# Appendix 12. Model Archival Summary for Total Suspended Solids 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 total suspended solids (TSS; U.S. Geological Survey [USGS] parameter code 00530) concentration model developed to compute 15-minute TSS 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, 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 TSS concentration and sensor-measured TBY during November 28, 2018, through June 21, 2021. Samples were collected throughout the range of continuously observed hydrologic conditions. Two samples had concentrations below laboratory minimum reporting limits; therefore, a Tobit regression model was developed to compute estimates of linear regression model parameters using the absolute maximum likelihood estimation approach (Hald, 1949; Cohen, 1950; Tobin, 1958; Helsel and others, 2020).

Potential outliers initially were identified using scatterplots of the TSS and TBY model-calibration data (Rasmussen and others, 2009). Additionally, outlier test criteria, including leverage and Cook's distance (Cook's D; Cook, 1977), were used to estimate potential outlier effect on the final Tobit regression model. 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.

## Total Suspended Solids 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 TSS concentration at the USGS National Water Quality Laboratory in Lakewood, Colorado.

## Model Development

Discretely collected TSS was related to sensor-measured TBY and other continuous sensor-measured data using stepwise regression analysis in R programming language (R Core Team, 2022). 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).

Censored results (less than the minimum reporting level) made up 5.9 percent of the model-calibration dataset. Tobit regression models were developed using absolute maximum likelihood estimation methods to relate discretely collected TSS concentration to sensor-measured TBY. Tobit model parameter estimates were calculated using the *smwrQW* (*v0.7.9*) package in R programming language (R Core Team, 2022).

TBY was selected as a good surrogate for TSS based on residual plots, pseudocoefficient of determination (pseudo-$R^2$), and estimated residual standard 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 TSS concentration at USGS site 06888990:

TSS concentration-based model:

$$\log TSS = 1.06(\log TBY) + 0.205$$

where

log = logarithm base 10,

*TSS* = total suspended solids concentration, in milligrams per liter, and

*TBY* = turbidity, in formazin nephelometric units.

TBY makes physical and statistical sense as an explanatory variable for TSS because of its positive correlation with suspended material.

The logarithmically (log) transformed model may be retransformed to the original units so that TSS 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 1.06. The retransformed model, accounting for BCF is as follows:

$$TSS = 1.06 \times (TBY^{1.06} \times 10^{0.205})$$

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 and Data

## Model

$\log TSS = 1.06(\log TBY) + 0.205$

Computation method: Absolute Maximum Likelihood Estimation (AMLE)

## Variable Summary Statistics

| | TSS | TBY |
|---|---|---|
| Minimum | <15.0 | 8.23 |
| 1st Quartile | 59.0 | 25.3 |
| Median | 136 | 60.6 |
| Mean | 526 | 230 |
| 3rd Quartile | 776 | 312 |
| Maximum | 3,480 | 1,240 |

## Basic Model Statistics

| | | |
|---|---|---|
| Estimated residual standard error (unbiased) | 0.165 | |
| Number of observations | 34 | |
| Number censored | 2 | (5.90 percent) |
| Log-likelihood (model) | 10.57 | |
| Log-likelihood (intercept only) | -36.42 | |
| Chi-square | 93.98 | |
| Degrees of freedom | 1 | |
| p-value | <0.0001 | |
| Pseudo-$R^2$ | 0.943 | |
| Akaike information criterion | -15.13 | |
| Bayesian information criterion | -10.56 | |
| Bias correction factor | 1.06 | |

## Explanatory Variables

Coefficients:

| | Estimate | Std. Error | z-score | p-value |
|---|---|---|---|---|
| (Intercept) | 0.2047 | 0.0942 | 2.173 | 0.0335 |
| logTBY | 1.0610 | 0.0465 | 22.816 | 0.0000 |

