Re-envisioning Carbon Monitoring at Water Resource Recovery Facilities: A Synergistic Look into Bio-Electrochemical Sensors and Artificial Intelligence

August 2022

2022 LIFT Intelligent Water Systems Challenge

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

HRSD

Team Members

Seyed Aryan Emaminejad – Civil and Environmental Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL, United States. **sae7@illinois.edu**

Expertise**:** Wastewater Treatment Process Modelling and Design, Microbial Electrochemical Technologies, Environmental Data Science

Roles and Responsibilities**:** Team Lead, Problem Statement and Scope Definition, Data Processing and Analysis, Methods development, Interpretation of Results, Technical Report

Samuel Aguiar[1] – Civil and Environmental Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL, United States. **saguiar2@illinois.edu**

Expertise: Wastewater Treatment Process Modelling and Operation, Nutrient Removal and Recovery, Data Integration

Roles and Responsibilities**:** Data Visualization and Integration, Interpretation of Results, Technical Report

Juliana Mejia-Franco – SENTRY™ Water Technologies Inc., Charlottetown, PE, Canada. **jmejiafranco@sentrywatertech.com**

Expertise: Environmental Engineering Modeling, Water/Wastewater Treatment

Roles and Responsibilities**:** Technology Provider, Data Collection and Pre-Processing

Jeffrey Sparks – Hampton Roads Sanitation District Nansemond Treatment Plant, Suffolk, VA, United States. **jsparks@hrsd.com**

Expertise: Water/Wastewater Treatment Process Engineer, Biological Nutrient Removal Processes

Roles and Responsibilities**:** Utility Partner, Problem Statement and Scope Definition, Interpretation of Results

Dr. Ro. D. Cusick – Civil and Environmental Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL, United States. **rcusick@illinois.edu**

Expertise: Wastewater Treatment, Resource Recovery from Waste, Microbial Electrochemical Technologies

Roles and Responsibilities**:** Academic Adviser, Problem Statement and Scope Definition, Communication of Results

[1] Replaced an existing team member and joined the team during the challenge solution period

1 Project Summary

The Hampton Roads Sanitation District (HRSD) pumps wastewater from the Town of Smithfield, Isle of Wight County, City of Suffolk, City of Chesapeake, and City of Portsmouth to its Nansemond Treatment Plant in Suffolk, Virginia, to undergo advanced treatment before discharge to the James River. The biological nutrient removal processes (BNR) at the Nansemond plant, especially the nitrogen (N) removal process, could be sensitive to carbon (C) availability and depend on control of recycle flow rates and methanol dosing to ensure satisfactory effluent quality. Additionally, the plant applies advanced treatment to 1 MGD of its treated water to also produce an effluent stream that meets drinking water quality standards. This advanced process does not include reverse osmosis (RO), and therefore can be impacted by organic loading variations in the influent stream. As the feedback control system could be adversely impacted by online sensor failures leading to increased operational costs and diminished process stability, an interdisciplinary university-utility-industry team was formed to propose and implement a feedforward component to one of the plant's existing feedback-only control systems to optimize and increase the resilience of their treatment and nutrient removal processes, and minimize the plant's operational costs associated with pumping and chemical requirements.

The proposed feedforward component presented in this report was based on a novel bio-electrochemical monitoring technology that has recently been developed for use in industrial scale applications. The signals obtained from these monitoring probes could be used to track influent C dynamics and metabolism in biological nutrient removal processes. Artificial intelligence (AI) models developed in this report were able to provide accurate predictions of the mass of nitrate removed at the plant's 1st stage anoxic zone based on a set of input variables with a focus on bio-electrochemical sensor (BES) signals, methanol dosing rates, and nitrified recycle (NRCY) flow rates. Additionally, the BES signal forecasting model developed in this report was able to provide accurate long-term predictions one month in advance. This offers significant operational advantages to the plant as it can be used to have a long-term estimate of expected BES signal values which is indicative of influent organic loading variations. With this tool, plant operators can predict and instantly identify plant shock loading events and make necessary time-sensitive operational decisions to ensure the stability of their biological nutrient removal processes and their satisfactory effluent quality.

The feedforward monitoring component developed in this report can change the way influent organic loading monitoring/control is currently conducted in water resource recovery facilities and be used in other public utilities that face similar operational challenges as the HRSD Nansemond plant. Finally, a conceptual software dashboard was designed to present an easy-to-use interface where all the models developed in this report could be used without requiring an extensive background in data science.

2 Problem Statement and Background

2.1 Operational Challenges

The HRSD Nansemond Treatment Plant is a 30-MGD BNR facility located in Suffolk, VA, and receives a mixture of domestic and industrial wastewater. The influent wastewater is highly treated prior to discharge to the James River; and 1 MGD of the treated water will undergo additional advanced treatment to meet drinking water quality standards [1]. Excessive nutrients including phosphorous (P) and N have been reported as one of the most serious water quality problems in the area. While the Nansemond plant has a major role in removing nutrients from wastewater in the area as a BNR plant, it is prone to influent characteristic imbalance events which can affect effluent quality. Moreover, the advanced water treatment process at the plant does not include an RO unit, therefore its operation is highly sensitive to variations in influent composition, more specifically the organic C content. The plant is currently using a 5-stage Bardenpho process for BNR and takes advantage of a variety of online instruments for process control.

The plant's BNR process currently employs a combination of feedforward and feedback controls to meet its total inorganic nitrogen (TIN) objectives for downstream advanced water treatment. However, denitrification in the first stage anoxic zone is based on a feedback-only controller that looks at the nitrate concentrations entering the downstream aerobic zone. The overarching strategy for first stage denitrification is to minimize the N load stress on the 2nd stage anoxic zone by leveraging increased NRCY flows and higher levels of N removal in the 1st stage as shown in Figure 1. The plant theorizes that this strategy helps with meeting the TIN objectives more reliably. During the wintertime, increasing the NRCY flow takes place at the expense of methanol addition to the 1st stage anoxic zone to support denitrification; while in the summertime, the primary clarifiers are stressed to push more carbon, or chemical oxygen demand (COD), to the 1st stage anoxic zone often allowing the suspension of 1st stage anoxic methanol dosing. Numerous controllers are used in the plant's biological processes, and to maintain process stability, each one of them should perform well. Within this controls framework, if any one of the controllers does not perform well, it could result in process upsets. For instance, if NRCY flow adjustment is not optimized, it would lead to imbalanced hydraulic retention time (HRT) values in the biological processes, causing the tuning parameters for other controllers to no longer be ideal and making them either too fast or too slow.

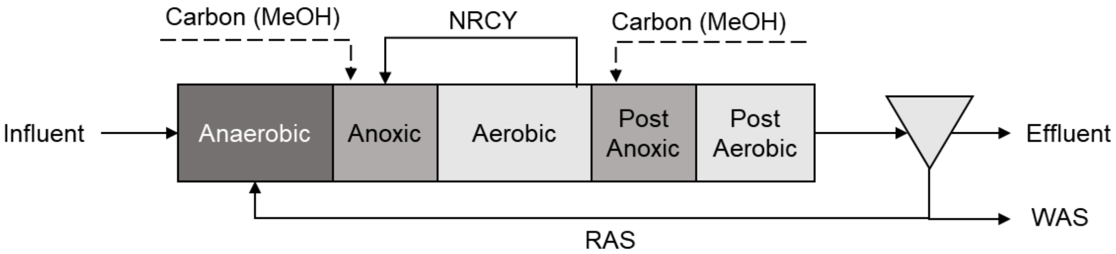


Figure 1. HRSD Nansemond plant BNR process overview

Additionally, these controllers usually have long delays, making process control more challenging even when all the controllers are performing as expected. Long delays are especially problematic for strictly feedback control systems as is the case at the HRSD Nansemond plant in the 1st stage anoxic zone since it can adversely impact the efficacy of the BNR process. Therefore, the plant is looking to add a feedforward component to their already existing feedback control system to have the ability to run a feedforward-feedback control strategy, which will allow them to optimize their process stability, operational costs, and BNR efficacy. Ideally, the plant is looking to build a controller that automates its 1st stage anoxic zone

NRCY flow adjustment/methanol addition based on a feedforward component that uses the predictive capabilities of AI.

Moreover, another one of the major factors impacting the efficacy of the plant's BNR process is the C content of influent wastewater. Sudden changes in the C content lead to an imbalanced C:N:P ratio which could adversely impact the BNR efficacy by creating an unfavorable environment for BNR microorganisms. Within this context, industrial discharge events and imbalanced C:N:P compositions have previously led to disruptions of BNR processes at the HRSD Nansemond plant. Therefore, efficient real-time C monitoring would offer significant advantages to the plant to monitor these discharge events and maximize its effluent quality. Additionally, a reliable real-time C monitoring tool could act as the feedforward component that the plant is looking to implement in the 1st stage anoxic zone to have a more efficient control strategy.

