INTELLIGENT WATER SYSTEMS CHALLENGE 2019

DEVELOP ADVANCED MODELS TO PREDICT AND OPTIMIZE CHEMICAL DOSAGE FOR ODOR AND CORROSION CONTROL

AT JAMES C. KIRIE WATER RECLAMATION PLANT

1 TEAM

A collaboration between the Metropolitan Water Reclamation District of Greater Chicago (MWRDGC) and Illinois State University (ISU).

i. **Fenghua Yang**, PE, BCEE, Senior Environmental Research Scientist, MWRDGC (Team leader and Utility Research), specializes in energy efficient wastewater treatment process, online process control, biological nutrient removal, and resource recovery. She currently serves in the WERF Low Carbon, Low Energy Nutrient removal project, and IWA Meta-Data Collection and Organization in Wastewater Resource Recovery Systems project. For the past few years, she led several MWRDGC research projects on process improvement for energy efficient nutrient removal. She is currently involved in evaluation of new and innovative technologies to improve water and resource recovery at the MWRDGC.

ii. **Thaís Pluth**, Environmental Research Scientist, MWRDGC (Utility Research and Data Management), evaluates technologies and processes for water resource recovery and solids treatment. She has extensive background in applying statistical models to understand environmental processes.

iii. **Matt Jurjovec**, PE, Operations Manager, MWRDGC (Utility Operation), specializes in treatment operations, plant automation and optimization, collection systems management, process control and instrumentation. Matt currently serves as Operations Manager at the Kirie WRP where he manages a team of operators for compliance with NPDES regulation and oversees maintenance needs at the facility. Matt has over 11 years of experience with the MWRDGC where he has worked in various roles related to wastewater treatment, stormwater collection and project development.

iv. **Dr. Xing Fang**, Assistant Professor in Computer Science, ISU (Data analytics leader), specializes in deep learning and big data analytics. He has an extensive background in deep learning related applications and machine learning algorithms.

v. **Dr. Yongning Tang,** Professor in Computer Science, ISU (Data Analytics Co-lead), has more than 20 years of intelligent system design experience. Among his previous research and industrial collaborations, he has successfully applied various artificial intelligence technologies into different problem domains, such as intelligent computing for sustainable energy and environment, and intelligent network operations and management. He currently serves as member of the Board of Directors for several local and international research and educational associations. http://www.itk.ilstu.edu/faculty/ytang/project-pub.html

vi. **Kyle Bradley Francq**, Student in Computer Science, ISU (Data Analytics), specializes in data analysis and computer programming.

2 PLAN

Our goal is to use 17 years of plant influent characteristic data, operational data, weather data, 6 months of H2S sensors data, and 10 weeks of NaOCl dose-response data, to train a series of properly selected machine learning models to predict and optimize the amount of chemical dose for corrosion/odor control without impacting downstream treatment process.

2.1 Problem Statement

The James C. Kirie Water Reclamation Plant (Kirie WRP) is one of seven treatment facilities owned and operated by MWRDGC. Kirie treats an average of 52 million gallons per day of municipal sewage from several communities in the northwestern Chicago metropolitan area. The Kirie WRP has historically dosed sodium hypochlorite (NaOCl) at its headwork to help curtail hydrogen sulfide (H_2S) production and consequently odor complaints and infrastructure corrosion. Routine NaOCl dosing at the headwork was suspended in May 2015 after the Kirie WRP initiated an enhanced biological phosphorus removal (EBPR) program to meet the anticipated phosphorus limits. The concern was that the oxidizing chemical could potentially reduce the volatile fatty acids (VFAs) in the influent, which are necessary in the EBPR process. However, the actual impact of NaOCl on the VFA level was never assessed. Currently, there is no online VFA monitoring instrument available and the waiting time for VFA laboratory results usually is 45 days or longer, which makes it difficult for the day-to-day decision making process for choosing the optimal NaOCl dosage.

Figure 1. Kirie WRP Headworks Infrastructure Corrosion

Since the suspension of NaOCl dosing, even though only a few odor complaints were received, the plant has observed increasing headworks infrastructure corrosion. Figure 1 shows the equipment corrosion on top of the flow split chamber. The real concern is the concrete corrosion inside the chamber could be worse, but unobservable from outside. H_2S spikes at the headwork structures were first recorded during the three weeks H_2S monitoring in 2017. The operators of the plant prefer to continue the use of existing NaOCl system for H_2S control because Kirie WRP maintains the chemical onsite for effluent disinfection. The necessary infrastructure is already installed, and the plant is satisfied with the effect of the NaOCl dosing in the past. The challenge here is how to reduce elevated H_2S emissions and potential corrosion issues at the Kirie WRP headworks by the use of NaOCl without the need for capital improvements.

The generation of H_2S and VFAs is impacted by many factors, including wastewater characteristics, plant flow, and operational conditions. H_2S emission is also impacted by factors such as pH, temperature, and turbulence caused by pumping, among others. Furthermore, the effectiveness of NaOCl dosing for H_2S control is impacted by the plant influent characteristic and operational conditions. While under-dosing will cause insufficient corrosion control, overdosing will waste chemicals, and consume VFAs which could negatively impact EBPR performance.

The purpose of this project is to develop a series of appropriate machine learning models based on the influent characteristics and plant operational conditions to predict H_2S and VFAs levels and then estimate the optimal NaOCl dose to maintain H_2S level lower than 5 ppm in the headspace of the influent flow split chamber, without significantly reducing the influent VFA levels. NaOCl dosing will be stopped when VFAs levels drop below 5 ppm in the plant influent channel.

2.2 Characterization of the Intelligent Water System

The data sets used during this challenge (detailed in Appendix B) include:

- ***Online Instrument Data***: routine plant monitoring data, include flow, Oxidation and Reduction Potential (ORP), pH, wastewater temperature, and elevation data for the connecting tunnel system.
- ***Water Quality Analysis Data:*** influent data regarding total solids (TS), suspended solids (SS), biochemical oxygen demand (BOD_5), total phosphorus (TP), ammonia (NH_3), sulfate (SO_4), organic nitrogen (Org-N), total Kjeldahl nitrogen (TKN), and VFAs.
- ***Odor and Corrosion Monitoring Data:*** H_2S monitoring data from an OdaLog sensor.
- ***Precipitation data:*** daily rainfall data.
- ***NaOCl dose:*** amount of NaOCl dosed per day during the dose-response test.

