Salinity App Methodology

Created by RTI International

November 2015

Overview

This application determines the probability that the salinity at four different monitoring stations within Sabine Lake will be outside the optimal monthly range of salinity for oyster health given user supplied conditions on streamflow, month, and lake level. The application works by fitting an ordinary least squares (OLS) linear regression model to the dataset of historic salinity concentrations at each station. The application then uses the fitted model to make predictions for new data. Rather than point predictions (i.e., single value predictions), the application creates simulated predictions; in other words, given the new data, what is the probabilistic distribution of salinity predictions conditional on the new data? A distribution of simulated predictions is created for each monitoring station. An empirical cumulative density function (CDF) is then calculated for each station from the distribution of simulated predictions. The probability of falling outside the optimal salinity range, which varies by calendar month, is then estimated from the CDF and returned within the application.

Data Source and Processing

All available post-2000 Texas Water Development Board (TWDB) salinity data were downloaded from https://www.twdb.texas.gov/index.asp for Sabine Lake between the dates 10/2000 and 9/2012. Four stations were available: SAB1, SAB2, SWBR, and USAB. The hourly salinity readings were averaged to produce mean daily salinity by Station ID.

A power transformation of 0.5168918 was then applied to the daily mean salinity data in order to make the response distribution approximately normal (**Figure 1**); response normality is not a requirement of linear modeling, but such a transformation often improves the diagnostics of the fitted model.

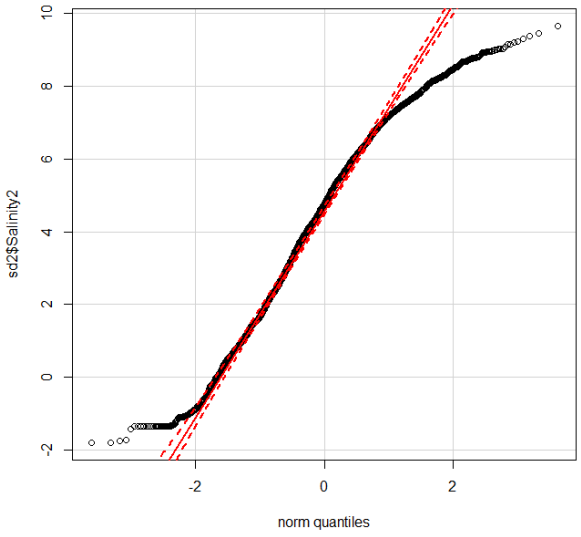

Figure 1. Q-Q plot of power-transformed daily mean salinity

Predictor Variables

Five predictor variables are used in the model. User inputs for new data are required for all of these except Station ID, since all stations are automatically modeled.

Station ID (Station_IDX): the TWDB salinity station ID

USGS flow at Neches Tidal (USGSFlow): http://nwis.waterdata.usgs.gov/tx/nwis/uv?site_no=08041780

Water level from NOAA buoy at Sabine Pass (WaterLevel): http://tidesandcurrents.noaa.gov/waterlevels.html?id=8770570&units=standard&bdate=19920101&edate=20131231&timezone=GMT&datum=MLLW&interval=hl&action=data

Sample Month (SampMonthXX): from TWDB time stamp

Toledo Bend Outflow (WFDam): from WaterFALL modeling simulation

Model Fitting

$$Mean\ Daily\ Salinity = \beta_0 + \beta_1 Toledo\ Bend\ Outflow + \beta_2 USGS\ Neches\ River\ Flow + \beta_{3i} Sample\ Month_i + \beta_{4i} Station_i + B_5 Sabine\ Pass\ NOAA\ Water\ Level + \varepsilon$$

Adjusted R-squared: 0.6348

Table 1. Model coefficients. P-value below 0.05 indicated statistical significance

