Appendix 6. Model Archival Summary for Chloride 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 chloride (Cl; U.S. Geological Survey [USGS] parameter code 00940) concentration model developed to compute 15-minute Cl 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.

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 (SC), dissolved oxygen, pH, turbidity, 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, SC, dissolved oxygen, pH, turbidity, 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 Cl concentration and sensor-measured SC 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.

Potential outliers initially were identified using scatterplots of the Cl and SC 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.

Chloride Sampling Details

During November 2018 through February 2019, samples were collected using the equal-width increment collection method (U.S. Geological Survey, variously dated). In March 2019, sample collection location changed to the southern bank of the Kansas River above Topeka Weir using the single-vertical collection method (U.S. Geological Survey, variously dated) to avoid safety risks caused by a nearby low-head dam. All samples were composited for analysis (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. A FISP US DH–81, DH–95, D–95, or D–96a depth integrating sampler was used. Samples were analyzed for Cl concentration at the USGS National Water Quality Laboratory in Lakewood, Colorado.

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 Cl concentration to sensor-measured SC. 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).

SC was selected as a good surrogate for Cl 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 Cl concentration at USGS site 06888990:

Cl concentration-based model:

$$\log Cl = 1.83(\log SC) - 3.48$$

where

log = logarithm base 10,

Cl = chloride concentration, in milligrams per liter, and

SC = specific conductance, in microsiemens per centimeter at 25 degrees Celsius.

SC makes physical and statistical sense as an explanatory variable for Cl because of its positive correlation with charged ionic species (Hem, 1985).

The logarithmically (log) transformed model may be retransformed to the original units so that Cl 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 1.01. The retransformed model, accounting for BCF is as follows:

