How to Solve Operational Challenges Using Data-Driven Modeling and Optimization

Kendra Walter[1], Sara Troutman[2]

[1]Xylem, South Bend, IN

[2]Xylem, Ann Arbor, MI

[1]Email: kendra.walter@xylem.com

[2]Email: sara.troutman@xylem.com

ABSTRACT
Real-time models, or digital twins, ingest operational data in real-time to simulate plant performance under actual system conditions. When coupled with optimization algorithms, a real-time model can be a powerful tool for process control. This paper outlines how real-time data-driven models and optimization algorithms can be applied to optimize process control at water resource recovery facilities (WRRFs). A technical approach and practical considerations for implementation are presented, followed by several case studies highlighting use cases and outcomes at different WRRFs. Both case studies demonstrate the opportunity to reduce energy usage while maintaining regulatory compliance. A discussion of challenges and lessons learned from implementations is also presented.

KEYWORDS
Digital twins, data-driven models, process control optimization, real-time decision support systems, water resource recovery facilities

INTRODUCTION | BACKGROUND

While traditional analytical models seek to replicate the complex processes that occur at water resource recovery facilities (WRRFs) using known biological and physical relationships, data-driven models are an alternative that leverage abundant data resources and learn from statistical relationships that are present in the data. One advantage of data-driven models is that they are well suited to real-time applications. Real-time models, or digital twins, ingest operational data in near real-time to simulate plant performance under current system conditions. When coupled with optimization algorithms, a real-time model can be a powerful tool for process control. Optimization algorithms can run a large number of model scenarios with varying operating setpoints, subsequently offering recommendations for setpoints that will achieve near optimal performance outcomes. While sensor data alone can provide operators with insights into the current state of the plant, this approach goes a step further to utilize data to generate strategic support for solving operational challenges. It allows for control actions to be made based on a quantitative understanding of expected future behavior. This "feed-forward" control approach offers greater stability than traditional feedback control. Another advantage of this approach is the ability to explicitly consider multiple, sometime competing, objectives. While traditional control logic may struggle to balance all of these operational objectives in a dynamic operating environment, these tools can provide near real-time analysis and recommendations that help operators fine tune operations.

In recent years, the water industry has seen increased interest in leveraging big data and artificial intelligence (Bahramian et al., 2023; Corominas et al., 2018; Newhart et al., 2019; Zhao et al., 2020). However, widespread adoption of innovative solutions at treatment facilities is slow. These types of technologies require very specific technical expertise and utilities may find it daunting to even know where to start. This paper aims to present a step-by-step approach to implementing real-time data-driven models for process control at water resource recovery facilities.

METHODOLOGY

Planning for an Implementation

The first step in this process should be the identification treatment process(es) that would benefit from real-time optimization of the process controls. It is important that the utility has a clear understanding of what benefits they are trying to achieve to evaluate whether this approach is a good fit. Typical use cases include balancing of multiple competing objectives — often, maximizing effluent water quality and minimizing operational expenses (e.g., power and chemical consumption). Other common objectives include increasing process stability and effluent quality consistency and supplementing operators' process control knowledge and limited bandwidth.

Once the objective is determined, the next step is evaluating whether the methodology is a good fit for the utility and for the use case. Digital twins offer many advantages over traditional process control strategies. They allow the control logic to consider expected future behavior early on in a feed-forward control fashion instead of waiting until the process reaction is actually measurable (i.e., traditional feedback control). However, they do require a coordinated effort to deploy and maintain. It is important to consider whether the utility has the right policies, systems, and people in place to make a digital twin project successful.

Several considerations should be made when evaluating a data-driven modeling approach. Data-driven models offer numerous advantages including extracting and learning from statistical relationships in the data, non-parametric structures that lend flexibility to model form, robustness to noisy and incomplete data, and fast simulations once trained (Abdallah et al., 2020; Bahramian et al., 2023; Cheng et al., 2020; Dürrenmatt & Gujer, 2012; Jawad et al., 2021; Zhao et al., 2020). Furthermore, these characteristics allow for data-driven models to be leveraged in a number of contexts, including soft sensors, data fault detection, variable prediction, and near real-time optimization and control (Corominas et al., 2018; Newhart et al., 2019). However, this approach will not be successful without sufficient data availability and data quality. The exact datasets required will depend on the process being optimized, and a technical expert can provide recommendations on the exact parameters, frequency, and volume of data needed for a given application. Generally speaking, water quality measurements both upstream and downstream of the selected treatment process are needed to train a model on the performance of the process. In certain cases, laboratory data may be sufficient for some of the parameters, as a soft sensor or virtual sensor can be developed. Consideration should also be given to whether the modeled process is anticipated to be subject to changes. Data-driven models are most reliable when replicating conditions they have seen previously. Significant changes to either influent conditions or the treatment process itself could necessitate a period of additional data collection and retraining of the model(s).

Another key step in the planning process is to identify utility staff to be involved with this project and a potential external implementation partner. On the utility side, it is important to generate buy-in from several parties early on. These may include:

- **Utility Leadership:** Commitment at a utility leadership level to implementing digital solutions will improve the project's chances of success. This will also help with development of a cross-functional team and with securing financial support for the project.
- **Engineering and/or Plant Management Staff:** A project manager should be identified to support the implementation.
- **Operations Staff:** Operators will ultimately be the ones who dictate whether and how the solution is used. By including them in the process early on, the solution can be designed to incorporate their knowledge of the system and operational pain points. This will go a long way in generating their buy-in. Their involvement will also be needed to facilitate data collection efforts and to support the model development process with their knowledge of how the system operates.
- **IT/OT Personnel:** Their support will be required for architecting a reliable system that meets the organization's security policies.

For most utilities, the process of developing a digital twin requires more expertise and capacity than is available in-house. Utilities can partner with technical experts with experience in development of data-driven models and optimization algorithms. When evaluating potential partners, consideration should also be given to how the utility wants the solution to be architected. A digital twin can be deployed as a custom-built package that is run on local servers. Another approach is to incorporate it into a proprietary software environment (cloud based or on-premises) that has been built for supporting these kinds of solutions. Use of these software platforms may offer benefits such as quicker deployment, more reliable support and maintenance, and improved ability to visualize recommendations and track adherence and key performance indicators (KPIs). Additionally, it can facilitate future deployments across other facilities that the utility operates.

Once a goal and approach are identified and a team is built, the analysis can begin. A good place to start is with a "discovery" analysis to evaluate existing datasets and identify data gaps as well as optimization potential.

Data Collection and Quality Control

To begin the development process of a data-driven model, data from the WRRF must be collected and reviewed. The locations of online sensors and samples for laboratory measurements should be documented; these should include locations across the WRRF, not just within the treatment process that is the focus since the behavior of other processes provide context for the focus area.

Through collaboration with utility data experts, the online and laboratory data across the identified locations should be compiled and reviewed for data availability and quality issues. In particular, focus should be given to:

- Periods in which data is missing and/or drops to a null value (this could be "false" zeros or negative values, depending on sensor convention),
- Sensor noise that substantially obscures the underlying behavior in the data signal,
- Periods in which data form a flatline, which could be an indication of an offline sensor, sensor maintenance issues, or real behavior from maintenance activities,
- Sudden changes in data magnitudes or behavior, which could be an indication of sensor calibration or maintenance activities

Questions about anomalous data can be reviewed as a group and can be used to:

- Exclude time periods due to lack of data availability or quality
- Identify maintenance activities in which system or process behavior deviates from typical operation
- Determine if additional data is needed from other process locations or other time periods

Throughout the data review process, it is important to remember that data do not have to be pristine for use in data-driven model development. Occasional outliers can often be removed through offline or real-time data pre-processing and minor sensor noise, where underlying system dynamics are still evident, can ensure that model training does not overfit to minute data features.

Once data is compiled and reviewed, a period of time must be selected for which all relevant data for the model are available and of reasonable quality for model development. Typically at least six months of data is required to develop a data-driven model for WRRF applications, while at least one year of data is ideal to fully capture seasonal behavior, a variety of wet-weather events, etc. Furthermore, the selection of input features, discussed further below, can change through data preparation for the model. For instance, a particular sensor may have initially been deemed irrelevant to the process that is being modeled, but after further investigation and model training, model performance improves by including the data because it contains some information about system behavior. Thus, it may be necessary to adjust the time period of data based on the need to include new data sources. This is part of the iterative process of developing a data-driven model.

