

Appendix 4. Model Archive Summary for Best-Fit Regression Developed to Estimate Fecal Coliform Concentration at Station 01472157; French Creek near Phoenixville, Pennsylvania

This model archive summary describes the regression model developed to estimate continuous instantaneous (15-minute) fecal coliform concentrations for the period of March 1 through October 31 each year starting in 2012.

Site and Model Information

U.S. Geological Survey (USGS) station number: 01472157

Station name: French Creek near Phoenixville, Pennsylvania

Location: Latitude 40°09'05", longitude 75°36'06", referenced to North American Datum of 1927, Chester County, PA, Hydrologic Unit 02040203, on right bank 70 (feet) ft downstream from two-span county bridge on French Creek Road, 4.5 mile (mi) northwest of Phoenixville, and 7.3 mi upstream from mouth.

Equipment: A Yellow Spring Instrument (YSI) 6920V2 monitor equipped with sensors for temperature and an Optical Monitoring System 600 series (YSI 6136) turbidity sensor. The YSI 6136 turbidity sensor has been in operation since November 7, 2006. Water temperature data has been collected for compensation of the turbidity sensor but not for official record and should be considered provisional for period of record through September 2016. The monitor is housed in a 4-inch perforated plastic pipe placed in the stream about 5 ft upstream of gage orifice and about 5 ft from the right bank. Readings from the sensors has been recorded every 15-minutes since 2012 and transmitted hourly by way of satellite.

Date regression model was created: October 2016.

Period of data for model calibration: March 1 – October 31, years 2012-15.

Model application date: October 2016 onward

Computed by: Lisa Senior, October 2016.

Reviewed by: Kirk White (Supervisory Hydrologist), Matt Gyves (Hydrologic Technician) April 6, 2017

Approved by: Joseph Duris (Pennsylvania Water Science Center Water Quality Specialist) May 26, 2017

Model Calibration Dataset

All data were collected using USGS protocols and are stored in the National Water Information System (NWIS) database. Linear regression models were developed using the TIBCO Spotfire S+ 8.1 program and open-source software package "R." Explanatory variables selected as inputs to linear regression were: water temperature, turbidity, and streamflow. Seasonal components (sine and cosine terms calculated using Julian day as a fraction of the year) were also evaluated as explanatory variables in the models to determine if seasonal changes affected the model. All combinations of physicochemical properties and seasonality components were evaluated to determine which combinations produced the best models.

The final regression model is based on 37 concurrent measurements of fecal coliform and turbidity concentrations from March 1 – October 31 of each year for years 2012-15, plus computed seasonal variables. Fecal coliform concentrations were determined from analysis of discrete samples, and turbidity concentrations were determined from continuous record of 15-minute values, interpolated when necessary to correspond with collection time of the discrete sample for bacteria analysis. Samples were collected through a range of hydrologic conditions during the March-October sampling period each year. Studentized residuals for final

model were inspected and considered for potential removal as outliers if residual values were greater than 3 or less than -3; however, no samples met these criteria, and no samples were removed from the dataset.

Fecal Coliform Data

Discrete grab samples for bacteria analysis were collected from midpoint of the stream near the gaging station and chilled until processed in the laboratory at the USGS office in Exton, Pa. within 6 hours of sample collection. The number of fecal coliform colonies in a sample was determined by membrane filtration using a 0.7 micron filter and subsequent plating and incubation using standard methods. At the laboratory, a range of dilutions was plated for each stream sample to obtain optimal counts (20-60 colonies) on at least one plate.

Model Development

Regression analysis was done using S+ and R by examining turbidity ($Turb$), streamflow (Q), and water temperature (parameter 00010 or $Temp$), in addition to computed seasonality terms ($\sin 2\pi JD$ and $\cos 2\pi JD$) as explanatory variables for estimating fecal coliform (FC) concentration. A variety of linear regression models that predict FC and $\log_{10}(FC)$ were evaluated. The distribution of residuals was examined for normality, and plots of residuals (the difference between the measured and computed values) as compared to computed FC were examined for homoscedasticity (meaning that their departures from zero did not change substantially over the range of computed values). This comparison led to the conclusion that the most appropriate and reliable model would be one that estimated $\log_{10}(FC)$.

$\log_{10}(Turb)$ and $\sin 2\pi JD$ explanatory variables were selected as the best predictors of $\log_{10}(FC)$ based on residual plots, relatively high adjusted coefficient of determination (adjusted R^2), and relatively low model residual standard error (or root mean square error, $RMSE$) and low standard percentage error ($MSPE$).

