Enhancing Struvite Recovery Systems through Top-Down Machine Learning Based Prediction Models

March 2025

2025 IWEA Intelligent Water Systems Challenge

Team Members:

Rishabh Puri – Civil and Environmental Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL, United States. **rp34@illinois.edu**

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

Roles and Responsibilities**:** Team Lead, Data Analysis, Model development

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

Expertise: Struvite Crystallization Process Optimization and Simulation, ML Based Modelling

Roles and Responsibilities**:** Crystallization Process Benchmarking, Data Preprocessing, Model development

Don Sim - Civil and Environmental Engineering Department, University of Illinois at Urbana-Champaign, Urbana, IL, United States. **dhsim2@illinois.edu**

Expertise: Struvite Crystallization Process, Deep Learning Model Architecture

Roles and Responsibilities: Literature Review, Deep Learning Model Architecture, Technical Report

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

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

Roles and Responsibilities**:** Academic Advisor

1. Project Summary

The Madison Metropolitan Sewerage District (MMSD) operates a moderately-sized treatment facility, the Nine Springs Wastewater Treatment Plant (NS WWTP), which serves a population of 360,000 residents within an area of 180 square miles (466 km²). This 42 MGD facility, located in Madison, WI, is equipped for phosphorus removal and recovery using mainline enhanced biological phosphorus removal (EBPR) and side stream struvite crystallization. The treatment facility faces significant challenges in optimizing the efficiency of its struvite-based nutrient recovery processes. Current precipitation process simulations are computationally intensive and require extensive data collection to accurately account for the impacts of solution chemistry, particle growth dynamics, and reactor hydrodynamics on overall reactor performance.

To address these challenges, a university team was formed to propose and implement a ML based model capable of predicting reactor performance from existing historical operations data. From the perspective of the NS WWTP, a major driver of the need to improve performance is to ensure consistent P removal and recovery, thereby reducing the effects of struvite scaling within the plant as well as reducing operational costs with consistent revenue generation from struvite sales. Secondly, developing insights into which features most impact the struvite recovery process would allow for the creation of a framework of minimum features required for viable performance prediction i.e. the minimum features to be considered for future data collection campaigns.

The major aim of this work is to develop a data driven model that optimizes the removal and recovery of P as struvite by analysing conversion and yield. This tool will enhance the prediction of conversion and yield enabling modification of the process controls necessary to improve process performance. This modelling framework can be easily deployed by other utilities that struggle to manage the performance of side stream struvite precipitation.

2. Background and Problem Statement

2.1 Operational Challenges

Since November 2013, NS WWTP has employed Ostara's Pearl and WASSTRIP technologies to manage its struvite ($MgNH_4PO_4*6H_2O$) recovery system. A simplified plant schematic is depicted in Figure 1 showing P flow throughout the NS WWTP. This system injects sodium hydroxide (NaOH) and magnesium chloride ($MgCl_2$) into EBPR sludge filtrate feed in a fluidized bed reactor seeded with struvite pellets (diameter ~ 1 mm) to drive struvite precipitation. Two reactors operate in parallel, each with a volume of 30,949 L and featuring a recycle stream (recycled to the reactor influent), with 24,113 L of reactor volume placed before the recycle line. The damp struvite product is collected and dried using recycled heated water from the facility's system, then organized, stored, and packaged in one-ton bags. Ostara collects and stores these "prills," the solid particles produced, and markets them as fertilizer. The facility is expected to produce approximately 2.5 metric tons per day. [1]

However, the NS WWTP still encounters challenges in optimizing its nutrient recovery efforts both in terms of P removal (conversion) and recovery (yield). First, the nutrient recovery system currently operates with a manual feedback control system for magnesium dosing based on Mg:P molar ratios. This primarily effects conversion, with periods of over and under dosing resulting from limited updates to the dosing systems and low frequency collection of influent P data. A ML based model could reduce overdosing costs by providing insights into the required magnesium dosing, controlling conversion and yield parameters of the reactors, and preventing excessive generation of fines in the effluent, which

could alter reactor performance. Secondly, the plant experiences limited control of "upset" events where yield can wildly deteriorate due to the washout of fine particles. While historical phosphorus flows and heuristic knowledge of struvite precipitation are currently used to operate the crystallization system, the complex dynamics of solution chemistry, particle generation, and reactor hydrodynamics necessitate a more sophisticated model-based control strategy to maximize yield. Existing precipitation process models require a complex dataset [2,3] not currently available based on the existing NS WWTP historical operations data.

