MWRD PAA - Preliminary Data Analysis

Version 2

Kate Newhart

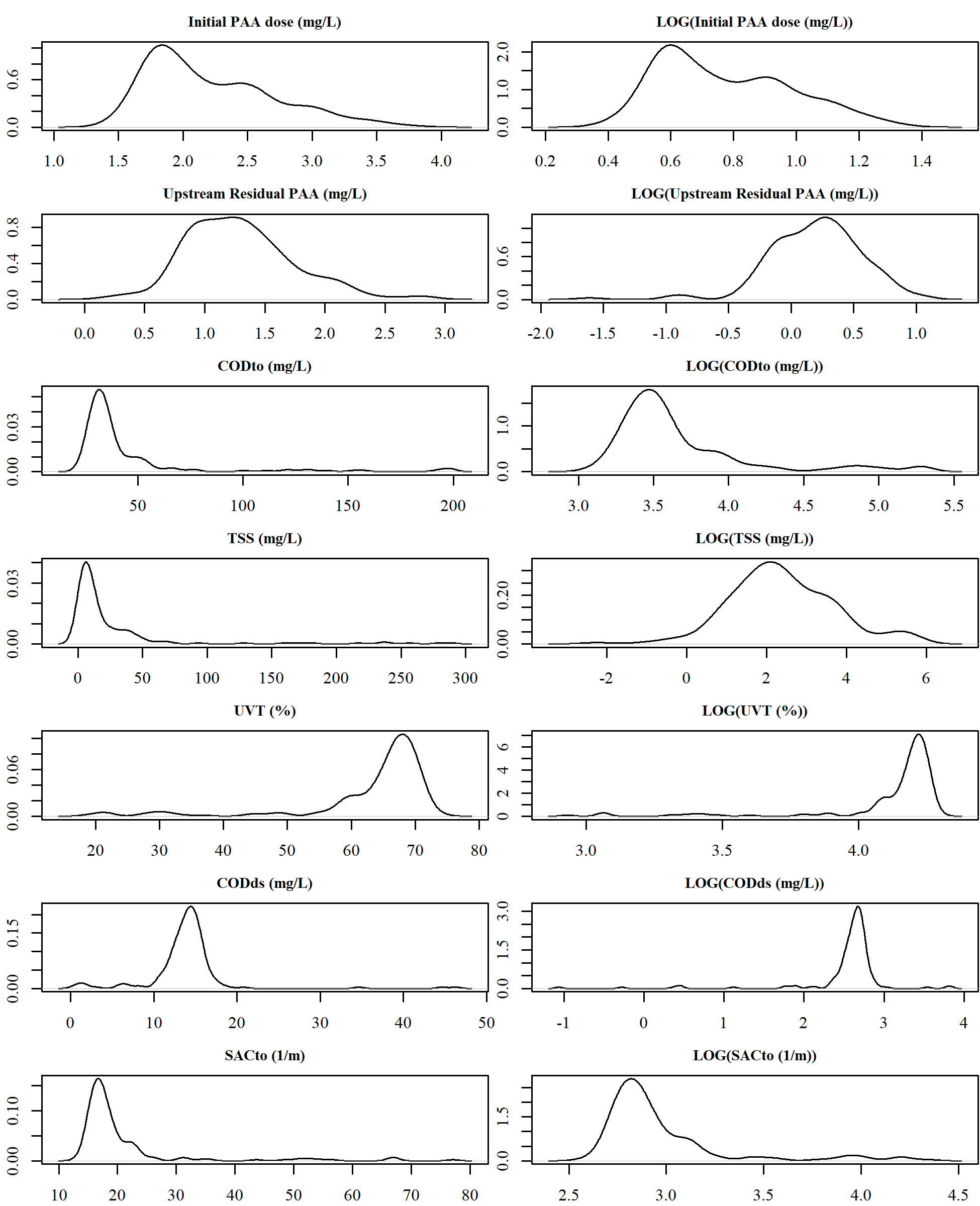
4/9/2019

# Introduction

Previous analyses (version 1, 2019-03-28) demonstrated no substantial link between *E. coli* concentrations entering the PAA disinfection system and other effluent water quality parameters measured online in the North Secondary system.

In this document, online water quality parameters are used to predict PAA residual immediately post-dosing (i.e., “upstream residual”). The initial PAA concentration is calculated based on a flow-mass balance from the PAA dosing pump rate and influent disinfection basin flowrate.

