LIFT Challenge 2019 - Preliminary Data Analysis

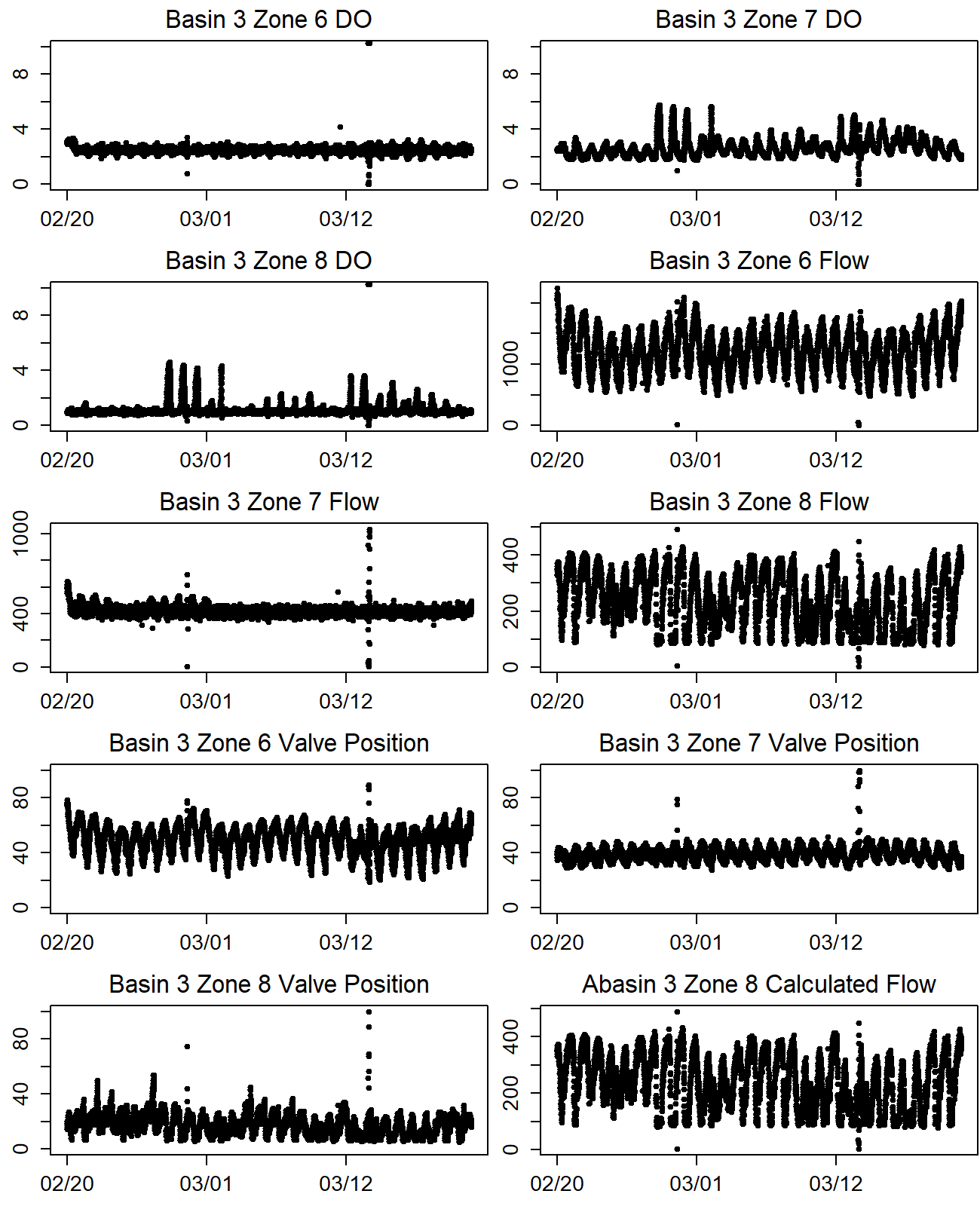
