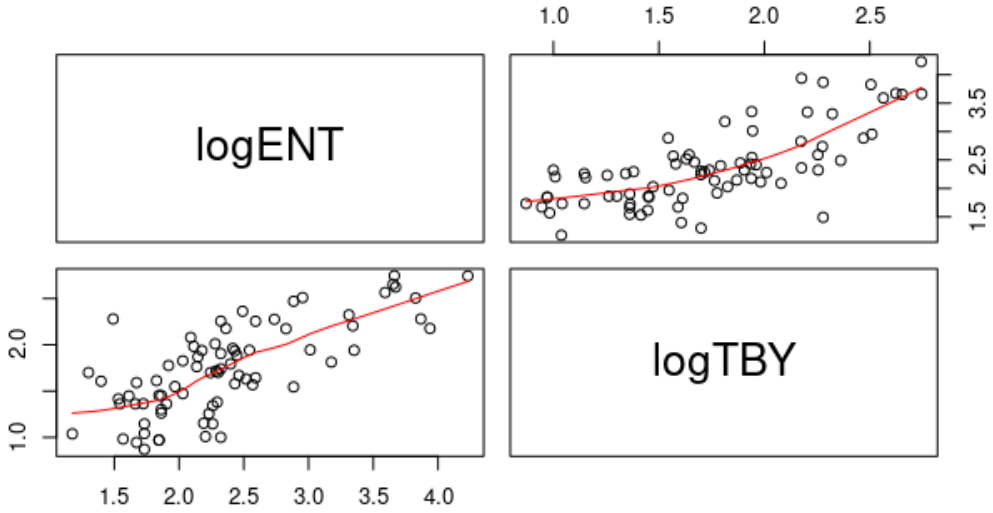
Variable Summary Statistics

	logENT	ENT	logTBY	TBY
Minimum	1.18	15	0.87	7.42
1st Quartile	1.86	73	1.36	23.10
Median	2.28	190	1.70	50.20
Mean	2.36	1030	1.74	101.00
3rd Quartile	2.59	390	2.08	120.00
Maximum	4.23	17000	2.75	557.00

Box Plots



Exploratory Plots



Red line shows the locally weighted scatterplot smoothing (LOWESS).

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

Basic Model Statistics

Number of Observations	77
Standard error (RMSE)	0.467
Average Model standard percentage error (MSPE)	129
Coefficient of determination (R^2)	0.536
Adjusted Coefficient of Determination (Adj. R^2)	0.53
Bias Correction Factor (BCF)	1.73

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	0.578	0.199	2.90	4.83e-03
logTBY	1.030	0.110	9.31	3.89e-14

