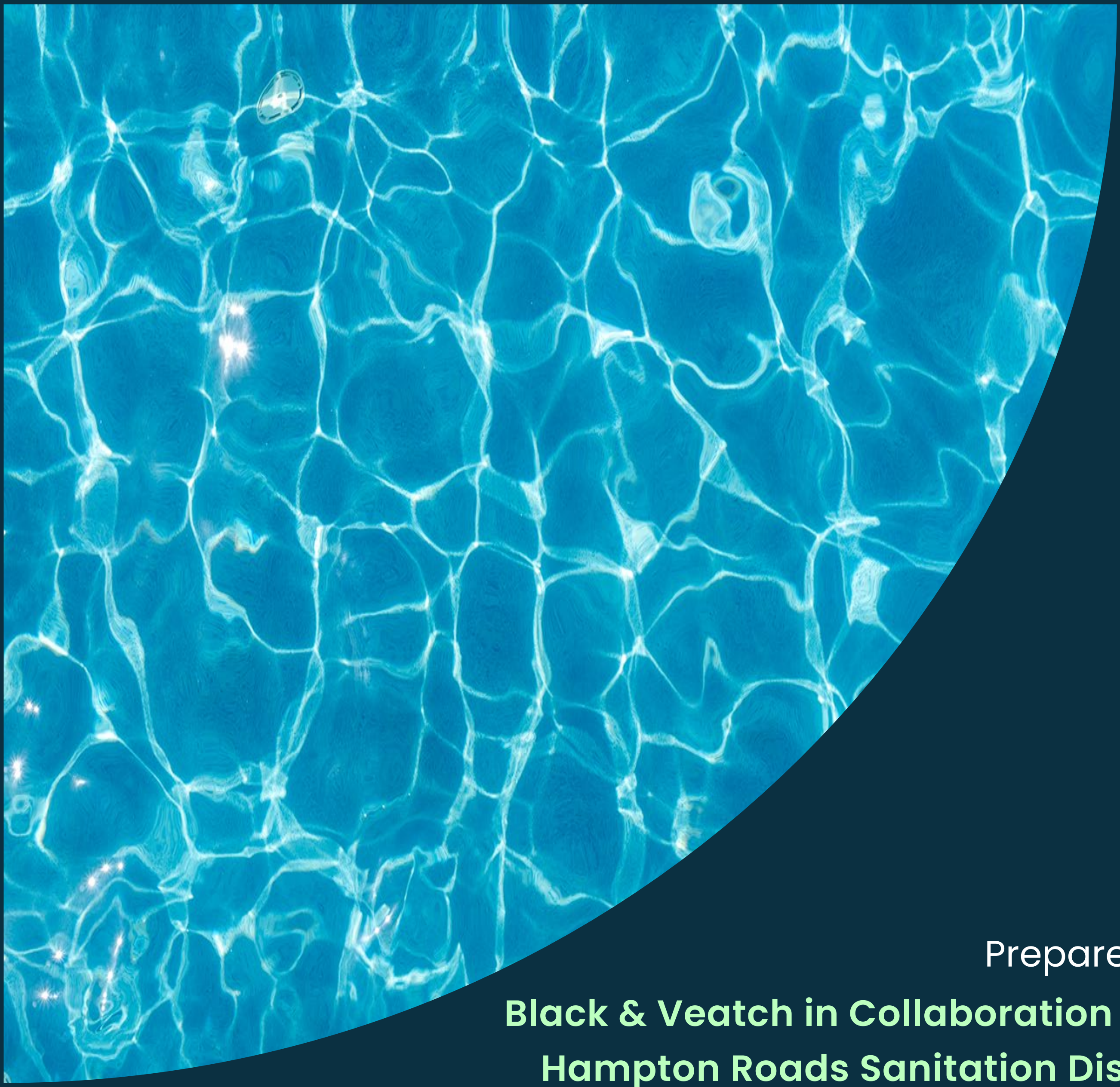
2024 LIFT Intelligent Water Systems Challenge

July 2024

A HYBRID INTELLIGENT CONTROL SYSTEM TO OPTIMIZE CHLORINATION PROCESSES: AN HRSD CASE STUDY

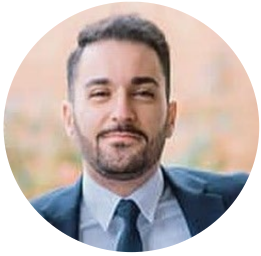


Prepared BY

Black & Veatch in Collaboration with Hampton Roads Sanitation District

Contents

Meet the team:



Aryan Emaminejad, Ph.D.
(Team Lead), AI/ML Engineer
Black & Veatch, Chicago, IL, USA
emaminejada@bv.com

Area of Expertise:

- AI/ML Modeling, Data Science
- Wastewater Process Eng.

Roles and Responsibilities:

- ML/Mechanistic Model Development
- Model Evaluation
- Team Coordination

Jeff Sparks, P.E., Utility Partner
Digital Water
HRSD, Nansemond, VA, USA
jsparks@hrsd.com

Area of Expertise:

- Plant-Wide Operations
- Wastewater Process Eng.

Roles and Responsibilities:

- Project Scope Definition
- Data Preparation, ML Model Deployment

Patrick Dunlap, P.E.
Wastewater Process Engineer
Black & Veatch, Philadelphia, PA, USA
dunlappj@bv.com

Area of Expertise:

- Plant-Wide Simulations
- Wastewater Process Eng.