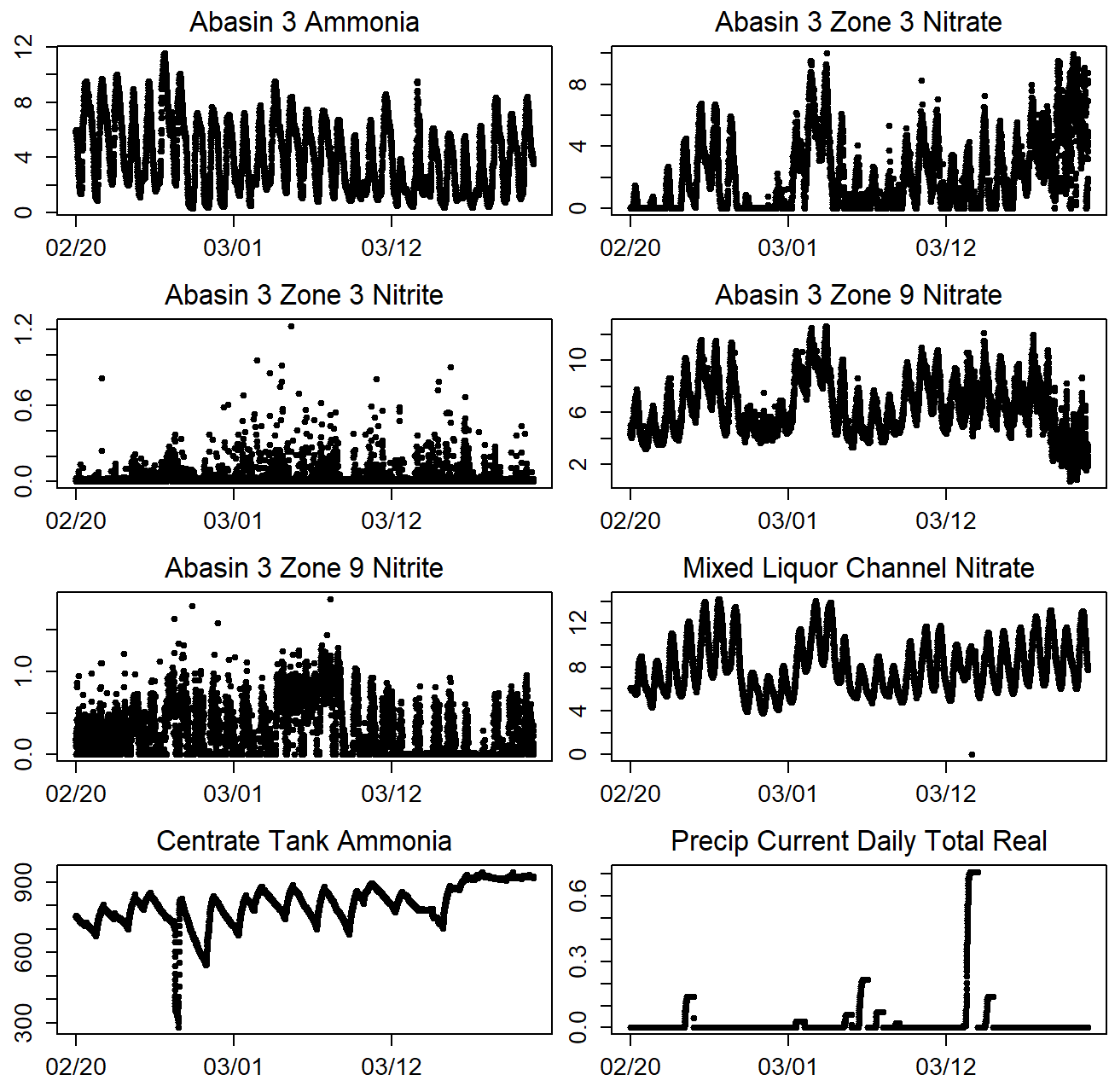
Kate Newhart