The next step is to pre-process the data from the selected time period to quality control what will be used in model development. These pre-process steps should be documented and may include:

- Resampling to have a uniform sampling frequency across all data
- Removing flatlines in the data
- Removing outliers using typical ranges of each data signal

As mentioned above, not all anomalous data need to be removed before continuing. Imperfect data is to be expected from collection in a WRRF and, while some anomalous and erroneous data can be removed offline or in real-time, in general, some imperfect data during model training can result in a more robust model for future use.

Data Preparation

Since all relationships in a data-driven model are derived from the input and output data provided during model training, the careful selection and preparation of data *before* model training begins is crucial. Just as model parameters, such as neural network hyperparameters, can be tuned to adjust model performance throughout the model training process, so can decisions around which and how much data to use in model training. The below sections summarize several key data selection and preparation decisions.

Data Preparation: Select Input Features, Output Target Variables, and Control Variables

The first step in preparing data for development of a data-driven model is to select input features, output target variables, and control variables. For any timestep t_i, the data-driven model will predict output target variable values in a forecast time window $(t_i, t_i + FW]$, where FW is the length of the forecast time window. These predictions will be made using the input features and control variables as model inputs from a past time window that includes the current simulation time $[t_i - PW, t_i]$, where PW is the length of the past time window. Output target variables are measurable or computable outputs of the treatment process that the model will predict, such as effluent ammonia concentration from the secondary aeration system. The input features and control variables are data sources that will inform the model predictions of the output target variables, such as influent water temperature, secondary influent ammonia concentration, and the air flow rate into the secondary aeration system. Control variables are specifically inputs into the treatment process and model that will be controlled to augment the system outputs, such as air flow rate into the secondary aeration system to control the secondary effluent ammonia concentration. These control variables will be iterated over in future control optimization to determine the best control values to produce output target variables that meet system objectives. The input features are all other system inputs. It is important to include output target variables for the past time window as input features — put simply, the current value of a target variable will impact the future values of the target variable.

In general, model accuracy can be improved by including as many input features that directly impact or are related to the output target variables as available. This ensures that any relationships between model inputs and outputs are available for the model to learn. For instance, if developing a data-driven model of a secondary aeration system to predict secondary effluent ammonia concentration, it may be beneficial to include chemical oxygen demand and phosphorous species data influent and/or effluent from the secondary treatment process as the transformation of these components influence available dissolved oxygen which, in turn, impacts ammonia oxidation. However, if there is a calculated data source that are directly derived from a measured data signal, including both will not likely improve model accuracy since the calculated data source will not be providing any *new information*.

While including any input features can be bolster model performance, each model should have a limited number of target variables. With each additional output target variable included in a single model, the model complexity and amount of data needed to adequately train the model increase, requiring longer time for model training and increasing the difficulty to provide high model accuracy. If more output target variables are needed, it may be appropriate to develop and train separate models.

Data Preparation: Select Past and Forecast Time Windows for Model Inputs and Outputs

Since the data-driven model will be informed by past timeseries of input feature and control variable data, it is important to select a past time window that encompasses the travel time between the hydraulically first input feature location and the location of the output target variables. The aggregated hydraulic retention time (HRT) of the treatment process(es) between these two locations represents this past time window. For instance, the WRRF influent temperature at time t will impact the secondary effluent ammonia concentration at future time $t + HRT$, where HRT is the hydraulic retention time between the WRRF influent measurement location and the secondary effluent ammonia measurement location.

The aim of the data-driven model is to predict output target variables in the future across a forecast time window. The forecast time window size can greatly impact model performance; since making predictions are inherently uncertain, making predictions far out into the future compounds the uncertainty and makes it more difficult to predict accurately. As such, the forecast time window should balance between the ability to act on the forecasts via control decisions and the ability to train an appropriately accurate model to produce such forecasts. For instance, if the HRT between a control variable location and the target variable location is HRT hours, then we would want to be able to forecast the target variable HRT hours in the future. However, we may not be able to train a model that can accurately predict this far into the future and so may need to reduce this forecast time window. The length of this forecast time window can be iterated on during model training.

Data Preparation: Divide Data into Training, Validation, and Testing Datasets

To ensure model performance can be verified with an independent dataset and to prevent overfitting of the model to data, the total dataset needs to be divided into three sets: training, validation, and testing. The training and validation datasets are for training the model (with the training dataset), evaluating performance (on the validation dataset), adjusting hyperparameters based on evaluation, and then re-training (with the training dataset), and so forth until satisfactory performance is achieved for the validation dataset. Once this is complete, the model is now no longer independent from the training and validation datasets. Thus, for final evaluation, the model is trained with the combined training and validation and then evaluated with the testing dataset (Fig 1).

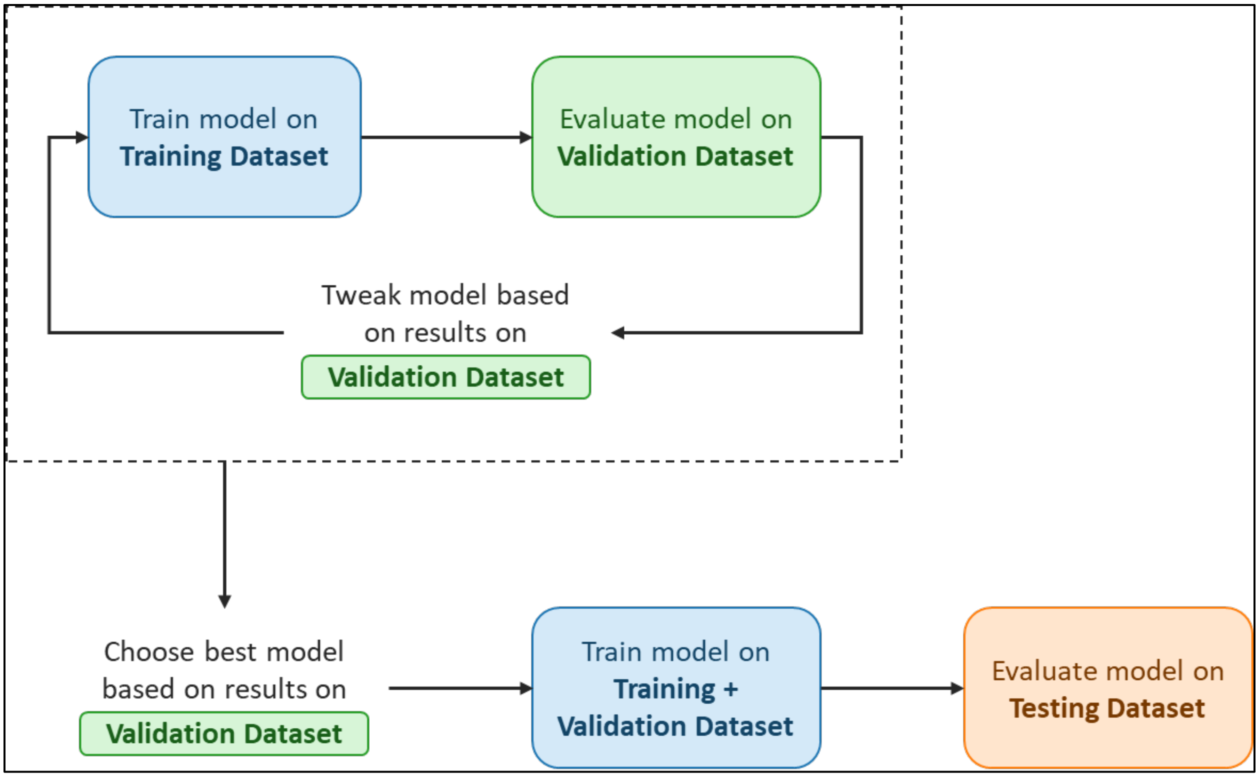

Fig 1: Schematic of model training and verification with datasets

The proportion of the total data in each dataset needs to ensure that sufficient data is provided for model training while reserving data for validation and testing. These requirements also underscore the need for a long time period of data; else, there will be insufficient data to undergo training and verification in this process. The training, validation, and testing datasets should also be distributed across different operating conditions, seasonal variation, etc. For instance, it is important that both model training *and* verification include samples of snow melt during the spring season, various wet-weather events, and dry periods throughout a calendar year, rather than being trained and validated during a wet spring season and then performing poorly when testing during a dry season.