2.2 Proposed Intelligent Water System Solution

C monitoring and management at water resource recovery facilities (WRRFs) can lead to significant energy savings and enhance the efficacy of biological wastewater treatment and nutrient recovery processes. Although the benefits of real-time C monitoring are clear, existing technologies often lack the ability to offer a robust and time-efficient monitoring tool. Recently, bio-electrochemical sensors (BESs) are gaining increasing attention as a novel approach for online C monitoring due to their high sensitivity and robustness, low maintenance requirements, and quick response time. BESs provide real-time amperometric data of the soluble C content of water/wastewater by utilizing electroactive biofilms that convert organic matter into electrical current. The generated current is a direct measurement of biological activity, which is proportional to the biodegradable organic content of water/wastewater and can be used to track C loading/concentration dynamics.

To explore the operational advantages of using BES technology for C monitoring as a feedforward component of a novel and modified control system at the HRSD Nansemond plant, two commercially available BESs (SENTRY, Canada) were installed at the plant's primary clarifier's effluent channels – one before Return Activated Sludge (RAS) blending and one after. The primary goal of using BES monitoring was to monitor the variations of biological activity in the primary clarifier effluent to detect influent wastewater composition imbalance events. As mentioned, BES technology for C monitoring at the HRSD Nansemond plant could also be used in a feedforward operational mode within a network of online sensors to determine the impact of different operational parameters (NRCY flow rate, methanol dosing rate, secondary clarifier ferric chloride addition, DO, etc.) on the plant's BNR efficacy and effluent quality. In addition to increasing process resilience, this approach could also offer financial benefits to the plant through optimization of pumping/blower energy requirements and chemical consumption rates.

The proposed feedforward-feedback control system leads to generation of massive datasets which warrants a need for a comprehensive framework of sensor signal processing and AI models to fully extract valuable process-related information from the obtained data and make time-sensitive operational decisions. The main objectives of this university-utility-industry collaboration are as follows:

- Providing a detailed analysis of the relationship between BES signals and BNR parameters to determine the applicability of wastewater quality monitoring using BES technology as a novel approach for real-time plantwide C monitoring.
- Development and comparison of various state-of-the-art machine learning (ML) and deep learning (DL) models to predict the BNR performance of the plant using BES signals as the core responding variable to C dynamics, in addition to other online sensor measurable parameters. This will be done

by using a variety of online sensor inputs and operational variables to predict the mass of nitrate removed in the 1st stage anoxic zone. The ideal goal is for the plant to use the best developed model as a feedforward component to automate its NRCY flow/methanol addition control system.

- Development of a signal processing/forecasting framework using DL and advanced mathematical algorithms to uncover the structural dynamics of the BES signals and monitor/detect/predict influent organic shock loading events which would offer significant operational benefits to the plant, ranging from increased BNR process resilience to higher effluent quality.
- Propose the intelligent water system incorporating every technical aspect of this new control strategy as an intuitive and user-friendly software dashboard, enabling plant operators to take advantage of the obtained information using a unified interface.

A general overview of the proposed solution is presented in Figure 2.

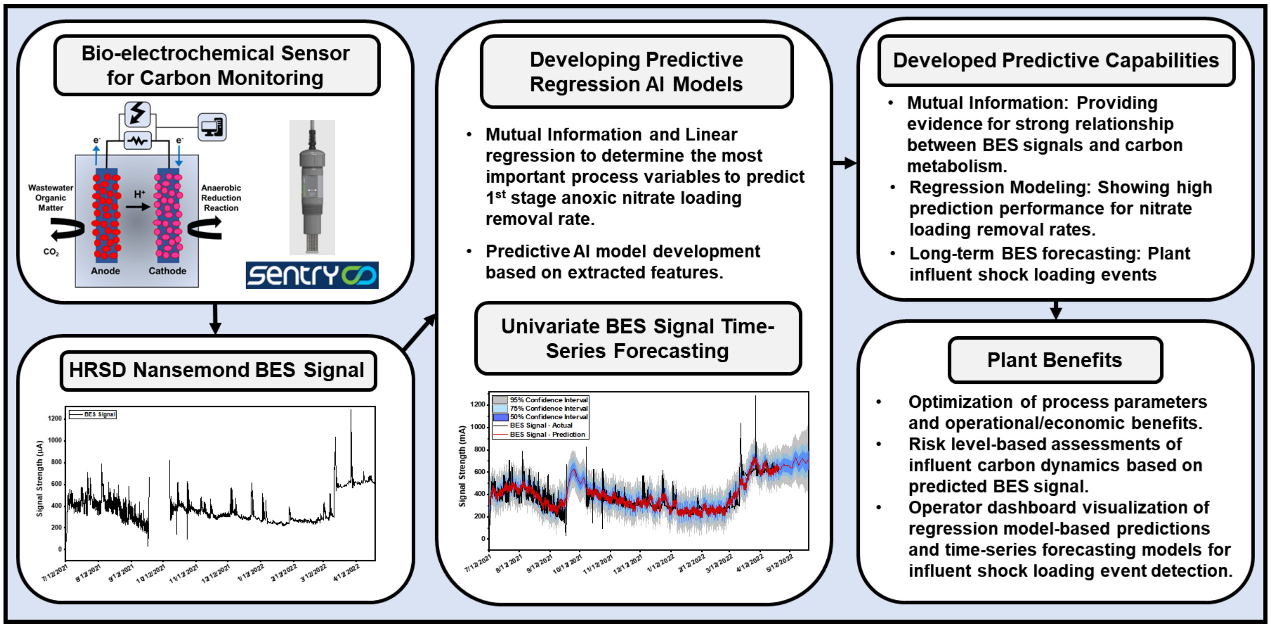

Bio-electrochemical Sensor for Carbon Monitoring
e⁻
e⁻
Wastewater Organic Matter
H⁺
Anaerobic Reduction Reaction
CO₂
Anode
Cathode
sentry
HRSD Nansemond BES Signal
BES Signal
Signal Strength (μA)
1200
1000
800
600
400
200
0
7/12/2021
8/12/2021
9/12/2021
10/12/2021
11/12/2021
12/12/2021
1/12/2022
2/12/2022
3/12/2022
4/12/2022
Developing Predictive Regression AI Models
• Mutual Information and Linear regression to determine the most important process variables to predict 1st stage anoxic nitrate loading removal rate.
• Predictive AI model development based on extracted features.
Univariate BES Signal Time-Series Forecasting
95% Confidence Interval
75% Confidence Interval
50% Confidence Interval
BES Signal - Actual
BES Signal - Prediction
Signal Strength (mA)
1200
1000
800
600
400
200
0
7/12/2021
8/12/2021
9/12/2021
10/12/2021
11/12/2021
12/12/2021
1/12/2022
2/12/2022
3/12/2022
4/12/2022
5/12/2022
Developed Predictive Capabilities
• Mutual Information: Providing evidence for strong relationship between BES signals and carbon metabolism.
• Regression Modeling: Showing high prediction performance for nitrate loading removal rates.
• Long-term BES forecasting: Plant influent shock loading events
Plant Benefits
• Optimization of process parameters and operational/economic benefits.
• Risk level-based assessments of influent carbon dynamics based on predicted BES signal.
• Operator dashboard visualization of regression model-based predictions and time-series forecasting models for influent shock loading event detection.

Figure 2. Proposed Intelligent Water System solution

3 Data Collection and Methodology

The BESs were installed at the primary clarifier effluent channels on 07/12/2021 and have been recording data since then with minute interval. The BES dataset used in this study started from 07/12/2021 and ended on 04/30/2022. 10-minute average values were calculated to reduce the size of the dataset used in time-series forecasting and regression models. For multivariate time-series forecasting, temperature and pH values of influent wastewater with matching timestamps with the BES 10-minute averaged values were used, whereas for univariate time-series forecasting, only the BES 10-minute averaged dataset was used.

The full regression dataset to predict the mass of nitrate removed in the 1st stage anoxic zone during the same time period (07/12/2021 – 04/30/2022) contained 42 different variables which included the plant's influent flow rate, recycle and NRCY flow rates, methanol dosing rates, online sensor measurements in different BNR unit processes (DO, ortho-phosphate (OP), nitrate, ammonia, etc.). Many of these sensor measurements were recorded in individual parallel tanks corresponding to different unit processes. Therefore, an additional data preprocessing step was taken by averaging sensor measurements across all parallel tanks of their corresponding unit process which reduced the number of variables from 42 to 16. The mass of nitrate removed in the 1st stage anoxic zone was calculated as follows:

$$MNR = (Q_{NRCY} * C_{2,inf}) - [(Q_{NRCY} + Q_{inf} + Q_{RAS}) * C_{1,eff}]$$

Where MNR is the mass of nitrate removed (kg N as NO_3/day); Q_{NRCY}, Q_{inf}, and Q_{RAS} are NRCY flow, plant influent flow, and RAS flow (L/day) rates, respectively; and $C_{2,inf}$ and $C_{1,eff}$ are average nitrate concentrations (kg N/L) in the 2nd stage anoxic zone influent and 1st stage anoxic zone effluent respectively. Since during the course of this study, a number of sensors went offline due to sensor failures and recorded zero or NaN (Not a Number) values, all rows containing zero/NaN values were deleted from the dataset. The regression dataset was then shuffled and ready for further analysis. Both the time-series and regression datasets were divided into training (70%) and testing/validation (30%) sub-datasets. The error metrics for all regression and forecasting models were R^2 and root mean square error (RMSE).