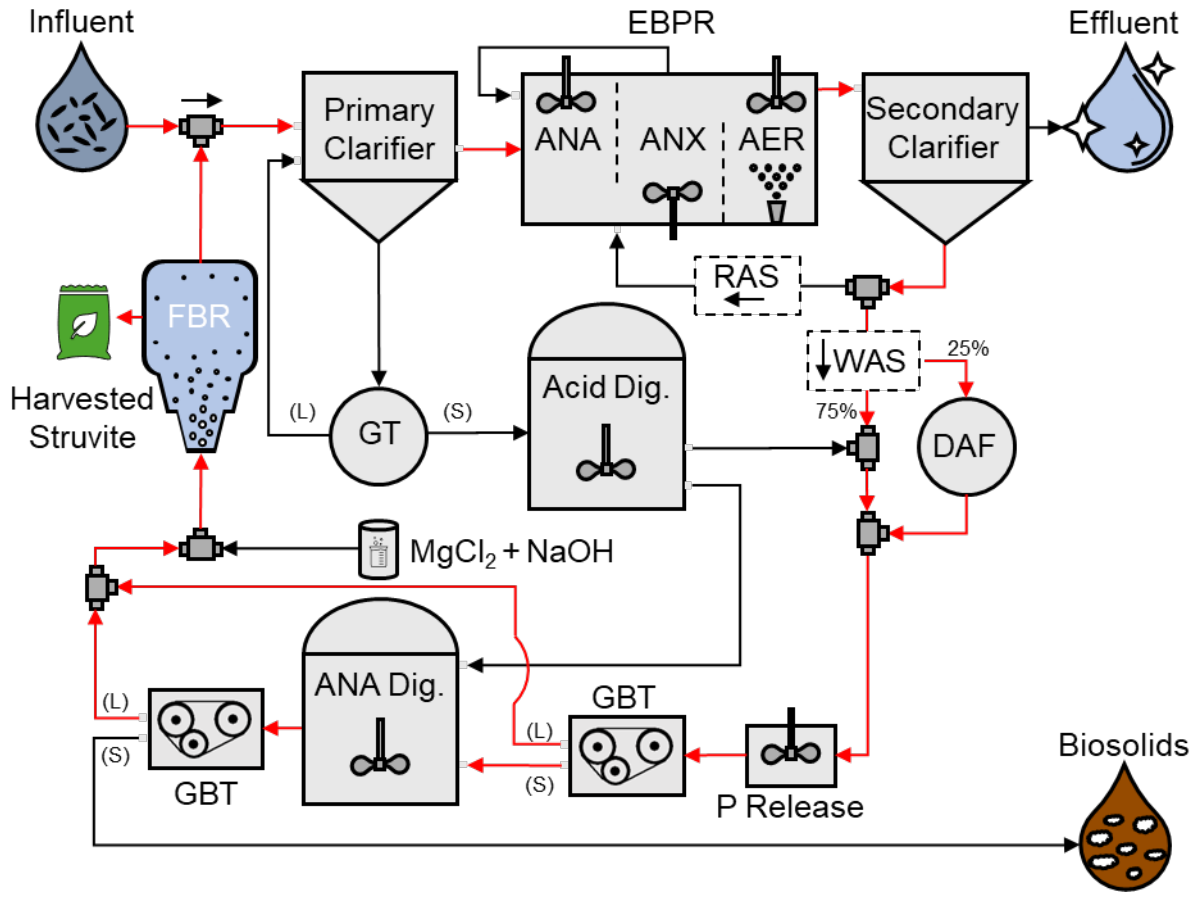

Figure 1. Simplified Plant Schematic with key P pathways in red. Influent P is removed by EBPR. P rich EBPR sludge is then fermented in a WASSTRIP to promote P release and separated by GBT where filtrate is sent directly to an Ostara Pearl struvite crystallizer (FBR). GBT solids are digested and separated again with GBT; the resulting liquids are sent to the struvite crystallizer. Struvite is formed in the crystallizer by a magnesium chloride and sodium hydroxide. Crystallizer effluent is returned to the plant headworks.

Before this report, these issues were in part investigated by the authors and findings were published in the dissertation titled "Understanding the Limits of Struvite Based Phosphorus Recovery Through Mechanistic Modelling and Data-driven Performance Evaluations". In that work the analysis focused on prediction of reactor performance based on feed flows only (i.e. the GBT filtrate flows shown in Figure 1). The analysis proposed here extends previous work by including the effects of recycle flow, which allows for a better accounting of the flow rate and solution composition directly entering the struvite crystallizer. The proposed work also expands the types of models considered for prediction.

2.2 Proposed Intelligent Water System Solution and Objectives

Existing data collection practices in the operation of full-scale struvite recovery systems have led to an abundance of operations data, but a limited ability to predict reactor performance or develop process optimization strategies. State of the art population balance models for mineral precipitation systems require an intensified and specialized data collection campaign that are outside of current collection practices and would require significant investment in advanced particle sensing instrumentation from WWTP lab facilities. This highlights the need to develop an alternative data driven approach to process evaluation as described in this report. The main objectives of this work are as follows:

- Generate an enhanced dataset that expands the raw operations to include thermodynamic features related to mineral solubility, metal to phosphorus molar ratios, cyclical encoding to include seasonal effects, and lagged performance indicators to account for unsampled but significant process drivers (i.e. seed bed height not directly measured, but whose effect is indirectly captured in previous reactor performance).
- Development and comparison of various machine learning based modelling strategies to predict orthophosphate removal (ΔOP), Conversion and Yield
- Propose a user interface that provides guidance on process sensitivity to operational controls and indicates scenarios where low performance is expected for current influent conditions

A general overview of the proposed solution is presented in Figure 2. By further enhancing the existing historical operations data with expert analysis and feature engineering methods we expect the development of a data driven modeling framework to provide utilities and practitioners an analysis framework that enables the development of predictive models for struvite recovery systems without the need for an increase in data collection practices. Further this will allow for the development of process optimization strategies, either increasing P removal or reducing operational costs, that are currently limited due to the complexity of existing process models.

