Forecasting total phosphorus in wastewater treatment plant effluent with machine learning – a case study in Ontario

Michael De Santi, Ciprian Panfilie, Dale Barker, Usman T. Khan,
Satinder Brar, Stephanie Gora

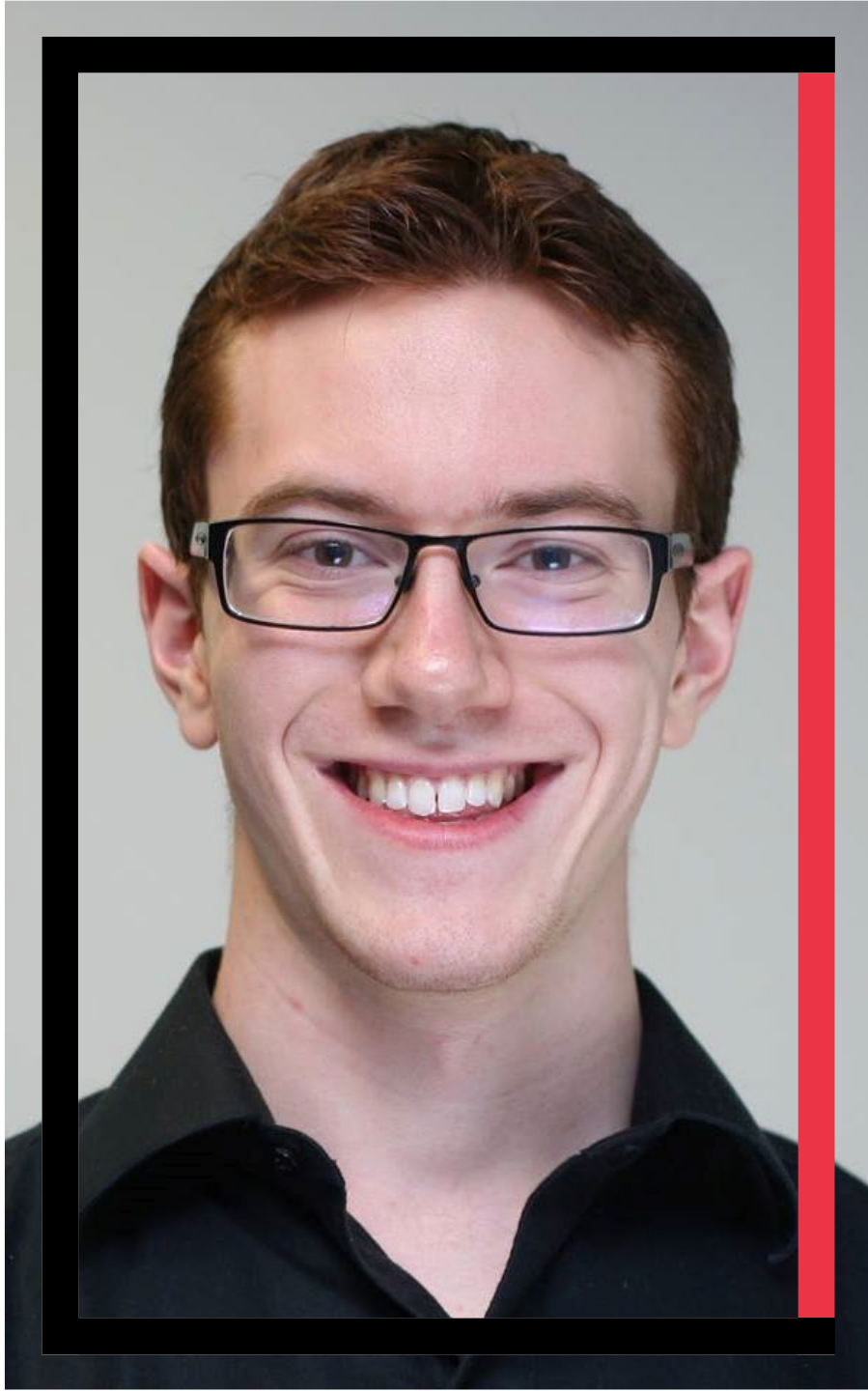
Collaboration

OCWA

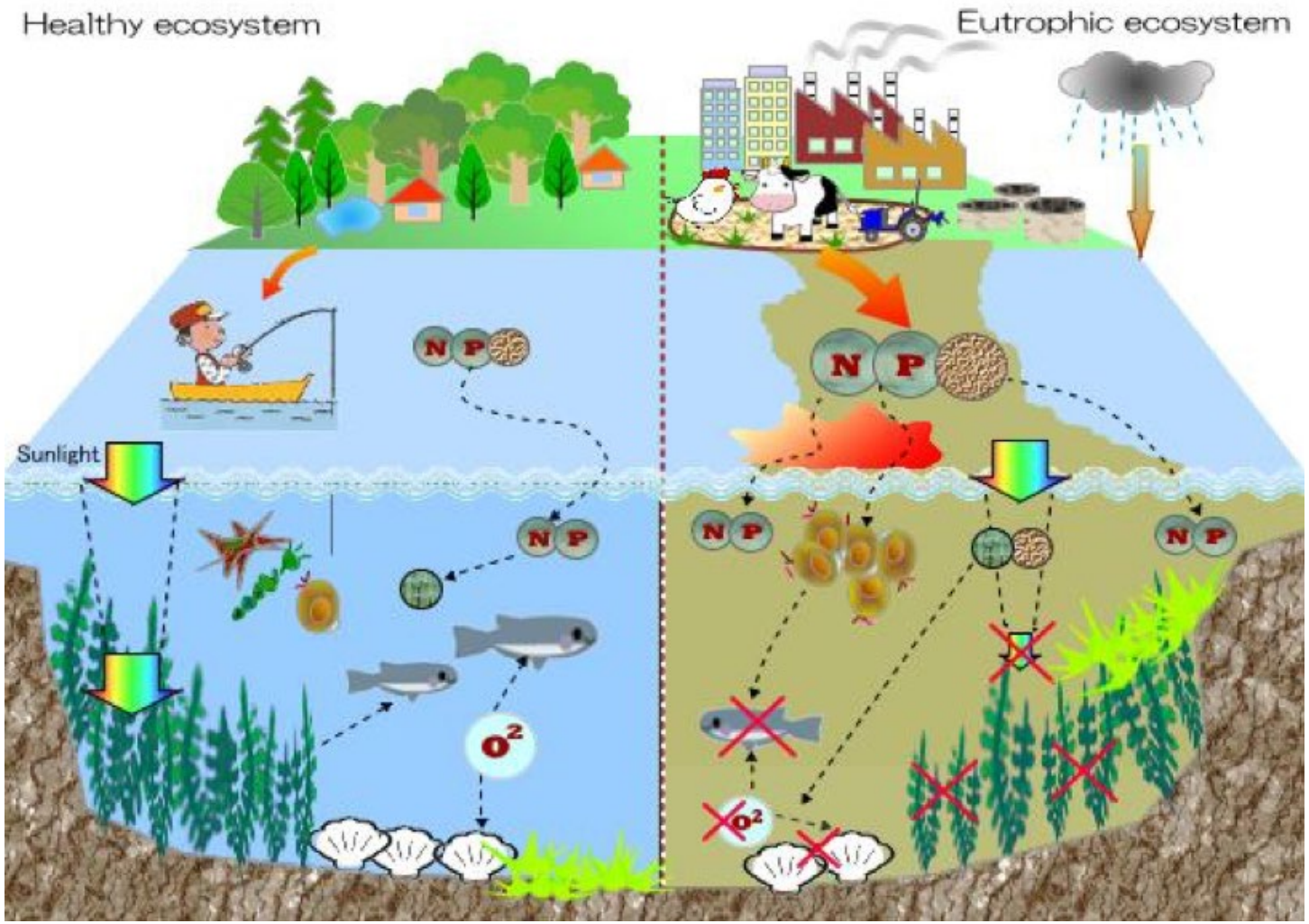
ONTARIO CLEAN WATER AGENCY
AGENCE ONTARIENNE DES EAUX

Mitacs

LASSONDE SCHOOL OF ENGINEERING | YORK U



Phosphorus in the environment



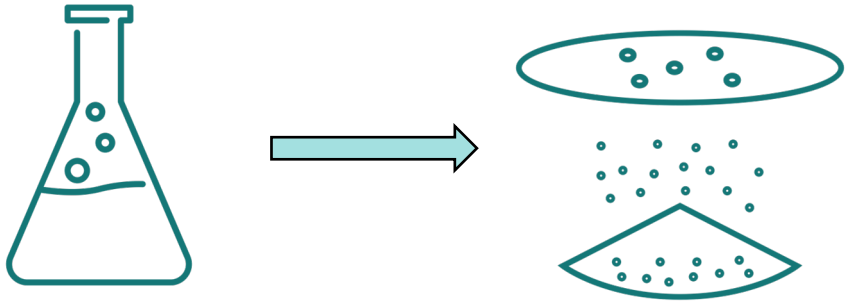
- Phosphorus in wastewater effluent has been linked to eutrophication and the development of cyanobacterial blooms
- Ontario provincial regulation: Limit total phosphorus discharge to a monthly average of 0.5 mg/L based on composite samples taken every two weeks
- Effluent objective at WWTP: 0.3 mg/L

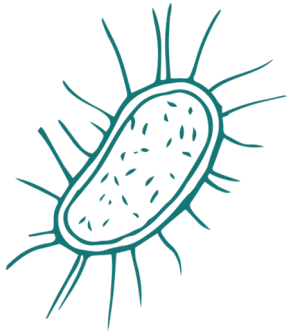
Image source: https://www.unep.org/nowpap/what-we-do/prevent-and-reduce-pollution/eutrophication

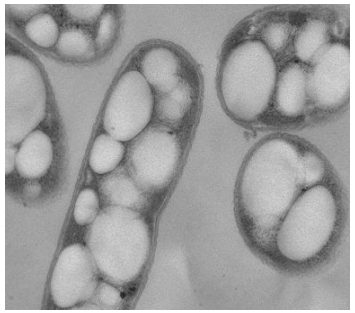
Phosphorus removal in wastewater treatment plants

Phosphorus is removed in wastewater treatment plants (WWTPs) via:

- Chemical coagulation followed by filtration
  - Lime, alum, ferric chloride
- Biological phosphorus removal
  - Enhanced biological phosphorus removal (EBPR) with phosphorus accumulating organisms (PAOs)
  - Incidental removal in secondary treatment process







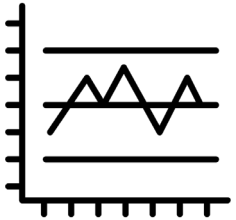
The WWTP in this study relies primarily on chemical coagulation with alum and pH adjustment at the headworks followed by filtration in the tertiary step for phosphorus removal

TEM image of PAOs used in EBPR : Yu, J., Porter, M., & Jaremko, M. (2013). Generation and utilization of microbial biomass hydrolysates in recovery and production of poly (3-hydroxybutyrate).Biomass Now: Cultivation and Utilization, InTech, 33-48.

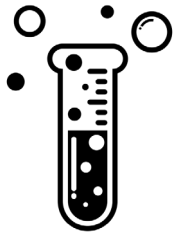
How can we leverage existing WWTP data sources to predict effluent TP?

Wastewater treatment plants are complex

Dynamic influent conditions

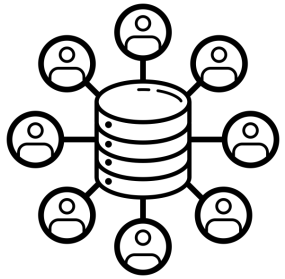


Complex unit processes



Uncertain process interrelation

Increased monitoring



COMPLEXITY

Research objectives

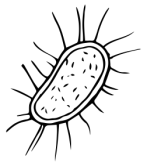
Objective 1: To identify the most useful process variables for predicting effluent total phosphorus (TP)

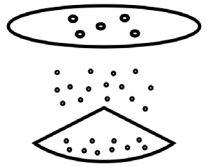
- Exploratory data analysis: Understand inputs and outputs and potential relationships
- Input variable selection: Formal process of establishing which parameters are important inputs to ML models

Objective 2: To develop and compare two approaches to modelling effluent TP using machine learning models:

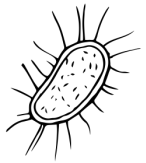
- Model 1: Predict effluent TP concentration directly
- Model 2: Predict the probability of exceeding the provincial effluent TP objective
- Model 3: Predict the probability of exceeding the provincial effluent TP limit

Methods: Plant and data description



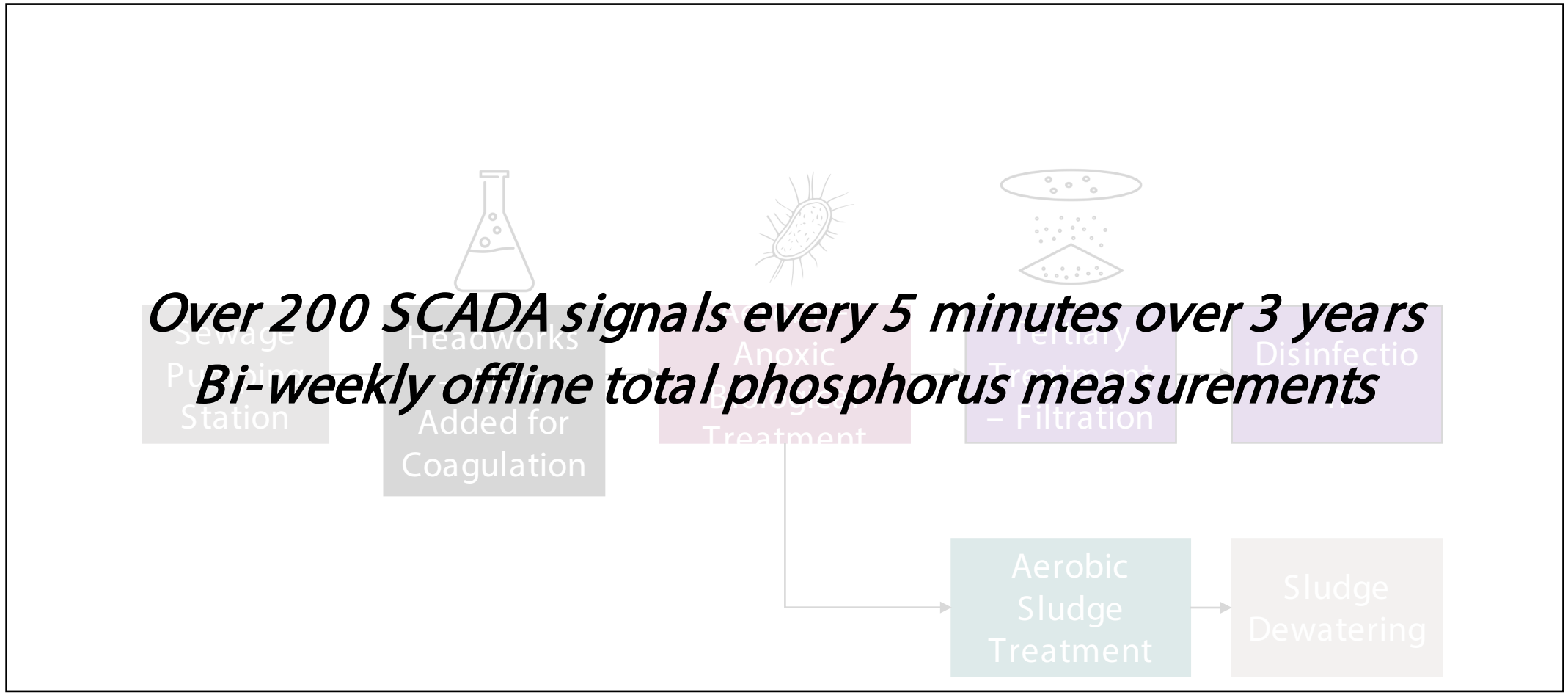


Sewage Pumping Station → Headworks Alum Added for Coagulation → Anoxic-Aerobic Biological Treatment → Filtration → Disinfection



Anoxic-Aerobic Biological Treatment → Aerobic Sludge Treatment → Sludge Dewatering

Methods: Plant and data description



Over 200 SCADA signals every 5 minutes over 3 years

Bi-weekly offline total phosphorus measurements

Modelling approach

Problem:

- What signals should we pay attention to?
- What lag period should be applied to different signals?
- Additional challenge – major discrepancy between number of SCADA measurements and number of TP measurements:
  - SCADA: once every 5 minutes = 105,120 measurements per year
  - TP: twice weekly measurements = 104 measurements per year
  - Ratio is 1000:1!