3.1 Correlation analysis

Different correlation analysis methods were used on the regression dataset to detect the relationship between different variables that impacted the mass of nitrate removed in the 1st stage anoxic zone as a measure of the plant's BNR performance. The selected methods were F-test for regression (f_regression) analysis and mutual information (MI). F-test captures the linear dependency between individual variables and the target variable. This method computes the linear correlation strength and the F-statistic to return the F-values. Larger values mean stronger linear relationships between the dependent and independent variables of interest.

Unlike F-test that reveals linear relationships, MI is used in information theory and presents an overview of general interactions (both linear and nonlinear) between the independent variable and the dependent target variable [2]. MI returns a value between 0 and 1 and is a measure of the reduction in uncertainty for one variable given a known value of the other variable. Intuitively, it reflects how much knowledge of one variable tells us about the other. The closer the MI coefficient is to 1, the more knowledge one can gain from a specific parameter knowing the other.

3.2 ML and DL Regression Models

A variety of state-of-the-art ML and DL methods were used to develop a set of models capable of predicting the mass of nitrate removed in the 1st stage anoxic zone. These models included support vector machine (SVM), random forest (RF), extreme gradient boosting (XGboost), and artificial neural network (ANN) and were all developed in Python using "scikit-learn [3]," "xgboost [4]," and "keras [5]" packages. Hyper parameter tuning of the models was conducted using scikit-learn "GridSearchCV" and learning curve plots, while cross-validation was used to increase the generalizability of the models. Overall, final models were selected out of more than 250 different tested models to ensure the robustness of the prediction performance. For ANN training, "early-stopping" functionality was used to prevent overfitting of the trained model on the training dataset. This was done by automatically stopping the model training process when the prediction error of the validation dataset did not significantly improve over a specific number of iterations. More details are provided in the Appendix.

3.3 Time-Series Forecasting Models

For time-series forecasting models, two different approaches were evaluated to determine the right approach in monitoring BES signals. (i) Univariate analysis where the BES data were used to predict future

BES values, and (ii) multivariate analysis where in addition to the BES data, other related process variables were used to make predictions for future BES values. A comparison was made to determine whether multivariate analysis presents a significant advantage over univariate forecasting models. The DL univariate forecasting model consisted of univariate long short-term memory recurrent neural networks (LSTM-RNN) due to their high prediction power. For multivariate analysis, a multivariate LSTM model was developed that included the BES signal and influent wastewater temperature and pH as inputs. Additionally, "Facebook Prophet" (FB prophet) which is an open-source time-series algorithm developed by Facebook was used to model the BES signal. More details regarding development of time-series forecasting models and their application-specific information are provided in the Appendix.

4 Results and Discussion

4.1 Correlation Analysis

The first logical step to propose and add BES monitoring as an additional feedforward component to the BNR control system of the HRSD Nansemond plant was to explore the relationship between the BES signal and BNR performance parameters, with a particular interest in whether a strong correlation between BES signal and nitrate loading removal exists for feedforward control. The selected methods (F-test regression and MI) were used to quantify the contribution of each variable to the overall variance of the mass of nitrate removed in the 1st stage anoxic zone. Figure 3 shows the correlation analysis results along with their respective coefficient values.

As demonstrated in Figure 3A, F-test results indicated that the BES signal did not show a strong linear relationship with mass of nitrate removed in the 1st stage anoxic zone. This indicated that the interaction between the BES signal and the mass of nitrate removed in the 1st stage anoxic zone could have a complex and nonlinear nature requiring further analysis. Similarly, MI results shown in Figure 3B indicated that the 2nd stage influent nitrate had the strongest general relationship with nitrate removal in the 1st stage, which was expected as this variable was part of the equation to calculate mass of nitrate removed in the 1st stage anoxic zone. Furthermore, the BES signal was the second strongest pair with the mass of nitrate removed in the 1st stage anoxic

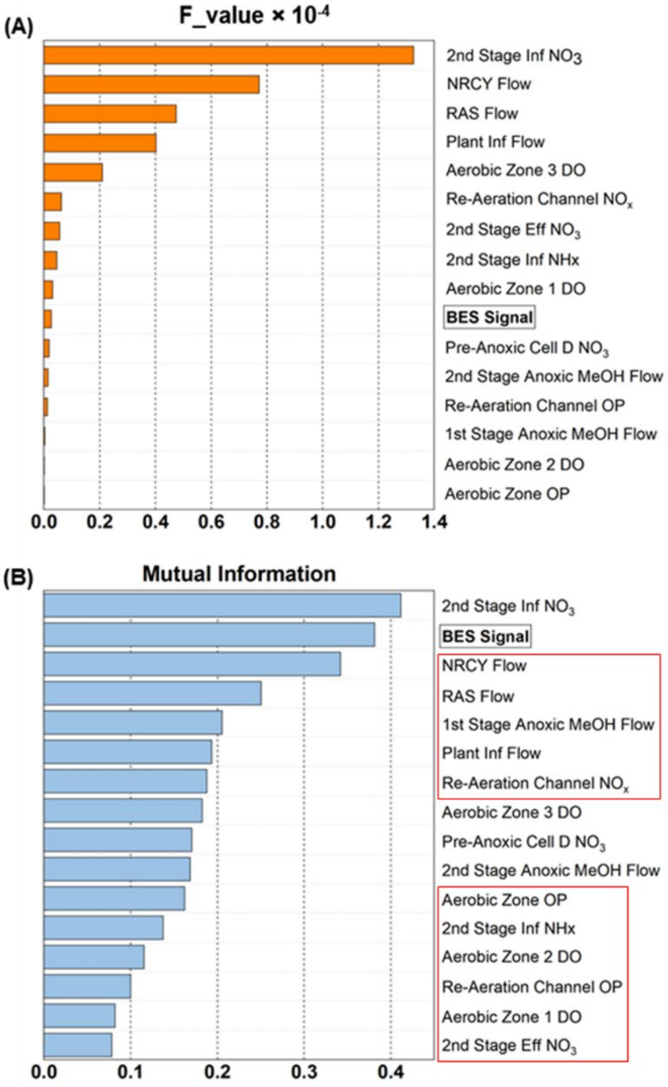


Figure 3. Linear correlation or F-test regression (A) and general information or MI analysis (B). The variables in the red box and the BES signal were selected as the final variables for development of the regression models.

zone which had two important implications: (i) the BES signal proved to be the strongest predictor of the nitrate mass removed in the 1st stage anoxic zone, and (ii) BES biomonitoring is heavily governed by complex nonlinear interactions and cannot be fully analyzed if a linear approach for parameter analysis is chosen. After discussions with the utility partner, the variables with the highest impact on the mass of nitrate removed in the 1st stage anoxic zone (and not included in MNR calculation) were selected to be used in the regression models.

These results correspond to the ability of using BESs installed in the plant's primary clarification effluent channel to track the C availability/metabolism of the system which would also impact C availability in later stages. The fact that the BES signal had the strongest general (linear + nonlinear) sensory relationship with the mass of removed nitrate in the 1st stage anoxic zone indicates the vital importance of real-time C monitoring for efficient BNR at WRRFs. Since the BES signal demonstrated the highest degree of general relationship with the mass of nitrate removed in the 1st stage anoxic zone, the plant could switch from a strict feedback NRCY-methanol control strategy to a feedforward-feedback BES-NRCY-methanol control strategy for a more representative and energy-efficient method of BNR monitoring and control. The ultimate goal would be to automate the NRCY flow adjustment and methanol addition to the 1st stage anoxic zone based on this feedforward component.

This section concludes that although linear correlation analysis between dependent and independent variables is very common in peer-reviewed articles and technical reports, it could fail to account for complex nonlinear interactions. As wastewater treatment and BNR processes are governed by highly nonlinear and complex interactions, a different approach to simultaneously account for linearity and nonlinearity in large datasets is suggested in this report to present a more comprehensive overview of wastewater-related variable interactions.