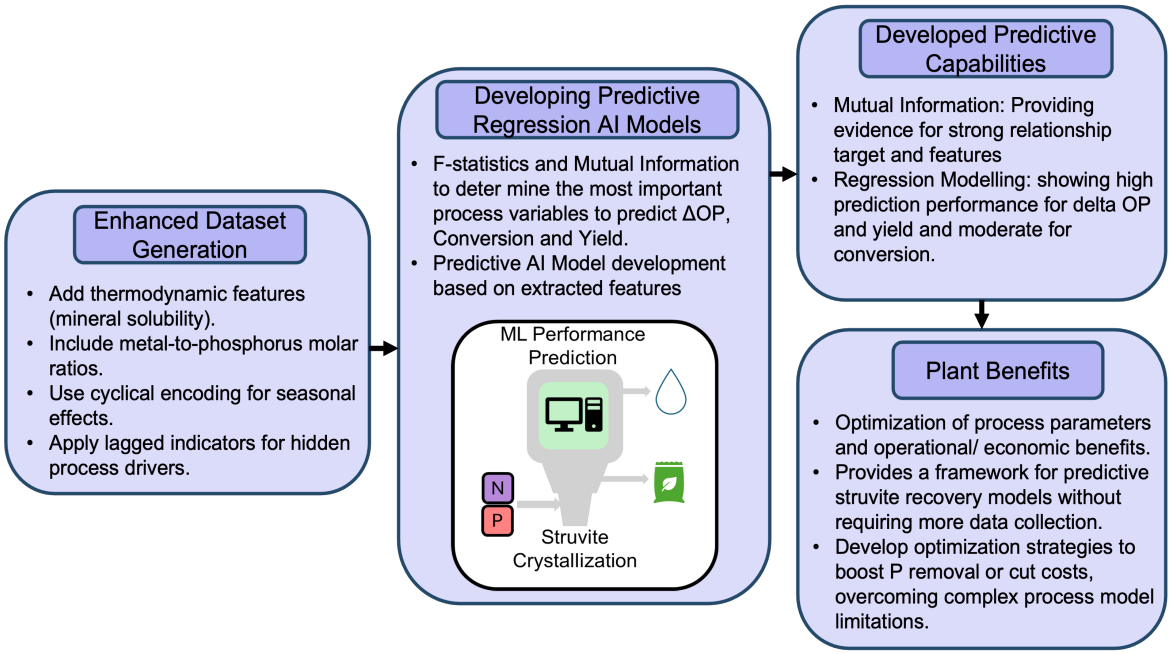

Figure 2. Proposed Solution Framework

3. Methodology

3.1. Data Collection

This study utilizes historical data (2018–2023) on flow rates, internal recycle rates, and crystallizer influent/effluent concentrations (TSS, FSS, TP, PO_4-P, NH_4-N, Mg, Na, Ca, K, Fe), along with effluent pH, pressure differential, and plant effluent temperature. Data collection frequency varies but includes daily recordings at most. Preliminary reactor analysis considered both influent and effluent parameters, while ML regression models focused only on influent data and real-time effluent pH to prevent data leakage. Earlier data (2013–2018) was excluded due to inconsistencies during the plant's initial learning phase. To assess struvite crystallizer performance, the change in orthophosphate (ΔOP), Conversion, and Yield were calculated for each observation. A mass balance was conducted on all phosphate-containing species within a system boundary encompassing influent, effluent, and harvested struvite flows, as illustrated in Figure 2.

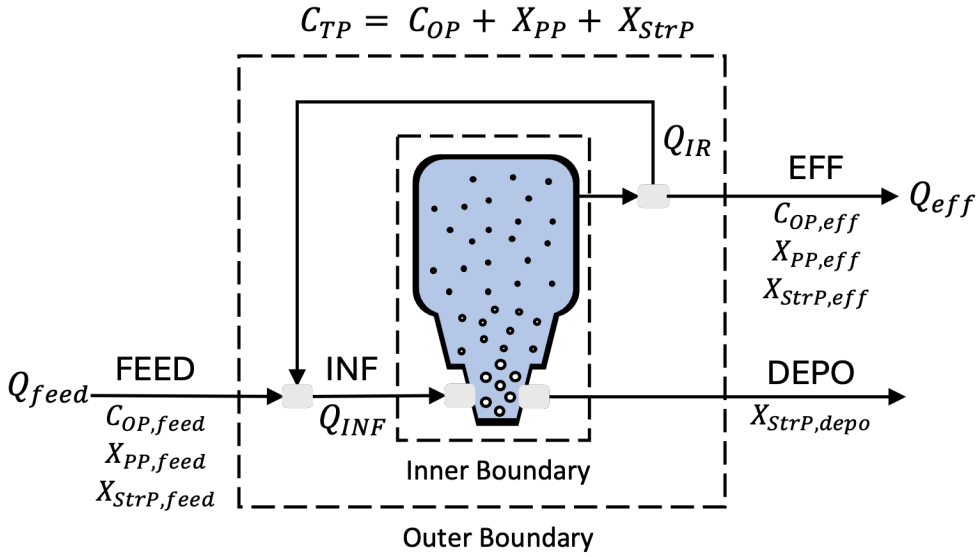

Figure 3. Simplified Flow Diagram for Struvite Reactor

The analysis of this reactor was categorized into two system boundaries: Inner Flows and Outer Flows. The outer features only consider the feed flow, dosed flows, effluent flow, and harvested struvite as well as each streams composition. In contrast inner features consider all the outer features as well as the internal recycle flow which are summed to generate an influent reactor flow and composition. The internal recycle composition is not used directly as a feature to not introduce data leakage. Features labeled as "effective" do not account for recycle flows. The nomenclature for target variables was determined based on inner and outer features, leading to classifications such as Delta OP mM inner and outer, conversion inner and outer, and yield inner and outer. The formulas for both inner and outer target variables are detailed below with a full mass balance given in appendix 9.2

$$\Delta OP\ (outer) = C_{OP,feed} - C_{OP,eff}\ ;\ \Delta OP\ (inner) = C_{OP,inf} - C_{OP,eff} \quad (1)$$

$$Conversion\ (outer) = 1 - \left(\frac{C_{OP,eff}}{C_{OP,inf}}\right);\ Conversion\ (inner)\ = 1 - \left(\frac{C_{OP,eff}}{C_{OP,inf}}\right) \quad (2)$$

$$Yield\ (outer) = \frac{C_{TP,feed} - C_{TP,eff}}{C_{OP,feed} - C_{OP,eff}}\ ;\ Yield\ (inner) = \frac{C_{TP,inf} - C_{TP,eff}}{C_{OP,inf} - C_{OP,eff}} \quad (3)$$

$$HRT = \frac{V_{rxtr}}{Q_{feed} + Q_{IR} + Q_{Mg\ Dose} + Q_{Na\ Dose}}\ ; Effective\ HRT = \frac{V_{rxtr}}{Q_{feed} + Q_{Mg\ Dose} + Q_{Na\ Dose}} \quad (4)$$

