Appendix 5. Model Archive Summary for Best-Fit Regression Developed to Estimate Fecal Coliform Concentration at Station 01478245; White Clay Creek near Strickersville, 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: 01478245

Station name: White Clay Creek near Strickersville, Pennsylvania

Location: Latitude 39°44'51", longitude 75°46'15" referenced to North American Datum of 1983, Chester County, PA, Hydrologic Unit 02040205, on right bank 0.1 mile (mi) downstream from West Branch White Clay Creek, in the White Clay Creek State Preserve, and 1.5 mi northeast of Strickersville.

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 5, 2011. 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 50 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. Seasonality 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 seasonality 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 (*sin2piJD* and *cos2piJD)* 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 lead to the conclusion that the most appropriate and reliable model would be one that estimated $\log_{10}(FC)$.

$\text{Log}_{10}(Turb)$,), *sin2piJD,* and *cos2piJD* 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 01478245,
FC concentration-based model:

$$\log_{10}(FC)=0.653\times\log_{10}(Turb)-0.281\text{x}\sin(2\pi JD/365)\ )-0.747\text{x}\cos(2\pi JD/365)+1.87$$

where

FC = fecal coliform in colonies per 100 milliliter (col/mL) (parameter 31625);
Turb = turbidity in formazine nephelometric units (FNU) (parameter 63680);
Sin & Cos = sine and cosine functions used to compute seasonality components; 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 summer months when stream temperatures are highest. The negative coefficients for seasonal variables sin2piJD and cos2piJD [computed as $\sin(2\pi JD/365)$ and $\cos(2\pi JD/365)$] has the effect to increase predicted *FC* the most during the peak of 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 1.60. The retransformed model, using the BCF, is:

$FC = 1.60x10^{\log_{10}(FC)}$ or $FC = (Turb^{\,0.653} \text{ x } 118.6)/[10^{\,0.281\sin(2\pi JD/365)} \text{ x } 10^{\,0.747\cos(2\pi JD/365}]$

