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

Victor Bertret, Roman Le Goff Latimier, Valérie Monbet

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

Victor Bertret[a,b,c,*], Roman Le Goff Latimier[b], Valérie Monbet[c]

[a] *Purecontrol, 68 Av. Sergent Maginot, Rennes, 35000, FRANCE*
[b] *SATIE, ENS Rennes, CNRS, Bruz, 35170, FRANCE*
[c] *IRMAR, Université Rennes 1, CNRS, Rennes, 35700, FRANCE*

Abstract

This study compares various modeling approaches to predict ammonium concentration in wastewater treatment plants (WWTPs), with a focus on integrating data assimilation techniques. It explores white-box, grey-box, and black-box models, evaluating their ability to capture the complex dynamics of WWTPs and manage uncertainties associated with limited data and sensor noise. The article highlights the importance of data assimilation for simultaneously calibrating model parameters, latent variables (such as unmeasured species concentrations), and quantifying prediction uncertainty. Simulation results demonstrate that the non-parametric black box model outperforms all other models in terms of predictive accuracy and uncertainty estimation. This finding underscores the effectiveness of machine learning when integrated with data assimilation techniques to extract insights from training datasets, even in the presence of limited data. Interestingly, the addition of an extra sensor, such as an oxygen sensor, did not enhance model performance. Experiments conducted in a real system showed that the non-parametric black box model could effectively capture the general dynamics of

Email addresses: victor.bertret@purecontrol.com (Victor Bertret), roman.legoff-latimier@ens-rennes.fr (Roman Le Goff Latimier), valerie.monbet@univ-rennes1.fr (Valérie Monbet)

ammonium concentration in an actual wastewater treatment plant. However, its performance was somewhat diminished compared to simulation results, likely due to variability in input concentrations that were not accounted for in the model.
Keywords: Activated Sludge, Data assimilation, Model Identification, Uncertainty Quantification, Expectation Maximization, Conditional Particle Filter

1. Introduction

Wastewater treatment plants (WWTPs) are complex, energy-intensive systems that must meet increasingly stringent regulations regarding effluent quality and energy efficiency. Meanwhile, energy prices and volatility have increased significantly in recent years (RTE, 2021). One of the highest energy demands at these facilities is the aeration process (Reardon, 1995), where, among other processes, nitrification takes place. Therefore, a significant challenge is to reduce the concentration of ammonium (NH_4^+), primarily originating from the incoming water, while simultaneously minimizing energy consumption. The Alternating Activated Sludge Plant (AASP) is a specific configuration of activated sludge plants (ASPs) that achieves complete nitrogen removal in a single bioreactor through nitrification and denitrification (Chachuat et al., 2005), as shown in Figure 1. It alternates between aeration (introducing oxygen (O_2) into the aeration tank with compressors) and non-aeration phases, requiring precise control to meet strict regulations, including daily average ammonium (NH_4^+) limits (Balku, 2007; Lukasse et al., 1999).

Given the complex dynamics and strict regulatory requirements of AASPs, modeling serves as a valuable approach for developing optimization strate-

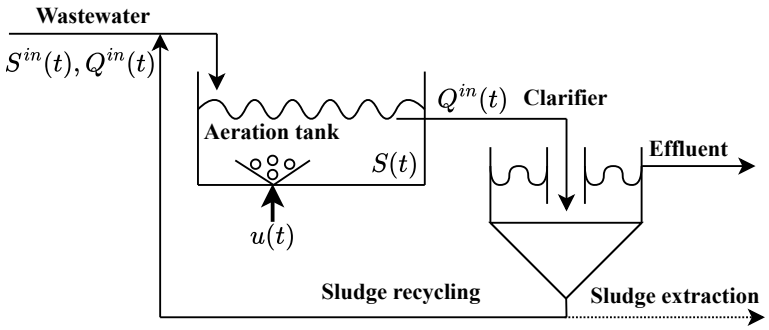

Figure 1: A schematic representation of the secondary treatment process in an AASP plant. The study focuses on quantifying the concentration of the species in the aeration tank, denoted by $S(t)$, which is affected by the influent wastewater concentration $S^{in}(t)$, the inflow rate $Q^{in}(t)$, and the control input $u(t)$, representing the state of the compressors.

gies. While deep reinforcement learning (DRL) has been explored as a control strategy for wastewater treatment plants (WWTPs) (Fu et al., 2022; Sutton and Barto, 2018), real-world applications often struggle with stability and reliability issues (Dulac-Arnold et al., 2019). In contrast, model-based approaches can provide more reliable control by providing a better understanding of system dynamics. However, these approaches encounter significant challenges in AASPs due to the complex interactions within the system and variability in influent composition, further complicated by limited and noisy sensor data (Lukasse et al., 1999). It is important to distinguish between estimating the varying influent composition and reconstructing the system's dynamics. Although influent estimation is well-documented in the literature (Andreides et al., 2022), it remains outside the scope of this study. Instead, our focus is on the reconstruction of system dynamics in ASPs, which presents several challenges:

- **Medium Dimensionality**: although there are many species to track, the number of variables is relatively moderate compared to fields like climate or oceanography.

- **High Interaction and Uncertainty**: the concentrations of species interact with each other in complex ways, influenced by unobserved factors such as pH, temperature, and sludge characteristics.