5 Prediction of the Mass of Nitrate Removed in the 1st Stage Anoxic Zone

After conducting correlation analysis on the regression input dataset and selecting sub-datasets with the final variables, various AI models were developed to predict the mass of nitrate removed in the 1st stage anoxic zone. Different models were chosen purposefully to accommodate the conditions of the input wastewater variables; it was evident that highly nonlinear and complex interactions exist in the data with outliers caused by different plant events. As shown in Figure 4, all the models had high prediction performances with low error values. The distribution of individual model predictions was in a close range as the actual observed nitrate loading removal around the median and 25th/75th percentiles, with no significant differences. Out of all the trained models, XGboost had the highest R^2 value of 0.92, and the lowest RMSE value of 142.6 kg N as NO_3/day. This was followed by ANN, RF, and SVM models with R^2 values of 0.91, and 0.89 (RF and SVM had a similar R^2) and RMSE values of 154.6, 159.8, and 178.5 kg N as NO_3/day, respectively.

The close prediction performance of the trained ML and DL models indicated that in relatively smaller datasets such as the one used in this study with around 10 months of recorded data, choosing the right model may not strictly be a function of prediction strength, but it would rather be a function of training/prediction speed, interpretability, and complexity of model development. Therefore, the recommended model would be XGboost as in addition to having an excellent prediction power with the highest R^2 and lowest RMSE values, it is relatively not prone to common overfitting problems, is highly customizable, and has a straightforward model development process.

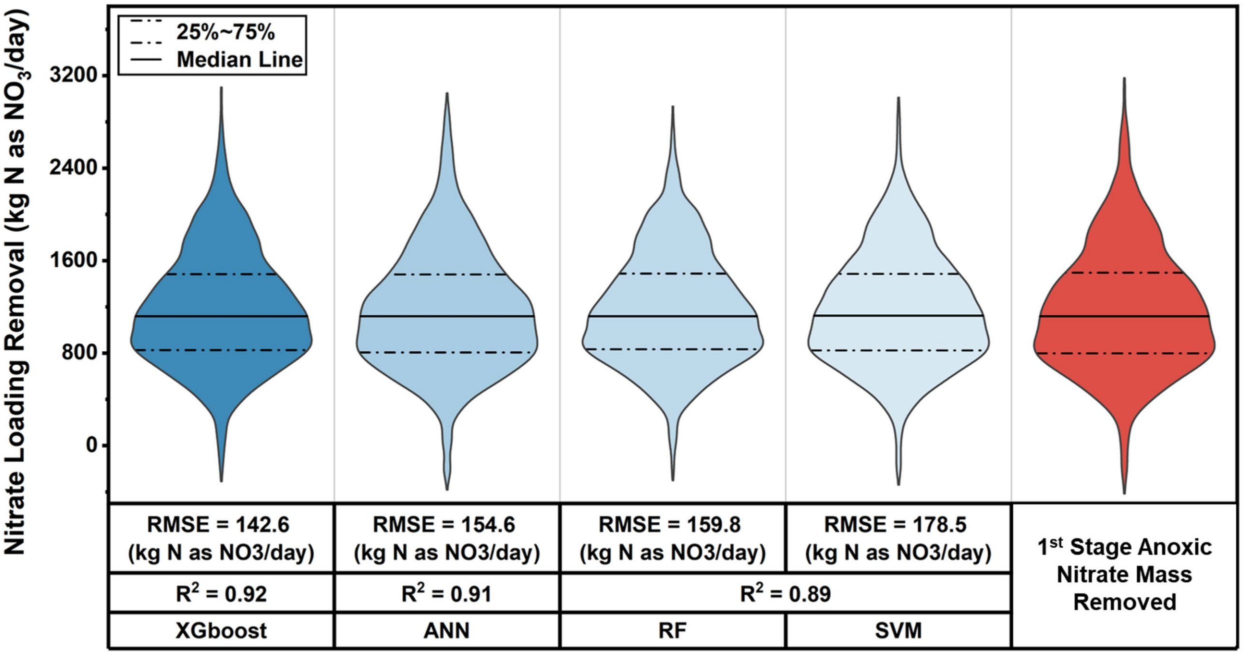


Figure 4. Model prediction performance of the mass of nitrate removed in the 1st stage anoxic zone.

Following discussions with the plant about the applicability of these models, two scenarios are envisioned as follows: (i) using these models, the plant operator can evaluate a variety of different scenarios with various NRCY flow rates with and without methanol dosing to predict the mass of nitrate removed in the 1st stage anoxic zone and easily convert these values to removal percentage. Depending on the plant's desired BNR goal, the NRCY flow and methanol dosing rate would be optimized and lead to increased BNR efficacy/effluent quality and minimized chemical consumption with both operational and economic advantages; (ii) modifying the regression model to directly predict the required NRCY flow to remove a certain mass of nitrate with available carbon (BES signal + optional methanol dosing).

The second scenario could be achieved by operating the 1st stage BNR with bounds on the NRCY flow (e.g., 250% to 500% of plant influent flow) and predicting the required NRCY flow rate to achieve a certain mass of nitrate removal without methanol addition. If the predicted NRCY flow was found to be less than the minimum bound (e.g., 250%), then the NRCY flow would be kept constant at 250% and the impact of methanol dosing would be investigated next. Given the presented evidence on the effectiveness of using BES probes to track C dynamics, these scenarios are all possible with minor modifications in the developed regression models. This team will keep working closely with the plant to evaluate all the possible scenarios.

6 BES Signal Prediction and Time-Series Forecasting Modelling

Methanol dosing and redirecting additional C from primary clarifiers play a vital role in effective BNR performance at the HRSD Nansemond plant. Given the evidence presented on the effectiveness of using BES technology as a proxy for biodegradable C dynamics and metabolism monitoring, accurately predicting the BES signal can help the plant to have a real-time influent organic shock loading event detection tool and prevent the adverse impacts of C breakthrough on their BNR processes. Detection/prediction of such plant events offers a significant process control advantage by enabling the plant operators to be one step ahead and make necessary operational decisions for maintaining BNR stability that would have otherwise been impacted by an organic shock loading event. These decisions can range from adjusting recycle flow rates, to methanol dosing for denitrification in anoxic zones or addition of ferric chloride in the secondary clarifiers to prevent high effluent P concentrations. While in a univariate time-series forecasting model, previous BES signal datapoints are used to predict the future values [6], during a multivariate analysis a combination of input variables is used to predict the BES signal [7]. As mentioned, for the multivariate LSTM model, influent wastewater temperature and pH were added to the BES signal as the three inputs. This combination was chosen based on the results of an earlier case study at the Stickney Water Reclamation Plant, Chicago, IL, where temperature and pH were found to also contribute to the variance of BES signals [8].

A comparison between the actual BES signal in the test dataset region and LSTM univariate and multivariate predictions is presented in Figure 5. Both LSTM models provided highly accurate predictions of the BES signal within an hourly time window, with R^2 and RMSE values of 0.94 and 34.5 mA for the univariate, and 0.95 and 28.1 mA for the multivariate model, respectively. This high prediction performance indicated that using a multivariate analysis in this case would not be justified as having a higher number of input variables did not significantly change the performance of the model. Therefore, a univariate analysis is suggested due to faster training/prediction speed and easier model development. As the input dataset contained hourly averaged BES signal values and since an LSTM model uses a sliding window to make the next predictions, the model outputs signal values one hour into the future. To make this prediction window larger e.g., daily predictions, the input dataset needs to have daily values for the BES signal. Given the relatively short period of this case study (9-10 months), this would have significantly decreased the number of available datapoints for model training and adversely impacted the prediction performance of the model. Therefore, although the univariate LSTM model can make extremely accurate predictions into the near future (1 hour in this case), another algorithm was also evaluated to make long-term predictions into the future which will be discussed in the next section.

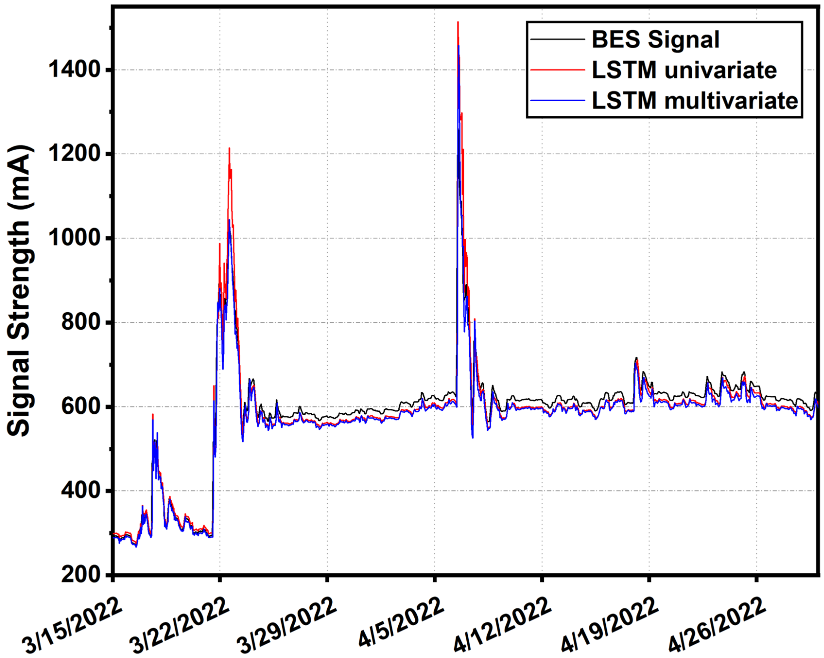


Figure 5. Comparison between univariate and multivariate LSTM prediction of the BES signal.

