Data assimilation for prediction of ammonium in wastewater treatment plant: from physical to data driven models

Victor Bertret [1, 2, 3], Roman Le Goff Latimier [2], Valérie Monbet [3]

[1] Purecontrol, 68 Av. Sergent Maginot, Rennes, FRANCE victor.bertret@purecontrol.com

[2] Univ Rennes, ENS Rennes, IETR CNRS UMR 6164

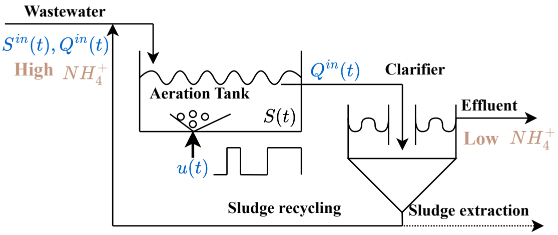
[3] IRMAR, Université Rennes 1, CNRS, Rennes, FRANCE

28 March 2025

Outline

1. Wastewater Treatment Plants
2. Possible models for ammonium concentration
3. Model's calibration with data assimilation
4. Simulated Results
5. First real results

Zoom on the secondary treatment of **Alternated Activated Sludge Plants** (AAS) :



- Municipal plants.
- **Input** : NH_4^+ (ammonium) high.
  - $Q^{in}(t)$, inflow.
  - $S^{in}(t)$, inlet concentration.
- **Output** : NH_4^+ low.

Objective : Reduce concentration of NH_4^+ coming from inlet water (nitrogen removal).

How ? : By **controlling** the addition of oxygen in the aeration tank with $u(t)$.

Aeration control

The removal of nitrogen is carried out by alternating between :

- **Aeration phases** (energy-intensive) : oxygen is added using a compressor.
- **Non-aeration** phases : complete absence of oxygen.

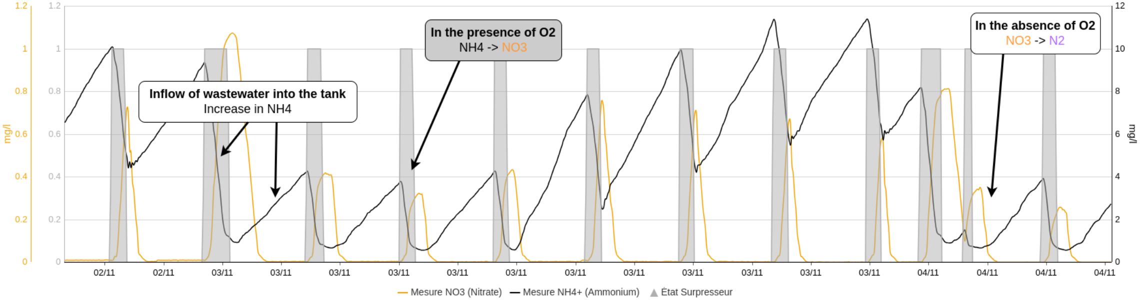


Figure – Nitrogen removal in the aeration tank.

Problems

Socio-economic Context

- **Increase** in environmental standards.
- Increase in **price** of energy.
- Increase in **variability** of energy.

Consequence : Minimize electric consumption.

Wastewater Treatment Plants

- **Complex** and **stochastic** system.
- **Expensive** and **unreliable** sensors.
- Controlled by **rudimentary** controllers.

Consequence : Finding new control algorithms.

Possible solution

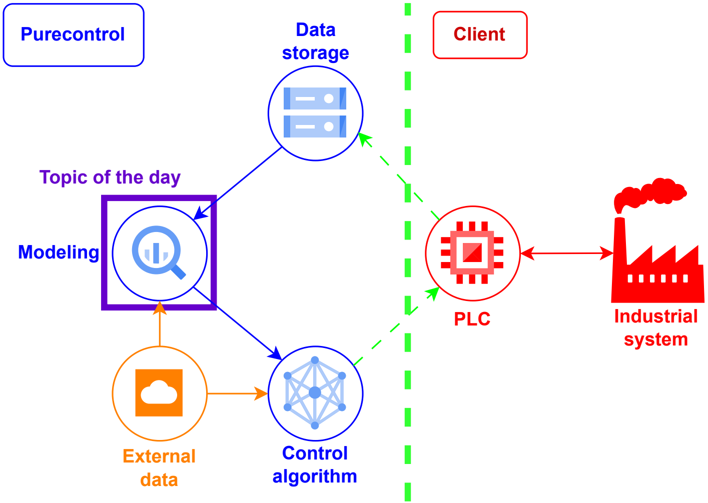
Feedforward/Scheduling Control.