- **Partial Observation**: the system is monitored by only a few sensors, which are prone to noise, characterizing it as partially observed.

Unfortunately, addressing these challenges complicates model calibration. Effective model-based control necessitates accurate predictions of system behavior, particularly ammonium concentration, over a 24-hour period, accounting for inherent uncertainties. This raises two critical questions:

- **Model Calibration**: how can the parameters of the model and uncertainties be determined?

- **Data Assimilation**: how can the hidden or latent space of the model be reconstructed and retrieved?

While existing literature often focuses on either parameter estimation or state reconstruction, there are few comprehensive studies that tackle both challenges simultaneously. Some theoretical papers explore these issues in linear cases (Holmes, 2013) or nonlinear cases (Kantas et al., 2015), but practical applications are lacking. Moreover, most studies present a single model without comparing different approaches, or they focus solely on comparing inference algorithms.

In wastewater treatment, models like the Activated Sludge Model No. 1 (ASM1) are widely used to describe nitrification processes, but parameter tuning is often neglected due to the complexity and number of parameters involved (Henze et al., 2006). Instead, efforts typically focus on reconstructing the latent state, such as unmeasured species concentrations, based on the few available measurements. To make calibration more manageable, some researchers simplify the physical models, reducing the number of parameters to be estimated (Chachuat, 2001). However, even with simplifications, accurately capturing system dynamics and uncertainties remains challenging (Boulkroune, 2008). Other studies (Gernaey et al., 2004; Sánchez et al., 2002) have developed linear models from ASM1, but these models are only effective for very short-term predictions (more or less 2 hours).

Data-driven methods, like neural networks, learn directly from data without requiring explicit parameter calibration, yet they often struggle with accurately estimating uncertainties and adapting to changing conditions (Alvi et al., 2023, 2022). This complexity limits their integration into data assimilation frameworks, which are vital for estimating uncertainties from both model predictions and observations. In meteorology and oceanography, researchers initially relied on physical models but later incorporated data-driven techniques for better performance. However, challenges with sparse and noisy data complicate direct training (Bocquet, 2023). Recent advancements suggest that combining physical models with data-driven approaches can leverage the strengths of both, enabling better state reconstruction and adaptability, although issues with well-calibrated uncertainty estimates persist (Bonavita, 2024; Lam et al., 2023; Lguensat, 2023). In the context of

wastewater treatment, hybrid modeling (HM), which combines mechanistic models with data-driven approaches, has shown potential to improve prediction accuracy. For example, Verhaeghe et al. (Verhaeghe et al., 2024) developed a parallel hybrid model for water and resource recovery facilities (WRRFs), where a neural network was trained on the residuals of a mechanistic model for effluent nitrate. While such hybrid approaches improve model predictions, one limitation remains: the inability to effectively quantify uncertainty arising from both the mechanistic and data-driven components, as well as from the observations themselves. In another direction, in meteorology and oceanography, some methods experimented using local linear regression improving performance by reducing complexity (Chau et al., 2021).

Grey-box models combine data-driven and physics-based approaches, effectively integrating both methodologies. One method involving reduced physical models to estimate dynamics is effective if simplifications are valid (Gómez-Quintero and Queinnec, 2002; Lukasse et al., 1999; Stentoft et al., 2018). In wastewater treatment plants (WWTPs), researchers like Stentoft et al. (Stentoft et al., 2018) and Lukasse et al. (Lukasse et al., 1999) have utilized these models; however, many studies overlook uncertainty calibration. For instance, Stare et al. (Stare et al., 2006) and Yang et al. (Yang et al., 2021) employ grey-box models to simulate the nitrification process. However, their main focus is on integrating new data to estimate time-varying parameters or unobserved concentrations, rather than explicitly quantifying uncertainties for the construction of confidence intervals. As a result, they do not calibrate both model parameters and uncertainties simultaneously, but

instead determine uncertainty values through a posteriori tuning. Stentoft et al. (Stentoft et al., 2018) stand out for successfully calibrating both model parameters and uncertainties. Another approach, primarily explored in meteorology and oceanography, simplifies data-driven models by integrating data-driven techniques into specific model components or throughout the data assimilation process (Fablet et al., 2021; Sacco et al., 2024), though these often rely on spatio-temporal data characteristics not directly applicable to WWTP dynamics.

This overview highlights a significant gap in the literature regarding the joint solution of model calibration and data assimilation. To address these gaps, this study aims to explore and compare a range of models, from purely data-driven to physics-based (mechanistic), to determine the most effective approach for 24-hour prediction of ammonium, limited to low-data-regime methods. The integration of data assimilation techniques is also pursued to estimate model parameters, uncertainties, and latent variables, thereby providing a robust framework for uncertainty quantification in WWTPs. In this study, an ammonium (NH_4^+) sensor will be continuously available, and an oxygen sensor will be provided for one experimentation. Although the focus is on WWTPs characterized by a medium-dimensional latent space, high interaction, and uncertainty, as well as partial observation, the methodologies developed in this study can be extended to other complex industrial systems with unknown models and uncertainties. The study first presents a detailed overview of the proposed models and the selected calibration techniques. Results are discussed, starting with simulated data to validate the models, followed by real-world data from a French wastewater treatment plant to

assess practical applicability.