Similar to previous analysis, very few of the measured water quality variables at the PAA dosing point (residual PAA, COD, TSS, UVT, and SAC) exhibit a normal distribution (**Figure S1**). This will be a source of error in many statistical methods. To minimize this error, a log transform of all variables was performed to better approximate a normal distribution (**Figure S1**). However, there limited improvement of the log transform in the distribution of each variable due to the highly skewed nature of the data (**Figure 1**). In future iterations of data analysis will include calculating the skewness and exploring additional transformations to account for the wide range of values for each process variable, or narrowing the range of values for each process variable to approximate a normal distribution.



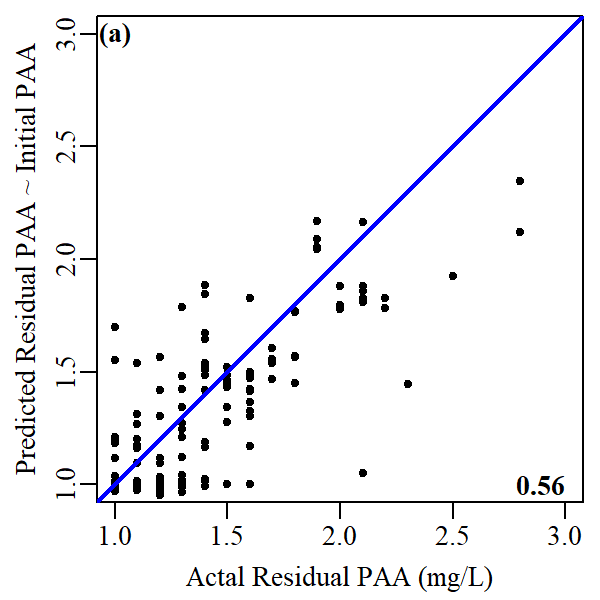
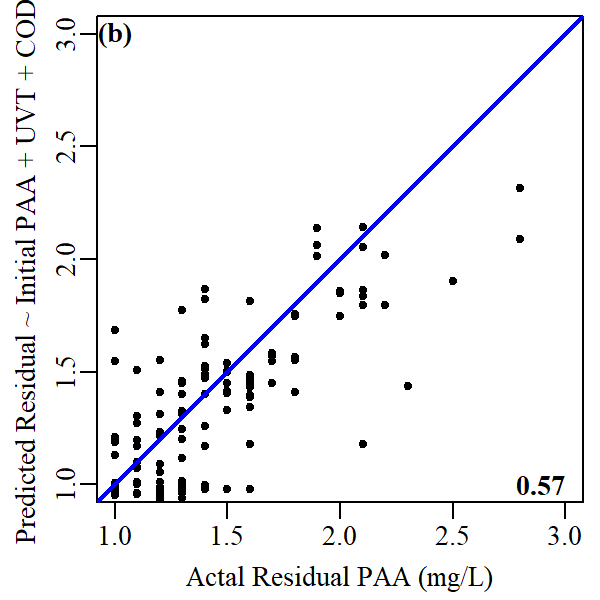
**Figure 1.** Kernel density plots of actual and log transformed online PAA data and Carbovis data. The vertical axis is frequency (i.e., density) of observations and the horizontal axis is the value of the observations.

# Methods

## Linear regression

In simple linear regression, a model is constructed of a response variable (**Y**) that is a linear function of other variables (*xi*). The linear regression model assumes that **Y** is normally distributed, errors are normally distributed and independent, and **X** has constant variance.

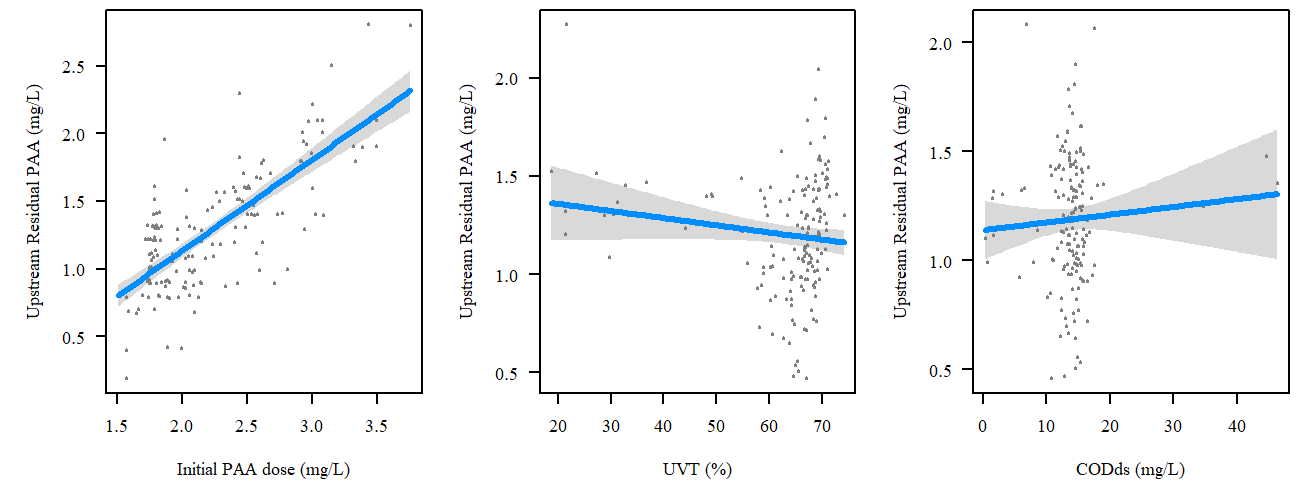
**Figure 2** shows a strong relationship between initial PAA dose and the upstream residual, but linear impacts of other online water quality variables are minimal. Ambient temperature was originally included, but did not improve model fit. Log transformed variables performed substanically worse for both linear model fit scenerios (from **Figure 2**, scenerio (a) r2 = 0.46 and (b) r2 =0.47)

