Appendix 11. Model Archival Summary for Chlorophyll *a* 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 chlorophyll *a* (Chla; U.S. Geological Survey [USGS] parameter code 70953) concentration model developed to compute 15-minute Chla 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, and chlorophyll (fCHL) 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 10, 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, turbidity, fCHL 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 Chla concentration and sensor-measured fCHL 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. Two sample concentrations were qualified as “estimated.”

Potential outliers initially were identified using scatterplots of the Chla and fCHL 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.

Chlorophyll *a* 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 Chla 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 Chla concentration to sensor-measured fCHL. 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). A possible model limitation of overestimation of Chla should be considered when interpreting model computations in the upper range based on irregular scatter in the plot of observed and regression computed Chla. Additional model-calibration data could improve this limitation in the future.

fCHL was selected as a good surrogate for Chla 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 Chla concentration at USGS site 06888990:

Chla concentration-based model:

$$\log Chla = 1.26(\log fCHL) + 0.687$$

where

log = logarithm base 10,

Chla = chlorophyll *a* concentration, in micrograms per liter, and

fCHL = chlorophyll fluorescence, in relative fluorescence units.

Although there is an unknown level of uncertainty inherent to fluorescence sensors (because of nonphotochemical quenching, matrix effects, and variable fluorescence responses of differing plankton communities; Foster and others, 2022), fCHL makes physical and statistical sense as an explanatory variable for Chla because chlorophyll *a* pigments fluoresce when irradiated by certain wavelengths of light emitted from the fCHL sensor.

The logarithmically (log) transformed model may be retransformed to the original units so that Chla 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 1.10. The retransformed model, accounting for BCF is as follows:

$$Chla = 1.10 \times (fCHL^{1.26} \times 10^{0.687})$$

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

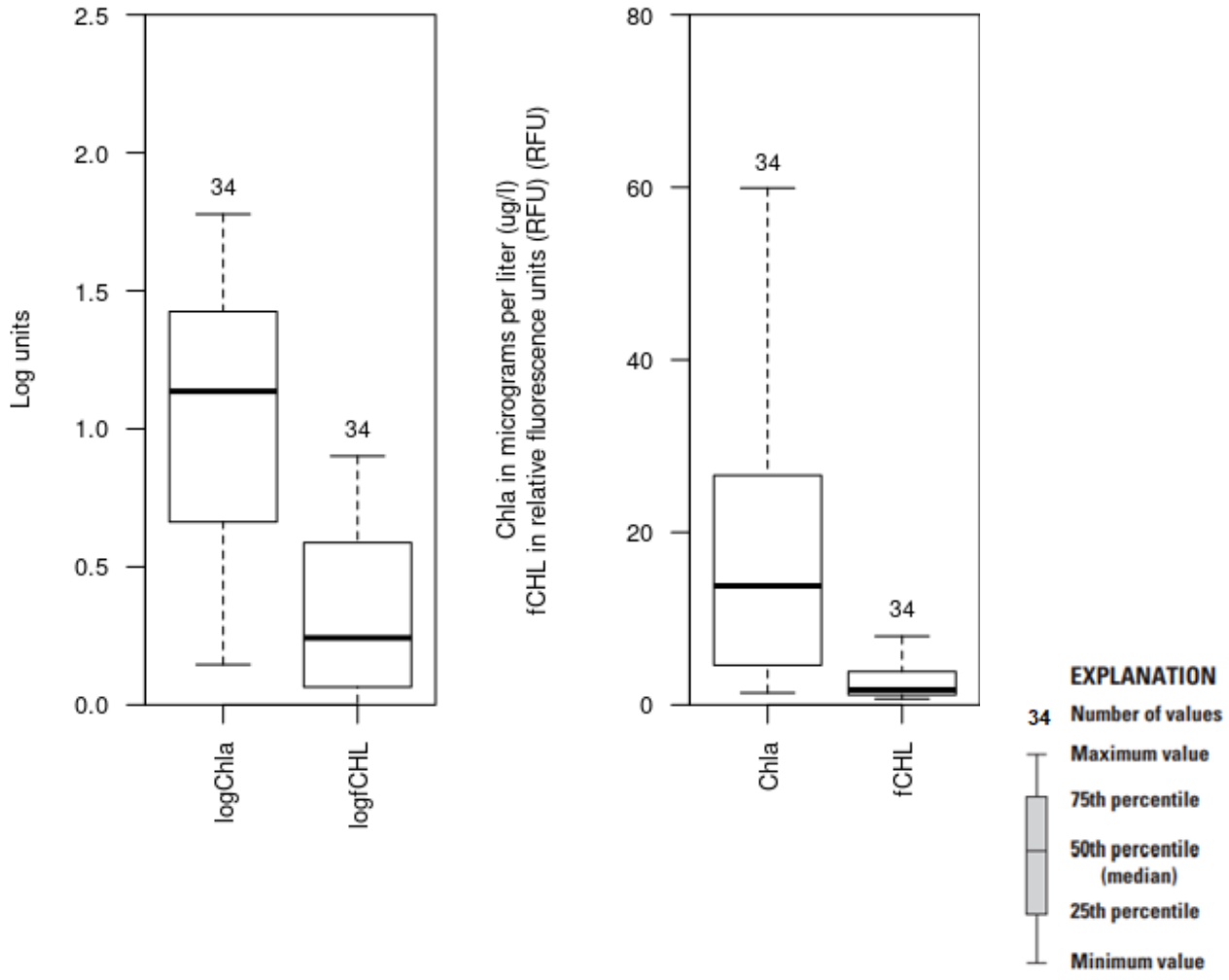
Model

$\log Chla = 1.26(\log fCHL) + 0.687$

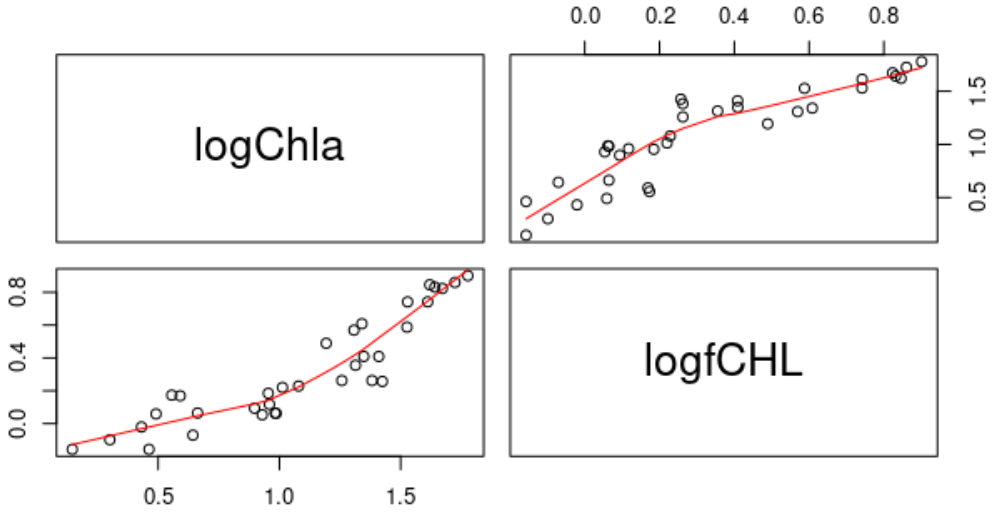
Variable Summary Statistics

	logChla	Chla	logfCHL	fCHL
Minimum	0.146	1.40	-0.1570	0.697
1st Quartile	0.663	4.6	0.0645	1.160
Median	1.140	13.8	0.2430	1.75
Mean	1.090	19.2	0.3220	2.78
3rd Quartile	1.420	26.6	0.5880	3.87
Maximum	1.780	59.9	0.9010	7.96