Model Statistics, Data, and Plots

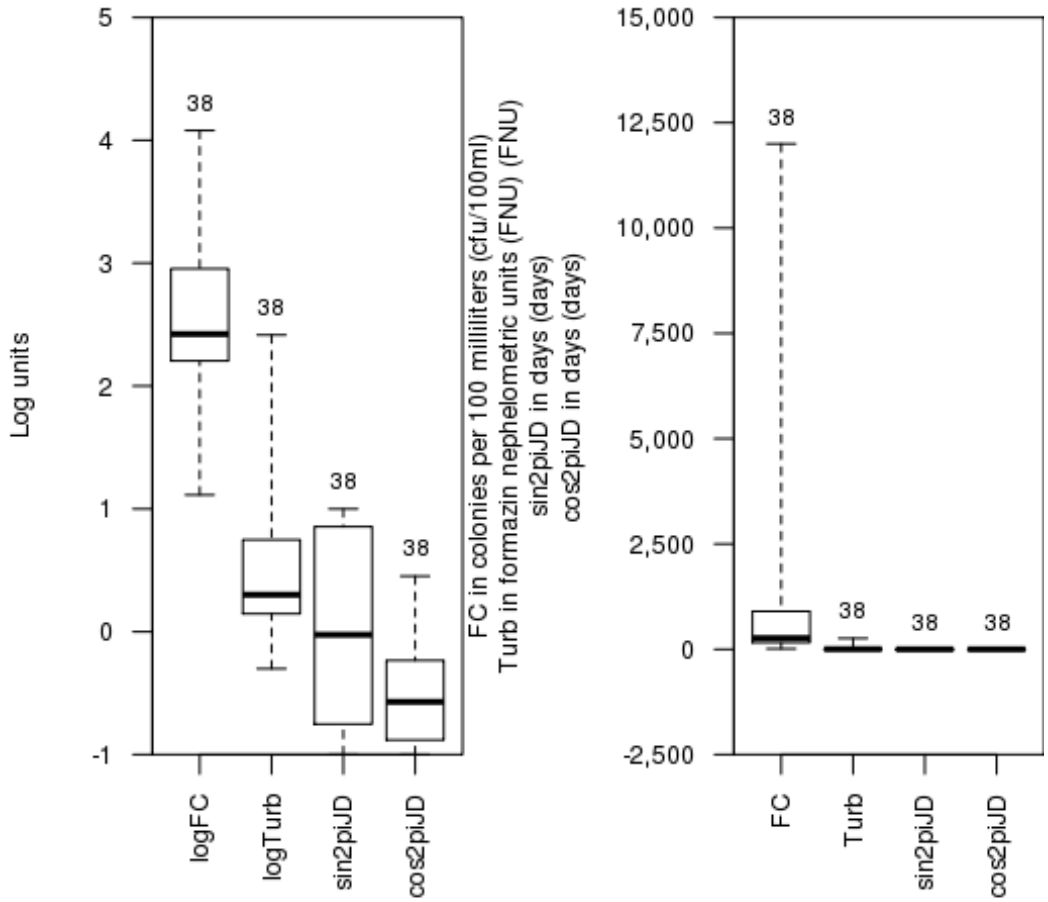
Model

logFC = + 0.653 * logTurb - 0.281 * sin2piJD - 0.747 * cos2piJD + 1.87

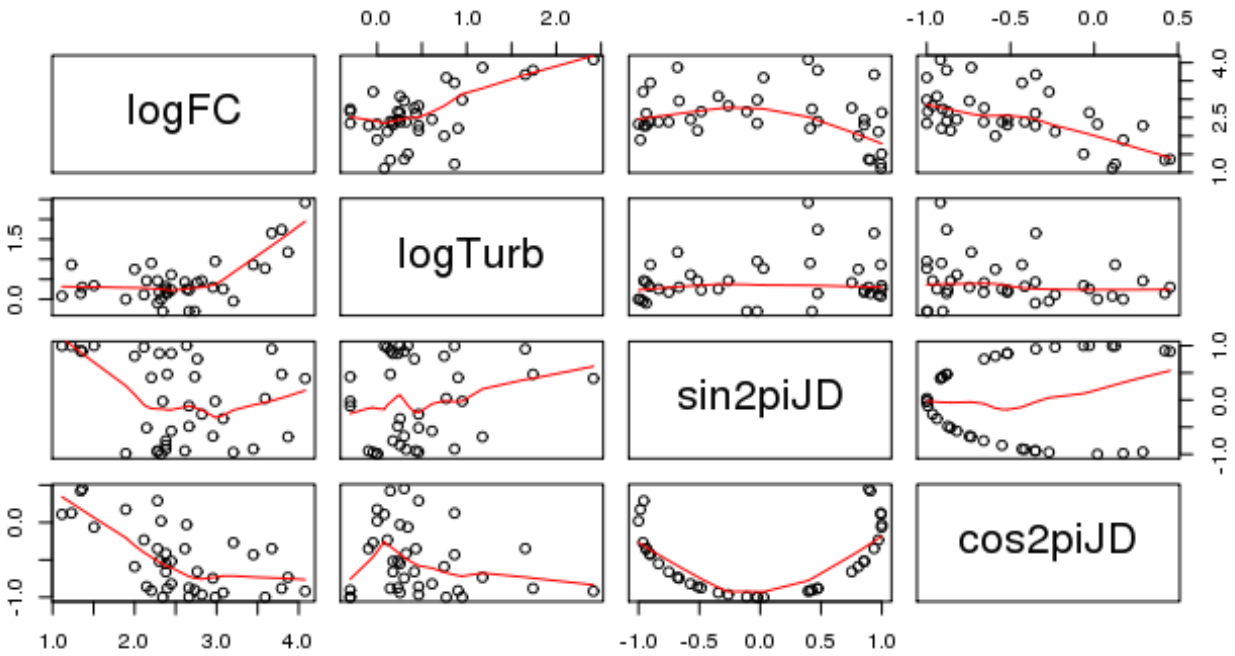
Variable Summary Statistics

	logFC	FC	logTurb	sin2piJD	cos2piJD	Turb
Minimum	1.11	13	-0.301	-1.00000	-1.000	0.5
1st Quartile	2.20	160	0.146	-0.75100	-0.881	1.4
Median	2.42	265	0.301	-0.02420	-0.570	2.0
Mean	2.54	1280	0.453	0.00814	-0.495	12.3
3rd Quartile	2.95	900	0.748	0.85500	-0.234	5.6
Maximum	4.08	12000	2.410	1.00000	0.453	260.0