4/15/2019

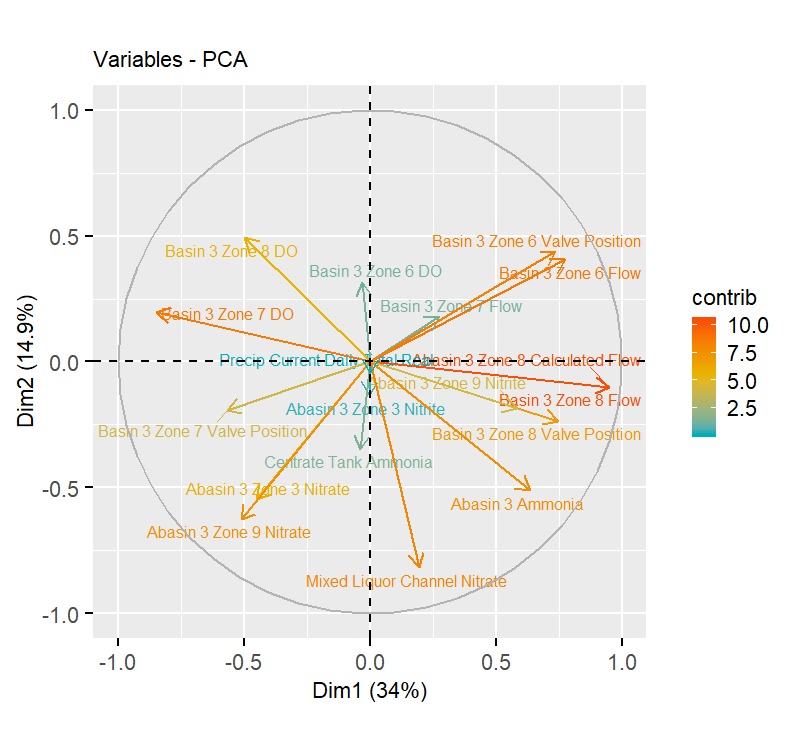
For data-driven analysis, a *clean* dataset is required to appropriately train a model in order to (i) identify relationships between variabels and (ii) define in-control and out-of-control conditions. The first step in data cleaning is the visual inspection of timeseries plots. Wastewater treatment data provided by City of Boulder, Colorado is plotted in Figure 1, minus process variables that did not change (e.g., variables from Zone 9) or did not have numerical data (e.g., influent flow) throughout the window of time provided (2019-01-01, 2019-03-21). However, a shift in flows at the end of February led to subsetting data to 2019-02-20, 2019-03-20. Is it okay to subset to this range? Or do we want to predict under February’s conditions as well?

Principal component analysis (PCA) shows that the majority of the variation present in the data is governed by flow and dissolved oxygen (DO). Which variable from the list are we most interested in predicting?

A strict 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 for models whose response variable follows different distributions. Generalized additive models (GAM) is an additive modeling technique where the impact of the predictive variables is captured through smooth functions which-depending on the underlying patterns in the data-can be nonlinear. Variables that appeared to trend with Abasin 3 Ammonia are in Figures 3 and 4.

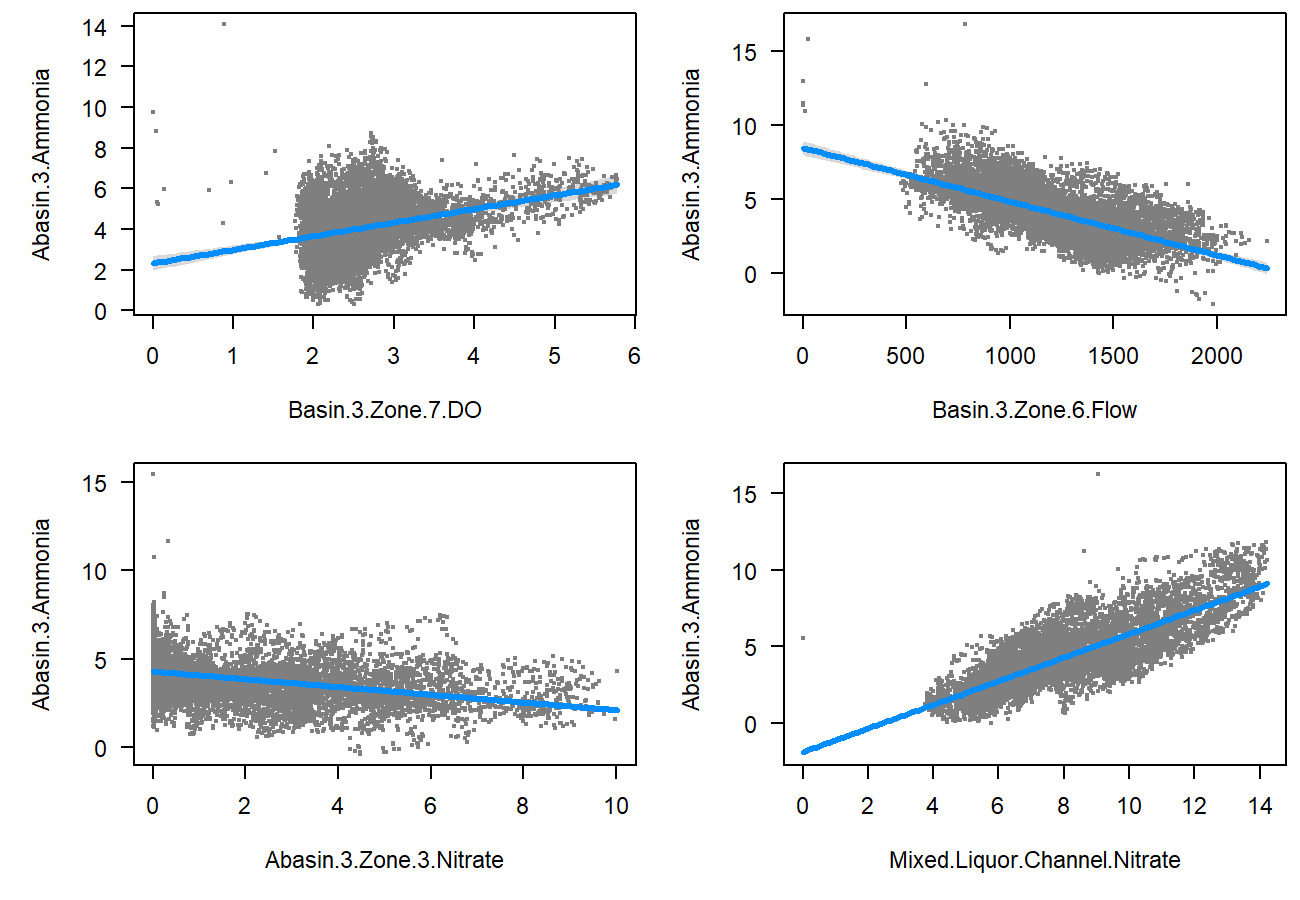
**Figure 1.** Process data from the City of Boulder, Colorado municipal wastewater treatment facility.