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)	0.198
Mean model standard percentage error (MSPE)	47.1
Coefficient of determination (R^2)	0.815
Adjusted coefficient of determination (Adj. R^2)	0.809
Bias correction factor (BCF)	1.1

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	0.687	0.0482	14.3	2.04e-15
logfCHL	1.260	0.1060	11.9	2.92e-13

Correlation Matrix

	Intercept	E.vars
Intercept	1.00	-0.71
E.vars	-0.71	1.00

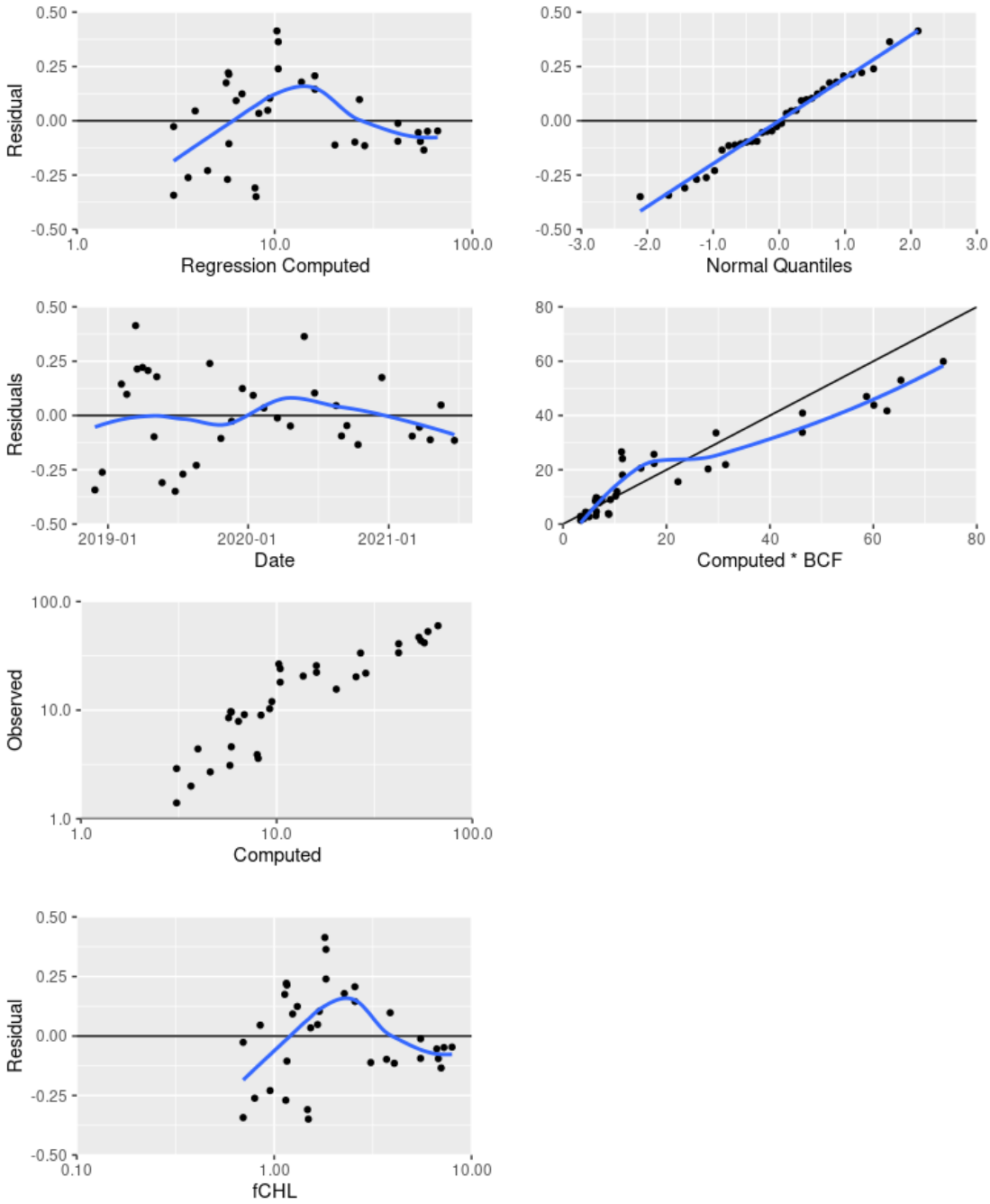