Roles and Responsibilities:

- Mechanistic Model Development
- Project Solution Analysis
- Project Scope Definition

Amirsalar Bagheri
Ph.D. Candidate
Kansas State University, KS, USA
amirbg@ksu.edu

Area of Expertise:

- AI/ML Modeling, Data Science

Roles and Responsibilities:

- ML Model Development
- Model Evaluation

1. Background and Problem Statement

Efficient control of chlorination and sodium hypochlorite dosage is challenging as it requires a balanced tradeoff between operational costs and effluent quality. Minimization of sodium hypochlorite dosage could reduce operational costs but potentially lead to insufficient chlorine residual levels in the effluent. On the other hand, although an elevated sodium hypochlorite dosage rate would ensure sufficient chlorine residual in the effluent, it will significantly increase operational costs related to sodium hypochlorite consumption. A new chlorine dose controller for effluent disinfection is desired at the HRSD Nansemond Treatment Plant to minimize sodium hypochlorite dosage to the degree possible while ensuring sufficient chlorine residual levels to reliably meet permitted disinfection requirements. The new controller will incorporate a feedforward element into the existing feedback controller. This feedforward element will utilize a hybrid modeling approach combining a mechanistic chlorine decay model and a data-driven model to predict mechanistic model errors. This novel approach will provide the facility with a robust system for chemical dosing that simultaneously minimizes costs and greenhouse gas emissions (GHGs) associated with the production and hauling of sodium hypochlorite.

1.1 Why a Hybrid Model Approach?

Although artificial intelligence (AI) and machine learning (ML) models provide exceptionally high predictive capabilities, their industrial-scale adoption is more justifiable if the underlying mechanisms involved cannot be fully explained by mechanistic insights. Mechanistic models have been around for decades and can provide real-time insights into the prediction process in a transparent manner (i.e., a "white box" as opposed to a "black box"). Therefore, if a wastewater process can already be characterized by mechanistic models, a full-scale replacement of the mechanistic model with an AI/ML tool may not be fully justified due to the complexities involved with AI/ML models including model development, deployment, and maintenance. However, a promising application of AI/ML in wastewater treatment processes with industrially accepted mechanistic models is to use the predictive capabilities of AI/ML models for mechanistic model error correction, and therefore, increase the reliability of model predictive control (MPC) practices for full-scale water resource recovery facility (WRRF) operations.

The selected hybrid model approach in this case study allows for simultaneous utilization of a widely used first-order parallel decay mechanistic model for the prediction of chlorination requirements, and an ML model to correct the mechanistic model errors as the backbone of a reliable control system. Additionally, the predictive nature of the developed ML model allows the operators to prepare for the necessary adjustments in the control system and optimize the plant's operational costs and effluent quality.

2. Project Outline

The project outline for the successful development/deployment of the hybrid MPC tool is summarized below:

1. Optimize the chlorination process to minimize sodium hypochlorite dosage while maintaining satisfactory effluent chlorine residual levels.
2. Develop a parallel first-order mechanistic model with a rolling window to frequently update the model parameters and predict the decay rate.
3. Develop advanced time-series forecasting ML models to predict and correct the mechanistic model error and return an updated chorine decay rate and sodium hypochlorite dosage.
4. Deploy the developed multi-step tool for full-scale operational control and report the savings associated with sodium hypochlorite operational costs, GHGs, and monitor the effluent quality.

3. Solution

3.1 Introduction

Disinfection of public water supplies has been one of the greatest and most successful examples of human intervention in promoting public health and eliminating waterborne diseases caused by microbial pathogens.[1] One of the most widely used disinfection methods is chlorination which has been utilized on an industrial scale for decades. Disinfection in the context of WRRF operations is especially important because the final discharge is typically released to natural water bodies and could further be used for irrigation and recreational purposes which may lead to direct or indirect human contact with discharged effluent.[1] Typical forms of chlorine compounds used in WRRFs are elemental chlorine, hypochlorite, and chlorine dioxide; and, each one of them has its own operational advantages and disadvantages.[2]

The Nansemond plant uses sodium hypochlorite as the disinfecting agent which is in liquid form and forms hypochlorous acid and sodium hydroxide when it reacts with water. Sodium hypochlorite dosing requirements are governed by chlorine demand which can be defined as the difference between the added chlorine and residual chlorine leaving the plant after a certain period of contact time.[4] This demand results from organic and inorganic compounds in wastewater which react with added chlorine. Since these reactants and the kinetics associated with these reactants are variable, the chlorine demand is not constant. Chlorine reacting with organic compounds present in wastewater form chlororganic compounds which have a limited disinfection capacity; while chlorine reacting with ammonia forms chloramine (i.e., monochloramine, dichloramine, and trichloramine), a strong disinfecting agent.[3]

