Appendix 30. Model Archival Summary for Fecal Coliform Bacteria Density at U.S. Geological Survey Site 06892350, Kansas River at De Soto, Kansas, during October 2013 through September 2019

This model archival summary summarizes the fecal coliform bacteria (FCB; U.S. Geological Survey [USGS] parameter code 31625) density model developed to compute 15-minute FCB densities from October 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 14, 2020

Model calibration data period: October 21, 2013, through September 24, 2019

Model application date: October 21, 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 75 concurrent measurements of FCB density, sensor-measured TBY and seasonal components (sine and cosine variables) during October 21, 2013, through September 24, 2019. Samples were collected throughout the range of continuously observed hydrologic conditions. No samples had densities below laboratory detection limits. Twenty-nine 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 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.

Fecal Coliform 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 FCB 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 FCB density to sensor-measured TBY and seasonal components (sine and cosine variables). 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 and seasonal components (sine and cosine variables) were selected as good surrogates for FCB 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 FCB density at USGS site 06892350:

FCB density-based model:

$$\log FCB = 1.28 \times \log TBY - 0.438 \times \sin(2\pi D) - 0.296 \times \cos(2\pi D) - 0.0872$$

where

log = logarithm base 10;

FCB = fecal coliform bacteria density, in colonies per 100 milliliters;

TBY = turbidity, in formazin nephelometric units;

sin = sine;

D = date, in decimal years; and

cos = cosine.

TBY makes physical and statistical sense as an explanatory variable for FCB because of its positive correlation with suspended material to which fecal indicator bacteria can physically bind. Increases in turbidity during precipitation runoff events also makes seasonality (sine and cosine variables) a logical explanatory variable for FCB.

The logarithmically (log) transformed model may be retransformed to the original units so that FCB 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:

$$FCB = 1.60 \times \left(TBY^{1.28} \times 10^{[-0.438 \times \sin(2\pi D)]} \times 10^{[-0.296 \times \cos(2\pi D)]} \times 10^{-0.0872}\right)$$

Previous Models

Start Year	End Year	Model Equation	Reference
2012	2019	logFCB = 0.877logTBY - 0.228sin(2πD) -0.59cos(2πD) - 0.481	Foster and Graham (2016)
1999	2003	logFCB = 1.53logTBY - 1.05	Rasmussen and others (2005)