In the past, NaOCl was dosed to maintain an ORP in the range of 0 to +50 mV, which was believed to be able to control raw sewage odors and corrosion. However, as discussed above, the H_2S generation and emission is impacted by many factors. Dosing based only on an ORP setup could cause under-dosing, leading to insufficient corrosion control; or overdosing, which will consume VFAs and could negatively impact EBPR performance and increase chemical use. Appendix A shows that even with an ORP above 50 mV, there were H_2S spikes. Additionally, when ORP was negative, the H_2S concentration was 0 and ≤5ppm in 87% and 97% of the time, respectively. This demonstrates that dosing based only on ORP is not effective. Between 2010 and 2014, the average NaOCl dosage at Kirie WRP was 33,000 gallons/year. If the system was set to dose when the H_2S level was greater than 5 ppm, the chemical saving during that time period would have been 97%, representing a cost saving of $24,330 per year.

2.3 Plan

The plan of the project is to design, develop, and deploy a series of machine learning models to forecast information that can be utilized to determine the amount of NaOCl that should be applied in order to control the odor/corrosion at Kirie WRP headwork without negatively impacting the downstream EBPR performance. Our approach consists of: (1) data preprocessing and quality assurance/quality control (QA/QC) from February to May 2019; (2) development of Module 1 to forecast wastewater characteristics in May 2019; (3) development of Module 2 to predict H_2S and VFAs from May to June 2019; (4) testing the data analysis models of Modules 1 and 2 in June 2019; (5) full scale NaOCl dose-response test to generate data for Module 3 from May to July 2019; (6) development of Module 3 to estimate the NaOCl optimal dosage from June to July 2019; (7) testing Module 3 in July and August 2019; (8) developing action plans and training Kirie WRP personnel to use the developed system to predict their NaOCl dosage for odor/corrosion control from September 2019 to February 2020; and (9) full scale implementation and deployment from November 2019 to October 2020. All models will be developed in Python, which is an open source free software, using free software packages and libraries (e.g. Tensor Flow, Keras).

3 SOLUTION AND IMPLEMENTATION

Data pre-processing and exploration are followed by training and testing of a series of supervised machine learning algorithms. Results will be used to determine the proper NaOCI dosage for odor/corrosion control.

3.1 Implementation of Module 1

3.1.1 Data, data preprocessing, and QA/QC

Detailed information can be found in Appendix B. In this module, only the daily wastewater dataset was used. Kirie WRP has daily wastewater data available since 1997. However, data visualization showed a significant wastewater characteristic change up to 2001 due to the reduction in industrial wastewater.

Therefore, we decide to use data from January 1, 2002 to December 31, 2018. Correlation analysis shows that certain wastewater parameters are positively correlated to H_2S, namely TS, SS, BOD_5, Org-N, TKN, and TP. For most of these parameters, there is a lag of 4 days to get the testing results back from the laboratory. For BOD_5 the lag can be up to 7 days. Because of these lags, the information is not available for the plant daily planning. To address this problem, Module 1 was developed to forecast current wastewater characteristics based on the 7-day (for BOD_5) and 4-day (for other parameters) lag periods. In this dataset, the missing values were filled by propagating the non-missing values backward along the time series. All the variables were normalized prior to be used in this module. Normalization is a good technique to use when the distribution of the data is not known or when the distribution is not Gaussian. In addition, data normalization can enhance the performance of various machine learning models.

3.1.2 Analysis and interpretation

Three data analysis models were developed to forecast wastewater characteristics: Recurrent Neural Network (RNN) using bidirectional long short-term memory, Autoregressive Integrated Moving Average (ARIMA), and Random Forest (RF). Table 1 compares the best results of each method through their corresponding Mean Absolute Error (MAE). MAE is the average difference between the predicted and the true values. The MAE for all three models are comparable, but the RNN model was selected for Module 1 to predict influent characteristics. The figures showing the actual and predicted values for BOD_5 (which had the lowest MAE using the RNN model), NH_3 (which had the highest MAE using the RNN model), and TS (the most important parameter predicting H_2S according to the model 2) can be seen in Appendix C.

Table 1 - Mean absolute error for different methods predicting module 1 variables

Variable	**RNN**	**ARIMA**	**RF**
NH_3	0.06	0.07	0.04
BOD_5	0.01	0.01	0.01
TS	0.03	0.03	0.03
Org-N	0.02	0.03	0.03
TP	0.02	0.03	0.03
SS	0.02	0.02	0.03
TKN	0.02	0.03	0.03
SO_4	0.03	0.02	0.03

3.2 Implementation of Module 2 – prediction of H_2S and VFA

3.2.1 Data, data preprocessing, and QA/QC

Detailed information can be found in Appendix B. To develop the H_2S and VFA prediction module, three datasets are used as input: (i) water quality analysis data (TS, SS, TP, NH_3, SO_4, BOD_5, Org-N, and TKN), (ii) online instrument data (flow, ORP, pH, wastewater temperature, tunnel pumping, and tunnel elevation), and (iii) precipitation data. The H_2S and VFA data were used as labels in both the H_2S model and the VFA model, respectively. The dataset used in the H_2S model was from the two periods with H_2S monitoring: July 14, 2017 to August 3, 2017 and from March 7 to April 30, 2019. To develop the VFA model, data from April 1, 2015 to May 13, 2019 were used.

One of the major challenges faced during this project was the compilation of disparate data types, formats, and frequencies into a single data matrix with correct information. For instance, wastewater

variables are sampled daily, whereas the online instrument data are available both daily and every 15 minutes. Therefore, two different solutions for predicting the H_2S concentration were developed: one in a 15-minute interval and another in a daily interval. To predict VFA, only solution 2 is used since VFA is only available daily.

- Solution 1 – 15-minute interval matrix: data only available in daily intervals were duplicated in order to allow them to pair with the data belonging to the other variables that were sampled every 15 minutes.
- Solution 2 – daily interval matrix: the daily maximum is used for variables available in 15-minute intervals.

3.2.2 Analysis and interpretation

Classifiers were developed to predict the H_2S and VFA concentration given certain input data. The H_2S data ranged from 0 to 140 ppm and were divided into 16 classes as shown in Appendix D. However, class 12 had no data and class 0 had 99% of the data. The VFA data ranged from 0 to 102 mg/L and was divided in 12 classes, with no data in class 9 (Appendix D). To deal with the imbalanced data, random oversampling was employed. This technique allows creating more data points for the classes that have a few points and overall improves the prediction.

