Appendix 14. Model Archival Summary for *Escherichia coli* Bacteria Concentration at U.S. Geological Survey Site 06888990, Kansas River above Topeka Weir at Topeka, Kansas, during November 2018 through June 2021

This model archival summary summarizes the *Escherichia coli* bacteria (ECB; U.S. Geological Survey [USGS] parameter code 90902) concentration model developed to compute 15-minute ECB concentrations from November 2018 onward. This model is specific to USGS site 06888990, the Kansas River above Topeka Weir at Topeka, Kansas, during this study period and cannot be applied to data collected from other sites on the Kansas River or data collected from other waterbodies.

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: 06888990
Site name: Kansas River above Topeka Weir at Topeka, Kans.
Location: Lat 39°04'19", long 95°42'58" referenced to North American Datum of 1927, in NW 1/4 sec.23, T.11 S., R.15 E., Shawnee County, Kans., hydrologic unit 10270102.

Equipment: A Xylem YSI EXO2 water-quality monitor equipped with sensors for water temperature, specific conductance, dissolved oxygen, pH, turbidity (TBY), and chlorophyll and phycocyanin fluorescence was installed during November 2018 through June 2021. Readings from the water-quality monitor were recorded every 15 minutes and transmitted by way of satellite, hourly.

Date model was created: December 8, 2021

Model-calibration data period: November 28, 2018, through June 21, 2021

Model-application date: November 28, 2018, 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 USGS National Water Information System (U.S. Geological Survey, 2022) database and available to the public. Ordinary least squares analysis was used to develop regression models using R programming language (R Core Team, 2022). 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 2 hours in the continuous data record resulted in missing interpolated data). Seasonal components (sine and cosine variables) also were evaluated as potential explanatory variables. Previously published explanatory variables (Rasmussen and others, 2005; Foster and Graham, 2016; Williams, 2021) at other Kansas River sites were strongly considered for continuity in model form.

The final selected regression model was based on 34 concurrent measurements of ECB concentration and sensor-measured TBY during November 28, 2018, through June 21, 2021. Samples were collected throughout the range of continuously observed hydrologic conditions. No samples had concentrations below laboratory minimum reporting limits. Thirteen sample densities were qualified as “estimated.”

Potential outliers initially were identified using scatterplots of the ECB and TBY model-calibration data (Rasmussen and others, 2009). Studentized residuals from the model were inspected for values greater than three or less than negative three (Pardoe, 2020). Values outside of that range were considered potential outliers and were investigated. Additionally, computations of leverage, Cook’s distance (Cook’s D), and difference in fits (DFFITS) statistics were used to estimate potential outlier effect on the final selected regression model (Cook, 1977; Helsel and others, 2020). Outliers were investigated for potential removal from the model-calibration dataset by confirming correct database entry, evaluating laboratory analytical performance, and reviewing field notes associated with the sample in question (Rasmussen and others, 2009). All potential outliers were not determined to have errors associated with sample collection, processing, or analysis and were therefore considered valid.

Escherichia coli Bacteria Sampling Details

Indicator bacteria samples were collected either from the southern bank of the Kansas River above Topeka Weir. The grab sample collection method with weighted basket was used for all indicator bacteria samples (contrary to the single vertical collection method used for all other analytes; U.S. Geological Survey, variously dated). During November 2018 through June 2020, samples were collected on a biweekly to monthly basis. During July 2020 through June 2021, samples were collected on a monthly to quarterly basis, depending on flow conditions. Samples occasionally were collected during targeted reservoir release and runoff events to get a more representative dataset. An open-mouth bottle with weighted-basket sampler was used. Samples were analyzed for ECB concentration at the USGS Kansas Water Science Center in Lawrence, Kans.

Model Development

Ordinary least squares regression analysis was done using the *stats* (*v4.3.0*) package in R programming language (R Core Team, 2022) to relate discretely collected ECB concentration to sensor-measured TBY. The distribution of residuals (the difference between the measured and computed values) was examined for normality, and the plots of residuals were examined for homoscedasticity (departures from zero did not change substantially over the range of computed values).