Model Summary

Final regression model for fecal coliform (FC) concentration at site number 01472157,
 FC concentration-based model:

$$\log_{10}(FC) = 1.25 \times \log_{10}(Turb) - 0.543 \times \sin(2\pi JD/365) + 1.40$$

where

FC = fecal coliform in colony-forming units per 100 milliliter (cfu/mL) (parameter 31625);

$Turb$ = turbidity in formazin nephelometric units (FNU) (parameter 63680);

$Temp$ = water temperature (parameter 00010); and,

\sin = sine function used to compute seasonality variables; and

JD = Julian day (day of year).

$Turb$ and seasonality terms make physical and statistical sense as explanatory variables for FC because previous studies showed bacteria concentrations were related to turbidity concentrations and were highest in late summer months when stream temperatures are highest. The negative coefficients for seasonal variable $\sin 2\pi JD$ [computed as $\sin(2\pi JD/365)$] has the effect to increase predicted FC the most during late summer. The transformed model may be retransformed to the original units so that FC concentrations can be calculated directly. A potential bias that is introduced because of retransformation can be corrected using Duan's bias correction factor (BCF). For this model the BCF is 1.56. The retransformed model, using the BCF, is:

$$FC = 1.56 \times 10^{\log_{10}(FC)} \quad \text{or} \quad FC = (Turb^{1.25} \times 39.2) / [10^{0.543 \sin(2\pi JD/365)}]$$