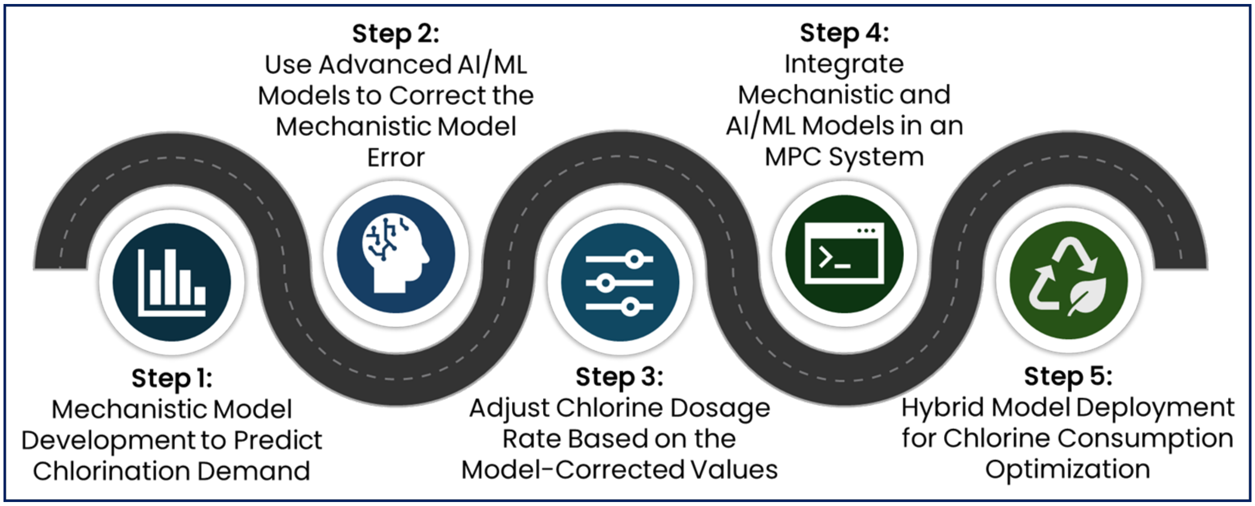


Figure 1. Overview of the chlorine optimization MPC deployment process using a mechanistic and AI/ML hybrid modeling approach.

Despite an understanding of chlorine demand mechanisms and the availability of mechanistic models for predicting chlorine residuals[6], the application of these models to chlorine dose control is limited in practice. This is likely because there is substantial heterogeneity in effluent composition and operating conditions which introduce an unacceptable degree of error to the model predicted chlorine demand. This error could lead to an under-dosing of chlorine and permit violations. This project aims to address this issue by utilizing a hybrid AI/ML model to correct the errors associated with the chlorine decay mechanistic model in real-time and optimize chlorine consumption using an MPC system (Figure 1). Following the deployment of this controller, operators at the HRSD Nansemond plant will be able to optimize the chlorine dosage MPC within an intelligent water framework to reduce dosages, and associated GHG emissions and operational costs, while consistently meeting disinfection requirements.

3.2 Mechanistic Model Development – Chlorine Decay

In an effort to systematically study the chlorine demand kinetics from real wastewater, a parallel first-order decay model was proposed and successfully demonstrated an adequate fit to real-world data according to Equation 1:[4]

$$\boldsymbol{C(t) = (C_0 \times x \times e^{-k_1 \times t}) + (C_0 \times (1 - x) \times e^{-k_2 \times t})} \qquad \text{Equation 1}$$

Where $C(t)$ and C_0 are the chlorine concentrations at time t and time 0, respectively, K_1 and K_2 are the decay constants for different chlorine decay pathways, where t is the chlorine contact reaction time, x (limited between 0 and 1) is the fraction of the initial chlorine concentration that decays with the K_1 rate constant, and (1-x) is the fraction of the initial chlorine concentration that decays with the K_2 rate constant.[6] The two reaction rate constants allow for differentiation of the slow and fast reacting portion of chlorine compounds. To account for the variations in parameter fitting values across different operational conditions, this mechanistic model was fitted over data collected from the HRSD Nansemond treatment plant using a real-time parameter fitting approach with a rolling two-week moving window.

The plant uses two chlorine contact tanks (CCTs) for disinfection. Therefore, this approach included calculating chlorine decay based on the dosed sodium hypochlorite and the measured chlorine concentration one CCT hydraulic retention time (HRT) into the future (approximately 30 minutes); this approach included calculating the difference between the initially dosed C_0 and the mechanics model's C_0 concentration given a C_t measurement at each time interval. The parameter fitting rolling window then moved over the chlorine decay data and used the previous two weeks of recorded data to generate a new set of fitting parameters every day. With this approach, a new mechanistic model with updated fitting parameters was obtained daily which ensured the accuracy of model parameters. Instead of using a typical curve fitting method for parameter fitting, the Nelder-Mead method was used to find the best model fitting parameters. Unlike curve fitting that uses least squares optimization and is sensitive to undefined or noisy derivatives, Nelder-Mead relies on a simplex-based optimization algorithm which does not require derivative computations, making it ideal for non-smooth functions.[5] Using fitted Nelder-Mead parameters, the mechanistic model predictions for chlorine concentration were compared with the recorded values (Figure 2).