Where, $C_{OP,feed}$, $C_{OP,inf}$, and $C_{OP,eff}$ are the concentration of PO_4 - P in the crystallizer feed, influent, and effluent respectively; $C_{TP,feed}$, $C_{TP,inf}$, and $C_{TP,eff}$ are the concentration of PO_4 - P in the crystallizer feed, influent, and effluent respectively. HRT and $Effective\ HRT$ are the hydraulic residence time with and without accounting for the recycle flow (i.e. true HRT and single pass HRT). Equation 3 is derived with an assumption that P phases other than orthophosphate and struvite (i.e. microbial P, organic P, condensed P) forms remain non-reactive and do not transition between liquid and solid phases in the struvite crystallizer.

3.2 Data QA/QC Considerations and Preprocessing

The frequency of data collection has varied over time and across different parameters, with daily recordings being the highest frequency since 2018. Initial preprocessing involved compressing asynchronous reporting by different users to a daily entry or average value if multiple measurements were made. Any positive or negative infinite values were removed from the dataset. All considered parameters were filtered to be within ±3 standard deviations from the mean value to account for extreme outliers or data entry errors. Only conversion between 0 and 1 and yield between 0 and 1 were considered as standard operation of the crystallizer lies in this range. The data were then ready for feature engineering and importance analysis.

3.3 Feature Engineering and Importance Analysis

The subsequent step involves generating features based on thermodynamic drivers of precipitation, autocorrelation, and other time-dependent factors. The molar ratios of influent Mg:P, N:P, and Mg:N are included as variables to determine the system's limiting reactant. Additionally, the inverse of OP concentration is incorporated as a variable to potentially improve the prediction of Conversion (see equation 2). The saturation index of struvite, brushite, amorphous calcium phosphate (ACP), hydroxyapatite, and vivianite in the influent (post-$MgCl_2$ and NaOH dosage) is calculated for each observation using Visual MINTEQ V4.05 [4] and included as features. To account for seasonal weather patterns and weekly or monthly variations (e.g., weekdays vs. weekends), cyclical encoding of the observation date is applied within the dataset.

In accordance with best practice guidelines [5] for sample-to-feature ratios (ranging from 10 to 100), feature importance was determined using two distinct metrics: the F-statistic, which evaluates linear relationships between features and the target variable, and Mutual Information (MI), which measures information gain, including both linear and non-linear dependencies [6]. Further description of how these two ranking metrics were used within each model are described in later sections.

3.4 ML Regression Models

A range of advanced machine learning methods was employed to develop models capable of predicting ΔOP, conversion, and yield. These models include Multiple Linear Regression, Extreme Gradient Boosting (XGBoost), Random Forest, and Gradient Boosting, all implemented in Python using the "scikit-learn [7]," "xgboost [8]," libraries. Manual hyperparameter tuning was performed using scikit-learn and learning curve plots, while 5-fold cross-validation was applied to enhance model generalizability. Ultimately, the final models were selected from various tested models to ensure robust predictive performance. The accuracy of the regression predictions was evaluated using the coefficient of determination (R^2) and the root mean square error (RMSE).

3.5 Proposed Timeline

The team set the following dates as internal completion milestones corresponding roughly to the sections included in this report.

- Data Preprocessing – December 24, 2024
- Feature Importance Analysis – January 10, 2025
- Model Development – January 25, 2025
- Analysis of Results – February 5, 2025
- Report Writing and Submission – February 17, 2025

4. Results and Discussion

4.1 Feature Importance Analysis

To determine which features are most likely to impact model sensitivity all features were ranked according to the F statistic and MI scores for each of the target labels as shown in Figure 3. For the first iteration of model development, the feature sets selected included only the top 10 features from both metrics. These sets were created without replacement where overlap existed between the F score and MI to reduce the inclusion of errant features. For ΔOP, the top four features across both metrics were feed OP and its inverse value, feed TP, and dosed NaOH flow suggesting P removal is independent of magnesium dosing and that overdosing is likely occurring under current operation. For conversion, the top two features across both metrics were dosed flows for NaOH and $MgCl_2$ indicating solution composition is a driver of performance agreeing with previous experimental [9] and simulation [10] work focused on struvite precipitation in upflow FBRs. Finally, for Yield the top three features were influent TP-OP (i.e. particulate P), TP, and FSS. These metrics all point towards the importance of particulates in the influent stream and suggest particle capture mechanisms related to aggregation of smaller freshly generated particles as a main driver of yield. In a system with high recycle flow, such as this one, this points towards a sensitivity towards high effluent solids which can further exacerbate fines generation and capture. Notably, across all labels there was good agreement between both the F score and MI values indicating linear models alone may be strong predictors. The following regression analysis considered both linear and nonlinear options to develop a better understanding of the nonlinear components and whether they could introduce substantial predictive power to any models relative to linear models alone.