→

Condition

Dynamic model (NH_4^+).

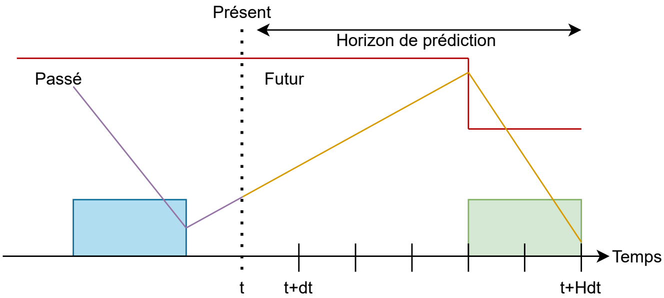
Purecontrol Solution for Energy Optimization



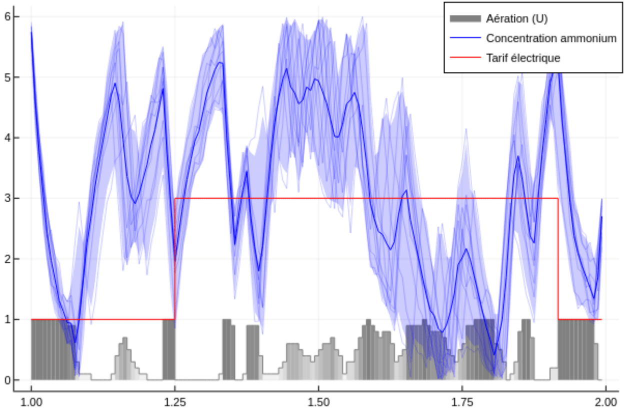
Model Predictive Control (MPC)

What is MPC ?

Model Predictive Control (MPC) predicts the system's future behavior using a dynamic model, allowing the generation of optimal control actions.



Recalculated at each time step *dt* $\Longrightarrow$ **dynamic adaptation** to system variations.



Prediction Objective

24-hour stochastic prediction of ammonium (NH_4^+).

Modeling Requirements

Calibration of parameters and uncertainties.

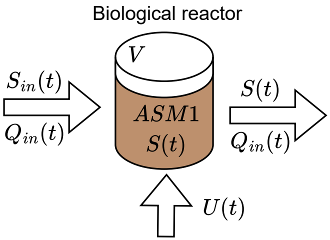
Outline

1. Wastewater Treatment Plants
2. Possible models for ammonium concentration
3. Model's calibration with data assimilation
4. Simulated Results
5. First real results

Well-known physic model

Development of family of **ASM's physic based** model

$$\frac{\partial S(t)}{\partial t} = \frac{Q^{in}(t)}{V}(S^{in}(t) - S(t)) + r(t, S(t), u(t))$$



- $S(t)$, **unknown** concentration of species.
- V, **unknown** volume of the reactor.
- $(Q^{in}(t), S^{in}(t))$, **known** inflow and inlet concentration.
- $u(t)$, **known** control applied to the system.

Our Case Study

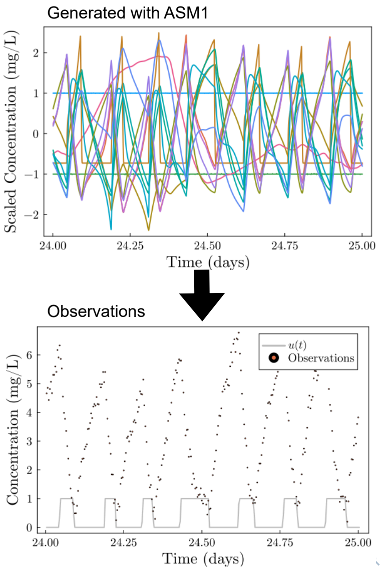
Main objective : High-precision NH_4^+ predictions over a 24-hour horizon.

Problem decomposition :

$$P(S(t), S_{in}(t)) = \overbrace{P(S(t)|S_{in}(t))}^{\text{NH}_4\text{ Dynamics}} \times \overbrace{P(S_{in}(t))}^{\text{Inlet Estimation}}$$

System characteristics :

- Medium Dimensionality
- Strong Interactions & High Uncertainty
- Partial Observability



White-box model : Reduced ASM1 model

$$\frac{\partial S(t)}{\partial t} = \overbrace{\frac{Q^{in}(t)}{V}(S^{in}(t) - S(t))}^{\text{Not modified}} + \overbrace{r(t, S(t), u(t))}^{\text{Modified part}}$$

- **Problem** :
  - Complex physic model.
  - 13 species.
  - Many parameters.
- **Solution** : Reduced ASM1 Model
  - 5 species
  - Still 21 parameters.