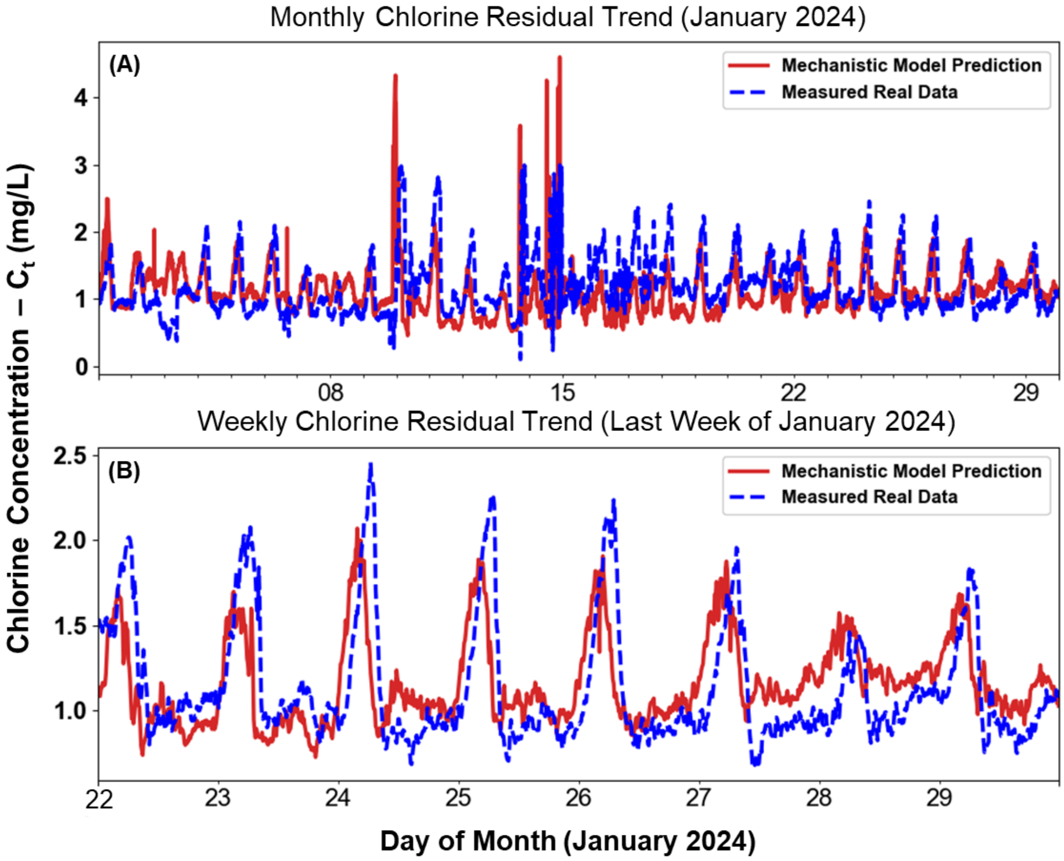


Figure 2.[1] Recorded chlorine residual vs. predicted values using a parallel first-order decay mechanistic model. The overall trend in January 2024 as a representative period (A) and its last week (B) demonstrate clear daily patterns.

The mechanistic model predictions closely followed the measured chlorine concentrations by chlorine sensors with a Root Mean Squared Error (RMSE) of 1.94 and a Mean Absolute Error (MAE) of 1.21 mg/L on the test dataset (Figure 2A). This close alignment indicated that the first-order parallel decay model was able to predict

[1] Figures with technical data were created using the Arial font.

chlorination requirements by accounting for both fast and slow decaying portions of chlorine compounds present in wastewater. Although chlorination requirement patterns were successfully captured by the model, a consistent model error was observed (Figure 2). This deviation from the recorded C_t values downstream was clearer when analyzing the individual daily patterns (Figure 2B). The observed error limits the ability of the MPC using only the mechanistic model to minimize sodium hypochlorite consumption while ensuring compliance with effluent disinfection and chlorine residual regulations. The next section will highlight how an AI/ML-based component to this control strategy allows the plant to predict and correct the daily mechanistic model error, thereby further optimizing its sodium hypochlorite dosage rates, maximizing operational cost and GHG emission savings.

3.3 Predictive AI/ML Models: Neural Prophet

Chlorination requirements at the HRSD Nansemond plant are dominated by repeating diurnal patterns (Figure 2). Therefore, the most suitable approach is to use an advanced time-series forecasting model that can uncover and predict these repeating patterns. One of the most promising time-series forecasting models is Facebook (FB) Prophet, which is an additive algorithm that models trend, seasonality, and holiday effects and can decompose any time-series data to reveal its underlying repeating patterns.[6] Despite its advantages, FB Prophet is a statistical model that has a limited predictive performance compared to artificial neural networks (ANNs) and does not incorporate autoregressive features into model scenarios where immediate past is a strong predictor of future values. To address these issues, Neural Prophet was developed as an extension to FB Prophet to incorporate autoregressive features in addition to utilizing ANN layers.[7] This unique combination allows a Neural Prophet model to uncover highly nonlinear and complex interactions within a time-series data while simultaneously using the immediate past values to predict into the future through autoregression.

Therefore, Neural Prophet was selected as the first algorithm to predict the chlorination mechanistic model errors (Figure 3). The date range for the raw data used for model development was from November 2022 through February 2024. The data was recorded with a 10-minute resolution, and 80% of the overall dataset was used for training while the remaining 20% was used for the test period. The Neural Prophet model in this study utilized a deep 4-layer neural network architecture with

12 lagged steps as the autoregression features (i.e., the past 2 hours of data) with a prediction horizon of 6 steps (i.e., 60 minutes into the future). Appropriate batch size and learning rate values were optimized by trial and error.

A mechanistic approach would involve predicting the chlorine concentration at a desired time interval (e.g., one HRT into the future) given a selected sodium hypochlorite dosing rate and C_o concentration. Following this approach, at each time interval, there was a difference between the mechanistically predicted C_t concentration given a selected C_0, and the measured C_t concentration approximately one HRT into the future.