Model Statistics, Data, and Plots

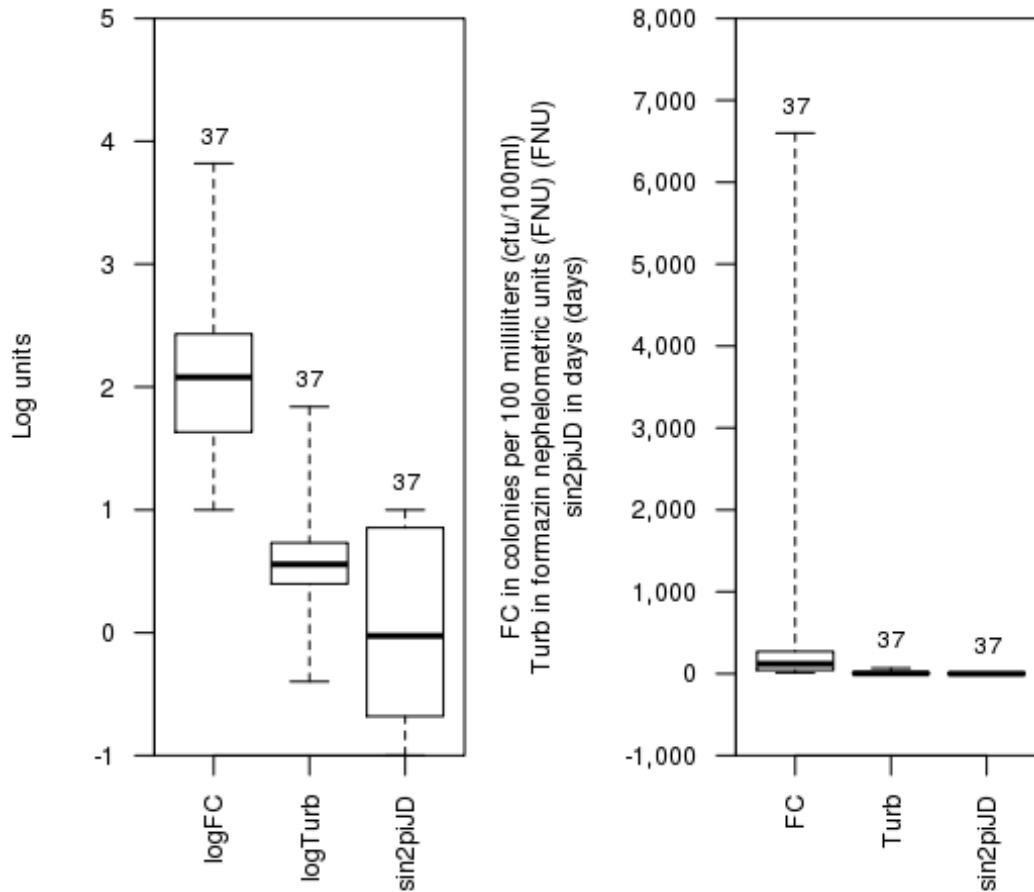
Model

$$\log FC = + 1.25 * \log Turb - 0.543 * \sin 2\pi jD + 1.4$$

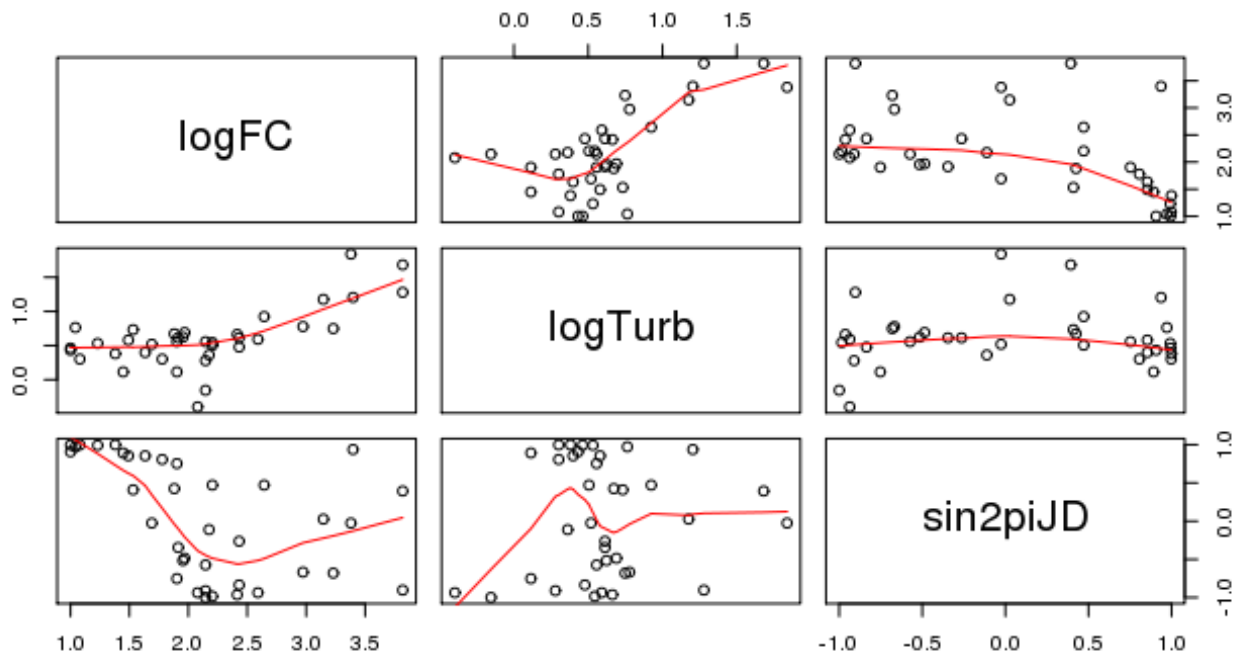
Variable Summary Statistics

	logFC	FC	logTurb	sin2piJD	Turb
Minimum	1.00	10	-0.398	-1.0000	0.40
1st Quartile	1.63	43	0.398	-0.6810	2.50
Median	2.08	120	0.556	-0.0258	3.60
Mean	2.14	692	0.607	0.0334	7.53
3rd Quartile	2.43	270	0.732	0.8540	5.40
Maximum	3.82	6600	1.840	1.0000	69.00

Box Plots



Exploratory Plots



Basic Model Statistics

Number of Observations 37
 Standard error (RMSE) 0.423
 Average Model standard percentage error (MSPE) 114
 Coefficient of determination (R^2) 0.708
 Adjusted Coefficient of Determination (Adj. R^2) 0.691
 Bias Correction Factor (BCF) 1.56

Variance Inflation Factors (VIF)

logTurb	sin2piJD
1.01	1.01

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(> t)
(Intercept)	1.400	0.1220	11.50	2.72e-13
logTurb	1.250	0.1650	7.54	9.25e-09
sin2piJD	-0.543	0.0939	-5.78	1.66e-06

Correlation Matrix

	Intercept	logTurb	sin2piJD
Intercept	1.0000	-0.8200	0.0558
logTurb	-0.8200	1.0000	-0.0991
sin2piJD	0.0558	-0.0991	1.0000