2. Materials and methods

2.1. Model's choice

To compare different approaches, models are categorized into three types: white-box, grey-box, and black-box. Before detailing these categories, the general structure of the models is outlined. Assume that N data points have been collected, denoted as $\forall i \in [1, \ldots, N]$:

- $S[i] = (s_1[i], \ldots, s_{n_x}[i]) \in \mathbb{R}^{n_x}$, representing the state of the system at time t_i, which, in this study is the species concentrations,
- $Y[i] = (y_1[i], \ldots, y_{n_y}[i]) \in \mathbb{R}^{n_y}$, representing the observations of the system at time t_i, which, in this study is the data collected from sensors,
- $U[i] = (u_1[i], \ldots, u_{n_u}[i]) \in \mathbb{R}^{n_u}$, representing the control signals at time t_i, which, in this study is the known aeration control,
- $E[i] = (e_1[i], \ldots, e_{n_e}[i]) \in \mathbb{R}^{n_e}$, representing the measured exogenous signals at time t_i, which, in this study is the inflow and inlet concentrations.

To separate observational noise from model noise in data assimilation, a discrete-time state space formulation is employed with two equations: the observation equation and the transition equation. By denoting $\mathcal{N}(\mu, \Sigma)$ as the multivariate Gaussian distribution with mean μ and covariance Σ, these equations can be expressed as:

$$S[i+1] = \mathcal{M}_i(S[i], U[i], E[i]) + w_i \tag{1a}$$

$$Y[i] = H_i(S[i]) + v_i \tag{1b}$$

$$v_i \sim \mathcal{N}(0, \Sigma_\epsilon), \quad w_i \sim \mathcal{N}(0, \Sigma_m) \tag{1c}$$

where Σ_ϵ and Σ_m are covariance matrices for the observation noise and the model noise, respectively. The timestep of the model and observation can differ, but for the sake of readability, it is considered to be the same. This state space approach allows for modeling any dynamical system generically, representing any system in this form (Durbin and Koopman, 2012). The purpose of this section is to define the functions $\mathcal{M}_i$ and H_i for each model. For the rest of the paper, Σ_ϵ and Σ_m are considered as diagonal matrices. However, the methodology can be extended to alternative covariance structures, including time-dependent formulations influenced by various factors (Cheng et al., 2023).

2.1.1. White-box model

The first model, introduced by Chachuat et al. (Chachuat, 2001) and used in various studies (Boulkroune, 2008; Chachuat, 2001), is a simplified version of the ASM1 model (Henze et al., 2006). This model, specifically developed for AASP, removes slow dynamics such as sludge weight variation and comprises five states:

$$S[i] = (x_{COD}[i], s_O[i], s_{NO}[i], s_{NH}[i], s_{ND}[i]).$$

These states represent the mass of COD (Chemical Oxygen Demand) and the concentrations of oxygen (O), nitrite/nitrate (NO), ammonium (NH),

and organic nitrogen (ND), respectively. The model maintains the structure of ASM1, which describes nitrogen processing in a Continuously Stirred Tank Reactor (CSTR) with a constant volume V and a continuous inflow rate $Q_{in}[i]$. Using the notation $\dot{S}[i]$ to define the time derivative of $S[i]$, the concentration evolution within the reactor is described by:

$$\dot{S}[i] = \frac{Q^{in}[i]}{V}(S^{in}[i] - S[i]) + r(i, S[i], U[i]) + w_i \tag{2}$$

where $E[i] = (Q_{in}[i], S^{in}[i])$ denotes the inflow and inlet water concentrations with $S^{in}[i] = (s^{in}_{NH}[i], \ldots, s^{in}_{ND}[i])$, and $U[i] \in \{0, 1\}$ represents the aeration control. The ASM1 model mathematically represents the function $(i, S, U) \mapsto r(i, S[i], U[i])$, detailing the reactions in the reactor tank. Detailed descriptions of this function can be found in the references (Henze et al., 2006, 2008; Metcalf & Eddy Inc. et al., 2013). Chachuat et al. (Chachuat, 2001) simplified the latent space $S[i]$ from 13 to 5 states and modified the structure of the function $r(i, S[i], U[i])$.

For the remainder of this article, this model will be referred to as the white-box model. Our objective is to estimate the model's parameters, initial conditions, and the concentrations within the latent space $S(t)$ to closely replicate the dynamics observed in the actual tank. However, as detailed in the previous section, this is not a straightforward task. In current literature, no attempts have been made to estimate all these components simultaneously using this model. For example, Boulkroune et al. (Boulkroune, 2008) and Chachuat et al. (Chachuat, 2001) estimated only a subset of model's parameters, fixing others within the latent space to simplify the problem. Moreover, as identified by Boulkroune et al. (Boulkroune, 2008), the model

is only observable under certain conditions, particularly when variables such as oxygen, nitrate, and ammonium concentrations are being monitored.

