Appendix 27. Model Archival Summary for *Escherichia coli* 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 *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: 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 78 concurrent measurements of ECB 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. Forty-two 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, 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.

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

Ordinary least squares regression analysis was done using R programming language (R Core Team, 2020) to relate discretely collected ECB 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 ECB 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 ECB density at USGS site 06887500:

ECB density -based model:

$$\log ECB = 1.27 \times \log TBY - 0.304$$

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 1.90. The retransformed model, accounting for BCF is as follows:

$$ECB = 1.90 \times (TBY^{1.27} \times 10^{-0.304})$$

Previous Models

Start Year	End Year	Model Equation	Reference
1999	2003	logECB = 1.15logTBY - 0.425	Rasmussen and others (2005)

Model Statistics, Data, and Plots

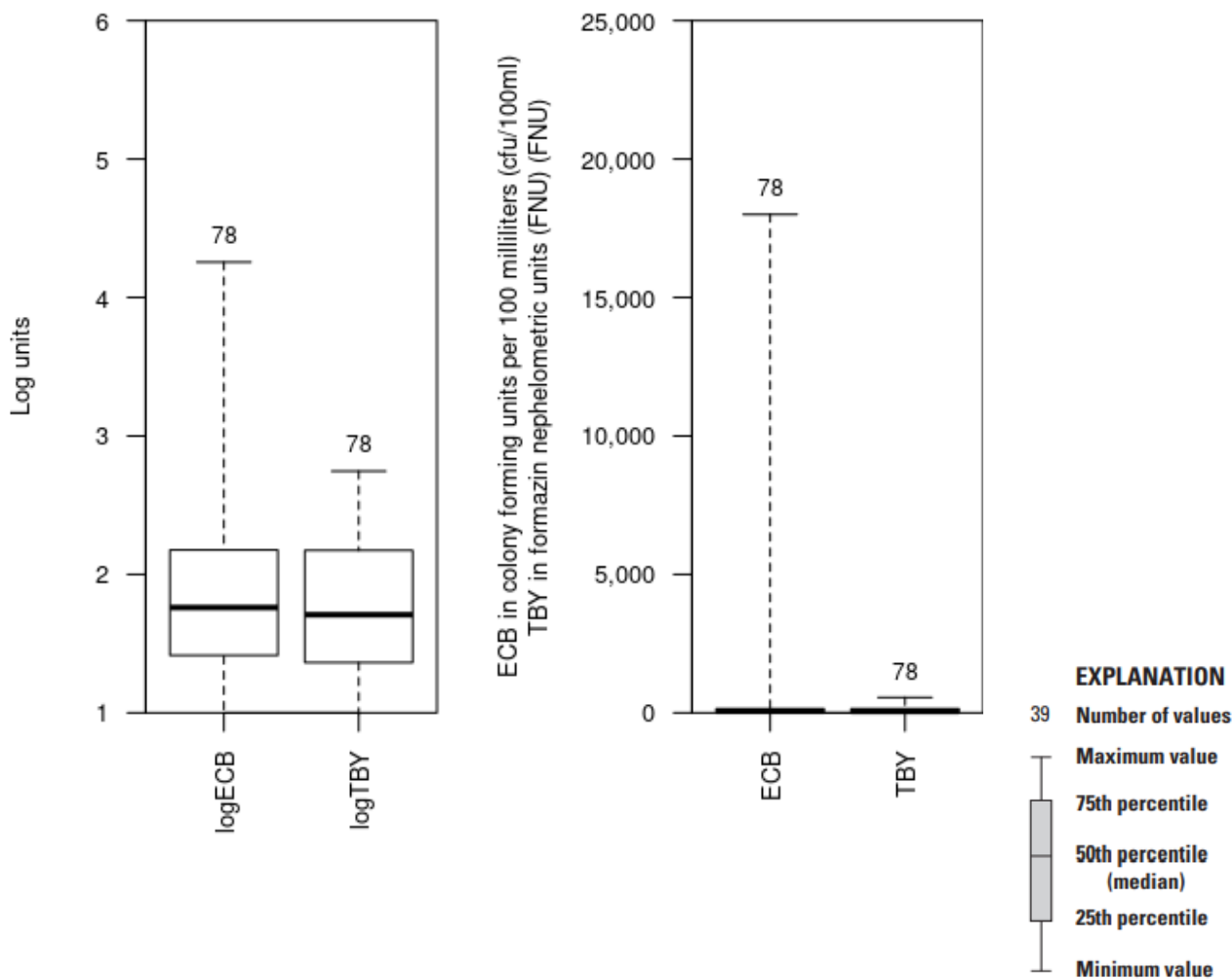## Model

logECB = + 1.27 * logTBY - 0.304

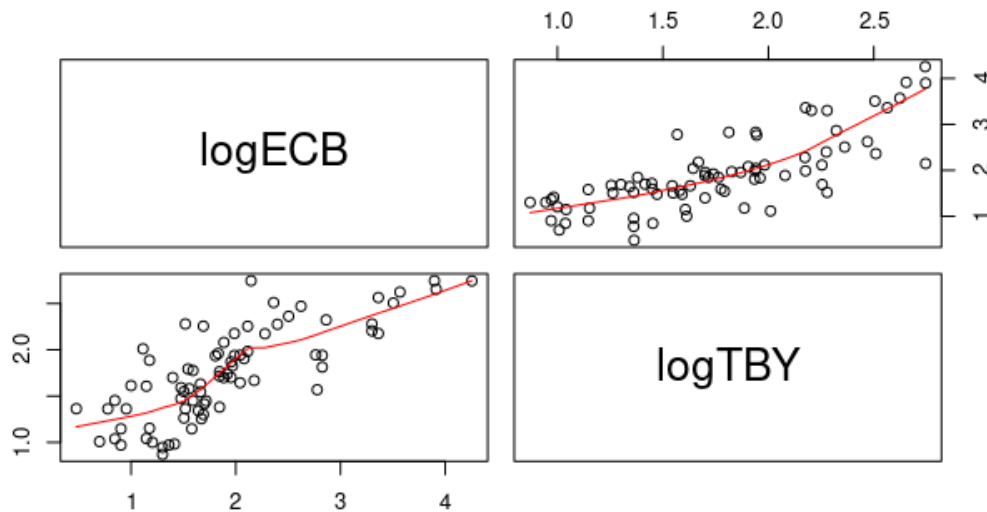## Variable Summary Statistics

```
              logECB    ECB logTBY     TBY
Minimum        0.477      3   0.87    7.42
1st Quartile   1.410     26   1.36   23.10
Median         1.760     58   1.71   51.10
Mean           1.920    734   1.75  107.00
3rd Quartile   2.180    150   2.17  149.00
Maximum        4.260  18000   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                              78
Standard error (RMSE)                            0.513
Average Model standard percentage error (MSPE)     148
Coefficient of determination (R²)                0.603
Adjusted Coefficient of Determination (Adj. R²) 0.598
Bias Correction Factor (BCF)                       1.9
```

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	-0.304	0.215	-1.42	1.61e-01
logTBY	1.270	0.118	10.70	6.76e-17