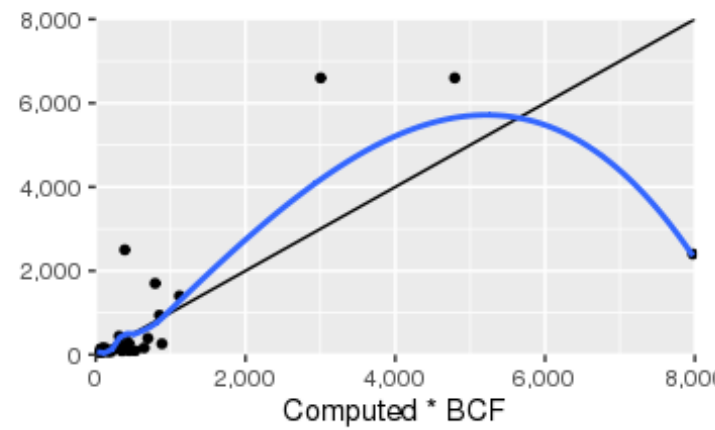
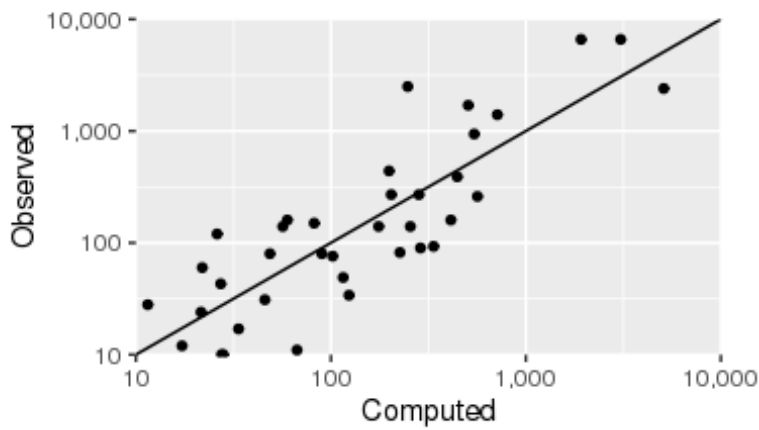
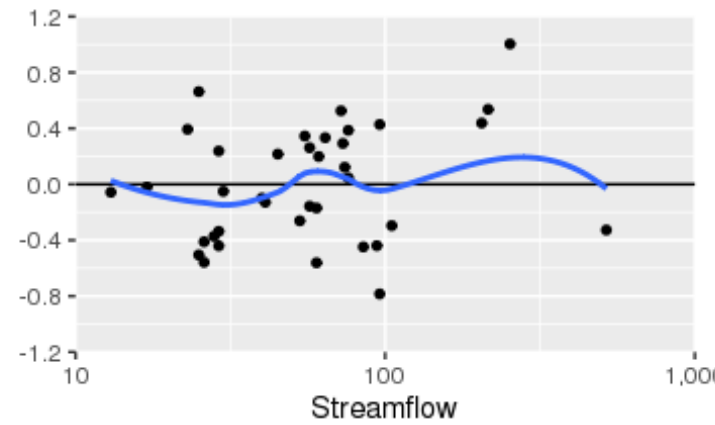
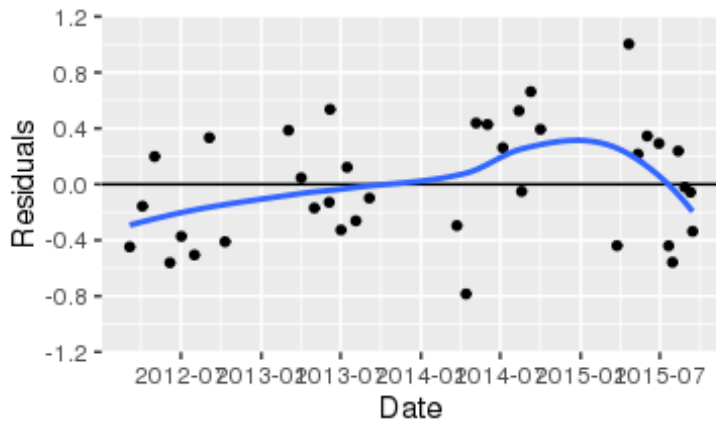
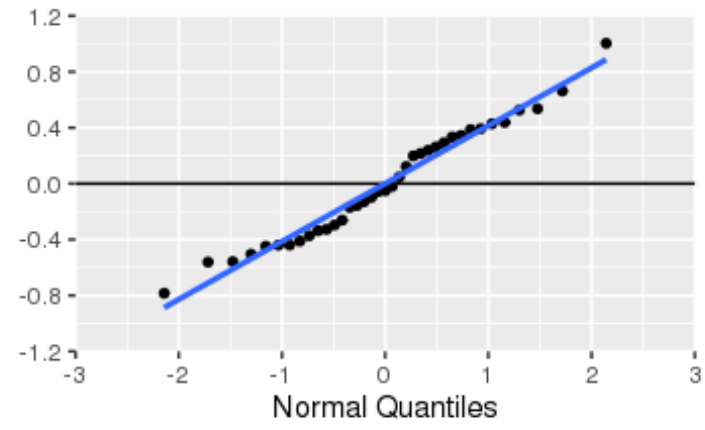
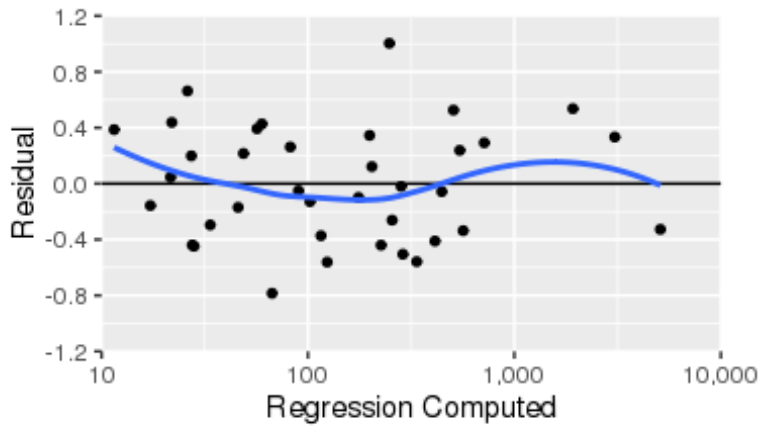
Outlier Test Criteria