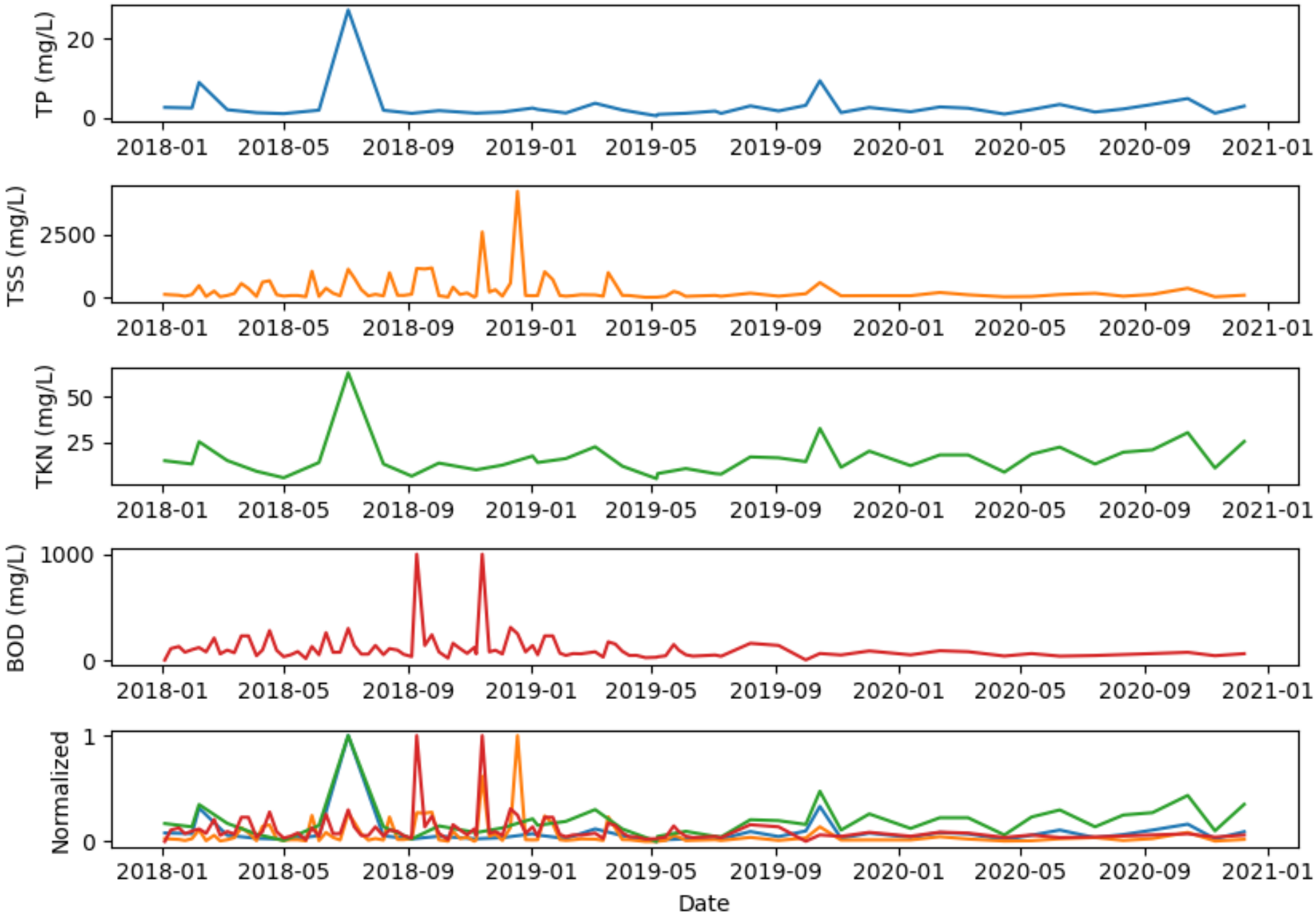
Modelling approach

Solution:

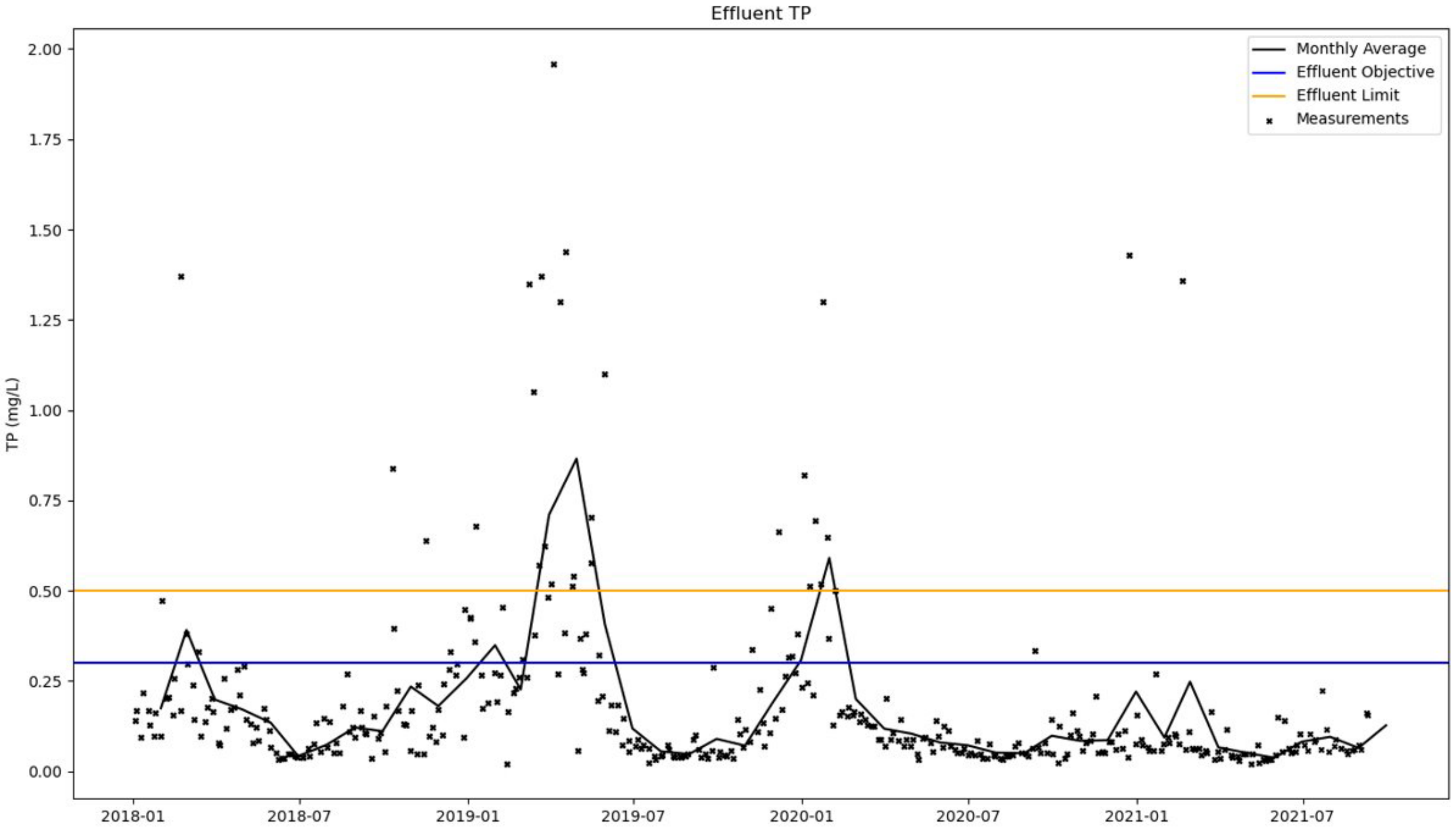
- Exploratory data analysis
- Prepare the data
  - Average SCADA inputs over the course of the day
- Build models
  - Regression
  - Classification
- Input variable selection (IVS)
  - Implement multi-method approach considering model-based and model-free methods with a final iterative backwards elimination approach to identify important input variables

Iterative

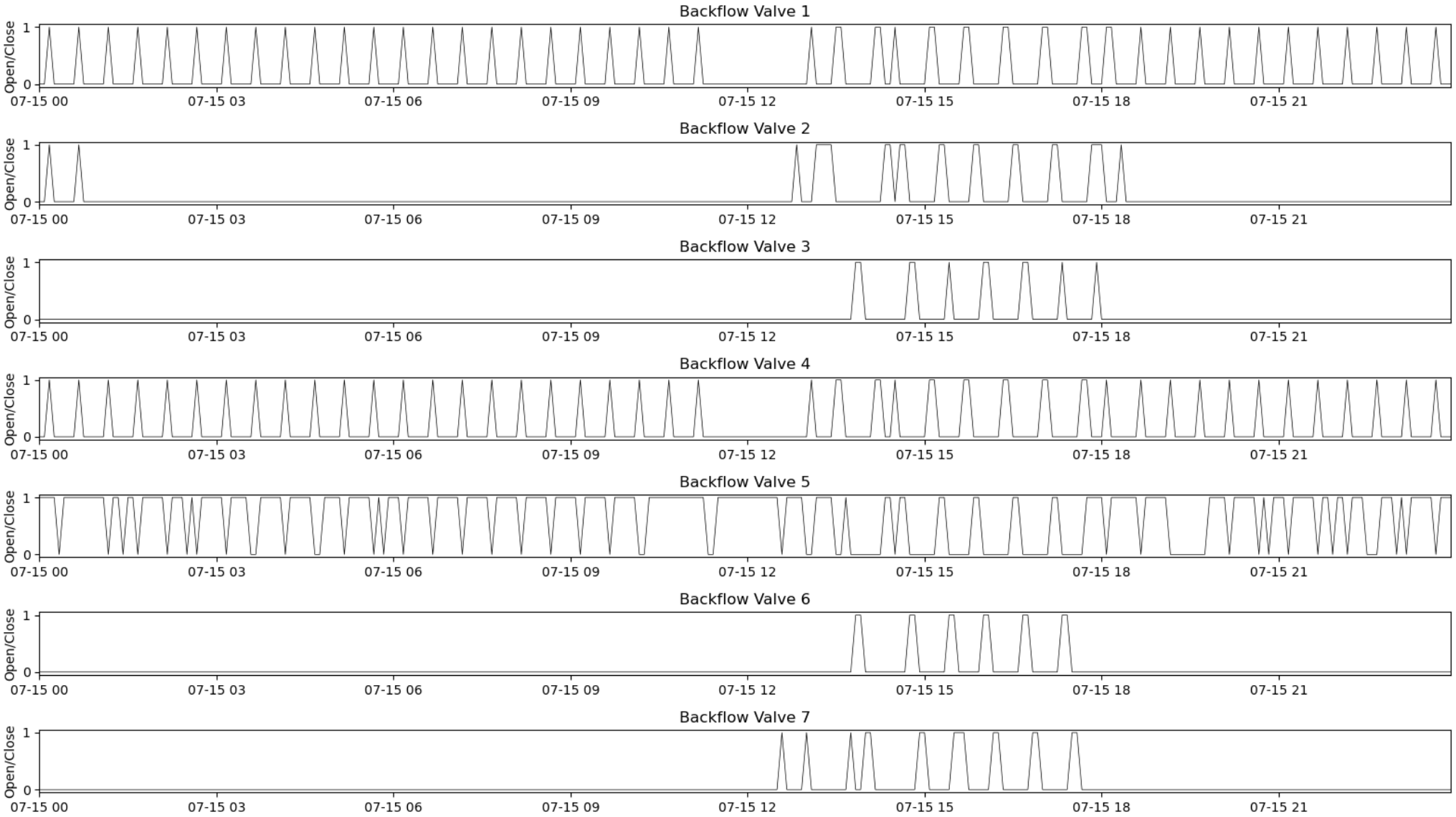
Preliminary data exploration: Influent water quality