6.1 Univariate Signal Prediction with Facebook Prophet

FB prophet is an in-house time-series forecasting algorithm developed by Facebook and was made public in 2017. This algorithm is a powerful and easy-to-use tool which was originally intended to make reliable forecasts for planning and goal setting based on nonlinear trends. According to the developers, this algorithm is optimized for applications with characteristics such as hourly, daily, or weekly observations with at least several months of history, strong human-scale seasonalities (day of week or time of year), holidays occurring at irregular intervals, missing observations or large outliers, historical trend changes, and trends that follow nonlinear growth curves [9].

These characteristics are well aligned with wastewater applications as many WRRFs demonstrate influent flow and organic loading rate daily/weekly/monthly patterns, are governed by nonlinear trends, and impacted by holiday events. Additionally, this algorithm can handle outliers and missing values which is ideal for sensor signal modeling as sensor signals can demonstrate such characteristics. Therefore, as shown in Figure 6, FB prophet was used to model the BES signal in this case study, analyze its capability to monitor the HRSD Nansemond plant's influent organic loading, and provide long-term BES signal predictions. The area shaded in light red in Figure 6 corresponds to a period where the biosensor was clogged by a layer of scum and grease and recorded very high signal values. The biosensor was cleaned after a number of days and went back to normal operational mode. The highlighted red area was therefore excluded from the graph for visualization purposes.

The FB prophet signal prediction (red line) closely follows the actual recorded BES signal (black line) except in the regions that sudden spikes were observed in the BES signal which corresponded to influent shock loading events at the plant. This feature was exploited to define three different prediction zones for monitoring the stability of the BES signal and thus providing a real-time plant event detection tool. These prediction ranges were divided into normal (50% confidence interval – dark blue), warning (75% confidence interval – light blue), and plant influent shock loading (95% confidence interval - grey) regions. This means that if the actual BES signal exceeds the 95% confidence interval region of the predicted signal, the user will be notified that a plant shock loading event is taking place. Similarly, if the actual signal is fluctuating within the 75% confidence interval region of the predictions, the user would receive a warning, and if this fluctuation is being observed in the 50% region, it would be marked as a normal operating mode.

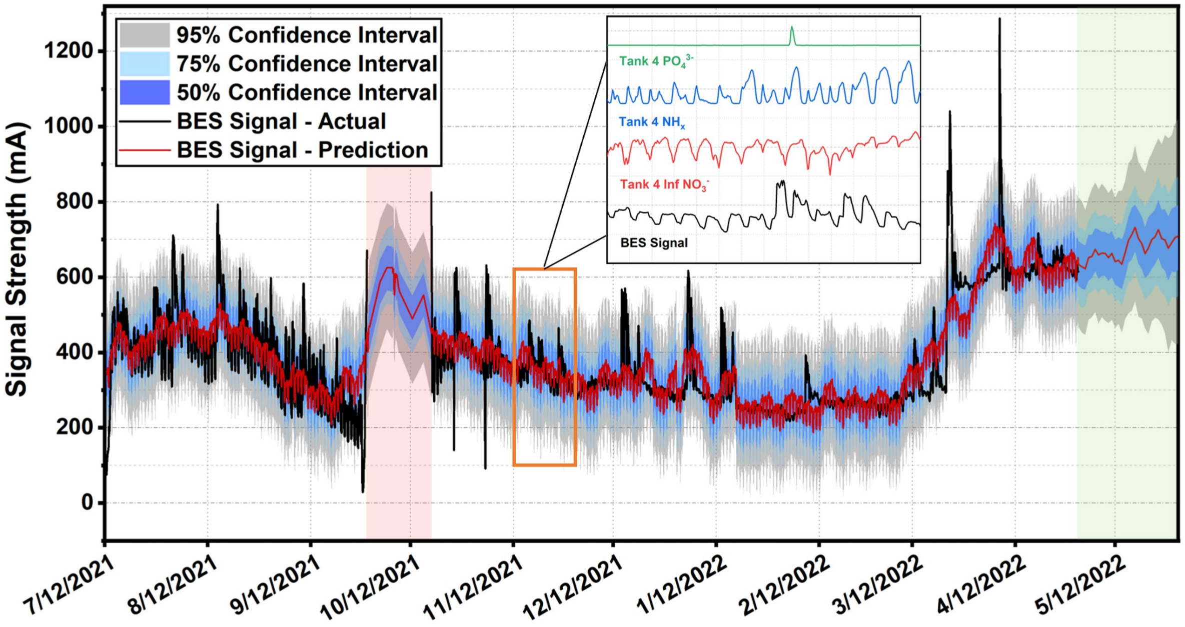


Figure 6. FB prophet prediction of the BES signal corresponding to different confidence interval ranges. The figure inset shows the structure of the BES signal plotted against other online sensors in the aeration channel as an example (P sensor undergoing a temporary failure in that time window).

In addition to developing this BES signal and influent C loading stability monitoring tool, the prediction feature of the FB prophet model was utilized to make long-term predictions into the future. At the time of the analysis, the last recorded datapoint obtained from this case study was on 04/30/2022. Therefore, the FB prophet model was used to predict the behavior of the BES signal one month into the future until 05/30/2022, which is shown as the light green shaded area in Figure 6. It is important to note that unlike the LSTM model that used a sliding window to predict future values based on previous readings, the FB prophet predictions were solely based on the algorithm, as there was no BES signal recorded after 04/30/2022 at the time of the analysis.

Figure 7 shows the actual BES signal daily averaged values that were recorded after the case study dataset, plotted against the FB prophet daily predictions. As demonstrated in the dark blue region, the model successfully provided a monthly prediction of the normal operational region within the 50% confidence interval, except in two spiked regions that were related to increased influent organic loading at the plant. Both recorded spikes in the BES signal, had surpassed the 75% confidence interval (warning region) and entered the 95% confidence interval region which was indicative of a plant influent shock loading event. This provides further evidence that the developed model can be used to successfully detect/predict plant influent shock loading events and prevent the implications that such events have on BNR efficacy.

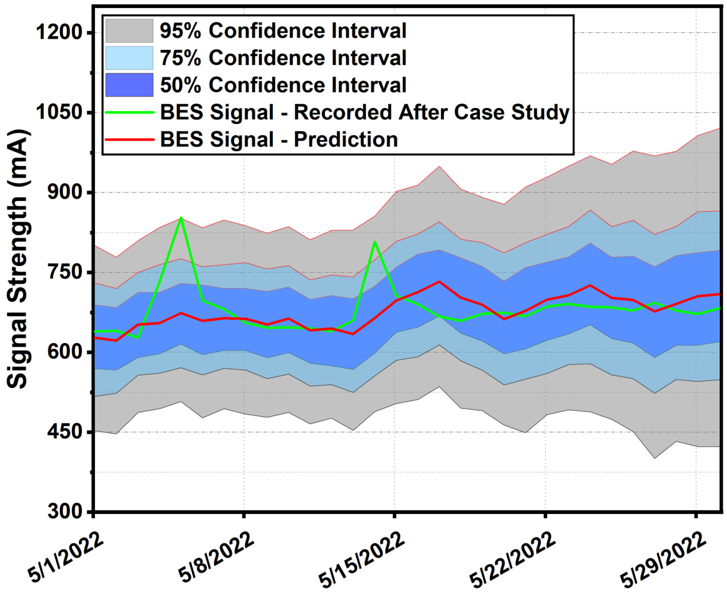


Figure 7. Comparison between the long-term monthly prediction of the BES signal and the actual recorded signal values after the case study and during May 2022.

Another feature of the FB prophet model was used to decompose the BES signal into its structural components to reveal the underlying repeating cycles that impacted the signal. Figure 8 demonstrates the repeating daily and weekly patterns that were detected by the model. The daily pattern shows a drop in signal intensity starting from around 09:30 PM and reaching its minimum at approximately 04:30 AM. At around 04:30 AM, the signal intensity starts to increase and reaches its maximum from 11:00 AM to 02:30 PM, and shows another peak at around 09:10 PM. Additionally, the weekly pattern shows an increase in signal intensity starting from mid-Monday which continues until Friday afternoon, and drops again during the weekend (Friday afternoon until Sunday morning).