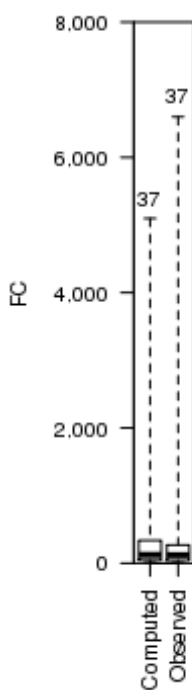
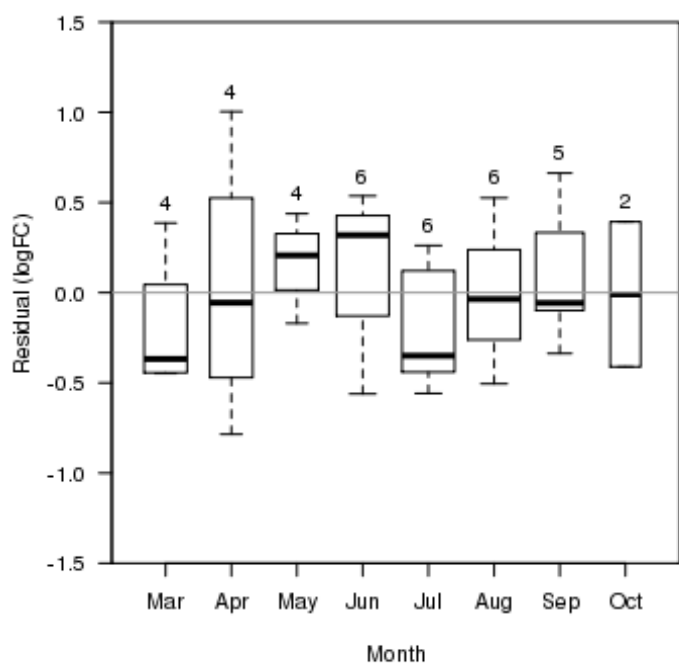
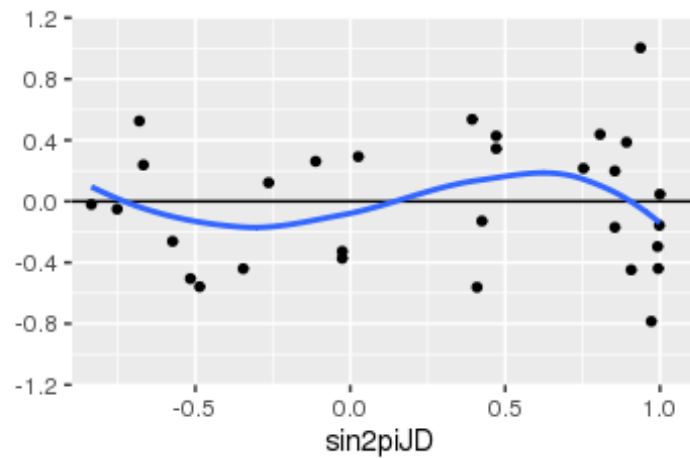
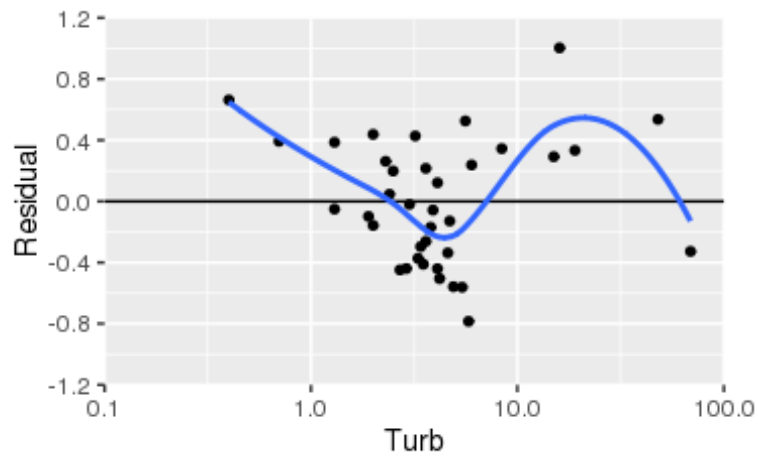
Leverage	Cook's D	DFFITS
0.162	0.194	0.465