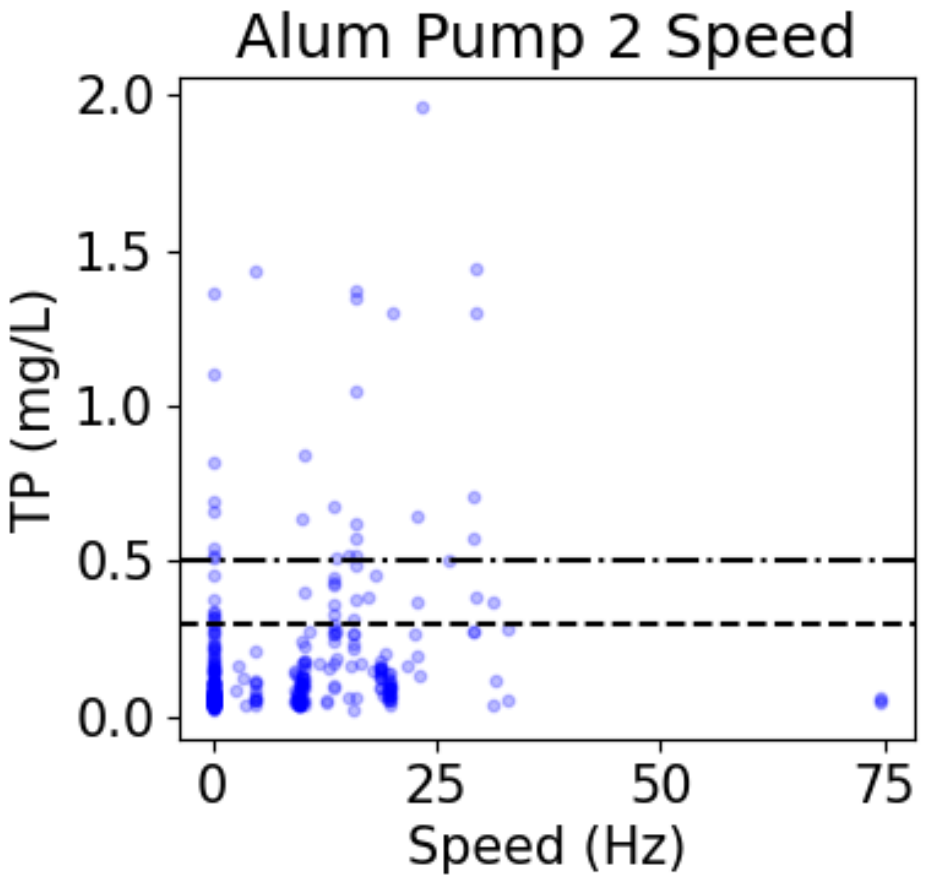
Preliminary data exploration: Effluent TP

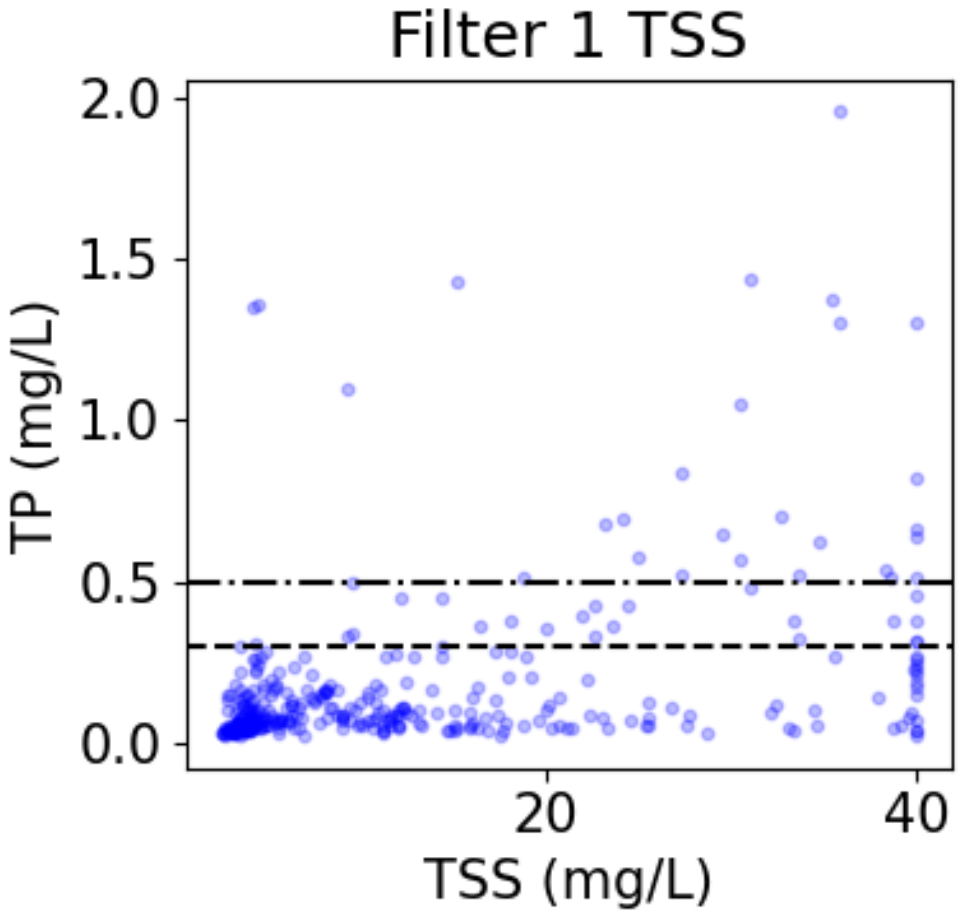


Exploratory data analysis: SCADA inputs



Preliminary data exploration: SCADA outputs vs. effluent TP





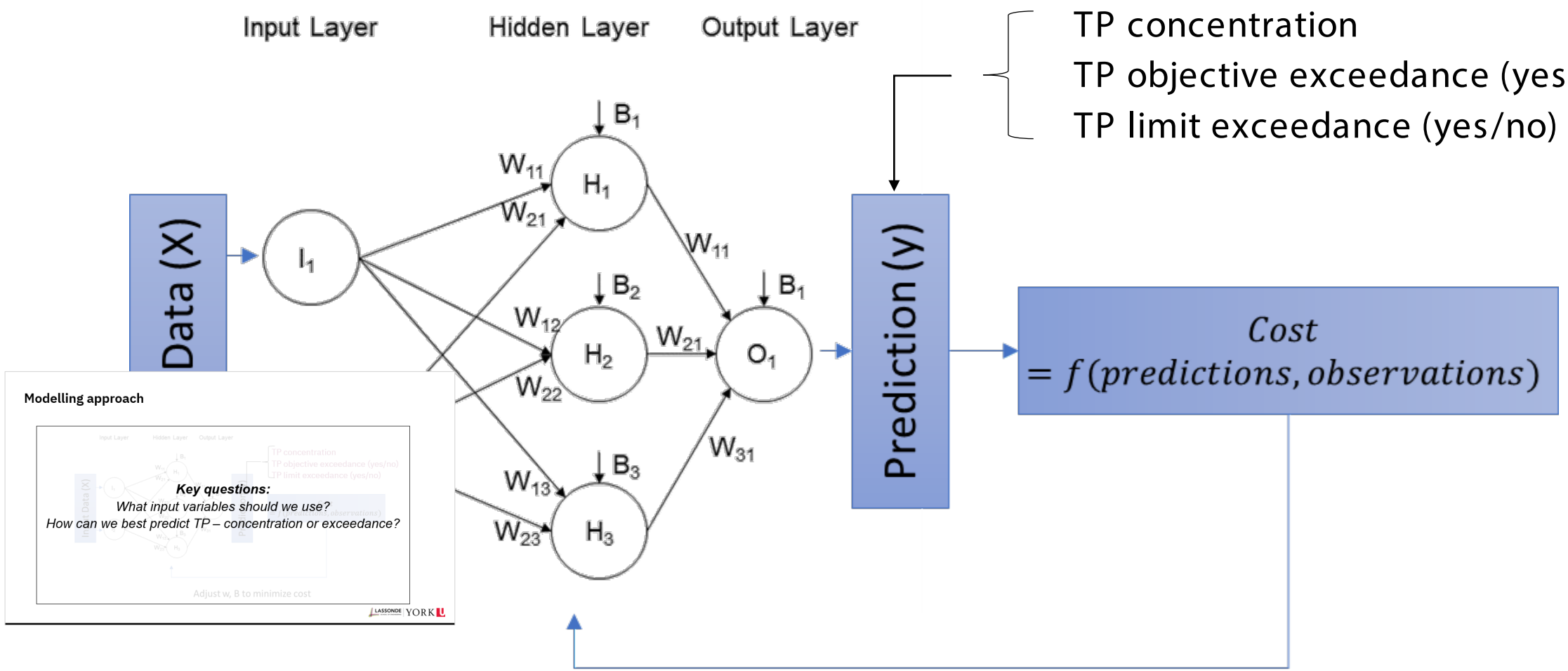
Solution: Machine learning

Machine learning:

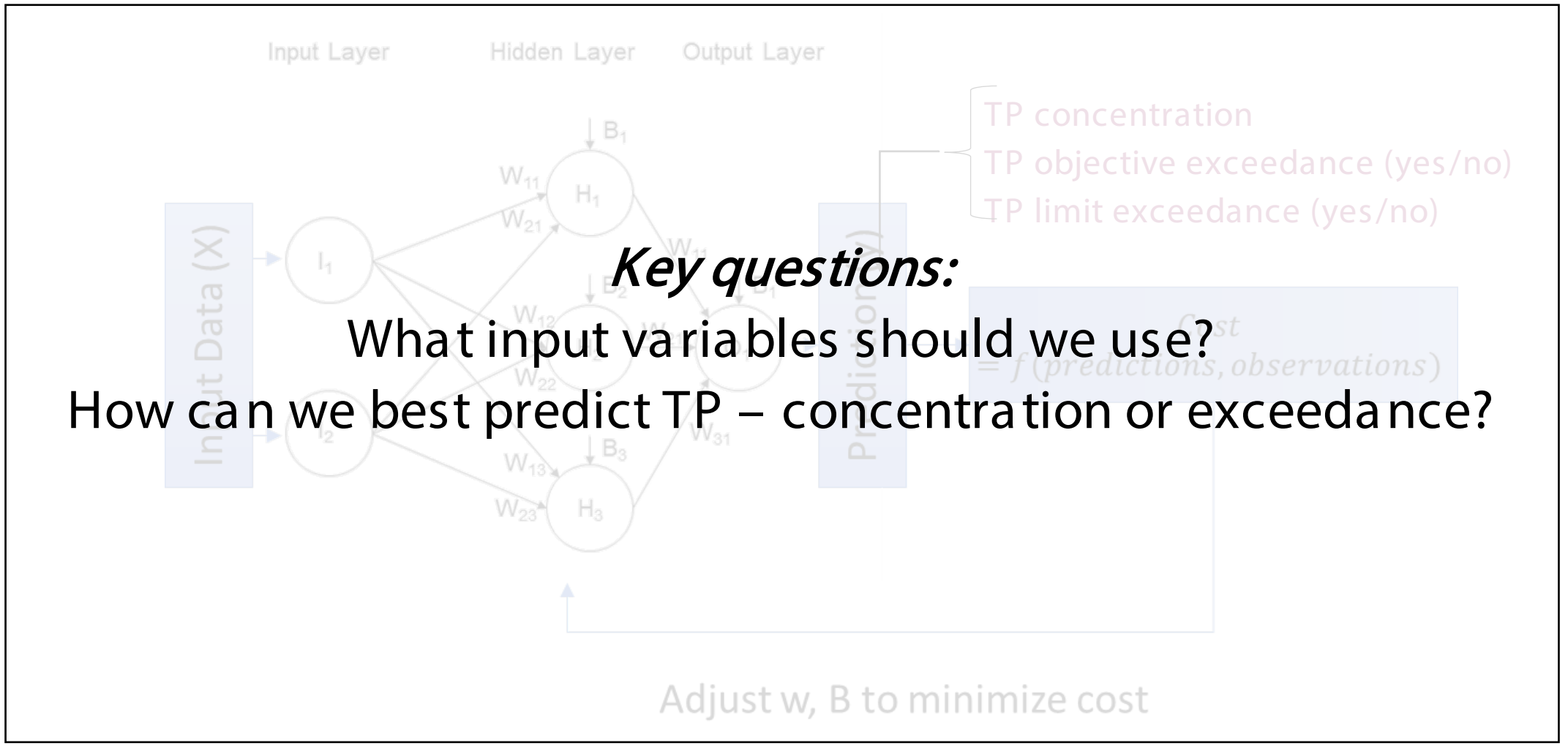
"The task of showing the inputs and outputs of a problem to an algorithm and letting it learn how to solve it"

- Serpa (2020) in Towards Data Science

Modelling approach



Modelling approach



Key questions:

What input variables should we use?

How can we best predict TP – concentration or exceedance?

Build initial models

Each model included multiple artificial neural network (ANN) multilayer perceptron (MLP) base learners that were combined to create ensemble ANNs

For each ensemble model, the response from each base learner is combined to create a confidence interval or probabilistic output

Model 1 — Predict TP concentration

- Regression model
- Cost function = mean squared error

Model 2 — Predict exceedance of TP objective

- Classification model
- Cost function = binary cross entropy

Model 3 — Predict exceedance of TP reg.

- Classification model
- Cost function = binary cross entropy

Evaluating model performance

Regression models: Root mean square error (RMSE) and R^2

$$\textbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\text{Predicted}_i - \text{Observed}_i)^2}{n}}$$

Classification models: Accuracy, Recall, Precision, Brier Score

$$\textbf{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n}$$

$$\textbf{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\textbf{Precision} = \frac{\text{True Positives} + \text{False Positives}}{\text{True Positives}}$$

$$\textbf{Brier score} = \sum_{i=1}^{n} \frac{\left(y_i^{obs} - p_i\right)^2}{n}$$

Confusion matrix

Predicted Condition	
Positive	Negative
True Positive	False Negative
False Positive	True Negative

Multi-method input variable selection (IVS)

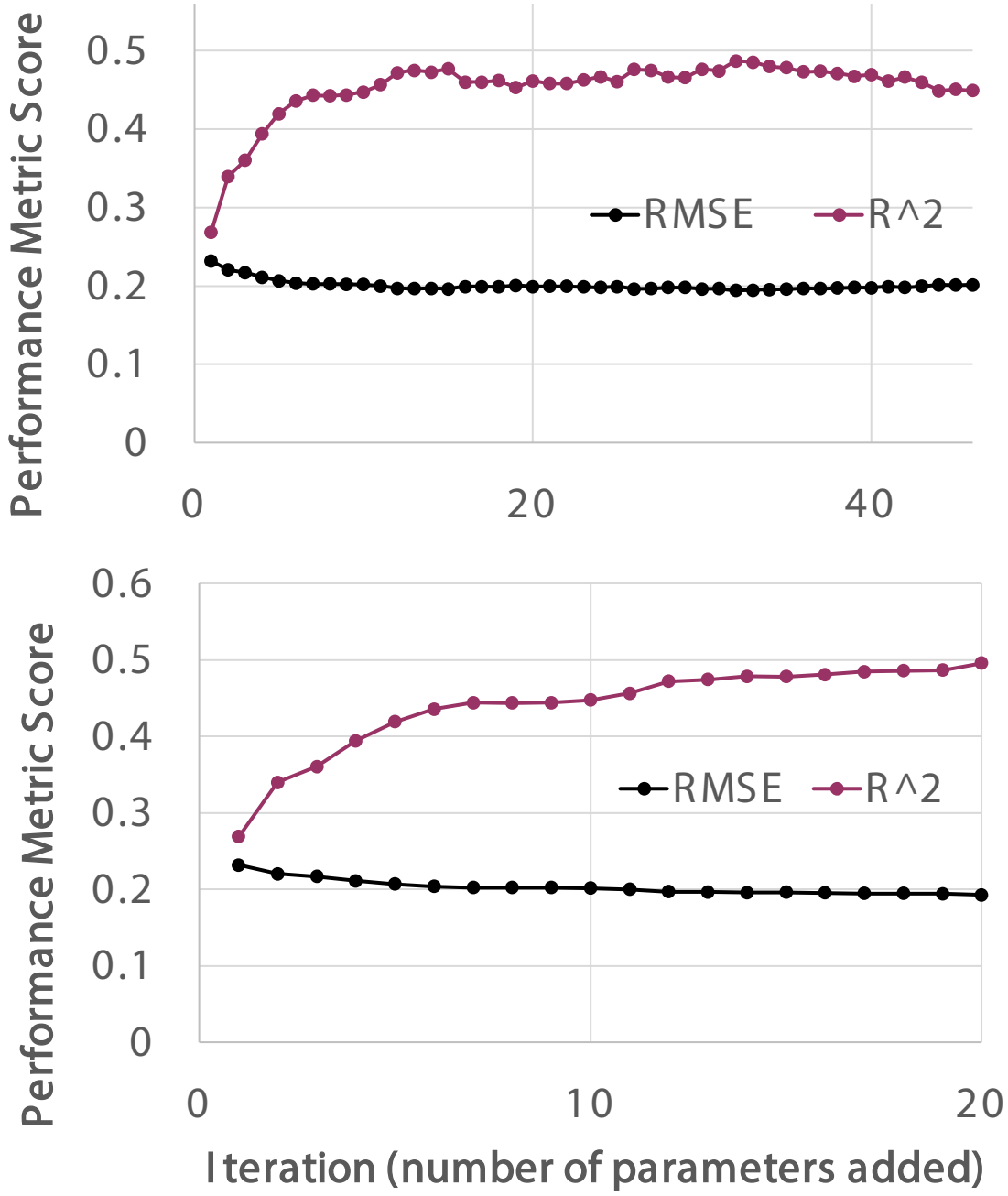
Step 1: Eliminate highly correlated variables