Two machine learning models were tested in Module 2: RF and Support Vector Machine (SVM). These two are chosen because they are reported as the best classifiers for not very large datasets. For both models, 80% of the data are randomly selected for training and the remaining 20% of the data are selected for testing. Table 2 shows the results of the models developed to predict H_2S and VFA. For H_2S, the RF model had a slightly better accuracy than the SVM model when using the 15-minute data. When using the daily data, the SVM model performed better and had an accuracy of 97.62%. For VFA, the RF model was more accurate (93.39%) when compared with the performance of the SVM model (91.03%).

Table 2 – Models for predicting H_2S and VFA

Output	**Method**	**Data interval**	**Number of actual classes**	**Classes without data**	**Sample size per class after oversampling**	**Accuracy in testing set (%)**	**Input attributes**
H_2S	RF	15 minute	15	12	7,503	87.56	ORP, pH, temperature, flow, rainfall, tunnel elevation, tunnel pumping, total solids, SS, BOD_5, TP, TKN, Org-N, NH_3, SO_4
	SVM					86.32	
	RF	daily	8	6 - 14	43	85.71	
	SVM					**97.62**	
VFA	RF	daily	11	9	55	**93.39**	
	SVM					91.03	

A problem faced with both the H_2S and VFA predictions was the lack of data in some predicting classes, which could be a problem since the models cannot be trained for those classes. To solve this issue, more data are expected to be collected and available in those missing ranges. Appendix E shows the order of importance of variables in the RF model predicting maximum daily H_2S and VFA levels.

3.3 Implementation of Module 3 – estimation of NaOCl optimal dosage

3.3.1 Data, data preprocessing, and QA/QC

Detailed information can be found in Appendix B. To develop the NaOCl prediction module, five datasets were used as input: (1) predicted H_2S data from module 2, (2) predicted VFAs data from module 2, (3) target H_2S levels, (4) target VFAs level, and (5) plant flow data. Two variables are introduced: delta H_2S (the difference between the predicted and the target H_2S level), delta VFAs (the difference between the predicted and the target VFAs level). The output of Module 3 is the predicted daily optimal NaOCl dosage, which was labeled during training using the results of the dose-response test.

In order to train the learning model in Module 3, a dose-response test was conducted between 5/14/19 – 7/25/19. We attempted to adjust dosage twice a week, then monitor the H2S and VFAs response data, and modify dosage accordingly. The NaOCl dosage ranged from 0 to 238 gallon per day. The major challenge faced by this module is the lack of monitored VFA data. While we have daily H_2S until July 25, 2019, VFAs data were only available until May 29, 2019 due to the long laboratory waiting time for VFAs. The missing monitored VFA data were estimated based on an observed correlation with the monitored H_2S and VFAs levels (Appendix F).

3.3.2 Analysis and interpretation

Several classifiers are developed to predict the NaOCl dosage. The NaOCl dosage is divided into 24 classes as shown in Appendix G. Two machine learning models were tested in Module 2: namely RF and SVM. For both models, 80% of the data are randomly selected for training and the remaining 20% of the data are selected for testing. Table 3 shows the results of the models developed to predict the NaOCl dosage. The RF model had a better accuracy than the SVM model. 89.75% of the testing data could accurately be predicted by the RF model. Appendix H shows the order of importance of variables in the RF model predicting the NaOCl dosage. Plant flow played the major role. We will re-test Module 3 when more monitored VFA data become available.

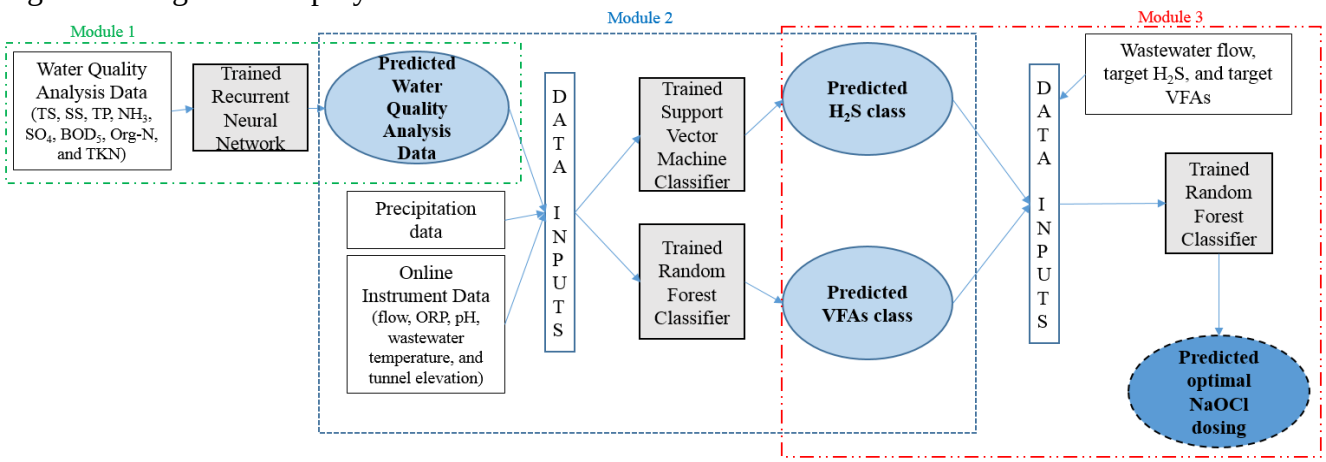
Table 3 – Models for predicting the NaOCl dosage

Output	**Model**	**Data interval**	**Number of actual classes**	**Classes without data**	**Sample size per class after oversampling**	**Accuracy in testing set (%)**	**Input attributes**
NaOCl Dosage	RF	daily	24	10	46	**89.75**	Predicted VFAs and H_2S data from module 2, target VFAs, target H_2S, and plant flow rate
	SVM					68.75	

3.4 Communication, use, and next steps

The final algorithm for Modules 1 through 3 will be deployed according to Figure 2. The predicted water quality analysis data (result of Module 1), together with online instrument data and rainfall data will be inputs to Module 2. The predicted H_2S and VFAs results of Module 2, combined with the target H_2S and VFAs levels, wastewater flow data, will be inputted to Module 3 to predict the optimal NaOCl dosing.