Model Statistics, Data, and Plots

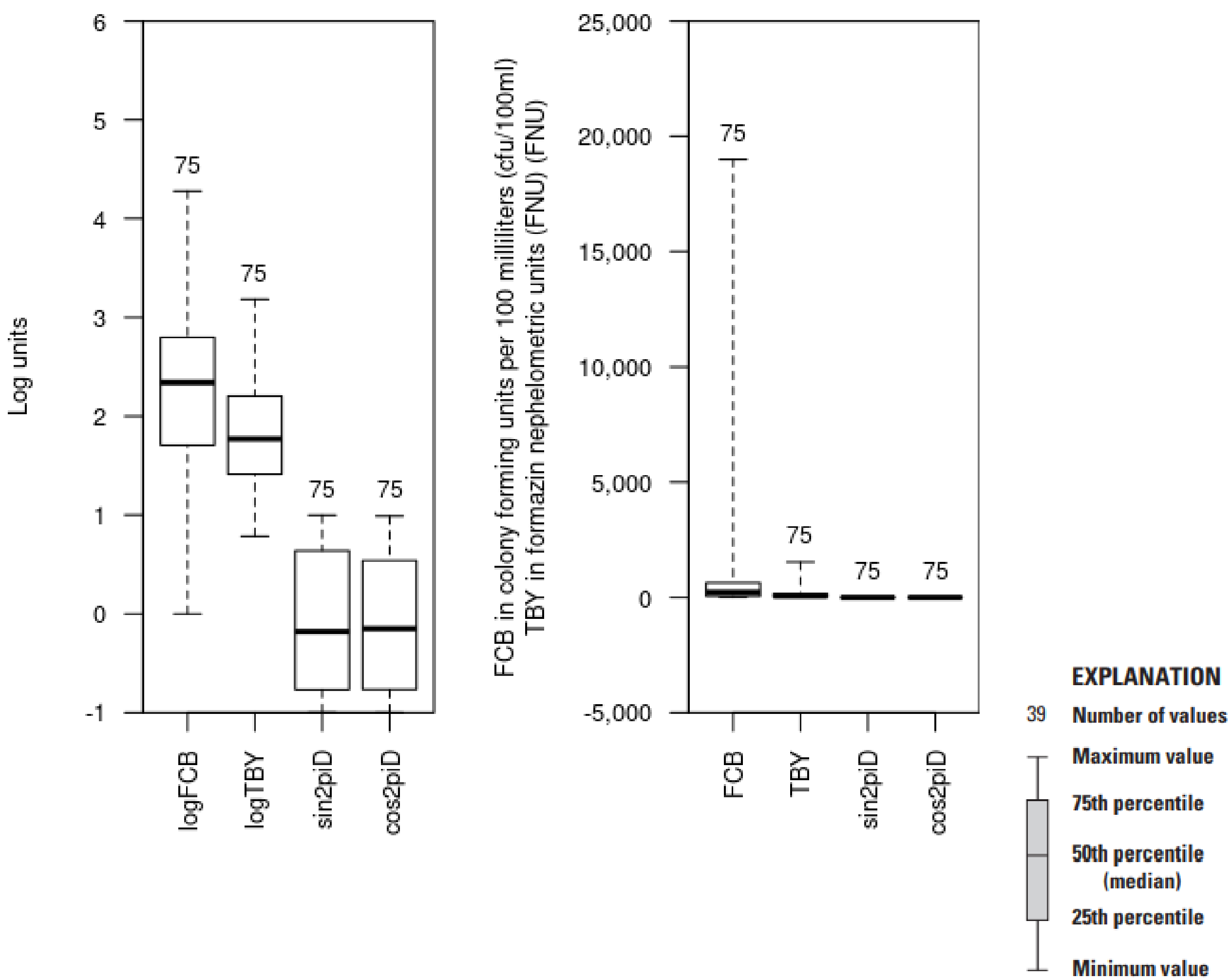
Model

logFCB = + 1.28 * logTBY - 0.438 * sin2piD - 0.296 * cos2piD - 0.0872

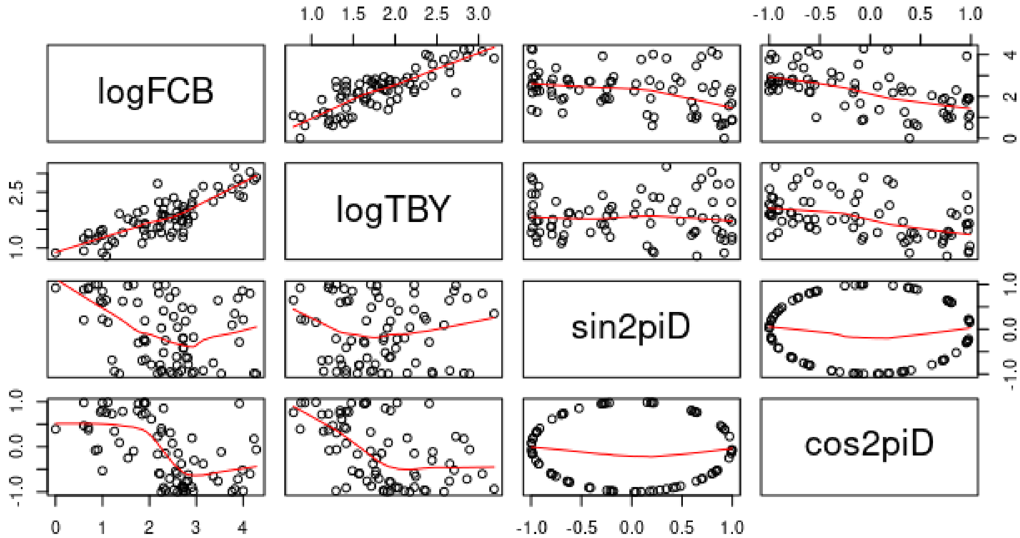
Variable Summary Statistics

	logFCB	FCB	logTBY	sin2piD	cos2piD	TBY
Minimum	0.00	1	0.785	-0.9980	-1.0000	6.1
1st Quartile	1.66	46	1.410	-0.7860	-0.7670	26.0
Median	2.34	220	1.770	-0.1800	-0.1500	59.0
Mean	2.30	1640	1.830	-0.0504	-0.0955	158.0
3rd Quartile	2.82	660	2.230	0.6420	0.6120	170.0
Maximum	4.28	19000	3.180	0.9980	0.9880	1530.0