Correlation Matrix

	Intercept	E.vars
Intercept	1.000	-0.964
E.vars	-0.964	1.000

Outlier Test Criteria

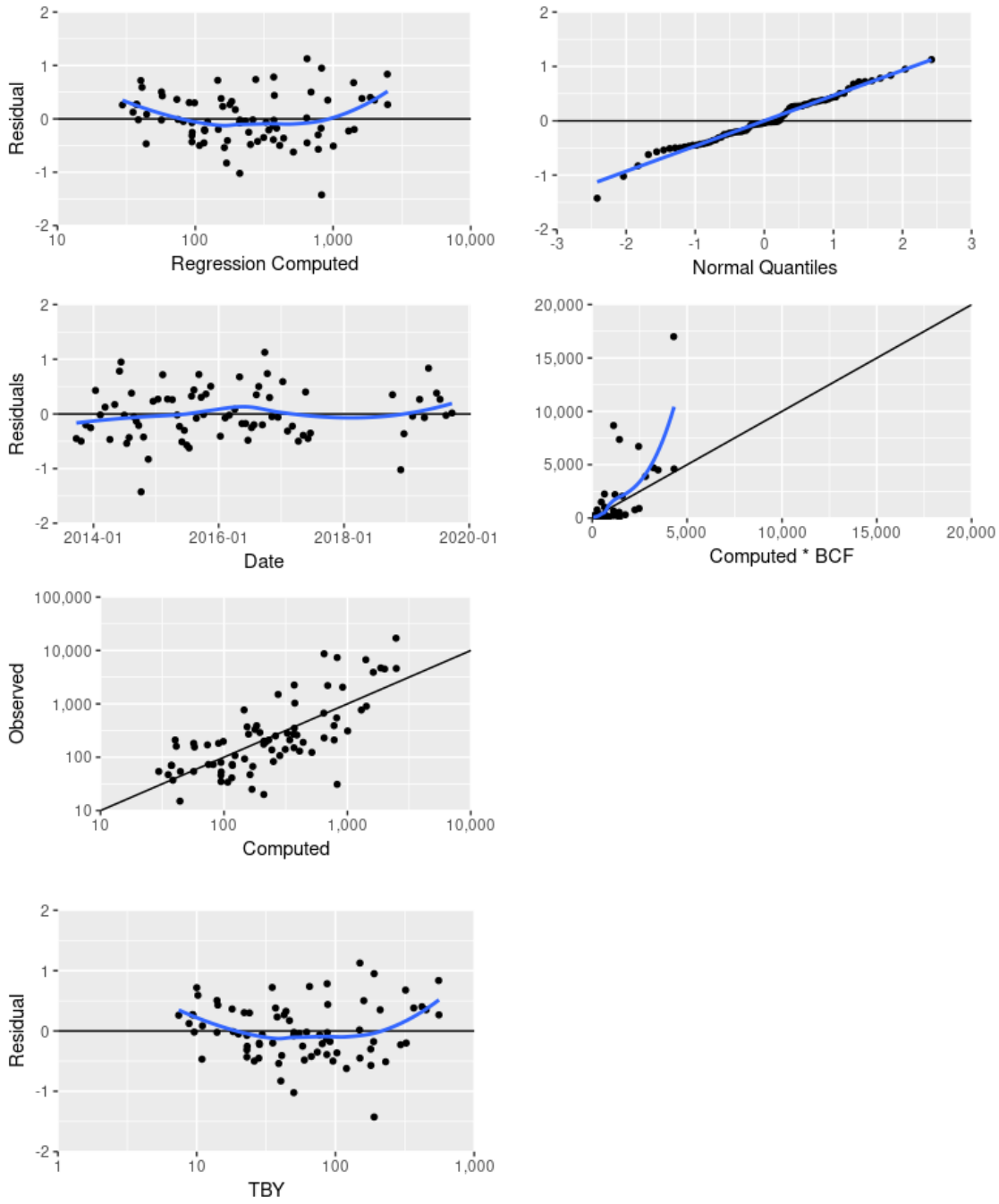
Leverage	Cook's D	DFFITS
0.0779	0.1943	0.3223

Flagged Observations

	logENT	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
201406110900	3.87	2.92	0.950	2.06	2.11	0.0292	0.0641	0.366
201410060900	1.49	2.92	-1.430	-3.10	-3.30	0.0292	0.1440	-0.571
201502091020	2.32	1.60	0.718	1.57	1.59	0.0435	0.0562	0.339
201605020920	3.83	3.15	0.677	1.49	1.50	0.0456	0.0528	0.328

201609260900	3.94	2.81	1.130	2.44	2.53	0.0236	0.0720	0.393
201905091520	4.23	3.39	0.836	1.86	1.89	0.0692	0.1280	0.515