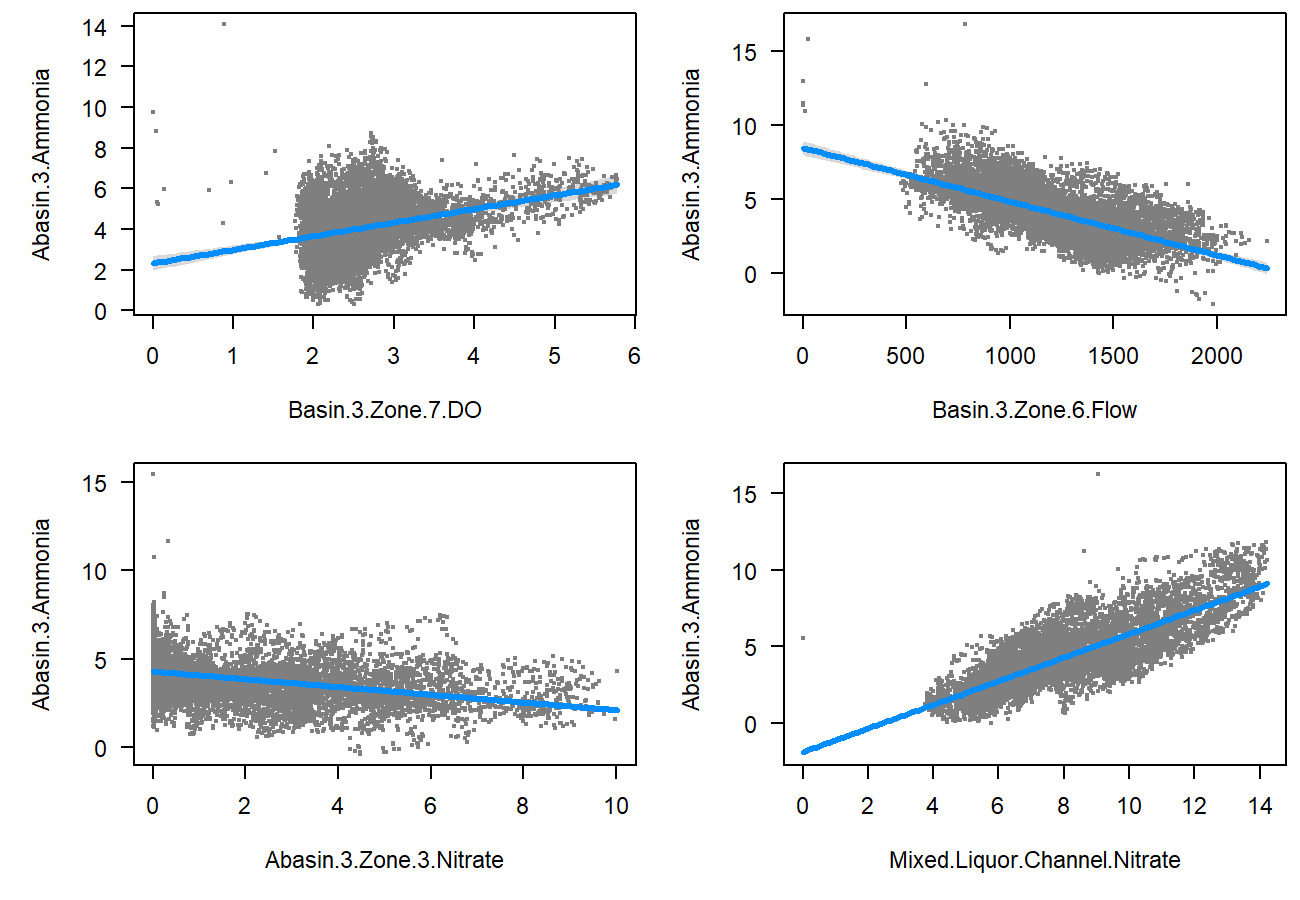
**Figure 2.** Principal component analysis of process data from the City of Boulder, Colorado municipal wastewater treatment facility between 2019-02-20, 2019-03-20.