Figure 2 – Algorithm deployment for the final solution



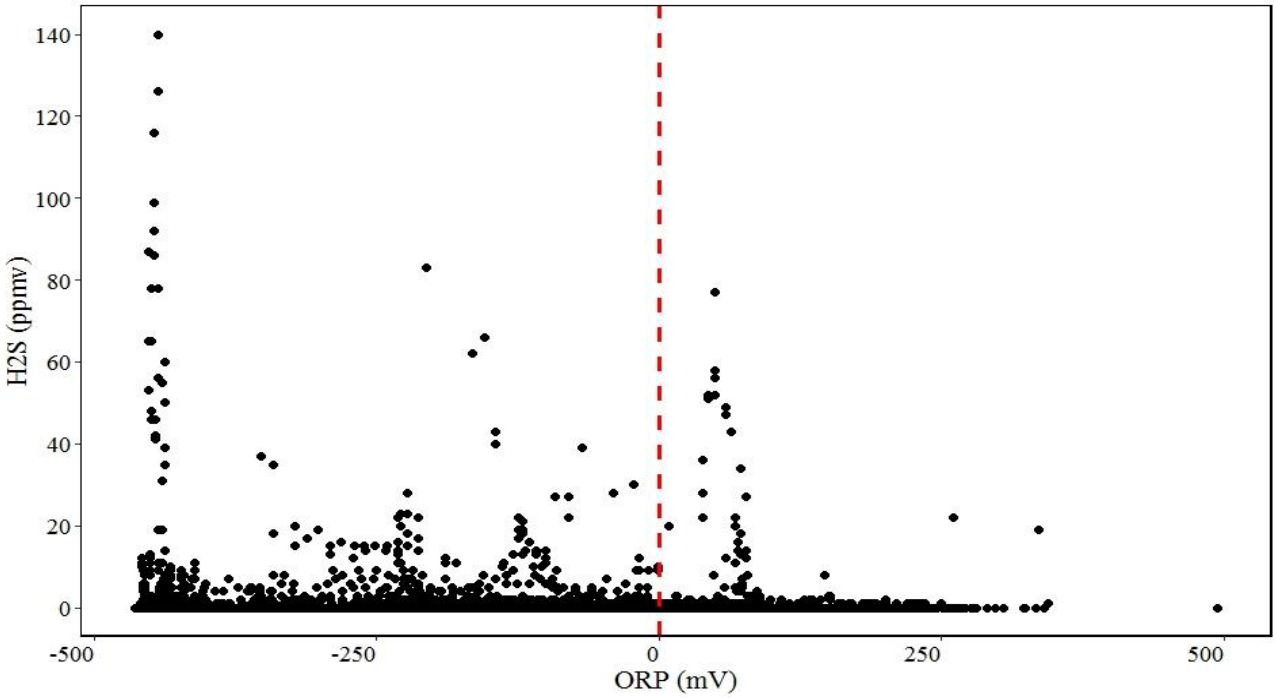
We had sufficient data to train models in Modules 1 and 2 to provide over 90% accuracy when predicting H_2S and VFAs. The predicted H_2S and VFA data will be used to guide the NaOCl dosing. After training Module 3 with the actual VFAs data, Kirie WRP will implement this approach to automate their daily NaOCl dosing. In addition to the benefit of predicting the NaOCl dosage for odor/ corrosion control, the predicted VFAs and H_2S levels can also be used to assist day-to-day operation. The predicted H_2S data, can be used to determine how to run the three existing exhausts and one purge fans in the coarse screen building. Additionally, we are evaluating an H_2S device that allows for online data transfer and provides 4-20 mA signal for control. Predicted H_2S data can be used to check the sensor technology. Furthermore, the VFAs predicted can be used to guide the EBPR operation, e.g., guiding mixer operation in fermentation zone. As part of the 2018 IWS challenge, we developed a data driven model to guide swing zone operation to balance EBPR and ammonia removal in Kirie WRP. Predicted VFAs this time can be used as an input to that model to improve its performance.

The data management and prediction approaches presented here are not specific to this project. It can be implemented at other treatment processes and in any other utility as long as sufficient data are available to train the model. Kirie WRP personnel expressed the strong interest to employ this approach throughout their treatment process when possible. We plan to continue working with the university, and possibly introducing a consultant company to combine this data driven modeling approach, together with process modeling, and other online monitoring and control program throughout the treatment processes at Kirie WRP to improve its operation efficiency and consequently reduce operational costs.

4 CONCLUSION

The team is challenged to handle a variety of data types and qualities. By combining appropriately designed data-driven machine learning approaches with the MWRDGC's specific domain knowledge on wastewater characteristic and process control, the three modules are able to jointly provide one-day prediction of influent characteristics, and accordingly, the VFA and H_2S levels to predict the optimal NaOCl dosage for corrosion control. We believe these efforts are relevant and informative for every water resource recovery utility that need to handle dynamic influent flow and characteristic to make operational changes to meet the treatment target. Our work provides a concrete example of how machine learning can be applied using existing sensors and operational data to solve water resource recovery facility problems. The project also shows the benefits of consolidating interdisciplinary knowledge through collaboration between academia and utility research and operation teams

Appendix A – H_2S and ORP relationship at Kirie WRP during all H_2S monitoring periods



When ORP was < 0

97% of H_2S readings were ≤ 5 ppm

87% of H_2S readings were less than detection limit

Appendix B – Data, data pre-processing, and QA/QC descriptions

Several QA/QC actions were taken during this challenge: (i) data visualization and spot-checks of data compiled from different sources; (ii) use of shared Dropbox to enable tracking of different file versions; (iii) comparison of H_2S data from different OdaLogs; and (iv) site visit and regular meetings between the MWRDGC and ISU members to share information and ideas. Detailed information is provided in Appendix B.