Box Plots



Exploratory Plots



Basic Model Statistics

```
Number of Observations                              38
Standard error (RMSE)                            0.452
Average Model standard percentage error (MSPE)     124
Coefficient of determination (R²)                0.642
Adjusted Coefficient of Determination (Adj. R²)   0.61
Bias Correction Factor (BCF)                       1.6

Variance Inflation Factors (VIF)
 logTurb sin2piJD cos2piJD
    1.09     1.07     1.08
```

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	1.870	0.119	15.70	3.73e-17
logTurb	0.653	0.136	4.79	3.17e-05
sin2piJD	-0.281	0.101	-2.78	8.82e-03
cos2piJD	-0.747	0.175	-4.26	1.52e-04

Correlation Matrix

	Intercept	logTurb	sin2piJD	cos2piJD
Intercept	1.0000	-0.346	-0.0397	0.608
logTurb	-0.3460	1.000	-0.2120	0.234
sin2piJD	-0.0397	-0.212	1.0000	-0.196
cos2piJD	0.6080	0.234	-0.1960	1.000

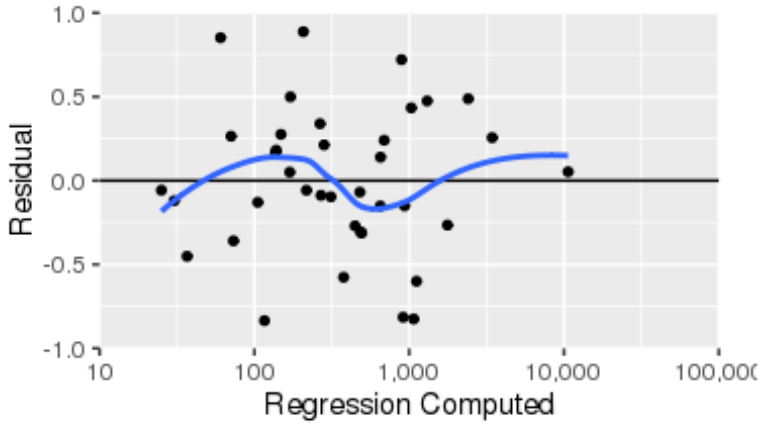
Outlier Test Criteria

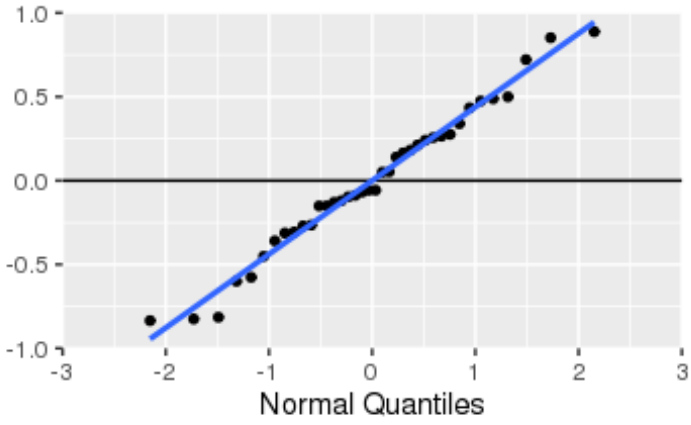
Leverage	Cook's D	DFFITS
0.237	0.262	0.562