Model Development

There are myriad data-driven model types and architecture options that can be considered and tested to build a representation of the wastewater treatment process. The different model types each have advantages, disadvantages, and nuances that can be weighed before a particular option is chosen for a particular application. This paper is not intended to provide a thorough list of those options, but rather will summarize some recent applications of data-driven models for wastewater treatment process modeling in literature and present some key considerations to support successful model development.

Model Development: Key Considerations

Many wastewater treatment processes are continuous in nature, and, at any given time, the behavior of the system and effluent quality depends not only on the system states at that time, but also previous system inputs and operations. Thus, wastewater processes are often best modeled with timeseries-based modeling, with inputs as timeseries and models that can retain some memory of the system's past behavior. This requirement doesn't not dictate a particular model type be used but does indicate an important consideration for selecting model structure and characteristics.

Another component of a data-driven model that should be considered before model selection is the potential benefit of including constraints in the model structure. These constraints can be defined to ensure that relationships that the neural network model learns align with an understanding of underlying system behavior. One example that demonstrates the importance of constraints is for a controlled secondary aeration system that is oxidizing ammonia. It is well understood that aerobic microorganisms consume oxygen to oxide ammonia to produce nitrate and nitrite. Thus, we would expect and want to ensure that a data-driven model of this process associates increasing dissolved oxygen concentration or air supply with effluent ammonia concentrations that do not increase. The air supply and/or dissolved oxygen in the real system may be actively controlled in response to the measured effluent ammonia concentration. In this case, the system will likely increase the air supply if the measured effluent ammonia concentration increases. This will likely manifest in the measured dataset to suggest the opposite relationship, that increasing ammonia is associated with increasing air supply. Thus, by constructing appropriate constraints, modeled relationships can be congruent to physical processes.

Construction of a suitable loss function is another crucial component in the development of a data-driven model. The loss function quantifies the performance of model predictions and thus is heavily used in model training. Model training aims to minimize the loss function, bringing the model predictions close to "fitting" the actual measurements by adjusting the model parameters. Choice of a loss function should be based on characteristics in the data that the predictions should capture well and can include a combination (sum) of multiple calculations. Common forms of loss functions include mean absolute error (MAE) and mean squared error (MSE), but can also focus on specific data features. For instance, peaks in WRRF effluent concentration are what can result in violation of regulatory limits. Thus, it might be beneficial for the loss function to include a measure of closeness between measurements and predictions at signal peak values. Further, certain output target variables may be more important in a regulatory sense and thus the prediction accuracy for those outputs may be weighted higher than others.

Model Development: Model Types and Architectures

With the two key considerations described above, focus can be applied to model architecture. Artificial neural networks (ANNs) are a popular choice in many data-driven modeling applications, WRRFs included, due to their structural flexibility. ANNs have been developed for use as soft sensors to predict WRRF influent quality components that are difficult, time-consuming, or expensive to measure, including BOD5 (Alsulaili & Refaie, 2021; Cheng et al., 2020), and COD and NH4 (Xu, et al., 2023). ANNs have also been developed extensively to predict WRRF effluent quality in contexts of soft sensors to provide more reliable real-time estimations of effluent measurements and/or to forecast effluent quality in the future to aid operations (Alsulaili & Refaie, 2021; Alvi et al., 2022; Cheng et al., 2020; Farhi et al., 2021; Guo et al., 2020; Jawad et al., 2021; Liu et al., 2021; Matheri et al., 2021; Xu et al., 2023; Zhang et al., 2022; Zhu et al., 2022).

In general, ANNs are an umbrella under which many architectures can be tailored. Given the importance of time-based behavior, one of the key considerations described above, the ability of a model to incorporate system "memory" is of particular importance for WRRF contexts. As a result, ANNs with long short-term memory (LSTM) architectures are frequently used. LSTMs are an expansion of recurrent neural networks (RNNs) that can capture time dependencies in the data by selectively retaining and discarding information through the use of gates: input, forget, and output gates (Cheng et al., 2020; Liu et al., 2021). A gated recurrent unit (GRU) architecture is another option that is similar to LSTMs. GRUs, however, have simpler gates, often described as update and reset gates. This results in GRUs having fewer parameters and being quicker to train compared to LSTMs. However, LSTMs may capture long-term dependencies better than GRUs (Cheng et al., 2020; Alvi et al., 2022).

Other types of data-driven models have also been demonstrated to model wastewater treatment processes, sometimes in conjunction with ANNs, including support vector machines (SVMs), Gaussian processes (GPs), regression, etc (Bahramian et al., 2023; Krivec et al., 2021; Newhart et al., 2019). Development of ANN models, however, has been predominant across literature (Bahramian et al., 2023).

Iterate

The process of data preparation, model training, and validation is iterative — regardless of experience with data-driven models, the best performing model is not to be expected within just one iteration. Once the performance of the trained model (with the training dataset) is evaluated (with the validation dataset), adjustments can be made to the data selection and preparation, and model development in an attempt to improve model predictive performance. These adjustments can include input features, past and forecast time windows, hyperparameter values, and even model architecture. Once an adjustment has been made, the model can be re-trained (with the training dataset) and re-evaluated (with the validation dataset), comparing model performance with previous iterations. It is best practice to make one change at a time to better understand how each change is impacting predictive performance to drive improvement in model accuracy. Thoroughly documenting this iterative process, including the data and model decisions along with model performance, will build understanding of the modeling process.

Set Up the Optimization System

Once a data-driven model is developed, it can then be used to build an optimization system. The goal of an optimization system is to run the data-driven model many times with various combinations of operational settings and the current system state (as represented by data) to determine the settings that provide the best system performance. An optimizer will adjust operational settings after each iteration of model simulations to determine the next combination of settings to evaluate. The use of a fast-simulating data-driven model in this application allows for this iterative optimization process in near real-time. Constraints on control settings (e.g., pump start times) can also be incorporated into the optimization process to prevent adverse operational conditions. The choice of optimizer will influence the search for optimal operational settings, including the convergence towards an optimal and the exploration of settings to prevent the optimization system from getting stuck in a local optimum. Examples of optimizers include genetic algorithms, particle swarm methods, and simple brute force in which the ranges of operational settings are discretized and every possible combination is tested.

As mentioned above, the optimization system should search for settings that provide the best system performance. What constitutes the best system performance is discussed and decided upon by stakeholders, and are often comprised of multiple objectives. Objectives should be measurable or calculable. Common objectives include:

- Maintain effluent quality within regulatory or internal limits
- Reduce the magnitude or occurrence of sudden changes in effluent quality
- Reduce air flow supplied to a secondary aeration system
- Reduce power supplied to a treatment process
- Reduce chemical supplied to a treatment process

In selecting multiple objectives that compete, objectives should be prioritized and balanced, similar to evaluation of a Pareto front.

Develop a Fallback Strategy

In the event of interruption to the optimization system, a fallback strategy to provide operational settings and guidance to operators is crucial. Interruptions can come in the form of absent critical data inputs into the data-driven model due to connection issues, unprecedented system behavior from equipment failure, etc. The fallback strategy should provide operators with recommendations for operational settings that allow continual WRRF operation and compliant effluent quality despite the interruptions.

Design of the fallback strategy is done in coordination with WRRF operators, managers, and engineers. Rather than being an all (full optimization system) or nothing (fallback strategy) approach, a cascade of fallback strategies can be developed that correspond to the degree of system interruption. For instance, if data are not available for a minorly important data signal, a first-tier fallback strategy can be designed to respond to the remaining online sensor data. This would represent a small step down from the near optimal operation considering the data available. A bare-bones fallback strategy could resemble typical operating protocols *before* the implementation of the optimization system and decision support system. This strategy is often simple, relies on few online measurements, and emphasizes treatment efficacy to stay safely below regulatory effluent limits. Each tier in a cascade of fallback strategies will not be as efficient or precise as the full optimization system but will scale efficiency of system performance based on data resources available.

Architect the Real-Time Decision Support System

When building a real-time decision support (RTDS) system on top of the optimization system, there are several key decisions that need to be made across the system stakeholders. One of the first decisions is if the RTDS will provide recommendations of operational settings to WRRF operators or automated control actions directly in the WRRF. This decision will be based on support of WRRF operators and management, feasibility of automated or manual controls, etc. Even for systems that will be receive automatic controls, implementations are generally started with operator recommendations to 1) validate optimization system performance and 2) build operator trust of the system. Another key component of an RTDS is the information that will be presented to operators. This could include real-time measurements from across the WRRF, predicted effluent quality (i.e., outputs of the trained data-driven model), outputs of the optimization system (i.e., recommendations or automated control actions), and alarms of disruption to normal system operation.