**Figure 2.** Linear model fit for upstream residual PAA given (a) the initial dosing concentration of PAA and (b) the initial dosing concentration of PAA, UV transmitance, and CODds. Black circles represent actual observations. Blue line represents a perfect model fit. R-squared value in lower right.

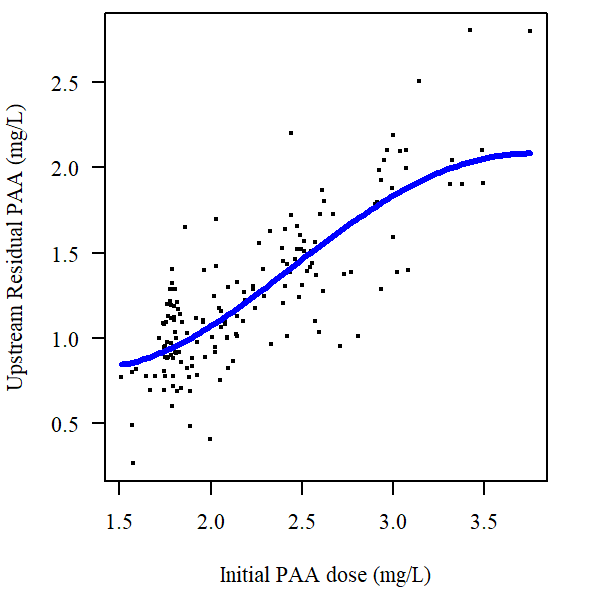
## Generalized Linear Model

The general linear model (LM) requires that the response variable follows the normal distribution whilst the generalized linear model (GLM) is an extension of the LM that allows the specification of models whose response variable follows different distributions. No difference was found between a LM fit with the initial calculated PAA dose, UVT, and CODds and a GLM fit with the same predictor variables.



**Figure 3.** GLM model constructed for upstream residual PAA concentration (mg/L) calculated from initial calculated PAA dose, UVT, and CODds.