2.1.2. Grey-box model

To further simplify the model and enhance observability, some research has focused specifically on the dynamics of interest (Bertret et al., 2024; Lukasse et al., 1999; Stentoft et al., 2018). For instance, if the primary concern is the concentration of ammonium (s_{NH}), the latent space can be reduced to include only ammonium concentration. This approach targets the relevant dynamics and eliminates dependencies on unmeasured variables. The models proposed by Lukasse et al. and Stentoft et al. (Lukasse et al., 1999; Stentoft et al., 2018) are based on ammonium and nitrate concentrations, while recent work focused solely on ammonium (Bertret et al., 2024). In our case, given the primary focus on ammonium, the model considered in this section will be the one proposed in recent work (Bertret et al., 2024), defined as:

$$\dot{s}_{NH}[i] = \frac{Q^{in}[i]}{V}\left[s^{in}_{NH}[i] - s_{NH}[i]\right] - \beta U[i]\frac{s_{NH}[i]}{s_{NH}[i] + K} + w_i \tag{3}$$

where K, β, and V are model's parameters. This model will be referred to as the grey-box model as it's a very simplified physical model where the model's parameters have no more physical meaning and will be determined by data.

However, while most studies attempt to estimate all the model's parameters (Lukasse et al., 1999; Stentoft et al., 2018), few consider calibrating uncertainties. In the literature, to the best of our knowledge, only Lukasse

et al. (Lukasse et al., 1999) and Bertret et al. (Bertret et al., 2024) have addressed uncertainty calibration.

2.1.3. Black-box model

Finally, the first two approaches are compared to models that are purely data-driven. Two types of data-driven models are explored. The first one involves using a linear state space model where the latent space dimension is equal to the number of observed variables:

$$S[i+1] = S[i] + AS[i] + BU[i] + CE[i] + w_i. \tag{4}$$

The objective of this model is to determine whether the dynamics of the system can be adequately represented by a linear model, as suggested in the literature (Gernaey et al., 2004; Sánchez et al., 2002). It will be called linear black-box model.

The second data-driven model is inspired by the recent work of Chau et al. (Chau et al., 2021; Lguensat et al., 2017). In this case, a local linear regression (LLR) is employed to model the reactor's dynamics. This approach has demonstrated strong performance when the latent space is small (Chau et al., 2021), effectively reconstructing the hidden states. Mathematically, this model is expressed as:

$$S[i+1] = S[i] + \mathrm{LLR}(S[i], U[i], E[i]) + w_i \tag{5}$$

where LLR represents the output of the local linear regression. Detailed information on the methods and hyperparameters can be found in Chau et al. (Chau et al., 2021; Lguensat et al., 2017). Not going into the details, the

methods identify the k nearest neighbors of the current point, make a linear regression on the selected point and predict the projection of the current point using the linear regression. It will be called non-parametric black-box model. The advantages of these two black-box methods are their proven utility in parameter fitting and latent space reconstruction. However, the second method has not yet been applied to wastewater treatment plants and may offer a promising middle ground between linear models and artificial neural networks (ANN) when data availability is limited. Other models, such as artificial neural networks or Gaussian processes (GP), could also be considered within the black-box approach. However, LLR can be viewed as a sparse GP (Stulp and Sigaud, 2015), and in scenarios with limited data, ANN tends to generalize less effectively than LLR (see Supplementary Material and (Otchere et al., 2021)).

This section has presented the different types of models. However, the simultaneous estimation of the model's parameters and reconstruction of the latent space has not yet been addressed. This will be the focus of the next section.

2.2. Model's calibration

In the context of data assimilation, several challenges arise concerning inference (Durbin and Koopman, 2012):

- reconstructing hidden states (the focus of most studies);
- calibrating model's parameters and uncertainties (addressed in fewer studies).

There are various solutions to the first challenge. Variational methods and sequential methods are commonly employed (Carrassi et al., 2018). Sequential methods, including the Kalman Filter, were developed first. However, these methods can become computationally expensive for high-dimensional systems, which led to the development of variational methods (Sasaki, 1970). Variational methods are primarily used in high-dimensional systems, such as those found in oceanography or meteorology, and work by minimizing a cost function (Carrassi et al., 2018; Sasaki, 1970). This study focuses on systems of medium to small dimensions, emphasizing sequential methods, which offer stronger theoretical guarantees.

For linear systems, the Kalman Filter and Kalman Smoother provide a closed-form solution. However, in nonlinear systems, adaptations are necessary. These adaptations fall into two categories:

- Linearization techniques, such as the Extended Kalman Filter (EKF) (Anderson and Moore, 2012) and the Unscented Kalman Filter (UKF).
- Sequential Monte Carlo methods, including the Ensemble Kalman Filter (EnKF) (Anderson and Moore, 2012; Evensen and Leeuwen, 2000) and the Particle Filter (PF) (Doucet and Johansen, 2011; Kantas et al., 2015).

While linearization methods can be faster, they may not be suitable when non-linearities between timesteps are significant, leading to less accurate solutions, as indicated by various studies (Dreano et al., 2017; Stentoft et al., 2018). Sequential Monte Carlo methods (Doucet et al., 2001; Doucet and Johansen, 2011) estimate the target distribution using particles. The EnKF

and Ensemble Kalman Smoother (EnKS) assume that the state is Gaussian at each timestep, while Particle Filters and their associated smoothers are not restricted to Gaussian distribution. Although Particle Filters is based on theorical results that provide more generality and flexibility than the EnKF, they usually require more particles than the EnKF for equivalent performance in estimating the posterior mean, making them slower (Chau et al., 2021). To address this, methods like the Conditional Particle Filter (CPF) have been developed to retain the benefits of Particle Filters while reducing the number of particles needed (Lindsten, 2013; Lindsten et al., 2014). The CPF involves constructing a Markov kernel whose asymptotic law is the posterior distribution of the hidden state. Instead of increasing the number of particles, this method iterates the kernel to obtain an accurate approximation of the distribution.