Outlier Test Criteria

Leverage	Cook's D	DFFITS
0.176	0.194	0.485

Flagged Observations

	logChla	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
201811281050	0.146	0.489	-0.343	-1.82	-1.9	0.0957	0.176	-0.616

Statistical Plots



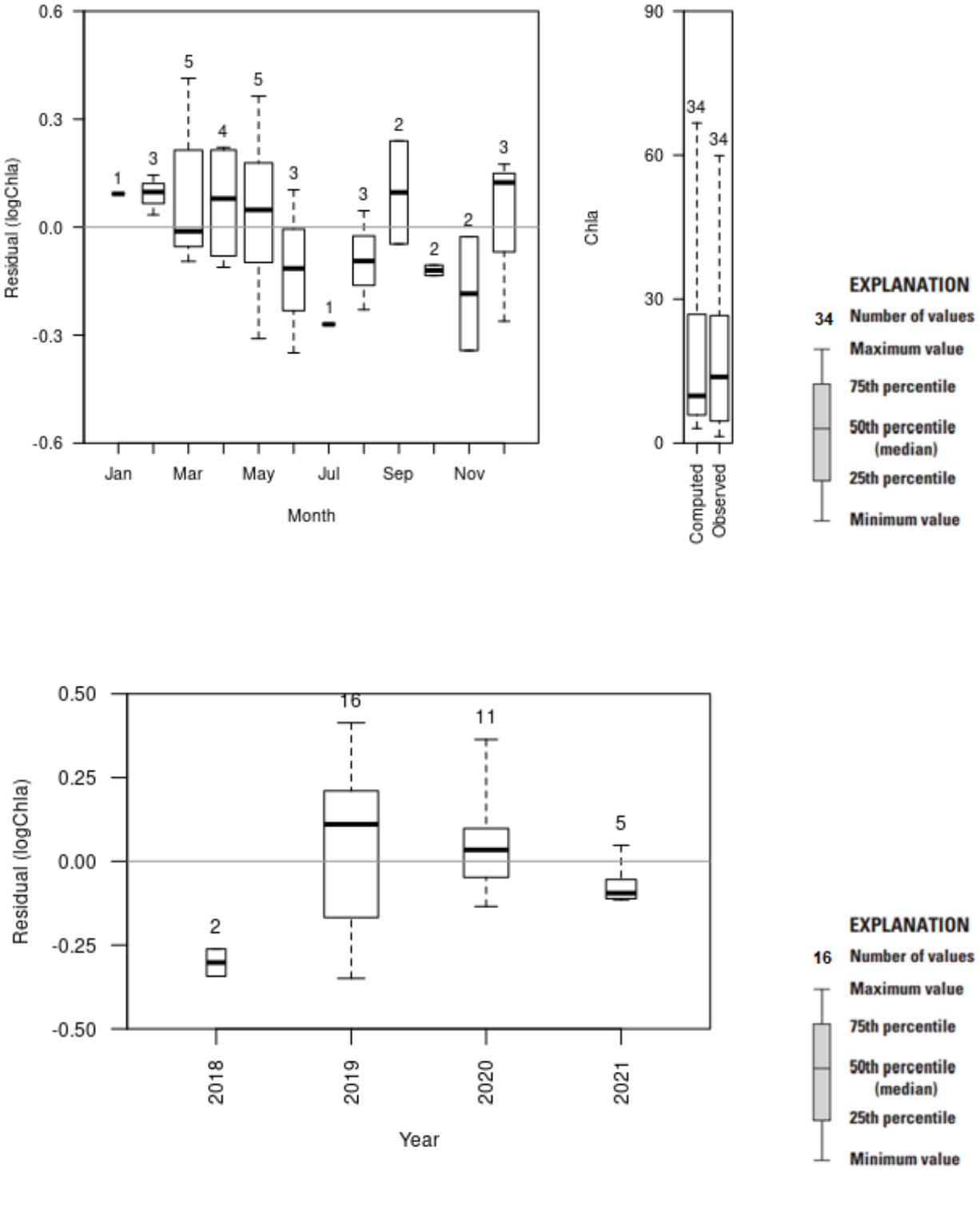
First row (left): Residual Chla related to regression computed Chla with local polynomial regression fitting, or locally estimated scatterplot smoothing (LOESS), indicated by the blue line.