Step 2: Determine best lag period using non-linear correlation

Step 3: Rank candidate input variables

- **Non-linear correlation**
- **ANN:** CNPSA and IO
- **RF:** MDI and MDA

Step 4: Backwards elimination



Multi-method input variable selection (IVS)

Step 1: Eliminate highly correlated variables

↓

Step 2: Determine best lag period using non-linear correlation

↓

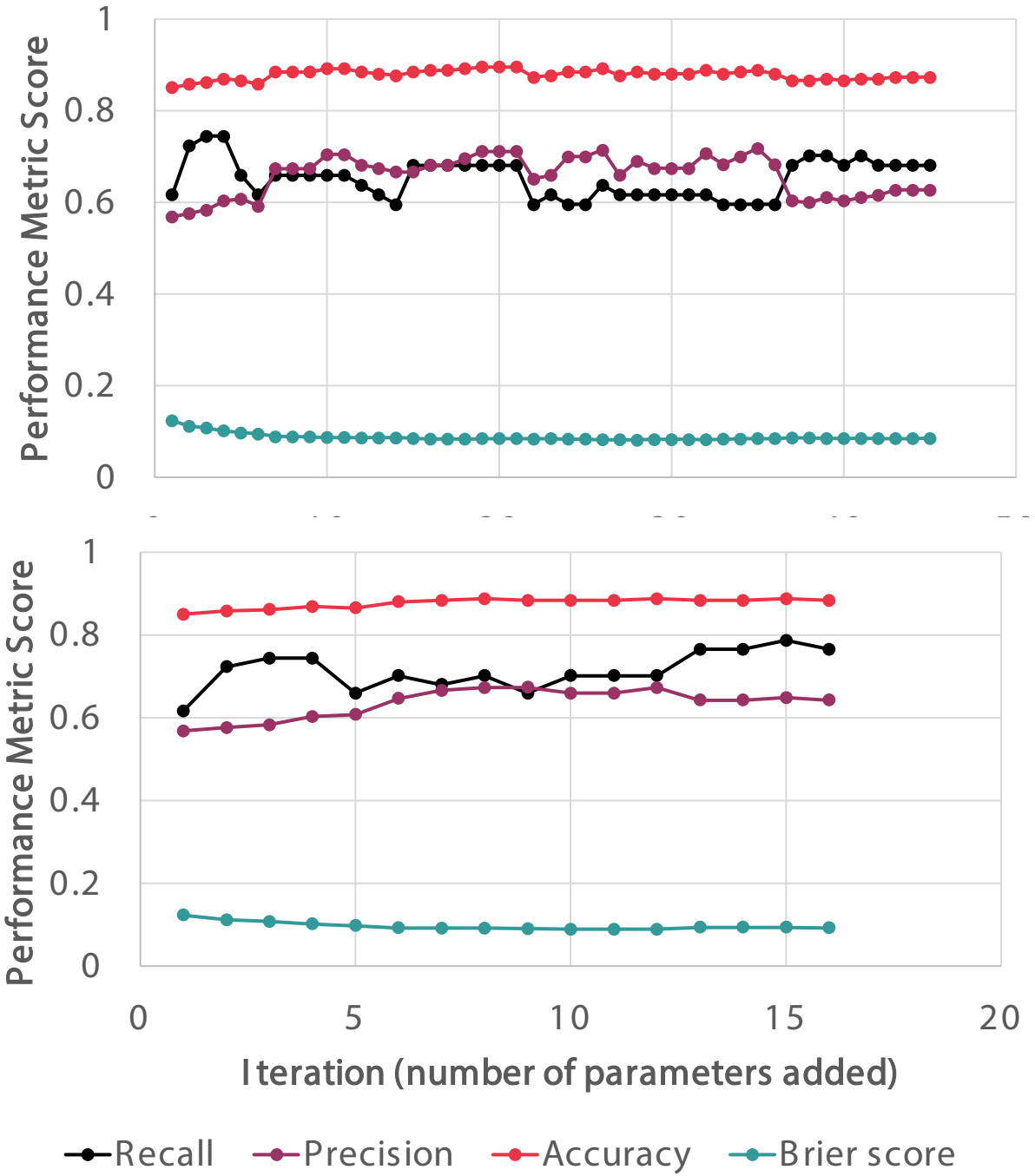
Step 3: Rank candidate input variables

↓

Non-linear correlation | **ANN:** CNPSA and IO | **RF:** MDI and MDA

↓

Step 4: Backwards elimination

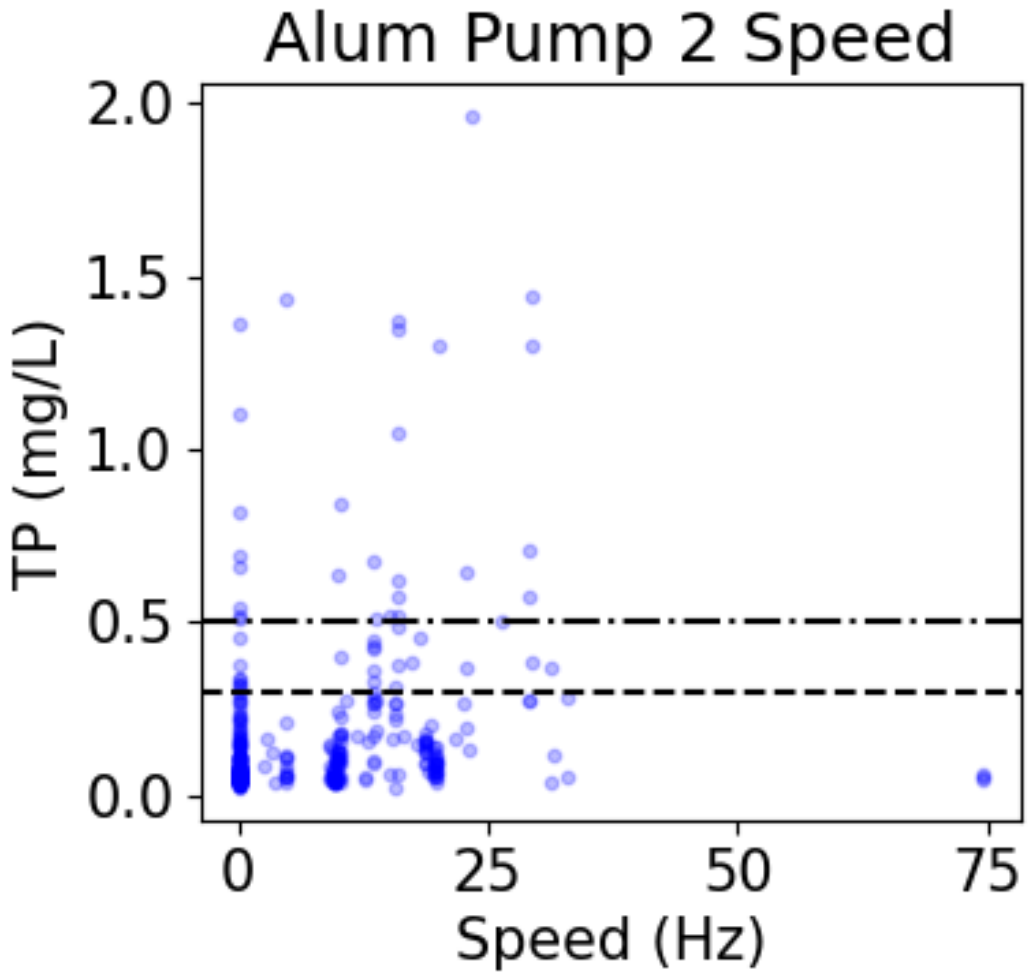


Results: Input variable selection

	SPS	Preliminary Treatment	Secondary Treatment	Tertiary Treatment and Disinfection	Sludge Digestion	Sludge Dewatering
Model 1:	Power Station Flow Rate	**Alum Pump 2 Speed** Headworks Compactor 1 Current	Blower 1 speed **Scum Chamber 1 Level** *RAS 1 Flow Rate* Aerobic Digester 1 Level Anoxic Mixer 1A Current	**Filter 1 TSS** *Filter 1 Level* **Filter 1 Waste Valve** *Filter 1 Waste Vacuum* *Filter 1 Backwash Valve* Filter 2 Waste Vacuum	*WAS 1 Flow Rate* **Sludge Room Methane**	Centrifuge Room H_2S Polymer Preparation Water Pressure Switch *Polymer Stock Tank Level* *Polymer Stock Tank Level*
Model 2:	*SPS Wet Well Level Switch*	**Alum Pump 2 Speed** *Headworks 1 Flow Rate*	**Scum Chamber 1 Level** *RAS 1 Flow*	**Filter 1 TSS** *Filter 1 Level* **Filter 1 Waste Valve** UV Bank 1A UV Intensity UV Bank 1B UV Intensity	**Sludge Room Methane**	Polymer Feed Pump 1 Speed *Polymer Containment Level Switch High* Centrifuge Pump 1 Speed
Model 3:	*SPS Wet Well Level Switch* SPS Wet Well Level	**Alum Pump 2 Speed** Caustic Containment Level *Headworks 1 Flow Rate* Alum Pump 1 Speed Grit Train 1A Blower Run Exhaust Fan 5 Current Exhaust Fan 6 Run	**Scum Chamber 1 Level** RAS Pump 2-1 Speed *Aerobic Digester 1 Level*	**Filter 1 TSS** Filter 2 Basin Level *Filter 1 Waste Vacuum* *Filter 1 Backwash Valve* **Filter 1 Waste Valve**	*WAS 1 Flow Rate* **Sludge Room Methane**	*Polymer Containment Level Switch High* Polymer Bag LSH