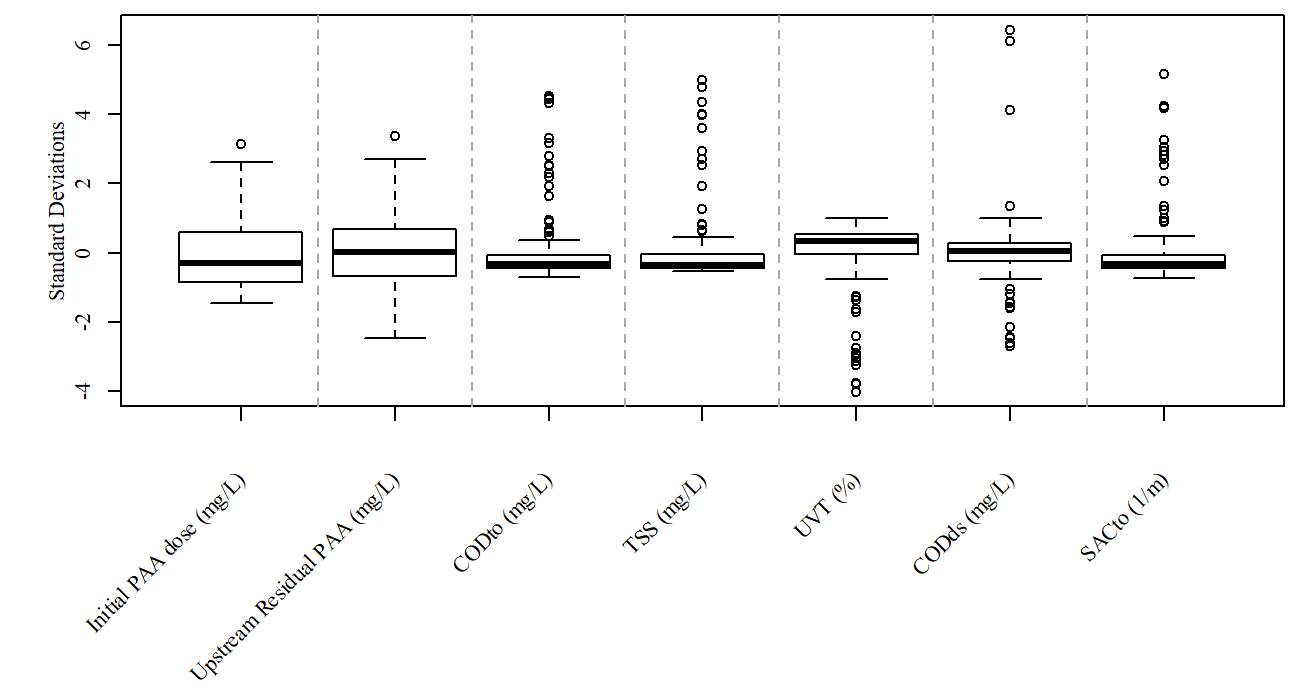
## Non-Regression Model



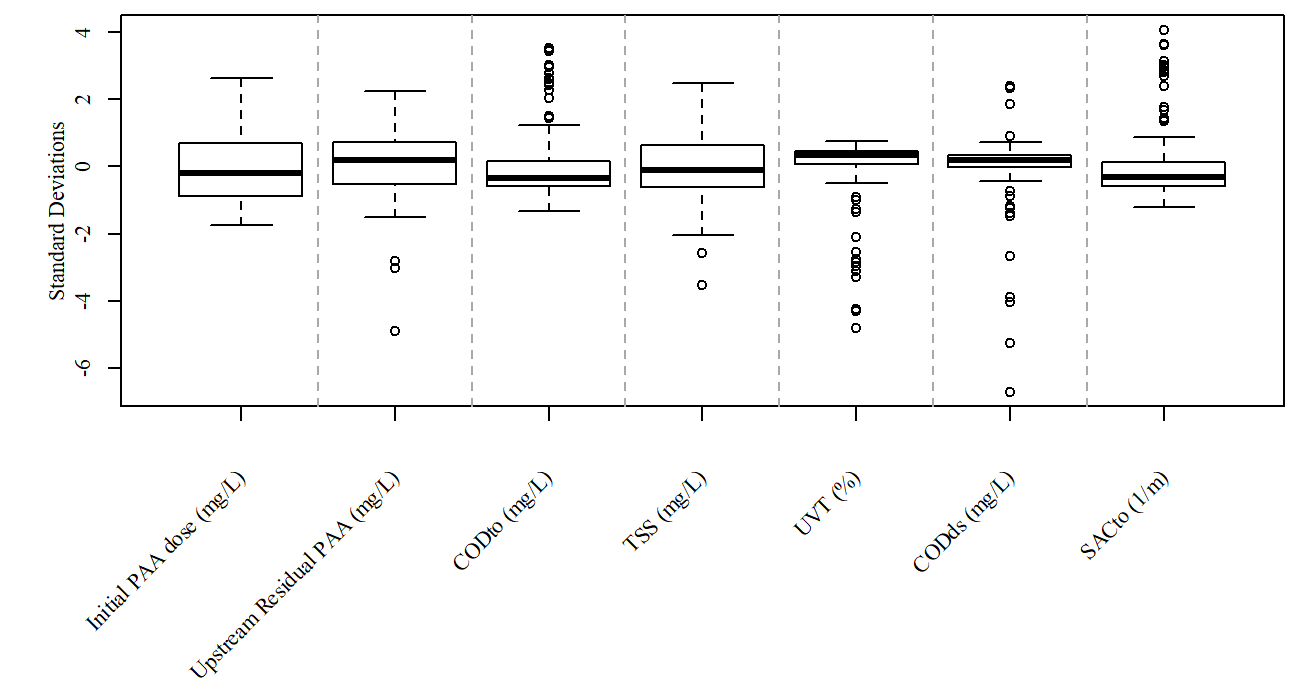
**Figure 4.** Nonregression model (support vector machine) for predicting upstream residual PAA shows a nonlinear behavior

# Supplementary Material

## Figures



**Figure S1.** Scaled boxplot of online PAA data and Carbovis data.



**Figure S2.** Scaled boxplot of log transformed online PAA data and Carbovis data.