After sharing these results with the HRSD Nansemond plant, it was concluded that these patterns aligned well with the influent organic loading variation patterns observed historically at the plant, except with a slight delay on Mondays. These results indicate that the BES signal obtained from the HRSD Nansemond plant is highly structured and contains valuable operational information. It was demonstrated that the BES signal not only provides an opportunity for a real-time monitoring tool of the plant's influent wastewater organic loading variations, but it can also be used as a strong predictor of BNR efficacy at the plant and act as a reliable and robust feedforward component within a unified control strategy. It is important to note that if more data were available in this case study (1-2 years of data), additional monthly, quarterly, and yearly patterns could have been presented too. Another figure showing the impact of holiday events on the plant's BES signal intensity and influent organic loading is presented in the Appendix.

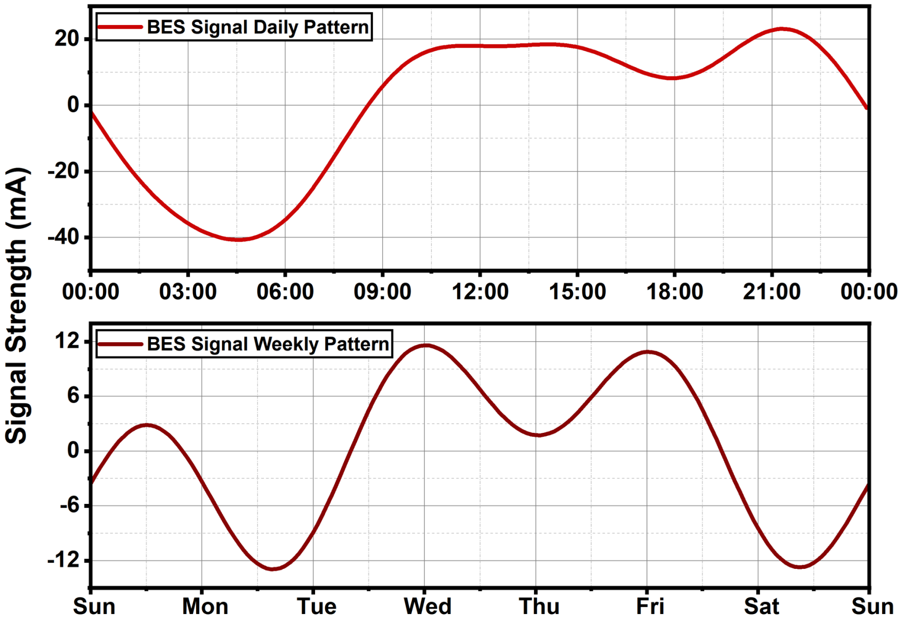
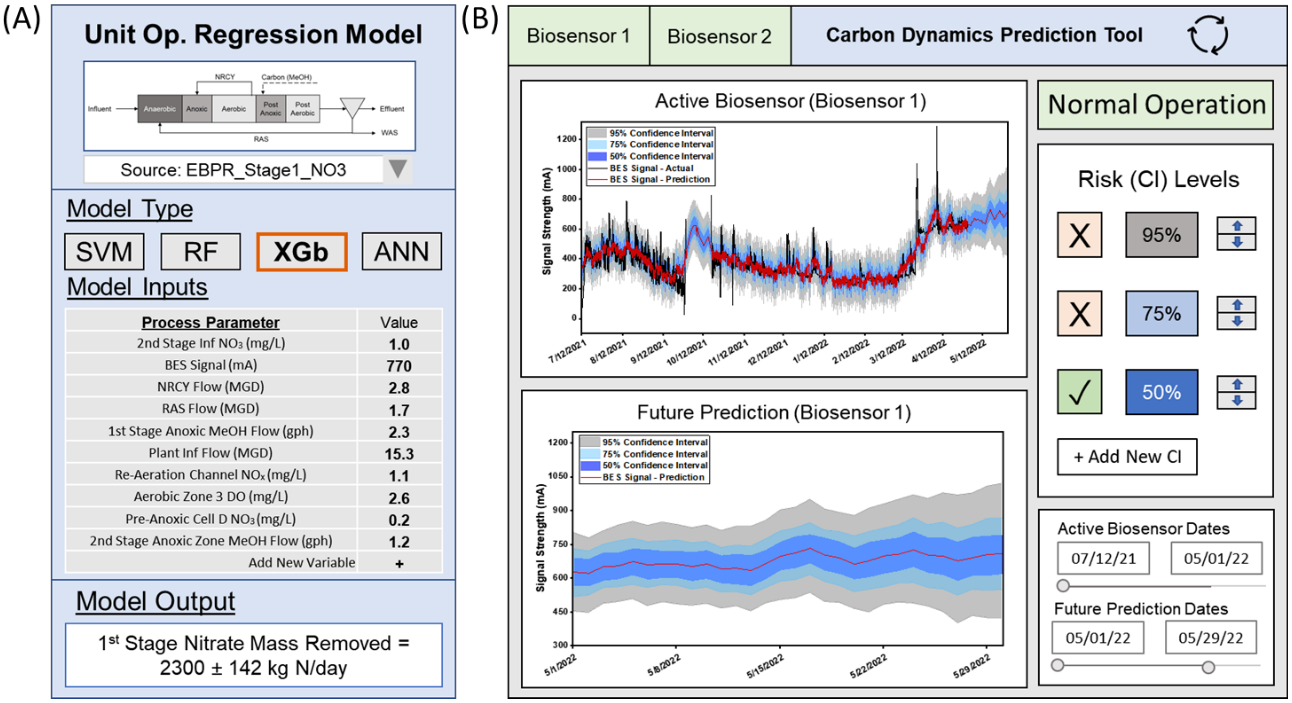


Figure 8. Daily/weekly cycles detected from the BES signal. The identified signal cycles followed the repeating daily/weekly influent organic loading patterns observed at the plant.

7 Software Dashboard Integration

Despite the immense potential that AI offers WRRFs to fully utilize their available data for optimized plantwide operation, developing functional models and using them for decision-making can be a challenging process for users who may not necessarily have a data science background and are experts in other fields. Therefore, an interface that integrates all the desired AI functionalities based on a WRRF's needs would be a crucial step in encouraging the application of AI technology in the water/wastewater sector. Thus, a conceptual software dashboard integrating all the features developed in this solution into an easy-to-use interface was developed and shown in Figure 9.

Figure 9A shows the regression software dashboard where a user can first choose their desired algorithm and then input a variety of different variables used in model development to predict the mass of nitrate removed in the 1st stage anoxic zone. This parameter can also be easily converted to removal percentage if intended, and the plant operator could try a variety of different scenarios to maximize the mass of removed nitrate while minimizing the NRCY flow rate and chemical requirements. Additionally, Figure 9B interface is specifically used for the BES signal. The main uses of this interface are to choose the active sensor screen, actively monitor the plant's influent organic loading rate using the BES signal, predict its expected values in the future within the desired timeframe, and adjust monitoring/prediction confidence intervals. Using this software dashboard enables everyone without any data science background to take advantage of the benefits that AI offers for process optimization and monitoring.



Source: EBPR_Stage1_NO3

Process Parameter	Value
2nd Stage Inf NO_3 (mg/L)	1.0
BES Signal (mA)	770
NRCY Flow (MGD)	2.8
RAS Flow (MGD)	1.7
1st Stage Anoxic MeOH Flow (gph)	2.3
Plant Inf Flow (MGD)	15.3
Re-Aeration Channel NO_x (mg/L)	1.1
Aerobic Zone 3 DO (mg/L)	2.6
Pre-Anoxic Cell D NO_3 (mg/L)	0.2
2nd Stage Anoxic Zone MeOH Flow (gph)	1.2
Add New Variable	+

1st Stage Nitrate Mass Removed = 2300 ± 142 kg N/day

☒ 95%
☒ 75%
☑ 50%

Active Biosensor Dates: 07/12/21 – 05/01/22

Future Prediction Dates: 05/01/22 – 05/29/22

Figure 9. Software dashboard integrating all the AI models developed in this analysis into a user-friendly and easy-to-use interface.

8 Implementation of AI and the Next Steps for This Solution

This solution provided the following insights for the utility partner and the team:

- **Team Formation**: Implementation of AI for an intelligent water system in WRRFs for decision making and process optimization requires an interdisciplinary approach. Developing an intelligent water system demands an extensive background in water/wastewater process engineering, data science and analysis, visualization, and data integration. The team presenting this solution could not have overcome the challenges faced in this analysis if it were not formed from members with different process engineering and data analytics backgrounds.
- **Main Challenges**: The challenges faced by this team mainly included finding the proper data collection/cleaning/preprocessing techniques, choosing the right algorithms that could help solve the problem statement, and ensuring that the developed models could help the plant. The interdisciplinary background of this team, extensive literature review, and regular meetings that were held with the utility partner and the technology provider were the key to overcoming these challenges
- **Implementation of AI for Decision-Making**: One of the biggest challenges that WRRFs face for fully utilizing AI could be the lack of a proper bridge between extensive AI-focused model development practices and a user-friendly and easy-to-use interface. Appropriate attention should be given to presenting AI benefits in a manner that can be utilized by individuals who may not necessarily have a data science background and are experts in other process-related fields. This is where individuals with both process engineering and data science backgrounds could help WRRFs to better communicate the advantages that AI offers.