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, all relevant sample data, and additional statistical information are included in the Model Statistics, Data, and Plots section of this appendix.

Model Summary

The following is a summary of the final regression analysis for ECB concentration at USGS site 06888990:

ECB concentration-based model:

$$\log ECB = 1.69(\log TBY) - 1.04$$

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

$$ECB\ = 1.85 \times (TBY^{1.69} \times 10^{-1.04})$$

This model was developed using continuous and discrete water-quality data collected during November 2018 through June 2021. These data were collected throughout the observed range of streamflow conditions during this time. However, a limitation in model accuracy during conditions outside of those observed during November 2018 through June 2021 should be considered when interpreting model computations beyond June 2021.

Previous Models

There are no previously published models at this site. However, similar models have been published at other Kansas River sites, as documented by Rasmussen and others (2005), Foster and Graham (2016), and Williams (2021).

Model Statistics, Data, and Plots

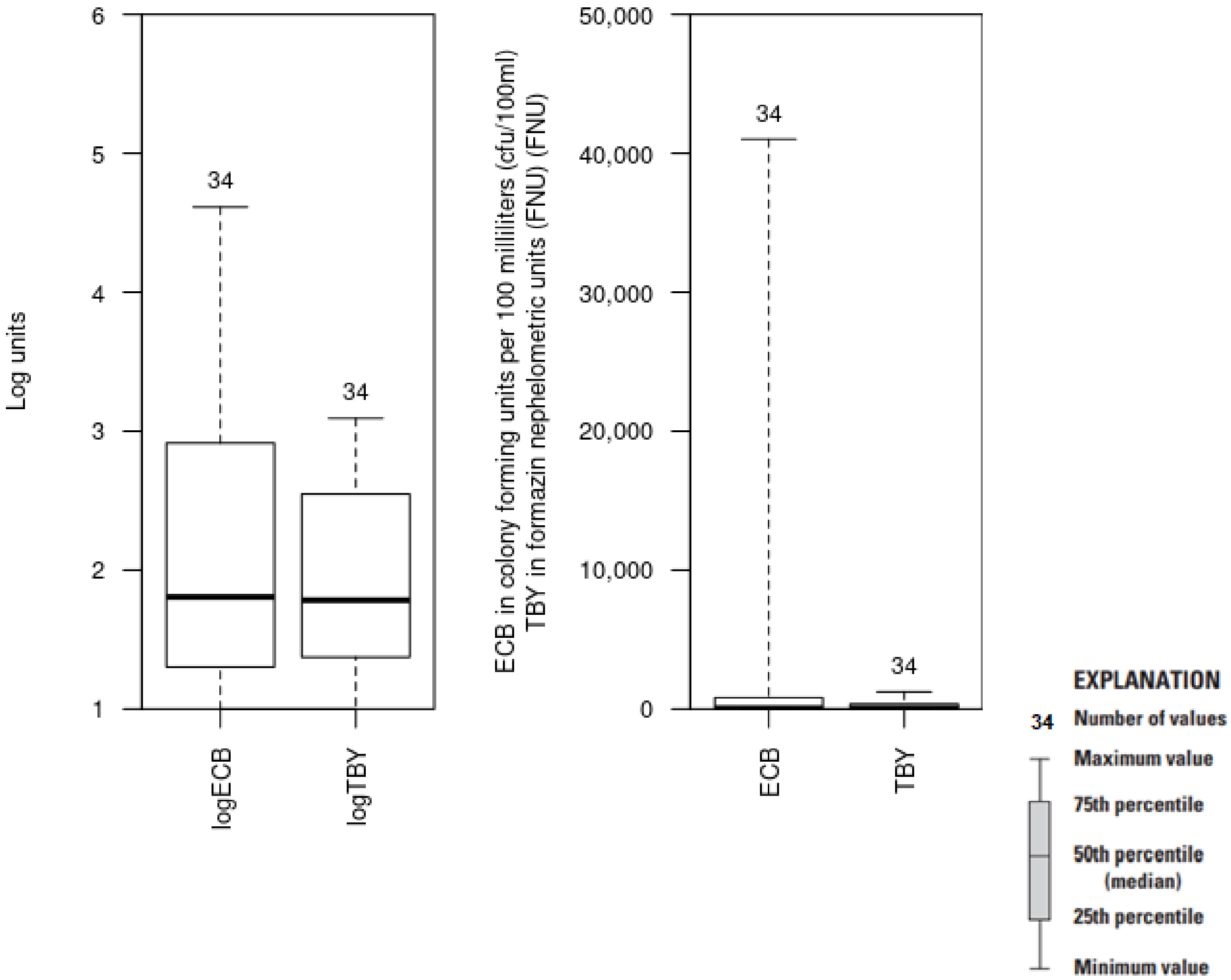
Model

$\log ECB = 1.69(\log TBY) - 1.04$

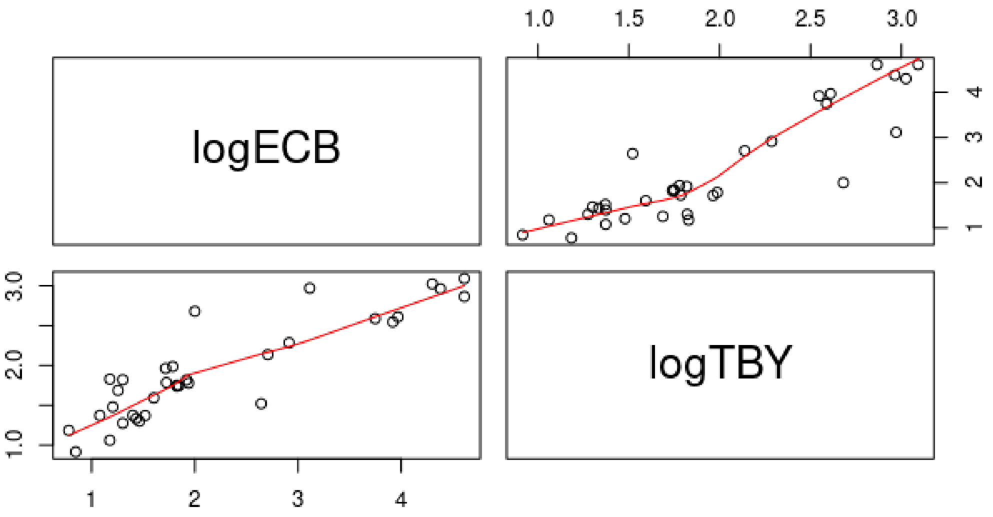
Variable Summary Statistics

```
             logECB    ECB  logTBY   TBY
Minimum       0.778   6.00   0.916  8.23
1st Quartile  1.300     20   1.370  23.6
Median        1.810   64.0   1.780  60.6
Mean          2.210  4,510   1.930   230
3rd Quartile  2.910    820   2.550   352
Maximum       4.610 41,000   3.090 1,240
```

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                              34
Standard error (RMSE)                            0.538
Mean model standard percentage error (MSPE)        158
Coefficient of determination (R²)                0.797
Adjusted coefficient of determination (Adj. R²) 0.791
Bias correction factor (BCF)                      1.85
```

Explanatory Variables

```
            Coefficients Standard Error t value Pr(>|t|)
(Intercept)        -1.04          0.304   -3.43 1.68e-03
logTBY              1.69          0.150   11.20 1.27e-12
```

Correlation Matrix