$$Cl = 1.01 \times (SC^{1.83} \times 10^{-3.48})$$

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

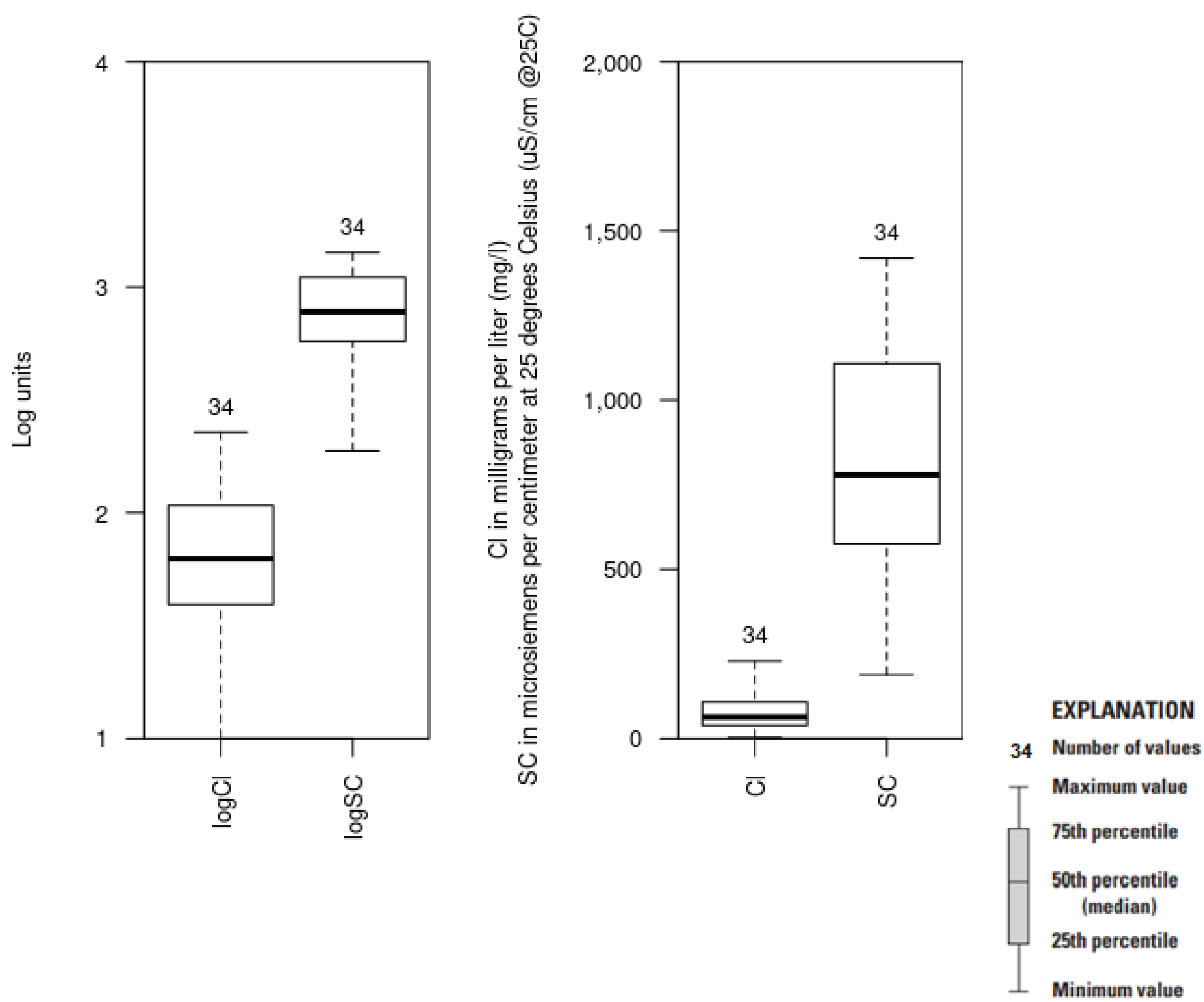
Model

log*Cl* = 1.83(log*SC*) - 3.48

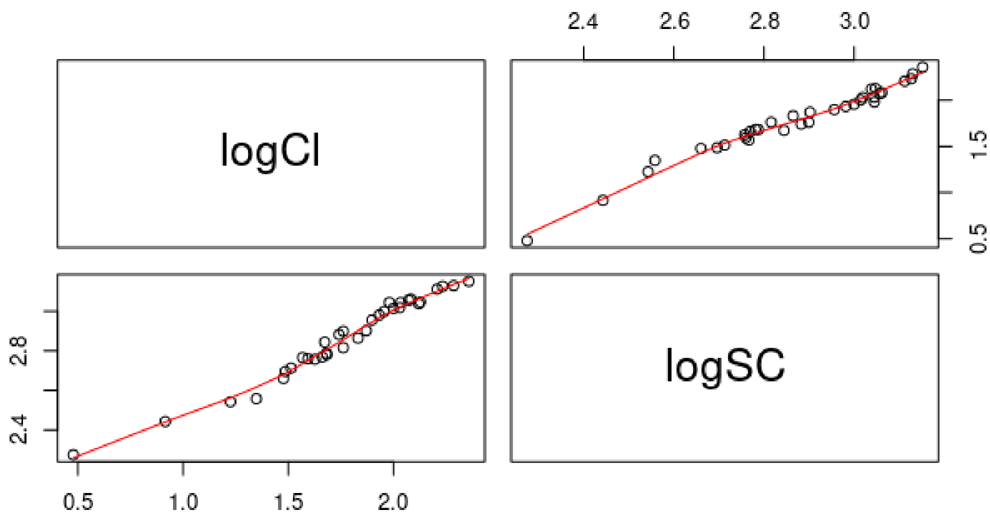
Variable Summary Statistics

	logCl	Cl	logSC	SC
Minimum	0.479	3.01	2.27	188
1st Quartile	1.590	39.2	2.76	575
Median	1.800	62.8	2.89	778
Mean	1.770	79.0	2.87	814
3rd Quartile	2.030	108	3.04	1,110
Maximum	2.360	228	3.15	1,420

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)	0.0706
Mean model standard percentage error (MSPE)	16.3
Coefficient of determination (R^2)	0.968
Adjusted coefficient of determination (Adj. R^2)	0.967
Bias correction factor (BCF)	1.01