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	75
Standard error (RMSE)	0.434
Average Model standard percentage error (MSPE)	117
Coefficient of determination (R^2)	0.814
Adjusted Coefficient of Determination (Adj. R^2)	0.806
Bias Correction Factor (BCF)	1.6

Variance Inflation Factors (VIF)

logTBY	sin2piD	cos2piD
1.31	1.00	1.31

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	-0.0872	0.1930	-0.453	6.52e-01
logTBY	1.2800	0.1040	12.300	2.87e-19
sin2piD	-0.4380	0.0709	-6.180	3.55e-08
cos2piD	-0.2960	0.0821	-3.600	5.77e-04

Correlation Matrix

	Intercept	logTBY	sin2piD	cos2piD
Intercept	1.0000	-0.9650	-0.0195	-0.4360
logTBY	-0.9650	1.0000	0.0407	0.4850
sin2piD	-0.0195	0.0407	1.0000	0.0502
cos2piD	-0.4360	0.4850	0.0502	1.0000

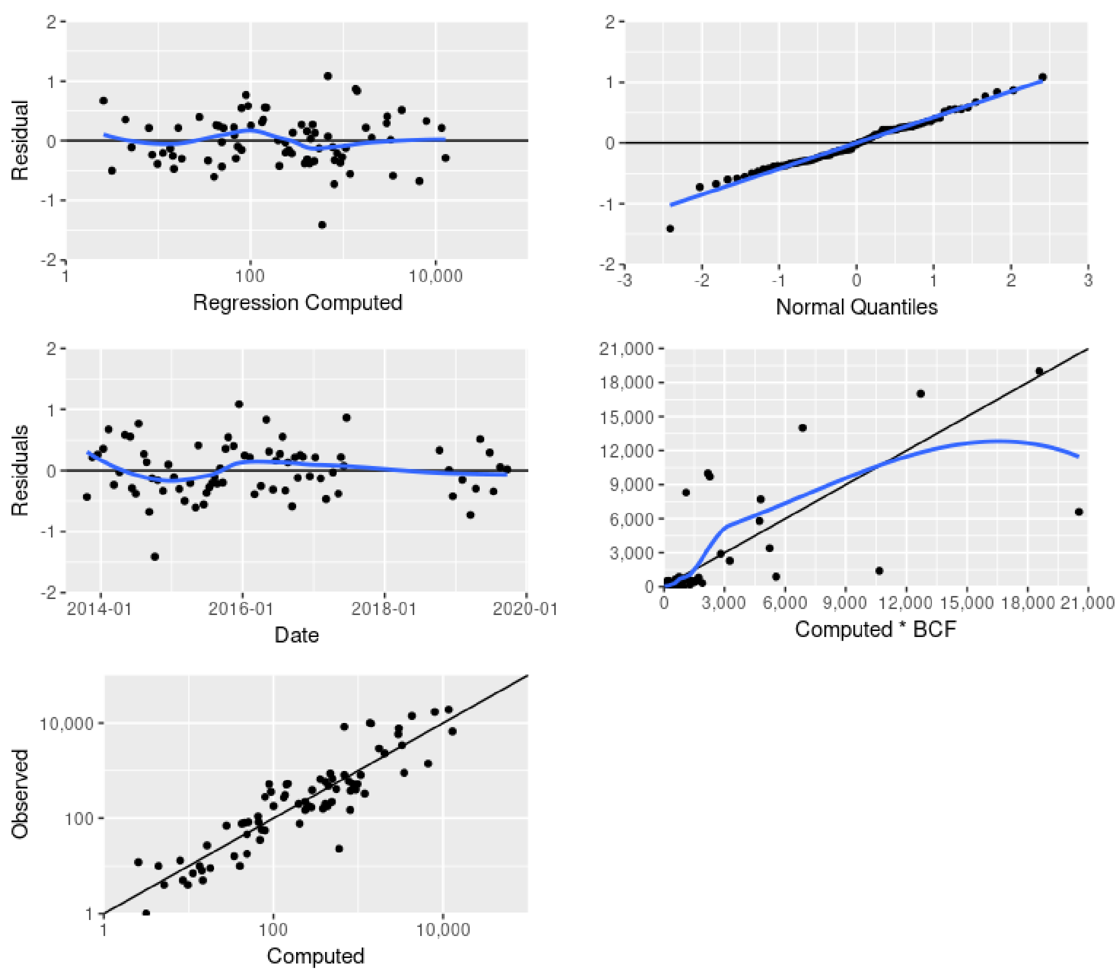
Outlier Test Criteria

Leverage	Cook's D	DFFITS
0.160	0.319	0.462

Flagged Observations