```
          Intercept E.vars
Intercept     1.000 -0.953
E.vars       -0.953  1.000
```

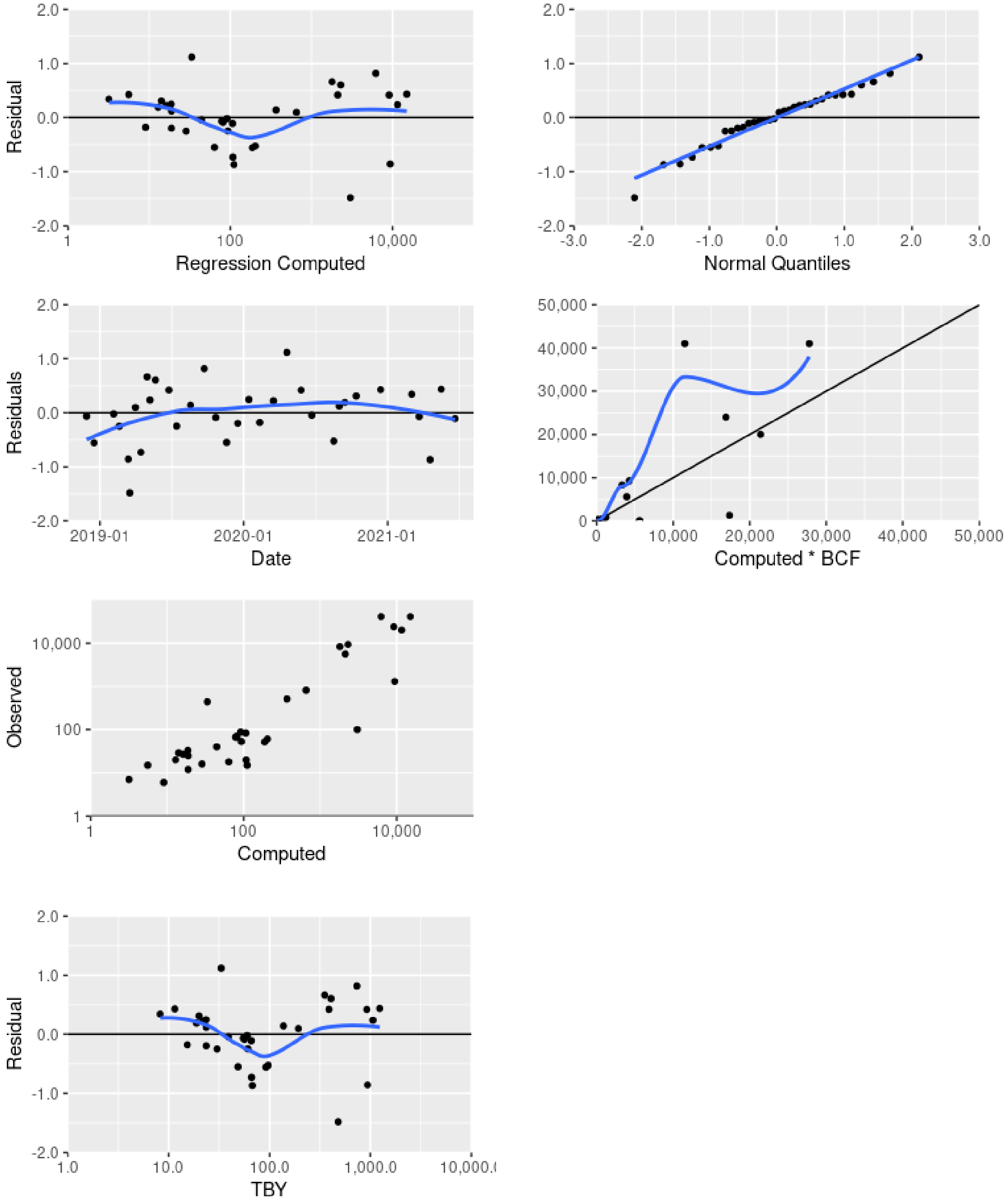
Outlier Test Criteria

```
Leverage Cook's D   DFFITS
   0.176    0.194    0.485
```

Flagged Observations