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	-3.48	0.1680	-20.6	4.67e-20
logSC	1.83	0.0586	31.2	1.52e-25

Correlation Matrix

	Intercept	E.vars
Intercept	1.000	-0.997
E.vars	-0.997	1.000

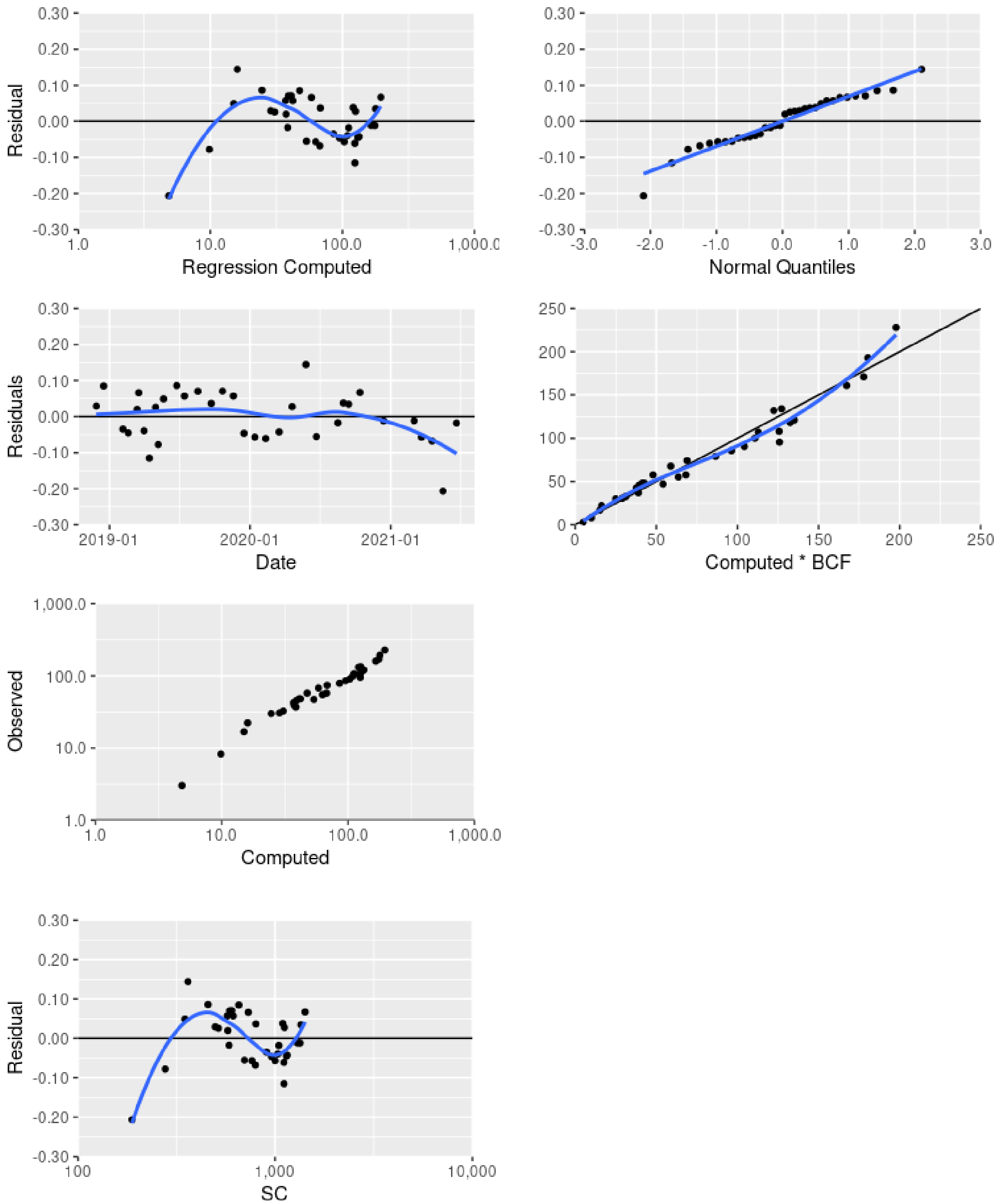
Outlier Test Criteria

Leverage	Cook's D	DFFITS
0.176	0.194	0.485

Flagged Observations

	logCl	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