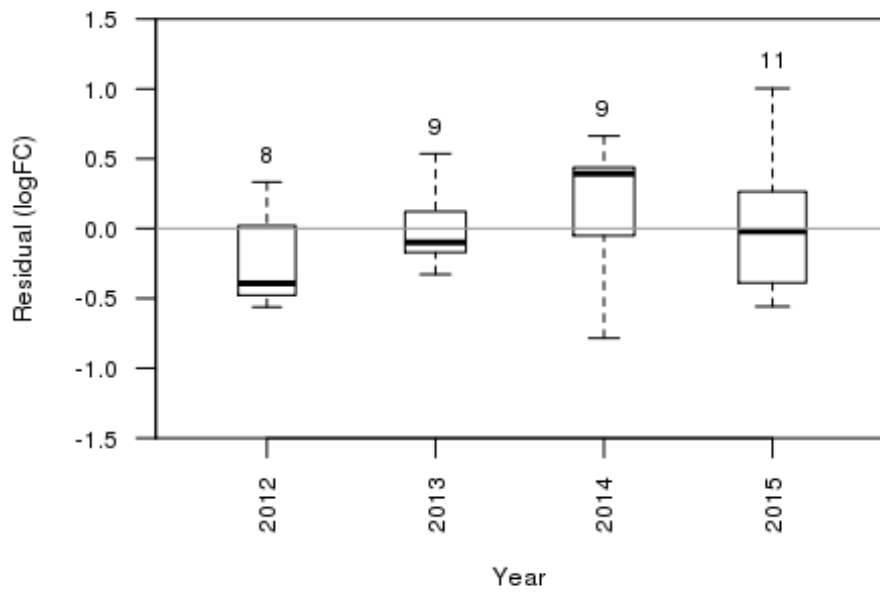
Flagged Observations

	logFC	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
06/07/2013 09:00	3.82	3.28	0.536	1.420	1.440	0.2020	0.1700	0.726
07/02/2013 09:45	3.38	3.71	-0.327	-0.898	-0.896	0.2590	0.0942	-0.530
04/14/2014 08:45	1.04	1.83	-0.785	-1.920	-2.010	0.0717	0.0953	-0.558
09/09/2014 08:15	2.08	1.42	0.663	1.760	1.820	0.2100	0.2760	0.940
04/21/2015 12:45	3.40	2.39	1.000	2.520	2.750	0.1120	0.2670	0.978