Table 2 **Process kinetics and stoichiometry for carbon oxidation, nitrification, and dentrification**

Component → / i j / Process ↓		1 S_I	2 S_S	3 X_I	4 X_S	5 $X_{B,H}$	6 $X_{B,A}$	7 X_P	8 S_O	9 S_{NO}	10 S_{NH}	11 S_{ND}	12 X_{ND}	13 S_{ALK}	Process Rate, ρ_j $[ML^{-3}T^{-1}]$
1	Aerobic growth of heterotrophs		$-\frac{1}{Y_H}$			1			$-\frac{1-Y_H}{Y_H}$		$-i_{XB}$			$-\frac{i_{XB}}{14}$	$\hat{\mu}_H\left(\frac{S_S}{K_S+S_S}\right)\left(\frac{S_O}{K_{O,H}+S_O}\right)X_{B,H}$
2	Anoxic growth of heterotrophs		$-\frac{1}{Y_H}$			1				$-\frac{1-Y_H}{2.86Y_H}$	$-i_{XB}$			$\frac{1-Y_H}{14\cdot 2.86Y_H} - i_{XB}/14$	$\hat{\mu}_H\left(\frac{S_S}{K_S+S_S}\right)\left(\frac{K_{O,H}}{K_{O,H}+S_O}\right) \times \left(\frac{S_{NO}}{K_{NO}+S_{NO}}\right)\eta_g X_{B,H}$
3	Aerobic growth of autotrophs						1		$-\frac{4.57-Y_A}{Y_A}$	$\frac{1}{Y_A}$	$-i_{XB}-\frac{1}{Y_A}$			$-\frac{i_{XB}}{14}-\frac{1}{7Y_A}$	$\hat{\mu}_A\left(\frac{S_{NH}}{K_{NH}+S_{NH}}\right)\left(\frac{S_O}{K_{O,A}+S_O}\right)X_{B,A}$
4	'Decay' of heterotrophs				$1-f_P$	-1		f_P					$i_{XB}-f_P i_{XP}$		$b_H X_{B,H}$
5	'Decay' of autotrophs				$1-f_P$		-1	f_P					$i_{XB}-f_P i_{XP}$		$b_A X_{B,A}$
6	Ammonification of soluble organic nitrogen										1	-1		$\frac{1}{14}$	$k_a S_{ND} X_{B,H}$
7	'Hydrolysis' of entrapped organics		1		-1										$k_h \frac{X_S/X_{B,H}}{K_X+(X_S/X_{B,H})}\left[\left(\frac{S_O}{K_{O,H}+S_O}\right) + \eta_h\left(\frac{K_{O,H}}{K_{O,H}+S_O}\right)\left(\frac{S_{NO}}{K_{NO}+S_{NO}}\right)\right]X_{B,H}$
8	'Hydrolysis' of entrapped organic nitrogen											1	-1		$\rho_7(X_{ND}/X_S)$
Observed Conversion Rates $[ML^{-3}T^{-1}]$		$r_i = \sum_j \nu_{ij}\rho_j$													
Stoichiometric Parameters: Heterotrophic yield: Y_H; Autotrophic yield: Y_A; Fraction of biomass yielding particulate products: f_P; Mass N/Mass COD in biomass: i_{XB}; Mass N/Mass COD in products from biomass: i_{XP}		Soluble inert organic matter $[M(COD)L^{-3}]$	Readily biodegradable substrate $[M(COD)L^{-3}]$	Particulate inert organic matter $[M(COD)L^{-3}]$	Slowly biodegradable substrate $[M(COD)L^{-3}]$	Active heterotrophic biomass $[M(COD)L^{-3}]$	Active autotrophic biomass $[M(COD)L^{-3}]$	Particulate products arising from biomass decay $[M(COD)L^{-3}]$	Oxygen (negative COD) $[M(-COD)L^{-3}]$	Nitrate and nitrite nitrogen $[M(N)L^{-3}]$	$NH_4^+ + NH_3$ nitrogen $[M(N)L^{-3}]$	Soluble biodegradable organic nitrogen $[M(N)L^{-3}]$	Particulate biodegradable organic nitrogen $[M(N)L^{-3}]$	Alkalinity – Molar units	Kinetic Parameters: Heterotrophic growth and decay: $\hat{\mu}_H$, K_S, $K_{O,H}$, K_{NO}, b_H; Autotrophic growth and decay: $\hat{\mu}_A$, K_{NH}, $K_{O,A}$, b_A; Correction factor for anoxic growth of heterotrophs: η_g; Ammonification: k_a; Hydrolysis: k_h, K_X; Correction factor for anoxic hydrolysis: η_h

Figure – Matrix representation of the ASM1 model.