201905080920	0.915	0.993	-0.0778	-1.20	-1.21	0.1540	0.131	-0.515
202005260810	1.350	1.200	0.1440	2.15	2.29	0.0955	0.244	0.743
202105170800	0.479	0.685	-0.2060	-3.43	-4.25	0.2720	2.200	-2.600

Statistical Plots



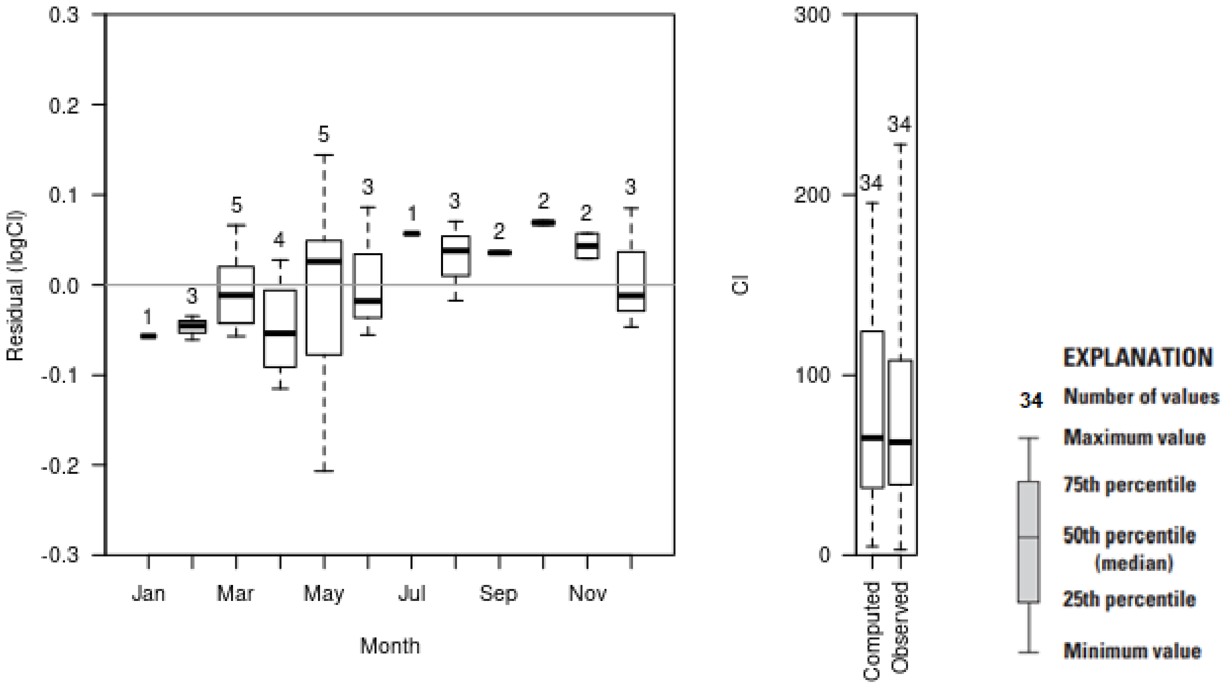
First row (left): Residual Cl related to regression computed Cl with local polynomial regression fitting, or locally estimated scatterplot smoothing (LOESS), indicated by the blue line.

