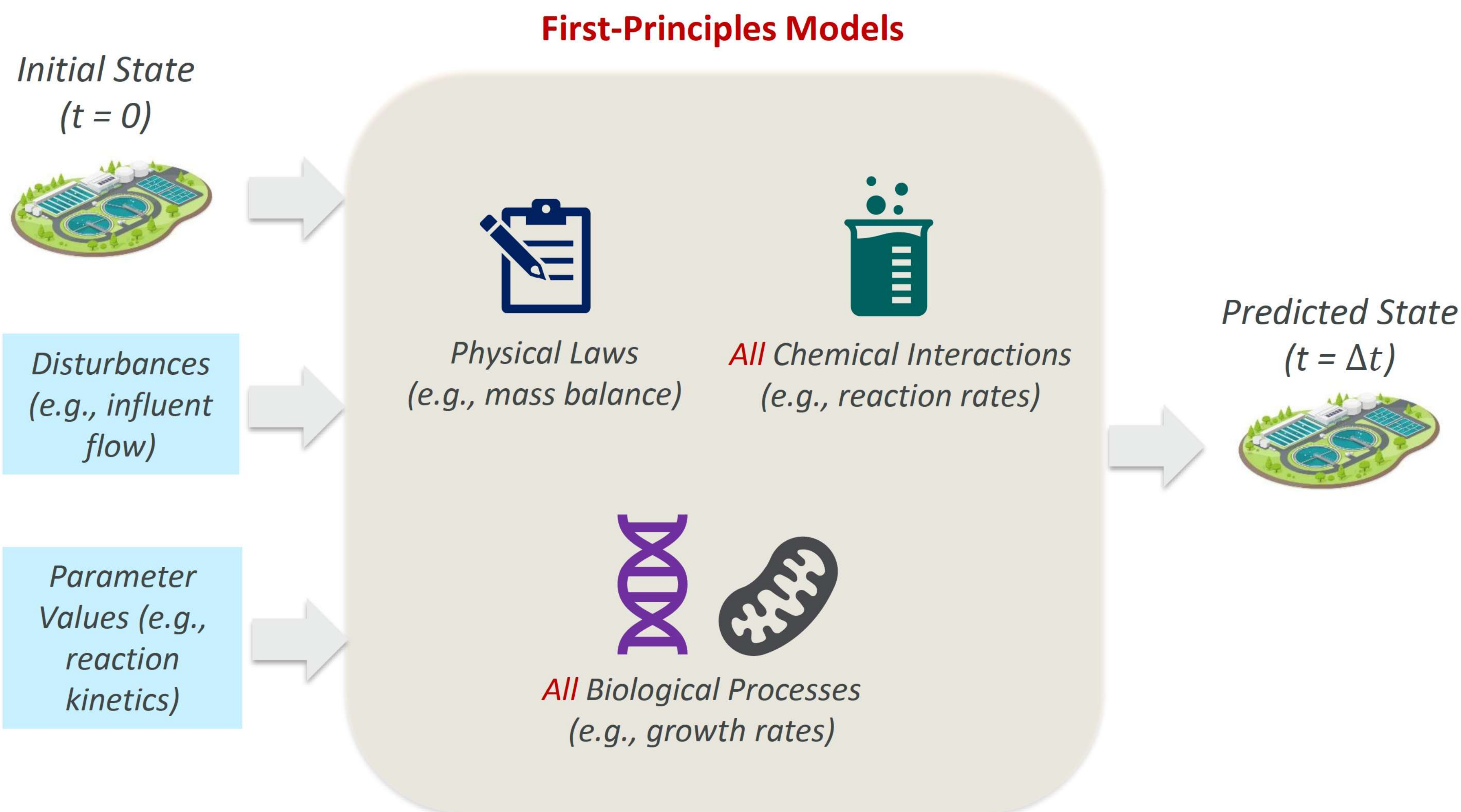
DATA-DRIVEN MODELING IN THE WATER SECTOR: A PARADIGM SHIFT IN PREDICTIVE METHODS

Mostafa Khalil, PhD

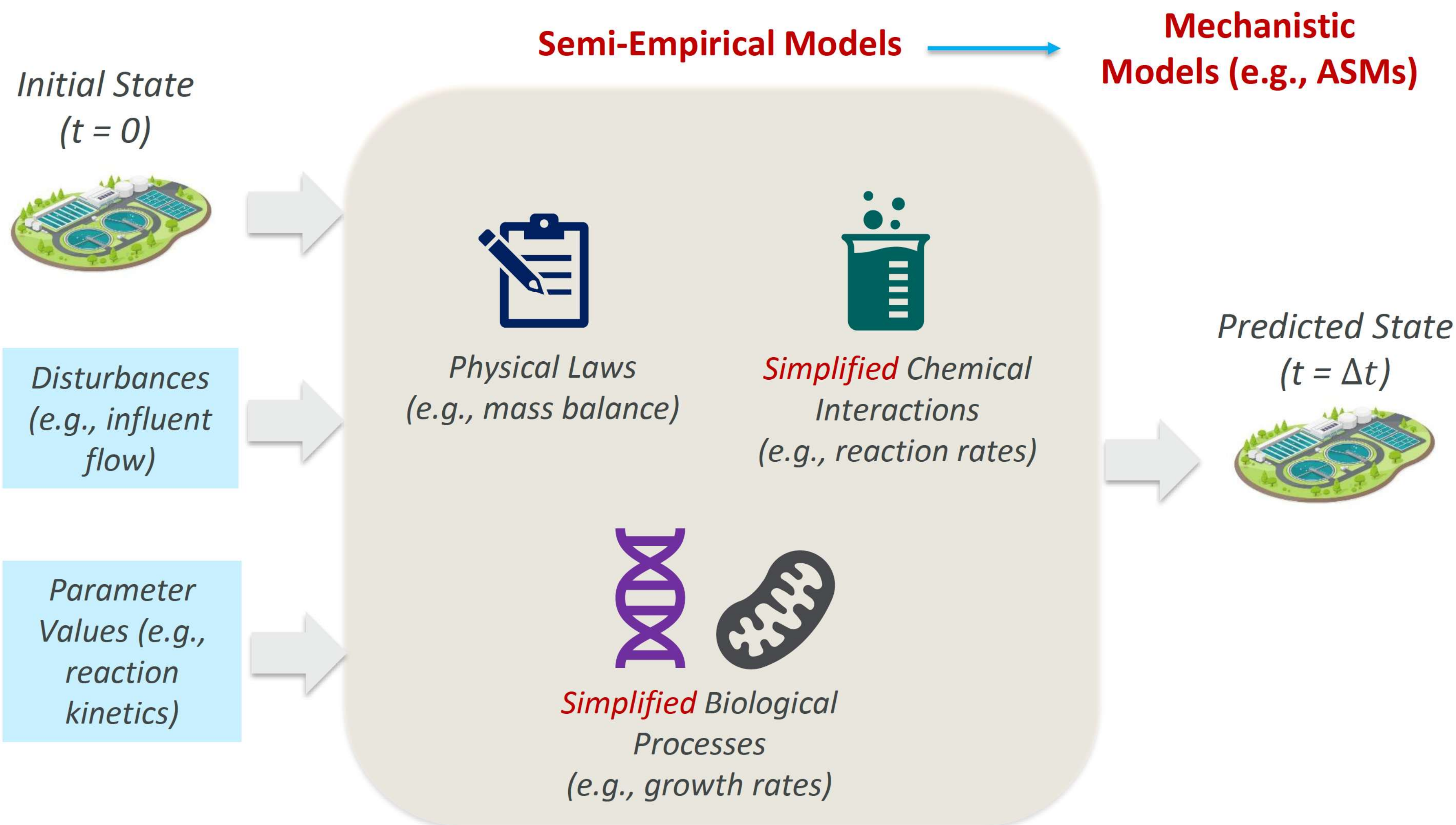
Data Scientist | Innovation Engineer, Stantec

Views are my own

MODELING IN THE WATER SECTOR



MODELING IN THE WATER SECTOR

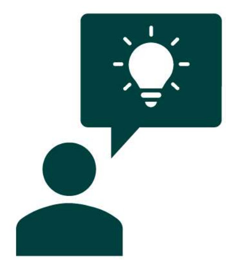


MECHANISTIC MODELS

- Gold standards in our field
- Model cause – effect relationships
- Can answer “what-if” questions
- Works outside historical range

$$\text{biomass growth} = \overset{\text{Max. Specific Growth Rate}}{\mu_h} \underset{\text{Biomass}}{X_{bh}} \left(\frac{S_O}{K_O + \underset{\text{Oxygen}}{S_O}}\right)\left(\frac{S_S}{\overset{\text{Half-Saturation Coefficient}}{K_S} + \underset{\text{Substrate}}{S_S}}\right)\left(\frac{S_{NH}}{K_{NH} + \underset{\text{Nutrient}}{S_{NH}}}\right)\left(\frac{S_{ALK}}{K_{ALK} + \underset{\text{Alkalinity}}{S_{ALK}}}\right)$$