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9.58E-01	1.51E+00	-0.637	0.524352
WaterLevel	4.35E-01	4.10E-02	10.616	< 2e-16
SampMonth02	1.23E-01	1.40E-01	0.877	0.38029
SampMonth03	-4.08E-01	1.32E-01	-3.089	0.002024
SampMonth04	-1.73E+00	1.34E-01	-12.938	< 2e-16
SampMonth05	-1.64E+00	1.36E-01	-12.067	< 2e-16
SampMonth06	-9.92E-01	1.33E-01	-7.436	1.31E-13
SampMonth07	-5.31E-01	1.31E-01	-4.046	5.33E-05
SampMonth08	1.21E-01	1.33E-01	0.909	0.363186
SampMonth09	4.12E-01	1.34E-01	3.07	0.00216
SampMonth10	2.04E-01	1.37E-01	1.486	0.137255
SampMonth11	5.81E-01	1.43E-01	4.076	4.69E-05
SampMonth12	2.33E-01	1.47E-01	1.581	0.113962
Station_IDSAB1	6.25E+00	1.50E+00	4.159	3.27E-05
Station_IDSAB2	7.97E+00	1.50E+00	5.308	1.18E-07
Station_IDSWBR	7.18E+00	1.50E+00	4.775	1.87E-06
Station_IDUSAB	5.62E+00	1.50E+00	3.733	0.000192
WFDam	-4.60E-05	6.97E-06	-6.603	4.65E-11
USGSFlow	-2.35E-04	5.35E-06	-43.799	< 2e-16

Table 2. Variance-inflation and generalized variance-inflation (GVIF) for model. Values indicate moderately low correlations. Common rule of thumb is a value <1 indicates no correlation; values between 1-5 indicate moderate correlation.

Variable	GVIF	Df	GVIF^(1/(2*Df))
WaterLevel	1.201	1	1.096
SampMonth	1.861	11	1.029
Station_ID	1.214	4	1.025
WFDam	1.389	1	1.179
USGSFlow	1.753	1	1.324