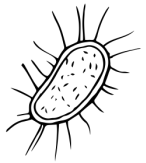
Important variables identified via IVS

Alum pump speed

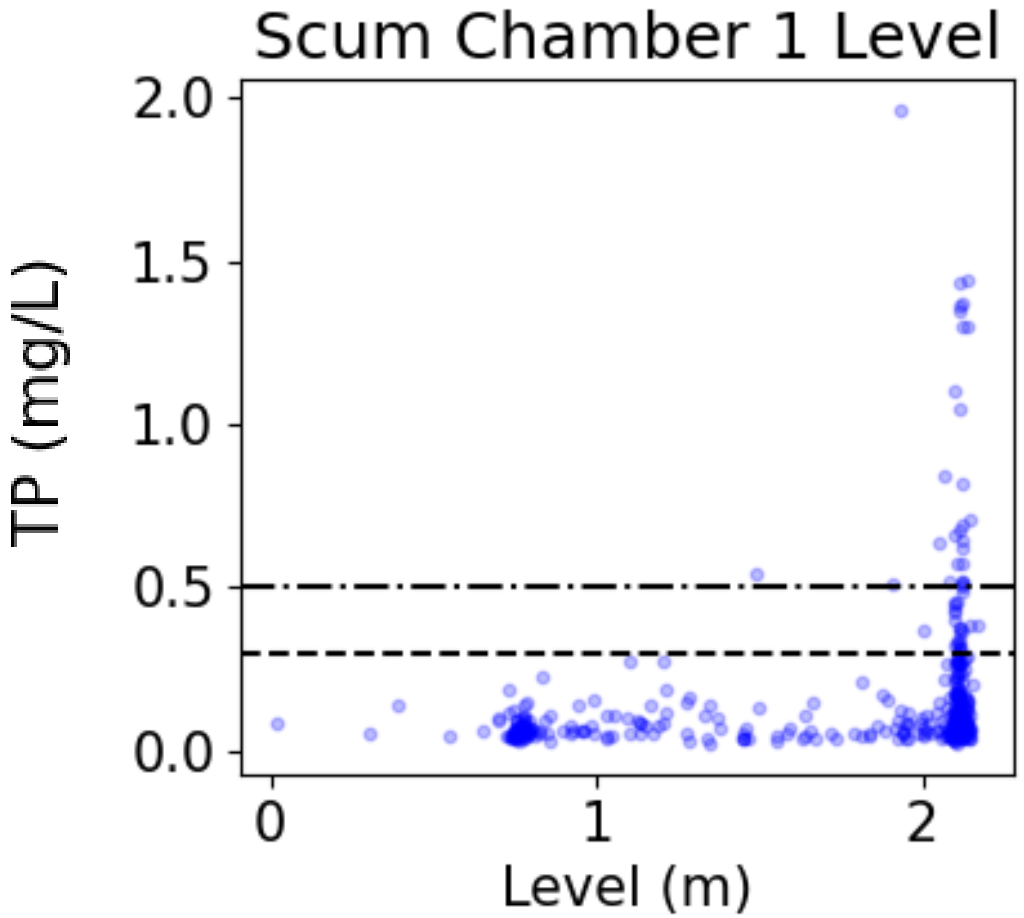


Important variables identified via IVS

Alum pump speed

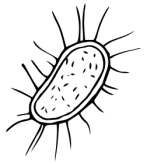


Scum chamber level

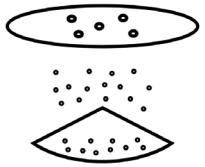


Important variables identified via IVS

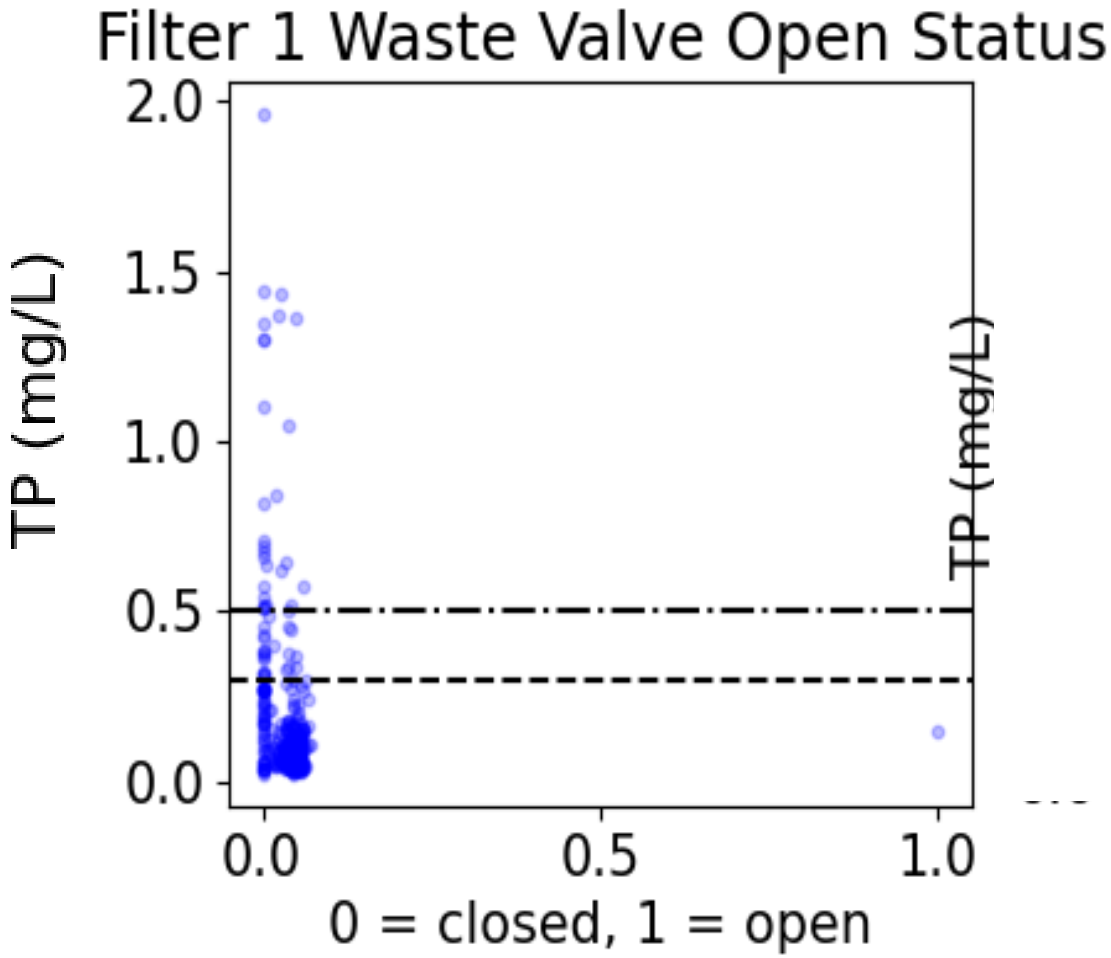
Alum pump speed



Scum chamber level

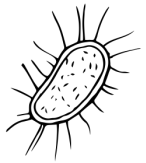


Filter waste valve status

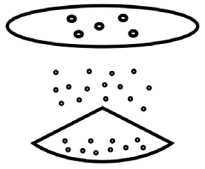


Important variables identified via IVS

Alum pump speed

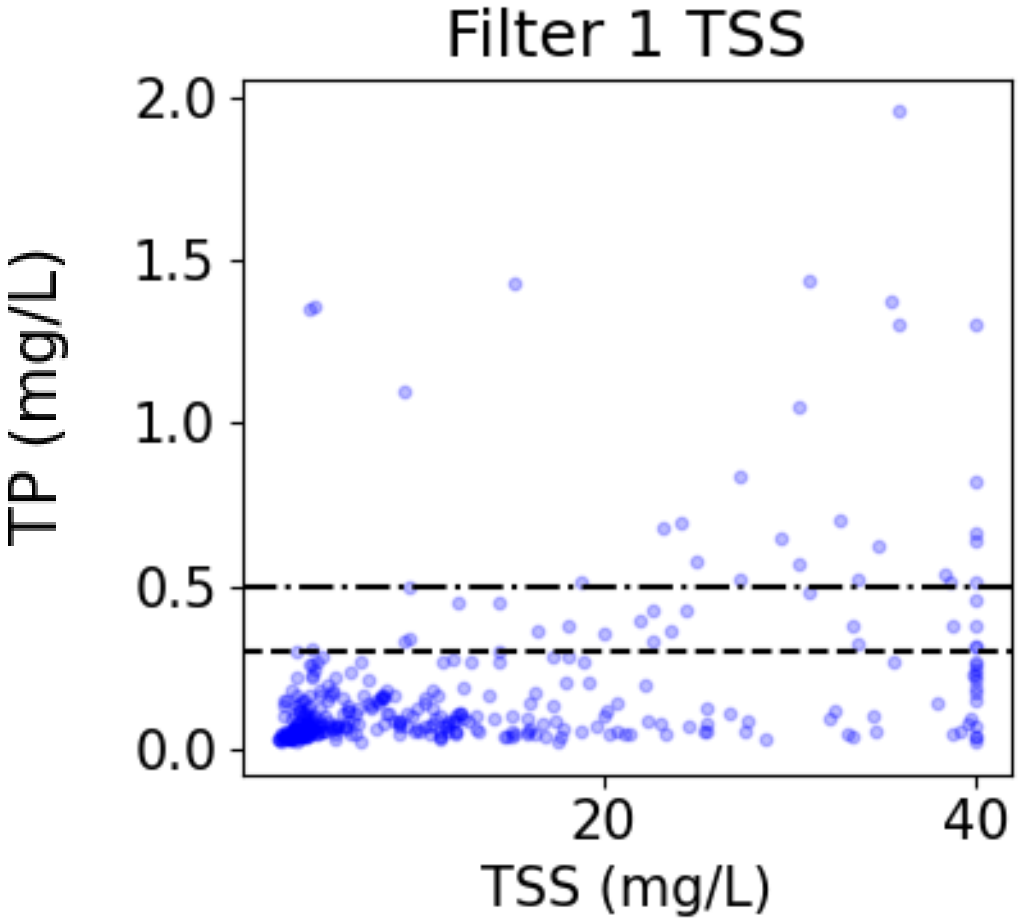


Scum chamber level



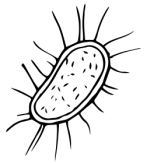
Filter waste valve status

Filter TSS

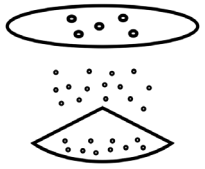


Important variables identified via IVS

Alum pump speed



Scum chamber level

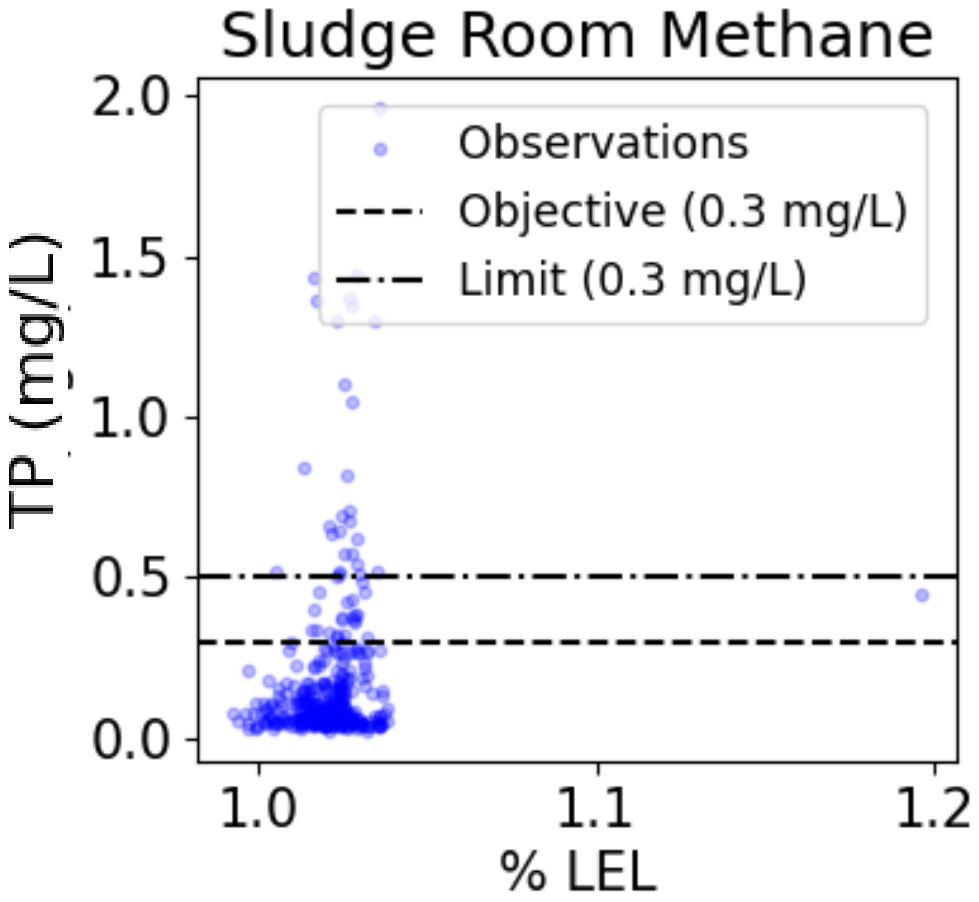


Filter waste valve status

Filter TSS

?

Sludge room methane



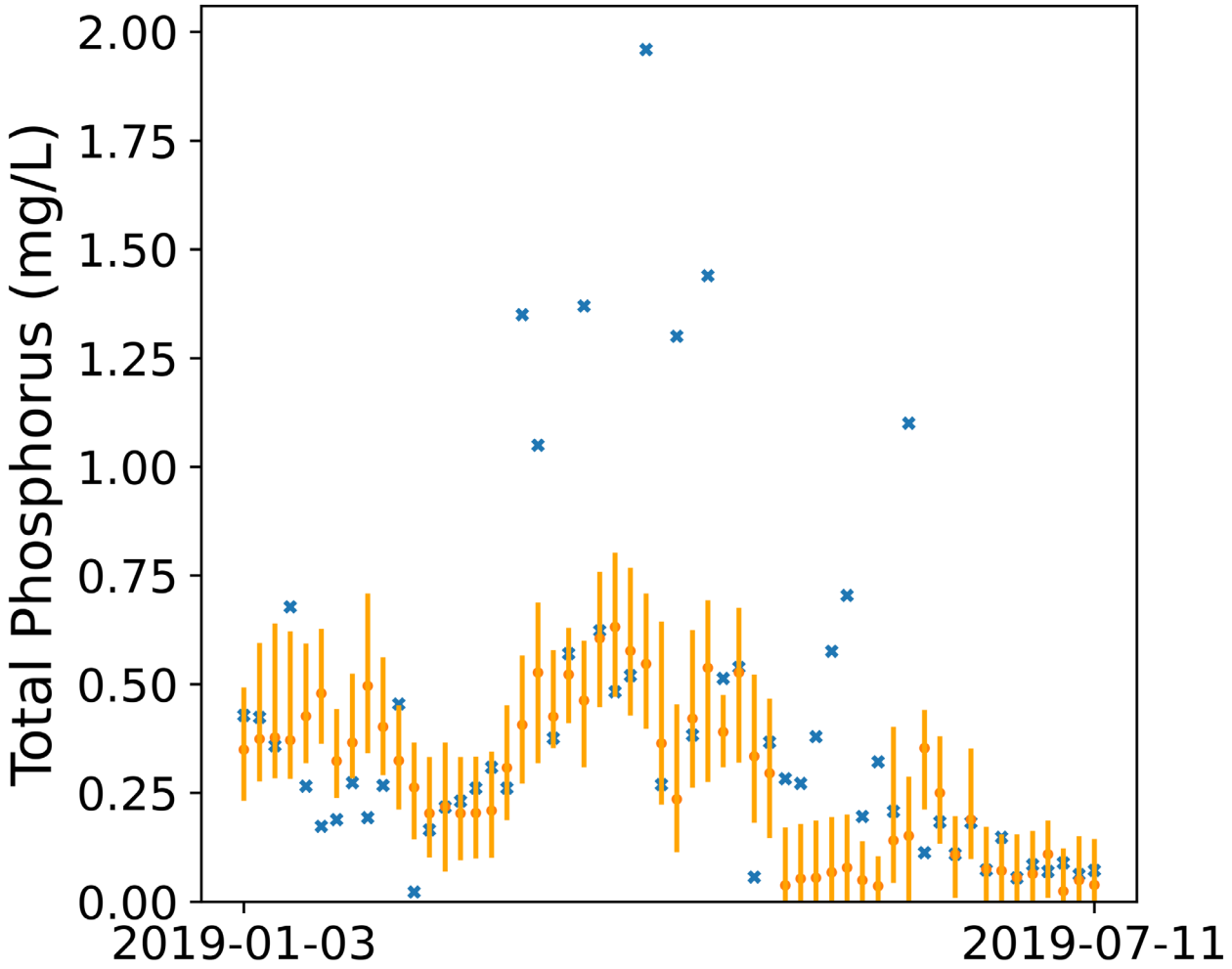
Results: Input variable selection

	SPS	Preliminary Treatment	Secondary Treatment	Tertiary Treatment and Disinfection	Sludge Digestion	Sludge Dewatering
Model 1:	Power Station Flow Rate	**Alum Pump 2 Speed** Headworks Compactor 1 Current	Blower 1 speed **Scum Chamber 1 Level** *RAS 1 Flow Rate* Aerobic Digester 1 Level Anoxic Mixer 1A Current	**Filter 1 TSS** *Filter 1 Level* **Filter 1 Waste Valve** *Filter 1 Waste Vacuum* *Filter 1 Backwash Valve* Filter 2 Waste Vacuum	*WAS 1 Flow Rate* **Sludge Room Methane**	Centrifuge Room H_2S Polymer Preparation Water Pressure Switch *Polymer Stock Tank Level* *Polymer Stock Tank Level*
Model 2:	*SPS Wet Well Level Switch*	**Alum Pump 2 Speed** *Headworks 1 Flow Rate*	**Scum Chamber 1 Level** *RAS 1 Flow*	**Filter 1 TSS** *Filter 1 Level* **Filter 1 Waste Valve** UV Bank 1A UV Intensity UV Bank 1B UV Intensity	**Sludge Room Methane**	Polymer Feed Pump 1 Speed *Polymer Containment Level Switch High* Centrifuge Pump 1 Speed
Model 3:	*SPS Wet Well Level Switch* SPS Wet Well Level	**Alum Pump 2 Speed** Caustic Containment Level *Headworks 1 Flow Rate* Alum Pump 1 Speed Grit Train 1A Blower Run Exhaust Fan 5 Current Exhaust Fan 6 Run	**Scum Chamber 1 Level** RAS Pump 2-1 Speed *Aerobic Digester 1 Level*	**Filter 1 TSS** Filter 2 Basin Level *Filter 1 Waste Vacuum* *Filter 1 Backwash Valve* **Filter 1 Waste Valve**	*WAS 1 Flow Rate* **Sludge Room Methane**	*Polymer Containment Level Switch High* Polymer Bag LSH