```
          logECB Estimate Residual Standard Residual Studentized Residual Leverage Cook's D DFFITS
201903141040   3.11     3.97   -0.858             -1.70                -1.75   0.1150    0.187 -0.631
201903181030   2.00     3.48   -1.480             -2.86                -3.27   0.0741    0.328 -0.925
201909230810   4.61     3.79    0.819              1.60                 1.65   0.0987    0.141  0.545
```

Statistical Plots



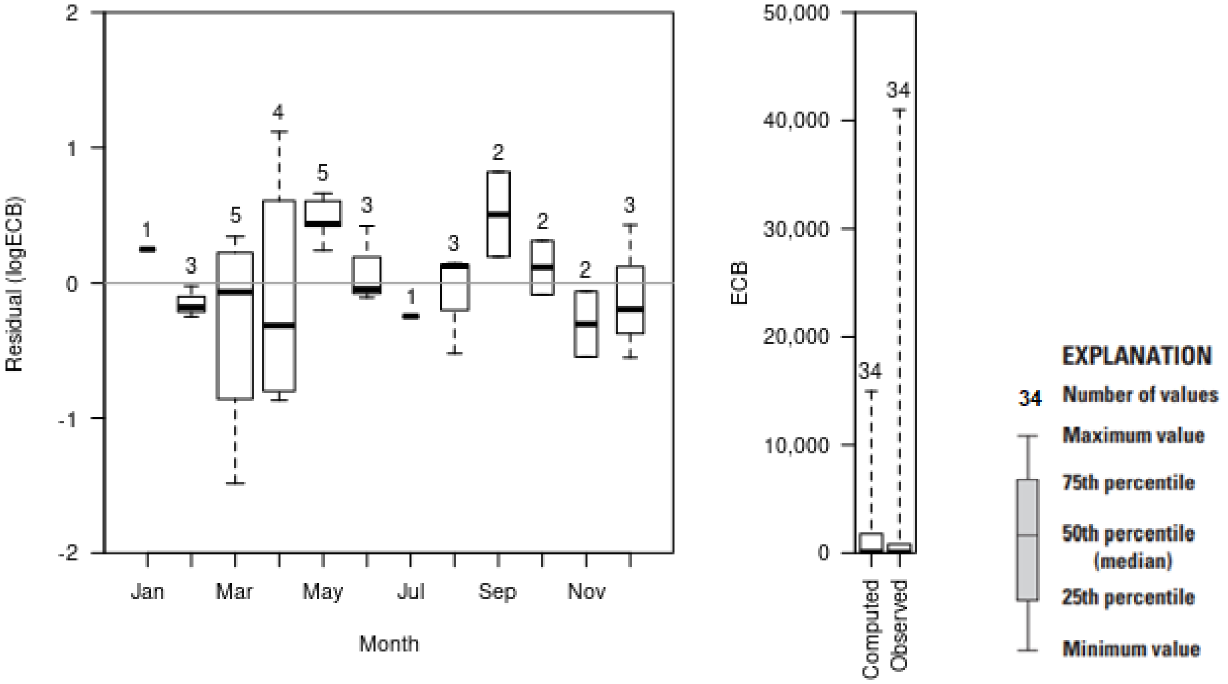