Correlation Matrix

	Intercept	E.vars
Intercept	1.000	-0.963
E.vars	-0.963	1.000

Outlier Test Criteria

Leverage	Cook's D	DFFITS
0.0769	0.1943	0.3203

Flagged Observations

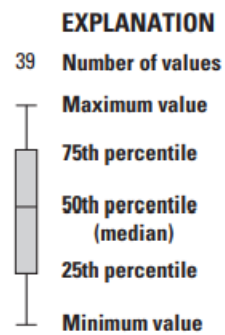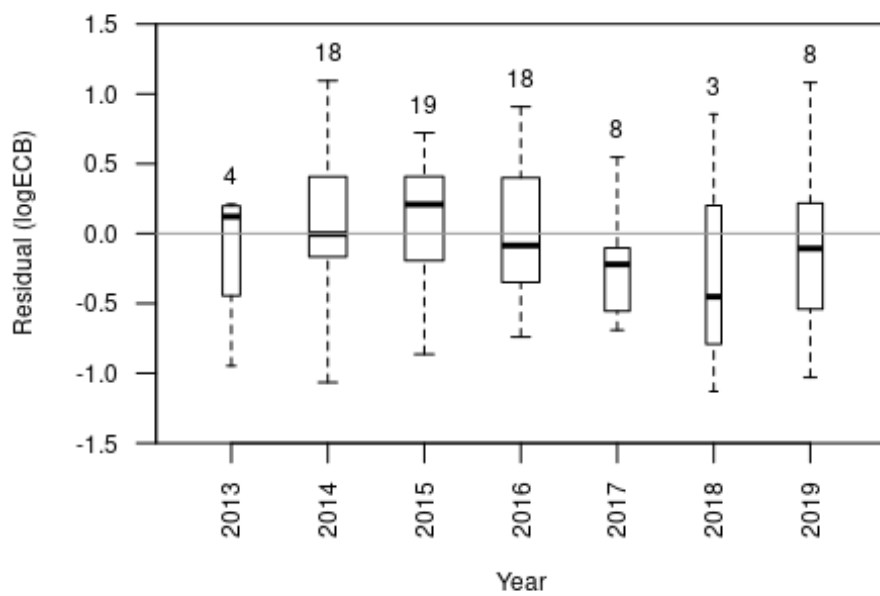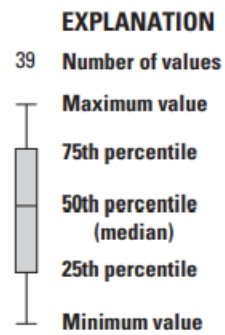|	logECB	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
201410060900	1.52	2.58	-1.070	-2.10	-2.15	0.0274	0.0624	-0.362
201512140910	3.90	3.18	0.722	1.45	1.46	0.0649	0.0733	0.386