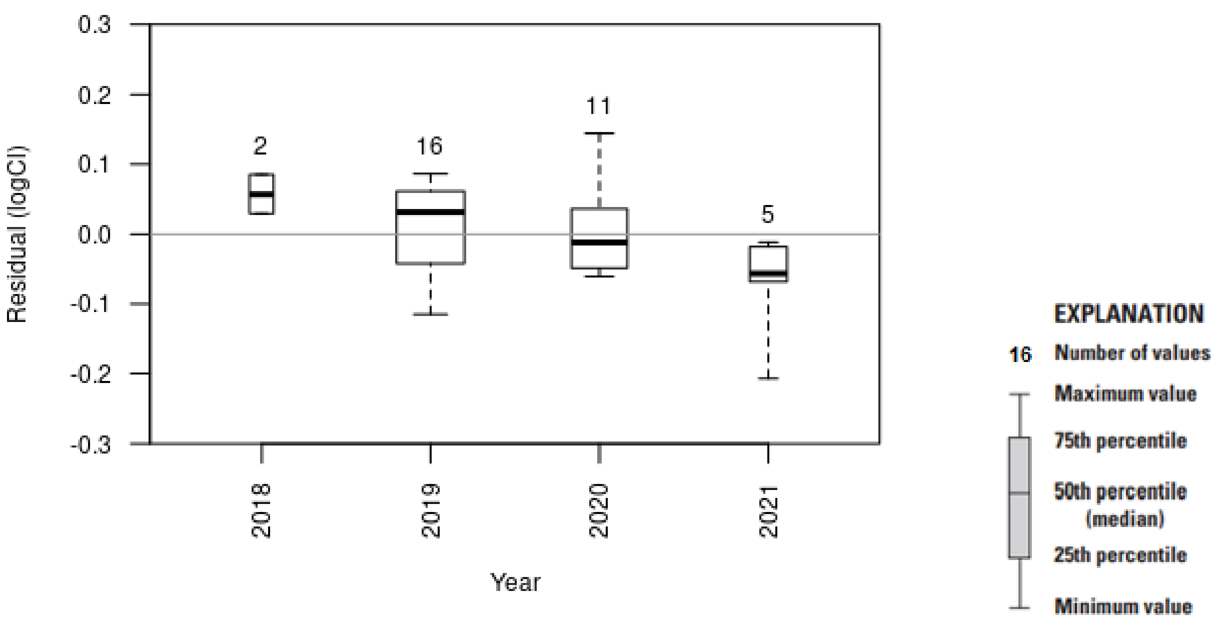
First row (right): Residual Cl related to the corresponding normal quantile of the residual with simple linear regression, indicated by the blue line.

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

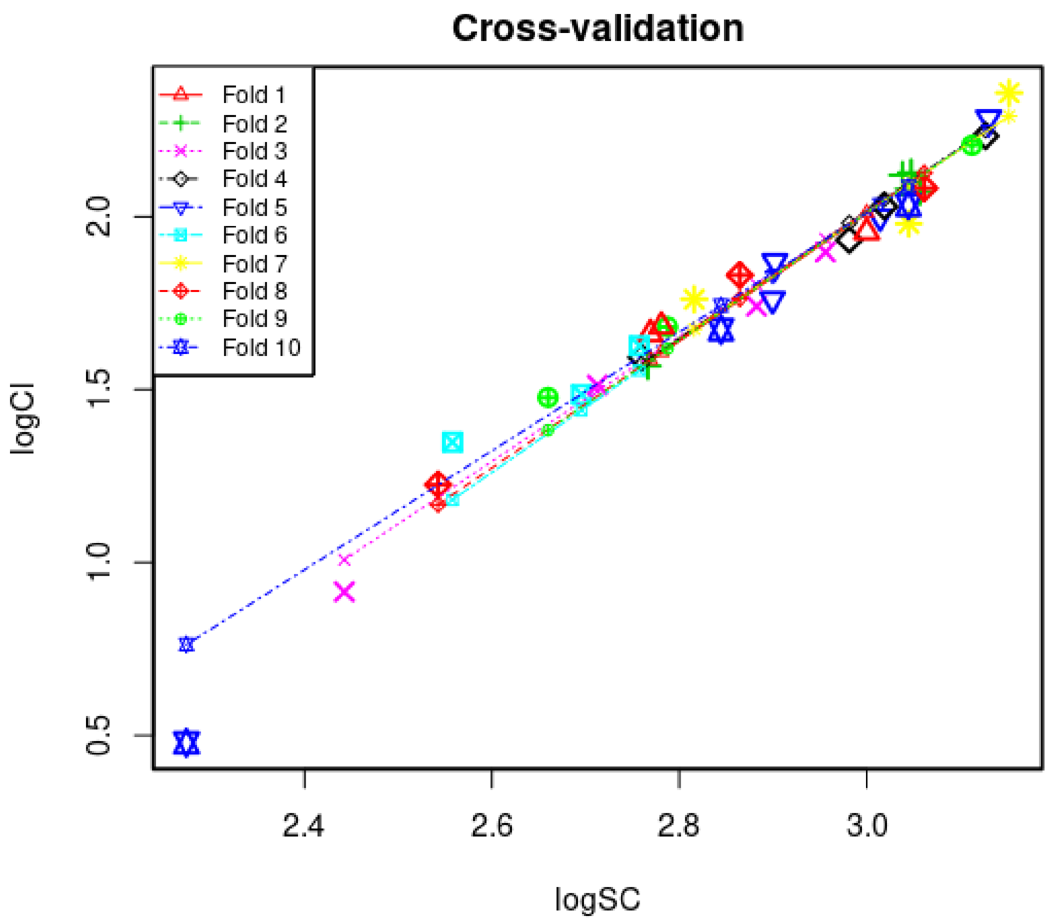
Third row: Observed Cl related to regression computed Cl.

