Appendix 4. 4-Year Logistic Regression Model Archival Summary for Geosmin Occurrence at Station 07144790, 2013–16

This model archival summary summarizes the logistic model for the probability of geosmin occurrence developed to compute hourly geosmin from January 1, 2013, onward.

Station and Model Information

Station number: 07144790
Station name: Cheney Re Nr Cheney, KS
Station location: Latitude 37°43'34", Longitude 97°47'38" referenced to the North American Datum of 1927, in SE¼NE¼NW¼ sec. 6, T. 27 S., R. 04 W., Sedgwick County, Kansas, Hydrologic Unit 11030014.

Equipment: From April 2001 through September 2014, a YSI 6600 water-quality monitor was installed equipped with sensors for water temperature, specific conductance, dissolved oxygen (YSI Clark cell [from April 2001 through January 2007] or YSI model 6150 optical [from February 2007 through September 2014]), pH, turbidity (YSI model 6026 [from April 2001 through September 2006] or YSI 6136 [from October 2006 through September 2014]), and chlorophyll. From October 2014 to the present (December 2016), a Xylem YSI EXO2 water-quality monitor has been used and is equipped with sensors for water temperature, specific conductance, dissolved oxygen, pH, turbidity, and chlorophyll fluorescence (YSI model 6025 sensor). The Xylem monitor is housed in a 4-inch diameter galvanized steel pipe. Readings from the water-quality monitor are recorded hourly and data are transmitted hourly by satellite.

Date model was created: August 16, 2016

Model calibration data period: January 15, 2013, through June 15, 2016

Model application date: August 2016 onward

Model-Calibration Dataset

All data were collected using U.S. Geological Survey (USGS) protocols (U.S. Geological Survey, variously dated; https://water.usgs.gov/owq/FieldManual/) and are stored in the National Water Information System database (https://doi.org/10.5066/F7P55KJN). Logistic model equations were developed using the multiple logistic regression routine in SigmaPlot® version 11.0 (Systat Software, Inc., 2008). Explanatory variables were evaluated individually and in selected combinations. Explanatory variables selected as inputs to logistic regression were physicochemical properties: specific conductance, pH, water temperature, dissolved oxygen, chlorophyll fluorescence, and elevation of the reservoir surface. Seasonal components (sine and cosine variables) also were evaluated as explanatory variables in the models to determine if seasonal changes affected the model. All combinations of physicochemical properties and a seasonal component were evaluated to determine which combinations produced the best models.

The final selected logistic regression model is based on 48 concurrent measurements of geosmin occurrence collected from January 15, 2013, through June 15, 2016, and models the probability of the presence or absence of geosmin. Samples were collected throughout the range of continuously observed hydrologic conditions. In total, seven samples were below the threshold for positive classification (5 nanograms per liter [ng/L]). Summary statistics and the complete model-calibration dataset are provided below. Studentized residuals were inspected for values outside the 95-percent confidence interval, and leverage values for independent variables were inspected for values greater than 2. Values outside of the specified ranges were considered potential outliers and were investigated. No outliers were identified in the model-calibration dataset.

Geosmin Sampling Details

Discrete water-quality samples were collected monthly to biweekly during January 2013 through June 2016. Samples were collected as integrated photic-zone (depth at which light is about 1 percent of that at the surface) samples using a double check-valve bailer; these samples were depth integrated. Geosmin was analyzed using solid phase microextraction gas chromatography/mass spectrometry by Engineering Performance Solutions, LLC, Gainesville, Florida.

Model Development

Logistic regression analysis was done using SigmaPlot by examining seasonality and other continuously measured data as explanatory variables for estimating geosmin presence. Seasonality was selected as the best predictor of geosmin based on a relatively low Pearson Chi-square Statistic, relatively high Likelihood Ratio Test Statistic, relatively low -2 Log Likelihood Statistic, relatively high Hosmer-Lemeshow Statistic, significant Wald Statistic, and relatively low Variance Inflation Factor. A model classification table with a threshold probability for positive classification (TPPC) of 0.5 also was used in final model selection. After the best model was selected, the TPPC for the model was adjusted based on the fraction of data classified as positive to make the model more conservative (more likely to overestimate a positive response) by guarding more strongly against false negatives. Values for all of the aforementioned statistics and metrics were computed for various models and are included below along with all relevant sample data and more in-depth statistical information.