Statistical Plots



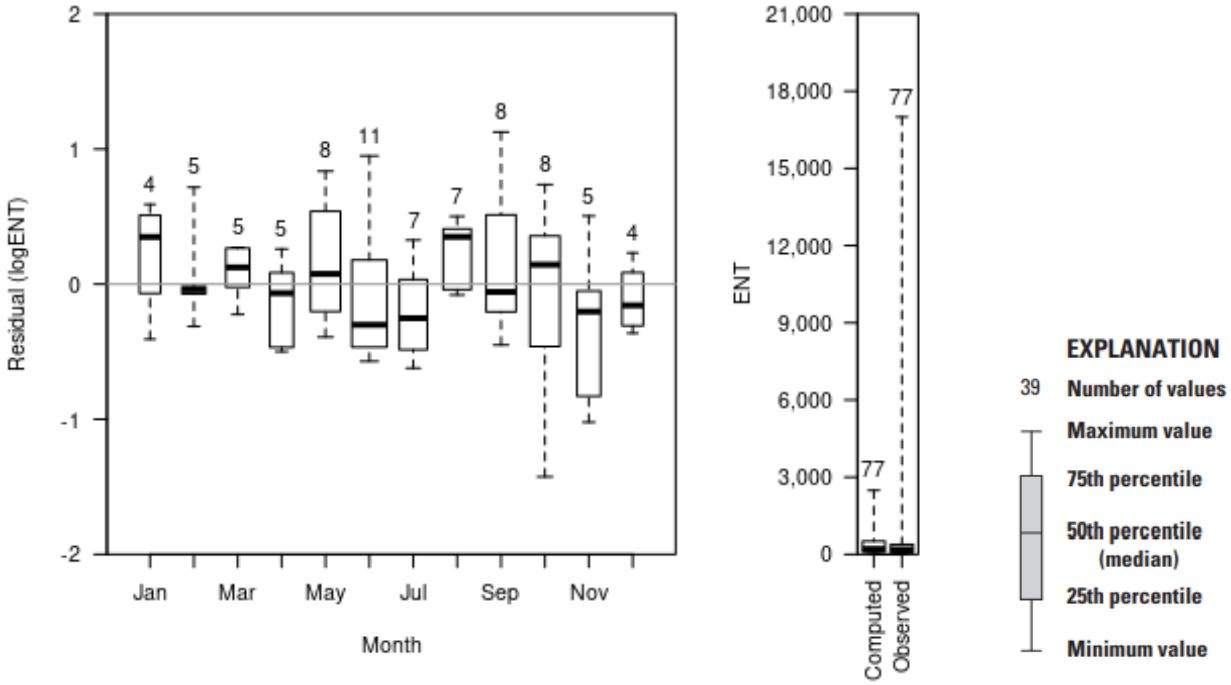
First row (left): Residual ENT related to regression computed ENT with local polynomial regression fitting, or locally estimated scatterplot smoothing (LOESS), indicated by the blue line.

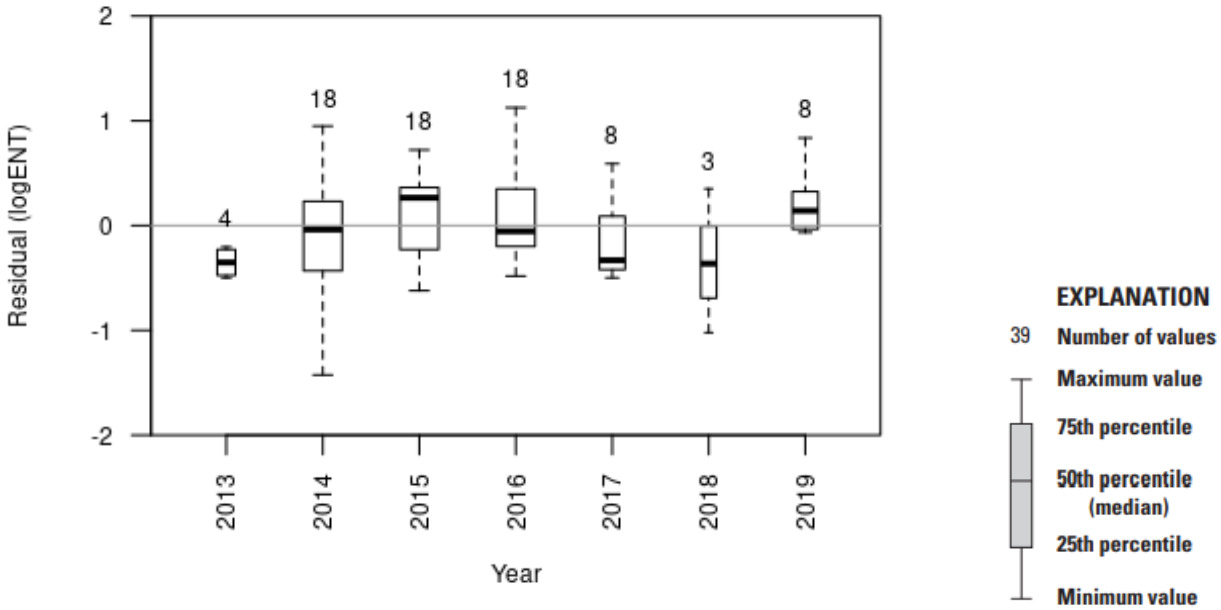
First row (right): Residual ENT related to the corresponding normal quantile of the residual with simple linear regression, indicated by the blue line.