	logFCB	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
201402100740	1.08	0.407	0.672	1.61	1.63	0.0751	0.0528	0.465
201407141250	2.72	1.950	0.768	1.84	1.88	0.0795	0.0735	0.552
201410061330	1.36	2.770	-1.410	-3.32	-3.59	0.0388	0.1110	-0.721
201512141510	3.92	2.830	1.080	2.64	2.76	0.0998	0.1930	0.918
201605020930	3.99	3.150	0.835	2.00	2.04	0.0711	0.0763	0.565
201903190920	2.18	2.900	-0.727	-1.77	-1.80	0.1080	0.0956	-0.628

Statistical Plots

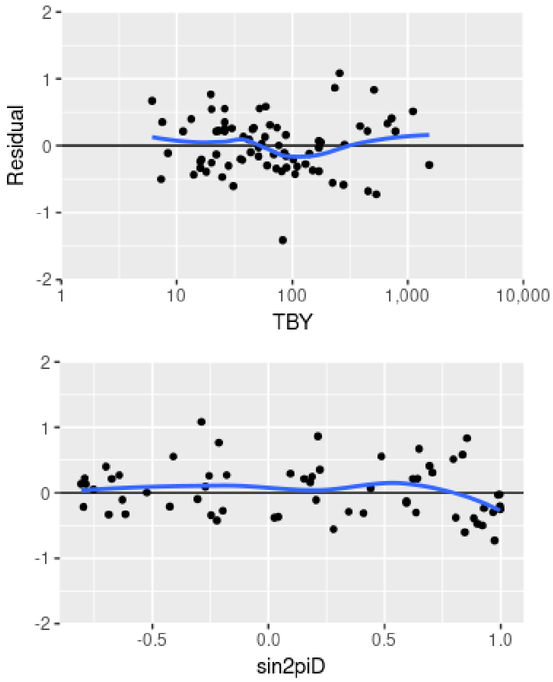


First row (left): Residual FCB related to regression computed FCB with local polynomial regression fitting, or locally estimated scatterplot smoothing (LOESS), indicated by the blue line.

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

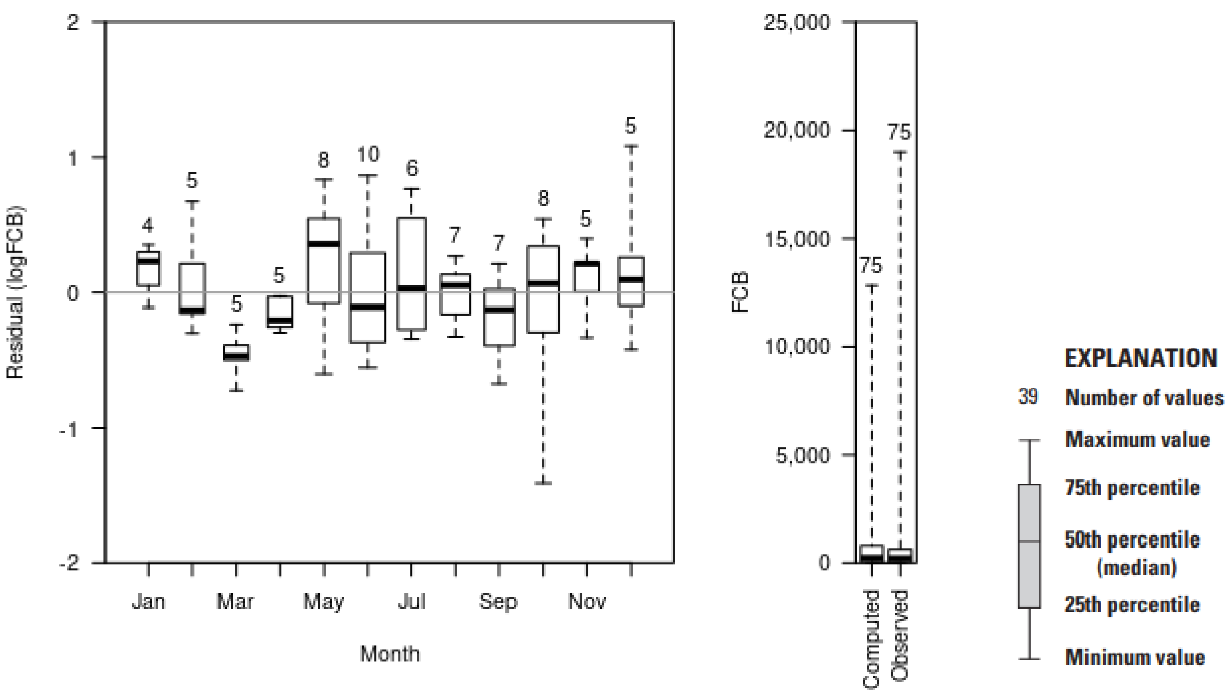
Second row: Residual FCB related to date (left) and regression computed FCB multiplied by the BCF (right) with LOESS, indicated by the blue line.