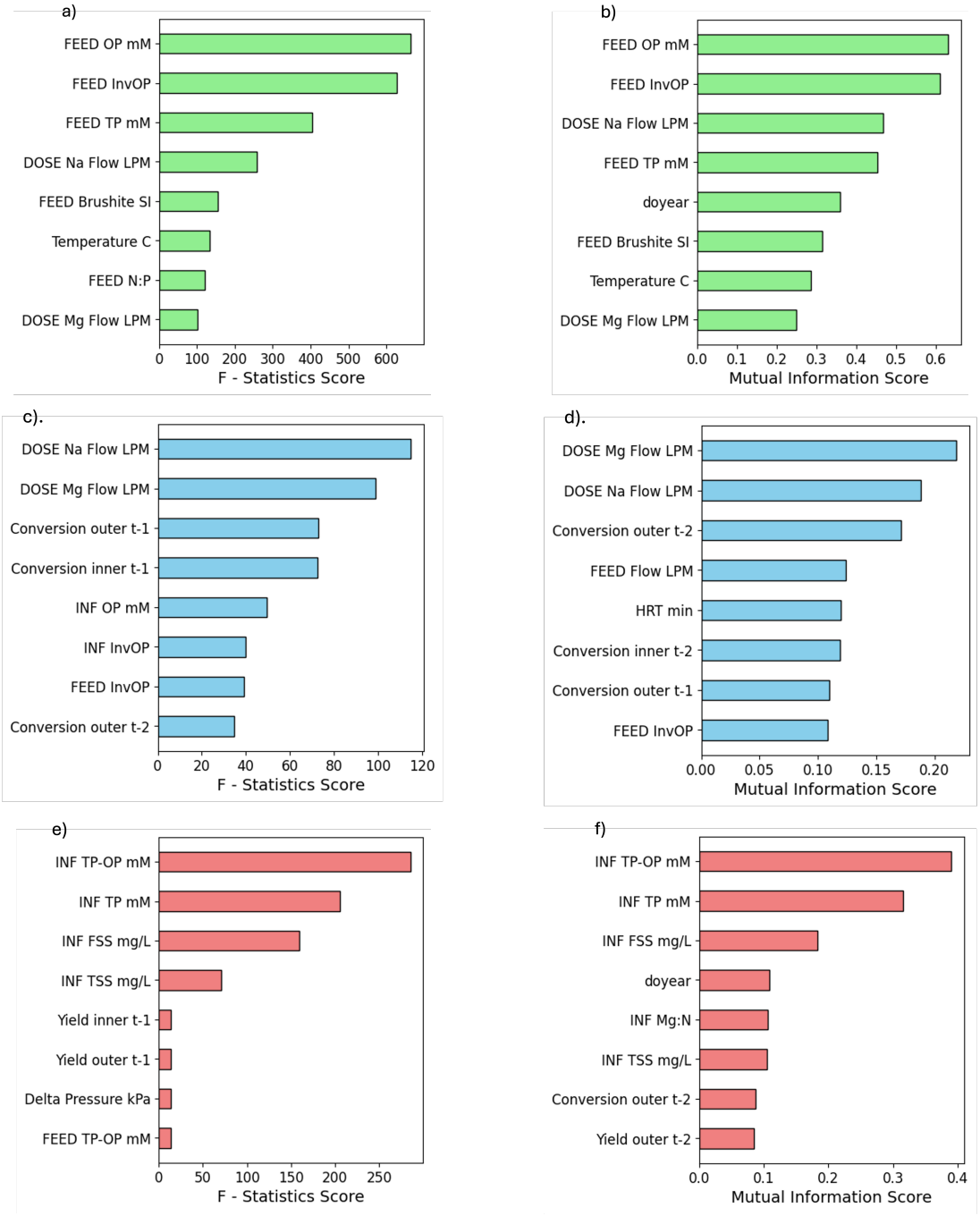

Figure 4. F statistic score for Delta OP (green), Conversion (blue) and Yield (red) are shown in Fig. 4a, c, and e respectively, and Mutual Information scores for Delta OP, Conversion and Yield are shown in Fig. 4b, d, and f respectively

4.3 Prediction of ΔOP, Conversion, and Yield

After conducting correlation analysis on the regression input dataset and selecting sub-datasets with the final variables, various AI models were developed to predict ΔOP, Conversion, and Yield for the struvite reactor. Different model types were chosen to accommodate both linear and nonlinear interactions existing within the dataset. The summarized results featuring the results from optimized models across all parameters are shown in Table 1. For ΔOP the best model was a gradient boosting model with an R^2 of 0.82 and RMSE of 0.04. Interestingly, the XGBoost model structure slightly underperformed the gradient boosting model which is likely due to the differences in the regularization methods included in XGBoost (and are not present in gradient boosting). These differences could cause overfitting of the gradient boosting method and since the dataset investigated here is limited, there could be unusual sampling occurring even across the 5-fold cross validation leading to overestimates of gradient boosting R^2 that would be evident with additional data. For Conversion the best model was a multilinear regression with an R^2 of 0.84 and RMSE of 0.04. All nonlinear methods underperformed the MLR indicating that at least for this label, the inclusion of nonlinear features causes model confusion through the incorporation of noisy and low information gain features. This is supported by the relatively weaker MI values shown in Figure 4 for Conversion relative to the other labels which feature stronger nonlinear correlations. Finally, for Yield the strongest model was XGBoost with an R^2 of 0.82 and RMSE of 0.07. Here nonlinear effects slightly improve predictive performance, though MLR only slightly underperformed indicating linear features were most important to overall predictive capability in agreement with the strong overlap of high ranking F statistic and MI features shown in Figure 4e and 4f. This dataset highlights the importance of testing both linear and nonlinear models as well as models with varying levels of complexity to best address the predictive task.