Simplified Physics-Based Model :

$$\frac{\partial S_{NH}(t)}{\partial t} = \underbrace{\frac{Q^{in}(t)}{V}\left[S_{NH}^{in}(t) - S_{NH}(t)\right]}_{\text{Not modified}} - \underbrace{\beta u(t)\frac{S_{NH}(t)}{S_{NH}(t)+K}}_{\text{Modified part}}$$

where V, β, and K are **unknown** parameters.

Key Remarks :

- **Simple** yet effective model.
- Few parameters, lacking direct **physical interpretation**.
- Based on multiple **assumptions** whose validity is **uncertain**.

Black-box Model : Linear State Space Model vs. Local Linear Regression

Linear State Space Model

$$S(t+1) = AS(t) + Bu(t) + CE(t) + \eta_t$$

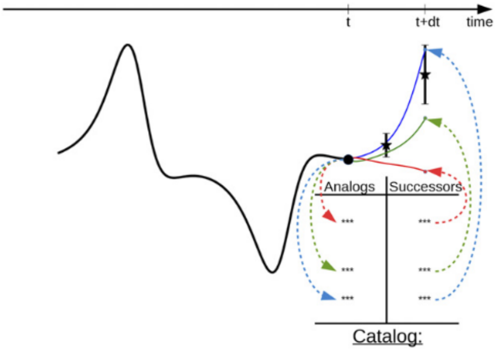
$$y(t) = HS(t) + \epsilon_t$$

Key Properties :

- **No physical meaning.**
- Assumes **global linearity**.
- Only based on **data-driven** methods.

Local Linear Regression

- Linear Regression on the k nearest neighbors.



Model Comparison

White-box

- **Based on physical laws.**
- **Requires many parameters** (often hard to estimate).
- Large number of species.
- High interpretability but may be too rigid.

Grey-box

- Uses simplified **mechanistic models** with data-driven corrections.
- **Reduces the number of parameters** compared to White-box.
- Trade-off between interpretability and flexibility.

Black-box

- **Fully data-driven.**
- Uses statistical or machine learning models.
- **No physical meaning**, purely empirical.
- Flexible but requires a lot of data and may lack generalization.

How to reconstruct the hidden concentrations and quantify the uncertainty ?

State Space Model structure

Separation of the model into two parts

- **Transition equation :**

$$S[t+h] = \text{ChoosenModel}(S[t], U[t], E[t]) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \Sigma_\eta)$$

- **Observation equation :**

$$Y_{NH}[t] = H(S[t]) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon)$$

How can I find **jointly** the **hidden** variables, the **parameters** of the model and the **uncertainties** ?

Methods for Uncertainty Quantification and Latent Space Reconstruction

White-box

- **Main interest :** Latent states reconstruction using observers (Gernaey et al., 2004).
- **Parameter Estimation :** Often neglected due to model complexity (Chachuat, 2001).
- **Uncertainty Handling :** Rarely integrated directly ; relies on a posteriori sensitivity analysis (Boulkroune, 2008).

Grey-box

- **Hybrid Calibration :** Combines mechanistic models with data-driven parameter tuning (Stare et al., 2006).
- **Uncertainty Handling :** Few models explicitly calibrate both parameters and uncertainty (Stentoft et al., 2018).
- **Key Challenge :** Ensuring consistency between physics-based and ML components.

Black-box

- **Fully Data-driven :** Requires large datasets ; lacks physical interpretability.
- **Uncertainty Estimation :** Difficult to quantify and integrate into assimilation frameworks (Alvi et al., 2023).
- **Recent Advances :** Local linear regression improves performance while reducing complexity (Chau et al., 2021).