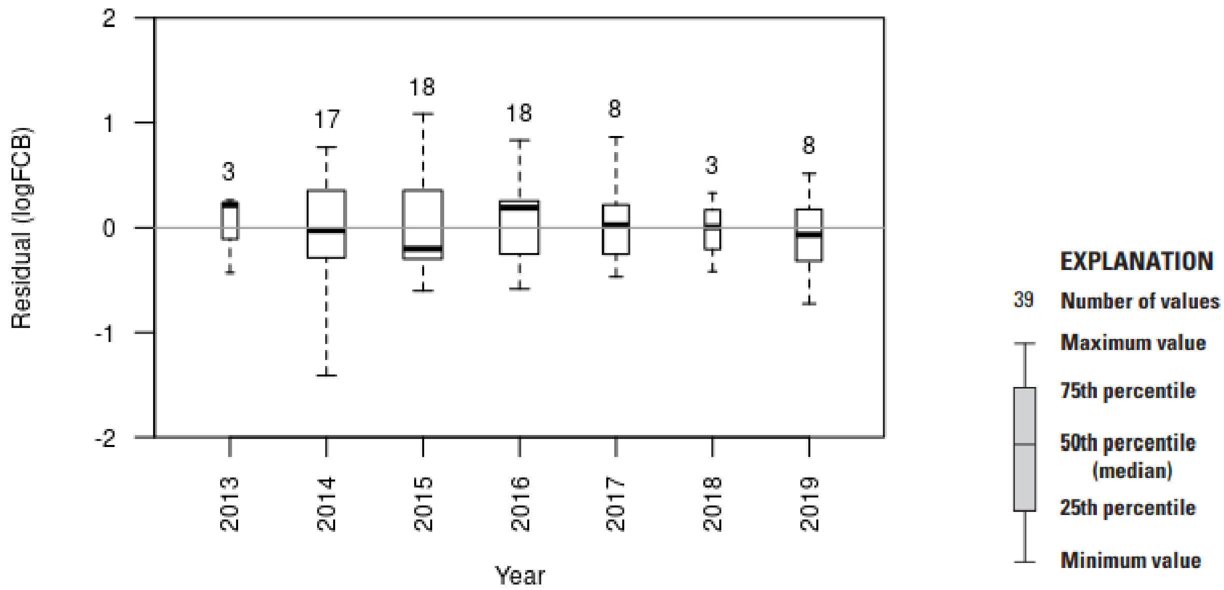
Third row: Observed FCB related to regression computed FCB.



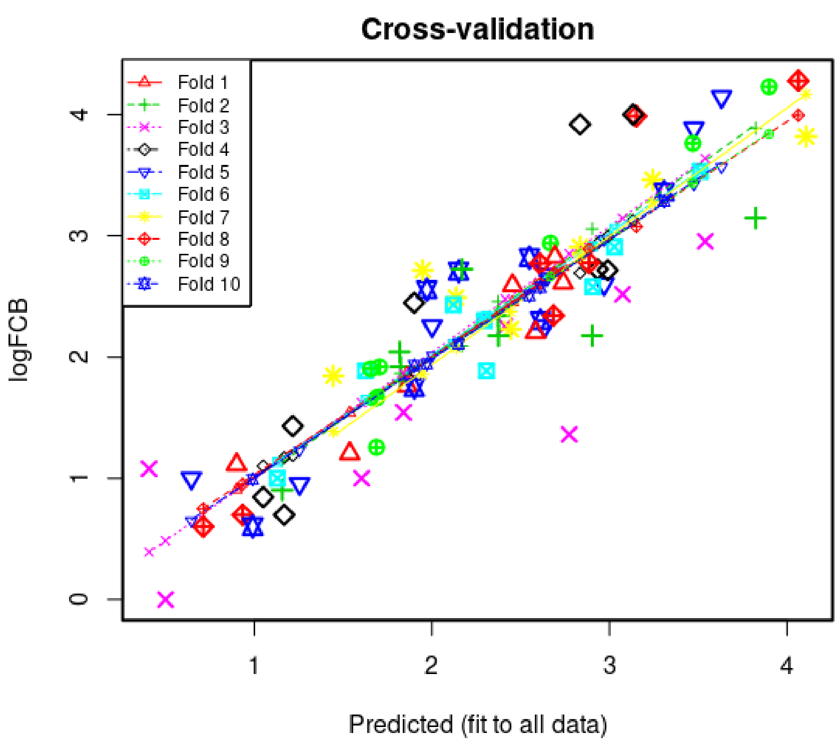
First row: Residual FCB related to TBY (left) and cos2piD (right) with LOESS, indicated by the blue line.