Second row: Residual ENT related to date (left) and regression computed ENT multiplied by the BCF (right) with LOESS, indicated by the blue line.

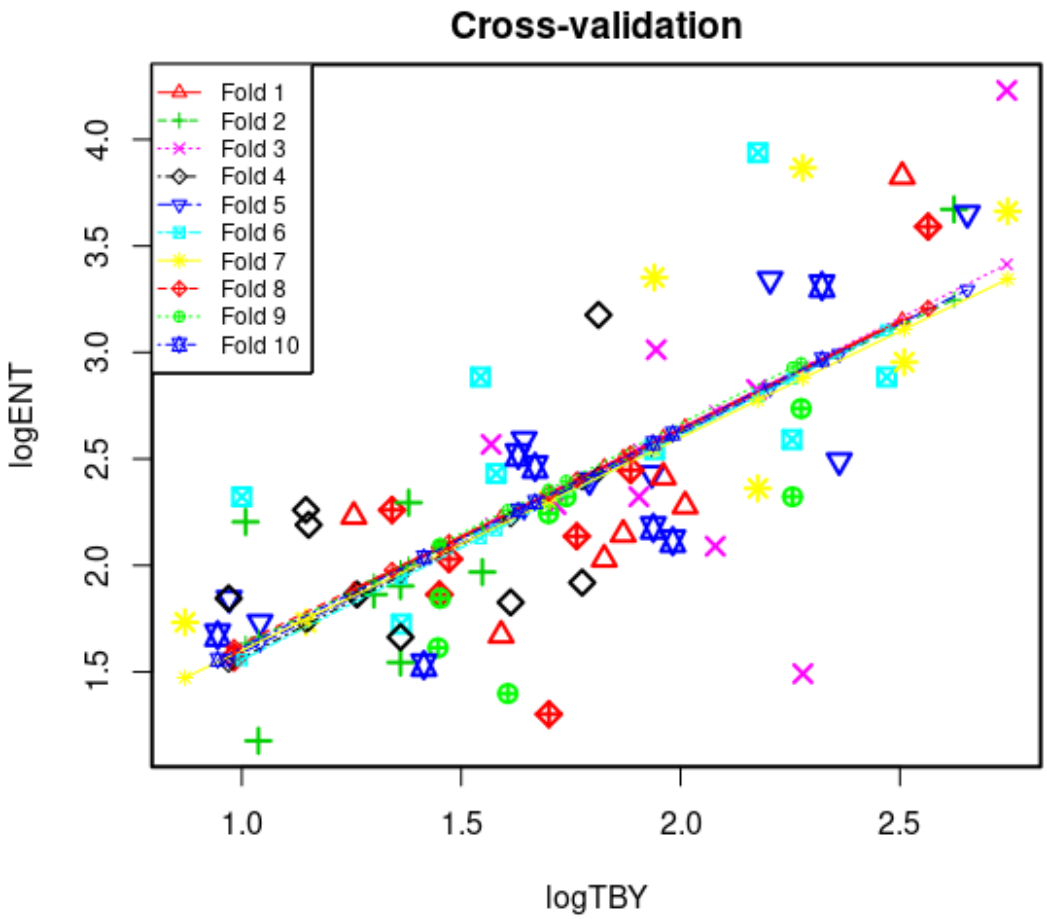
Third row: Observed ENT related to regression computed ENT.