## Outlier Test Criteria

| Leverage | Cook's D |
|---|---|
| 0.08824 | 0.70791 |

## Flagged Observations

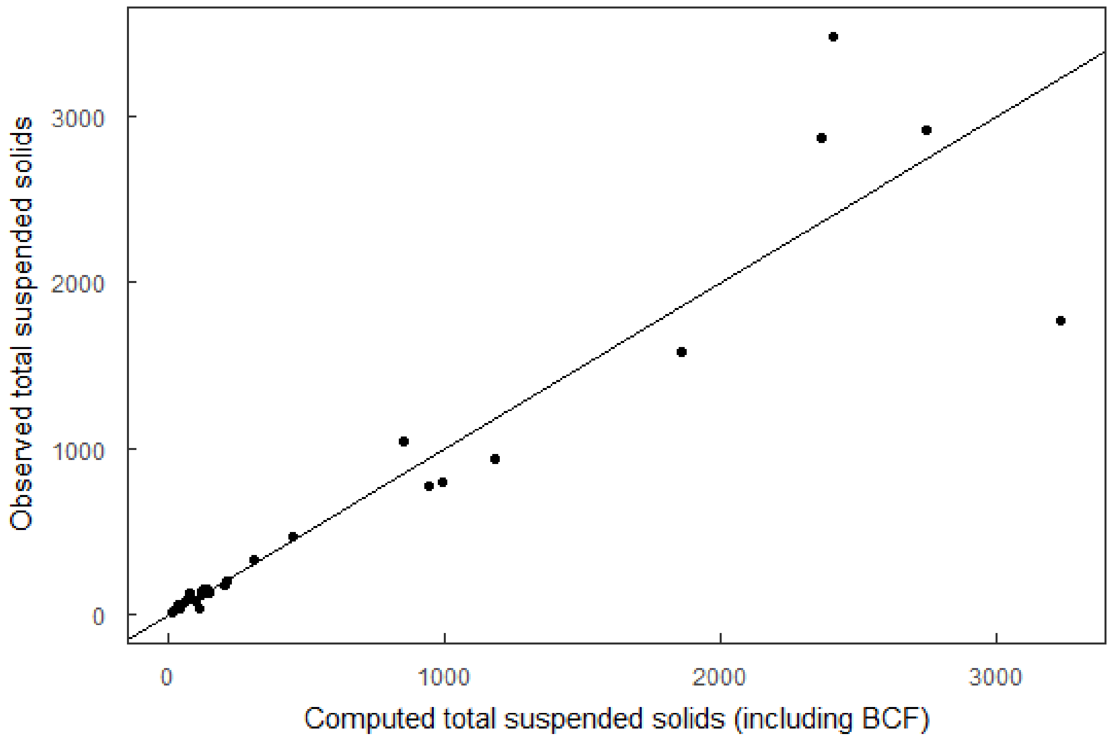
Observations exceeding at least one test criterion:

| | logTSS | ycen | yhat | resids | leverage | Cook’s D |
|---|---|---|---|---|---|---|
| 5 | 3.542 | FALSE | 3.357 | 0.1846710 | 0.11507 | 9.236e-02 |
| 10 | 3.465 | FALSE | 3.414 | 0.0511162 | 0.12416 | 7.794e-03 |
| 15 | 3.199 | FALSE | 3.245 | -0.0465298 | 0.09869 | 4.847e-03 |
| 23 | 3.458 | FALSE | 3.349 | 0.1084437 | 0.11392 | 3.145e-02 |
| 30 | 1.176 | FALSE | 1.176 | -0.0000246 | 0.10927 | 1.536e-09 |
| 33 | 3.248 | FALSE | 3.485 | -0.2374081 | 0.13606 | 1.894e-01 |