Second row: Residual FCB related to sin2piD (right) with LOESS, indicated by the blue line.





Cross-Validation



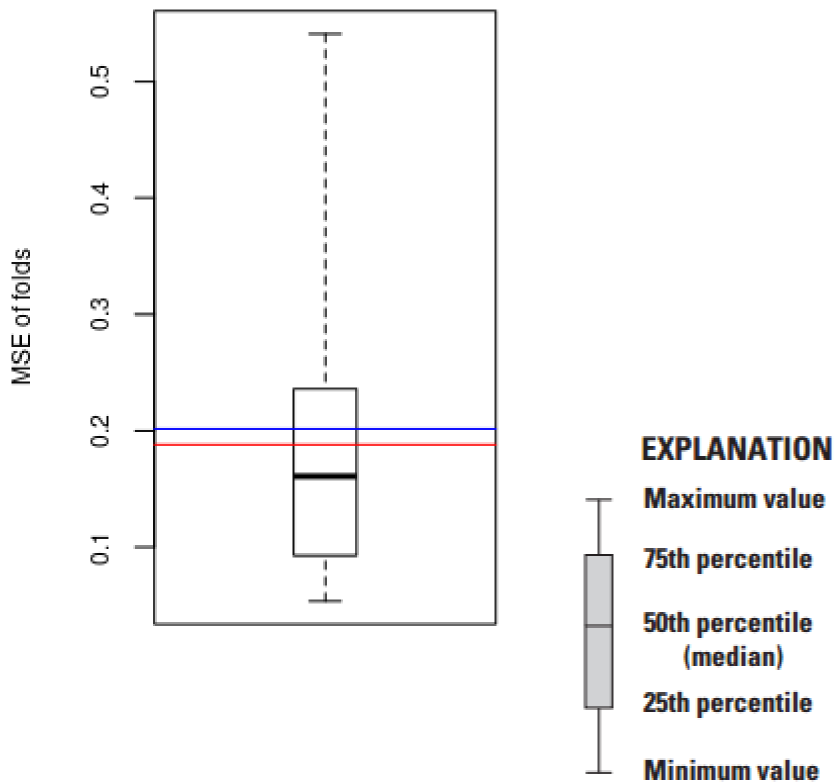

```
Fold - equal partition of the data (10 percent of the data).

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:  0.0535
                 Mean MSE of folds:  0.2020
               Median MSE of folds:  0.1610
              Maximum MSE of folds:  0.5410
 (Mean MSE of folds) / (Model MSE):  1.0700
```




```
Red line - Model MSE

Blue line - Mean MSE of folds
```