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

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

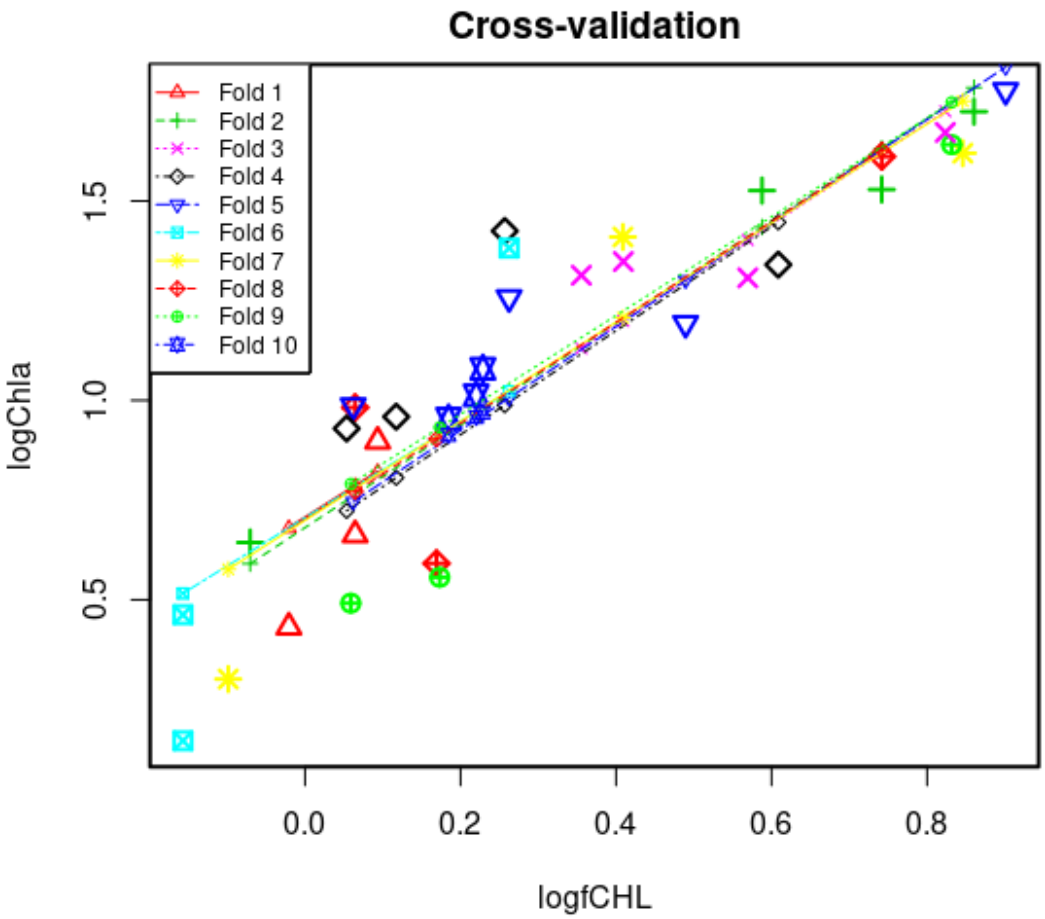
Third row: Observed Chla related to regression computed Chla.