- **Implementation of This Solution by Other Utilities and Next Steps**: The team from the University of Illinois at Urbana-Champaign is currently conducting other BES case studies in several WRRFs across the United States and is planning to employ the knowledge gained from the HRSD Nansemond project in other case studies. This includes similar signal processing and AI model development with a focus on BES monitoring technology to offer a unified and robust C monitoring framework that could be readily utilized by other utilities.

9 Conclusions and Benefits to Utilities

Many WRRFs utilize feedback control system to optimize their wastewater treatment and resource recovery processes. These feedback control systems are used to adjust different operational parameters which include but are not limited to recycle flow rates, chemical dosing requirements, etc. Online sensors usually require frequent calibrations or may provide delayed process-related information. This could complicate an optimized control of treatment and BNR processes leading to diminished effluent quality and increased operational costs associated with an overuse of chemicals or aeration energy requirements.

One promising solution is to add a robust feedforward component to a strict feedback control system which does not experience delays or require frequent downtimes for calibration and maintenance. BES technology offers all these characteristics as a novel biomonitoring tool. This technology, which uses electroactive biofilms to provide real-time amperometric data of readily available organic compounds in wastewater, has recently been developed to be used in an industrial scale setting and does not require frequent calibration/maintenance, has a low operational cost, and can be used to monitor and track the C dynamics at WRRFs.

AI and advanced statistical models are ideal tools to fully utilize BES technology as a feedforward control system component within an intelligent water system. This detailed analysis on the HRSD Nansemond case study was the first to provide the following insights:

- MI provided evidence for existence of a direct and strong relationship between the mass of nitrate removed in the 1st stage anoxic zone and BES signal measurements. The BES signal had the strongest general relationship out of all online sensors with the mass of nitrate removed in the 1st stage anoxic zone. This points to the ability of BES signals to track C availability/metabolism within a treatment/recovery process.
- DL and ML regression models demonstrated high predictive performance for the mass of nitrate removed in the 1st stage anoxic zone. Lack of a significant difference between the predictive performance of the different models indicated that in such applications with low-medium sized input datasets, choosing the right model depends primarily on training/prediction speed and straightforwardness of model development rather than prediction efficiency.
- In addition to providing evidence in support of univariate BES signal forecasting over a multivariate model, it was demonstrated that the FB algorithm can very well track the dynamic behavior of the signal and make accurate long-term predictions into the future. The daily averaged monthly predictions presented in this study aligned well with the normal operational range of the recorded BES signal. This feature can directly be used by plant operators to predict future influent organic loading events and have a real-time BES signal/organic loading stability monitoring tool to detect plant influent events in a timely manner.

This proposed intelligent water system solution can help the HRSD Nansemond plant evaluate a variety of different scenarios to predict their nitrate loading removal and optimize/automate their recycle flow and chemical dosing rates accordingly. The plant operators will be able to observe the changes in the mass of removed nitrate given specific BES signal readings (indicative of readily biodegradable C in influent), methanol dosing, and NRCY flow rates and therefore, minimize chemical and pumping energy requirements while ensuring a high effluent quality. Additionally, using the developed BES signal forecasting model, the plant will be able to make long-term predictions for the normal operational range of their influent BES signal and have a rapid plant event detection tool that could offer significant operational advantages. The developed intelligent water system in this study can be used in other utilities across the country for an optimized feedforward control strategy that can be integrated in already existing control systems.

10 Disclosures

The results and analysis presented in this report have been shared with the R&D department of the provider of the BES probes (SENTRY, Canada) prior to submission. The obtained original data used in this study is the result of a set of ongoing case studies that is being conducted by this team in several WRRFs across the United States to develop an intelligent influent organic monitoring system. No financial support from any person(s), institution, and/or company was received to prepare this report and participate in the 2022 LIFT Intelligent Water Systems Challenge.

11 References

[1] "Nansemond Treatment Plant Pipe Installation | HRSD.com." https://www.hrsd.com/nansemond-treatment-plant-pipe-installation (accessed May 10, 2022).

[2] P. E. Latham and Y. Roudi, "Mutual information," *Scholarpedia*, vol. 4, no. 1, p. 1658, Jan. 2009, doi: 10.4249/scholarpedia.1658.

[3] "scikit-learn: machine learning in Python — scikit-learn 1.1.2 documentation." https://scikit-learn.org/stable/ (accessed Aug. 15, 2022).

[4] "XGBoost Python Package — xgboost 1.6.1 documentation." https://xgboost.readthedocs.io/en/stable/python/index.html (accessed Aug. 15, 2022).

[5] "Keras: the Python deep learning API." https://keras.io/ (accessed Aug. 15, 2022).

[6] F. Hamami and I. A. Dahlan, "Univariate Time Series Data Forecasting of Air Pollution using LSTM Neural Network," in *2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, Oct. 2020, pp. 1–5. doi: 10.1109/ICADEIS49811.2020.9277393.

[7] S.-Y. Shih, F.-K. Sun, and H. Lee, "Temporal pattern attention for multivariate time series forecasting," *Mach Learn*, vol. 108, no. 8–9, pp. 1421–1441, Sep. 2019, doi: 10.1007/s10994-019-05815-0.

[8] S. A. Emaminejad *et al.*, "Statistical and microbial analysis of bio-electrochemical sensors used for carbon monitoring at water resource recovery facilities," *Environ. Sci.: Water Res. Technol.*, Feb. 2022, doi: 10.1039/D1EW00653C.

[9] "Prophet: forecasting at scale - Meta Research," *Meta Research*. https://research.facebook.com/blog/2017/02/prophet-forecasting-at-scale/ (accessed Aug. 05, 2022).

[10] "What is XGBoost?," *NVIDIA Data Science Glossary*. https://www.nvidia.com/en-us/glossary/data-science/xgboost/ (accessed Aug. 04, 2022).

[11] "Time Series - LSTM Model." https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm (accessed Aug. 04, 2022).

[12] M. Krieger, "Time Series Analysis with Facebook Prophet: How it works and How to use it," *Medium*, Feb. 02, 2022. https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a (accessed Aug. 10, 2022).

12 Appendix

12.1 Data QA/QC Considerations

The BES data was collected with a minute resolution, and later converted to hourly data for the regression and time-series forecasting models. Initial data cleaning strategy was chosen based on discussions with the utility partner and mainly included identifying and deleting rows of data that corresponded to online sensor failures and mis-readings (e.g., zero and NaN inputs in the regression dataset). More specifically, online sensor measurements with negative values were the result of sensor malfunction and were deleted form the dataset. Additionally, if an online sensor recorded the exact same value over a long period of time, it was another indicator of sensor malfunction, and their corresponding values were deleted from the final dataset.

The original dataset included 42 different variables, many of which were related to individual parallel tank measurements within a unit process. To reduce the number of variables, individual tank measurements of specific unit processes were averaged to have a final dataset with 16 variables. The correlation analysis step was conducted on the cleaned and averaged dataset. A final preprocessing step was done where all the selected input features were standardized by removing the mean and scaling to unit variance using Python scikit-learn "StandardScaler." After this step, the input dataset was ready to be used for regression model development. The goal was to choose data quality over quantity, and these steps for taken to ensure a proper data QC.

12.2 Regression Models

SVM was the first choice as a supervised ML algorithm as it is less prone to overfitting and has a high capability in modelling complex nonlinear interactions, but it can be a slow performer on very large datasets. The second model was based on an RF algorithm. RF models use ensemble methods by constructing a number of decision trees during training time and making final predictions based on the average prediction of the individual trees. In other words, an RF model outputs an optimal result based on the aggregation of the results of many individual decision trees. This makes RF models an accurate predictor, but a large number of trees can make the model slow in making predictions once it is trained; therefore, this hyper parameter should be carefully optimized during model training. Unlike an RF model that uses a technique known as "bagging" to build full decision trees in parallel, XGboost which was the third developed model, is an implementation of gradient boosted decision trees that uses "boosting" to combine weak learners sequentially so that a new tree corrects the errors of its predecessor [10].