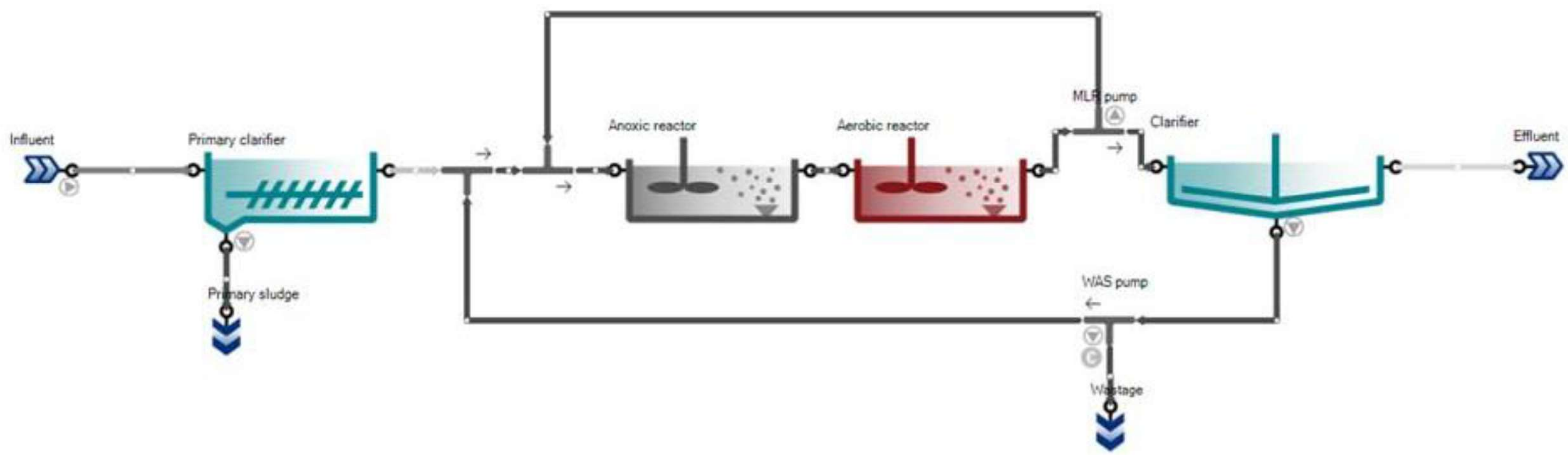
Cause 1 → Formula → Effect 1



MECHANISTIC MODELS

Conditional trust and confidence:

- No Unknown or poorly understood relationships
- Parameters are accurate and (somewhat) fixed over time
- Same formulas will hold at all conditions



WHY DO WE BUILD MODELS?

t | t+Δt

System States u(t)

Inputs u(t) → System States → **Outputs u(t)**

Influent Generator

Monitoring & Understanding

Planning & Decision Support

WHY DO WE BUILD MODELS?

t | t+Δt

Inputs u(t)

System States u(t)

Outputs u(t)

Resource Planning

Asset Health & Maintenance

Process Optimization & Control

WHY DO WE BUILD MODELS?

Process Monitoring & Understanding

Process Optimization & Control

Planning and Decision Support

Asset Health & Maintenance

Forecasting

UTILIZING THE POWER OF DATA

Artificial Intelligence

Machine Learning

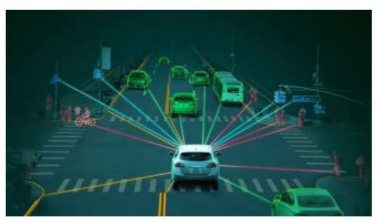
Natural Language Processing

Computer Vision

Robotics

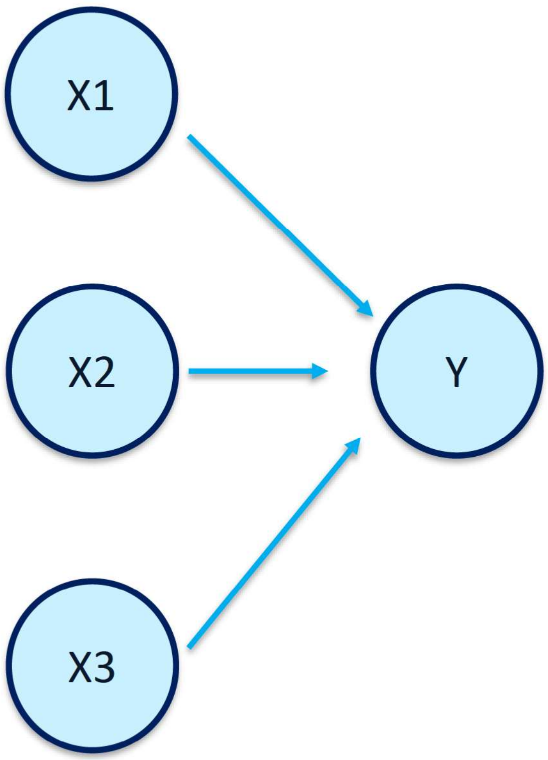
Speech Recognition

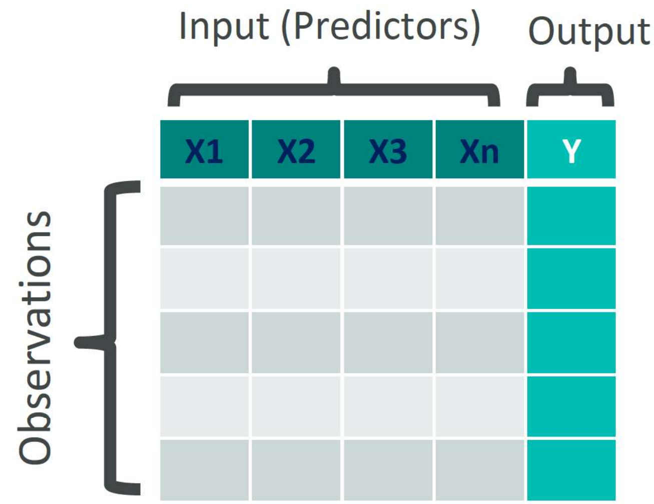
ChatGPT

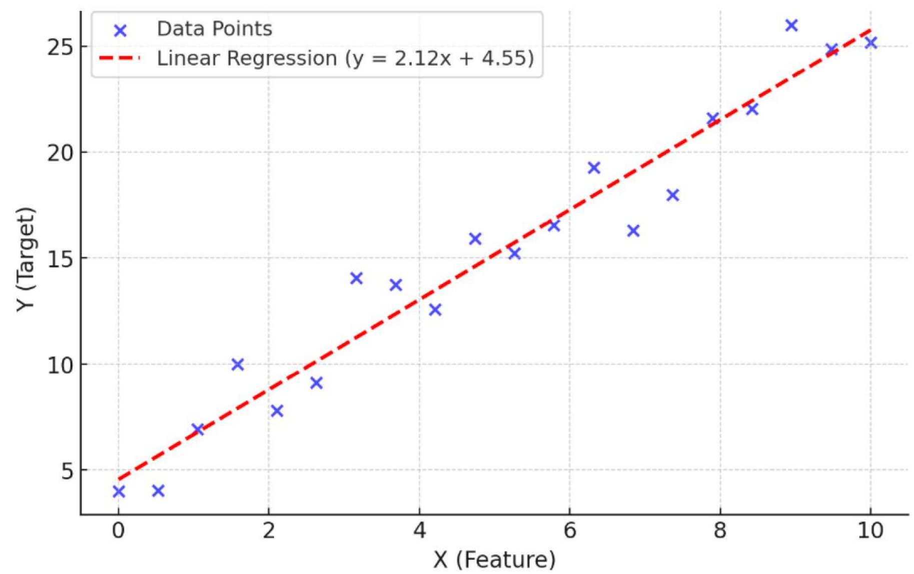


Hey Siri

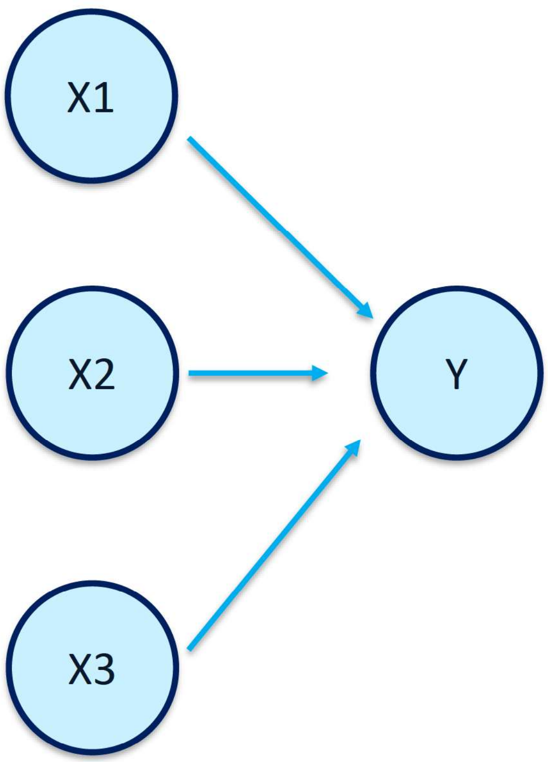
UTILIZING THE POWER OF DATA: MACHINE LEARNING