Outline

1. Wastewater Treatment Plants
2. Possible models for ammonium concentration
3. Model's calibration with data assimilation
4. Simulated Results
5. First real results

Grey-box calibration example

- **Transition equation :**

$$S_{NH}[t+h] = S_{NH}[t] + h\left[\frac{Q^{in}[t]}{V}\left[S_{NH}^{in}[t] - S_{NH}[t]\right] - \beta u[t]\frac{S_{NH}[t]}{S_{NH}[t]+K}\right] + w_t, \quad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$$

- **Observation equation :**

$$Y_{NH}[t] = S_{NH}[t] + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

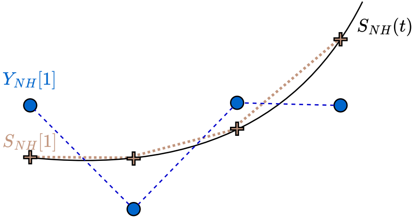
Majority of studies assume no observation noise during training.

Standard training : Prediction Error Minimization (PEM)

Training Dataset : $(Y_{NH}[i], \dots, Y_{NH}[N])$

Prediction Error Minimization : replace $S_{NH}[i]$ by observations $Y_{NH}[i]$

$$min_{V,\beta,K} \sum_{t=1}^{N-1} \left[Y_{NH}[t+1] - \left(Y_{NH}[t] + h \left[\frac{Q^{in}[t]}{V} \left[S^{in}_{NH}[t] - Y_{NH}[t] \right] - \beta u[t] \frac{Y_{NH}[t]}{Y_{NH}[t] + K} \right] \right) \right]^2$$



How to **adapt** to **hidden** variables and **observation** noise ?

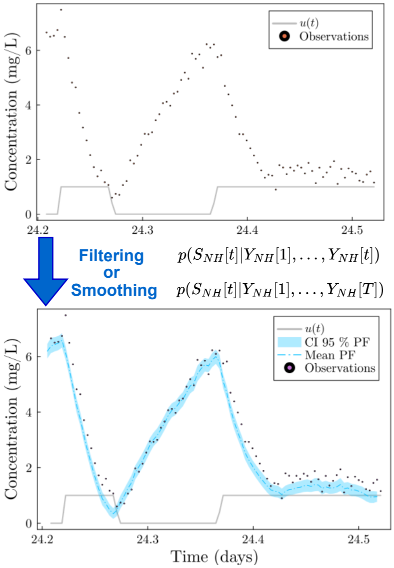
Data Assimilation

Data assimilation integrates real-world observations into models.

Key Challenges :

- **Reconstructing Hidden States :**
  - **Sequential Methods :** More time-consuming but provide stronger theoretical guarantees.
  - Variational Methods : Minimization of a cost function.
- **Calibrating Model Parameters and Uncertainties :**
  - Direct Likelihood Maximization : Involves challenging steps such as differentiating over sequential methods (which can be unstable).
  - Maximizing the complete likelihood with the **Expectation-Maximization** algorithm.

Filtering and Smoothing : Estimation of the latent variables

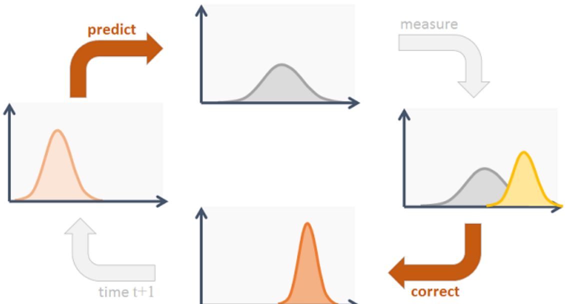


Condition : Known parameters.

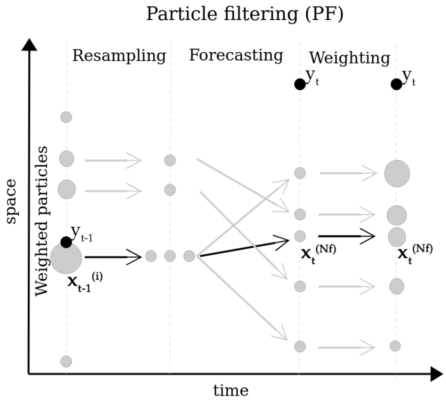
- **Linear model**
  - Kalman Filter and Smoother
- **Nonlinear model**
  - ◇ Linearization
    - Extended Kalman Filter and Smoother
  - ◇ Sequential Monte Carlo
    - Ensemble Kalman Smoother and Filter
    - Particle Filter and Backward Smoother