Fourth row: Residual ENT related to TBY with LOESS, indicated by the blue line.


Residual (logENT)
2
1
0
-1
-2
4
5
5
5
8
11
7
7
8
8
5
4
Jan
Mar
May
Jul
Sep
Nov
Month
ENT
21,000
18,000
15,000
12,000
9,000
6,000
3,000
0
77
77
Computed
Observed
EXPLANATION
39 Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value


Residual (logENT)
2
1
0
-1
-2
4
18
18
18
8
3
8
2013
2014
2015
2016
2017
2018
2019
Year
EXPLANATION
39 Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value

Cross-Validation



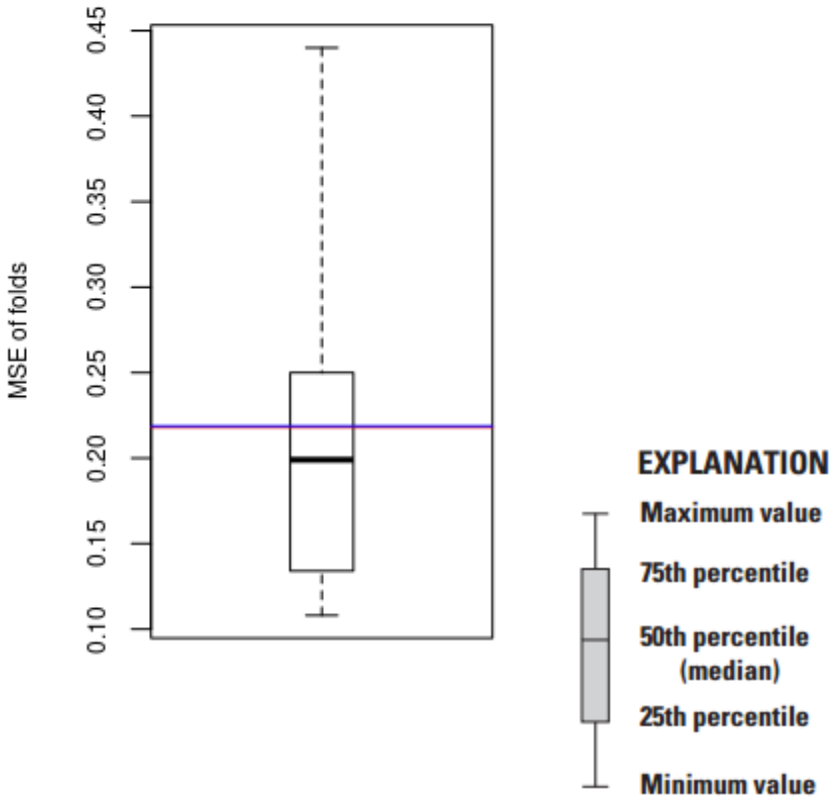

```
Fold - equal partition of the data (10 percent of the data).

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.


              Minimum MSE of folds:  0.108
                 Mean MSE of folds:  0.219
               Median MSE of folds:  0.199
              Maximum MSE of folds:  0.440
 (Mean MSE of folds) / (Model MSE):  1.000
```