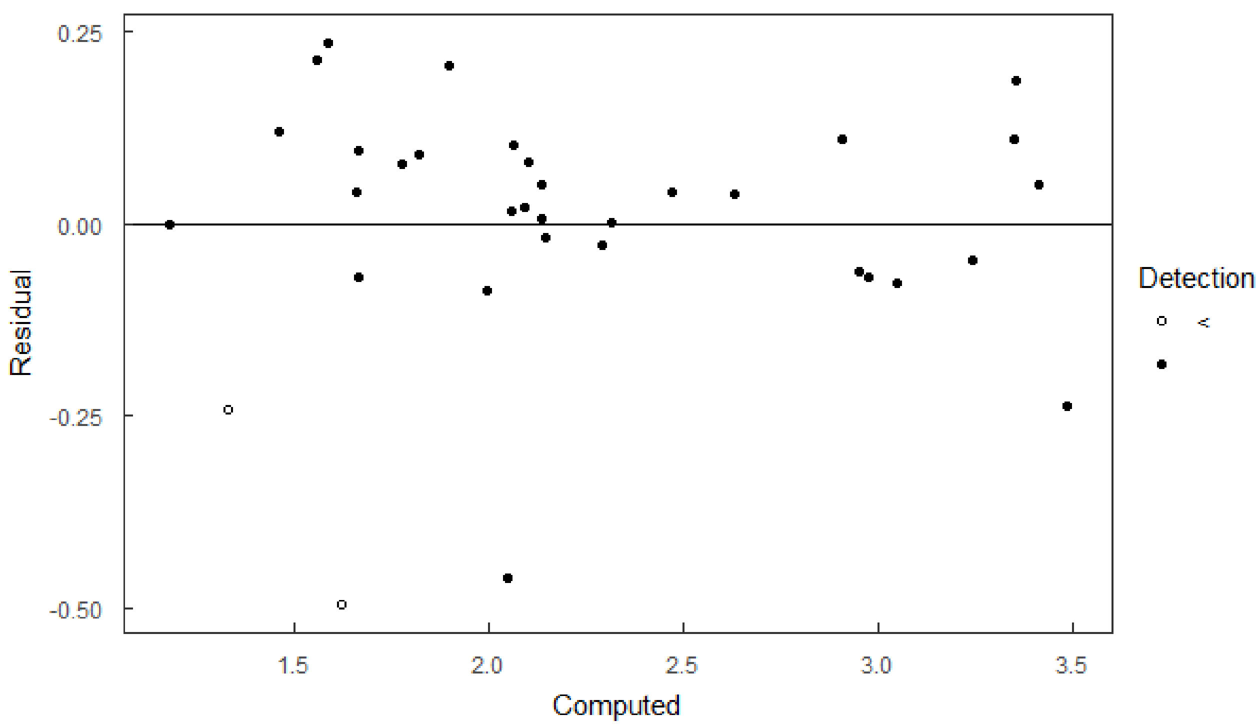
## 95 Percent Confidence Interval

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 0.02004369 | 0.3892998 |
| logTBY | 0.96985932 | 1.1521429 |