Training the model

Experiment	Input data period	Number of samples	Type of training dataset
1	All data from previous calendar year	~100	Static
2	365 days of data preceding the first day of the week being forecasted	103-104	Moving window
3	14 days of data preceding week being forecasted	4-6 samples	Moving window
4	28 days of data preceding the first day of the week being forecasted	10-12 samples	Moving window
5	All data starting from the first week of the preceding year	>100-200	Increasing window
6	All data starting from the last week of the preceding year	>2-100	Increasing window

Model 1: TP concentration modelling (regression model)



Model 1 applied to test data set

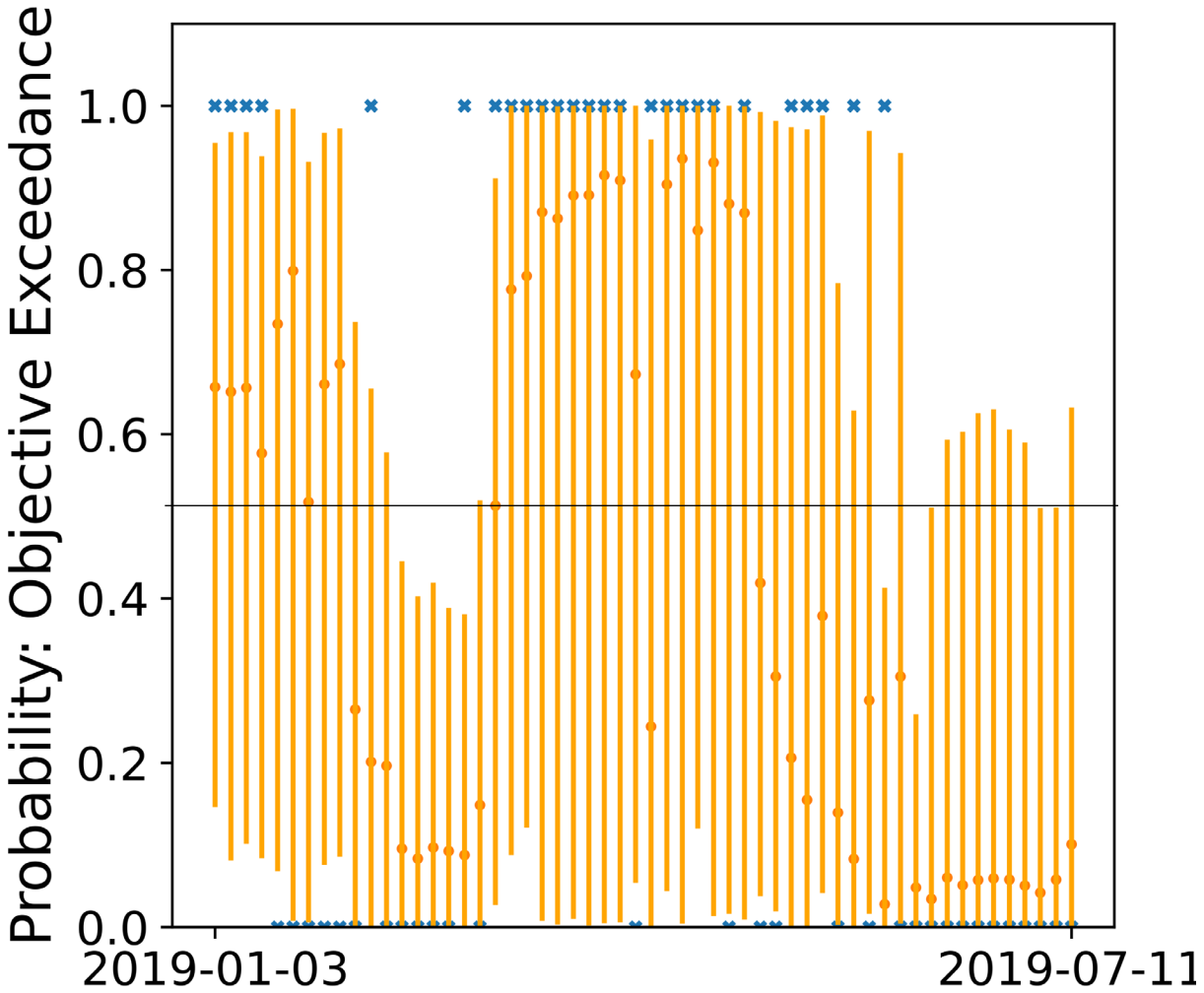
Blue dots: Observed TP values

Yellow bands: Confidence interval for the mean predicted effluent TP concentration

RMSE: 0.299 mg/L

Performance: Mean error was about 0.3 mg/L +/- the predicted value – this is pretty good!

Model 2: TP objective exceedance modelling (classification model)



Model 2 applied to test data set

Blue dots: Observed exceedance/non-exceedance

Yellow bands: Probabilistic predictions of exceedance

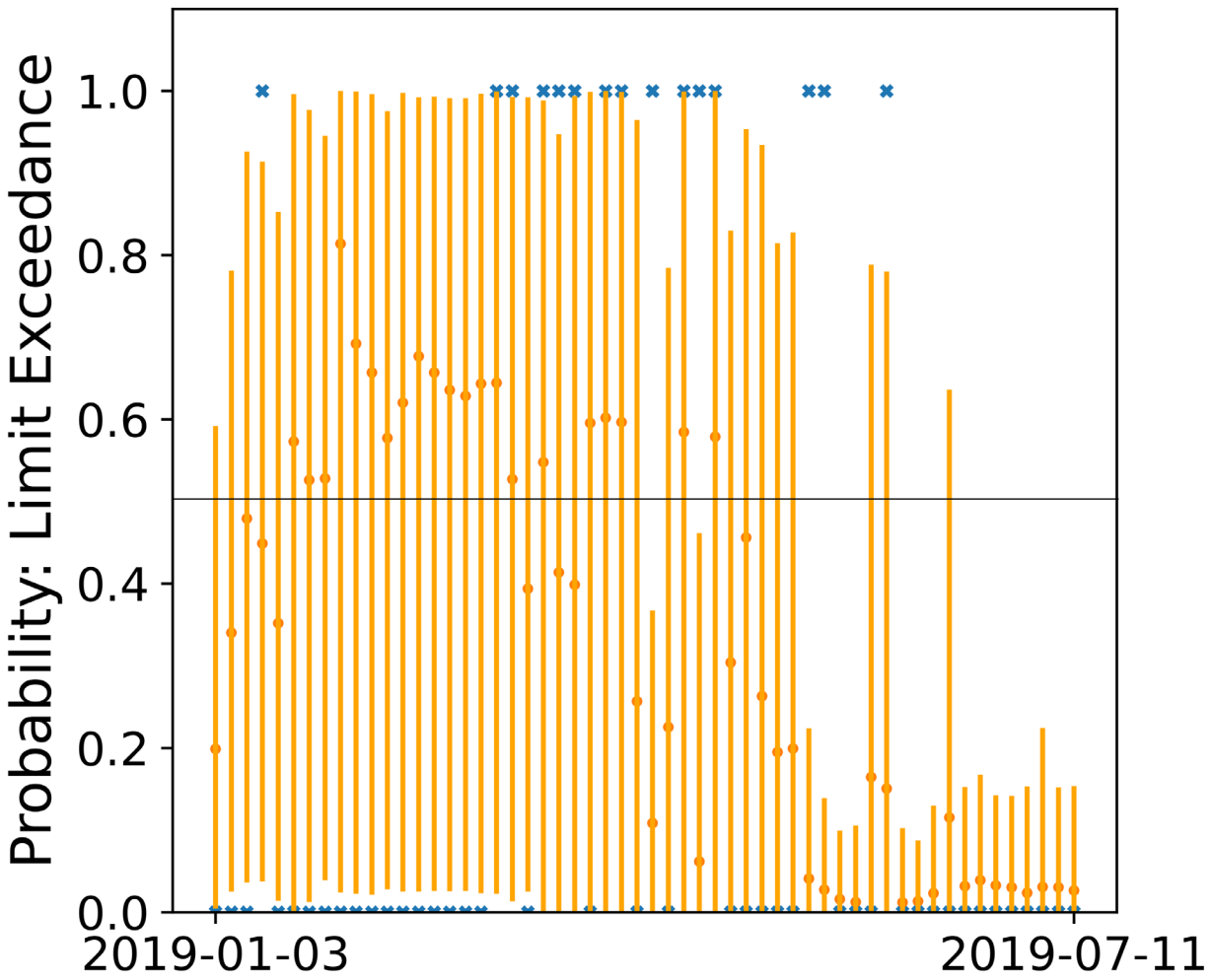
Accuracy: 71%

Recall: 73%

Precision: 68%

Performance: Correctly predicted exceedance vs. non-exceedance 71% of the time and correctly identified exceedances 73% of the time

Model 3: TP limit exceedance modelling (classification model)



Model 3 applied to test data set

Blue dots: Observed exceedance/non-exceedance

Yellow bands: Probabilistic predictions of exceedance

Accuracy: 71%

Recall: 53%

Precision: 47%

Performance: Correctly predicted exceedance vs. non-exceedance 71% of the time and correctly identified exceedances 53% of the time

Summary of model results

Model	Type	Performance	Notes
Model 1: TP concentration	Regression	RMSE: 0.299 mg/L	Mean error was about 0.3 mg/L +/- the predicted value – this is pretty good!
Model 2: Exceedance of TP objective	Classification	Accuracy: 71% Recall: 73% Precision: 68%	Correctly predicted exceedance vs. non-exceedance 71% of the time and identified exceedances 73% of the time
Model 3: Exceedance of TP regulatory limit	Classification	Accuracy: 71% Recall: 53% Precision: 47%	Correctly predicted exceedance vs. non-exceedance 71% of the time and identified exceedances 53% of the time

Takeaways

1. ML can accurately forecast wastewater effluent quality using routinely collected data
2. ML can select unintuitive but useful relationships between process variables and target outputs
3. The highly correlated and non-linear relationships between process variables require advanced IVS

Next steps

Short term:

1. Continue engagement with OCWA process engineers, data specialists, and operators
2. Deploy models for live testing on OCWA SCADA systems
3. Expand modelling to include other effluent quality parameters
4. New models to simulate and optimize individual unit processes

Long term: YorkU and OCWA are building a long term collaboration to explore the application of these methods to:

1. Optimizing wastewater plants for energy savings and resource recovery
2. Water safety monitoring in water treatment plants
3. Drinking water distribution system monitoring

Acknowledgements

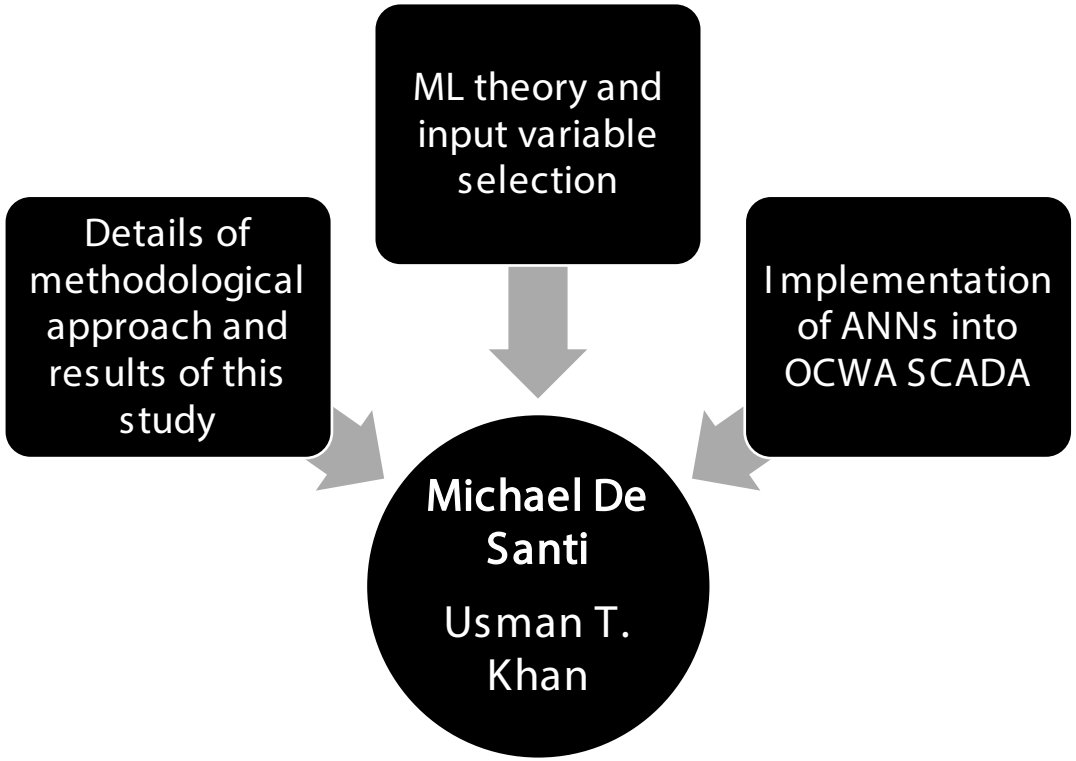
OCWA
ONTARIO CLEAN WATER AGENCY
AGENCE ONTARIENNE DES EAUX

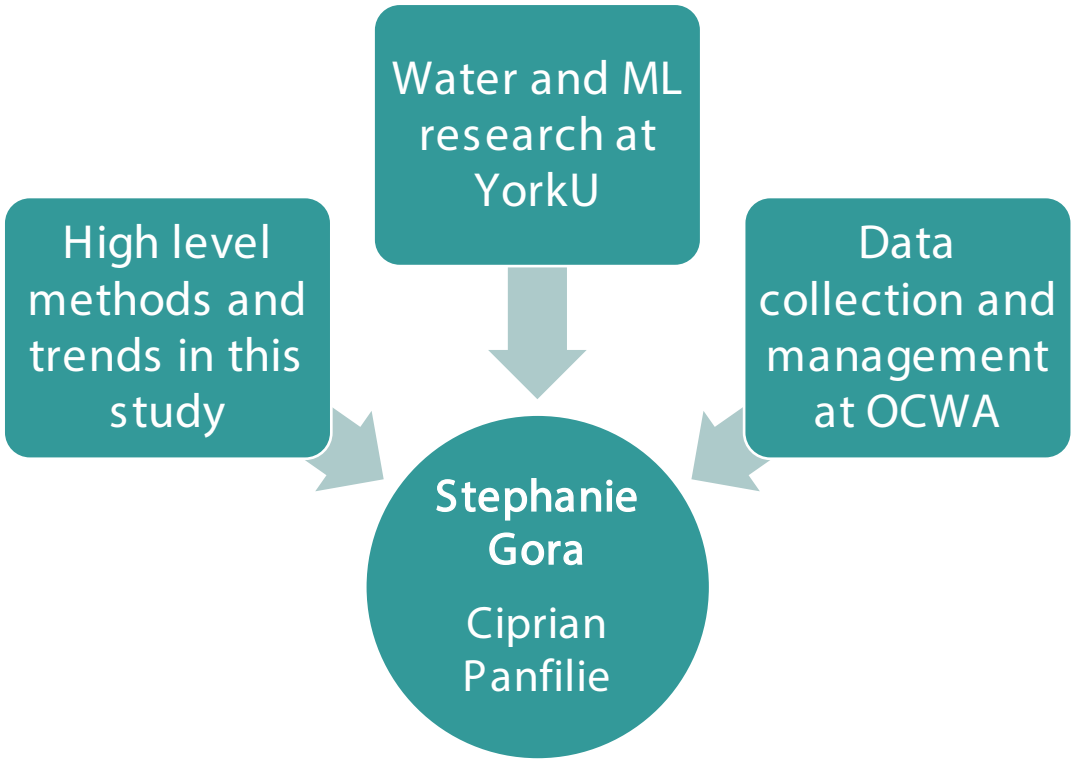
Mitacs

LASSONDE
SCHOOL OF ENGINEERING

YORK U

Questions? Comments?