Kalman Filter and Particle Filter

Kalman Filter



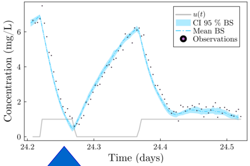
Particle Filter



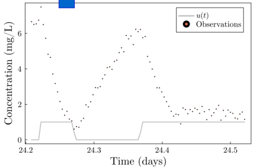
Monte-Carlo Expectation Maximization (MCEM)

$$\hat{S}_{NH}[1], \ldots, \hat{S}_{NH}[T]$$

E-Step



Filtering => Smoothing



M-Step

$$\hat{w}[t] = \left[\hat{S}_{NH}[i+1] - \left(\hat{S}_{NH}[i] + h\left[\frac{Q^{in}[t]}{V}[S^{in}_{NH}[t] - \hat{S}_{NH}[i]] - \beta u[t]\frac{\hat{S}_{NH}[i]}{\hat{S}_{NH}[i] + K}\right]\right)\right]$$

$$\hat{v}[t] = Y_{NH}[t] - \hat{S}_{NH}[t]$$

Residual's computation using smoothed values

$$min_{V,\beta,K,\sigma_m,\sigma_\epsilon} \sum_{i=1}^{N-1} \left[\log(\sigma_m^2) + \frac{\hat{w}[i]^2}{\sigma_m^2}\right] + \left[\log(\sigma_\epsilon^2) + \frac{\hat{v}[i]^2}{\sigma_\epsilon^2}\right]$$

Maximization of likelihood

$$V, \beta, K, \sigma_\epsilon, \sigma_m$$

Outline

1. Wastewater Treatment Plants
2. Possible models for ammonium concentration
3. Model's calibration with data assimilation
4. Simulated Results
5. First real results

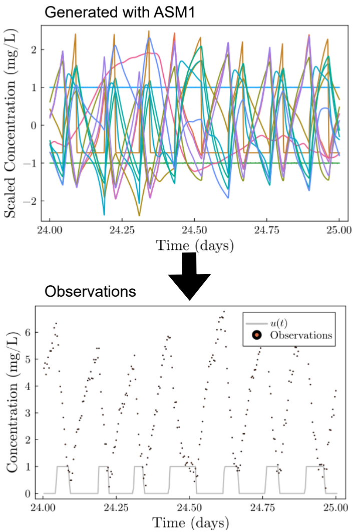
Experimental Setup

Objective :

- **Evaluate** model performance.
- Assess the **quality** of **uncertainty calibration**.

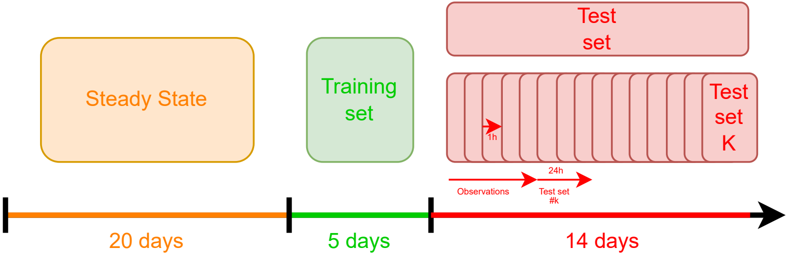
Experimental Design :

- Data generated from a simulator based on **ASM1**.
- ORP-based aeration control.
- **Diverse** external conditions.
- **Training :** 5 days | **Testing :** next 14 days.
- **Repeated** 10 times for each condition.



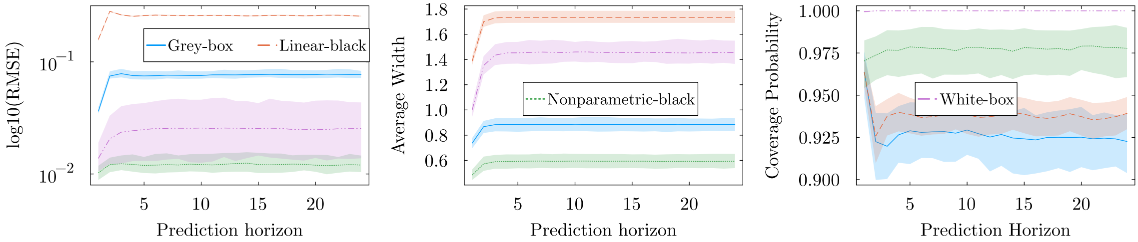
Metrics

New **observations** are assimilated **every hour**, and **24-hour predictions** are made.