Model Summary

Summary of final logistic regression analysis for geosmin occurrence at USGS station 07144790.

Probability of geosmin occurrence model:

$$\text{logit}(P) = -34.118 - 3.279\sin\left(\frac{2\pi D}{365}\right) + 0.393\cos\left(\frac{2\pi D}{365}\right) + 3.995\left(pH\right) \qquad (4\text{–}1)$$

where

logit(P)	is the logistic probability of geosmin occurrence (concentrations greater than or equal to 5 nanograms per liter);
D	is the Julian day of the year;
pH	is pH, in standard units.

Seasonality (the information contained in the sine [sin] and cosine [cos] component of the equation; Helsel and Hirsch, 2002) and *pH* make physical and statistical sense as explanatory variables for geosmin.

Previously Published Model

$$\text{logit}(P) = 0.829 + 0.825\sin\left(\frac{2\pi D}{365}\right) - 0.262\cos\left(\frac{2\pi D}{365}\right) - 0.102\left(TBY\right) \qquad (4\text{–}2)$$

Model author: Stone and others (2013)

Model data period: May 2001 through December 2009

Probability of Geosmin Occurrence Record

The geosmin record is computed using this regression model, and the complete water-quality record is stored at the National Real-Time Water Quality website: https://nrtwq.usgs.gov/ks. Data are computed at 60-minute intervals.

SigmaPlot® Output for Geosmin at Station 07144790

4-Year Model Form

$$\text{logit}(P) = -34.118 - 3.279\sin\left(\frac{2\pi D}{365}\right) + 0.393\cos\left(\frac{2\pi D}{365}\right) + 3.995\left(pH\right) \qquad (4\text{–}3)$$

Variable Summary Statistics

[µg/L, microgram per liter, *pH*, pH in standard units; <, less than; --, not measured]

Summary statistic	Geosmin (µg/L)	Geosmin binary	*pH*
Minimum	<1	0	8
1st quartile	2.2	0	8.3
Median	4.8	1	9
Mean	--	1	--
3d quartile	13	1	9
Maximum	54	1	9

Model Calibration Using Multiple Logistic Regression

See the model form in equation 4–3 above.

Number of samples=48
Missing observations=182
Estimation criterion: Maximum likelihood
Dependent variable: Geosmin (abbr) plus or minus (±)
Positive response=1
Reference response=0
Number of unique independent variable combinations=48

Pearson Chi-square Statistic=48.980 (probability value [*p*-value]=0.246)
Likelihood Ratio Test Statistic=26.395 (*p*-value=less than 0.001)
-2*Log(Likelihood)=40.147
Hosmer-Lemeshow Statistic=9.352 (*p*-value=0.313)

TPPC=0.5

Classification table	Predicted reference responses	Predicted positive responses	Total actual responses	Percent correctly classified responses
Actual reference responses	19	5	24	79
Actual positive responses	3	21	24	88
Total	22	26	48	83

Details of the logistic regression equation:

[*p*-value, probability value; VIF, Variance Inflation Factor; --, not measured; sin, sine of the seasonality component; cos, cosine of the seasonality component; *pH*, pH in standard units; <, less than]

Independent variable	Coefficient	Standard error	Wald statistic	*p*-value	VIF
Constant	–34.118	17.635	3.743	0.053	--
sin	–3.279	0.943	12.079	<0.001	1.149
cos	0.393	0.513	0.587	0.444	1.029
pH	3.995	2.072	3.717	0.054	1.180

Independent variable	Odds ratio	Lower 5-percent confidence interval	Upper 95-percent confidence interval
Constant	1.523×10^{-15}	1.486×10^{-30}	1.560
sin	0.0377	0.00593	0.239
cos	1.481	0.542	4.048
pH	54.303	0.936	3,151.114

Data Used in Model Development

[sin, sine of the seasonality component; cos, cosine of the seasonality component; ng/L, nanogram per liter; ≥, greater than or equal to; *pH*, pH in standard units; <, less than]