Table 1. Regression Performance on Test Sets

Model	R2	RMSE
	Delta OP mM inner	
MLR	0.75	0.06
XGBoost	0.78	0.05
Random forest	0.72	0.05
Gradient Boosting	**0.82**	**0.04**
	Conversion inner	
MLR	**0.84**	**0.04**
XGBoost	0.81	0.04
Random forest	0.79	0.05
Gradient Boosting	0.81	0.05
	Yield inner	
MLR	0.79	0.07
XGBoost	**0.82**	**0.07**
Random forest	0.70	0.09
Gradient Boosting	0.78	0.07

5 Software Dashboard Integration

Despite the immense potential that AI offers WWTPs to fully utilize their available data for optimized plantwide operation, developing functional models and using them for decision-making can be a challenging process for users who may not necessarily have a data science background and are experts in other fields. Therefore, an interface that integrates all the desired AI functionalities based on a WWTP's needs would be a crucial step in encouraging the application of AI technology in the water/wastewater sector. Thus, a conceptual software dashboard integrating all the features developed in this solution into an easy-to-use interface was developed and shown in Figure 5. The tool first (left panel) allows a user to select a label to predict (ΔOP, Conversion, and Yield, etc.), a model type (though XGB is selected as the default trained model), shows the model input features ranked by their output sensitivity, and the current output prediction assuming no changes are made to the system. The center panel features a series of windows which show the model outputs as a function of a user selected feature. This is helpful to visualize an optimization strategy, or the impact of a potential operator led process change. Finally, the right panel acts as a SCADA system control updater by showing the current system state for a single selected model feature, and a cursor mark for an update state (movable by user). The performance table below is used to understand the overall performance impact on ΔOP, Conversion, and Yield simultaneously. A button at the bottom of the right panel is then used to send the update state to the SCADA system. This dashboard focuses on linking operator decisions to process performance without a need for direct control of the ML models developed in this analysis.

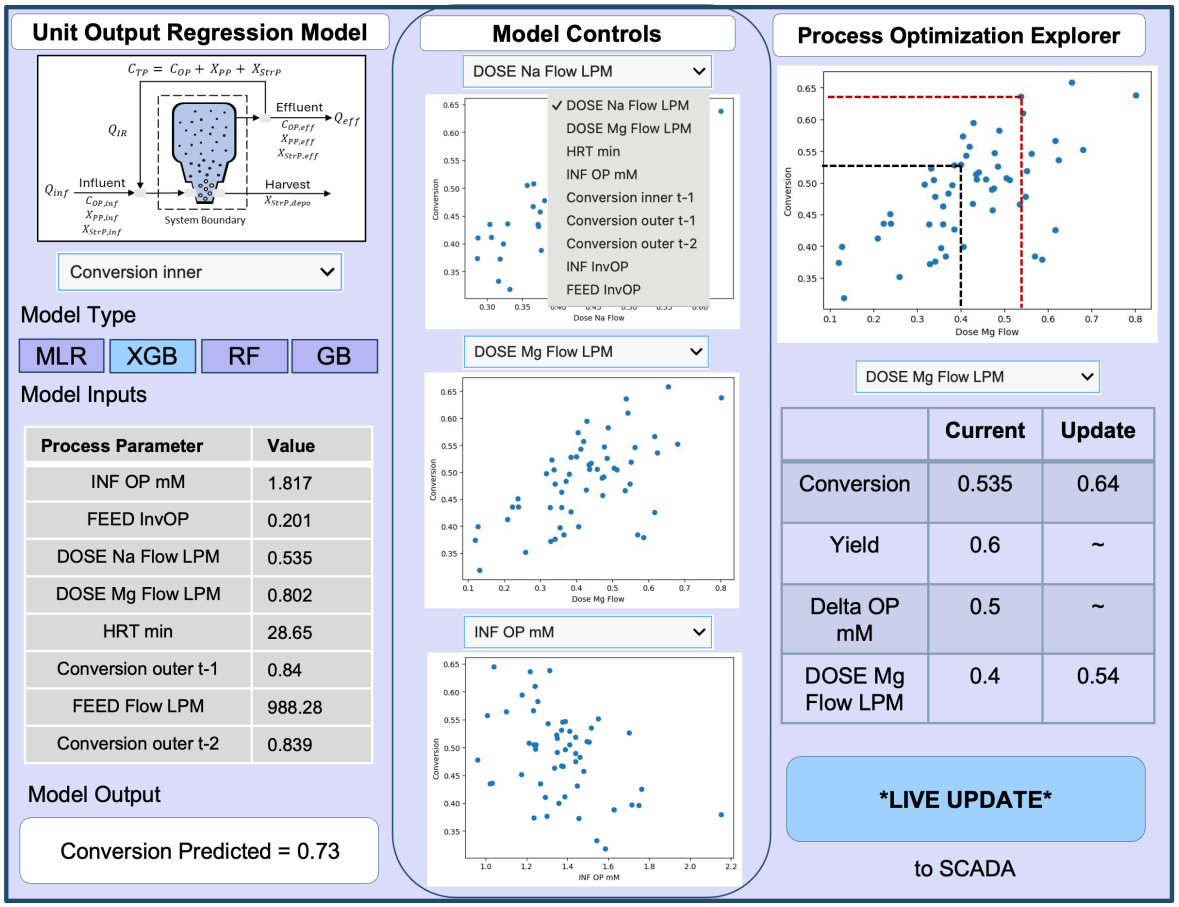

Figure 5. Software dashboard integrating all the AI models developed in this analysis into a user-friendly and easy-to-use interface.