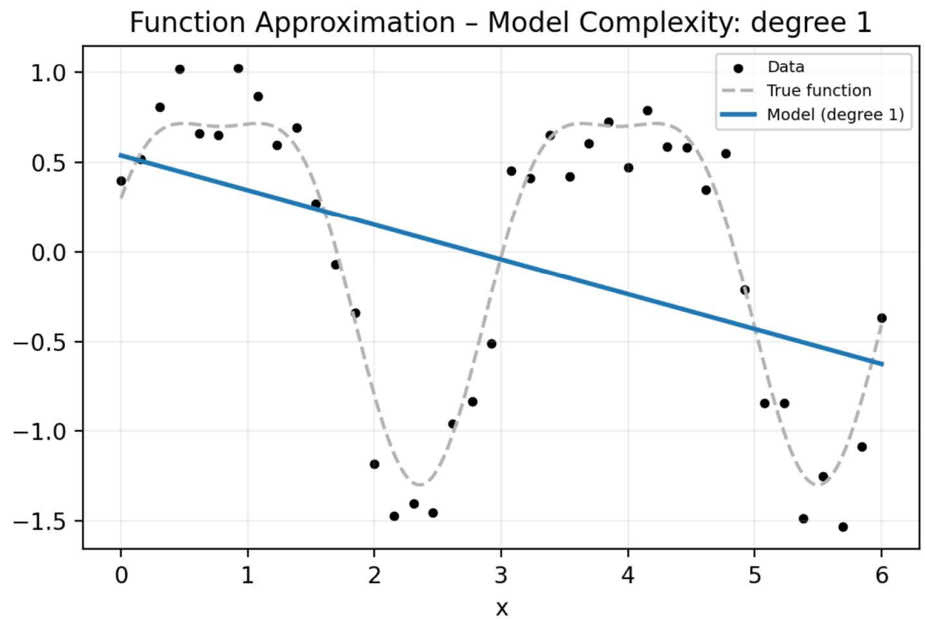




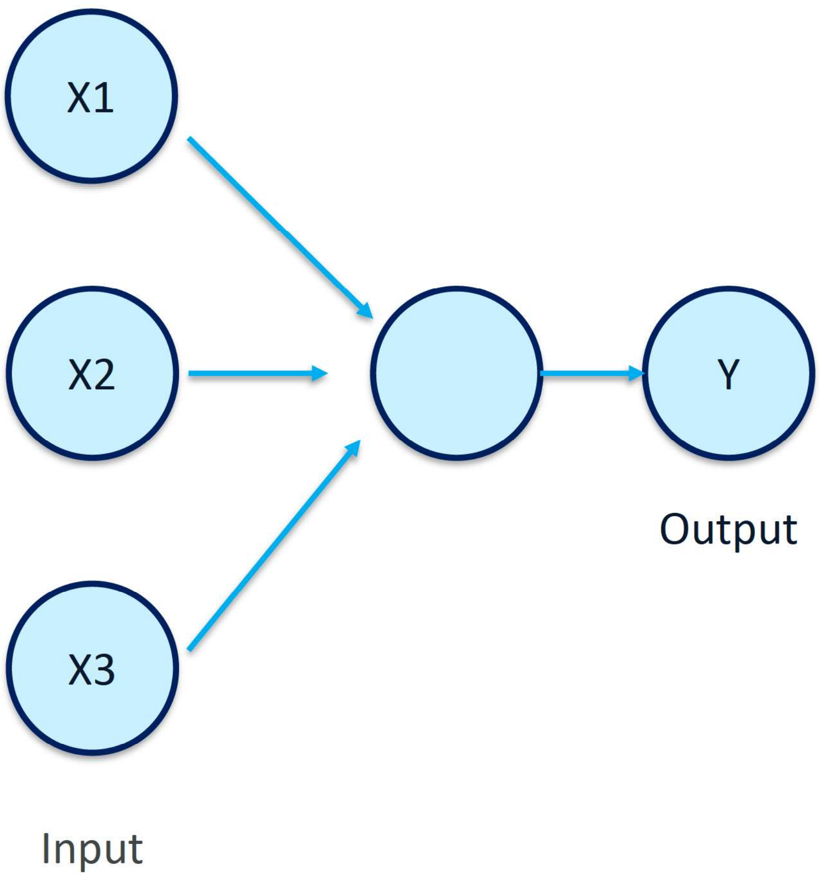


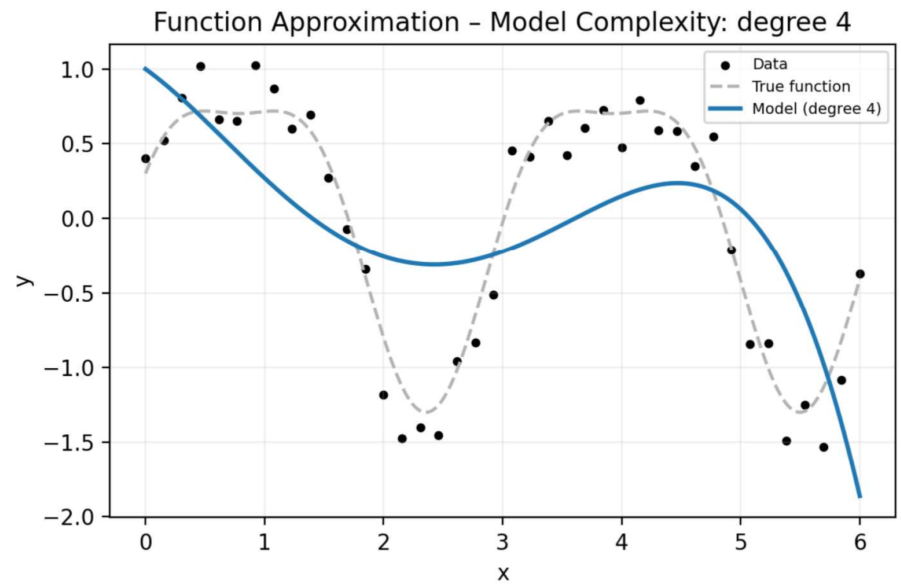
UTILIZING THE POWER OF DATA



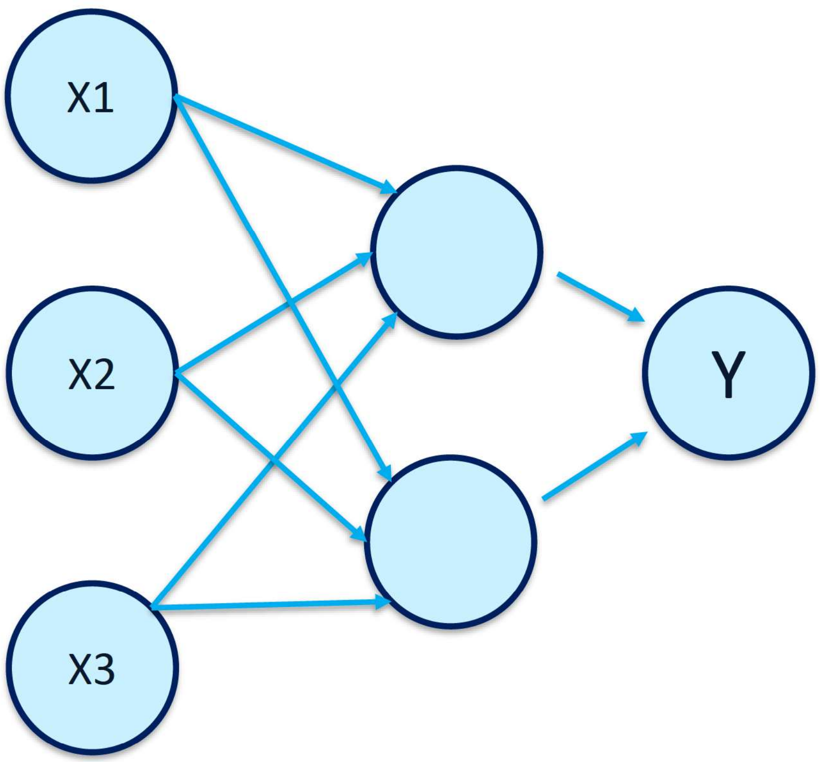


UTILIZING THE POWER OF DATA

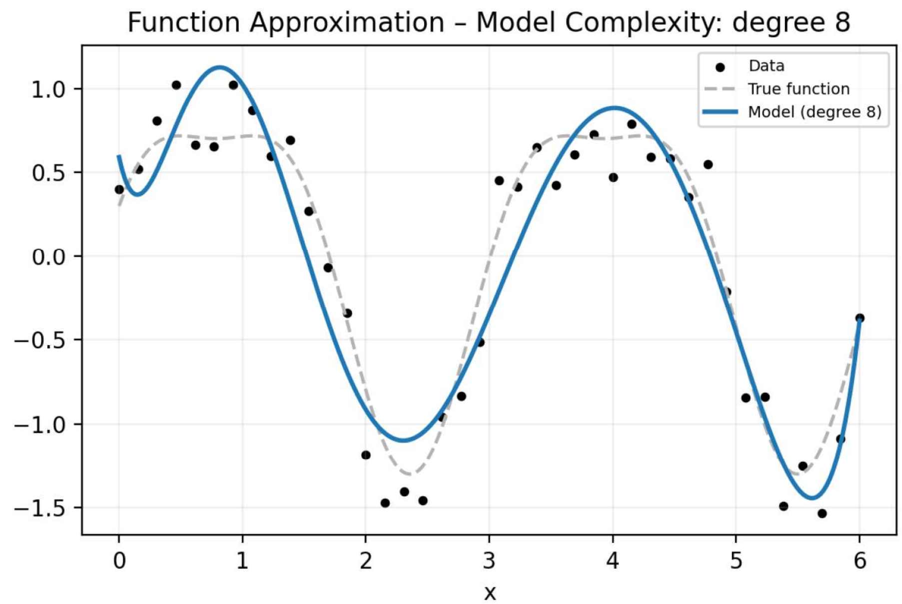




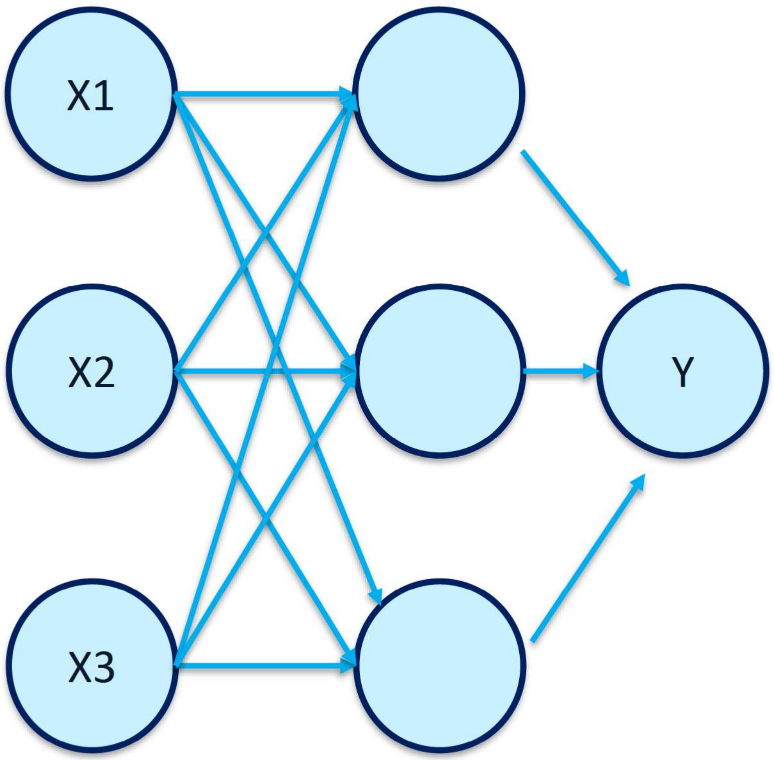
UTILIZING THE POWER OF DATA



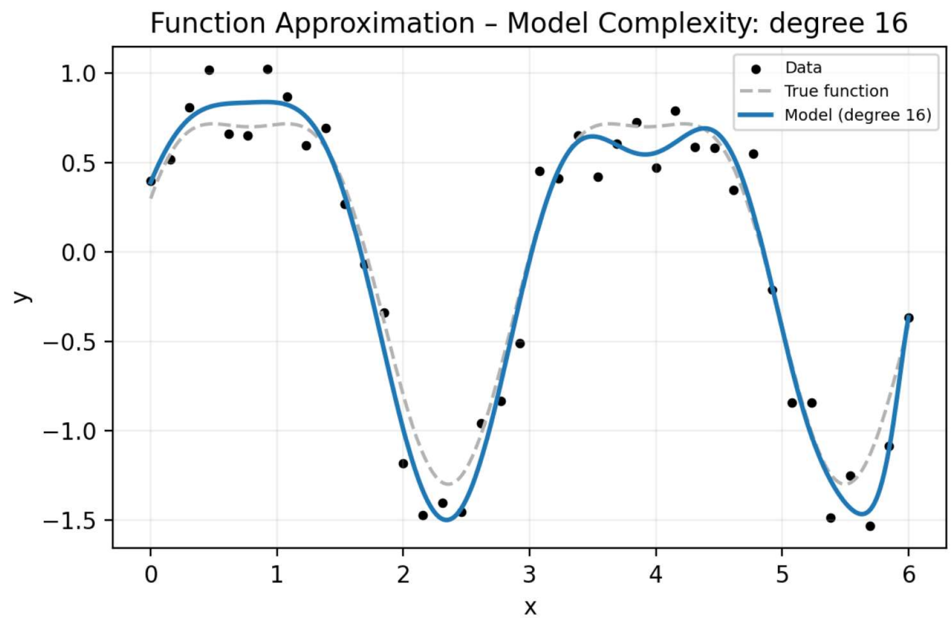
Input



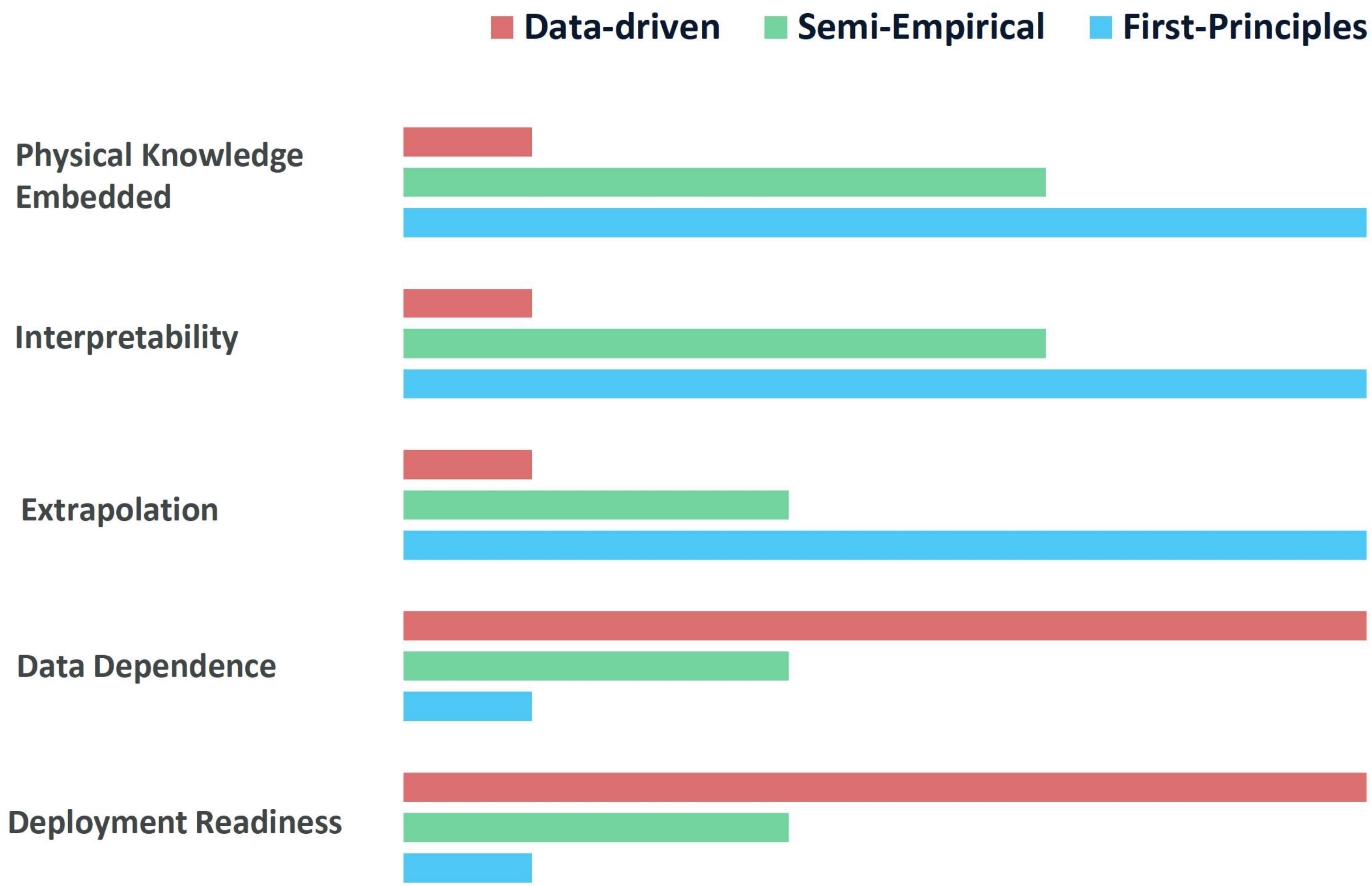
UTILIZING THE POWER OF DATA



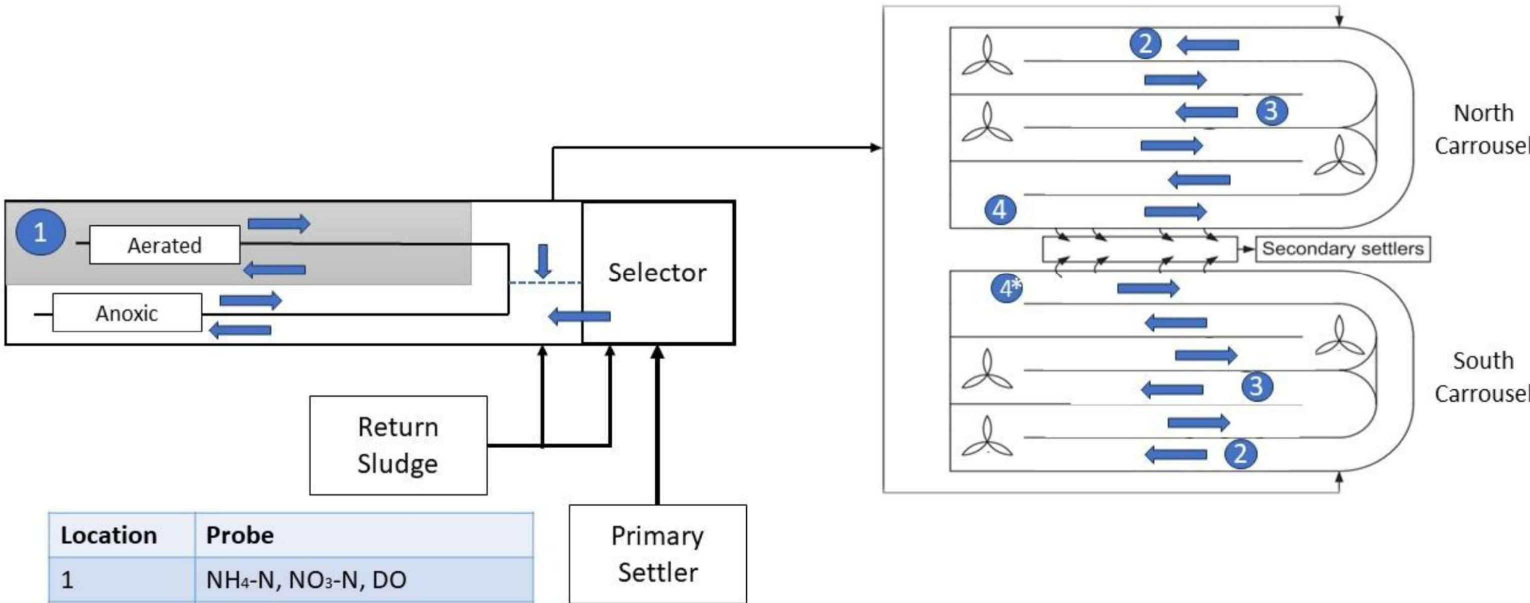
Input



INTRODUCTION TO CHALLENGES

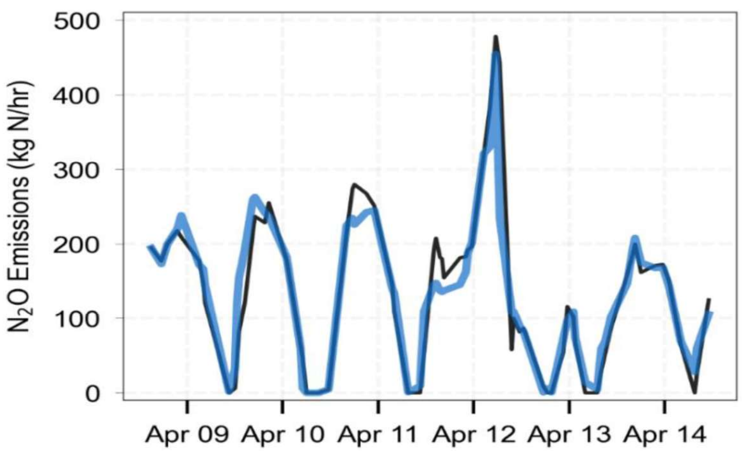


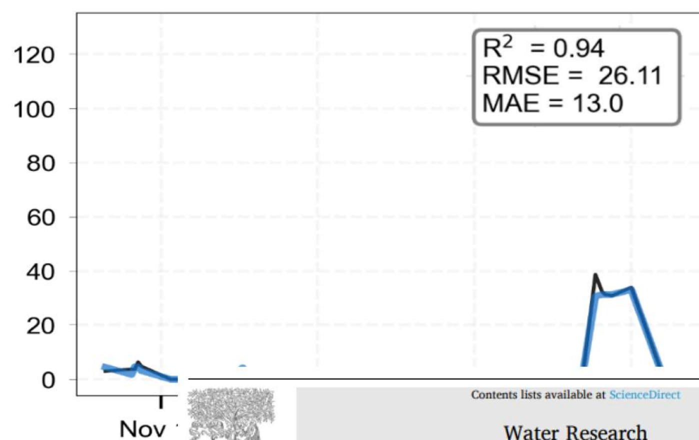
EXAMPLE: N_2O EMISSIONS MODELING