First row (left): Residual ECB related to regression computed ECB with local polynomial regression fitting, or locally estimated scatterplot smoothing (LOESS), indicated by the blue line.

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

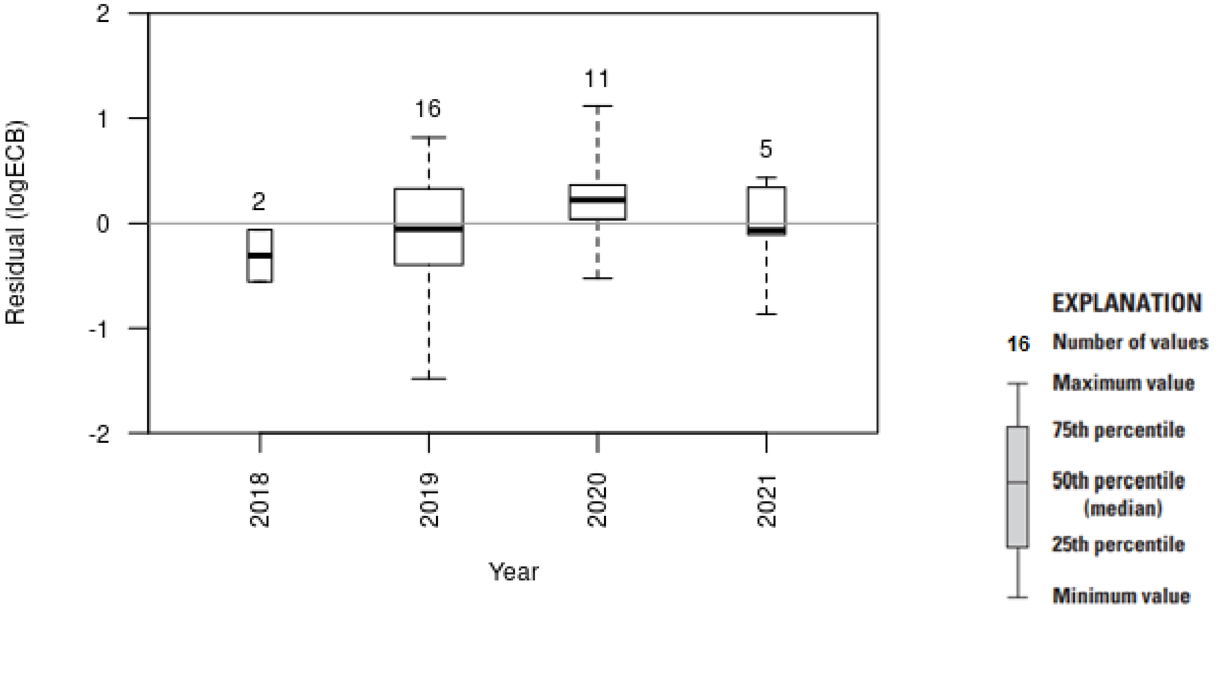
Second row: Residual ECB related to date (left) and regression computed ECB multiplied by the BCF (right) with LOESS, indicated by the blue line.

Third row: Observed ECB related to regression computed ECB.

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