In summary, various tools for addressing the problem of reconstructing hidden states have been explored. However, the calibration of model parameters and uncertainties remains to be addressed. The goal is to calibrate these parameters in order to maximize the likelihood of the observations, expressed as:

$$L(Y;\theta) = p(Y[1],\ldots,Y[n];\theta) = p(Y[1])\prod_{i=2}^{N} p(Y[i]|Y[1:i-1];\theta) \quad (6)$$

where $Y[1:i-1] = (Y[1],\ldots,Y[i-1])$. Almost any problem can be framed as a differentiable program. Thus, one potential approach is to differentiate Sequential Monte Carlo methods to find the model's parameters that maximize the likelihood (Durbin and Koopman, 2012; Stentoft et al., 2018).

This approach, however, can be unstable and may result in challenging optimization problems. An alternative is to maximize the complete likelihood, incorporating both observations and hidden states. Since the hidden states are unknown, a common solution is to use the 'Expectation-Maximization' (EM) algorithm. This algorithm divides the problem into two steps: the 'E-Step,' where the posterior distribution of the hidden state is reconstructed using filters and smoothers, and the 'M-Step,' where the complete likelihood is optimized using observations and the predicted hidden states from the E-Step (Chau et al., 2021; Dreano et al., 2017; Lindsten, 2013). This approach is known as the Expectation-Maximization (EM) algorithm for linear systems (Dempster et al., 1977; Durbin and Koopman, 2012). Extensions of this method using Sequential Monte Carlo methods for the "E-Step", such as Stochastic Approximation Expectation-Maximization (SAEM) and Monte Carlo Expectation-Maximization (MCEM), have been developed for nonlinear systems (Celeux et al., 1995; Lindsten, 2013).

Chau et al. (Chau et al., 2021) compared EKF, EnKF, PF, CPF, and associated smoothers for parameter estimation in nonlinear models. Based on their conclusions, the focus will be on the Particle Filter (PF) and Conditional Particle Filter (CPF) associated with the Backward Smoother (BS) for nonlinear models due to their efficiency, generality, and performance. For linear models, the Kalman Filter and Smoother will be employed. These smoothers and filters will be combined with the EM or MCEM methods, depending on whether the model is linear or nonlinear, as these methods have demonstrated strong performance in parameter calibration (Chau et al., 2021). Table 1 summarizes the methods chosen for each model presented in

the previous section.

Table 1: Summary of calibration methods for different model types in data assimilation. The table presents the approaches used for hidden state estimation and parameter calibration across various models considered in this study. The number of parameters for each model, including both model parameters and uncertainties, is also provided, where n_Y represents the number of observations.

Model	**Hidden State Estimation**	**Parameter Calibration**	**# parameters**
Linear Black-Box	Kalman Filter and Smoother	EM	$2n_Y^2 + 4n_Y$
Grey-Box	PF and BS	MCEM	5
White-Box	CPF and BS	MCEM	$24 + n_Y$
Non-Parametric Black-Box	CPF and BS	MCEM	$2n_Y^2 + 4n_Y$

The methods outlined in this section allow us to reconstruct hidden states and calibrate model's parameters and uncertainties, a task that is rarely undertaken in the literature on wastewater treatment plants. These methods enable the calibration of any type of state-space model, followed by the refinement of hidden state estimates using observations. This approach is more robust and powerful than standard Prediction Error Minimization (PEM) methods (Ljung, 1998), as demonstrated in recent work (Bertret et al., 2024).

3. Results and discussion

As detailed, the primary objective of this study is to develop a model capable of making rolling predictions of ammonium concentration. Various strategies, ranging from purely physical models to data-driven models, have been presented. To achieve this, certain physical aspects of the system are

simplified to isolate the most relevant dynamics for modeling. These simplifications are necessary to ensure tractability and solvability given the limited data available, without compromising the system's essential chemical properties. Consequently, the models are trained on sparse datasets, such as one week of observations. This approach is driven by two key factors: first, simplifying the models by assuming constant concentrations over short time periods, such as the size of the dataset, allows for calibration on recent data; second, recent data more accurately captures the system's immediate dynamics, particularly when external conditions, such as influent composition or temperature, change rapidly. This strategy reflects practical constraints, where extensive experimentation and long-term data collection are often infeasible, highlighting the need for adaptable models that perform well with minimal data.