Fourth row: Residual Cl related to SC with LOESS, indicated by the blue line.

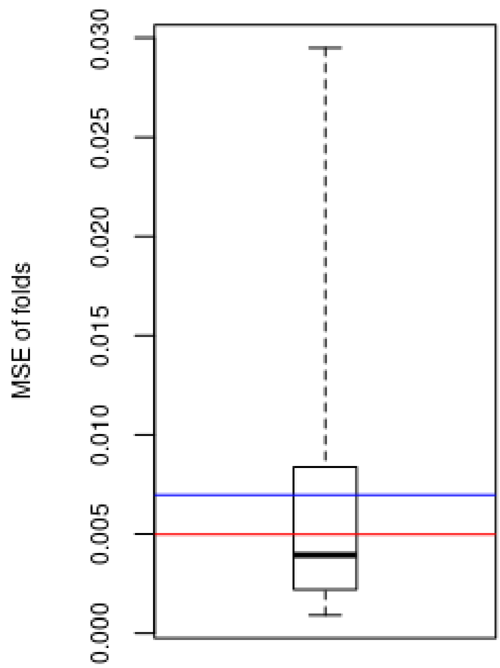

Residual (logCl)
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
Jan
Mar
May
Jul
Sep
Nov
Month
1
3
5
4
5
3
1
3
2
2
2
3
Cl
300
200
100
0
34
34
Computed
Observed
EXPLANATION
34 Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value

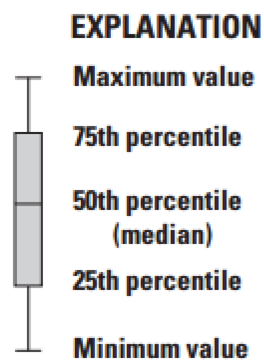

Residual (logCl)
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
2
16
11
5
2018
2019
2020
2021
Year
EXPLANATION
16 Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value

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:  0.000892
                 Mean MSE of folds:  0.006960
               Median MSE of folds:  0.003920
              Maximum MSE of folds:  0.029500
 (Mean MSE of folds) / (Model MSE):  1.400000
```