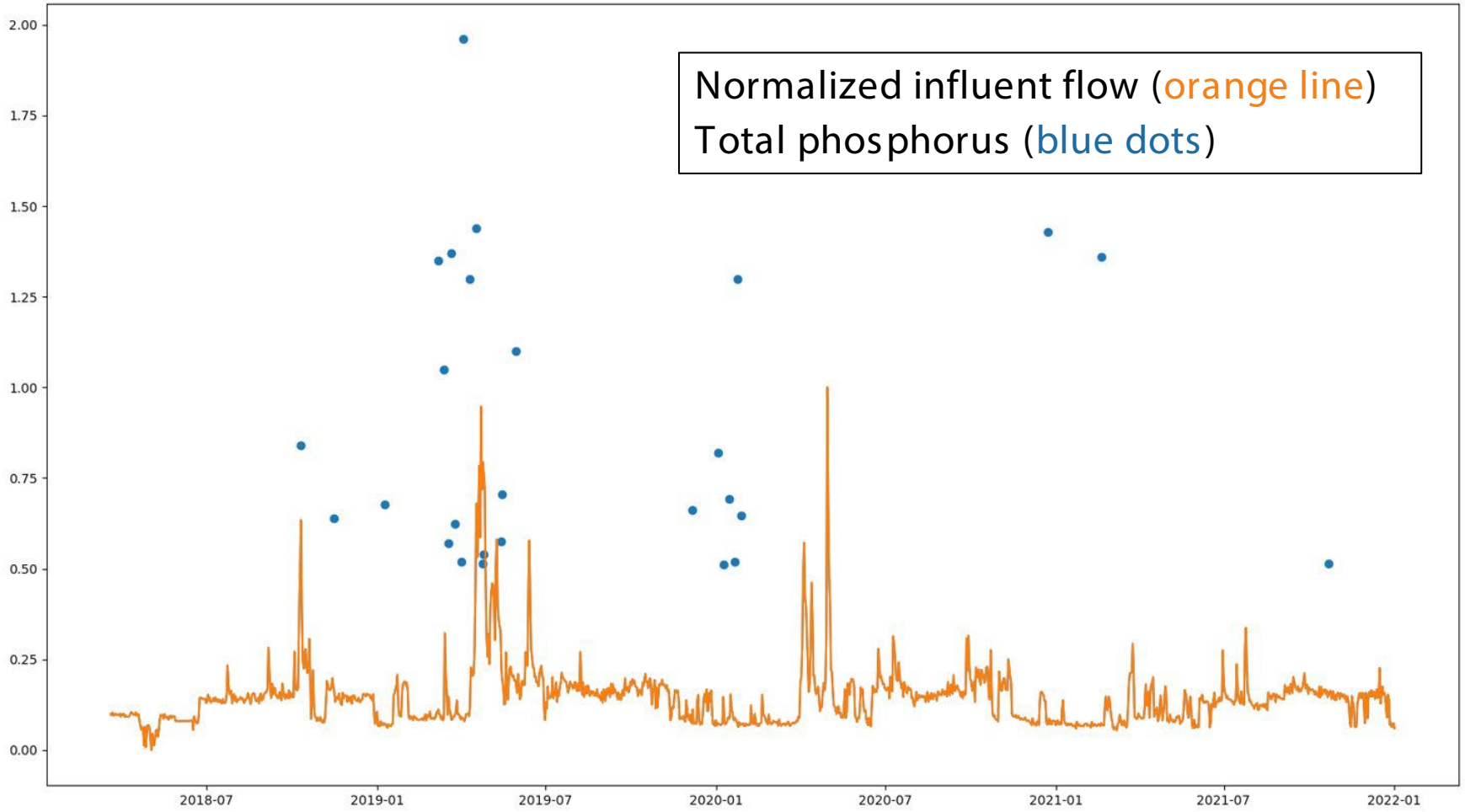


Email: desantim@yorku.ca
Twitter: @michaelvdesanti
LinkedIn: www.linkedin.com/in/michaeldesanti/

Email: stephanie.gora@lassonde.yorku.ca
Twitter: @goraradio
LinkedIn: www.linkedin.com/in/stephaniegora/

LASSONDE SCHOOL OF ENGINEERING | YORK U

Climate effects



Including the climate variables (temperature, precipitation, barometric pressure) did not improve performance

Bonus slides

Multi-method input variable selection (IVS)

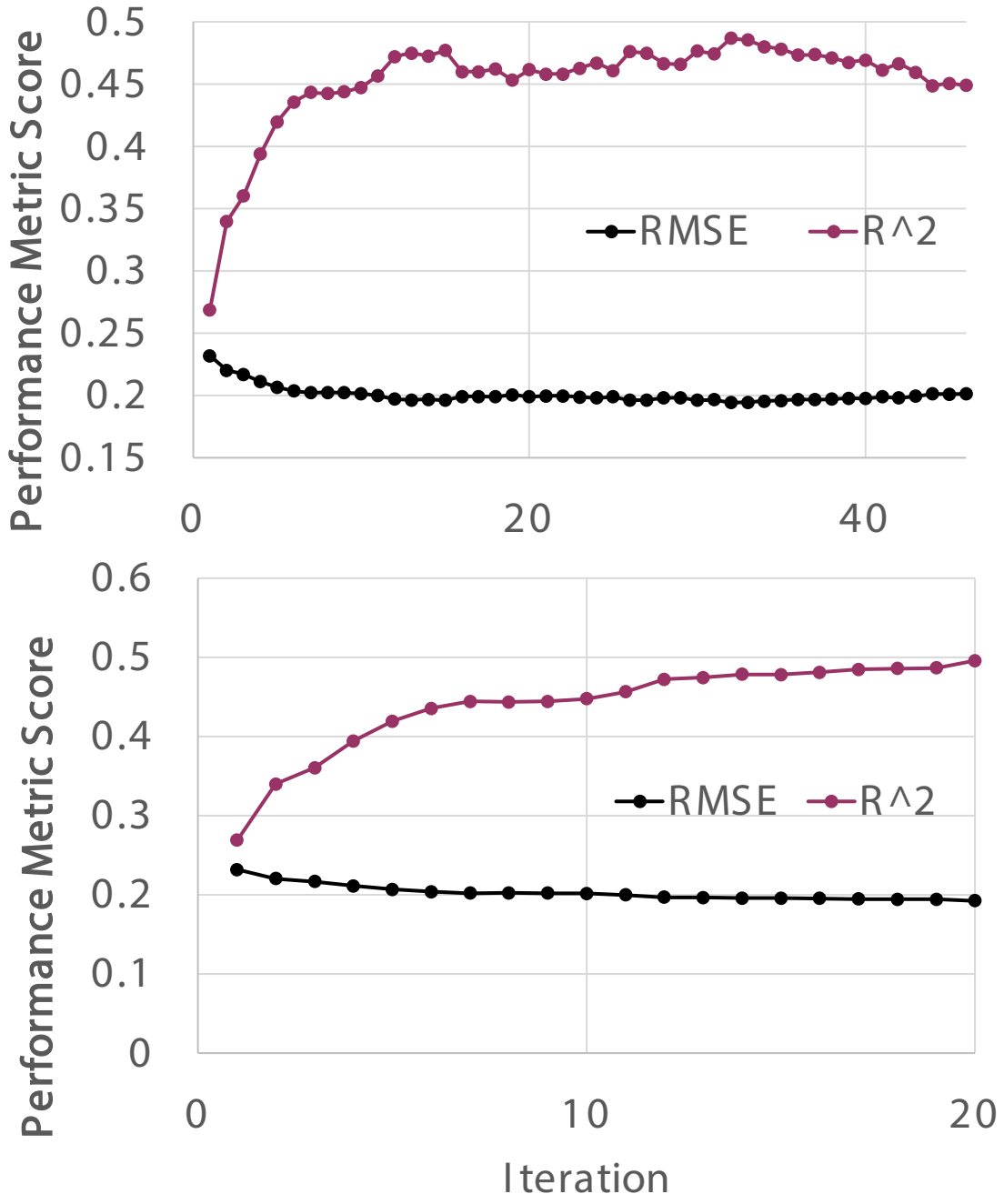
Step 1: Eliminate highly correlated variables

Step 2: Determine best lag period using non-linear correlation

Step 3: Rank candidate input variables

- **Non-linear correlation**
- **ANN:** CNPSA and IO
- **RF:** MDI and MDA

Step 4: Backwards elimination



Step 1

- Correlation between variables assessed with Pearson, Spearman, and Kendall
- If two variables had correlation(s) greater than 0.5 one was removed
  - When selecting which of the correlated pair of variables to retain or eliminate, we sought to retain candidate variables that were correlated with the most other variables as this ensured that the largest number of correlated variables possible were eliminated.

Step 2

- Correlation between variables assessed with Pearson, Spearman, and Kendall
- If two variables had correlation(s) greater than 0.5 one was removed
  - When selecting which of the correlated pair of variables to retain or eliminate, we sought to retain candidate variables that were correlated with the most other variables as this ensured that the largest number of correlated variables possible were eliminated.

Step 3

Six different IVS methods to rank each candidate variable's effect on the output variable. A multi-method approach allows us to assess patterns in the results across multiple methods, identifying which variables are consistently identified as useful input variables and which are not.

There are two main types of IVS methods: model-free (statistical measures) and model-based methods. Model-based approaches instead train a machine learning model on the full set of candidate variables and then measure the importance of each candidate variable to the model based either on the contribution to the model's prediction (for example, the coefficients in a linear regression model), or using a sensitivity analysis to assess the impact of the variable on model performance. The model-free methods we used were the Spearman and Kendall rank correlation coefficients, and the model-based approaches were Combined Network Pathway Strength Analysis (CNPSA) and Input Omission (IO), both measured using artificial neural networks (ANNs), and Mean Decrease in Impurity (MDI) and Mean Decrease in Accuracy (MDA) using random forests (RFs).

We used each of these methods to score each candidate variable only at the best lag identified in Step 2. Thus, for the ranking of variables based on Spearman and Kendall rank correlation, we used the correlation scores obtained for the best lag in Step 2 and ranked each candidate variable based on the absolute value of the correlation coefficient. The model-based methods are described below. After each IVS method had been used to rank each candidate variable for all three models, an overall ranking was generated for each variable for each model by taking the sum of the rank for each IVS method

Step 4

Train an ANN ensemble using the past two lagged TP measurements. Sequentially add each candidate variable at its best lag based on the overall ranking developed in Step 3, tracking the performance of the model with each variable added. Eliminate two candidate variables which contribute to a decrease in performance. Repeat until either there are no more variables that decrease performance or until the best performance is reached.

Testing the model

Regression models: Root mean square error (RMSE)

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\text{Predicted}_i - \text{Observed}_i)^2}{n}}$$

Classification models: Accuracy, Recall, Precision

$$\mathbf{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{n}$$

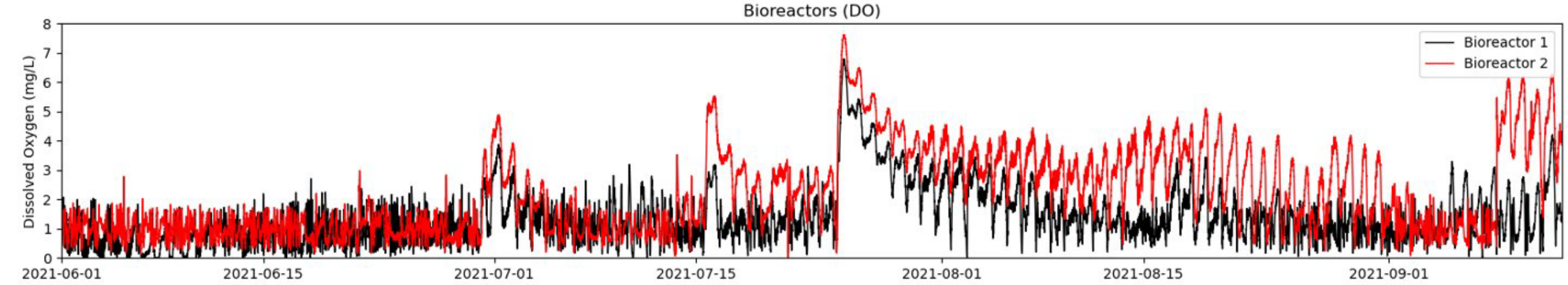
$$\mathbf{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