- **Root Mean Squared Error (RMSE)** : Quantifies the average magnitude of errors between predicted and actual values.
- **Average Width (AW)** : Represents the mean width of prediction intervals.
- **Coverage Probability (CP)** : Measures the proportion of times the true value lies within the predicted interval.

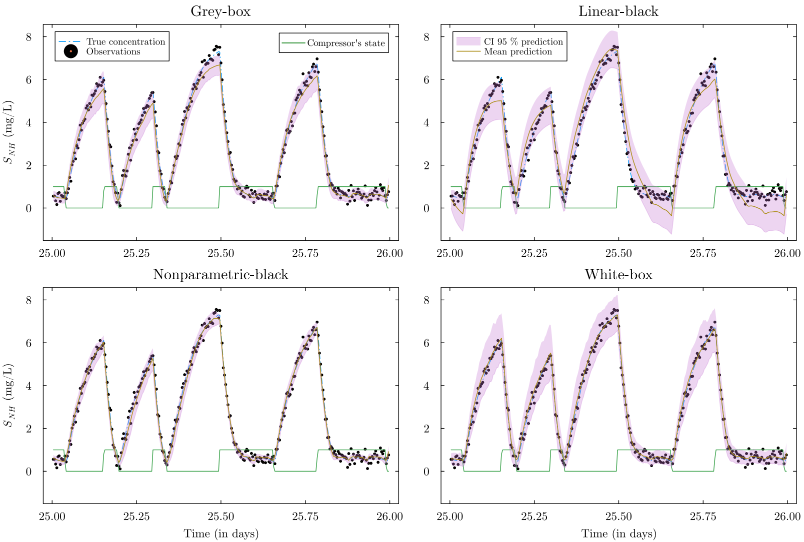
RMSE, AW, and CP Over Time Horizon



Key Observations :

- The **linear model** has the highest RMSE, while **LLR** achieves the best RMSE and AW.
- **White-box model** shows low RMSE but wide CI, indicating training instability.
- Most models are well-calibrated, except for the **white-box model**.

24h Model Predictions



Key Observations :

- **Linear model** fails to capture saturation at 0.
- **Grey-box model** shows bias at high concentrations.
- **White-box model** is well-calibrated but has wide confidence intervals.
- **LLR** achieves the best overall performance.

Estimated Parameters

Simulated observation standard deviation : 0.2 mg/L.

Model	Training Time (seconds)	Estimated Observation Standard Deviation	Estimated s_{NH} Standard Deviation
Grey-box	813	0.184	0.103
Linear Black-box	**162**	0.162	0.180
Nonparametric Black-box	3990	**0.197**	**0.062**
White-box	294000	0.189	0.109

Key Observations :

- All models provide accurate estimates of the observation standard deviation.
- **Linear model** is the fastest, while the **white-box** is extremely computationally expensive.
- **Nonparametric Black-box** achieves the lowest standard deviation for s_{NH}.

General Observations

Grey-box Model

- **Balanced** performance according to simplifications.
- Slight bias in high ammonium concentrations.
- Low computational cost.

Black-box Model

- **Best RMSE and AW** and well-calibrated confidence intervals for LLR.
- Linear version lacks expressivity (fails to model saturation).

White-box Model

- **Low RMSE** but large uncertainty intervals.
- Computationally expensive (long training time).
- **Unstable** training.

Outline

1. Wastewater Treatment Plants
2. Possible models for ammonium concentration
3. Model's calibration with data assimilation
4. Simulated Results
5. First real results

First Real-World Experiment

Experimental Setup :

- **Period :** April 12 – May 8, 2023.
- **Location :** Acigné, France.
- **Facility :** Wastewater treatment plant (14,000 PE).
- **Sampling Frequency :** Every minute.

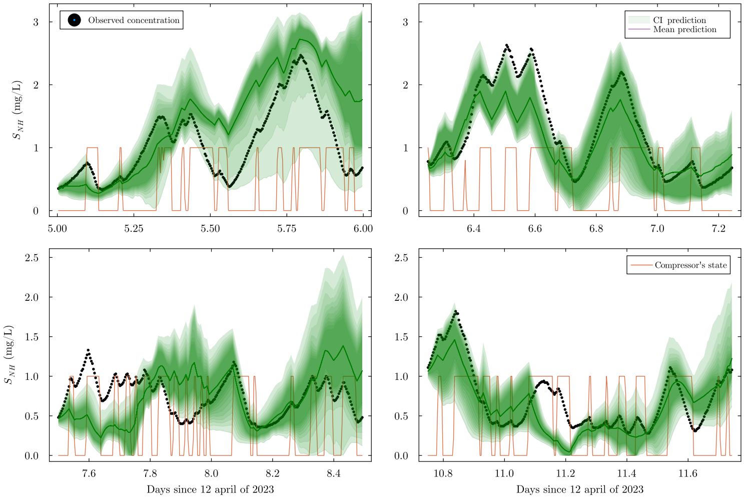
Available Data :

- **Inflow rate.**
- **Ammonium sensor measurements** (NH_4^+).
- No direct inlet concentration data.