Variable	Source	Description	Data pre-processing & QA/QC
Flow, oxidization reduction potential (ORP), pH, wastewater temperature (Temp), and tunnel elevation	Kirie WRP's Distributed Control System (DCS)	15-minute interval data reported by online instrument. All instruments function as a network with a Modbus Transmission Control Protocol or serial connection communication protocol and connect to the Kirie's DCS for monitoring and data acquisition. Sensor locations: pH, ORP, and Temp sensors are placed at raw sewage sampling tank; flow meter are on the discharge side of raw sewage pumps; tunnel elevation is taken upstream of the Kirie WRP in drop shaft using a bubble system.	Data were checked for duplicate entries and out-of-range readings. The variable "tunnel pumping" was created with information from the tunnel elevation.
Total solids, suspended solids, biochemical oxygen demand, total phosphorus, ammonia, sulfate, organic nitrogen, and total Kjeldahl nitrogen	MWRDGD's laboratory information management system (LIMS)	Daily 24-hour composite sample of influent wastewater. Data from 01/01/2002 until 12/31/2018 were used in module 1. For module 2, data were used for the same period that H_2S or VFA data were available.	Data were checked for out-of-range results. The missing values were filled by propagating the non-missing values backward along a series (for module 1), or interpolated (for module 2). Variables were normalized
Precipitation	National Oceanic and Atmospheric Administration (NOAA)	Daily rainfall data at O'Hare International Airport were collected from the same periods that H_2S or VFA data were available	Data were compared with rainfall data from different sources (USGS and Kirie WRP)

Appendix B (Continued) – Data, data pre-processing, and QA/QC descriptions

Variable	Source	Description	Data pre-processing & QA/QC
Volatile Fatty Acids (VFAs)	MWRDGD's LIMS	Twice a week grab sample of influent wastewater from influent channel available from March 2 2014 to May 29 2019. Samples testing for VFA were collected in irregular frequencies. In module 2, VFA models were trained with data from the period without NaOCl dosing (prior to May 14 2019).	Data were checked for out-of-range results. The missing monitored VFA data were estimated based on an observed correlation with the H_2S and VFAs levels (Appendix F).
Hydrogen Sulfide (H_2S)	OdaLog L2 high range gas logger	15-minute interval data manually downloaded weekly. OdaLog was placed temporarily at the headspace of the influent flow split box at the Kirie WRP during two periods: 07/14/2017-08/03/17 and 03/07/19-07/25/19.	Two side by side OdaLogs were installed and readings compared for three weeks. Day-level summary of H_2S readings was compiled using the maximum reading of the day.
NaOCl dosage for the dose response test	Kirie WRP	Data from dose-response test carried out between 5/14/19 – 7/25/19. We attempted to adjust dosage twice a week, monitor H_2S and VFA response data, and then modify dosage accordingly. Dosage data were used as label in module 3 and ranged between 0 gpd and 238 gpd	Data were checked for out-of-range results.