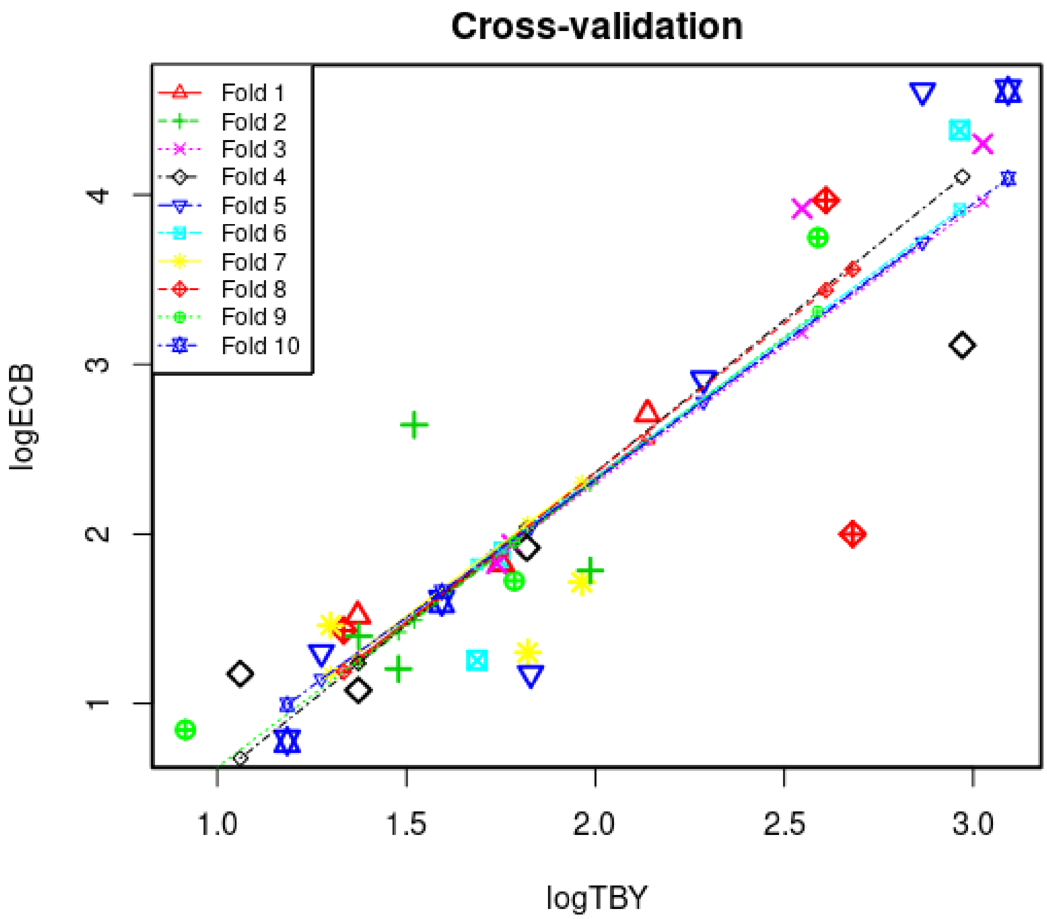


Residual (logECB)
2
1
0
-1
-2
1
3
5
4
5
3
1
3
2
2
2
3
Jan
Mar
May
Jul
Sep
Nov
Month
ECB
50,000
40,000
30,000
20,000
10,000
0
34
34
Computed
Observed
EXPLANATION
34 Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value



Residual (logECB)
2
1
0
-1
-2
2
16
11
5
2018
2019
2020
2021
Year
EXPLANATION
16 Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value

Cross Validation



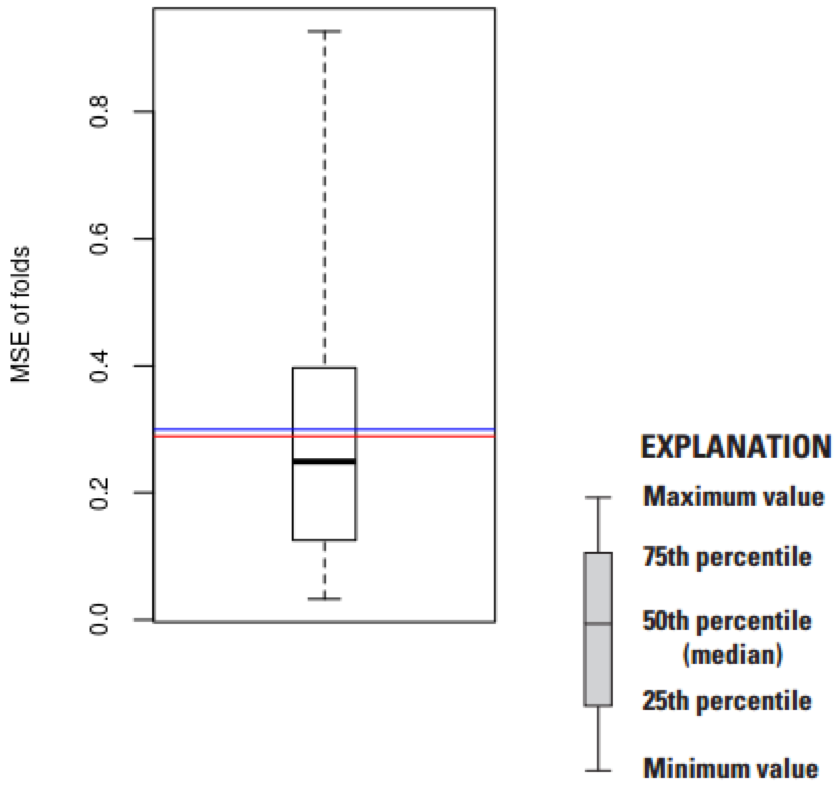
```
Fold - equal partition of the data (10 percent of the data).