3.1. Simulation results

This section aims to compare all the presented models and demonstrate how data assimilation can effectively handle latent space estimation and provide uncertainty quantification. First, the models are compared on simulated data, and then the best-performing model is tested in a real installation in France in section 3.2. The analysis was conducted using the Julia programming language, and the experimental code is available [1]. As previously mentioned, real wastewater treatment plants are typically monitored with a limited number of noisy sensors. Most municipal plants are equipped with

[1]The code for reproducing results of this paper can be accessed at `https://github.com/vbertret/WWTP-DataAssimilation-ModelComparison`.

only one oxygen and/or redox sensor, and occasionally an ammonium sensor, all placed at a single location in the reactor tank. Consequently, the sensor readings may be noisy and might not accurately represent the conditions throughout the entire reactor. This study focuses on quantifying ammonium concentration, assuming the availability of an ammonium sensor, along with an additional oxygen sensor in some cases.

3.1.1. Experimental Setup

The primary goal of this study is to compare the models in terms of 24-hour horizon prediction, which is crucial for effective control. A fair comparison is ensured by testing the models under various conditions, including different types of training and test data, to provide robust results. To achieve this, the ASM1 model, as described earlier, is used to generate synthetic data. As previously noted, the estimation of inlet concentrations is left for a future study, so inlet concentrations and inflow data from the Benchmark Simulation Model (BSM) are utilized. The control of the aeration system is determined by a strategy resembling ORP (Oxydo Reduction Potential) control, where aeration is activated or deactivated based on thresholds for NH_4^+ and NO_3^- (Paul et al., 1998). To make the simulation results more realistic, noise is added to the observations, and training is performed on 50 independent realizations of the observation time series. The system is first simulated for 20 days to reach a steady state, followed by a training period ($N_{training}$ points) and a rolling test period (N_{test} points), with observations collected every 5 minutes.

The test period lasts 14 days, during which 24-hour horizon predictions are made iteratively. After each 1-hour interval, new observations are assim-

ilated, and the predictions are updated for the next 24 hours. This approach allows us to evaluate the models across a wide range of inflow and concentration conditions as illustrated on the Figure 2.

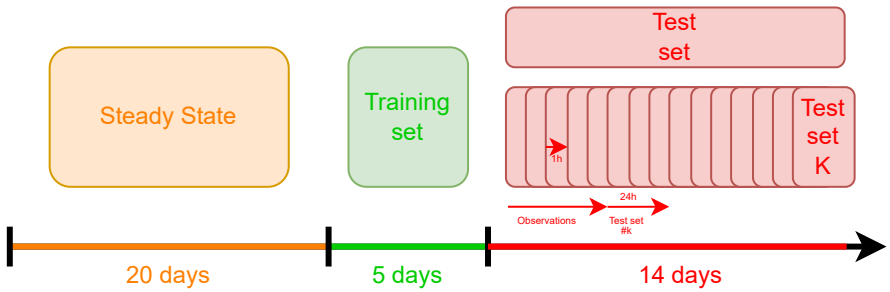

Figure 2: Dataset construction used in model training and testing. The system is first simulated to reach a steady state, followed by a training period of 5 days and a rolling test period of 14 days. Observations are assimilated every hour, and 24-hour horizon predictions are made iteratively.

3.1.2. Metrics

In this study, three distinct metrics are used to evaluate the model's performance: Root Mean Squared Error (RMSE), Average Width (AW), and Coverage Probability (CP). Each metric serves a specific purpose in assessing different aspects of the model's predictions, providing a comprehensive evaluation of both accuracy and reliability. Let $\hat{Y}[i|i-h]$, $\hat{Y}^{\alpha}_{LB}[i|i-h]$, and $\hat{Y}^{\alpha}_{UB}[i|i-h]$ represent, respectively, the mean prediction, the lower bound, and the upper bound of the model's $\alpha\%$ confidence interval at time i, given that the system is observed up to time $i-h$. Since predictions can be made at different horizons, all metrics are computed over a range defined by a starting horizon H_1 and an ending horizon H_2 (with $H_2 > H_1$), and the results are

averaged over this range.

RMSE (Root Mean Squared Error). This metric evaluates the accuracy of the model by measuring the difference between the predicted and true values. Specifically, it quantifies how well the model predicts concentrations at each time step. A lower RMSE indicates that the model is closer to the actual system behavior. For rolling predictions, the RMSE is computed as:

$$\mathcal{L}_{RMSE}(H_1, H_2) = \frac{1}{H_2 - H_1} \sum_{h=H_1}^{H_2} \sqrt{\sum_{i=H_2}^{N_{test}} \frac{||\hat{Y}[i|i-h] - Y[i]||_2^2}{N_{test} - H_2}} \tag{7}$$

RMSE provides a direct measure of the model's predictive accuracy by focusing on the deviation between predicted and observed concentrations.

AW (Average Width). This metric evaluates the precision of the model's predictions by assessing the width of the confidence intervals. A narrower interval indicates that the model is more certain about its predictions, while a wider interval implies more uncertainty. For rolling predictions, the AW is computed as:

$$\mathcal{L}_{AW}(H_1, H_2) = \frac{1}{H_2 - H_1} \sum_{h=H_1}^{H_2} \sum_{i=H_2}^{N_{test}} \frac{[\hat{Y}^{\alpha}_{UB}[i|i-h] - \hat{Y}^{\alpha}_{LB}[i|i-h]]}{N_{test} - H_2} \tag{8}$$

AW assesses the model's confidence in its predictions. A smaller interval indicates higher precision, while still accounting for uncertainty.