Appendix C – Forecast of water quality analysis data – Module 1

Figure C1 – Predicted and actual BOD_5 data

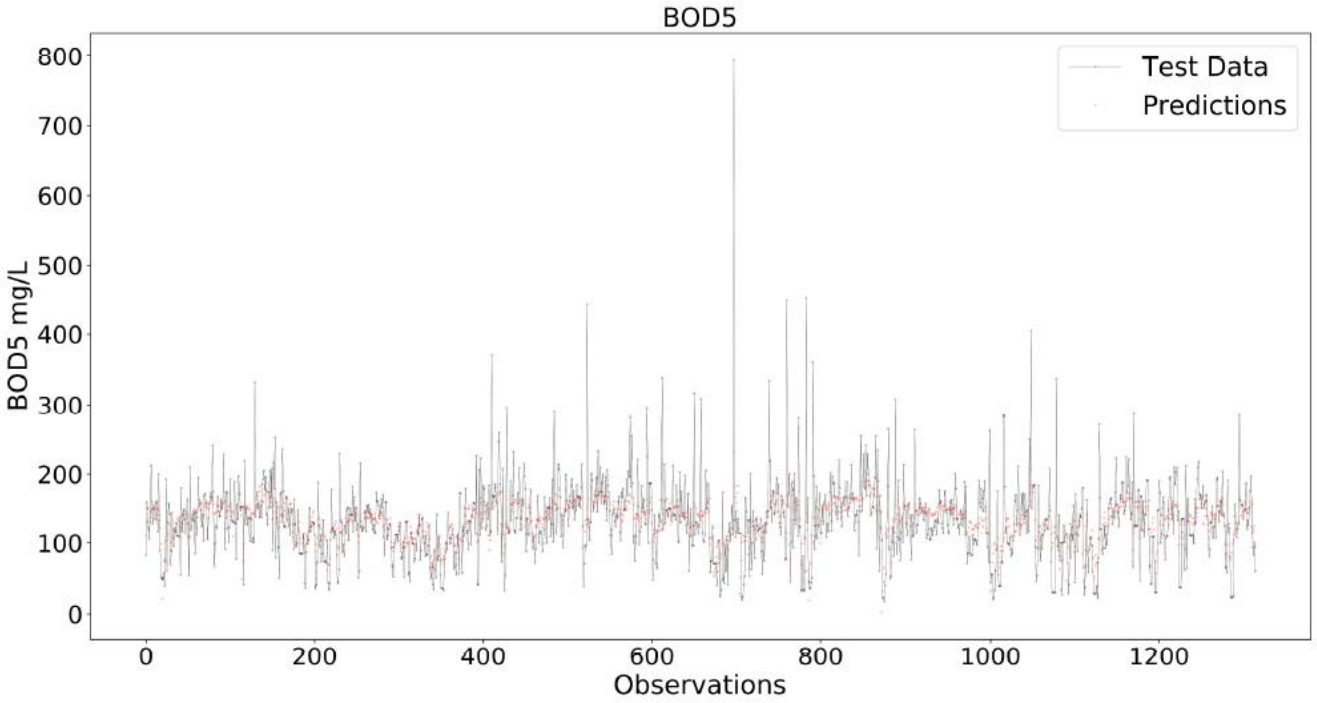


Figure C2 – Predicted and actual ammonia data

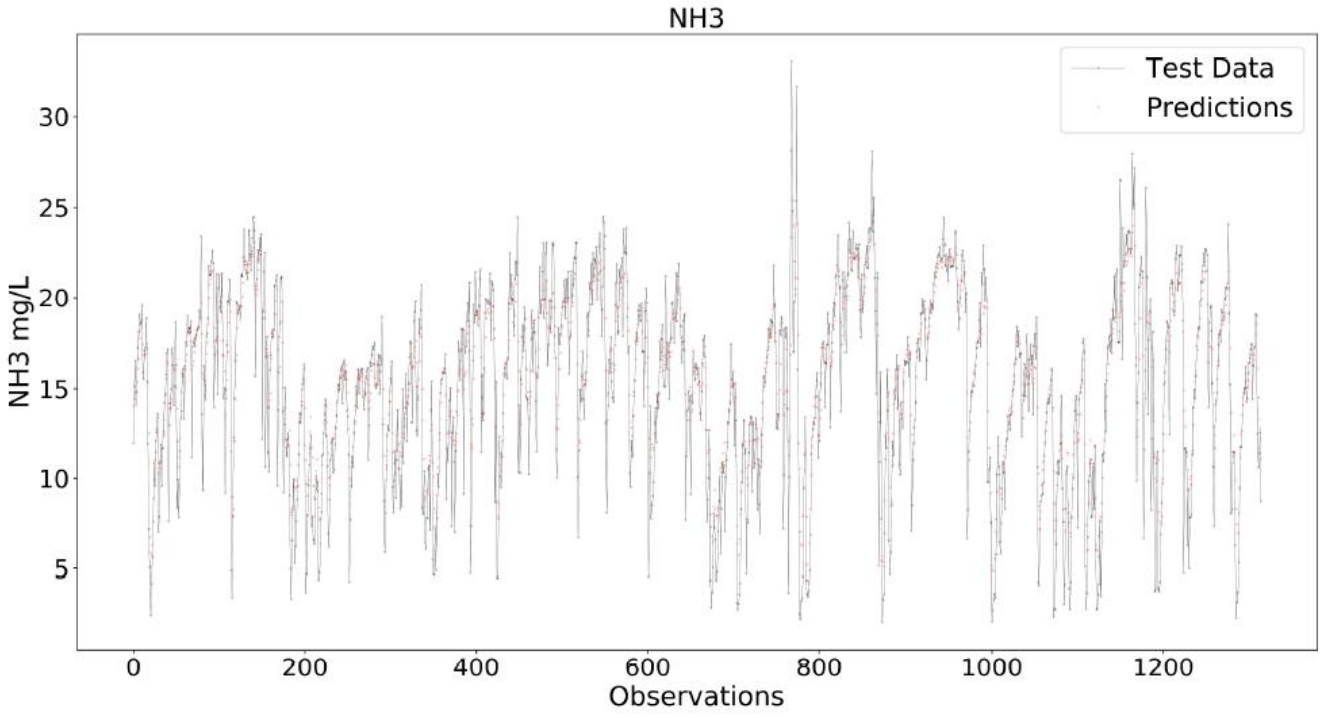
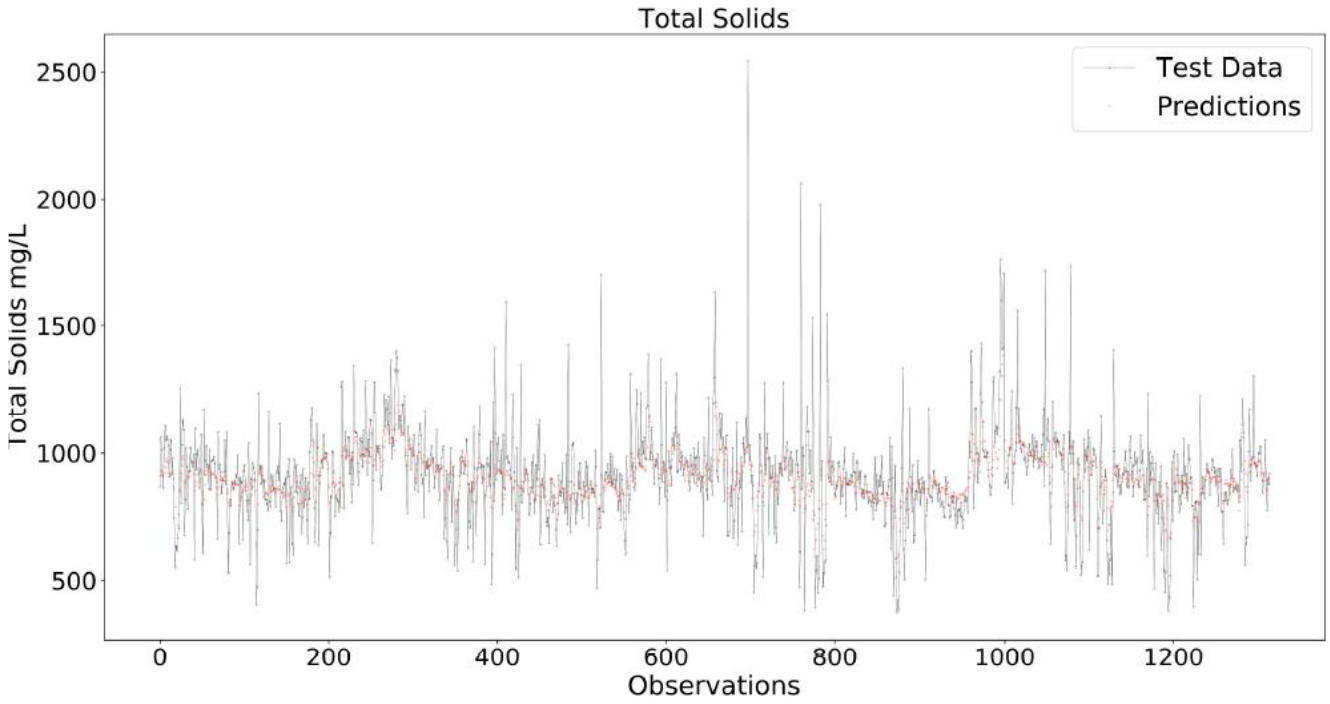


Figure C3 – Predicted and actual total solids data



Appendix D – H_2S and VFAs classes being predicted in module 2

Class Number	H_2S concentration range (ppm)	VFA concentration range (mg/L)
0	0	[0, 5]
1	[1, 5]	[6, 10]
2	[6, 10]	[11, 20]
3	[11, 20]	[21, 30]
4	[21, 30]	[31, 40]
5	[31, 40]	[41, 50]
6	[41, 50]	[51, 60]
7	[51, 60]	[61, 70]
8	[61, 70]	[71, 80]
9	[71, 80]	[81, 90]*
10	[81, 90]	[91, 100]
11	[91, 100]	[101, 110]
12	[101, 110]*	
13	[111, 120]	
14	[121, 130]	
15	[131, 140]	

*class with no data

Appendix E – Order of importance of variables in the Random Forest models predicting maximum daily H_2S and VFAs levels

Figure E.1 – Order of importance of variables in the RF model (Module 2) predicting maximum daily H_2S