201810110940	3.91	3.06	0.855	1.71	1.74	0.0557	0.0865	0.421
201903180940	2.15	3.18	-1.030	-2.07	-2.12	0.0649	0.1490	-0.559
201905091520	4.26	3.17	1.080	2.18	2.24	0.0647	0.1640	0.588

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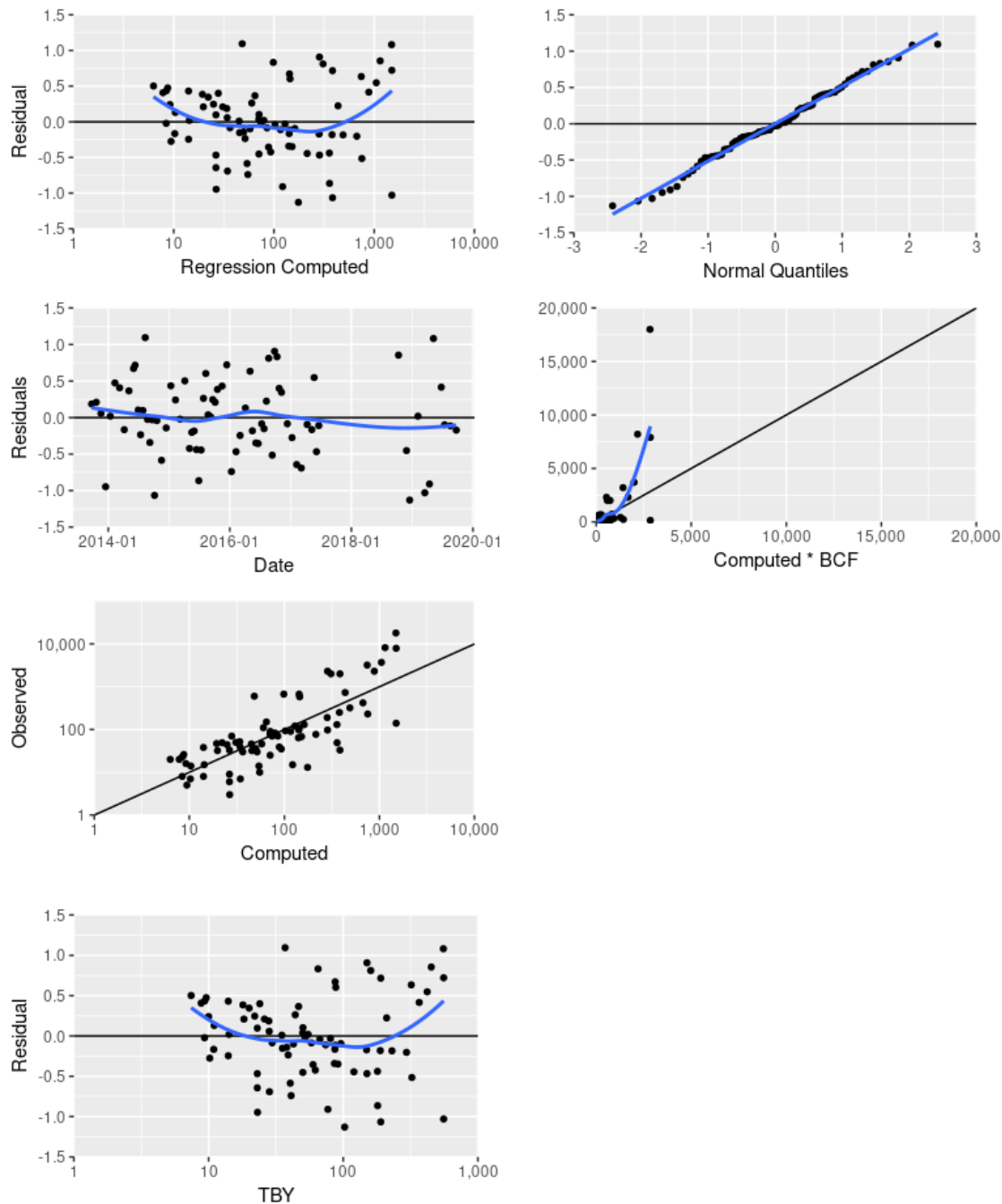
First row (left): Residual ECB related to regression computed ECB with local polynomial regression fitting, or locally estimated scatterplot smoothing (LOESS), indicated by the blue line.