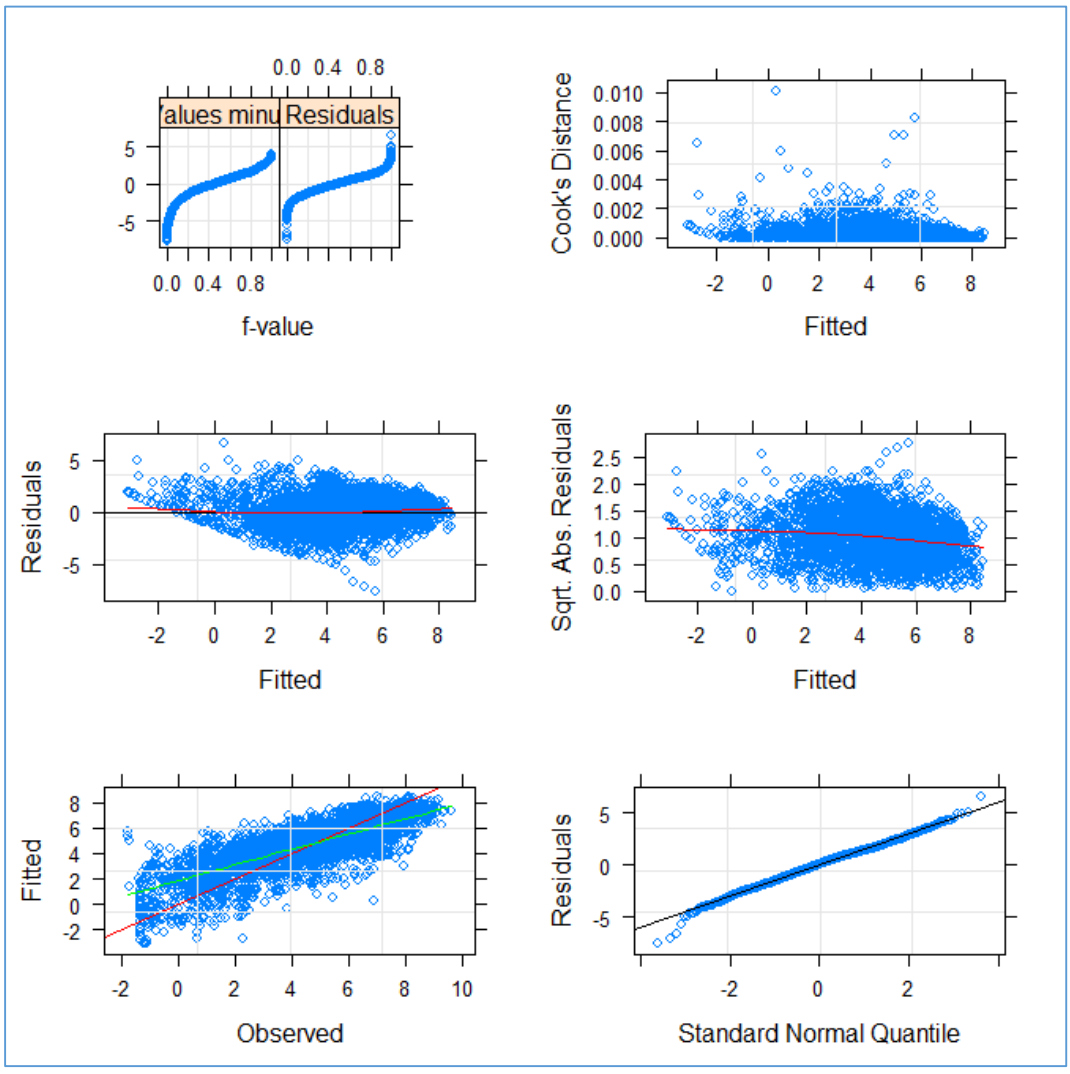

Figure 2. Diagnostic plots for fitted model. Checks for residual normality (bottom right); linear trend (bottom left); homoscedasticity and independence of errors (middle row); influential variables (upper right); residual-fit spread plot to visual R^2 (upper left).

Simulation steps

1. New data

The user specifies new data for each predictor variable in the model: USGS flow, Toledo Bend outflow, water level, and sample month. All TWDB Station IDs are included automatically. The upper and lower limits on user inputs are drawn from the observed data in order to minimize the risk of extrapolating beyond the data that was used to fit the regression (**Table 3**). Note: the min/max values for Toledo Bend outflow are calculated from the average min and max observed values for USGS tailwater gage at Toledo Bend rather than modeled estimates: http://waterdata.usgs.gov/usa/nwis/uv?site_no=08025350.

Table 3. Upper and lower limits of predictor variables

	Toledo Bend			Water Level			USGS 08041780 Neches River		
Month	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**
1	2354.08	7893.77	5815.46	-0.066	2.221	1.129	330	17400	6960
2	2422.62	9476.54	5735.54	-0.679	2.434	1.191	136	33300	10900
3	2892.31	17586.92	8840.08	-1.414	2.907	1.1005	242	34100	10150
4	1081.38	12049.00	4400.77	-0.896	3.077	1.1765	160	26200	6245
5	201.69	6676.92	4315.62	-0.361	2.333	1.065	38	21400	2895
6	198.92	5216.15	2926.58	-0.643	2.651	0.748	11	26800	2065
7	193.54	7764.62	5018.46	-0.732	2.552	0.7345	24	26800	2040
8	208.23	7239.23	4289.23	-1.181	2.651	0.719	305	8310	2560
9	227.85	7186.62	3663.15	-0.866	6.896	1.411	19	18100	2205
10	207.46	7321.23	1315.85	-0.348	2.838	1.319	32	15400	2730
11	1203.54	9068.62	2126.89	-0.472	3.061	1.4455	31	30600	4540
12	758.00	6485.31	3377.69	-0.892	2.943	1.281	208	27800	5815

2. Simulation

A modified version of the simulate() function in base R is then used to make 1000 estimates of daily mean salinity for each TWDB station given the predictor variable values specified by the user. In other words, given the following inputs:

Month: August
Water Level: 0.7
Toledo Bend Outflow: 4000
Neches River: 6000

what is the distribution of possible salinity values conditional on the observed relationships in the fitted linear model? Each station is therefore associated with a distribution of predicted daily mean salinity values.