The hardware and location of the optimization is another aspect of the RTDS that requires special consideration. Regardless of location, the optimization system will receive online sensor measurements (e.g., via OPC) and laboratory measurements (e.g., via LIMS), run the data-driven model, run the optimization system, and return operational settings to the WRRF to display via dashboard and/or enact via SCADA. The decision of the computer(s) location that executes these actions will need to balance the feasibility of system maintenance (e.g., periodic model retraining, troubleshooting), costs, and security policies of the WRRF. Some examples of hardware arrangements include:

- Cloud-based server
- On-premise desktop
- Hybrid between cloud and on-premise with myriad combinations (e.g., on-premise desktop to receive measurements, run model and optimization system, and then push operational recommendations to a cloud-based visualization platform)

Deployment, Monitoring, and Maintenance

Following deployment, conducting an initial observation period (in which the system is still in "recommendation only" mode) allows for verification that the solution is performing as expected. Adjustments can be made during this time to improve how the solution operates. During this time, operators can also begin to develop confidence in the solution. Following this review period, utilities may move forward with automating implementation of the recommendations. A strength of this type of control strategy comes from the ability to dynamically react to current system conditions. Automation allows for setpoints to be adjusted in near real-time, whereas operations staff would typically not be able to make changes manually at the same frequency.

Ongoing performance monitoring is critical to ensuring that the system continues to perform. Model performance should be evaluated annually at a minimum. Over time, retraining of models may be required due to changes in influent conditions or the treatment process itself. Additionally, monitoring and maintenance may be needed to maintain system connectivity and security.

RESULTS

Case Study #1

The first case study focuses on an application for Gold Bar Wastewater Treatment Plant (WWTP), which is owned and operated by EPCOR in Edmonton, Alberta, Canada. Xylem was engaged by EPCOR to conduct a preliminary analysis evaluating the opportunity to optimize energy usage using Xylem's digital twin solution — Plant Real Time Decision Support (P-RTDS). The Gold Bar WWTP is designed for an average flow of approximately 283 MLD (75 mgd) and a maximum flow of approximately 1,200 MLD (317 mgd). The main treatment processes are shown in the below schematic (Fig 2).

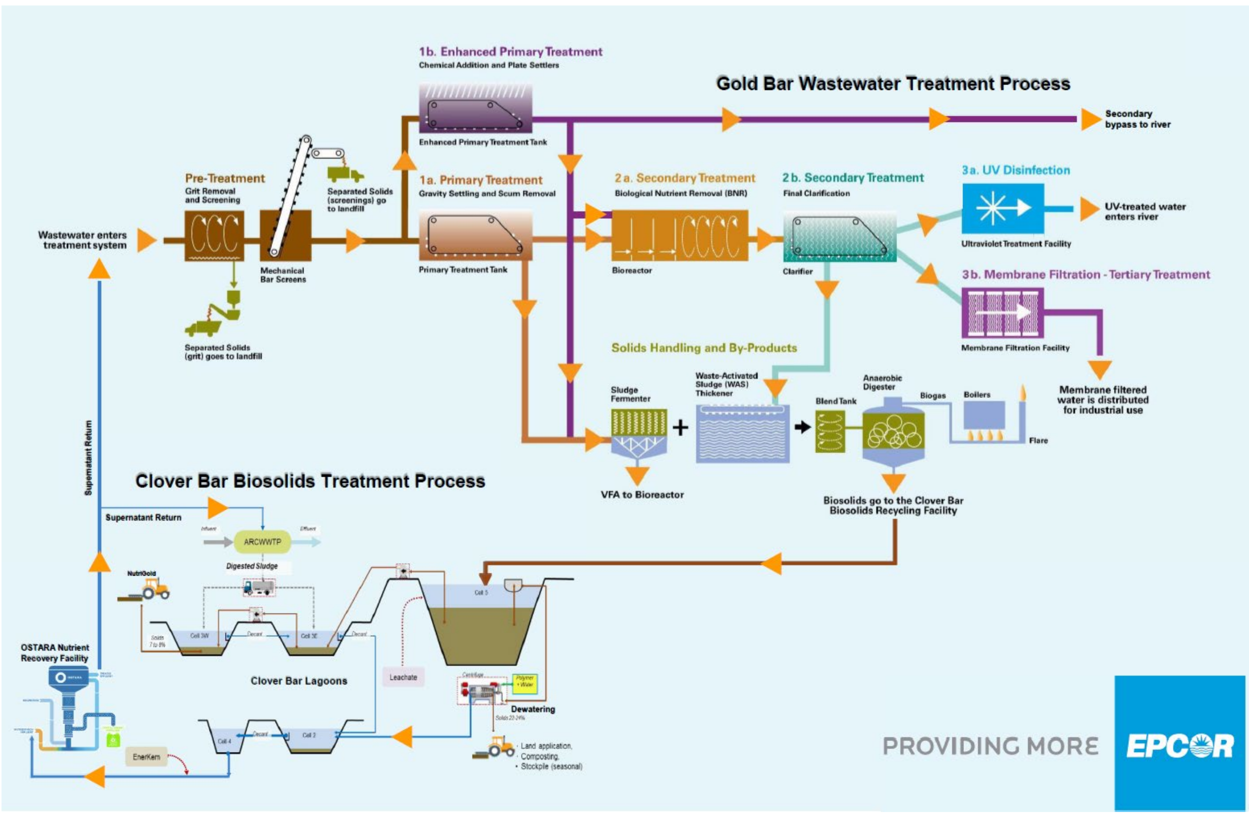

Fig 2: Schematic of Gold Bar WWTP

Gold Bar's operating permit sets effluent limits for BOD, TSS, *E. coli*, total phosphorous, and ammonia. Additionally, EPCOR sets internal metrics for effluent quality that are even more stringent than those set by the permit. The secondary treatment process provides biological nutrient removal for ammonia and phosphorous. It includes 11 parallel trains with 8 cells in each train providing different levels of oxygenation. A focus for this analysis was on predicting nitrogen species concentrations in the secondary effluent, especially NH4.

The treatment processes throughout the Gold Bar WWTP are heavily monitored via online sensors and samples taken for laboratory measurements to verify treatment performance. The ample data across the WWTP and the extensive time period of data collection made this Gold Bar WWTP a prime candidate for data-driven model development.

Ten months of sensor data for flow and water quality, spanning November 2022 to September 2023, was used to develop a data-driven model of the secondary aeration process for one of the 11 secondary treatment trains. The one secondary treatment train was selected based on consistent data availability and quality. As described in the Methodology section, data preparation and model setup was iterated over for this case study; the description included below represents the best performing model developed based on prediction of the testing data. The input features include:

- WWTP influent: Flow rate, temperature
- Primary effluent: NH4 concentration, chemical oxygen demand (COD) concentration
- Secondary aeration system: Influent flow rate, mixed liquor suspended solids (MLSS) concentration in the cell 8, air flow rates into cells 3–8, recycle pump speed, return activated sludge (RAS) flow rate
- Secondary effluent: NH4 concentration, NO3 concentration, NO2 concentration, PO4 concentration
- Time of day of sensor measurements

Of these input features, the control variables are the air flow rates into cells 3–8. The output target variables are the secondary aeration system NH4, NO3, and NO2 concentrations (Fig 3).