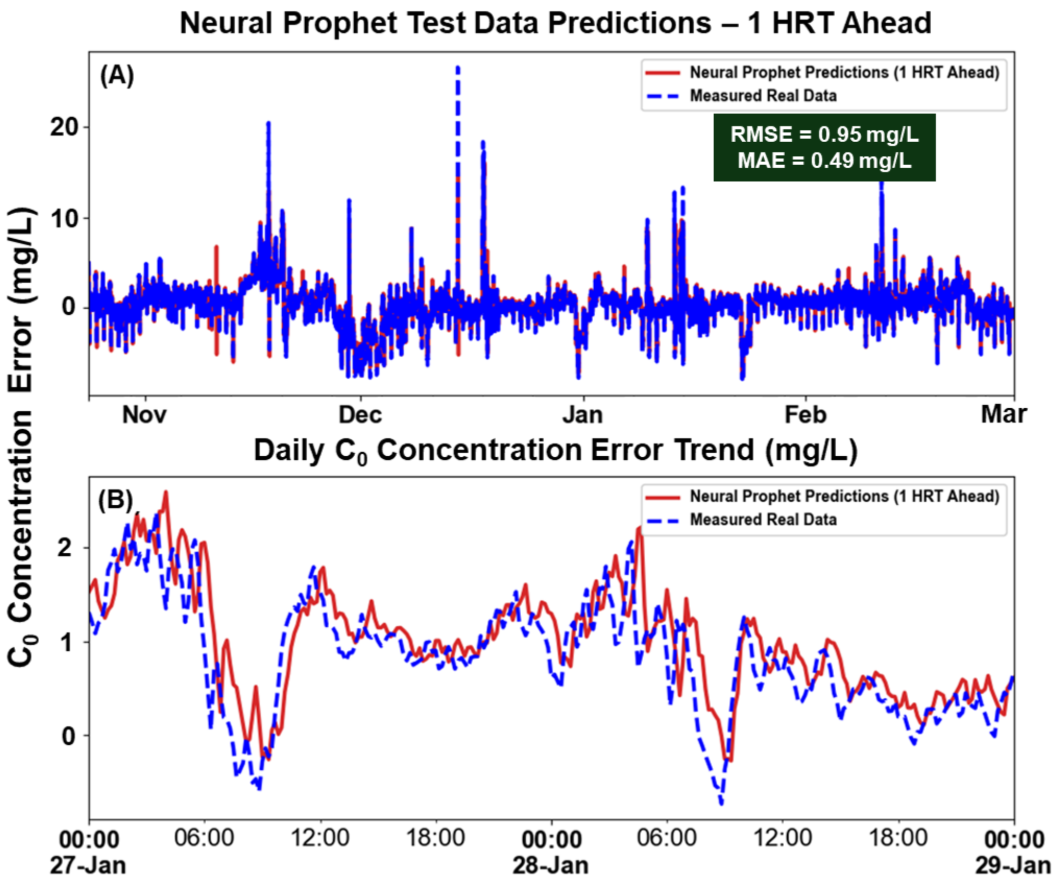


Figure 3. Neural prophet predictions 1 HRT into the future on the overall test data (A) and daily C_0 concentration error trends over a selected representative period (B).

Through back-calculation with the mechanistic chlorine decay fraction, it was possible to determine what C_0 should have been to achieve the desired C_t concentration, and therefore, correcting the initial sodium hypochlorite dose.

Therefore, the Neural Prophet model was trained to predict the mechanistic model's error in selecting the correct C_0 concentration at each 10-minute interval. Once this error between C_0 concentrations is accurately predicted, operators can adjust the target chlorine residual (C_t) concentration accordingly to optimize MPC performance for minimal dose and compliance risk.

The RMSE and MAE values were 0.19, 0.36 mg/L for the training dataset, and 0.68, 0.40 mg/L for the test dataset, respectively. The performance metrics demonstrated that the model had a high predictive performance both on the training and test datasets (Figure 3A). Despite this high predictive performance on the test dataset, visual inspection of the predictions indicated that when the chlorination requirements and the subsequent mechanistic model errors went through sudden peaks or dips (e.g., in early morning hours), the model made the predictions with a slight delay (Figure 3B). This delay could have either been caused by a lack of enough training data around peaks/dips to capture these sudden variation events or reflect the model's limitations in response to sudden peaks/dips in data. Although the overall accuracy of the Neural Prophet model to correct the mechanistic model's prediction was high, another advanced time-series forecasting approach was evaluated to determine the best modeling strategy for this project to minimize the prediction delay. The final evaluated state-of-the-art approach was based on Transformers which are the building blocks of large language models (LLMs) including ChatGPT and have recently gained attention for time-series forecasting due to their ability in modeling long-range and sequential dependencies using self-attention mechanisms.[8]

3.4 Predictive AI/ML Models: Transformers

The emergence of Transformers as a new deep learning architecture revolutionized Natural Language Processing which advanced generative AI.[9] Unlike Recurrent Neural Networks (RNNs), which processed sequential data in a recursive order to observe temporal patterns, Transformers use the self-attention method to capture the temporal patterns between all data points in a sequence. It is important to note that the self-attention method ignores the chronological order of the data points in a sequence; therefore, positional encoding is implemented in Transformers to keep track of the temporal order after extracting the attention scores. Overall, the encapsulation of positional encoding, the self-attention method, and feedforward neural networks (FNNs) in an encoder-decoder architecture forms a Transformer block (Figure 4).[10] The employment of FNNs reduces the training computational cost, allowing for the training of large Transformer-based models on larger datasets and very long sequences.[11]