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

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

Third row: Observed ECB related to regression computed ECB.

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

EXPLANATION

39   Number of values

Maximum value

75th percentile

50th percentile
(median)

25th percentile

Minimum value
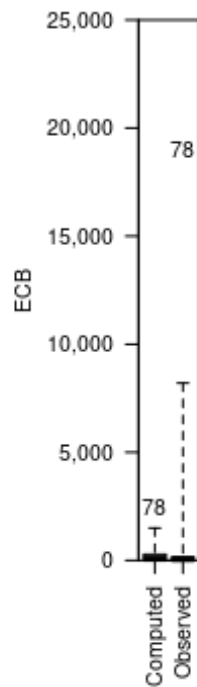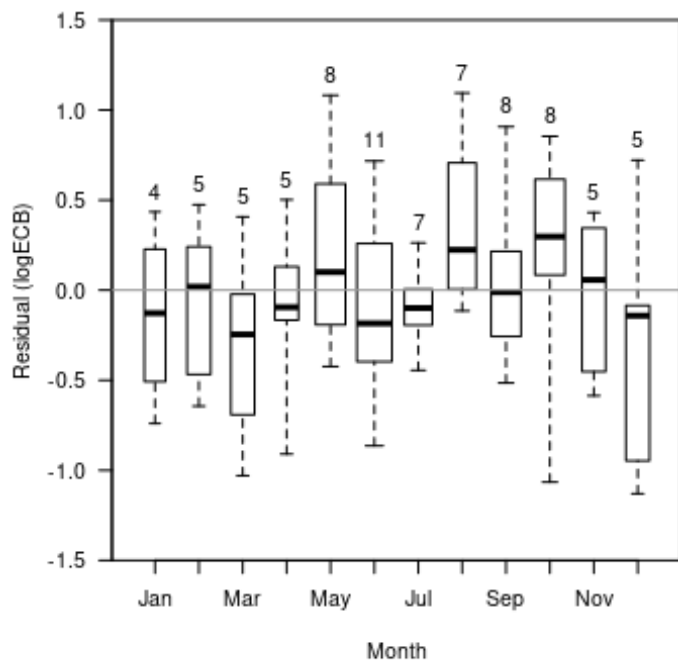
Cross-Validation
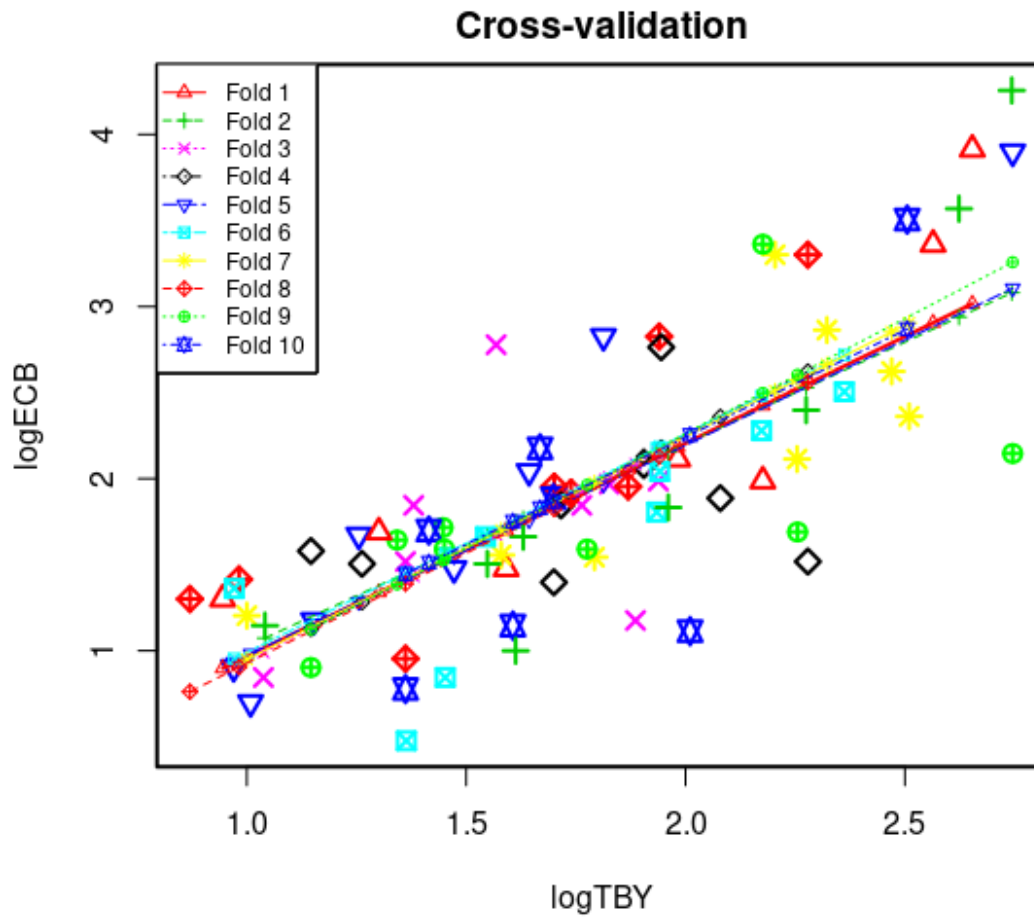
Fold - equal partition of the data (10 percent of the data).