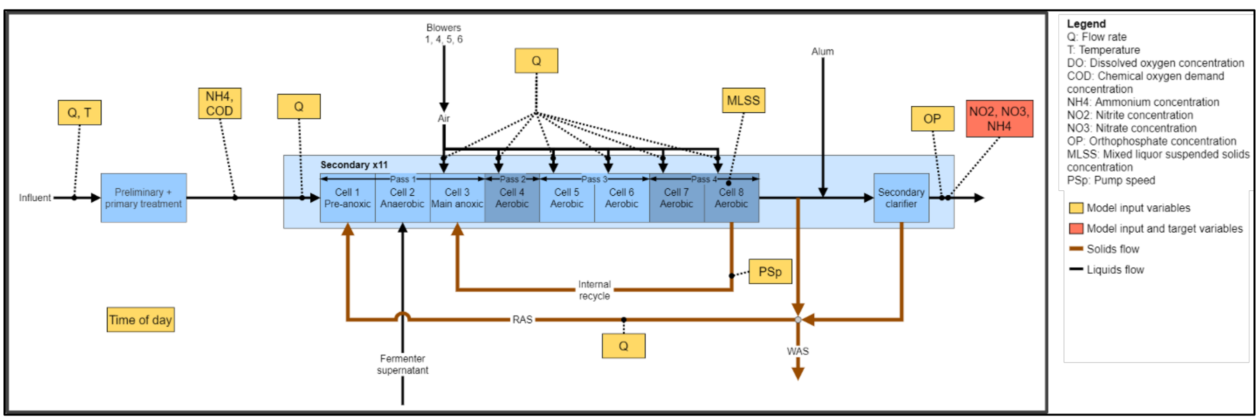

Fig 3: Schematic of Gold Bar WWTP and data inputs and outputs for the data-driven model

The median hydraulic retention time (HRT) of the secondary treatment process, including bioreactors and secondary clarifiers, is approximately 20 hours. With this consideration, a past time window of 36 hours at a 1-hr resolution was used for all input features. Balancing actionability of model predictions and model accuracy, a forecast time window of 6 hours at a 2-hr resolution was determined to perform best — thus the model is predicting NH4, NO3, and NO2 concentrations in the secondary effluent at 2, 4, and 6 hours into the future from the current time (Fig 4). The ten months of data for model development was split between training, validation, and testing datasets with proportions of 40%, 30%, and 30%, respectively. The total dataset was divided into lengths of 7 days and each length was randomly assigned to each dataset based on these proportions.

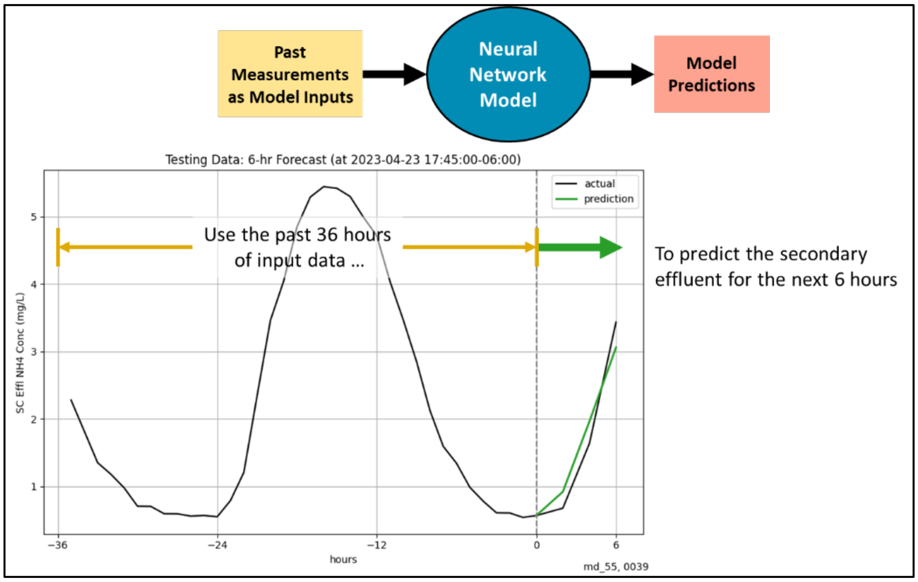

Fig 4: Schematic of data-driven model architecture highlighting past and forecast time windows

A neural network was used to develop this data-driven model. The model architecture uses a gated recurrent unit (GRU) structure. The layer dimensions and other hyperparameter values were tuned specifically to the data in this case study and so are unique to this application; they are not shared here since they would not directly translate to other applications. Constraints are included that dictate that the sum of air flow into the secondary aeration system must be negatively correlated to the secondary effluent NH4 concentration. This enforces the well understood causality between increasing dissolved oxygen and reducing NH4 concentration in the model. The loss function for model training and verification is the mean squared error (MSE) between the predicted and actual measured values of the output target variables of NH4, NO3, and NO2 concentrations of the secondary effluent. Because prediction of NH4 is the most important of these three outputs for regulatory purposes, the MSE of the secondary effluent NH4 concentration is weighted 50% higher than that of NO3 and NO2.

The final data-driven model for this EPCOR case study, trained on the training and validation datasets and evaluated on the testing dataset, achieves accurate predictions for secondary effluent NH4, NO3, and NO2 concentrations under a wide range of operating conditions. Fig 5 shows the comparison of measured and 6-hr forecasted secondary effluent concentrations for the testing dataset. Dry-weather fluctuations in effluent quality are represented well. Of critical consideration in a regulatory context, the timing and magnitude of many wet-weather concentration peaks are predicted by the trained model. The model does underpredict the NH4 concentration peak during the largest storm in the testing dataset occurring on March 28, 2023 and between testing timesteps 2500 and 2700, which is also the largest event in the entire provided data period, by 2.5 mg/L. However, with additional training data, especially including events that are similar in magnitude, the model will increase prediction performance.

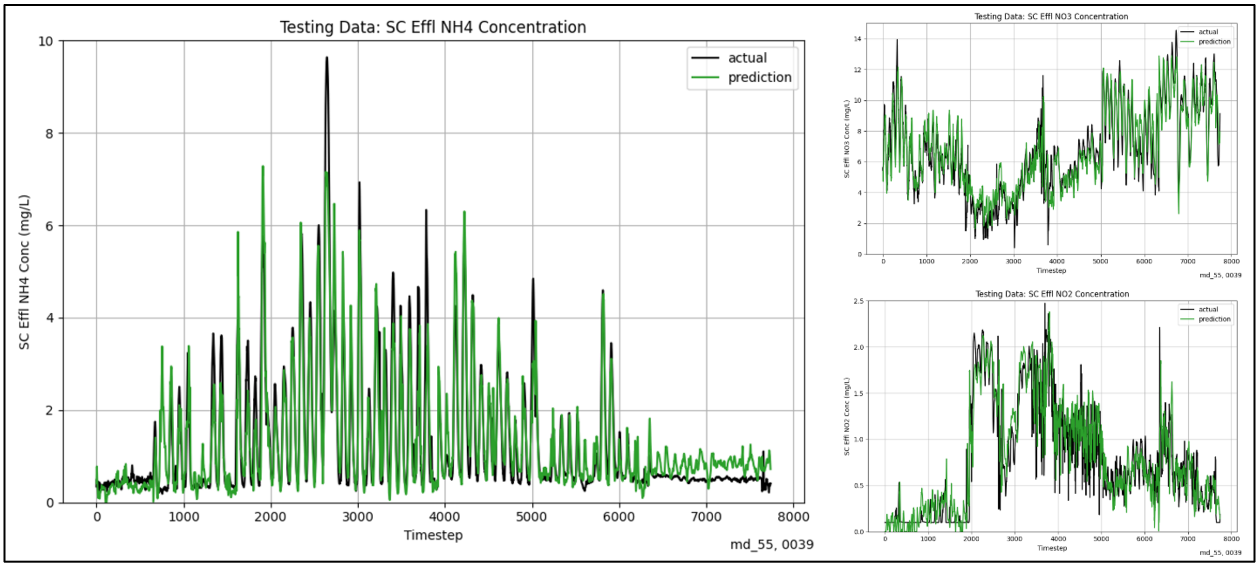

Fig 5: Comparison of measured vs. predicted effluent NH4, NO3, NO2 concentrations for testing data

The overall performance of the model is also quantified in the mean absolute error (MAE) and Pearson correlation coefficient for each point in the forecast horizon: 2, 4, and 6 hours in the future from the current time (Fig 6). As expected, the 2-hour forecasts have lower error values and higher correlation with the measured values, and error increases as the forecast extends to 6 hours. Even for the 6-hour forecasts, the MAE remains low relative to the magnitude of the measurements (<0.4 mg/L for NH4 concentrations) and the predictions are highly correlated to the measured values (>0.9).