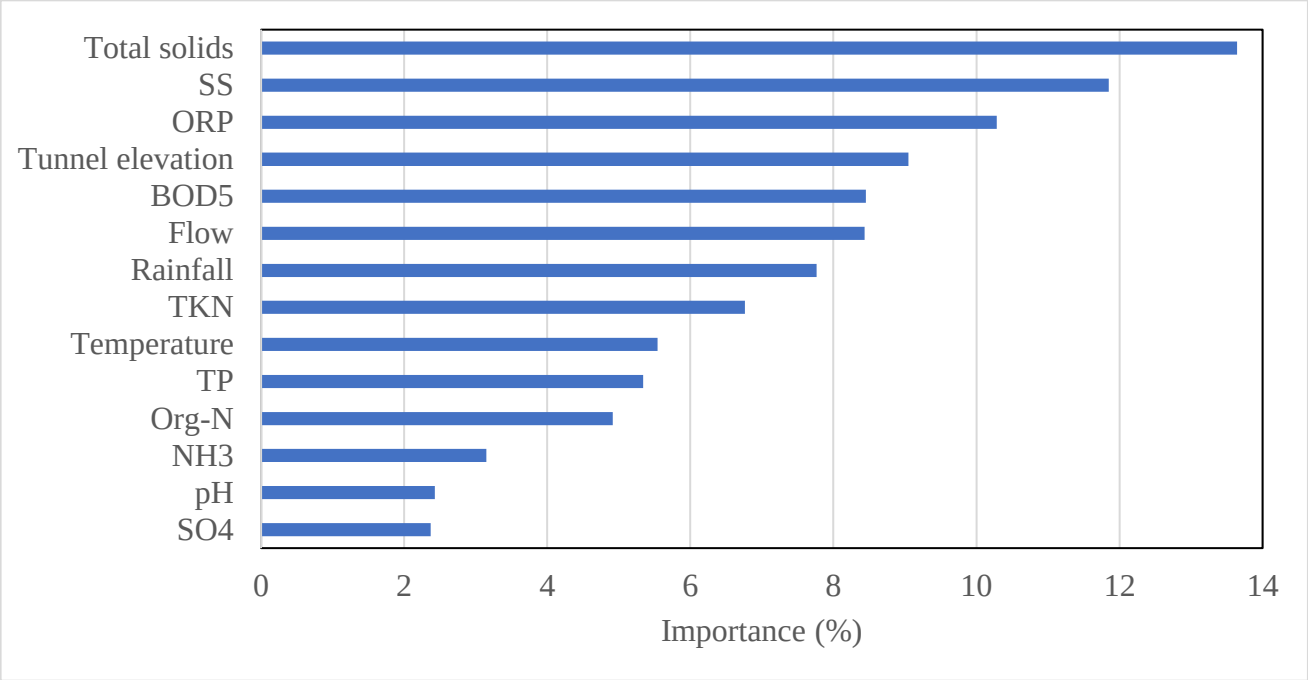
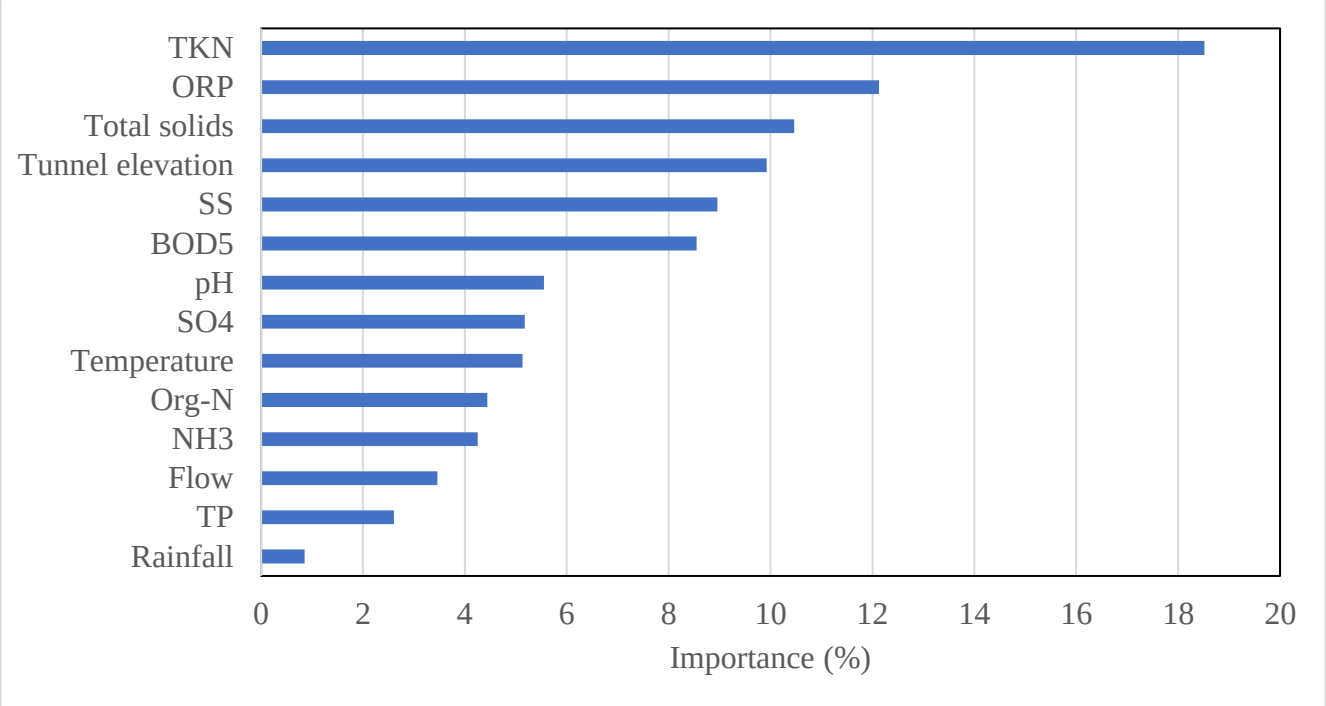
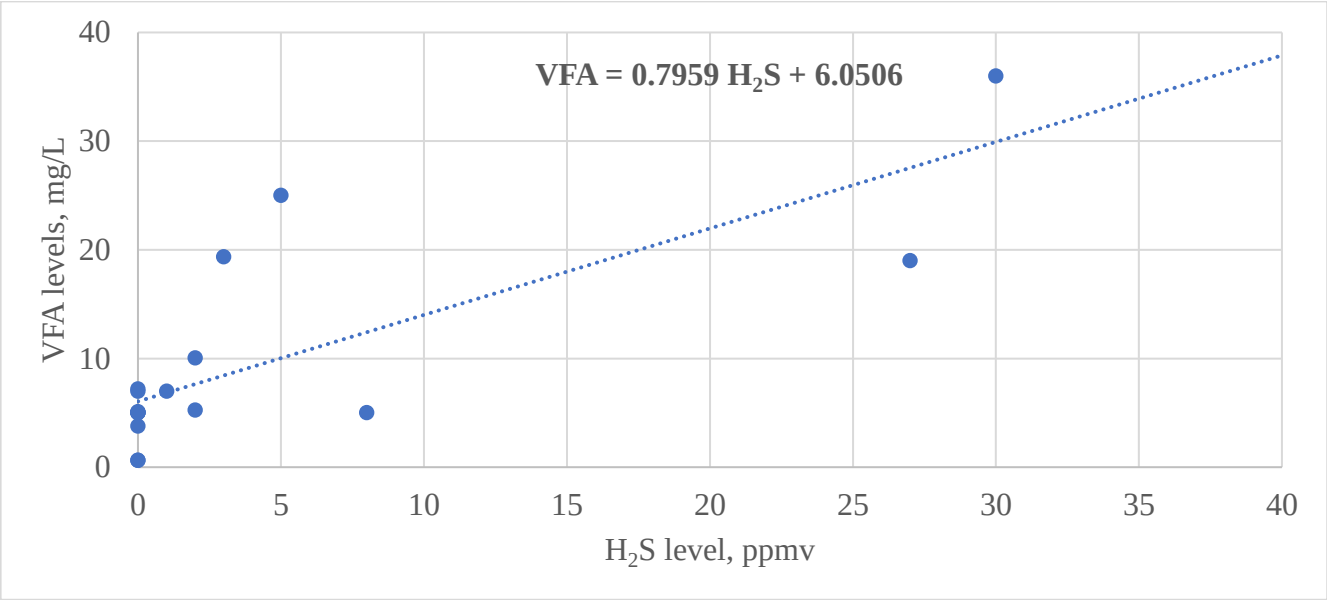