Red line - Model MSE

Blue line - Mean MSE of folds

Model-Calibration Dataset

	Date	logENT	logTBY	ENT	TBY	Computed logENT	Computed ENT	Residual	Normal Quantiles	Censored Values
1	2013-09-23	1.61	1.45	41	28	2.06	201	-0.451	-1.04	--
2	2013-10-21	1.53	1.41	34	26	2.03	186	-0.499	-1.29	--
3	2013-11-18	1.86	1.45	73	28.2	2.07	202	-0.203	-0.4	--
4	2013-12-16	1.72	1.36	53	23.1	1.98	165	-0.253	-0.583	--
5	2014-01-13	2.19	1.15	155	14.2	1.76	99.9	0.429	0.981	--
6	2014-02-10	1.57	0.982	37	9.6	1.59	66.9	-0.0182	0.163	--
7	2014-03-10	1.67	0.945	47	8.81	1.55	61.2	0.124	0.297	--
8	2014-04-07	1.18	1.04	15	10.9	1.64	76.2	-0.467	-1.09	--
9	2014-05-05	2.46	1.67	290	46.6	2.29	338	0.172	0.331	--
10	2014-06-02	3.35	1.94	2250	87.1	2.57	643	0.783	1.68	--
11	2014-06-11	3.87	2.28	7360	190	2.92	1430	0.95	2.04	--
12	2014-06-30	2.3	1.7	200	50.2	2.32	365	-0.0227	0.065	--
13	2014-07-14	1.67	1.59	47	39	2.21	282	-0.539	-1.46	--
14	2014-07-28	1.54	1.36	35	23	1.98	164	-0.432	-0.93	--
15	2014-08-11	2.57	1.57	370	37	2.19	267	0.38	0.834	--
16	2014-08-25	2.29	1.72	193	52	2.34	379	-0.0538	-0.065	--
17	2014-09-08	2.43	1.93	270	86	2.56	635	-0.132	-0.229	--
18	2014-09-22	2.32	1.9	210	80.3	2.53	592	-0.211	-0.435	--
19	2014-10-06	1.49	2.28	31	190	2.92	1430	-1.43	-2.42	--
20	2014-10-20	2.03	1.83	107	67	2.45	491	-0.423	-0.881	--