Model-Calibration Dataset

0	Date	logFCB	logTBY	sin2piD	cos2piD	FCB	TBY	Computed logFCB	Computed FCB	Residual	Normal Quantiles	Censored Values
1	2013-10-21	1.26	1.15	-0.94	0.342	18	14	1.69	78	-0.433	-1.2	--
2	2013-11-18	1.92	1.34	-0.674	0.738	83	22	1.71	81.2	0.214	0.375	--
3	2013-12-16	1.89	1.48	-0.255	0.967	77	30	1.63	67.7	0.261	0.641	--
4	2014-01-13	1	0.875	0.222	0.975	10	7.5	0.645	7.07	0.355	0.912	--
5	2014-02-10	1.08	0.785	0.649	0.761	12	6.1	0.407	4.08	0.672	1.55	--
6	2014-03-10	0.699	1.2	0.928	0.374	5	16	0.934	13.8	-0.235	-0.485	--
7	2014-04-07	1.66	1.71	0.995	-0.0988	46	51	1.69	78	-0.025	0	--
8	2014-05-05	2.56	1.77	0.836	-0.549	360	59	1.97	150	0.585	1.44	--
9	2014-06-02	2.71	1.72	0.486	-0.874	510	52	2.15	227	0.557	1.35	--
10	2014-06-11	3.82	3.18	0.346	-0.938	6600	1530	4.11	20500	-0.289	-0.601	--
11	2014-06-30	2.26	1.91	0.0258	-1	180	81	2.64	692	-0.38	-1.02	--
12	2014-07-14	2.72	1.29	-0.214	-0.977	520	19.7	1.95	142	0.768	1.67	--
13	2014-08-11	2.82	1.67	-0.642	-0.767	660	46.3	2.55	567	0.27	0.683	--
14	2014-08-25	2.59	1.58	-0.806	-0.591	390	37.7	2.45	456	0.136	0.27	--
15	2014-09-08	3.15	2.66	-0.924	-0.382	1400	452	3.82	10700	-0.677	-1.82	--
16	2014-09-22	2.61	1.84	-0.989	-0.15	410	69	2.74	880	-0.127	-0.201	--
17	2014-10-06	1.36	1.92	-0.996	0.0903	23	82.8	2.77	950	-1.41	-2.41	--
18	2014-10-20	2.26	1.71	-0.946	0.325	180	51	2.41	414	-0.157	-0.305	--
19	2014-11-17	1.2	1.2	-0.687	0.727	16	16	1.54	55.2	-0.333	-0.815	--
20	2014-12-15	1.92	1.63	-0.272	0.962	83	42.2	1.82	107	0.0952	0.201	--
21	2015-01-12	0.602	0.923	0.205	0.979	4	8.37	0.712	8.25	-0.11	-0.134	--
22	2015-02-09	0.954	1.45	0.635	0.772	9	28	1.25	28.8	-0.301	-0.683	--
23	2015-03-09	0	0.865	0.921	0.39	1	7.33	0.5	5.06	-0.5	-1.35	--
24	2015-04-06	0.845	1.21	0.997	-0.0817	7	16.3	1.05	18	-0.205	-0.375	--