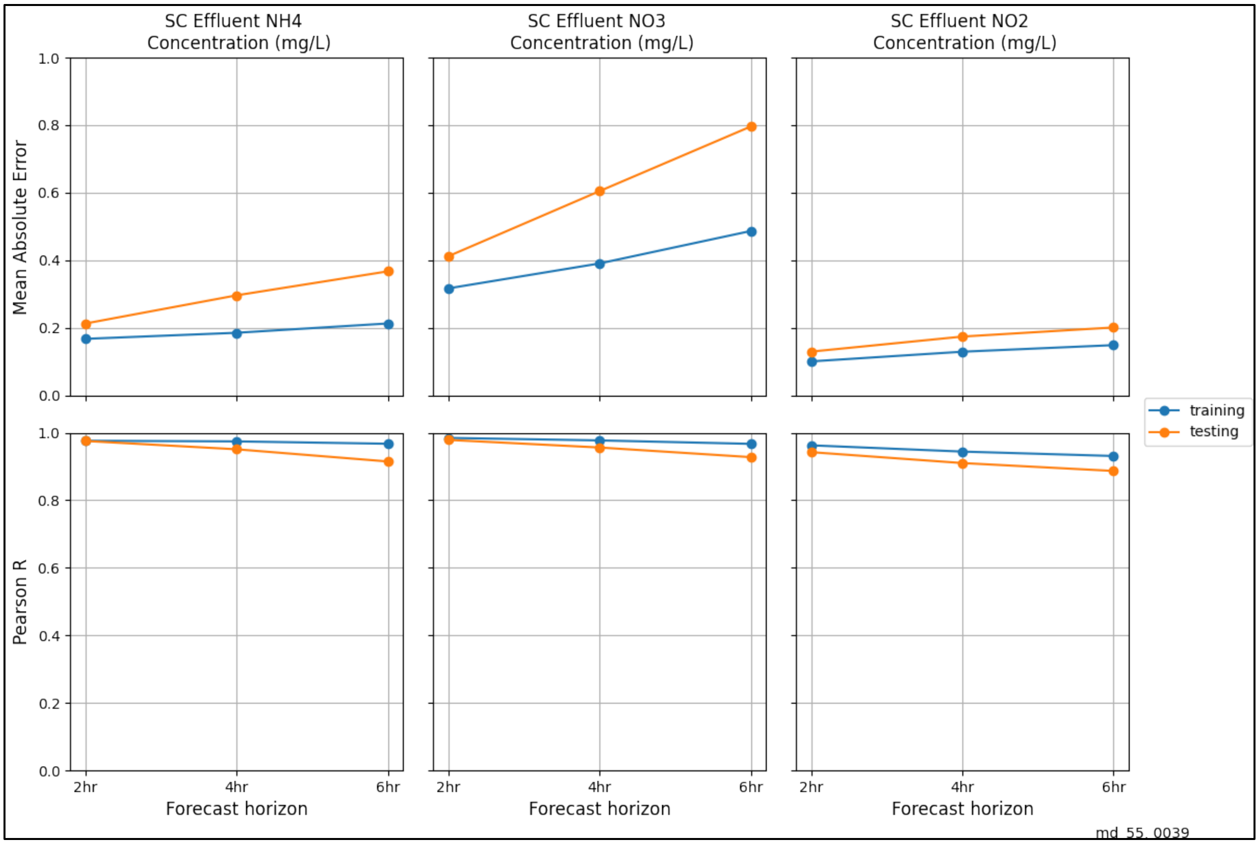

Fig 6: MAE and Pearson correlation coefficient for each forecast horizon timestep

The model was then used to evaluate the potential to reduce energy consumption by reducing air flow while maintaining effluent quality that meets the EPCOR permit limits and internal metrics. To do this, a sensitivity analysis was conducted to study the impacts of changing the secondary aeration air flow rates, which are the control variables for the model, on the secondary effluent NH4 concentration. Uncertainty in model predictions increases as air flow rates reduce significantly below what was typical in the data used for model training. Thus, this sensitivity analysis is not used to evaluate the impact of atypically low air flow rates and rather focuses on air flow rates that are 0–40% reductions of actual air flow rates observed in the measured data. Also, for brevity and clarity, the comparison shown here is made only to the permit limit for NH4, which changes seasonally.

By making reductions in the air flow rate model input, the model predicted that effluent NH4 concentration would increase, aligning with the understood relationship between dissolved oxygen and NH4 oxidation (Fig 7). As expected, peaks in NH4 concentration increase and approach the limits as air flow decreases. However, there are extended time periods where the model predicts that secondary effluent NH4 concentrations would remain substantially below the operating permit limit even with a sizeable decrease in air flow. This demonstrates that opportunity exists to reduce air flow, and thus reduce energy consumption of the secondary aeration system while maintaining compliance.

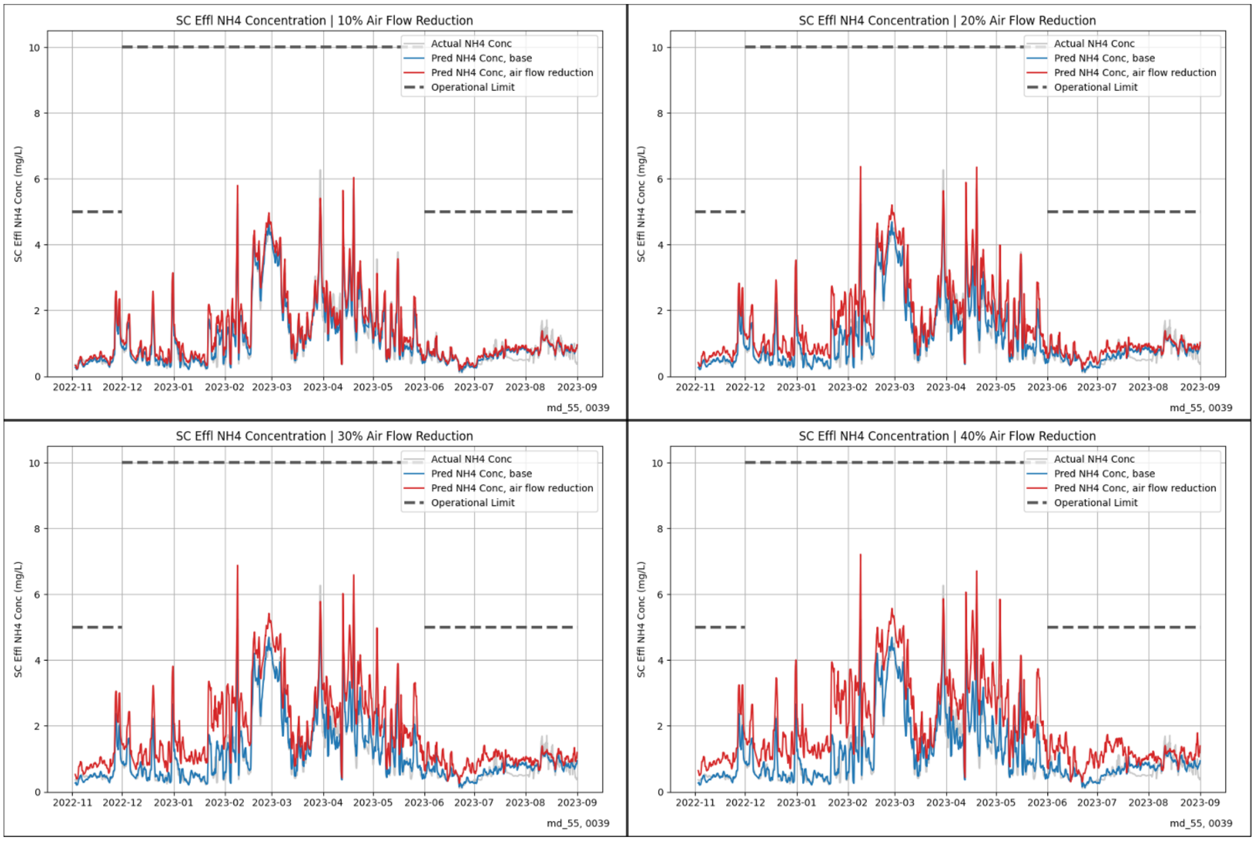

Fig 7: Sensitivity analysis for the impact of secondary air flow on effluent NH4 concentration

The predictive power of this data-driven model will enable operators to know when these requirements of higher air supply vs. flexibility to reduce air flow occur based on real-time predictions of effluent quality. By coupling the real-time model with an optimization engine, different airflow setpoints couple be evaluated in near real-time to determine the setpoints that will provide the best outcomes with respect to both objectives (i.e., energy reduction and effluent quality).

Case Study #2

The second case study is for a WRRF utility in Germany. Faced with significant energy costs associated with aeration, the utility wanted to optimize energy consumption as well as improve safety with better control of chemical usage. This facility did not have online sensors for the inflow available; however, by first building "virtual sensors" to estimate incoming loads, the project team was still able to develop a process model. After several months of operating in a recommendation mode, the solution was ultimately deployed in an automated control mode. Since implementation, a 30% reduction in aeration energy usage has been achieved, corresponding to 1.2 million kWh annually — enough energy to power 321 homes using 3,500 kWh per year.