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:  0.187
           Mean MSE of folds:  0.273
         Median MSE of folds:  0.258
        Maximum MSE of folds:  0.389
(Mean MSE of folds) / (Model MSE):  1.040
```

EXPLANATION

Maximum value

75th percentile

50th percentile
(median)

25th percentile

Minimum value

Red line - Model MSE

Blue line - Mean MSE of folds

Model-Calibration Dataset

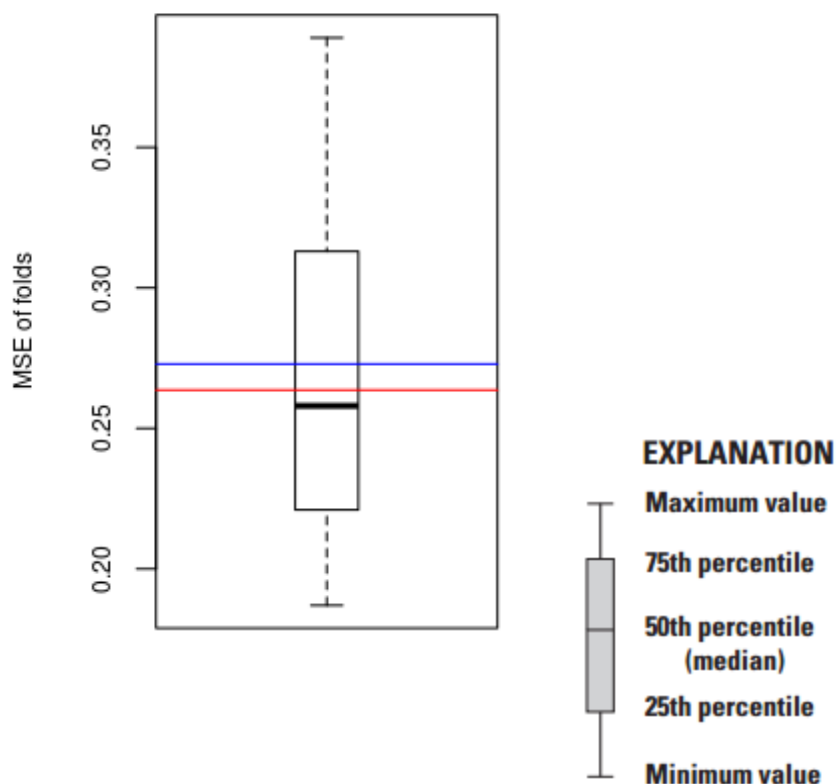|	Date	logECB	logTBY	ECB	TBY	Computed logECB	Computed ECB	Residual	Normal Quantiles	Censored Values
0										
1	2013-09-23	1.72	1.45	52	28	1.53	64.4	0.186	0.343	--
2	2013-10-21	1.7	1.41	50	26	1.49	58.6	0.21	0.412	--
3	2013-11-18	1.59	1.45	39	28.2	1.53	64.9	0.0571	0.21	--
4	2013-12-16	0.477	1.36	3	23.1	1.42	50.4	-0.947	-1.68	--
5	2014-01-13	1.18	1.15	15	14.2	1.16	27.2	0.0198	0.145	--
6	2014-02-10	1.41	0.982	26	9.6	0.941	16.6	0.474	0.939	--
7	2014-03-10	1.3	0.945	20	8.81	0.893	14.9	0.408	0.755	--
8	2014-04-07	0.845	1.04	7	10.9	1.01	19.5	-0.166	-0.377	--
9	2014-05-05	2.18	1.67	150	46.6	1.81	123	0.366	0.633	--
10	2014-06-02	2.83	1.94	670	87.1	2.15	271	0.671	1.23	--
11	2014-06-11	3.3	2.28	2000	190	2.58	729	0.717	1.3	--
12	2014-06-30	1.95	1.7	90	50.2	1.85	135	0.103	0.276	--
13	2014-07-14	1.48	1.59	30	39	1.71	98	-0.235	-0.556	--
14	2014-07-28	1.52	1.36	33	23	1.42	50.2	0.0968	0.243	--
15	2014-08-11	2.78	1.57	600	37	1.68	91.6	1.09	2.42	--
16	2014-08-25	1.85	1.72	70	52	1.87	141	-0.0257	0.016	--
17	2014-09-08	1.81	1.93	64	86	2.15	267	-0.342	-0.672	--
18	2014-09-22	2.08	1.9	120	80.3	2.11	245	-0.031	-0.016	--
19	2014-10-06	1.52	2.28	33	190	2.58	729	-1.07	-2.04	--
20	2014-10-20	1.97	1.83	93	67	2.01	194	-0.0418	-0.0481	--

21	2014-11-17	1.15	1.61	14	40.4	1.73	102	-0.586	-1.16	--
22	2014-12-15	1.56	1.58	36	38	1.7	94.8	-0.142	-0.276	--
23	2015-01-12	1.36	0.971	23	9.36	0.927	16	0.435	0.89	--
24	2015-02-09	1.2	1	16	10	0.963	17.5	0.241	0.483	--
25	2015-03-09	0.903	0.97	8	9.33	0.925	16	-0.0222	0.0481	--
26	2015-04-06	1.3	0.87	20	7.42	0.799	12	0.502	0.99	--
27	2015-05-04	1.54	1.79	35	62	1.97	176	-0.424	-0.799	--
28	2015-05-18	2.62	2.47	420	295	2.83	1270	-0.203	-0.519	--
29	2015-06-01	2.51	2.36	320	230	2.69	929	-0.184	-0.483	--
30	2015-06-15	2.11	2.25	130	180	2.55	678	-0.439	-0.843	--
31	2015-06-29	1.69	2.26	49	180	2.55	681	-0.864	-1.46	--
32	2015-07-13	1.89	2.08	77	120	2.33	407	-0.445	-0.89	--
33	2015-07-27	2.04	1.64	110	44	1.78	114	0.263	0.556	--
34	2015-08-10	2.76	1.94	580	88	2.16	275	0.603	1.1	--
35	2015-08-24	1.89	1.7	77	50	1.85	134	0.0373	0.177	--
36	2015-09-08	1.66	1.54	46	35	1.65	85.4	0.00992	0.0802	--
37	2015-09-21	1.64	1.34	44	22	1.4	47.4	0.246	0.519	--