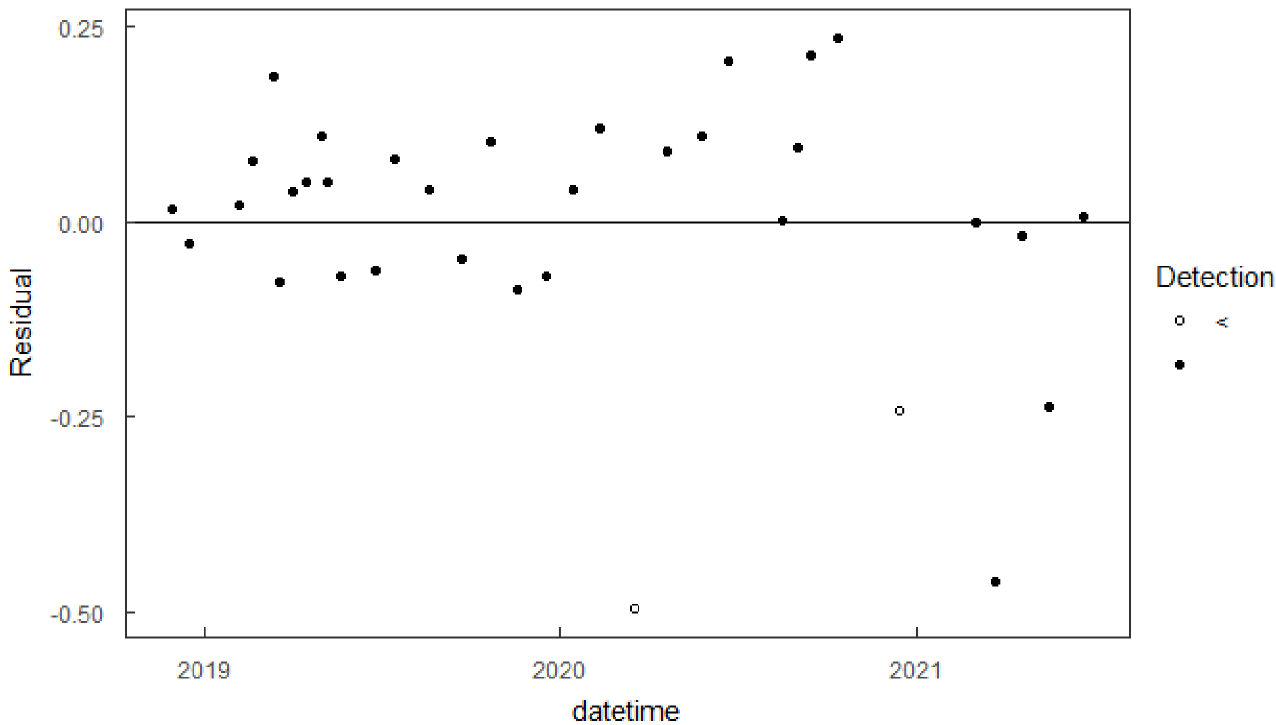
## Plots

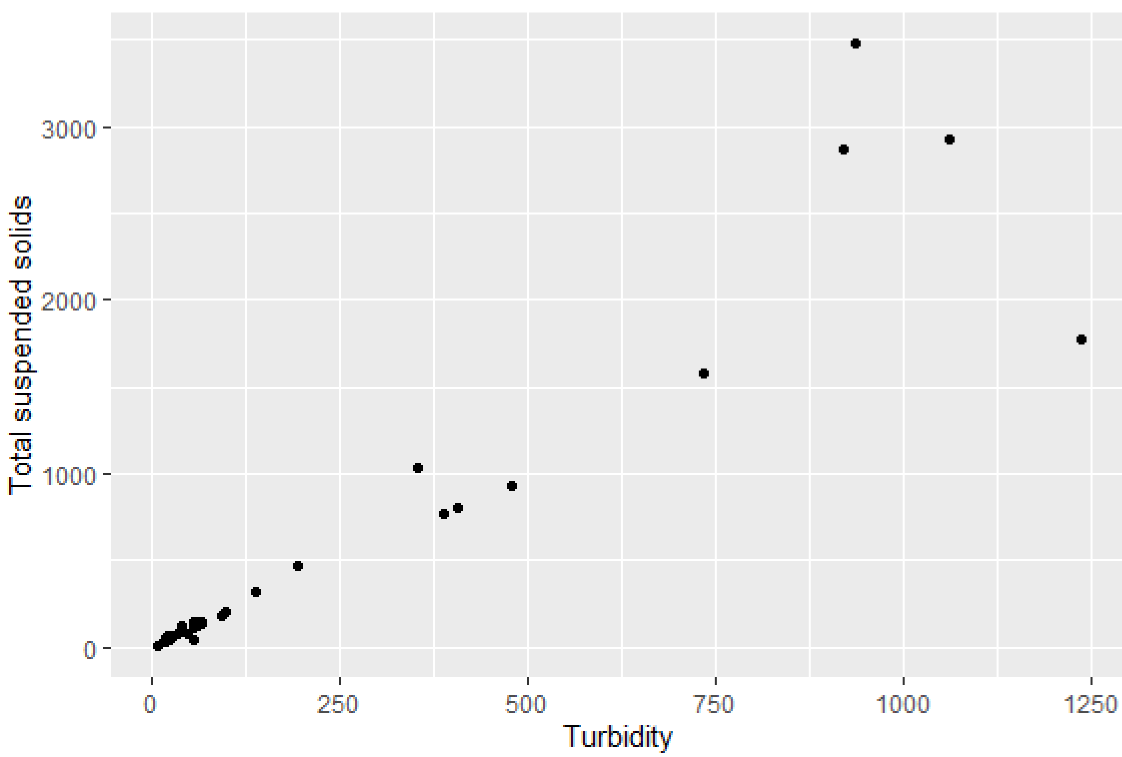


Observed total suspended solids
3000
2000
1000
0
0
1000
2000
3000
Computed total suspended solids (including BCF)