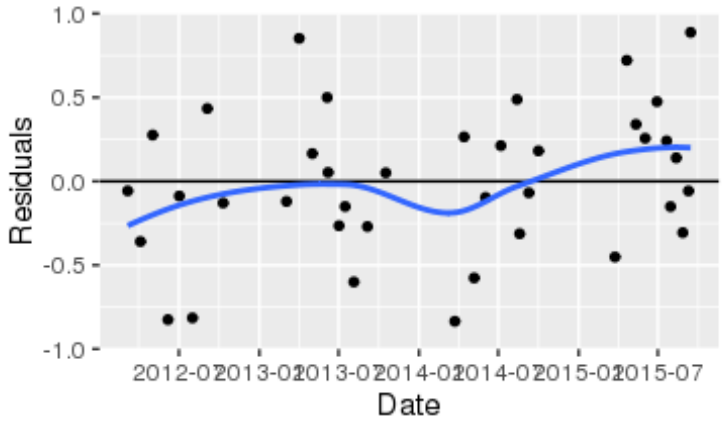
Flagged Observations

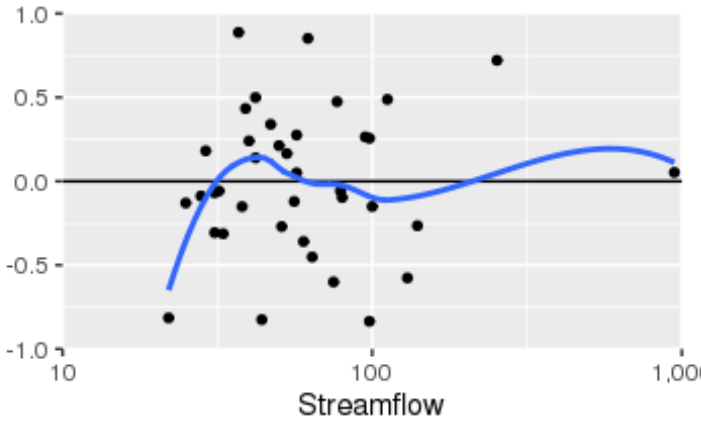
	logFC	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
04/02/2013 09:15	2.63	1.78	0.8520	1.980	2.080	0.0965	0.10500	0.678
06/07/2013 10:45	4.08	4.03	0.0527	0.144	0.142	0.3480	0.00278	0.104
03/24/2014 12:30	1.23	2.07	-0.8350	-1.980	-2.070	0.1290	0.14500	-0.799
04/21/2015 11:15	3.67	2.95	0.7210	1.760	1.820	0.1770	0.16500	0.841
09/14/2015 11:00	3.20	2.32	0.8880	2.060	2.170	0.0911	0.10600	0.686