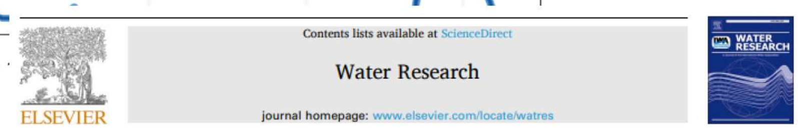


Location	Probe
1	NH_4-N, NO_3-N, DO
2	DO1, TSS
3	DO2, Temperature
4	DO3, NH_4-N, NO_3-N,NO_2-N
4*	DO3, NH_4-N, NO_3-N

EXAMPLE: N_2O EMISSIONS MODELING







Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

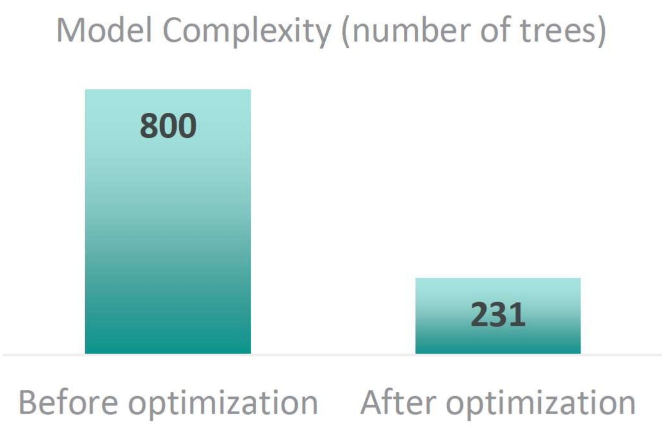
ELSEVIER

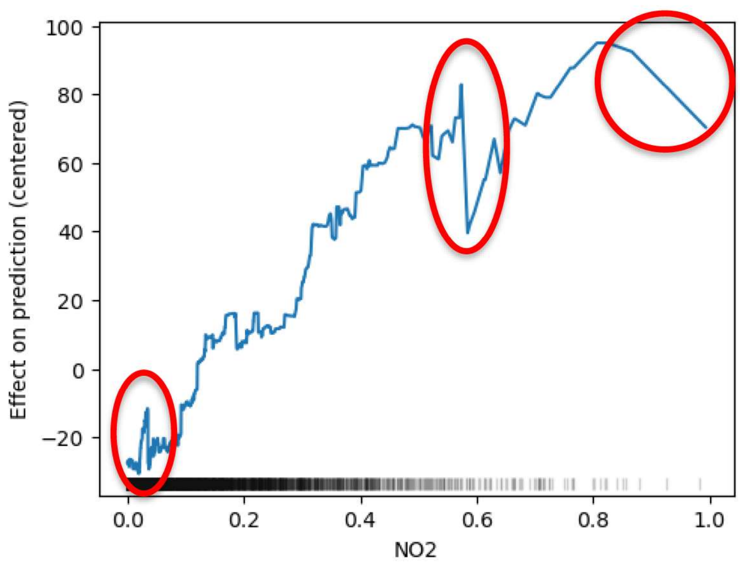
Machine learning for modeling N_2O emissions from wastewater treatment plants: Aligning model performance, complexity, and interpretability

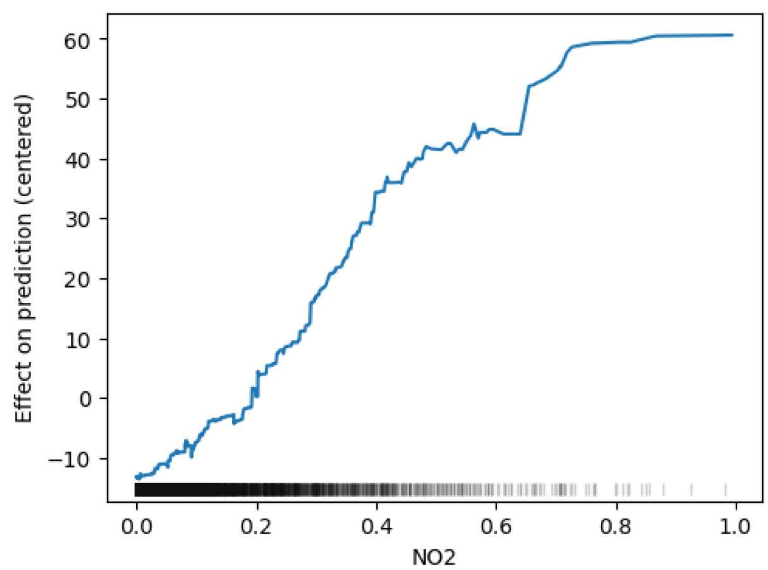
Mostafa Khalil [a], Ahmed AlSayed [b], Yang Liu [a,c,*], Peter A. Vanrolleghem [d]

[a] Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada
[b] Department of Civil and Environmental Engineering, McCormick School of Engineering, Northwestern University, United States
[c] School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, Queensland, Australia
[d] modelEAU, Département de génie civil et génie des eaux, Université Laval, 1065 av. de la Médecine, Québec, QC G1V 0A6, Canada

ALIGNMENT WITH DOMAIN KNOWLEDGE

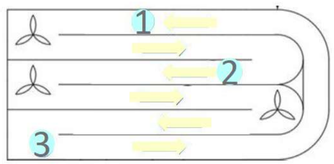




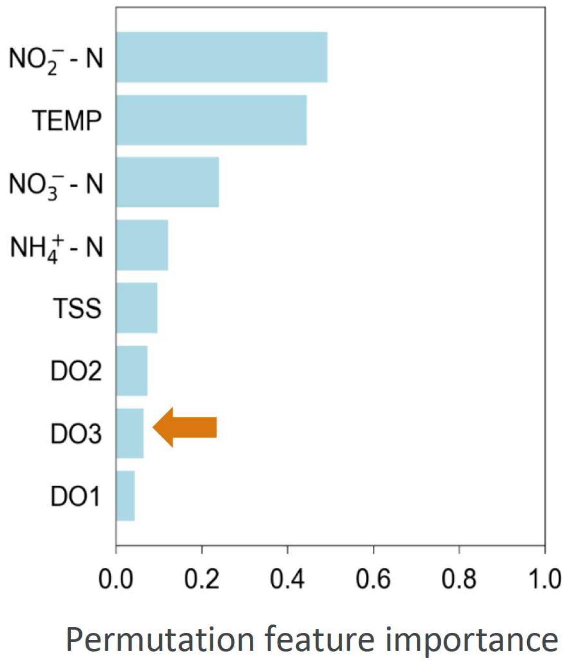


ALIGNMENT WITH DOMAIN KNOWLEDGE

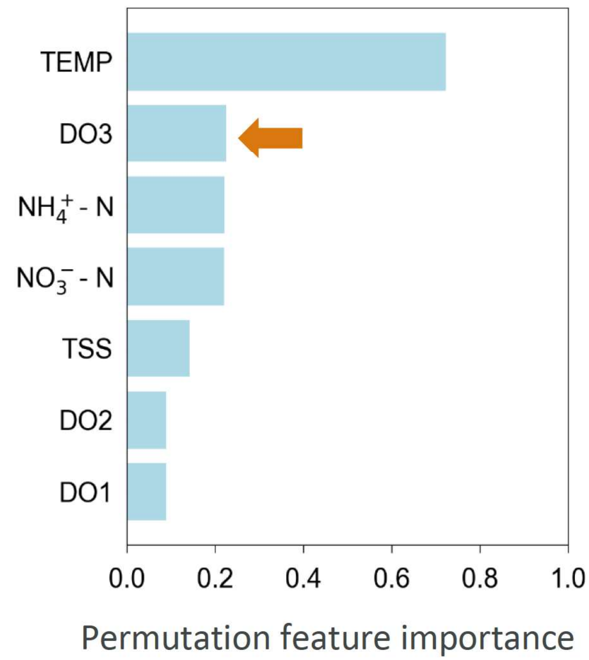
DO Measurement Locations



What features are the most important
for the model to make prediction?