Fourth row: Residual Chla related to fCHL with LOESS, indicated by the blue line.


Residual (logChla)
0.6
0.3
0.0
-0.3
-0.6
1
3
5
4
5
3
1
3
2
2
2
3
Jan
Mar
May
Jul
Sep
Nov
Month
Chla
90
60
30
0
34
34
Computed
Observed
EXPLANATION
34 Number of values
Maximum value
75th percentile
50th percentile (median)
25th percentile
Minimum value

Residual (logChla)
0.50
0.25
0.00
-0.25
-0.50
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



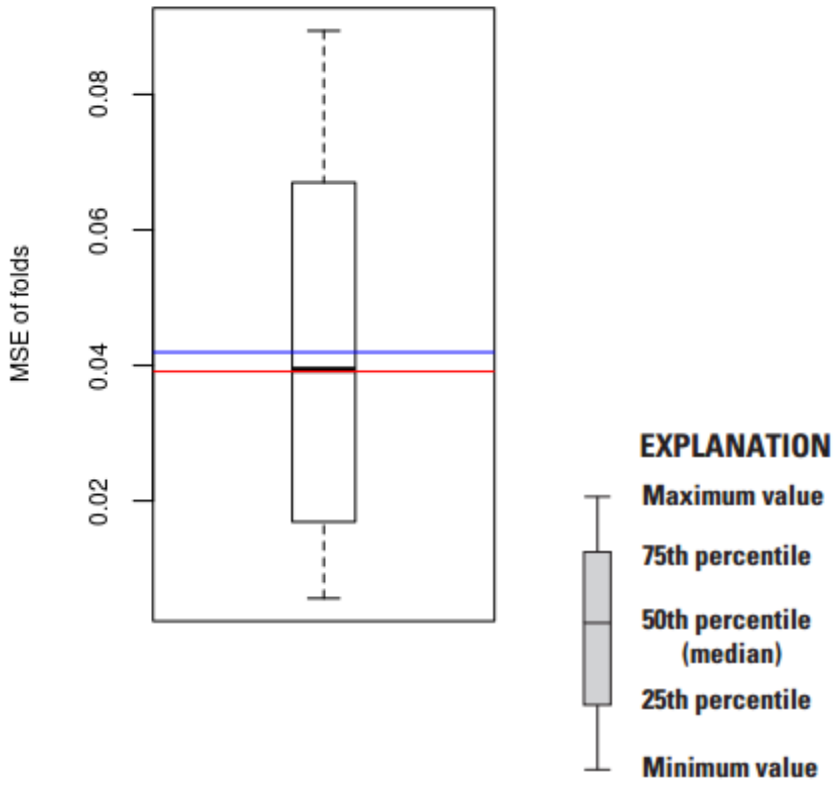

```
Fold - equal partition of the data (10 percent of the data).

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:  0.0056
                 Mean MSE of folds:  0.0419
               Median MSE of folds:  0.0393
              Maximum MSE of folds:  0.0894
 (Mean MSE of folds) / (Model MSE):  1.0700
```




```
Red line - Model MSE

Blue line - Mean MSE of folds
```