6. Implementation of AI and the Next Steps for This Solution

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

- **Team Formation:** The successful implementation of AI in WWTPs for decision-making and process optimization necessitates a multidisciplinary team. Expertise in water and wastewater process engineering, data science, analysis, visualization, and data integration is essential. This project benefited from the diverse skills of team members with backgrounds in process engineering, struvite reactor knowledge, and machine learning.
- **Primary Challenges:** The team faced several significant challenges, including developing effective data collection, cleaning, and preprocessing methods, selecting the most appropriate algorithms, and ensuring the models' practical application for the plant. Overcoming these challenges was possible due to the team's interdisciplinary expertise, thorough literature review, and regular collaborative meetings.
- **AI for Decision-Making:** A major obstacle in fully leveraging AI in WWTPs is the disconnect between complex AI model development and the need for user-friendly interfaces. It is crucial to present AI benefits in a way that is accessible to professionals without a data science background. Experts with dual knowledge in process engineering and data science can help translate AI advancements into practical tools for WWTPs.
- **Broader Implementation and Future Directions:** The research team from the University of Illinois at Urbana-Champaign is reaching out to other utilities across the United States to acquire data from various struvite reactors. This data will be used to apply and adjust the model, enhancing its generalizability and making it applicable to a broader range of struvite reactor data sets.

A helpful next step would be to determine if low yield events have influenced the performance of EBPR. Specifically, this would involve examining the data to see if these low yield events have affected key indicators such as the phosphorus levels in the plant effluent.

7. Conclusions, Advantages for Utilities, and Next Steps

The development of this prediction tool will offer utilities and industry practitioners a comprehensive analytical framework for developing predictive models tailored to struvite recovery systems. This framework aims to eliminate the necessity for extensive additional data collection, making model development more efficient and accessible. Furthermore, it will facilitate the creation of process optimization strategies that enhance phosphorus removal efficiency or lower operational costs. These advancements address current limitations stemming from the complexity of existing process models, ultimately contributing to more sustainable and cost-effective struvite recovery operations. The model's ability to decompose values into structural components and reveal underlying patterns further enhances its utility, offering deeper insights into the dynamics of the struvite recovery system.

This analysis on the NS WWTP struvite crystallizer can be summarized to the following insights:

- Though struvite crystallization is a complex collection of crystallization phenomena paired with specific reactor hydrodynamics and nonlinear interactions, the primary analysis developed for future predictive systems should focus first on linear interactions. Across all labels tested linear interactions were strong and performed nearly as well or in some cases outperformed more complex models.

- General heuristic development should focus on influent OP, Mg/NaOH dosing, and particulate P for ΔOP, Conversion, and Yield, respectively.
- Significant differences exist between the predictive capabilities of models developed for inner and outer (not shown here) predictions. Future automated control systems should look to act based on inner predictions as they better incorporate the role of recycle flows which can be strong sources of particulate P and cause a dilution effect on fresh feed.
- Though some models were developed with sufficient accuracy across all prediction tasks ($R^2 >$ 0.8) to provide guidance to operators, a better performing model is needed for long-term confidence in the potential costs savings that could be achieved through operational optimization. Better datasets should be sought both in terms of more controlled conditions (i.e. pilot scale testing with a specific goal to create a ML focused dataset) and across different system types (various struvite crystallizers have been commercialized and feature significant differences in capture mechanisms).

This proposed intelligent water system solution can assist the NS WWTP in evaluating various scenarios to predict their influent OP loadings. This capability allows for better prediction of conversion and yield and operator intervention to optimize reactor performance. Future work should seek to incorporate operational costs into the suggested process updates as economic factors ($ per kg of struvite recovered resulting from chemical and pumping energy requirements or reductions in maintenance costs related to struvite scaling) may be more important than reactor performance alone. The intelligent water system developed in this study can be applied to other utilities nationwide, providing an optimized control strategy that can be integrated into existing control systems.

7. References

1. Grooms A, Reusser S, Dose A, Britton A, Prasad R: **Operating Experience with Ostara Struvite Harvesting Process**. *proc water environ fed* 2015, **2015**:2162–2177.

2. Elduayen-Echave B, Lizarralde I, Larraona GS, Ayesa E, Grau P: **A New Mass-Based Discretized Population Balance Model for Precipitation Processes: Application to Struvite Precipitation**. *Water Research* 2019, **155**:26–41.

3. Galbraith SC, Schneider PA, Flood AE: **Model-driven experimental evaluation of struvite nucleation, growth and aggregation kinetics**. *Water Research* 2014, **56**:122–132.

4. Gustafsson JP: **Visual MINTEQ 3.0 user guide**. *KTH, Department of Land and Water Recources, Stockholm, Sweden* 2011, **550**.

5. **Machine Learning in Environmental Research: Common Pitfalls and Best Practices | Environmental Science & Technology**. [date unknown],

6. **Phys. Rev. E 69, 066138 (2004) - Estimating mutual information**. [date unknown],

7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al.: **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research* 2011, **12**:2825–2830.

8. **XGBoost Python Package — xgboost 2.1.3 documentation**. [date unknown],

9. Ye X, Ye Z-L, Lou Y, Pan S, Wang X, Wang MK, Chen S: **A comprehensive understanding of saturation index and upflow velocity in a pilot-scale fluidized bed reactor for struvite recovery from swine wastewater**. *Powder Technology* 2016, **295**:16–26.

10. Ye X, Gao Y, Cheng J, Chu D, Ye Z-L, Chen S: **Numerical simulation of struvite crystallization in fluidized bed reactor**. *Chemical Engineering Science* 2018, **176**:242–253.

8. Disclosures

No financial support from any person(s), institution, and/or company was received to prepare this report and participate in the 2024 LIFT Intelligent Water Systems Challenge.