CP (Coverage Probability). The CP metric evaluates the reliability of the confidence intervals by checking how frequently the true concentration values fall within the predicted intervals. Ideally, the actual values should fall within

the interval as frequently as the interval was calibrated to (e.g., 95%). For rolling predictions, the CP is computed as:

$$\mathcal{L}_{CP}(H_1, H_2) = \frac{1}{H_2 - H_1} \sum_{h=H_1}^{H_2} \sum_{i=H_2}^{N_{test}} \frac{\mathbf{1}_{\{\hat{Y}^{\alpha}_{LB}[i|i-h] \leq Y[i] \leq \hat{Y}^{\alpha}_{UB}[i|i-h]\}}}{N_{test} - H_2} \tag{9}$$

CP measures how well-calibrated the model is in terms of its uncertainty estimates, ensuring the intervals contain the correct proportion of true values.

3.1.3. Forecasting performance

The primary goal of this experiment is to evaluate the forecasting performance of each model when only ammonium concentration is observed, reflecting a real-world scenario where sensor availability is limited due to costs and calibration requirements. The models were trained for 5 days using a fixed inlet concentrations and then tested over a 14-day period, as described earlier. Figure 3 illustrates the performance metrics across various prediction horizons.

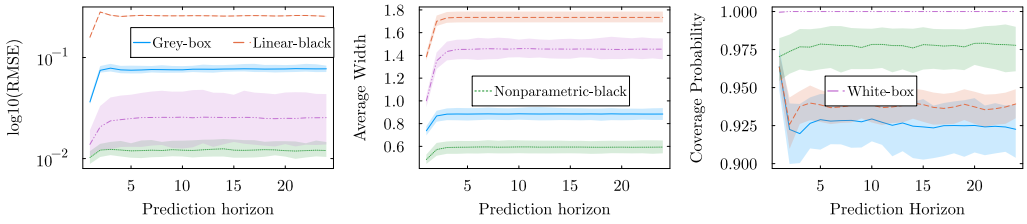

Figure 3: Comparison of model performance metrics. This figure presents the root mean squared error (RMSE), average width (AW), and coverage probability (CP) for different models across various prediction horizons (in hours). The value at $x = 5h$ corresponds to the average metric computed over the interval $\mathcal{L}_{\text{RMSE}}(4h, 5h)$, and similarly for the other metrics. The solid lines represent the mean values, while the shaded areas denote the 95% confidence intervals.

The results indicate that the linear model performs the worst in terms of RMSE, corroborating findings from previous studies (Gernaey et al., 2004; Sánchez et al., 2002). In contrast, the non-parametric black-box model significantly outperforms all others across all metrics. This counter-intuitive result will be discussed later, as it may seem surprising given that grey-box and white-box models contain more information about the system. The comparison between the white-box and grey-box models is less straightforward: while the white-box model shows a lower RMSE, its confidence intervals are considerably wider, suggesting higher variability and instability. This instability likely stems from the white-box model being initialized with parameter values close to those used in the simulator, resulting in good RMSE but poor training stability. Interestingly, no error accumulation is observed under these controlled conditions, likely due to the use of a simulator with limited and controlled sources of uncertainty. However, in real-world applications, error accumulation may arise due to higher uncertainty relative to ammonium dynamics as shown in Section 3.2. As seen in Figure 3, all models are well-calibrated with a coverage probability close to 95%, except for the white-box model, which suffers from overly broad confidence intervals. This finding underscores the effectiveness of our parameter and uncertainty calibration approach. Table 2 provides detailed information on the estimated standard deviations for the models and observations, as well as training times.

The linear model is the fastest to train, followed by the grey-box and non-parametric models. Although the non-parametric model requires 20 times more training time than the linear model and 5 times more than the grey-box, the difference remains manageable. However, the white-box model is

Table 2: Model training time in seconds and standard deviation estimates in mg/L. This table summarizes the training time, estimated standard deviation of observations, and estimated standard deviation of the model for each of the four models. The known standard deviation of observations was set to **0.2**.

Model	Training Time (seconds)	Observation Standard Deviation (mg/L)	s_{NH} Standard Deviation (mg/L)
Grey-box	8.13e+02	1.84e-01	1.03e-01
Linear-black	**1.62e+02**	1.62e-01	1.80e-01
Nonparametric-black	3.99e+03	**1.97e-01**	**6.23e-02**
White-box	2.94e+05	1.89e-01	1.09e-01

extremely time-consuming, taking 1850 times longer than the linear model and 375 times longer than the grey-box, making it impractical for real-world use in parameter and uncertainty calibration.

Regarding the estimation of observation noise, all models perform well, with values close to the true standard deviation of 0.2. Notably, the non-parametric model achieves a value of 0.197 and also shows the lowest standard deviation for the model itself. As highlighted in Table 2, the non-parametric model outperforms the others overall. Figure 4 showcases a single 24-hour prediction realization for each model.