Model-Calibration Dataset

	Date	logChla	logfCHL	Chla	fCHL	Computed logChla	Computed Chla	Residual	Normal Quantiles	Censored Values
1	2018-11-28	0.146	-0.157	E1.4	0.697	0.489	3.4	-0.343	-1.68	--
2	2018-12-17	0.301	-0.0987	E2	0.797	0.562	4.02	-0.261	-1.11	--
3	2019-02-05	1.35	0.409	22.3	2.57	1.2	17.6	0.145	0.67	--
4	2019-02-19	1.53	0.588	33.6	3.87	1.43	29.6	0.0977	0.415	--
5	2019-03-14	1.42	0.257	26.6	1.81	1.01	11.3	0.414	2.11	--
6	2019-03-18	0.982	0.0645	9.6	1.16	0.768	6.47	0.214	1.11	--
7	2019-04-01	0.987	0.062	9.7	1.15	0.765	6.42	0.222	1.25	--
8	2019-04-15	1.41	0.409	25.7	2.56	1.2	17.6	0.207	0.979	--
9	2019-05-01	1.31	0.569	20.3	3.71	1.41	28	-0.0981	-0.496	--
10	2019-05-08	1.31	0.355	20.6	2.27	1.14	15.1	0.178	0.867	--
11	2019-05-22	0.591	0.169	3.9	1.48	0.9	8.76	-0.309	-1.43	--
12	2019-06-25	0.556	0.173	3.6	1.49	0.906	8.87	-0.349	-2.11	--
13	2019-07-15	0.491	0.0589	3.1	1.15	0.761	6.36	-0.27	-1.25	--
14	2019-08-19	0.431	-0.0208	2.7	0.953	0.661	5.05	-0.229	-0.979	--
15	2019-09-23	1.26	0.262	18.1	1.83	1.02	11.5	0.239	1.43	--
16	2019-10-22	0.663	0.0645	4.6	1.16	0.768	6.47	-0.106	-0.581	--
17	2019-11-19	0.462	-0.157	2.9	0.697	0.489	3.4	-0.0265	-0.0367	--
18	2019-12-17	0.959	0.117	9.1	1.31	0.835	7.54	0.124	0.581	--
19	2020-01-14	0.898	0.0934	7.9	1.24	0.805	7.03	0.0927	0.336	--
20	2020-02-11	0.954	0.185	9	1.53	0.92	9.17	0.0342	0.11	--

21	2020-03-17	1.61	0.742	40.9	5.52	1.62	46.3	-0.0116	0.0367	--
22	2020-04-20	1.72	0.86	53	7.25	1.77	65.3	-0.0485	-0.184	--
23	2020-05-26	1.38	0.262	24.1	1.83	1.02	11.5	0.364	1.68	--
24	2020-06-22	1.08	0.229	12	1.69	0.976	10.4	0.104	0.496	--
25	2020-08-17	0.643	-0.0706	4.4	0.85	0.598	4.37	0.0455	0.184	--
26	2020-08-31	1.53	0.742	33.8	5.52	1.62	46.3	-0.0941	-0.336	--
27	2020-09-14	1.78	0.901	59.9	7.96	1.82	73.5	-0.0467	-0.11	--
28	2020-10-13	1.62	0.846	41.7	7.01	1.75	62.6	-0.134	-0.867	--
29	2020-12-14	0.929	0.0535	8.5	1.13	0.755	6.26	0.175	0.765	--
30	2021-03-03	1.64	0.832	43.8	6.79	1.74	60.1	-0.0951	-0.415	--
31	2021-03-22	1.67	0.823	47	6.66	1.73	58.7	-0.0541	-0.259	--
32	2021-04-19	1.19	0.489	15.6	3.09	1.3	22.2	-0.112	-0.67	--
33	2021-05-17	1.01	0.22	10.3	1.66	0.965	10.2	0.0481	0.259	--
34	2021-06-21	1.34	0.609	21.9	4.06	1.46	31.4	-0.115	-0.765	--

E: estimated

Definitions

Chla: Chlorophyll *a*, in micrograms per liter (USGS parameter code 70953).

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

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

E.vars: Explanatory variables.

fCHL: Chlorophyll fluorescence, in relative fluorescence units (USGS parameter code 32320).

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).

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.

Foster, G.M., Graham, J.L., Bergamaschi, B.A., Carpenter, K.D., Downing, B.D., Pellerin, B.A., Rounds, S.A., and Saraceno, J.F., 2022, Field techniques for the determination of algal pigment fluorescence in environmental waters—Principles and guidelines for instrument and sensor selection, operation, quality assurance, and data reporting: U.S. Geological Survey Techniques and Methods, book 1, chap. D10, 34 p. [Also available at https://doi.org/10.3133/tm1D10.]

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.