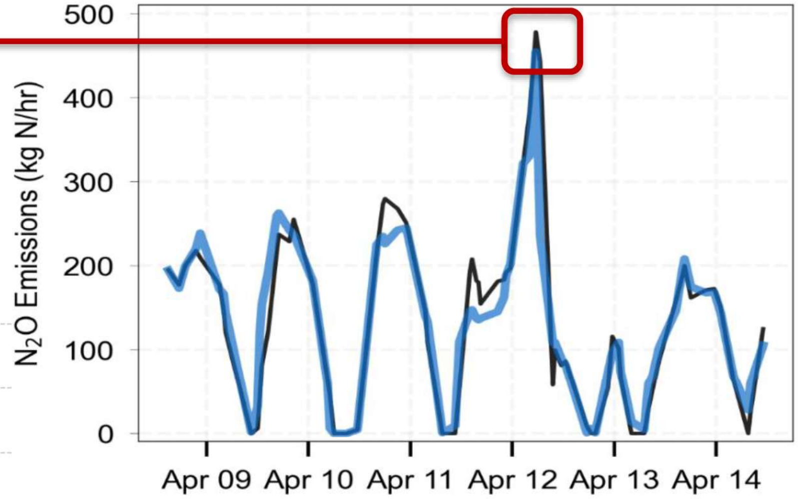


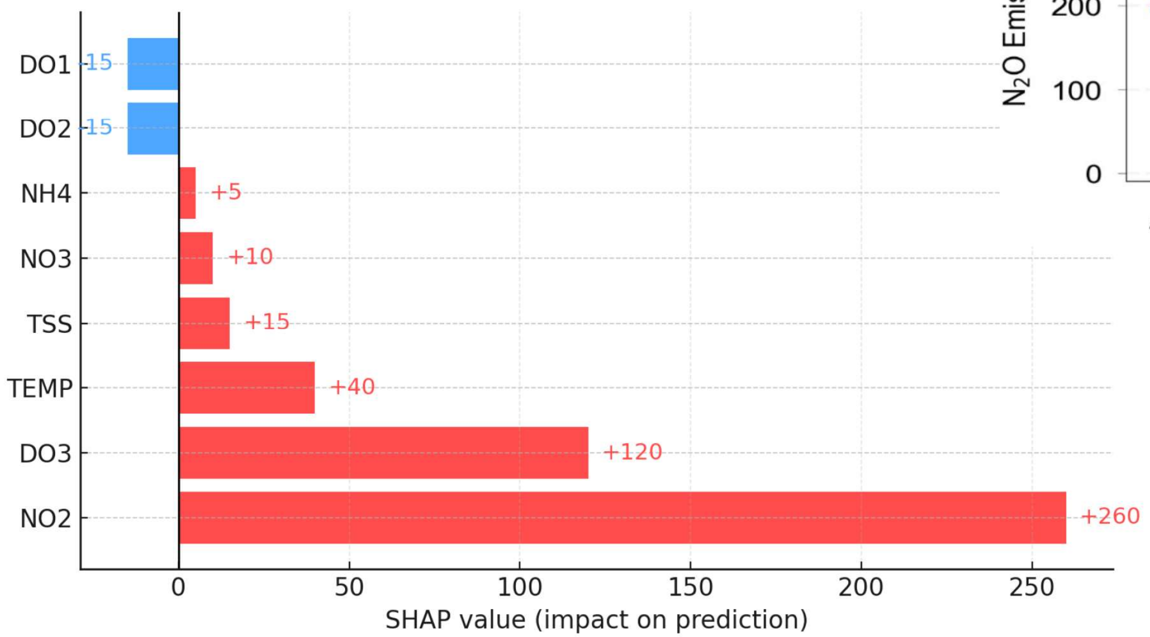
Remove NO_2 from input features



INTERPRETABILITY OF MODEL PREDICTIONS

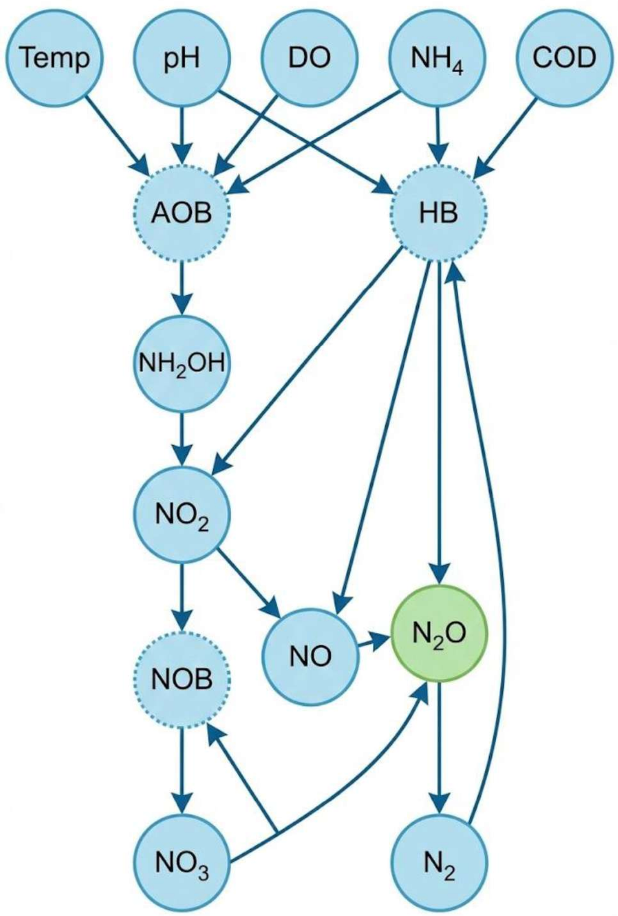
Why did the model predict this value?



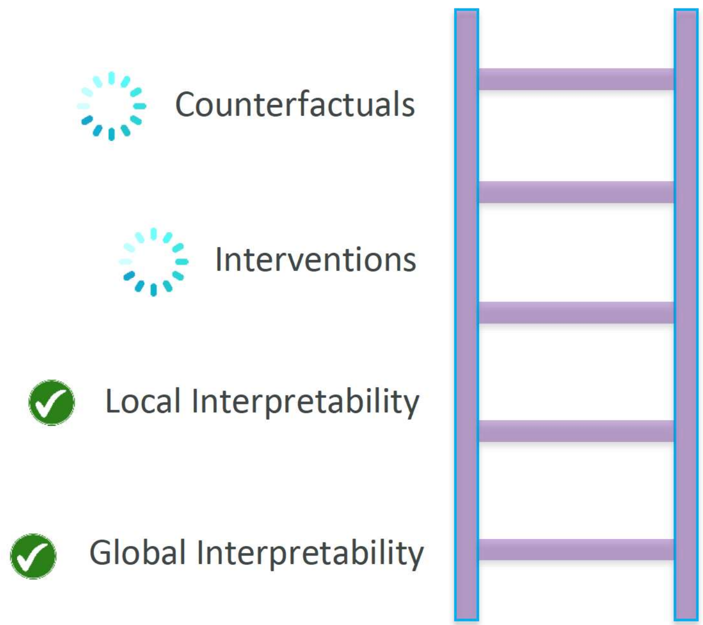


This is not a causal effect!

ADVANCING TOWARDS DECISION SUPPORT



Interpretability ≠ Causality

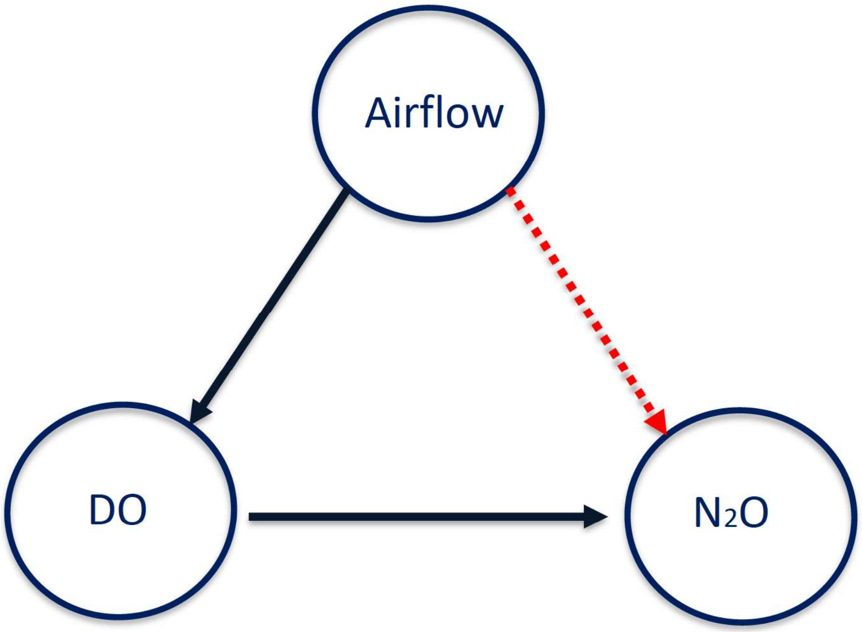


THE CAUSALITY PROBLEM

- Correlation can be misleading when an unmeasured factor (confounder) influences both variables