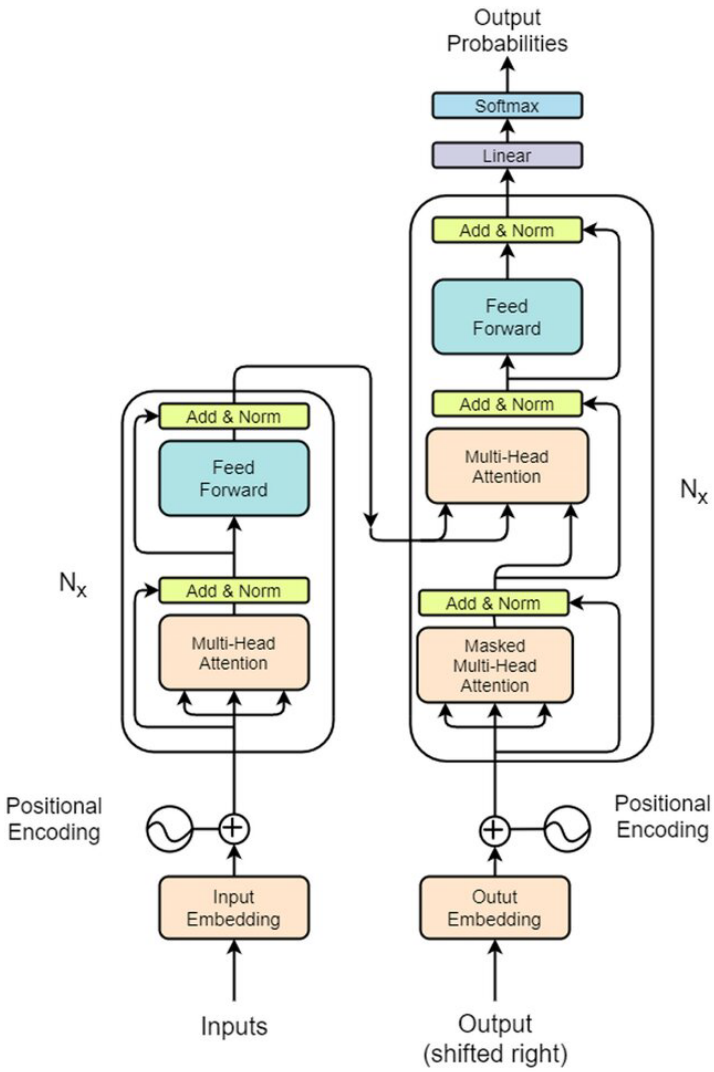


Figure 4.[10] Graphical representation of a Transformer architecture.[2]

The advantages of Transformers led to the formation of famous LLMs like ChatGPT and Llama. In this context, Transformers consider each sentence as a word sequence, finding the temporal patterns between words regardless of the chronological word order in the encoder part. This information is further used in the

[2] Figure sourced from reference 10.

decoder part to carry out language tasks like machine translation or text generation in a generative AI chatbot based on previous sentences.[12,13] By taking advantage of this perspective, Transformers were evaluated as the second strategy for predicting the mechanistic model's error in selecting the correct C_0 concentration.

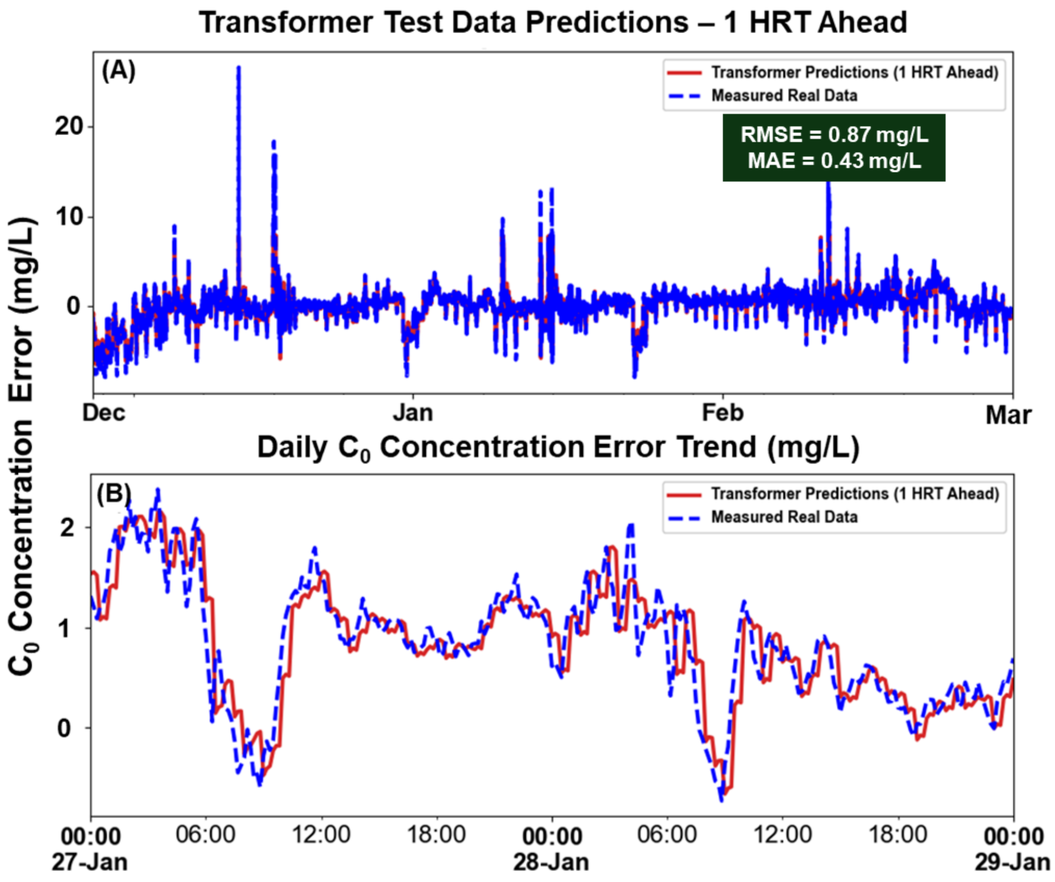


Figure 5. Transformer predictions 1 HRT into the future on the overall test data (A) and daily C_0 concentration error trends over a selected representative period (B).