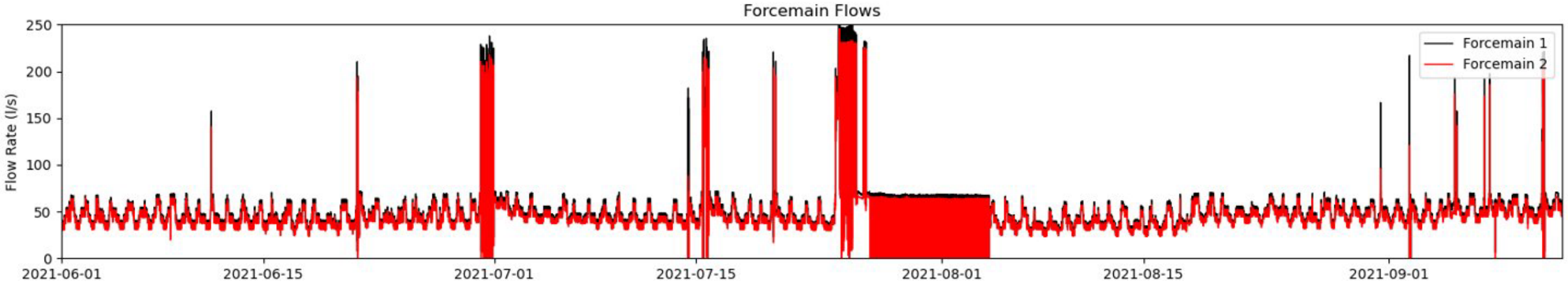
$$\mathbf{Precision} = \frac{\text{True Positives} + \text{False Positives}}{\text{True Positives}}$$

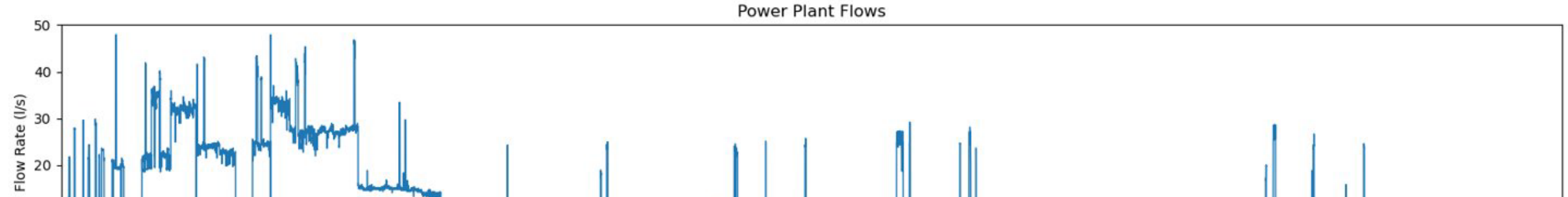
Confusion matrix

Predicted Condition	
Positive	Negative
True Positive	False Negative
False Positive	True Negative

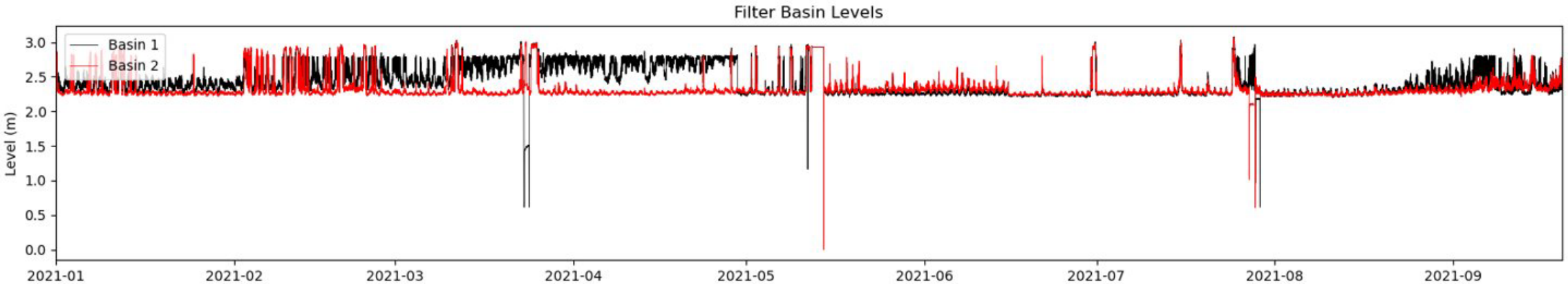
Preliminary data exploration

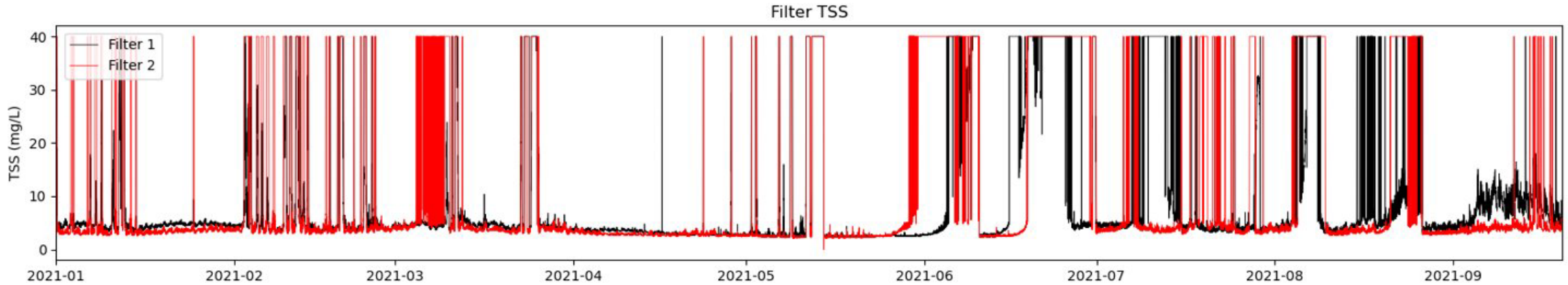


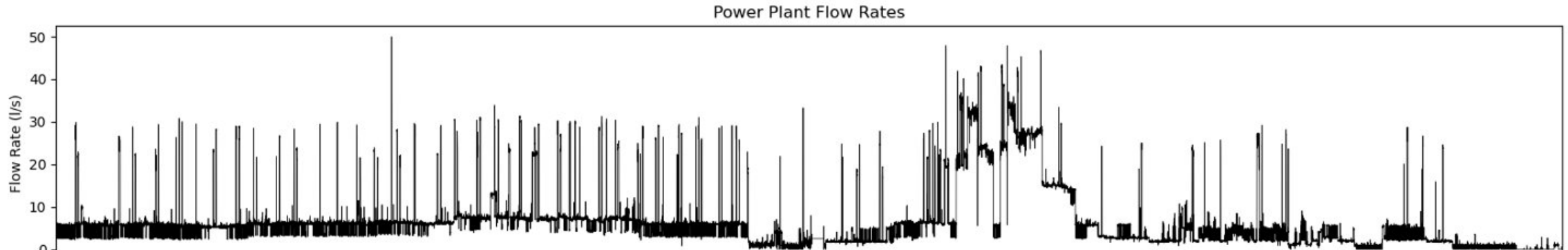




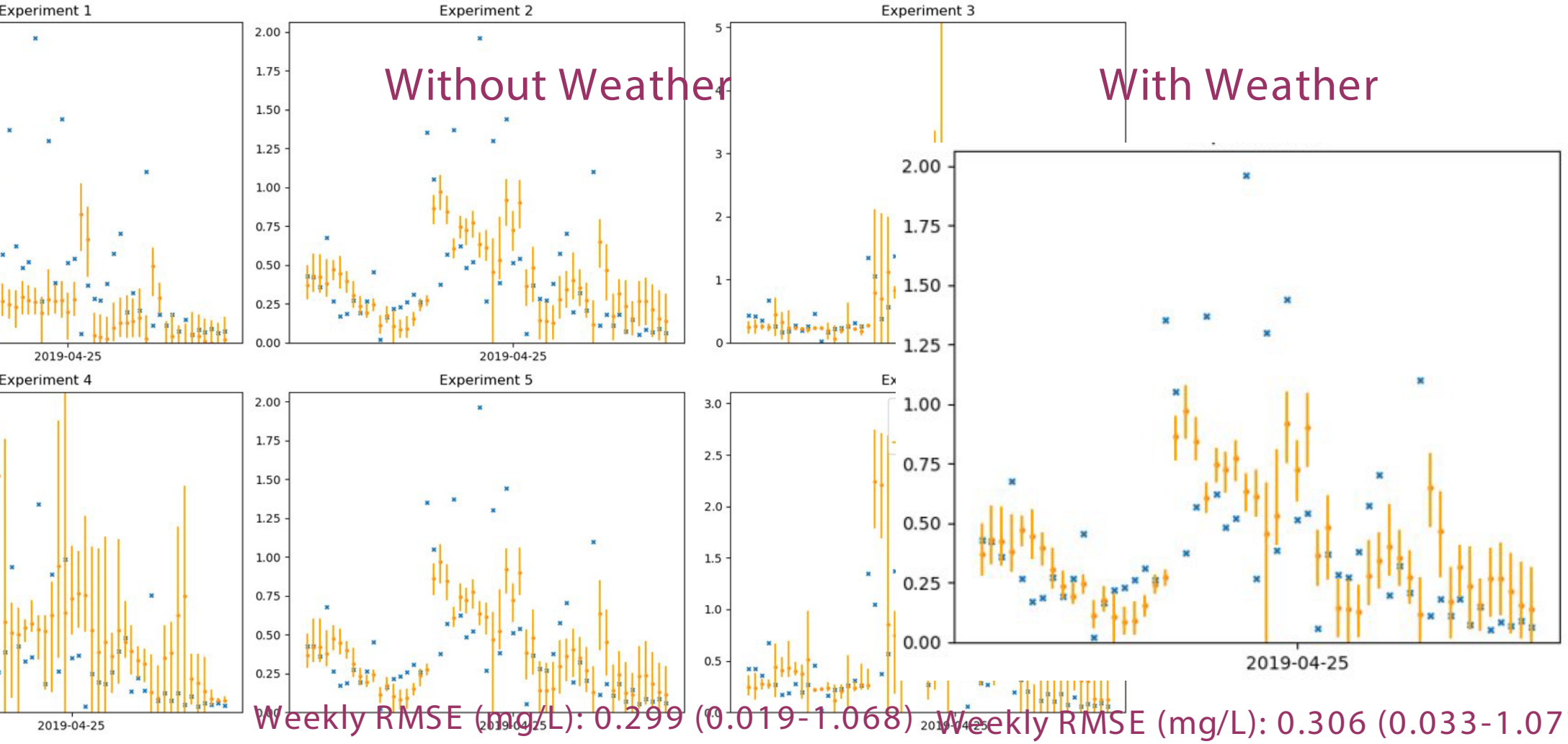
Preliminary data exploration







Climate effects (barometric pressure and rainfall)



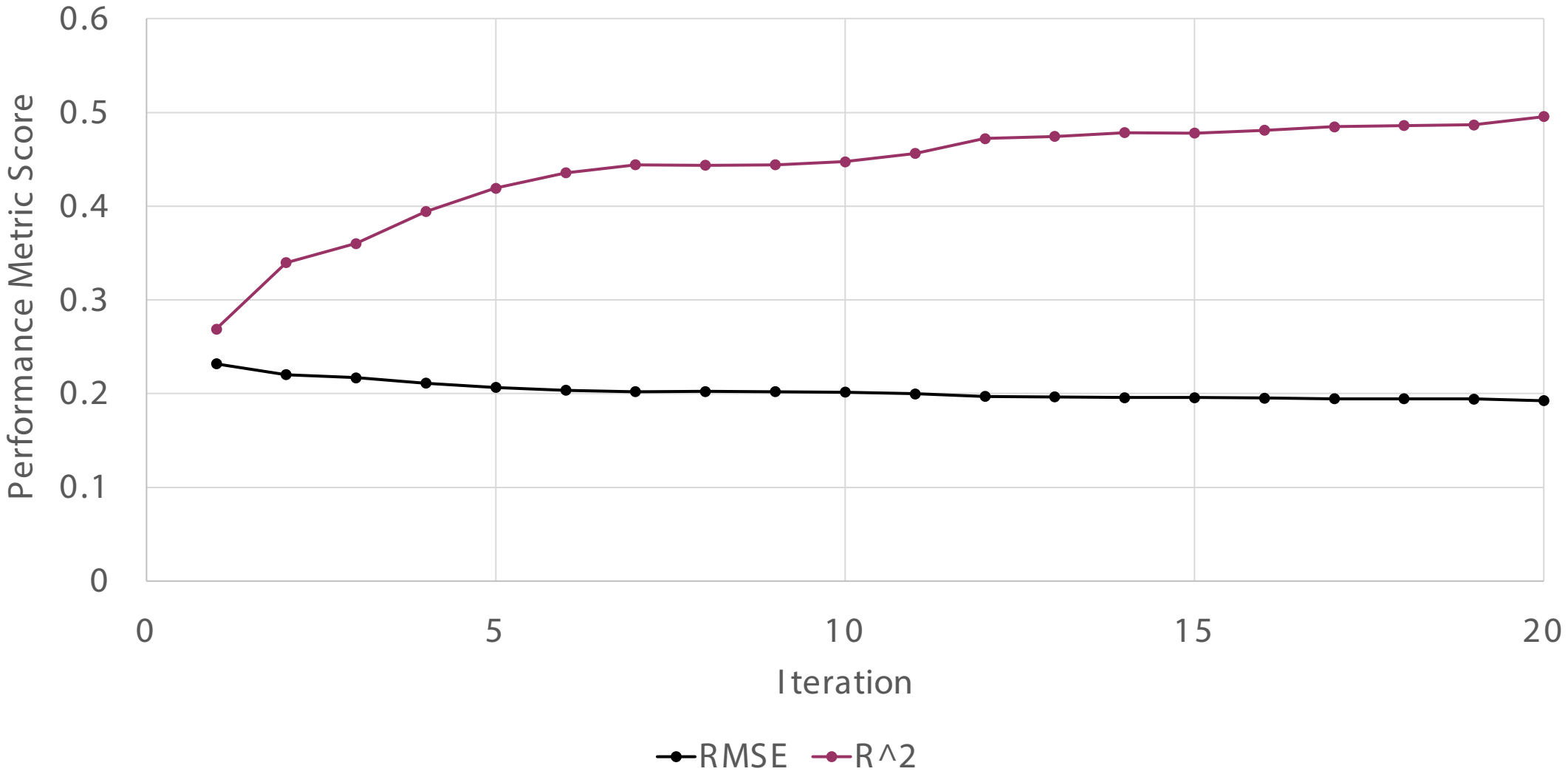
Without Weather

Weekly RMSE (mg/L): 0.299 (0.019-1.068)

With Weather

Weekly RMSE (mg/L): 0.306 (0.033-1.077)

Backwards elimination: Final round



Methods: Plant and Data Description

Anonymized WWTP operated by Ontario Clean Water Agency (OCWA)

No primary clarification – solids removal in headworks and tertiary treatment

Phosphorus removal:

- Primary treatment via chemical precipitation (with alum)
- Secondary treatment
- Tertiary treatment (filtration)

Dataset

- 200+ SCADA variables collected every 5 minutes
- Biweekly total phosphorus