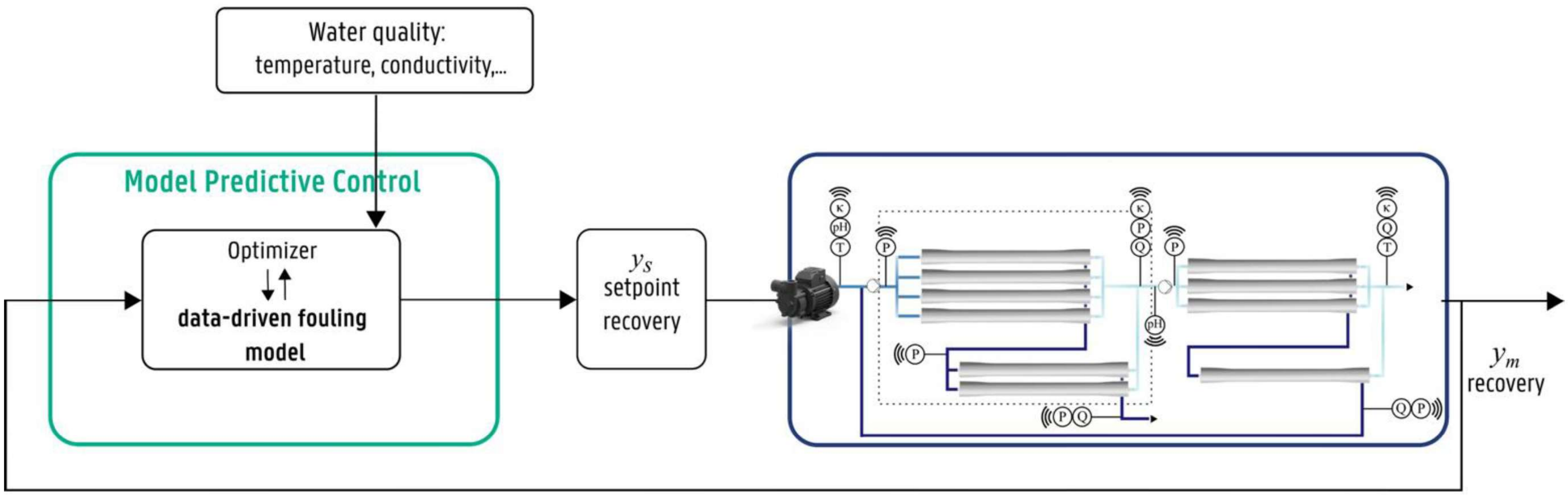
- P (N_2O | DO): What we observe in the data (correlation)

- P (N_2O | do(DO) = d): What would happen if we *intervene* on DO (causal effect)



ML models can easily learn correlations
but extracting *causation* requires extra work

EXAMPLE: RO MEMBRANE FOULING FORECASTING

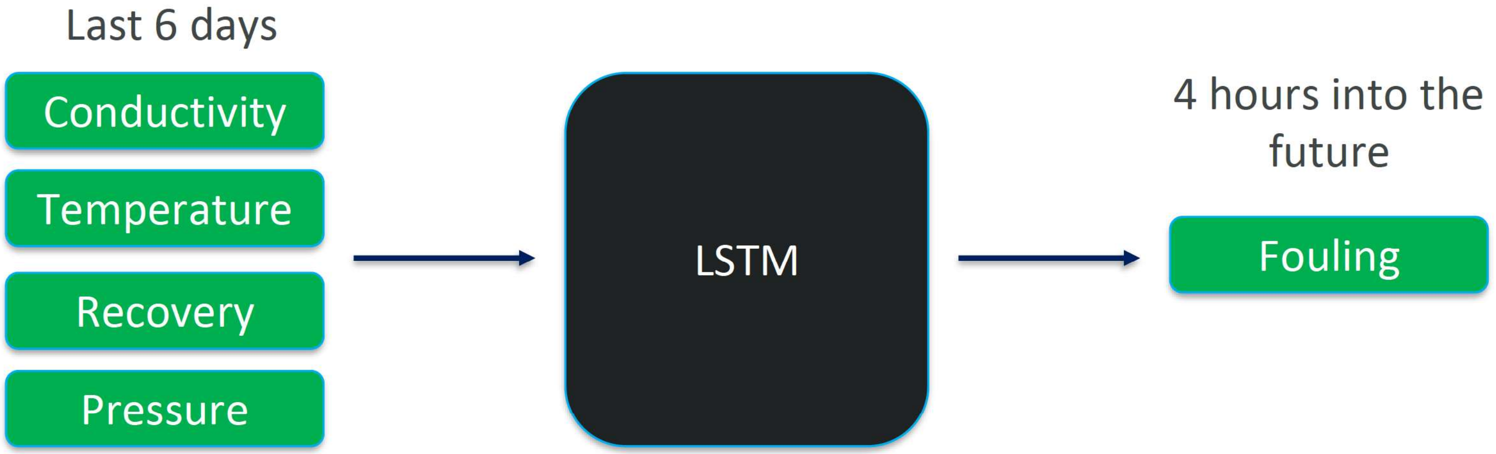


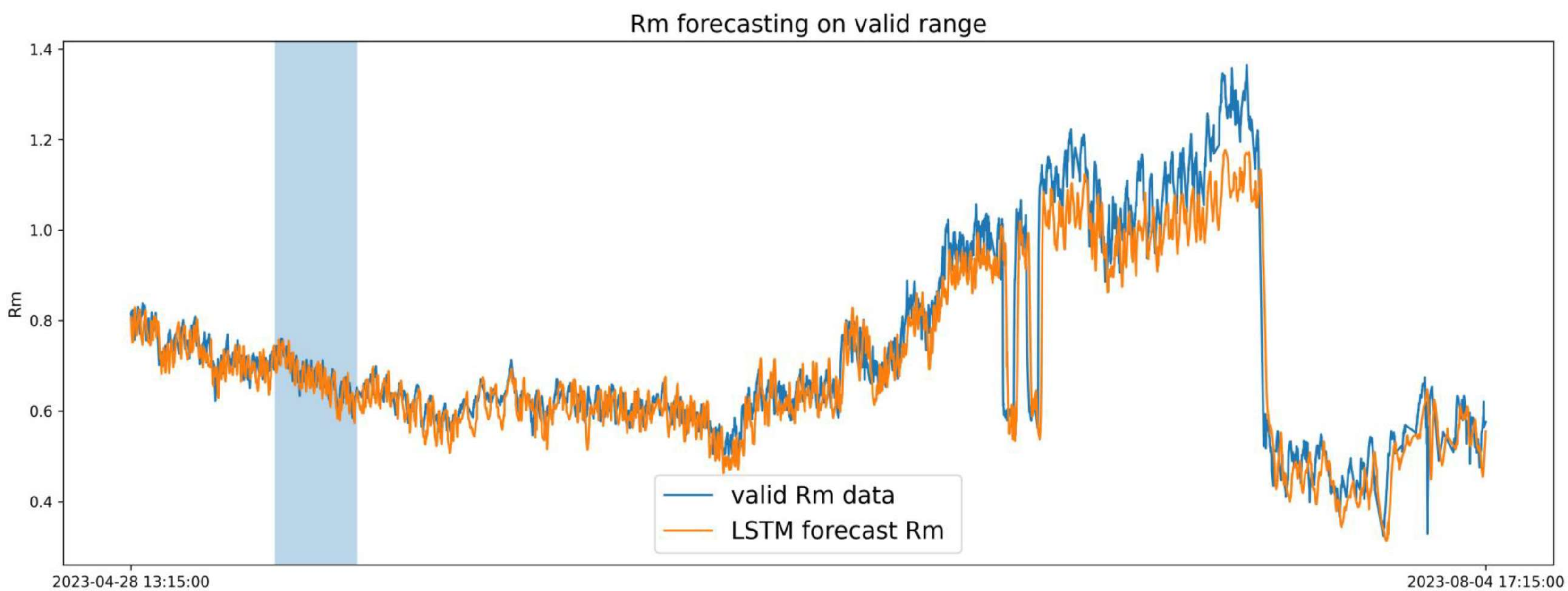
Objective function:

minimize fouling,

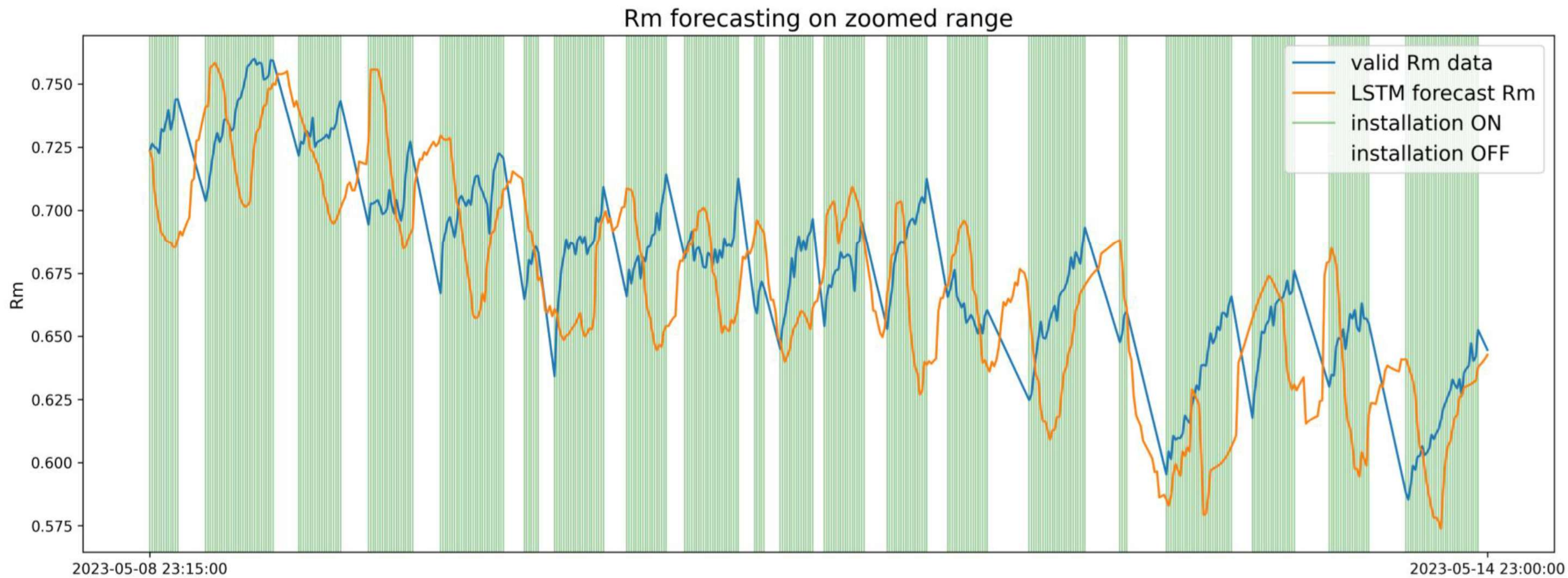
maximize recovery

EXAMPLE: RO MEMBRANE FOULING FORECASTING



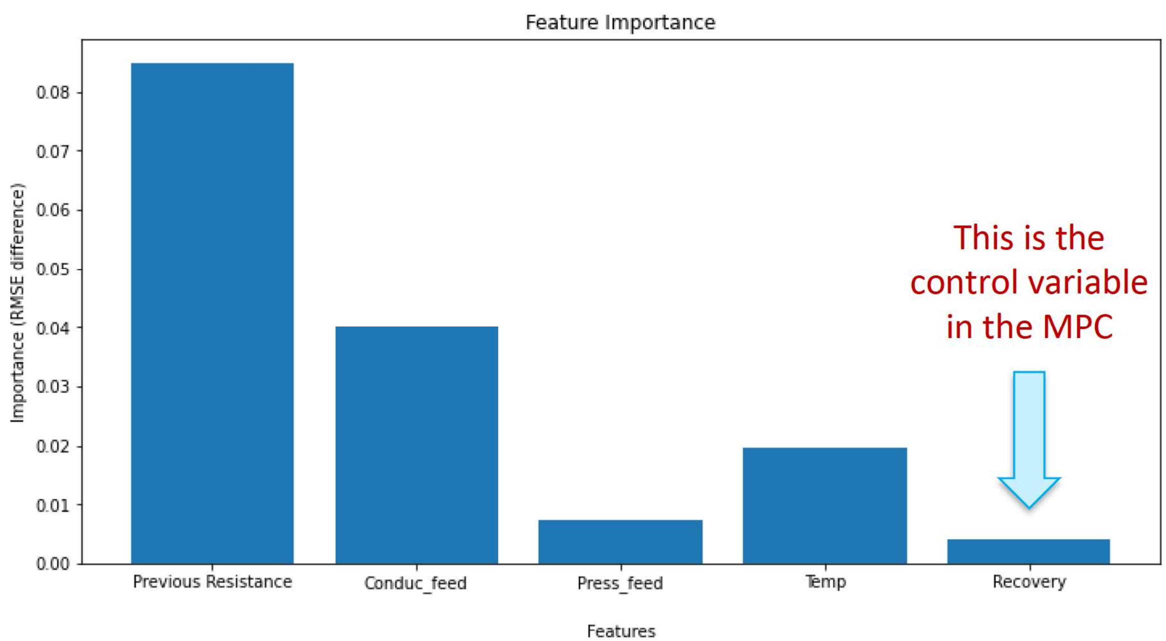


EXAMPLE: RO MEMBRANE FOULING FORECASTING

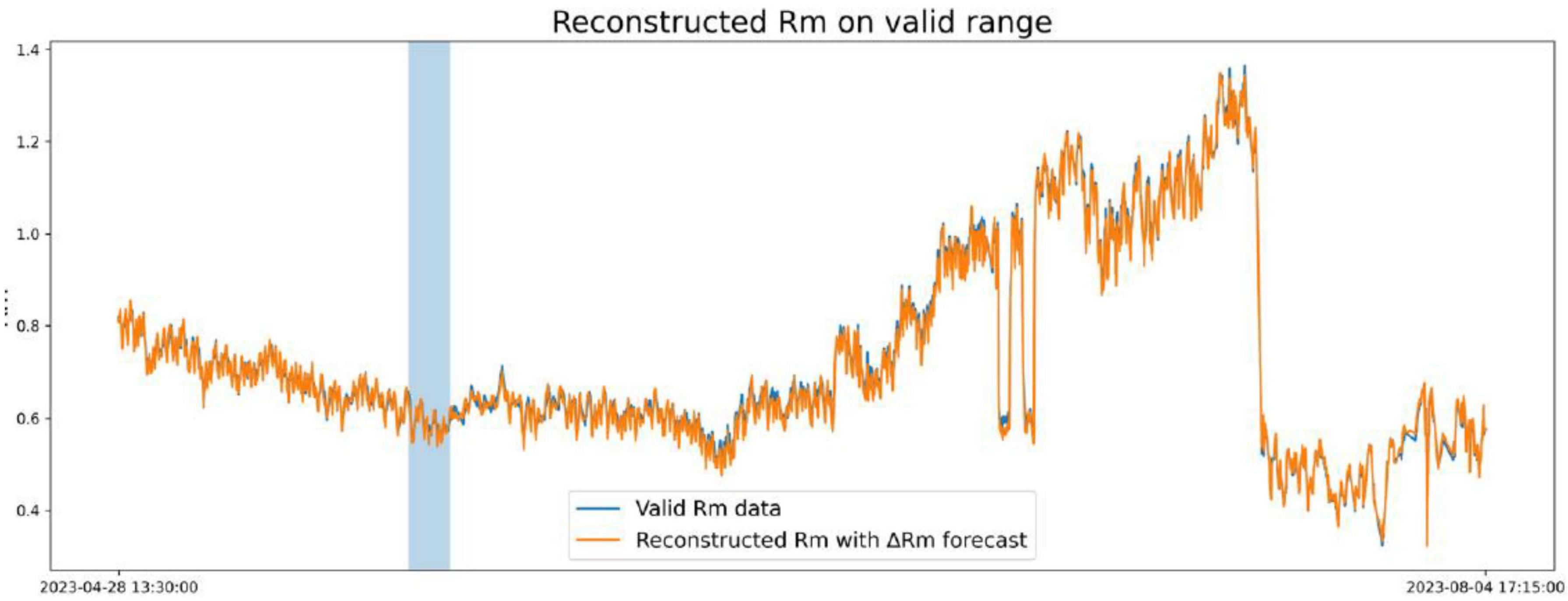


Lag in forecasted fouling (Rm) was caused by frequent OFF periods in the installation

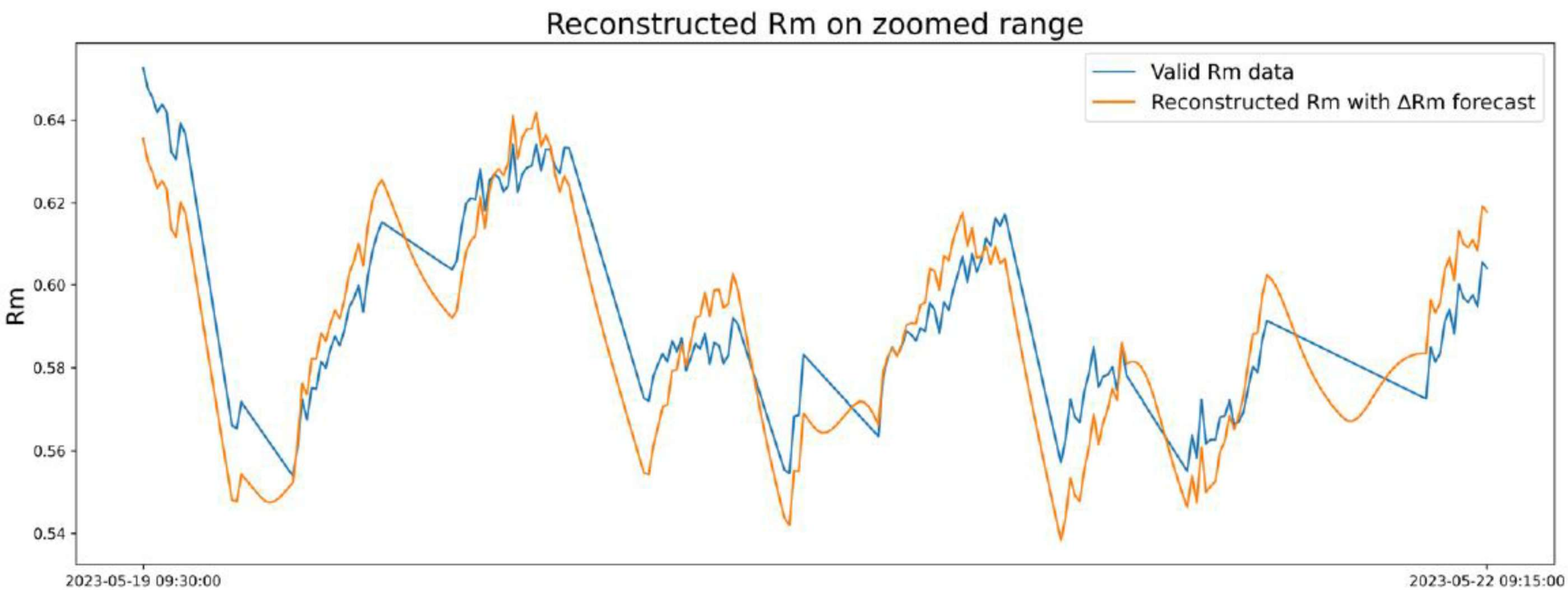
EXAMPLE: RO MEMBRANE FOULING FORECASTING



EXAMPLE: RO MEMBRANE FOULING FORECASTING



EXAMPLE: RO MEMBRANE FOULING FORECASTING



Predicting the *change* in fouling makes the model learn system dynamics directly, eliminating lag and improving responsiveness.

EXAMPLE: RO MEMBRANE FOULING FORECASTING

- Even for a powerful model, forecasting Rm directly made it slower to react to sudden changes (e.g., frequent on – off)
- Predictions were lagged during fast transitions
- ΔRm prediction is simpler as it removed slow trends and noise (stationary signal)
- Final Rm is reconstructed by adding predicted ΔRm, leading to better accuracy

KEY TAKEAWAYS!

Thank you!

- Data-driven methods expand our modeling toolbox — they don't replace physics or expertise
- ML is powerful but could be fragile: performance depends more on data and context than on algorithms
- Accuracy can be misleading: a good fit does not mean the model is correct
- ML is not magic. It introduces new challenges (drift, retraining, explainability) that must be managed deliberately
- A model can look right and be wrong. Accuracy is not the whole story