In the context of Transformers and their role in LLMs, sequential chlorine requirement predictions ordered in time were regarded in a similar manner as individual words ordered in a sentence. The combination of positional encoding and the self-attention method not only captured the seasonality and trend of the mechanistic model's error but also observed the contextual essence of the chlorination process, leading to even more accurate predictions with less delay. The Transformer proposed in this study was trained on the dataset produced by the Nelder-Mead algorithm that contained updated mechanistic model predictions and back-calculated C_0 concentrations for each day in 10-minute frequencies. The dataset was split into 70% training, 10% validation, and 20% testing purposes. Benefiting from long sequence learning, the Transformer model took 144 previous steps as the historical sequence (i.e., the past day) with a prediction horizon of 3 steps (i.e., 30 minutes into the future, 1 HRT ahead). Furthermore, the architecture of the proposed Transformer model contained 4 encoder and 4 decoder layers. The FNNs in the Transformer were activated by the Gaussian Error Linear Unit (GELU) function followed by dropout layers to prevent overfitting. Furthermore, using hyperparameter optimization, an optimal batch size and learning rate were used for training.

The RMSE and MAE values were 0.42, 0.28 mg/L for the training dataset, and 0.60, 0.33 mg/L for the test dataset, respectively. Similar to the Neural Prophet model, these metrics demonstrated an exceptionally high predictive performance by the Transformer model (Figure 5A). Additionally, visual inspection of model predictions in the test dataset indicated minimal prediction delay compared to the Prophet model (Figure 5B). This emphasized the capability of Transformers in modeling the dynamic nature of C_0 concentration error trends and reflecting it in the hybrid model. Training Transformer models is known to be a data-intensive task and the current training dataset with around 50,000 datapoints is relatively small. It is safe to assume that if more data were available (e.g., several years instead of 2), an even more accurate Transformer model However, it should be noted that the self-attention method is computationally intensive, which creates a trade-off between accuracy and training time — a crucial consideration for MPC deployment.

3.5 Dashboard Integration

A fully functional online dashboard was created using Streamlit in Python to test the real-time performance of the hybrid chlorination control system (Figure 6). The online dashboard integrates the individual components of this solution by i) outputting the daily chlorine decay function parameters using the last two weeks of data, ii) creating live figures representing the recorded/predicted C_t concentrations and corrected C_0 concentrations using the mechanistic and ML models, and iii) outputting the current sodium hypochlorite dosage rates vs. the mechanistic and ML-corrected dosage rates. Figure 6 represents the real-time performance of this control system by depicting the model predictions (blue lines) compared with the recorded values in the test dataset (red lines). One of the main observations during this study was related to the impact of sudden chlorine demand spikes on the performance of the predictive models. When the system was relatively stable, although the 1-step-ahead predictions from the mechanistic model were relatively close with the recorded C_0 values in the test dataset, the ML models outperformed the mechanistic model predictions and offered much more accurate predictions. On the other hand, during spiked chlorine demand events when the system was going through a disturbance, the mechanistic model clearly deviated from the recorded C_0 values, but the ML models still provided accurate predictions with a significant improvement (Figure 6). This close alignment between the ML-predicted and recorded C_0 values indicated a more robust control system during both stable and spiked chlorine demand conditions, and therefore, would've led to more savings in sodium hypochlorite consumption.

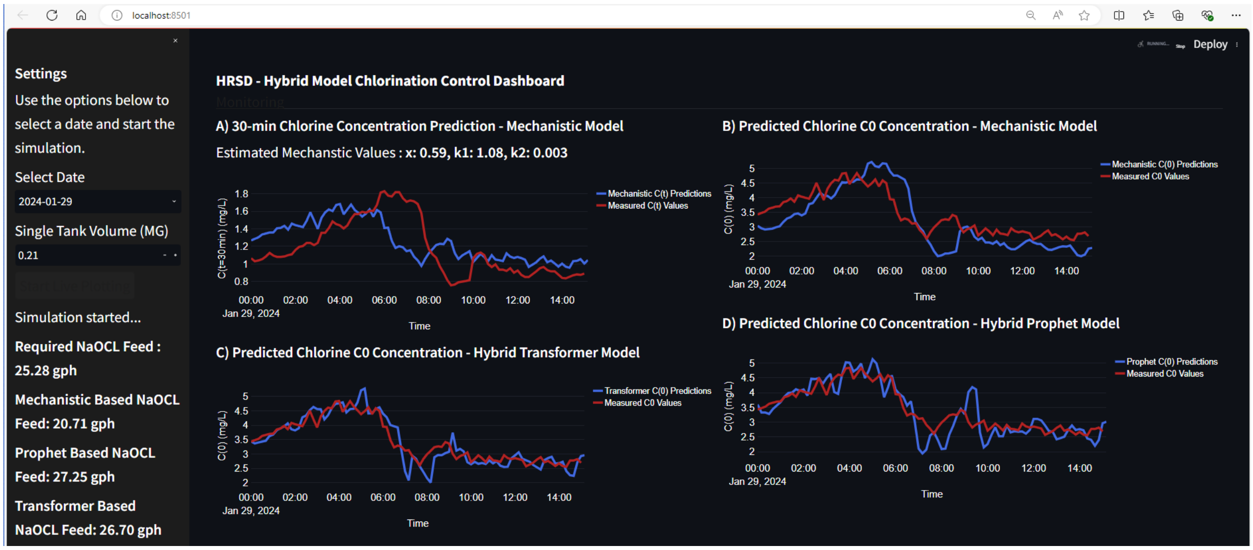


Figure 6. Online dashboard demonstrating the real-time performance of the models under normal and spiked chlorine demand conditions.