21	2014-11-17	1.4	1.61	25	40.4	2.23	292	-0.829	-1.83	--
22	2014-12-15	2.43	1.58	270	38	2.2	274	0.232	0.365	--
23	2015-01-12	1.85	0.971	70	9.36	1.58	65.2	0.27	0.508	--
24	2015-02-09	2.32	1	210	10	1.6	69.7	0.718	1.37	--
25	2015-03-09	1.85	0.97	70	9.33	1.57	65	0.271	0.545	--
26	2015-04-06	1.73	0.87	54	7.42	1.47	51.3	0.261	0.4	--
27	2015-05-04	2.4	1.79	250	62	2.42	454	-0.0198	0.13	--
28	2015-05-18	2.88	2.47	767	295	3.11	2250	-0.228	-0.508	--
29	2015-06-01	2.49	2.36	310	230	3	1740	-0.511	-1.37	--
30	2015-06-15	2.59	2.25	390	180	2.89	1350	-0.301	-0.622	--
31	2015-06-29	2.32	2.26	210	180	2.89	1350	-0.571	-1.56	--
32	2015-07-13	2.09	2.08	123	120	2.71	893	-0.622	-1.68	--
33	2015-07-27	2.59	1.64	390	44	2.26	319	0.326	0.662	--
34	2015-08-10	3.01	1.94	1030	88	2.57	650	0.439	1.04	--
35	2015-08-24	2.24	1.7	175	50	2.32	364	-0.0789	-0.196	--
36	2015-09-08	2.88	1.54	767	35	2.16	252	0.722	1.46	--
37	2015-09-21	2.26	1.34	182	22	1.96	157	0.304	0.622	--
38	2015-10-05	1.86	1.26	73	18.3	1.87	130	-0.0106	0.196	--
39	2015-10-19	2.23	1.26	170	18	1.87	127	0.364	0.789	--
40	2015-11-16	2.26	1.15	182	14	1.75	98.5	0.506	1.15	--
41	2016-01-11	1.83	1.61	67	41	2.23	297	-0.407	-0.834	--
42	2016-02-08	1.9	1.36	80	23	1.98	164	-0.0727	-0.163	--
43	2016-03-03	1.73	1.15	54	14	1.75	98.5	-0.0222	0.0976	--
44	2016-04-04	1.73	1.04	54	11	1.65	76.9	0.0853	0.263	--
45	2016-05-02	3.83	2.51	6710	320	3.15	2440	0.677	1.29	--
46	2016-05-16	2.74	2.27	545	188	2.91	1420	-0.177	-0.297	--
47	2016-06-06	2.41	1.96	260	91.3	2.59	675	-0.175	-0.263	--
48	2016-06-20	1.92	1.78	83	59.7	2.4	436	-0.482	-1.15	--
49	2016-07-11	2.14	1.76	137	58	2.39	424	-0.251	-0.545	--
50	2016-07-25	1.97	1.55	93	35.3	2.17	255	-0.199	-0.331	--
51	2016-08-08	3.31	2.32	2050	210	2.96	1590	0.35	0.703	--
52	2016-08-22	3.34	2.2	2200	160	2.84	1200	0.502	1.09	--
53	2016-09-12	2.95	2.51	900	323	3.15	2470	-0.2	-0.365	--
54	2016-09-26	3.94	2.18	8670	150	2.81	1120	1.13	2.42	--
55	2016-10-11	3.18	1.81	1500	65	2.44	476	0.737	1.56	--
56	2016-10-24	2.29	1.38	197	24	1.99	171	0.3	0.583	--
57	2016-11-07	1.86	1.3	73	20	1.91	142	-0.0502	-0.0325	--
58	2016-12-12	2.03	1.47	107	29.7	2.09	213	-0.0599	-0.0976	--
59	2017-01-09	2.2	1.01	160	10.2	1.61	71.2	0.591	1.22	--
60	2017-02-06	1.66	1.36	46	23	1.98	164	-0.313	-0.662	--
61	2017-03-06	1.85	1.45	70	28.3	2.07	203	-0.224	-0.471	--
62	2017-04-10	2.11	1.98	130	96	2.61	710	-0.499	-1.22	--
63	2017-05-08	2.18	1.94	150	86.7	2.57	640	-0.391	-0.789	--
64	2017-05-22	3.67	2.62	4700	420	3.27	3230	0.402	0.93	--
65	2017-06-05	2.36	2.18	230	150	2.81	1120	-0.45	-0.981	--
66	2017-06-19	2.15	1.87	140	74	2.5	544	-0.351	-0.703	--
67	2018-10-11	3.65	2.65	4500	450	3.3	3470	0.352	0.746	--
68	2018-11-28	1.3	1.7	20	50.1	2.32	364	-1.02	-2.04	--
69	2018-12-17	2.28	2.01	190	102	2.64	759	-0.363	-0.746	--
70	2019-02-05	2.32	1.74	210	54.8	2.36	400	-0.0408	0	--
71	2019-03-18	3.66	2.75	4600	557	3.4	4310	0.267	0.435	--
72	2019-04-15	2.45	1.89	280	76.9	2.51	566	-0.0666	-0.13	--
73	2019-05-09	4.23	2.74	17000	554	3.39	4290	0.836	1.83	--
74	2019-06-25	3.59	2.56	3900	367	3.21	2810	0.381	0.881	--
75	2019-07-15	2.52	1.63	330	42.7	2.25	309	0.267	0.471	--

```
 76 2019-08-19   2.54   1.94   350 87.3     2.57       644  -0.0262    0.0325       --
 77 2019-09-23   2.83   2.17   670  149     2.81      1120   0.0173     0.229       --
```

Definitions

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

DFFITS: Difference in fits statistic (Helsel and others, 2020).

E.vars: Explanatory variables.

ENT: Enterococci, in colonies per 100 milliliters (90909).

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

LOESS: Local polynomial regression fitting, or locally estimated scatterplot smoothing (Helsel and others, 2020).

LOWESS: Locally weighted scatterplot smoothing (Cleveland, 1979; Helsel and others, 2020).

MSE: Model standard error (Helsel and others, 2020).

MSPE: Model standard percentage error (Helsel and others, 2020).

Probability(>|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).

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

References Cited

Cleveland, W.S., 1979, Robust locally weighted regression and smoothing scatterplots: Journal of the American Statistical Association, v. 74, no. 368, p. 829-836.

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.]

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.]

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.]

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.]