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:  0.0322
                 Mean MSE of folds:  0.3000
               Median MSE of folds:  0.2500
              Maximum MSE of folds:  0.9270
 (Mean MSE of folds) / (Model MSE):  1.0400
```



```
Red line - Model MSE

Blue line - Mean MSE of folds
```

Model-Calibration Dataset

```
          Date logECB logTBY   ECB  TBY Computed  Computed Residual    Normal Censored
 0                                        logECB       ECB          Quantiles   Values
 1 2018-11-28   1.85   1.75    70 55.9     1.91       149  -0.0614    -0.184       --
 2 2018-12-17   1.72   1.96   E52   92     2.27       346   -0.556     -1.11       --
 3 2019-02-05   1.94   1.78    87 60.2     1.96       169  -0.0211   -0.0367       --
 4 2019-02-19    1.2   1.48   E16 30.2     1.45      52.7    -0.25    -0.765       --
 5 2019-03-14   3.11   2.97  1300  935     3.97     17400   -0.858     -1.43       --
 6 2019-03-18      2   2.68  E100  479     3.48      5620    -1.48     -2.11       --
 7 2019-04-01   2.91   2.29   820  193     2.82      1210   0.0983    0.0367       --
 8 2019-04-15    1.3   1.82   E20 66.3     2.03       199    -0.73     -1.25       --
 9 2019-05-01   3.92   2.55  8300  352     3.26      3340    0.663      1.43       --
10 2019-05-08    4.3   3.03 20000 1060     4.06     21400    0.238     0.415       --
11 2019-05-22   3.97   2.61  9300  407     3.36      4270    0.606      1.25       --
12 2019-06-25   3.75   2.59  5600  388     3.33      3930    0.422     0.867       --
13 2019-07-15   1.72   1.79   E53   61     1.97       173   -0.247     -0.67       --
14 2019-08-19   2.71   2.14   510  137     2.57       681    0.142     0.184       --
15 2019-09-23   4.61   2.87 41000  734     3.79     11500    0.819      1.68       --
16 2019-10-22   1.83   1.75    67 56.3     1.91       151  -0.0861    -0.336       --
17 2019-11-19   1.26   1.69   E18 48.7      1.8       118    -0.55    -0.979       --
18 2019-12-17   1.08   1.37   E12 23.6     1.27      34.8   -0.195    -0.581       --
19 2020-01-14   1.52   1.37    33 23.5     1.27      34.6    0.247     0.496       --
20 2020-02-11  0.778   1.18    E6 15.3    0.957      16.8   -0.179    -0.496       --
```

```
21 2020-03-17   1.43   1.33    27 21.6     1.21        30    0.222     0.336       --
22 2020-04-20   2.64   1.52   440 33.2     1.52        62     1.12      2.11       --
23 2020-05-26   4.38   2.96 24000  920     3.96     16900     0.42     0.765       --
24 2020-06-22    1.6   1.59    40 39.2     1.65        82  -0.0444     -0.11       --
25 2020-08-17   1.79   1.99   E61   97     2.31       379   -0.525    -0.867       --
26 2020-08-31    1.4   1.37    25 23.6     1.28      34.9    0.122      0.11       --
27 2020-09-14    1.3   1.28   E20 18.9     1.11      23.8    0.191     0.259       --
28 2020-10-13   1.46    1.3    29   20     1.15      26.3    0.311     0.581       --
29 2020-12-14   1.18   1.06   E15 11.5    0.747      10.4    0.429     0.979       --
30 2021-03-03  0.845  0.916    E7 8.23    0.502      5.89    0.343      0.67       --
31 2021-03-22   1.83   1.74    67 54.8     1.89       145  -0.0666    -0.259       --
32 2021-04-19   1.18   1.83   E15 67.5     2.04       205   -0.868     -1.68       --
33 2021-05-17   4.61   3.09 41000 1240     4.18     27800    0.436      1.11       --
34 2021-06-21   1.92   1.82    83 65.9     2.03       197   -0.108    -0.415       --

E: estimated
```

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 (USGS parameter code 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: Mean square 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 (USGS parameter code 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.

Cook, R.D., 1977, Detection of influential observations in linear regression: Technometrics, v. 19, no. 1, p. 15–18. [Also available at https://doi.org/10.2307/1268249.]

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 February 2022 at https://doi.org/10.3133/ofr20161040.

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.]

Pardoe, I., 2020, Applied Regression Modeling: United Kingdom, John Wiley & Sons, 336 p.

R Core Team, 2022, R—A language and environment for statistical computing, version 4.1.2: Vienna, Austria, R Foundation for Statistical Computing, accessed December 2021 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, 2022, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed February 2022 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 February 2022 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.]

Williams, T.J., 2021, Linear regression model documentation and updates for computing water-quality constituent concentrations or densities using continuous real-time water-quality data for the Kansas River, Kansas, July 2012 through September 2019: U.S. Geological Survey Open-File Report 2021–1018, 18 p., accessed April 2022 at https://doi.org/10.3133/ofr20211018.