The observed improvement in the prediction performance was consistent with the RMSE and MAE values obtained from the different models as discussed in the previous sections. To fully capture sudden variation dynamics, ML models and especially Transformers need to have enough training data to observe such peak events on multiple occasions so they can better predict these events on unseen data. This observation indicates that such predictive models become more and more accurate as their deployment process goes on where they have access to more training data. Therefore, it is possible to further increase the performance of the control system during disturbance events by i) having access to more training data that represent irregular spikes in chlorine demand and ii) developing a separate disturbance detection model to predict the irregular spikes in advance and select the best control strategies accordingly. The integrated dashboard demonstrates how advance AI/ML models can correct mechanistic model predictions and provide access to a sophisticated control system with a meaningful potential for significant savings in chemical consumption and GHG emissions while maintaining satisfactory effluent quality.

4. Conclusions and Next Steps

This case study demonstrated a novel application of advanced AI/ML models to support and complement the prediction accuracy of mechanistic models that have been widely used in the industry. Using AI/ML models as a prediction correction tool for the chlorination mechanistic model can lead to a more efficient and robust chlorine dosing operation at the HRSD Nansemond plant. This strategy helps the plant reduce its sodium hypochlorite consumption and increase its resiliency in meeting effluent chlorine residual permit requirements. Some key takeaway points are summarized below:

- What are the main benefits to utilities? Given the complexities and dynamic nature of disinfection processes using chlorination, a feedback control system may fail to provide accurate adjustments in a timely manner. The main benefit of using AI/ML models as a predictive monitoring tool for chlorination processes is to use them as a feedforward component in the control system to account for unexpected variations in chlorine demand. A feedback-feedforward control strategy optimizes resource use while ensuring regulatory compliance with chlorine residual levels in the effluent.
- Risk-based control and probabilistic modeling: The HRSD Nansemond plant aims to ideally maintain a setpoint of 0.5 mg/L chlorine residual in its effluent. Chlorine residuals below 0.5 mg/L are regarded as exceptions, and levels below 0.1 mg/L are regarded as violations. Therefore, there is a tradeoff between minimizing sodium hypochlorite usage and maintaining adequate effluent chlorine residual levels. In a risk-based approach with probabilistic modeling, model predictions would be presented within a confidence interval with lower and higher probability bands. In this scenario, at the beginning of the month when higher exceptions are allowed, operators would adjust the

dosing rate following the lower band of the model prediction's confidence interval; while towards the end of the month with lower exceptions allowed, they would take a more conservative approach and move on the higher end of the model prediction's confidence interval. Implementing this approach is one of the additional planned steps in the deployment process.

- Identification of irregular spikes in chlorine demand: Results obtained from this project indicated that the prediction of accurate dosing requirements would become more challenging when there was a sudden spike in chlorine demand. Therefore, predicting when the plant is going through a sudden chlorine demand spike is crucial to successfully deploy this hybrid control approach. To achieve this capability, a classification model would be trained using water quality parameters collected from the plant's SCADA system to act as a "disturbance" model to offer protections against sudden spikes in chlorine demand and identify regions where model predictions could be negatively impacted.

The final step to fully implement this approach is under development and focuses on the integration of all these individual components in a monitoring and control framework that is both easy to use and meets the data security considerations at the plant. The overarching goal is to create a user-friendly interface for plant operators to efficiently optimize the chlorination process at the HRSD Nansemond Treatment Plant.

5. References

1. Pignata, C. *et al.* Chlorination in a wastewater treatment plant: acute toxicity effects of the effluent and of the recipient water body. *Environ. Monit. Assess.* **184**, 2091–2103 (2012).
2. Spellman, F. R. *Choosing Disinfection Alternatives for Water/Wastewater Treatment Plants*. (CRC Press, 1999).
3. Lee, W. & Westerhoff, P. Formation of organic chloramines during water disinfection – chlorination versus chloramination. *Water Res.* **43**, 2233–2239 (2009).
4. Haas, C. N. & Karra, S. B. Kinetics of Wastewater Chlorine Demand Exertion. *J. Water Pollut. Control Fed.* **56**, 170–173 (1984).
5. Gao, F. & Han, L. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Comput. Optim. Appl.* **51**, 259–277 (2012).
6. Prophet. *Prophet* http://facebook.github.io/prophet/.

7. Quick Start Guide - NeuralProphet 1.0.0rc8 documentation. https://neuralprophet.com/quickstart.html.
8. Merritt, R. What Is a Transformer Model? *NVIDIA Blog* https://blogs.nvidia.com/blog/what-is-a-transformer-model/ (2022).
9. Bengesi, S. *et al.* Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. Preprint at https://doi.org/10.48550/arXiv.2311.10242 (2023).
10. Tripathy, J. K., Chakkaravarthy, S. S., Satapathy, S. C., Sahoo, M. & Vaidehi, V. ALBERT-based fine-tuning model for cyberbullying analysis. *Multimed. Syst.* **28**, 1941–1949 (2022).
11. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).
12. Liu, X. *et al.* GPT Understands, Too. Preprint at https://doi.org/10.48550/arXiv.2103.10385 (2023).
13. Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).