25	2015-05-04	1	1.49	0.845	-0.534	10	30.9	1.6	64.3	-0.604	-1.67	--
26	2015-05-18	3.89	2.86	0.693	-0.721	7700	724	3.48	4790	0.41	1.08	--
27	2015-06-15	2.52	2.35	0.28	-0.96	330	223	3.07	1900	-0.556	-1.44	--
28	2015-06-29	2.6	2.18	0.043	-0.999	400	150	2.97	1490	-0.368	-0.912	--
29	2015-07-13	2.72	2.11	-0.197	-0.98	520	130	2.99	1560	-0.274	-0.561	--
30	2015-07-27	2.72	2.01	-0.425	-0.905	530	102	2.93	1370	-0.209	-0.411	--
31	2015-08-10	2.78	1.93	-0.629	-0.778	600	85.3	2.89	1230	-0.108	-0.1	--
32	2015-08-24	2.23	1.57	-0.796	-0.605	170	37	2.44	445	-0.214	-0.448	--
33	2015-09-08	2.68	1.74	-0.924	-0.382	480	54.3	2.65	711	0.0336	0.1	--
34	2015-09-21	2.18	1.55	-0.986	-0.167	150	35.5	2.37	380	-0.199	-0.34	--
35	2015-10-05	2.49	1.41	-0.997	0.0731	310	26	2.14	219	0.355	0.964	--
36	2015-10-19	2.45	1.3	-0.951	0.309	280	20	1.9	127	0.547	1.2	--
37	2015-11-16	1.85	1.12	-0.699	0.715	70	13.3	1.45	44.6	0.4	1.02	--
38	2015-12-14	3.92	2.41	-0.288	0.957	8300	257	2.83	1090	1.08	2.41	--
39	2016-01-11	1.9	1.66	0.188	0.982	80	45.3	1.66	72.5	0.247	0.561	--
40	2016-02-08	1.43	1.41	0.622	0.783	27	26	1.22	26.4	0.215	0.448	--
41	2016-03-03	0.602	1.26	0.884	0.467	4	18	0.991	15.7	-0.389	-1.08	--
42	2016-04-04	0.903	1.3	0.998	-0.0645	8	20	1.16	23	-0.254	-0.523	--
43	2016-05-02	3.99	2.71	0.854	-0.52	9700	510	3.15	2270	0.835	1.82	--
44	2016-05-16	2.43	1.81	0.706	-0.709	270	64	2.12	212	0.31	0.815	--
45	2016-06-06	2.3	2.04	0.409	-0.912	200	110	2.61	655	-0.311	-0.726	--
46	2016-06-20	2.77	1.94	0.18	-0.984	590	88	2.61	652	0.161	0.305	--
47	2016-07-11	2.94	1.87	-0.18	-0.984	870	73.7	2.67	746	0.271	0.726	--
48	2016-07-25	2.72	1.41	-0.409	-0.912	530	26	2.17	237	0.554	1.28	--
49	2016-08-08	2.58	1.95	-0.615	-0.788	380	89	2.91	1290	-0.327	-0.77	--
50	2016-08-22	2.83	1.76	-0.786	-0.619	670	58	2.69	790	0.133	0.235	--
51	2016-09-12	2.95	2.44	-0.954	-0.301	900	277	3.54	5550	-0.586	-1.55	--
52	2016-09-26	4.28	2.89	-0.998	-0.0645	19000	780	4.06	18600	0.214	0.411	--
53	2016-10-11	2.91	2.15	-0.981	0.192	810	140	3.03	1710	-0.119	-0.167	--
54	2016-10-24	2.26	1.41	-0.914	0.405	180	26	2	161	0.254	0.601	--
55	2016-11-07	2.04	1.36	-0.791	0.612	110	23	1.82	105	0.223	0.523	--
56	2016-12-12	1.76	1.64	-0.305	0.952	58	43.7	1.86	116	-0.0968	-0.0667	--
57	2017-01-09	1.11	1.05	0.154	0.988	13	11.3	0.9	12.7	0.214	0.34	--
58	2017-02-06	1	1.34	0.595	0.804	10	22	1.13	21.6	-0.13	-0.235	--
59	2017-03-06	0.699	1.39	0.9	0.437	5	24.7	1.17	23.6	-0.469	-1.28	--
60	2017-04-10	2.34	2.23	0.989	-0.15	220	170	2.37	379	-0.0314	-0.0333	--
61	2017-05-08	2.2	2.23	0.806	-0.591	160	170	2.58	615	-0.38	-0.964	--
62	2017-05-22	3.46	2.65	0.642	-0.767	2900	447	3.24	2810	0.218	0.485	--
63	2017-06-05	2.91	2.23	0.441	-0.898	810	170	2.84	1100	0.0732	0.167	--
64	2017-06-19	4	2.37	0.214	-0.977	10000	233	3.13	2180	0.866	2.03	--
65	2018-10-11	4.23	2.82	-0.984	0.176	17000	666	3.9	12700	0.331	0.863	--
66	2018-11-29	2.3	1.88	-0.523	0.852	200	76.4	2.3	316	0.00522	0.0333	--
67	2018-12-18	1.89	2.02	-0.222	0.975	77	106	2.31	325	-0.422	-1.14	--
68	2019-02-06	1.75	1.95	0.595	0.804	56	88.3	1.9	127	-0.153	-0.27	--
69	2019-03-19	2.18	2.73	0.974	0.226	150	533	2.9	1280	-0.727	-2.03	--
70	2019-04-16	1.54	1.78	0.968	-0.251	35	60.6	1.84	111	-0.296	-0.641	--
71	2019-05-09	4.15	3.04	0.796	-0.605	14000	1110	3.63	6860	0.514	1.14	--
72	2019-06-26	3.76	2.59	0.0945	-0.996	5800	385	3.47	4730	0.293	0.77	--
73	2019-07-16	2.34	1.86	-0.247	-0.969	220	72.6	2.69	776	-0.343	-0.863	--
74	2019-08-20	3.36	2.25	-0.753	-0.658	2300	177	3.31	3260	0.0533	0.134	--
75	2019-09-24	3.53	2.45	-0.993	-0.116	3400	282	3.51	5220	0.018	0.0667	--

Definitions

D: Date, in decimal years.

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

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

E.vars: Explanatory variables.

FCB: Fecal coliforms, in colonies per 100 milliliters (31625).

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.]