Figure E.2 – Order of importance of variables in the RF model (Module 2) predicting VFAs level



Appendix F – Correlation between VFAs levels and H_2S levels

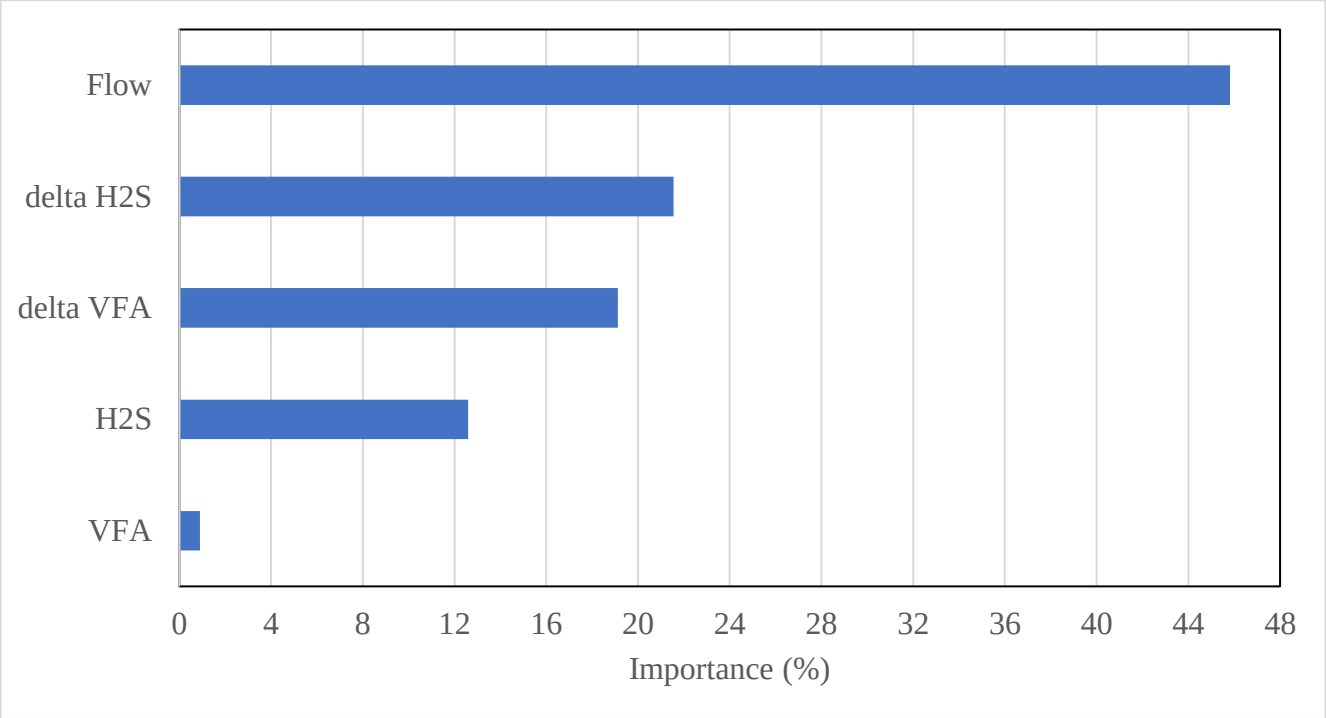


Appendix G – NaOCl classes being predicted in module 3

Classes	NaOCl dosage (gallons/day)
0	0
1	[1,10]*
2	[11,20]*
3	[21,30]*
4	[31,40]
5	[41,50]
6	[51,60]*
7	[61,70]
8	[71,80]
9	[81,90]*
10	[91,100]
11	[101,110]*
12	[111,120]
13	[121,130]
14	[131,140]*
15	[141,150]
16	[151,160]
17	[161,170]*
18	[171,180]
19	[181,190]*
20	[191,200]
21	[201,210]
22	[211,220]*
23	221and above

*class with no data

Appendix H – Order of importance of variables in the RF model (Module 3) predicting NaOCl dosage



Note: delta H_2S is the difference between the predicted H_2S level vs the target H_2S level and delta VFAs is the difference between the predicted VFAs level vs the target VFAs level