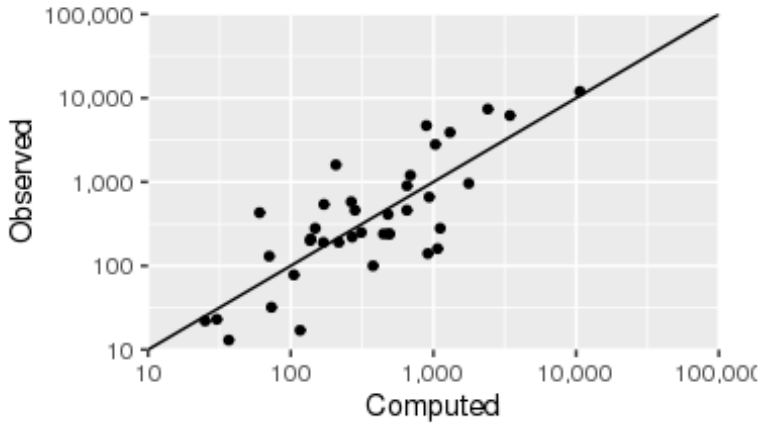
Statistical Plots

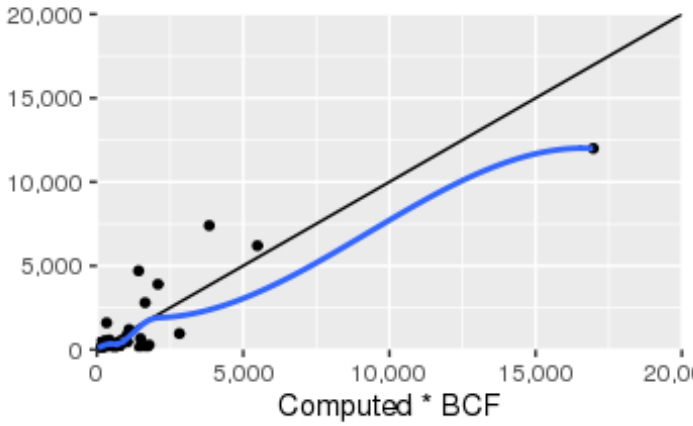


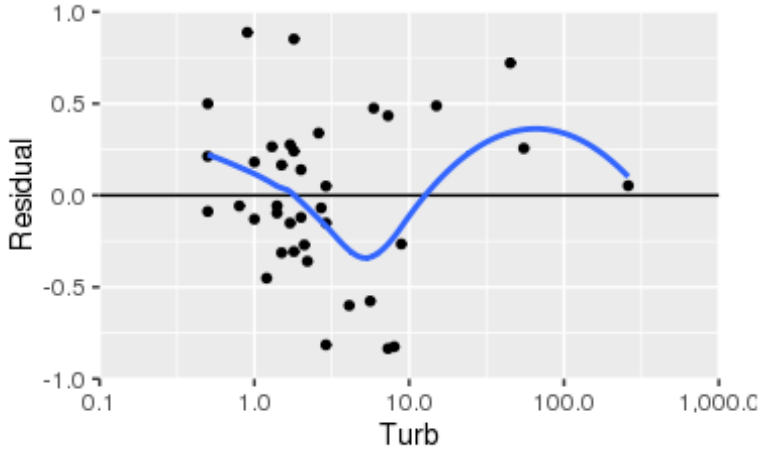


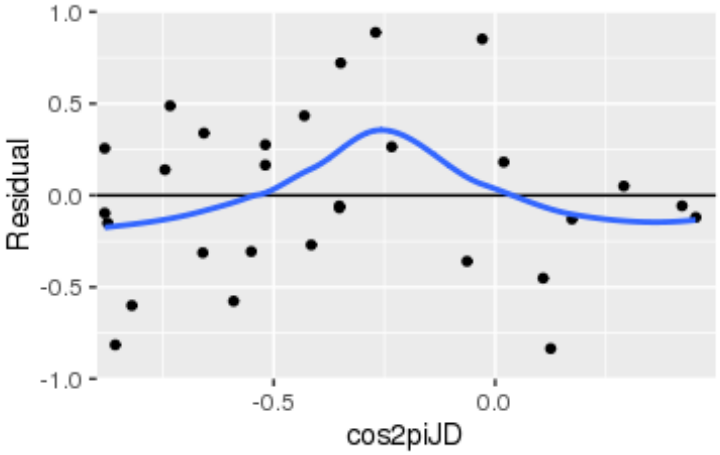


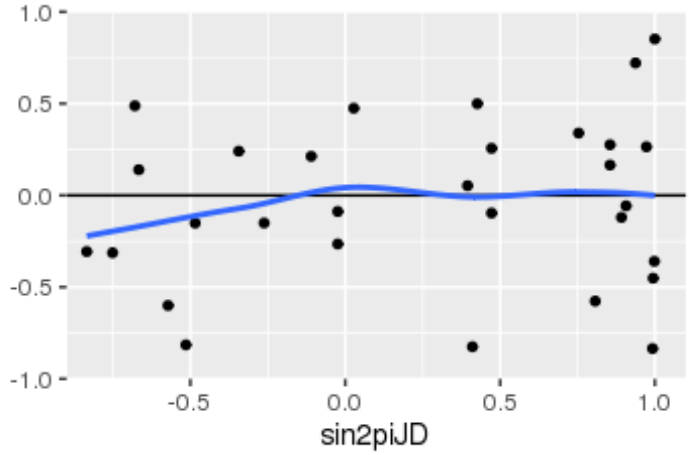


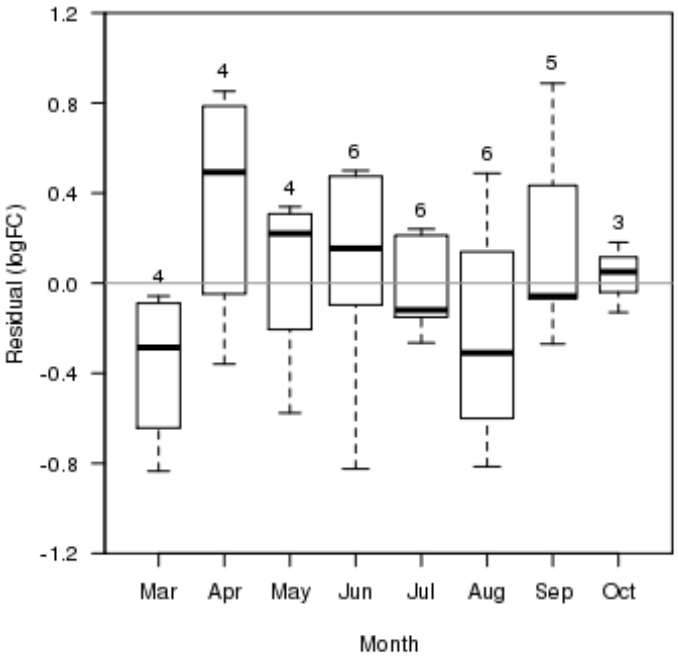


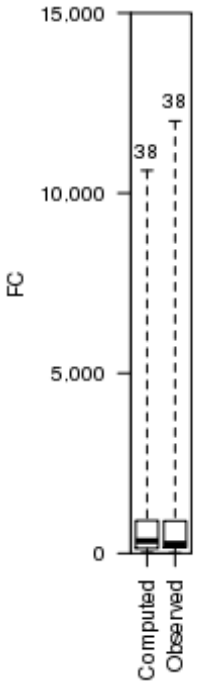


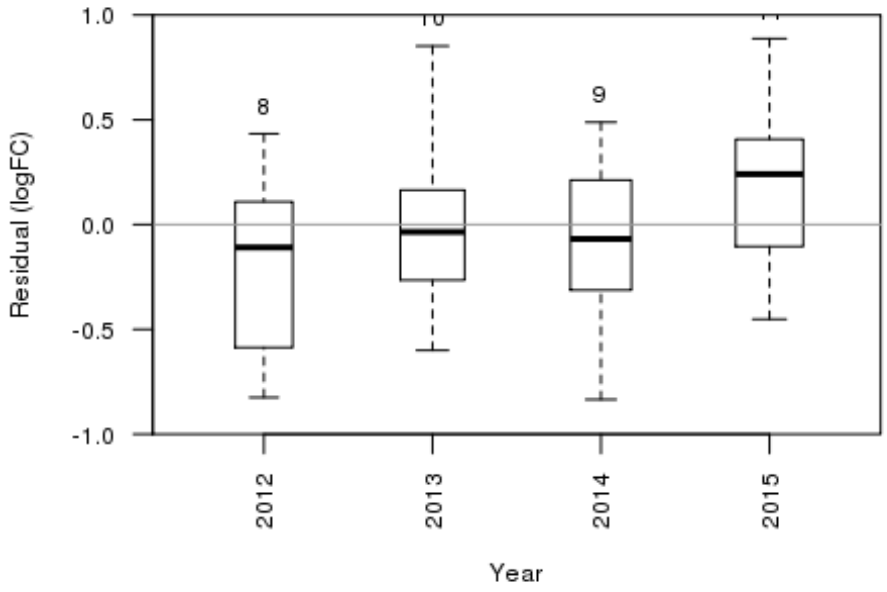

Residual
Turb
1.0
0.5
0.0
-0.5
-1.0
0.1
1.0
10.0
100.0
1,000.0


Residual
cos2piJD
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0

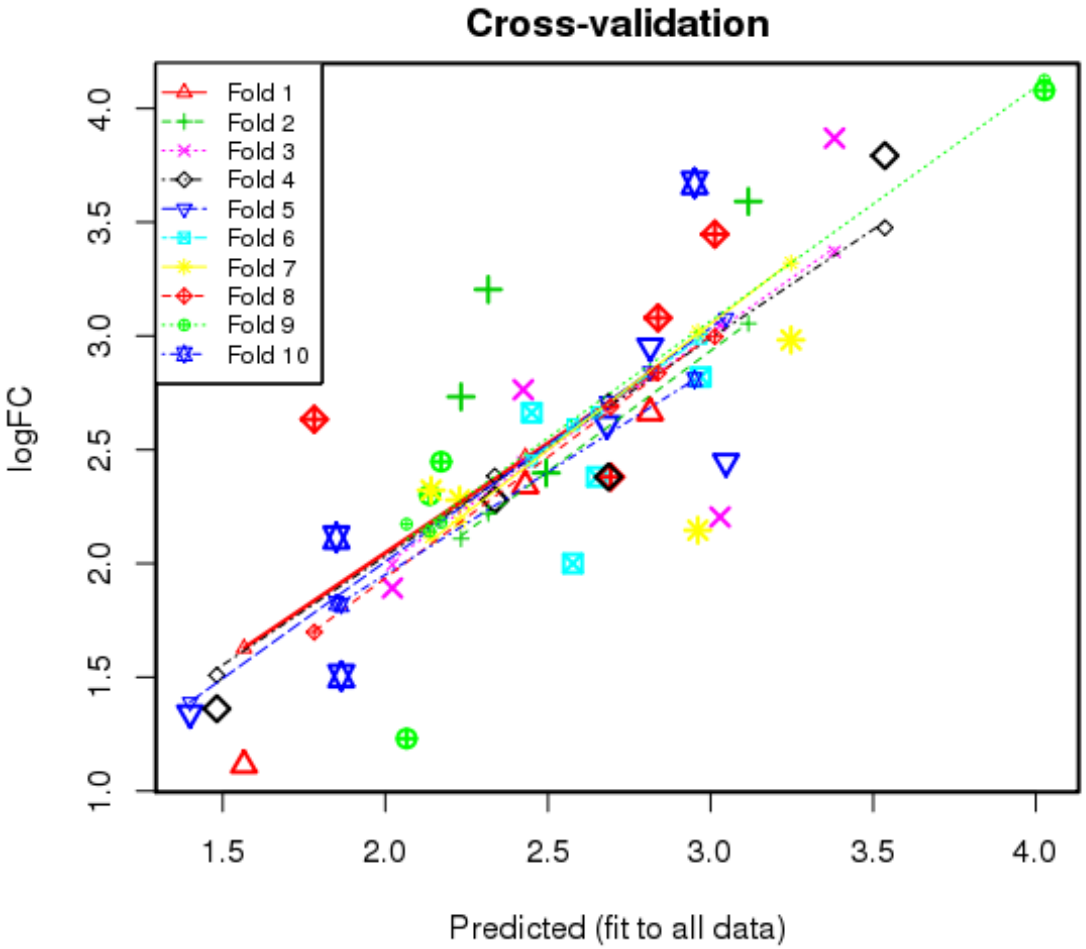

sin2piJD
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0


Residual (logFC)
Month
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