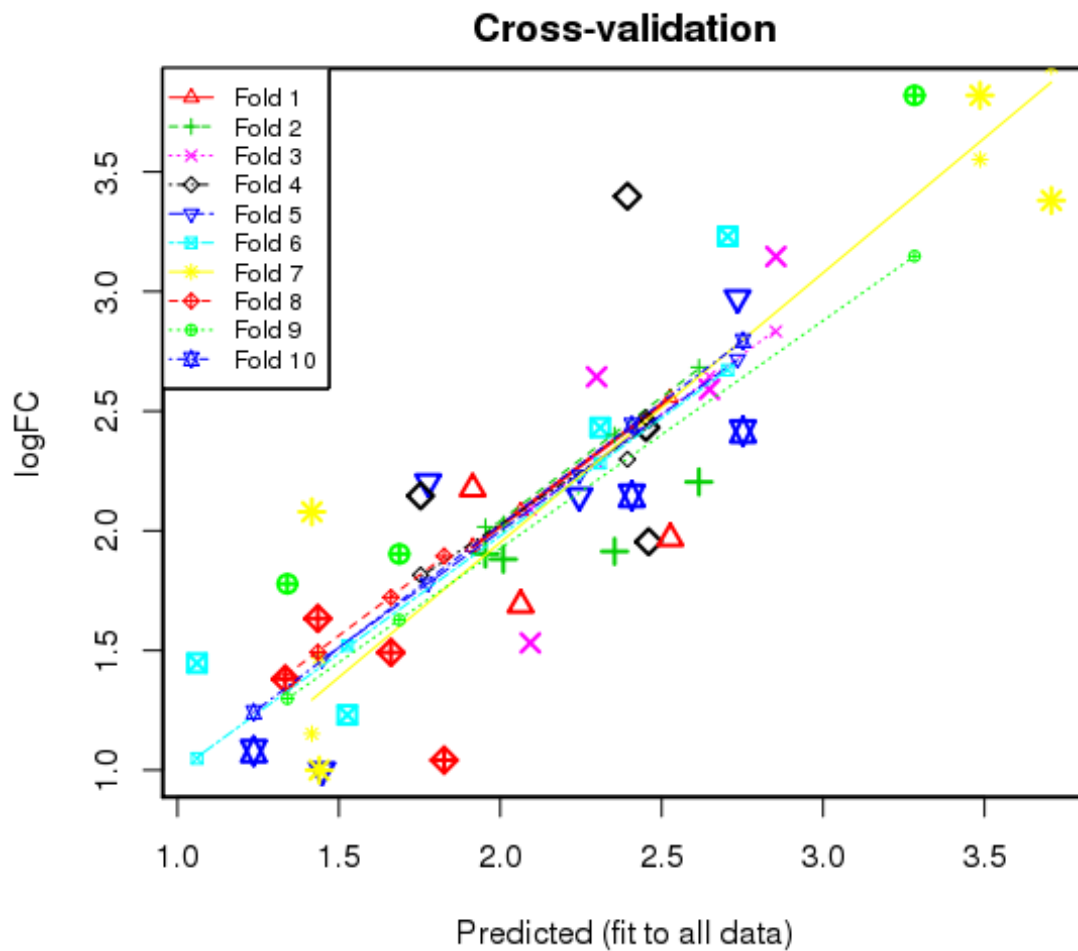
Statistical Plots



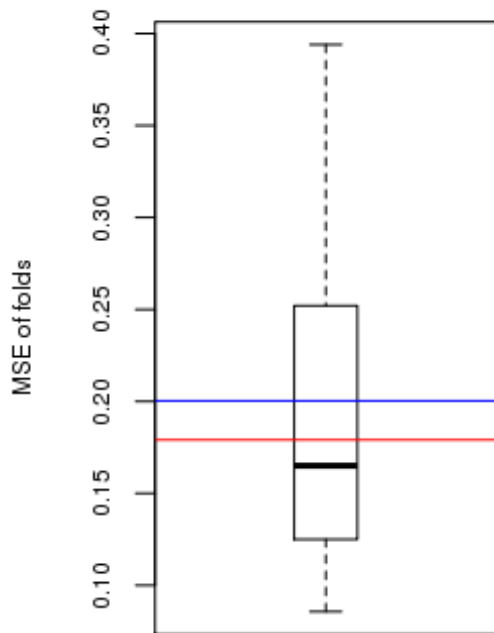




Cross Validation



Minimum MSE of folds: 0.0857
 Mean MSE of folds: 0.2000
 Median MSE of folds: 0.1650
 Maximum MSE of folds: 0.3940
 (Mean MSE of folds) / (Model MSE): 1.1200