Date	Julian date	sin	sin	Geosmin (ng/L)	Geosmin binary (≥ 5 ng/L)	*pH*	Computed probability	Correct classification
01/15/2013	015	0.262	0.965	2.8	0	8.6	0.5179	No
01/23/2013	023	0.392	0.920	<1	0	8.5	0.5141	No
02/12/2013	043	0.679	0.734	2.3	0	8.6	0.4667	Yes
03/19/2013	078	0.976	0.220	2.1	0	8.6	0.3738	Yes
04/09/2013	099	0.990	–0.139	<1	0	8.6	0.2761	Yes
05/07/2013	127	0.813	–0.583	1.7	0	8.5	0.3528	Yes
06/06/2013	157	0.419	–0.908	<1	0	8.4	0.1849	Yes
07/08/2013	189	–0.119	–0.993	12.5	1	8.6	0.1618	No
07/23/2013	204	–0.368	–0.930	4.5	0	8.5	0.1907	Yes
08/06/2013	218	–0.579	–0.815	17.5	1	8.5	0.2594	No
08/19/2013	231	–0.746	–0.666	3.9	0	8.2	0.2597	Yes
09/06/2013	250	–0.915	–0.404	11.1	1	8.2	0.2918	No
09/09/2013	252	–0.933	–0.359	12.6	1	8.3	0.3033	No
09/12/2013	255	–0.951	–0.310	7.6	1	7.8	0.2358	No
09/25/2013	268	–0.996	–0.092	50	1	7.9	0.2196	No
09/30/2013	273	–1.000	–0.006	33.7	1	8.3	0.2297	No
10/21/2013	294	–0.937	0.348	54.1	1	8.5	0.2664	No
10/28/2013	301	–0.888	0.459	33.8	1	8.45	0.2523	No
11/06/2013	310	–0.808	0.590	27.9	1	8.5	0.2471	No
11/13/2013	318	–0.729	0.685	18.1	1	8.5	0.2589	No
11/18/2013	322	–0.669	0.743	16.5	1	8.5	0.2806	No
12/12/2013	346	–0.314	0.949	5.8	1	8.4	0.3837	No
01/15/2014	015	0.262	0.965	2.8	0	8.4	0.4421	Yes
02/19/2014	050	0.763	0.646	2.5	0	8.3	0.4763	Yes
03/19/2014	078	0.976	0.220	1.9	0	9	0.3993	Yes
04/16/2014	106	0.966	–0.257	2.1	0	8.2	0.3934	Yes
05/20/2014	140	0.663	–0.748	<1	0	8.24	0.2930	Yes
06/25/2014	176	0.105	–0.995	22.2	1	8.6	0.3231	No
07/10/2014	191	–0.153	–0.988	4.3	0	8.3	0.2451	Yes
07/22/2014	203	–0.352	–0.936	16.6	1	8.4	0.2692	No
08/05/2014	217	–0.565	–0.825	4.3	0	8.2	0.2942	Yes
09/16/2014	259	–0.970	–0.245	11.8	1	8.6	0.3196	No

10/28/2014	301	–0.889	0.458	5	1	8.2	0.3222	No
11/20/2014	324	–0.643	0.766	13.1	1	8.7	0.4713	No
12/16/2014	350	–0.255	0.967	11.1	1	8.5	0.3518	No
01/13/2015	013	0.222	0.975	7.6	1	8.7	0.5342	Yes
02/10/2015	041	0.649	0.761	5	1	8.6	0.5375	Yes
03/10/2015	069	0.928	0.374	2.7	0	8.6	0.5446	No
04/15/2015	105	0.972	–0.234	<1	0	8.3	0.2611	Yes
05/06/2015	126	0.826	–0.563	<1	0	8.4	0.3670	Yes
06/09/2015	160	0.378	–0.926	8.4	1	8.8	0.4544	No
07/07/2015	188	–0.095	–0.996	3.4	0	8.4	0.3305	Yes
08/04/2015	216	–0.545	–0.838	4.2	0	8.5	0.3556	Yes
09/08/2015	251	–0.924	–0.382	8.1	1	8.7	0.3061	No
11/09/2015	313	–0.780	0.625	2.7	0	8.6	0.2958	Yes
02/17/2016	048	0.735	0.678	1.3	0	8.9	0.4267	Yes
05/17/2016	138	0.693	–0.721	<1	0	8.6	0.3583	Yes
06/15/2016	167	0.264	–0.965	13.7	1	8.9	0.4640	No

References Cited

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources—Hydrologic analysis and interpretation: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p. [Also available at https://pubs.usgs.gov/twri/twri4a3/.]

Systat Software, Inc., 2008, SigmaPlot® 11.0 statistics user's guide: Systat Software, Inc., 564 p.

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–A10, accessed September 2016 at https://pubs.water.usgs.gov/twri9A.