1.2
0.8
0.4
0.0
-0.4
-0.8
-1.2
4
4
4
6
6
6
5
3

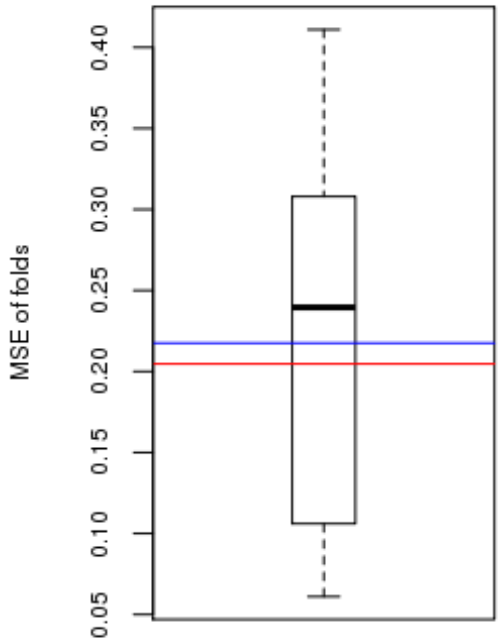

FC
15,000
10,000
5,000
0
38
38
Computed
Observed



Cross Validation



```
             Minimum MSE of folds:  0.061
                Mean MSE of folds:  0.218
              Median MSE of folds:  0.240
             Maximum MSE of folds:  0.411
(Mean MSE of folds) / (Model MSE):  1.060
```




```
Red line - Model MSE

Blue line - Mean MSE of folds
```

Model-Calibration Data Set

	Date	logFC	logTurb	sin2piJD	cos2piJD	FC	Turb	Computed logFC	Computed FC	Residual	Normal Quantiles	Censored Values
1	2012-03-06	1.34	0.146	0.907	0.422	22	1.4	1.4	40.1	-0.0572	0.0328	--
2	2012-04-04	1.51	0.342	0.998	-0.0637	32	2.2	1.86	117	-0.359	-0.943	--
3	2012-05-02	2.45	0.23	0.855	-0.519	280	1.7	2.17	237	0.276	0.755	--
4	2012-06-06	2.2	0.903	0.411	-0.912	160	8	3.03	1710	-0.825	-1.73	--
5	2012-07-02	2.34	-0.301	-0.0242	-1	220	0.5	2.43	430	-0.088	-0.165	--
6	2012-08-01	2.15	0.462	-0.514	-0.858	140	2.9	2.96	1460	-0.815	-1.49	--
7	2012-09-04	3.45	0.863	-0.902	-0.431	2800	7.3	3.01	1650	0.434	0.943	--
8	2012-10-10	1.89	0	-0.985	0.173	78	1	2.02	168	-0.13	-0.369	--
9	2013-03-04	1.36	0.301	0.892	0.453	23	2	1.48	48.4	-0.12	-0.3	--
10	2013-04-02	2.63	0.255	1	-0.0293	430	1.8	1.78	96.6	0.852	1.73	--
11	2013-05-02	2.3	0.176	0.855	-0.519	200	1.5	2.14	218	0.165	0.3	--