Using data collected across the WRRF, neural network models were developed to predict removal of NH4, TN, and PO4 in the effluent. Inputs to those models include reactor HRT, computed dynamically from flow measurements; influent flow from industrial and municipal sources; influent carbon to nitrogen ratio (C/N) from measured COD, NH4, and NO3; return sludge ratio; food to microorganism ratio (F:M); and specific airflow relative to influent loading rate.

Once trained, the final models were evaluated on the testing dataset (Fig 8). The standard deviation of prediction error as a percent of the value range for the output target variable was one metric to quantify performance; these values are 6.6%, 4.4%, and 9.6% for the NH4, TN, and PO4 models, respectively. The Pearson correlation coefficient was also used for model evaluation; these values are 0.89, 0.89, and 0.82 for the NH4, TN, and PO4 models, respectively. Both metrics indicate that the neural network models perform well to predict the nitrogen and phosphorous species degradation in the effluent and are suitable for further use to inform the control of the secondary treatment process.

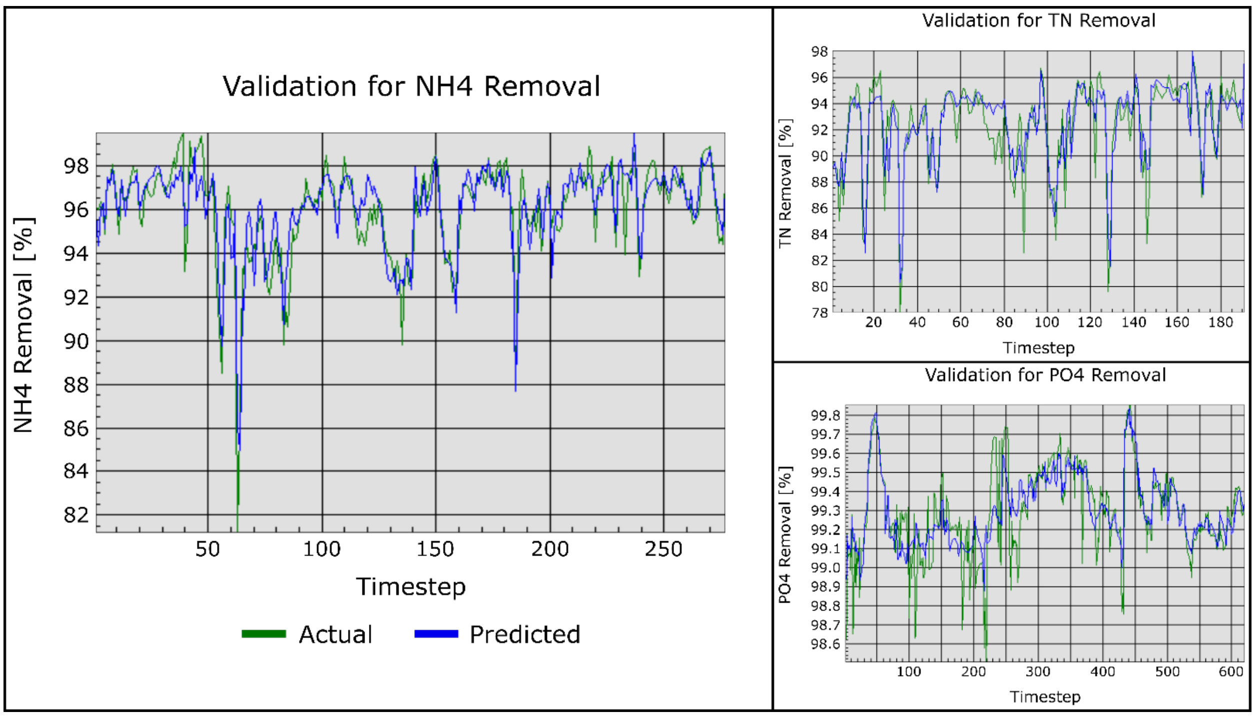

Fig 8: Comparison of measured vs. predicted NH4, TN, PO4 removal for testing data

Leveraging the neural network model, an optimization system is established to provide operational settings that dynamically respond to the predicted effluent quality. The basis of the optimization system is a genetic algorithm that evaluates combinations of operational settings, including air supply and recirculation pumping, using the model to select the best settings that will maximize energy savings while maintaining effluent compliance. In this strategy, the optimization system determines whether NH4 or TN will reach the respective limits first using the effluent model predictions. Based on this driver, the system will either focus on increasing aeration to reduce the effluent NH4 concentration or will adjust aeration settings to provide more denitrification volume to reduce the nitrate concentration, thereby driving down the effluent TN concentration.

Further flexibility was added to the optimization system via two operating modes: dry-weather (aka economy mode, EM) and wet-weather (aka load-driven mode, LM). EM occurs when load and flow rates are normal and the goal of the plant and optimization system is to maximize energy reduction. LM, on the other hand, is characterized by increased influent loads and effluent concentrations, where the goal is to maximize reduction of effluent pollutants to ensure compliance. Based on the current and predicted state of the system, the optimization system will prioritize the strategy dictated by the current operating mode and make adjustments to the operational settings.

The optimization system has safeguards put in place to manage system issues or unexpected events. For instance, if the operating situation is significantly different from all situations previously available during model training, the optimization system will switch to a fallback strategy. This strategy does not rely on the neural network forecasts and rather is based on measured WRRF loads. While the fallback strategy is not as precise in determining and applying optimal operational settings, it still ensures reliable compliance of effluent quality. In the event of lost critical input signals (e.g., from broken sensors or failure of the data connection to the SCADA system), the system reverts to conventional control that is automated, but based on simple DO control.

DISCUSSION

As illustrated by the two case studies, this methodology has significant potential for improving operational efficiency at WRRFs. Through these two case studies, several themes of lessons learned were encountered pertaining to project planning and implementation.

Data Availability and Quality

The first category of lessons learned centers on data availability and quality. Data is the backbone of this process. As such, abundant data is a requirement, both in terms of having long periods of data collection to represent a variety of process behavior and events, and having key variables measured to fully represent the treatment process and operation, its influent characteristics, and the effluent quality. At the beginning of this data-driven journey, it is imperative to understand what this data requirement looks like for the given context. While some WRRFs may already meet these requirements by means of online sensors, and frequent and consistent laboratory measurements, many WRRFs may need to undergo preliminary planning to understand where data gaps are, deploy and maintain sensors, and collect data for an extended time period. In these instances, it is valuable to have an implementation partner can support with sensor evaluation and deployment.

The availability of data must also be coupled with data quality. While the deployment of sensors in a WRRF may be a one-time activity, maintenance of the sensors is a continuous commitment to ensure that sensors are calibrated, are returning reliable data, and sensor issues are troubleshooted in a timely manner. As iterated in the Data Collection and Quality Control section, it is not feasible for online sensor measurements to be perfect, nor is it a requirement of data-driven modeling and optimization approaches. However, system behavior should be reasonably discernable from data signals and issues such as sensor drift, offline status, and clogs should be actively addressed. In coordination with operator, management, and maintenance teams, a sensor maintenance plan and schedule should be drafted and consistently adhered to in order to mitigate and preemptively address sensor issues.

Key questions to begin addressing data availability and quality include:

- Does your utility do online monitoring for relevant parameters already?
- What new sensors may be needed to fully visualize and inform a treatment process?
- What is the maintenance protocol for existing and new sensors for data quality?
- How will data gaps be accounted for and what is the strategy to QA online data?

Operator Buy-In

Another theme of lessons learned pertains to obtaining and building operator buy-in to an optimization process. Just as data is the backbone of a data-driven process, operators comprise the backbone of WRRF functionality. Operators ensure continuous day-to-day operation of the WRRF, and problem-solve in real-time during storm events and facility malfunction. The success of new or augmented WRRF operations depends heavily on the participation and trust of the operators in such changes. As a result, operators should be included from the beginning of the data-driven journey, in activities including, but not limited to:

- Identifying treatment processes that would be viable candidates for an optimization system
- Verifying the understanding of WRRF system behavior
- Giving context to behavior observed in data
- Developing fallback control strategies
- Assessing feasibility and efficacy of system recommendations or actions throughout system deployment

By working with operations staff to understand existing systems and operational protocols, stakeholders can design a system that will work in tandem with operations staff, embedding both operational experience and the insights that can be gained through a data-driven system.