0.25
0.00
-0.25
-0.50
Residual
1.5
2.0
2.5
3.0
3.5
Computed
Detection
<




## Model-Calibration Dataset

| | datetime | logTSS | logTBY | TSS | TBY | Computed_logTSS | Computed_TSS |
|---|---|---|---|---|---|---|---|
| 2 | 2018-11-28 10:50:00 | 2.08 | 1.747 | 119 | 55.89 | 2.06 | 121.1 |
| 3 | 2018-12-17 09:40:00 | 2.26 | 1.964 | 182 | 92.02 | 2.29 | 205.6 |
| 4 | 2019-02-05 10:10:00 | 2.11 | 1.779 | 130 | 60.18 | 2.09 | 131.0 |
| 5 | 2019-02-19 08:30:00 | 1.85 | 1.479 | 71 | 30.16 | 1.77 | 63.0 |
| 6 | 2019-03-14 10:40:00 | 3.54 | 2.971 | 3480 | 935.41 | 3.36 | 2407.5 |
| 7 | 2019-03-18 10:30:00 | 2.97 | 2.681 | 935 | 479.39 | 3.05 | 1184.5 |
| 8 | 2019-04-01 07:50:00 | 2.67 | 2.286 | 467 | 193.14 | 2.63 | 451.5 |
| 9 | 2019-04-15 10:10:00 | 2.19 | 1.821 | 154 | 66.28 | 2.14 | 145.1 |
| 10 | 2019-05-01 11:00:00 | 3.02 | 2.547 | 1040 | 352.16 | 2.91 | 853.9 |

```
11 2019-05-08 09:20:00   3.47  3.025  2920 1059.41            3.41       2747.4
12 2019-05-22 09:50:00    2.9  2.610   803  407.29            2.97        996.4
13 2019-06-25 08:30:00   2.89  2.589   776  387.86            2.95        946.1
14 2019-07-15 08:10:00   2.18  1.786   151   61.04            2.10        133.0
15 2019-08-19 09:40:00   2.51  2.138   326  137.33            2.47        314.4
16 2019-09-23 08:10:00    3.2  2.866  1580  734.01            3.25       1861.4
17 2019-10-22 09:50:00   2.16  1.751   146   56.33            2.06        122.1
18 2019-11-19 08:40:00   1.91  1.687    81   48.67            1.99        104.6
19 2019-12-17 09:30:00   1.59  1.373    39   23.60            1.66         48.5
20 2020-01-14 09:20:00    1.7  1.371    50   23.50            1.66         48.3
21 2020-02-11 08:30:00   1.58  1.185    38   15.30            1.46         30.6
22 2020-03-17 08:30:00  <1.18  1.334   <15   21.60            1.62         44.2
23 2020-04-20 08:00:00   1.91  1.521    81   33.20            1.82         69.7
24 2020-05-26 08:10:00   3.46  2.964  2870  920.37            3.35       2366.5
25 2020-06-22 07:40:00    2.1  1.593   126   39.20            1.90         83.1
26 2020-08-17 08:00:00   2.31  1.987   206   97.00            2.31        217.4
27 2020-08-31 07:20:00   1.76  1.374    57   23.63            1.66         48.6
28 2020-09-14 08:10:00   1.77  1.275    59   18.85            1.56         38.2
29 2020-10-13 08:20:00   1.82  1.300    66   19.96            1.58         40.6
30 2020-12-14 08:20:00  <1.18  1.061   <15   11.50            1.33         22.6
31 2021-03-03 08:20:00   1.18  0.916    15    8.23            1.18         15.9
32 2021-03-22 08:00:00   1.59  1.739    39   54.85            2.05        118.7
33 2021-04-19 08:20:00   2.13  1.829   134   67.47            2.15        147.9
34 2021-05-17 08:00:00   3.25  3.092  1770 1236.20            3.49       3236.3
35 2021-06-21 08:20:00   2.14  1.819   138   65.87            2.13        144.2
```

## Definitions

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

**Leverage:** An outlier's measure in the x direction (Helsel and others, 2020).

**p-value:** The probability that the independent variable has no effect on the dependent variable (Helsel and others, 2020).

**Pseudo-$R^2$:** Pseudocoefficient of determination. An estimation of the proportion of varianc e in the response variable explained by the model (McKelvey and Zavoina, 1975).

**TBY:** Turbidity, in formazin nephelometric units (USGS parameter code 63680).

**TSS:** Total suspended solids, in milligrams per liter (USGS parameter code 00530).

**z-score:** The estimated coefficient divided by its associated standard error (Helsel and others, 2020).

## References Cited

Cohen, A.C., Jr., 1950, Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples: Annals of Mathematical Statistics, v. 21, no. 4, p. 557–569, accessed February 2022 at https://doi.org/10.1214/aoms/1177729751.

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.

Hald, A., 1949, Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point: Scandinavian Actuarial Journal, v. 1949, no. 1, p. 119–134. [Also available at https://doi.org/10.1080/03461238.1949.10419767.]

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.]

McKelvey, R.D., and Zavoina, W., 1975, A statistical model for the analysis of ordinal level dependent variables: The Journal of Mathematical Sociology, v. 4, no. 1, p. 103–120. [Also available at https://doi.org/10.1080/0022250X.1975.9989847.]

R Core Team, 2022, R—A language and environment for statistical computing, version 4.0.3: 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.]

Tobin, J., 1958, Estimation of relationships for limited dependent variables: Econometrica, v. 26, no. 1, p. 24–36. [Also available at https://doi.org/10.2307/1907382.]

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 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.]

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.