This algorithm is designed for speed and accuracy and has recently been dominating the field of applied ML. Finally, an ANN model, as a DL algorithm, was also developed and tested. ANN is a flexible alternative for datasets with high nonlinearity but can sometimes be difficult to interpret as a black-box model. Different strategies were used to ensure the robustness of the developed models. In ML models, grid search with cross validation was used to find the optimum model parameters. The model parameters that we considered for model development are as follows (final parameters were bolded):

- SVM:
    - "Kernel" as the problem-solving function: 'rbf'
    - "C" or the penalty parameter of the error term: [1, 10, 50, 100, 200, 400, **500**]
    - "gamma" or the parameter for non-linear hyperplanes: [0.01, 0.05, 0.1, 0.5, **1**, 2]
- RF:
    - "max_depth" to determine the depth of each tree in the forest: [18, 19, 20, **21**, 22, 23]

  - "n_estimators" or the number of trees to build before taking the average of predictions: [300, 400, 500, **600**, 700]
- XGboost:
  - "n_estimators" or the number of trees to build before taking the average of predictions: [1000, 1500, 2000, **2500**, 3000]
  - "eta" as the learning rate or shrinkage which shrinks the feature weights to make the boosting process more conservative and prevent overfitting: [0.01, **0.05**, 0.1, 0.2]
  - "gamma" which specifies the minimum loss reduction required to make a split: [**0**, 0.05, 0.1]
  - "max_depth" specifying the maximum depth of each tree: [6, **8**, 10]
- ANN:
  - Activation function: "ReLU"
  - "Sequential" model
  - Dense layer 256, 128, 64, 32, 16 (ReLU), 1 (Linear)
  - Final learning rate: 0.01
  - Early stopping was used to prevent overfitting

12.3 Time-Series Forecasting Models

LSTM-RNN was used as a powerful DL architecture. RNN is a special kind of neural network that has a short-term memory functionality which is a very useful feature for time-series data as past inputs leave a footprint. LSTM adds to this functionality by implementing both short-term and long-term memory functionalities which allow it to learn long-term dependencies in sequential data [11]. The raw BES signal was used for development of time-series forecasting models, showcasing one of the advantages of using BES probes as they provide continuous and uninterrupted measurements with minimum maintenance requirements. Similarly, uninterrupted pH and temperature data was provided by the plant and used as additional variables in the multivariate time-series forecasting model. For time-series forecasting models, hourly averaged data were used as inputs. The DL models had the following characteristics:

- Univariate LSTM:
  - Sliding window with a window size of 6, meaning the model would go over the previous 6 datapoints to predict the 7th value. This process was repeated until the model was trained. As data frequency was hourly, the model provided hourly predictions. This time window could be larger as desired (e.g., 10-20 hours or daily), but that would also require more data to ensure a satisfactory prediction performance. These types of models can provide very accurate predictions (daily, monthly, etc.) especially if large historical datasets consisting of years of recorded time-series data are available, which was not available in this case study.
  - "Sequential" model.
  - Input layer (6,1)
  - Layers: LSTM 64, Dense 8 (ReLU), Dense 1 (Linear)
  - Trainable parameters: 17,425
  - Learning rate: 0.001, epoch: 35
- Multivariate LSTM:
  - Input dataset consisted of hourly BES signal datapoints, plant influent wastewater pH, and temperature values with matching timestamps.
  - "Sequential" model.
  - Input layer (6,3)

  - Layers: LSTM 64, Dense 8 (ReLU), Dense 1 (Linear)
  - Trainable parameters: 17,937
  - Learning rate: 0.01, epoch: 35

Moreover, the FB prophet algorithm which was chosen as the final approach for univariate analysis is the sum of three functions of time and error at its core as follows:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$$

Where g(t) corresponds to the growth term and change points which models the overall trend of the data, s(t) is a Fourier Series as a function of time which is the seasonality function, h(t) considers holiday effects and allows the FB prophet model to adjust forecasting when a major national event or holiday might change the forecast and has built-in customizable functions to account for holidays in different countries, and lastly, ε_t is the error term [12]. The FB prophet model input dataset was prepared by changing the timestamp and BES signal column names to "ds" and "y", so they become recognizable by the model package. The "Prophet" package was the imported, and the model was developed with the following characteristics:

- FB Prophet:
  - Confidence interval: 0.95, 0.75, 0.50
  - Seasonality 1: "quarterly"/period = 91.5/Fourier order = 10
  - Seasonality 2: "monthly"/period = 30.5/Fourier order = 5
  - Future dates: periods = 30, freq = 'D'
  - Holiday effect: country_name = "US"

Another important feature of the FB prophet algorithm is its ability to capture holiday effects. The package itself has a complete list of countries with their respective official holidays. The official holidays list of the United States was used in this report. It is important to note that these holidays are customizable, and the user can add different dates that are not listed as official holidays such as major sporting events (not considered in this report). After providing evidence about the strong relationship between the BES signal and C dynamics at the HRSD Nansemond plant, the holiday effects was on the BES signal to provide the plant with a quantitative/qualitative analysis of the impact of each holiday on the influent organic loading variations, as shown in Figure A1.

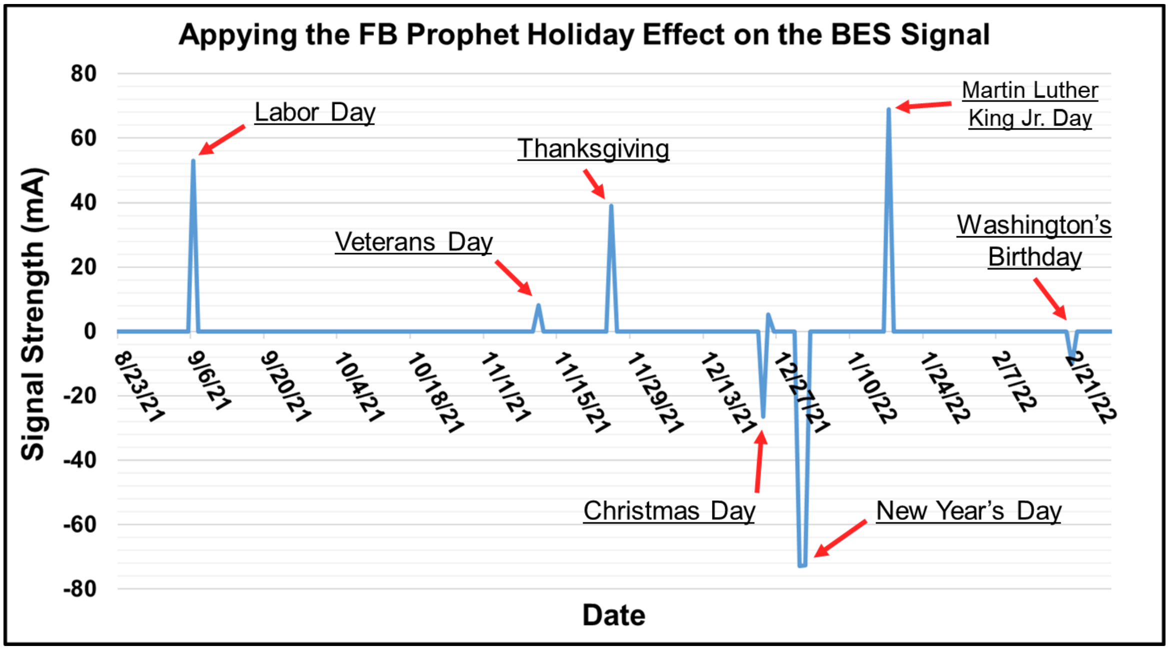


Figure A1. BES signal variations impacted by discharge events during United States holidays with potential implications on the HRSD Nansemond plant's influent organic loading.

Analyzing holiday impacts was not part of the problem statement of this report, and therefore this figure was only presented in the Appendix to show the utility of this method. If other utilities find any operational interest in analyzing the impact of holidays on their influent characteristics, this methodology could provide them with useful insights.

12.4 Abbreviations

A full list of abbreviations used in this study can be found in Table A1:

Table A1. List of abbreviations used in the study

Abbreviation	Description	Abbreviation	Description
ANN	Artificial Neural Network	ML	Machine Learning
AI	Artificial Intelligence	MNR	Mass of Nitrate Removed
BNR	Biological Nutrient Removal	N	Nitrogen
BES	Bio-Electrochemical Sensor	NaN	Not a Number
COD	Chemical Oxygen Demand	NRCY	Nitrified Recycle
C	Carbon	RO	Reverse Osmosis

CI	Confidence Interval	OP	Orthophosphate
DO	Dissolved Oxygen	P	Phosphorous
DL	Deep Learning	RAS	Return Activated Sludge
FB	Facebook	RF	Random Forest
HRSD	Hampton Roads Sanitation District	RNN	Recurrent Neural Network
HRT	Hydraulic Retention Time	RMSE	Root Mean Square Error
LSTM	Long Short-Term Memory	SVM	Support Vector Machine
MGD	Million Gallons Per Day	TIN	Total Inorganic Nitrogen
MI	Mutual Information	WRRF	Water Resource Recovery Facility