Red line - Model MSE

Blue line - Mean MSE of folds

Model-Calibration Data Set

	Date	logFC	logTurb	sin2piJD	FC	Turb	Computed logFC	Computed FC	Residual	Normal Quantiles	Censored Values
0											
1	2012-03-06	1	0.431	0.907	10	2.7	1.45	43.9	-0.448	-1.16	--
2	2012-04-04	1.08	0.301	0.998	12	2	1.24	27	-0.157	-0.273	--
3	2012-05-02	1.63	0.398	0.854	43	2.5	1.43	42.6	0.199	0.273	--
4	2012-06-06	1.53	0.732	0.409	34	5.4	2.09	194	-0.562	-1.72	--
5	2012-07-02	1.69	0.519	-0.0258	49	3.3	2.06	181	-0.373	-0.735	--
6	2012-08-01	1.95	0.623	-0.516	90	4.2	2.46	451	-0.505	-1.3	--
7	2012-09-04	3.82	1.28	-0.903	6600	19	3.49	4790	0.333	0.649	--
8	2012-10-10	2.2	0.544	-0.984	160	3.5	2.62	645	-0.411	-0.826	--
9	2013-03-04	1.45	0.114	0.892	28	1.3	1.06	18	0.386	0.826	--
10	2013-04-02	1.38	0.38	1	24	2.4	1.33	33.8	0.0462	0.135	--
11	2013-05-02	1.49	0.58	0.854	31	3.8	1.66	71.7	-0.17	-0.344	--

12	2013-06-05	1.88	0.672	0.425	76	4.7	2.01	160	-0.129	-0.204	--
13	2013-06-07	3.82	1.68	0.394	6600	48	3.28	3000	0.536	1.48	--
14	2013-07-02	3.38	1.84	-0.0258	2400	69	3.71	7970	-0.327	-0.568	--
15	2013-07-16	2.43	0.613	-0.264	270	4.1	2.31	319	0.122	0.204	--
16	2013-08-05	2.15	0.556	-0.574	140	3.6	2.41	400	-0.261	-0.416	--
17	2013-09-05	2.15	0.279	-0.911	140	1.9	2.24	275	-0.0987	-0.135	--
18	2014-03-24	1.23	0.531	0.992	17	3.4	1.53	52.6	-0.296	-0.491	--
19	2014-04-14	1.04	0.763	0.972	11	5.8	1.83	105	-0.785	-2.14	--
20	2014-05-07	1.78	0.301	0.806	60	2	1.34	34.2	0.438	1.16	--
21	2014-06-02	2.2	0.505	0.471	160	3.2	1.78	93.5	0.428	1.03	--
22	2014-07-07	2.18	0.362	-0.112	150	2.3	1.91	128	0.262	0.491	--
23	2014-08-13	3.23	0.748	-0.681	1700	5.6	2.7	793	0.526	1.3	--
24	2014-08-19	1.9	0.114	-0.753	80	1.3	1.95	141	-0.0508	0	--
25	2014-09-09	2.08	-0.398	-0.937	120	0.4	1.42	40.8	0.663	1.72	--
26	2014-10-01	2.15	-0.155	-1	140	0.7	1.75	88.6	0.393	0.925	--
27	2015-03-25	1	0.462	0.994	10	2.9	1.44	43	-0.439	-0.925	--
28	2015-04-21	3.4	1.2	0.937	2500	16	2.39	388	1	2.14	--
29	2015-05-12	1.9	0.556	0.753	80	3.6	1.69	76.2	0.216	0.344	--
30	2015-06-02	2.64	0.924	0.471	440	8.4	2.3	311	0.345	0.735	--
31	2015-06-29	3.15	1.18	0.0258	1400	15	2.85	1120	0.292	0.568	--
32	2015-07-21	1.91	0.613	-0.346	82	4.1	2.35	354	-0.44	-1.03	--
33	2015-07-30	1.97	0.69	-0.486	93	4.9	2.53	526	-0.558	-1.48	--
34	2015-08-12	2.97	0.778	-0.668	940	6	2.74	850	0.238	0.416	--
35	2015-08-27	2.43	0.477	-0.836	270	3	2.45	442	-0.02	0.0674	--
36	2015-09-09	2.59	0.591	-0.937	390	3.9	2.65	696	-0.057	-0.0674	--
37	2015-09-14	2.41	0.663	-0.963	260	4.6	2.75	883	-0.337	-0.649	--

Definitions

FC: Fecal coliforms in cfu/100ml (31625)

Turb: Turbidity in FNU (63680)

sin2piJD: $\sin(2\pi JD/365)$ in day/days