**Table 1.** PCA variable contributions to City of Boulder, Colorado municipal wastewater treatment facility

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 | Dim.6 |
| Basin 3 Zone 8 Flow | 14.7 | 0.4 | 0.2 | 0.2 | 0.9 | 0.2 |
| Abasin 3 Zone 8 Calculated Flow | 14.7 | 0.4 | 0.2 | 0.2 | 0.9 | 0.2 |
| Basin 3 Zone 7 DO | 11.8 | 1.4 | 0.3 | 0.1 | 0.9 | 1.9 |
| Basin 3 Zone 6 Flow | 9.8 | 6.1 | 8.9 | 1.5 | 1.2 | 1.6 |
| Basin 3 Zone 8 Valve Position | 9.1 | 2.1 | 7.9 | 1.1 | 0.1 | 0.4 |
| Basin 3 Zone 6 Valve Position | 8.9 | 7.1 | 3.8 | 0.0 | 1.8 | 0.2 |
| Abasin 3 Ammonia | 6.6 | 9.8 | 2.7 | 0.2 | 3.3 | 0.2 |
| Abasin 3 Zone 9 Nitrite | 5.6 | 1.3 | 5.1 | 8.3 | 0.1 | 0.2 |
| Basin 3 Zone 7 Valve Position | 5.3 | 1.4 | 14.7 | 5.3 | 0.1 | 1.7 |
| Abasin 3 Zone 9 Nitrate | 4.3 | 14.7 | 0.1 | 0.1 | 1.3 | 0.1 |
| Basin 3 Zone 8 DO | 4.1 | 9.1 | 0.8 | 2.5 | 0.6 | 1.3 |
| Abasin 3 Zone 3 Nitrate | 3.3 | 11.2 | 11.1 | 1.4 | 1.3 | 1.0 |
| Basin 3 Zone 7 Flow | 1.2 | 1.2 | 0.6 | 45.7 | 9.6 | 24.9 |
| Mixed Liquor Channel Nitrate | 0.6 | 24.8 | 3.4 | 0.0 | 0.1 | 0.4 |
| Centrate Tank Ammonia | 0.0 | 4.5 | 28.0 | 0.0 | 12.1 | 3.1 |
| Basin 3 Zone 6 DO | 0.0 | 3.7 | 2.8 | 6.3 | 5.4 | 2.4 |
| Abasin 3 Zone 3 Nitrite | 0.0 | 0.7 | 2.3 | 18.4 | 4.1 | 60.1 |
| Precip Current Daily Total Real | 0.0 | 0.1 | 7.1 | 8.6 | 56.4 | 0.0 |



**Figure 3.** GLM model for predicting Abasin 3 Ammonia.



**Figure 4.** GAM model for predicting Abasin 3 Ammonia.