38	2015-10-05	1.51	1.26	32	18.3	1.3	37.5	0.209	0.377	--
39	2015-10-19	1.67	1.26	47	18	1.29	36.8	0.385	0.672	--
40	2015-11-16	1.58	1.15	38	14	1.15	26.7	0.431	0.843	--
41	2015-12-14	3.9	2.75	7900	557	3.18	2850	0.722	1.38	--
42	2016-01-11	1	1.61	10	41	1.74	104	-0.74	-1.38	--
43	2016-02-08	0.954	1.36	9	23	1.42	50.2	-0.467	-1.04	--
44	2016-03-03	0.903	1.15	8	14	1.15	26.7	-0.245	-0.594	--
45	2016-04-04	1.15	1.04	14	11	1.02	19.7	0.13	0.309	--
46	2016-05-02	3.51	2.51	3200	320	2.87	1410	0.634	1.16	--
47	2016-05-16	2.4	2.27	250	188	2.58	721	-0.181	-0.447	--
48	2016-06-06	1.83	1.96	68	91.3	2.18	288	-0.348	-0.713	--
49	2016-06-20	1.59	1.78	39	59.7	1.95	168	-0.355	-0.755	--
50	2016-07-11	1.85	1.76	70	58	1.93	162	-0.0858	-0.112	--
51	2016-07-25	1.51	1.55	32	35.3	1.66	86.4	-0.153	-0.309	--
52	2016-08-08	2.86	2.32	730	210	2.64	828	0.224	0.447	--
53	2016-08-22	3.3	2.2	2000	160	2.49	586	0.812	1.46	--
54	2016-09-12	2.36	2.51	230	323	2.88	1430	-0.515	-1.1	--
55	2016-09-26	3.36	2.18	2300	150	2.45	540	0.908	1.84	--
56	2016-10-11	2.83	1.81	670	65	1.99	187	0.832	1.56	--
57	2016-10-24	1.85	1.38	70	24	1.45	52.9	0.4	0.713	--
58	2016-11-07	1.69	1.3	49	20	1.34	42	0.345	0.594	--
59	2016-12-12	1.48	1.47	30	29.7	1.56	69.3	-0.0847	-0.0802	--
60	2017-01-09	0.699	1.01	5	10.2	0.974	17.9	-0.275	-0.633	--
61	2017-02-06	0.778	1.36	6	23	1.42	50.2	-0.644	-1.23	--
62	2017-03-06	0.845	1.45	7	28.3	1.54	65.3	-0.691	-1.3	--
63	2017-04-10	2.11	1.98	130	96	2.21	307	-0.0943	-0.145	--
64	2017-05-08	1.99	1.94	97	86.7	2.15	270	-0.165	-0.343	--
65	2017-05-22	3.57	2.62	3700	420	3.02	1990	0.547	1.04	--
66	2017-06-05	1.99	2.18	97	150	2.45	540	-0.467	-0.99	--
67	2017-06-19	1.95	1.87	90	74	2.06	221	-0.111	-0.21	--
68	2018-10-11	3.91	2.65	8200	450	3.06	2180	0.855	1.68	--
69	2018-11-28	1.4	1.7	25	50.1	1.85	135	-0.452	-0.939	--
70	2018-12-17	1.11	2.01	13	102	2.24	333	-1.13	-2.42	--
71	2019-02-05	1.92	1.74	83	54.8	1.9	151	0.0191	0.112	--
72	2019-03-18	2.15	2.75	140	557	3.18	2850	-1.03	-1.84	--
73	2019-04-15	1.18	1.89	15	76.9	2.09	232	-0.91	-1.56	--
74	2019-05-09	4.26	2.74	18000	554	3.17	2830	1.08	2.04	--
75	2019-06-25	3.36	2.56	2300	367	2.95	1680	0.416	0.799	--

```
76 2019-07-15   1.66   1.63    46 42.7    1.76    110  -0.0991   -0.177    --
77 2019-08-19   2.04   1.94   110 87.3    2.16    272  -0.115    -0.243    --
78 2019-09-23   2.28   2.17   190  149    2.45    536  -0.172    -0.412    --
```

Definitions

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

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

E.vars: Explanatory variables.

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

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.]