Objective

Evaluate the performance of the non-parametric black-box model on real-world data.

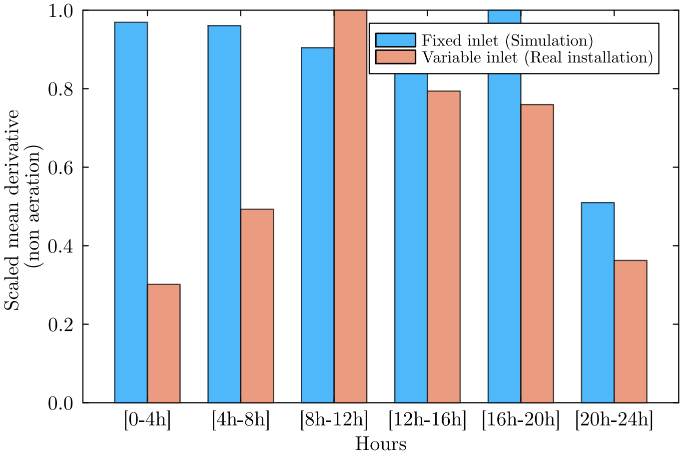
24h Model Predictions on Real Data



Key Observations :

- **Trend vs Amplitude :** Model struggles more with amplitude than trend.
- **Uncertainty Calibration :** Increased uncertainty, indicating good calibration.
- **Limitations :** Assumption of constant inlet concentrations is **violated** in real data.

Inlet Diurnal Variation in Catalog



Keys observations :

- **High Variability :** Significant fluctuations in the derivative within LLR catalogs.
- **Model Adjustment :** Need to incorporate **additional covariates** to improve performance.

Conclusion

- **Research Contributions** :
  - **Comparison** of various model types for ammonium prediction within a **data assimilation** framework.
  - The **non-parametric** black-box model outperformed all other models in terms of predictive **accuracy** and **uncertainty estimation**.
  - General trends were captured in **real data**, but performance still **needs improvement**.
- **Research Perspectives** :
  - Integration of heteroscedastic noise.
  - Enhancement of LLR on real data using covariates (e.g., rain, time of day, etc.).
  - Inlet concentration estimation.

Thank you for your attention

Questions ?

References I

Alvi, M., Batstone, D., Mbamba, C. K., Keymer, P., French, T., Ward, A., Dwyer, J., and Cardell-Oliver, R. (2023). Deep learning in wastewater treatment : a critical review. *Water Research*, 245 :120518.

Boulkroune, B. (2008). *Estimation de l'état des systèmes non linéaires à temps discret. Application à une station d'épuration*. PhD thesis, Université Henri Poincaré - Nancy I.

Chachuat, B. (2001). *Methodologie d'Optimisation Dynamique et de Commande Optimale des Petites Stations d'Epuration a Boues Activees*. PhD thesis, Institut National Polytechnique de Lorraine (INPL).

Chau, T. T. T., Ailliot, P., and Monbet, V. (2021). An algorithm for non-parametric estimation in state–space models. *Computational Statistics & Data Analysis*, 153 :107062.

Gernaey, K. V., van Loosdrecht, M. C., Henze, M., Lind, M., and Jørgensen, S. B. (2004). Activated sludge wastewater treatment plant modelling and simulation : state of the art. *Environmental Modelling and Software*, 19(9) :763–783.

Stare, A., Hvala, N., and Vrečko, D. (2006). Modeling, identification, and validation of models for predictive ammonia control in a wastewater treatment plant—a case study. *ISA Transactions*, 45(2) :159–174.

Stentoft, P. A., Munk-Nielsen, T., Vezzaro, L., Madsen, H., Mikkelsen, P. S., and Møller, J. K. (2018). Towards model predictive control : online predictions of ammonium and nitrate removal by using a stochastic asm. *Water Science and Technology*, 79(1) :51–62.