```
Red line - Model MSE

Blue line - Mean MSE of folds
```

Model-Calibration Dataset

0	Date	logCl	logSC	Cl	SC	Computed logCl	Computed Cl	Residual	Normal Quantiles	Censored Values
1	2018-11-28	1.49	2.7	30.6	496	1.46	29	0.0293	0.259	--
2	2018-12-17	1.76	2.82	57.7	655	1.68	48	0.0849	1.43	--
3	2019-02-05	1.9	2.96	79	904	1.93	86.6	-0.0348	-0.336	--
4	2019-02-19	2.07	3.06	118	1140	2.12	133	-0.0452	-0.581	--
5	2019-03-14	1.59	2.76	39.2	575	1.57	37.9	0.02	0.0367	--
6	2019-03-18	1.83	2.86	67.8	732	1.77	58.9	0.0661	0.867	--
7	2019-04-01	2	3.01	99.9	1030	2.04	111	-0.0394	-0.415	--
8	2019-04-15	1.98	3.04	95.5	1110	2.1	126	-0.115	-1.68	--
9	2019-05-01	1.51	2.71	32.6	516	1.49	31.1	0.0259	0.11	--
10	2019-05-08	0.915	2.44	8.23	277	0.993	9.96	-0.0778	-1.43	--
11	2019-05-22	1.23	2.54	16.8	349	1.18	15.2	0.0491	0.581	--
12	2019-06-25	1.48	2.66	30	457	1.39	24.9	0.0861	1.68	--
13	2019-07-15	1.68	2.79	48	613	1.62	42.6	0.057	0.67	--
14	2019-08-19	1.66	2.77	45.8	588	1.59	39.4	0.0703	1.11	--
15	2019-09-23	1.87	2.9	74.2	798	1.83	69	0.0367	0.415	--
16	2019-10-22	1.68	2.78	48.2	604	1.61	41.5	0.0707	1.25	--
17	2019-11-19	1.63	2.76	42.3	572	1.57	37.5	0.0572	0.765	--
18	2019-12-17	1.93	2.98	85.6	958	1.98	96.4	-0.0464	-0.67	--
19	2020-01-14	1.96	3	90.4	1000	2.01	104	-0.0568	-0.867	--
20	2020-02-11	2.03	3.04	108	1110	2.09	126	-0.061	-1.11	--

21	2020-03-17	2.08	3.06	121	1150	2.13	135	-0.0426	-0.496	--
22	2020-04-20	2.13	3.05	134	1120	2.1	127	0.0276	0.184	--
23	2020-05-26	1.35	2.56	22.3	361	1.2	16.2	0.144	2.11	--
24	2020-06-22	1.67	2.84	47.1	699	1.73	54.2	-0.0554	-0.765	--
25	2020-08-17	1.57	2.77	37	584	1.59	39	-0.0174	-0.184	--
26	2020-08-31	2.12	3.04	132	1090	2.08	122	0.0379	0.496	--
27	2020-09-14	2.29	3.13	193	1350	2.25	180	0.0346	0.336	--
28	2020-10-13	2.36	3.15	228	1420	2.29	198	0.0669	0.979	--
29	2020-12-14	2.23	3.13	171	1340	2.24	178	-0.0118	-0.11	--
30	2021-03-03	2.21	3.11	161	1290	2.22	167	-0.0115	-0.0367	--
31	2021-03-22	1.74	2.88	55.1	763	1.8	63.6	-0.0569	-0.979	--
32	2021-04-19	1.76	2.9	57.7	793	1.83	68.3	-0.0678	-1.25	--
33	2021-05-17	0.479	2.27	3.01	188	0.685	4.9	-0.206	-2.11	--
34	2021-06-21	2.03	3.02	107	1040	2.05	113	-0.018	-0.259	--

Definitions

Cl: Chloride, in milligrams per liter (USGS parameter code 00940).

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

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

E.vars: Explanatory variables.

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).

SC: Specific conductance, in microsiemens per centimeter at 25 degrees Celsius (USGS parameter code 00095).

t value: Student’s t value; the coefficient divided by its associated standard error (Helsel and others, 2020).

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.]

Hem, J.D., 1985, Study and interpretation of the chemical characteristics of natural water (3d ed.): U.S. Geological Survey Water-Supply Paper 2254, 264 p. [Also available at https://doi.org/10.3133/wsp2254.]

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.