9. Appendix

9.1 Abbreviations

Abbreviation	Description	Abbreviation	Description
ACP	Amorphous Calcium Phosphate	MMSD	Madison Metropolitan Sewerage District
AI	Artificial Intelligence	Na	Sodium
Ca	Calcium	NaOH	Sodium Hydroxide
CSTR	Continously Stirred Tank Reactor	NH_4-N	Ammoniacal Nitrogen
EBPR	Enhanced Biological Phosphorus Removal	OP	Orthophosphate
Fe	Iron	P	Phosphorous
FSS	Fixed Suspended Solids	PO_4-P	Soluble Phosphorous
GB	Gradient Boosting	PSD	Particle Size Distribution
GBT	Gravity Belt Thickener	RF	Random Forest
HRT	Hydraulic Retention Time	RMSE	Root Mean Square Error
K	Potassium	SCADA	Supervisory Control and Data Acquisition
Mg	Magnesium	SI	Saturation Index
$MgCl_2$	Magnesium Chloride	TP	Total Phosphorous
MI	Mutual Information	TSS	Total Suspended Solids
ML	Machine Learning	WWTP	Wastewater Treatment Plant
MLR	Multilinear Regression	XGBoost	Extreme Gradient Boosting
mM	Millimolar		

9.2 Mass Balance Expressions, Conversion, and Yield

Total Phosphorus Expressions

$$TP = OP + Struvite\ P + Organic\ P$$

Total Phosphorus Difference Across Reactor

$$\Delta TP = C_{TP,inf} - C_{TP,eff} = \left(C_{OP,inf} + C_{StrP,inf} + C_{OrgP,inf}\right) - \left(C_{OP,eff} + C_{StrP,eff} + C_{OrgP,eff}\right)$$

Assuming no removal of organic P during crystallization (ΔOrganicP $= 0$)

$$\Delta TP = \Delta OP + \Delta StrP$$

Struvite Mass Balance Expression

Assume no consumption from dissolution, generation by OP precipitation without co-precipitants, SS.

$$IN - OUT + GEN - \cancel{CONS} = \cancel{ACC}$$

$$Q_{inf}X_{StrP,inf} - Q_{eff}X_{StrP,eff} + Q_{inf}\left(C_{OP,inf} - C_{OP,eff}\right) - M_{StrP,depo} = 0$$

Given $Q_{inf} = Q_{eff}$,

$$Q_{inf}(\Delta StrP + \Delta OP) = M_{StrP,depo}$$

$$Q^{-1} * M_{StrP,depo} = X_{StrP,depo} = \Delta StrP + \Delta OP$$

Conversion

$$Conversion = \left(\frac{P\ converted\ to\ Struvite}{Influent\ P}\right) = \left(\frac{\Delta OP}{C_{OP,inf}}\right)$$

$$Conversion = \left(\frac{C_{OP,inf} - C_{OP,eff}}{C_{OP,inf}}\right)$$

$$Conversion = 1 - \left(\frac{C_{OP,eff}}{C_{OP,inf}}\right)$$

Yield

$$Yield = \left(\frac{deposited\ struvite}{total\ formed\ struvite}\right) = \frac{X_{StrP,depo}}{\Delta OP}$$

Recall, $X_{StrP,depo} = \Delta StrP + \Delta OP$,

$$Yield = \frac{\Delta StrP + \Delta OP}{\Delta OP}$$

Recall, $\Delta TP = \Delta OP + \Delta StrP$,

$$Yield = \frac{\Delta TP}{\Delta OP}$$

9.3 Regression model hyperparameter tuning

Initially, MLR was tested as a baseline model, but its assumption of linearity limited the potential for incorporation of nonlinear interactions. RF was then implemented to improve accuracy using ensemble learning, but its reliance on bagging made it computationally expensive. To further enhance predictive power, GB was applied, leveraging boosting to sequentially refine weak learners. Finally, XGBoost, an optimized gradient boosting algorithm with regularization and speed improvements, was developed and tested, showing strong predictive capabilities. Different strategies were used to ensure model robustness. In ML models, manual search with cross-validation was applied to optimize parameters. The model parameters considered for development are as follows:

- **RF**
  - n_estimators: [10, 20, 30, 40, 50, 60, 70, 80]
  - max_features: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10 ,11, 12, 13, 14, 15]
  - max_leaf_nodes: [5, 10, 15, 20, 25, 30, 35, 40, 45, 50 , 55, 60, 65, 70, 75]
- **GB**
  - n_estimators: [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
  - learning_rate: [0.01, 0.5] with 10 linearly spaced values
  - max_depth: [2, 4, 6, 8, 10, 12, 14, 16, 18, 20]
- **XGBoost**
  - n_estimators: [1,40] with increments of 1
  - learning_rate: [0,20]*0.1
  - gamma: [0,10]*0.1
  - max_depths: [1,20] with increments of 1

Optimized Hyperparameters Across All Model Types

	XGBoost Hyperparameters			
Predicted Label	**n_estimators**	**max_depth**	**learning_rate**	**gamma**
ΔOP	37	3	0.1	0.0
Conversion	39	3	0.19	0.0
Yield	39	5	0.15	0.0
	RF Hyperparameters			
Predicted Label	**n_estimators**	**max_features**	**max_leaf_nodes**	**-**
ΔOP	30	9	30	-
Conversion	50	12	50	-
Yield	30	9	50	-
	GB Hyperparameters			
Predicted Label	**n_estimators**	**max_depth**	**learning_rate**	**-**
ΔOP	40	8	0.12	-
Conversion	90	14	0.45	-
Yield	60	8	0.23	-