12	2013-06-05	2.73	-0.301	0.426	-0.905	540	0.5	2.23	273	0.5	1.32	--
13	2013-06-07	4.08	2.41	0.395	-0.919	12000	260	4.03	17000	0.0527	0.165	--
14	2013-07-02	2.98	0.949	-0.0242	-1	960	8.9	3.25	2820	-0.265	-0.59	--
15	2013-07-16	2.82	0.462	-0.262	-0.965	660	2.9	2.97	1490	-0.151	-0.44	--
16	2013-08-05	2.45	0.613	-0.572	-0.82	280	4.1	3.05	1780	-0.6	-1.32	--
17	2013-09-05	2.38	0.322	-0.91	-0.415	240	2.1	2.65	714	-0.27	-0.67	--
18	2013-10-17	2.28	0.462	-0.957	0.29	190	2.9	2.23	270	0.0503	0.0986	--
19	2014-03-24	1.23	0.863	0.992	0.125	17	7.3	2.07	186	-0.835	-2.15	--
20	2014-04-14	2.11	0.114	0.972	-0.234	130	1.3	1.85	113	0.265	0.67	--
21	2014-05-07	2	0.748	0.807	-0.59	100	5.6	2.58	602	-0.576	-1.17	--
22	2014-06-02	2.4	0.146	0.472	-0.881	250	1.4	2.49	499	-0.0967	-0.232	--
23	2014-07-07	2.66	-0.301	-0.11	-0.994	460	0.5	2.45	450	0.213	0.44	--
24	2014-08-13	3.87	1.18	-0.679	-0.734	7400	15	3.38	3840	0.488	1.17	--
25	2014-08-19	2.38	0.176	-0.751	-0.66	240	1.5	2.69	787	-0.313	-0.845	--
26	2014-09-09	2.61	0.431	-0.936	-0.352	410	2.7	2.68	767	-0.0688	-0.0986	--
27	2014-10-01	2.32	0	-1	0.0191	210	1	2.14	221	0.181	0.369	--
28	2015-03-25	1.11	0.0792	0.994	0.108	13	1.2	1.57	58.7	-0.451	-1.05	--
29	2015-04-21	3.67	1.65	0.937	-0.349	4700	45	2.95	1430	0.721	1.49	--
30	2015-05-12	2.76	0.415	0.753	-0.658	580	2.6	2.42	424	0.339	0.845	--
31	2015-06-02	3.79	1.74	0.472	-0.881	6200	55	3.54	5490	0.256	0.59	--
32	2015-06-29	3.59	0.771	0.0274	-1	3900	5.9	3.12	2090	0.475	1.05	--
33	2015-07-21	3.08	0.255	-0.344	-0.939	1200	1.8	2.84	1100	0.241	0.514	--
34	2015-07-30	2.66	0.23	-0.485	-0.875	460	1.7	2.81	1040	-0.151	-0.514	--
35	2015-08-12	2.95	0.301	-0.667	-0.745	900	2	2.81	1040	0.14	0.232	--
36	2015-08-27	2.38	0.255	-0.835	-0.551	240	1.8	2.69	776	-0.306	-0.755	--
37	2015-09-09	2.28	-0.0969	-0.936	-0.352	190	0.8	2.34	347	-0.0577	-0.0328	--
38	2015-09-14	3.2	-0.0458	-0.963	-0.27	1600	0.9	2.32	331	0.888	2.15	--

Definitions

```
FC: Fecal coliforms in cfu/100ml (31625)
Turb: Turbidity in FNU (63680)
sin2piJD: sin(2πJD/365) in day/days
cos2piJD: cos(2πJD/365) in day/days
```

App Version 1.0