To begin addressing operator buy-in, consider the following questions:

- Are operators already asking for these types of advanced control strategies?
- What challenges and concerns do operators face during day-to-day operations? How will an optimization system be designed to address these challenges?
- In what ways will operators be engaged throughout the process of building a new optimization system?
- How will the new system be monitored and its performance be assessed and shared?
- How will adherence to the system recommendations be measured to better understand system performance and operator buy-in?

- How will permits and/or internal operating procedures align with the proposed modifications to system operation?

IT Security

Another lesson learned is ensuring IT security standards are met in the design and implementation of the real-time optimization system. Whether the computational system is cloud-based, on-premise, or a creative hybrid approach, there are techniques to ensure that data communication is secure, safeguards are in place, and utility staff have confidence in the reliability and security of the system. From the two case studies and other applications, it has been found that:

- Based on IT security standards, agreement should be reached on which data can be communicated across internal and external systems, and in what direction data flow is permitted. For instance, some utilities consider movement of sensor data out of the internal utility network acceptable, but not movement of direct control actions into the utility network. These decisions will vary across utilities.
- Cloud-based systems typically have lower costs due to limited travel requirements for system updates and maintenance. This also applies to systems in which external technical support can remotely access the system.
- For on-premise deployments, consensus needs to be reached regarding ownership and accountability of computer hardware.
- If utility personnel supervision is required for external support on-premise, then dedicated time and resources need to be allocated to facilitate in such support.

To initiate the conversation around IT security, consider the following questions:

- What are the security policies and structures the utility IT department has in place regarding communication of data into and out of the WRRF computer system?
- Is there existing precedence in the utility for cloud-based vs. on-premise vs. hybrid architectures for real-time data and operations systems?
- What new security policies and/or structures would need to be put in place to accommodate a new real-time data and operations system that is cloud-based vs. on-premise vs. hybrid?
- For your context, what architecture option is most likely to succeed in terms of balancing security policies, cost, maintenance logistics, and function?

CONCLUSIONS

Digital twins have the potential to revolutionize how treatment facilities are operated in the coming years. As illustrated in the case studies, they can be used to achieve significant savings in operational expenditures while maintaining compliance with permitted effluent limits. The highly flexible nature of this approach allows for optimization of performance with respect to multiple KPIs, allowing for improved effluent quality and reduced operational costs. While both case studies presented focus on optimization of aeration control for secondary treatment, this methodology can be applied to other use cases such as chemical dosing or solids handling processes. However, this approach is not a quick fix. Among other factors, successful implementation requires high quality monitoring data, technical expertise, and support from operations and IT staff. But by engaging in the process, utility staff can gain not only a process control tool, but also a better understanding of how their plant operates. Through detailed review of existing monitoring data and discussions on known process dynamics and operational constraints, an implementation team can work together to build a platform that embeds and augments existing operational knowledge.

As the demand for digital solutions within the water industry grows, Xylem's team is engaged in ongoing research and development to continually improve our offerings. Active areas of research for the Plant Real-Time Decision Support solution include new use cases. One such example is modeling and optimization to reduce WRRF greenhouse gas emissions. New modeling approaches are also being evaluated to leverage ongoing advances in the field of machine learning. Additionally, new RTDS architectures are being developed to make deployment and maintenance easier, and to increase the security of the solution.

REFERENCES

Abdallah, M., Talib, M. A., Feroz, S., Nasir, Q., Abdalla, H., Mahfood, B. (2020). Artificial intelligence applications in solid waste management: A systematic research review. *Waste Management*, 109, 231–246. https://doi.org/10.1016/j.wasman.2020.04.057

Alsulaili, A., Refaie, A. (2021). Artificial neural network modeling approach for the prediction of five-day biological oxygen demand and wastewater treatment plant performance. *Water Supply*, 21(5), 1861–1877. https://doi.org/10.2166/ws.2020.199

Alvi, M., French, T., Cardell-Oliver, R., Keymer, P., Ward, A. (2022). Cost Effective Soft Sensing for Wastewater Treatment Facilities. *IEEE Access*, 10, 55694–55708. https://doi.org/10.1109/ACCESS.2022.3177201

Bahramian, M., Dereli, R. K., Zhao, W., Giberti, M., Casey, E. (2023). Data to intelligence: The role of data-driven models in wastewater treatment. *Expert Systems with Applications*, 217, Article 119453. https://doi.org/10.1016/j.eswa.2022.119453

Cheng, T., Harrou, F., Kadri, F., Sun, Y., Leiknes, T. (2020). Forecasting of Wastewater Treatment Plant Key Features Using Deep Learning-Based Models: A Case Study. *IEEE Access*, 8, 184475–184485. https://doi.org/10.1109/ACCESS.2020.3030820

Corominas, Ll., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., Poch, M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling & Software*, 106, 89–103. https://doi.org/10.1016/j.envsoft.2017.11.023

Dürrenmatt, D. J., Gujer, W. (2012). Data-driven modeling approaches to support wastewater treatment plant operation. *Environmental Modelling & Software*, 30, 47–56. https://doi.org/10.1016/j.envsoft.2011.11.007

Farhi, N., Kohen, E., Mamane, H., Shavitt, Y. (2021). Prediction of wastewater treatment quality using LSTM neural network. *Environmental Technology & Innovation*, 23, Article 101632. https://doi.org/10.1016/j.eti.2021.101632

Guo, Z., Du, B., Wang, J., Shen, Y., Li, Q., Feng, D., Gao, X., Wang, H. (2020). Data-driven prediction and control of wastewater treatment process through the combination of convolutional neural network and recurrent neural network. *RSC Advances*, 10, 13410–13419. https://doi.org/10.1039/D0RA00736F

Jawad, J., Hawari, A. H., Zaidi, S. J. (2021). Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review. *Chemical Engineering Journal*, 419, Article 129540. https://doi.org/10.1016/j.cej.2021.129540

Krivec, T., Hvala, N., Kocijan, J. (2021). Integrated theoretical and data-driven Gaussian Process NARX Model for the Simulation of Effluent Concentrations in Wastewater Treatment Plant. *IFAC-PapersOnLine*, 54(7), 714–719. https://doi.org/10.1016/j.ifacol.2021.08.445

Liu, X., Shi, Q., Liu, Z., Yuan, J. (2021). Using LSTM Neural Network Based on Improved PSO and Attention Mechanism for Predicting the Effluent COD in a Wastewater Treatment Plant. *IEEE Access*, 9, 146082–146096. https://doi.org/10.1109/ACCESS.2021.3123225

Matheri, A. N., Ntuli, F., Ngila, J. C., Seodigeng, T., Zvinowanda, C. (2021). Performance prediction of trace metals and cod in wastewater treatment using artificial neural network. *Computers & Chemical Engineering*, 149, Article 107308. https://doi.org/10.1016/j.compchemeng.2021.107308

Newhart, K. B., Holloway, R. W., Hering, A. S., Cath, T. Y. (2019). Data-driven performance analyses of wastewater treatment plants: A review. *Water Research*, 157, 498–513. https://doi.org/10.1016/j.watres.2019.03.030

Wang, G., Jia, Q.-S., Zhou, M. Bi, J., Qiao, J., Abusorrah, A. (2022). Artificial neural networks for water quality soft-sensing in wastewater treatment: a review. *Artificial Intelligence Review*, 55, 565–587. https://doi.org/10.1007/s10462-021-10038-8

Xu, B., Pooi, C. K., Tan, K. M., Huang, S., Shi, X., Ng, H. Y. (2023). A novel long short-term memory artificial neural network (LSTM)-based soft-sensor to monitor and forecast wastewater treatment performance. *Journal of Water Process Engineering*, 54, Article 104041. https://doi.org/10.1016/j.jwpe.2023.104041

Zhang, S.-Z., Chen, S., Jiang, H. (2022). A back propagation neural network model for accurately predicting the removal efficiency of ammonia nitrogen in wastewater treatment plants using different biological processes. *Water Research*, 222, Article 118908. https://doi.org/10.1016/j.watres.2022.118908

Zhao, L., Dai, T., Qiao, Z., Sun, P., Hao, J., Yang, Y. (2020). Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. *Process Safety and Environmental Protection*, 133, 169–182. https://doi.org/10.1016/j.psep.2019.11.014

Zhu, J., Jiang, Z., Feng, L. (2022). Improved neural network with least square support vector machine for wastewater treatment process. *Chemosphere*, 308(1), Article 136116. https://doi.org/10.1016/j.chemosphere.2022.136116