Monthly optimum salinity thresholds are adopted from Chatry et al. (1983) and reported in **Table 4**.

Table 4. Original scale and power-transformed optimum monthly salinity ranges (Chatry et al., 1983).

Month	Lower	Upper	Lower Transformed	Upper Transformed
Jan	15	19	5.745967	6.717798
Feb	11	17	4.63325	6.246211
Mar	7	15	3.291503	5.745967
Apr	2	13	0.8284271	5.211103
May	4	12	2	4.928203
Jun	9	16	4	6
Jul	11	15	4.63325	5.745967
Aug	13	17	5.211103	6.246211
Sept	14	23	5.483315	7.591663
Oct	13	18	5.211103	6.485281
Nov	11	18	4.63325	6.485281
Dec	13	17	5.211103	6.246211

3. Empirical cumulative density function (CDF)

A CDF is then created for each station ID. A CDF is a curve that shows the probability that a random variable will have a value above or below a specified threshold (**Figure 3**). X axis is power transformed salinity values. Y axis is the probability of an observations being at or below a given salinity value. For instance, the probability that a salinity value will be below the maximum observed/predicted value is 1, or 100%.

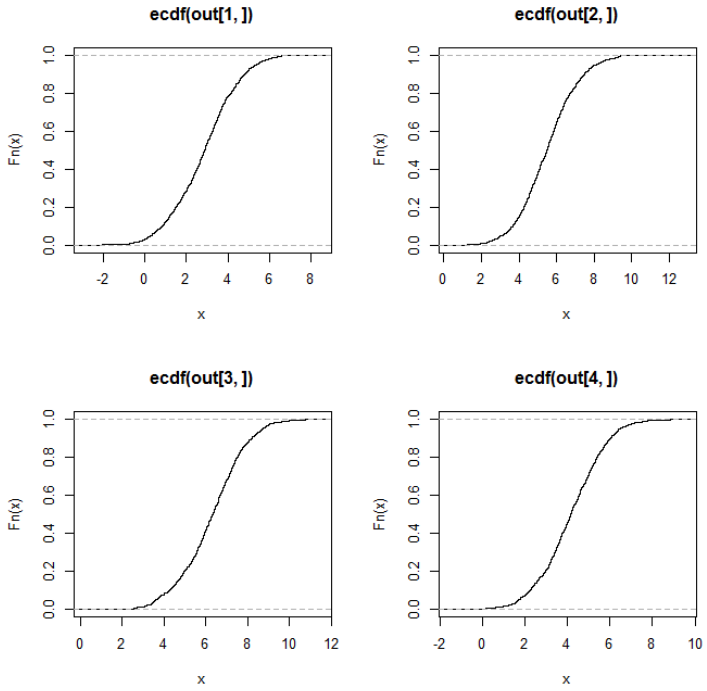

Figure 3. Empirical CDFs for simulated salinity predictions at each TWDB station.

The CDFs are used to calculate the proportion of observations that fall above and below the optimum monthly salinity range. For instance, if the optimum salinity range in a given month is 2-4 (in transformed units), the CDF can be used to find the proportion of observations above 4 (1 – cdf(4)) and below 2 (cdf(2)). The proportion of observations above and below the monthly optimum is then returned in a table.

For example, for the following user inputs reflect very high inflows and low tidal water levels to the lake in the month of October:

Month=10
WL= -0.3
Dam = 7300
USGS = 15000

The simulated salinity predictions indicate that under these high flow conditions, lake salinity conditions are likely to be below the lower optimum salinity range; this reflects dilution from high inflows and lower salinity inputs from the coast due to lower water levels (**Table 5**).

Table 5. Example output of the application.

Station	Upper Exceedance	Lower Exceedance
USAB	0.000	0.997
SWBR	0.004	0.974
SAB2	0.013	0.928
SAB1	0.000	0.993