The linear-black model, as expected, struggles with complex dynamics, such as saturation near zero concentration. The grey-box model delivers accurate predictions and reasonable confidence intervals, though it tends to be biased at high ammonium levels, possibly due to oversimplifications in the physical model. This bias is somewhat compensated by wider confidence intervals. The white-box model is well-calibrated in terms of mean predic-

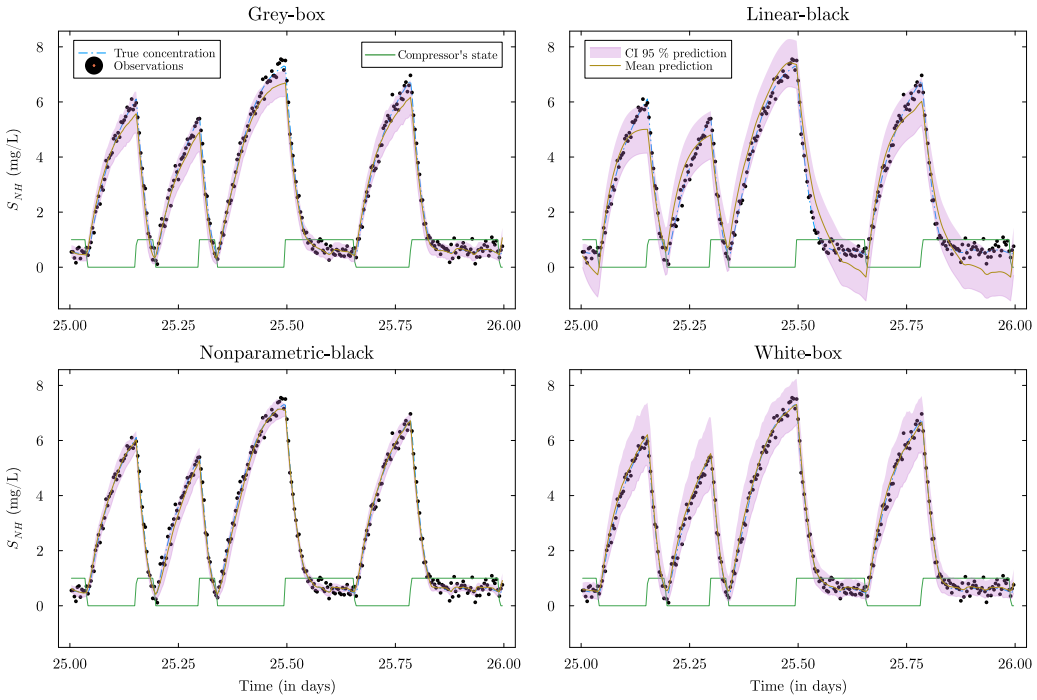

Figure 4: 24-hour prediction realizations for each model. This figure compares the predicted ammonium concentrations (solid lines) with the true concentrations (dots) for one realization of each model. The shaded areas represent the 95% confidence intervals.

tions but suffers from excessively broad confidence intervals, reflecting its instability. Lastly, the non-parametric black-box model demonstrates strong performance in both mean predictions and confidence intervals, effectively capturing the information from the training set.

In summary, the linear black-box model is fast and well-calibrated in terms of uncertainty but lacks the flexibility needed to capture the complex dynamics of a wastewater treatment reactor. The white-box model, while potentially powerful, is difficult to train, time-consuming, and highly dependent on good initial model's parameters, making it less generalizable. The

grey-box model offers a good balance between accuracy and computational efficiency, though some simplifications may degrade its performance. Finally, the non-parametric black-box model excels in both predictive accuracy and uncertainty estimation, highlighting its potential for robust modeling in wastewater treatment applications.

3.1.4. Added value of other concentration

Having compared the models using only ammonium concentration data, this section explores whether incorporating additional sensors, such as an oxygen sensor, can enhance model performance. Based on the findings from the previous section, the non-parametric black-box model emerged as the best-performing model, so this analysis will focus on that model alone. The experimental setup remains consistent with the first experiment, but now the model has access to both oxygen and ammonium concentration data. Table 3 presents a comparison of the model's performance when observing ammonium alone versus both ammonium and oxygen.

Table 3: Comparison of model performance with different observed species. This table compares the root mean squared error (RMSE), average width (AW), and coverage probability (CP) for the non-parametric black-box model when observing only ammonium (NH_4) versus observing both ammonium and oxygen (NH_4, O_2). The results are shown for ammonium prediction.

Observed species	Observed Standard Deviation	s_{NH} Standard Deviation	$\mathcal{L}_{RMSE}(1H, 24H)$	$\mathcal{L}_{AW}(1H, 24H)$	$\mathcal{L}_{CP}(1H, 24H)$
$[Y_{NH}]$	1.97e-01	**6.23e-02**	**1.21e-02**	**5.88e-01**	9.77e-01
$[Y_{NH}, Y_O]$	1.98e-01	9.05e-02	3.47e-02	7.94e-01	9.69e-01

The results indicate that the observation noise standard deviation is accurately estimated in both scenarios, validating the robustness of our un-

certainty estimation methodology. However, the model that observes only ammonium consistently outperforms the model that observes both ammonium and oxygen across all metrics. Specifically, the root mean squared error (RMSE), confidence interval width (AW), the coverage probability (CP) and model noise standard deviation are all lower when the model focuses solely on ammonium.

This counterintuitive outcome might be explained by the additional complexity introduced when the model must simultaneously reconstruct both species. The added noise in estimating oxygen appears to interfere with the accuracy of ammonium predictions. Figure 5 illustrates a 24-hour prediction realization for the model observing both oxygen and ammonium.

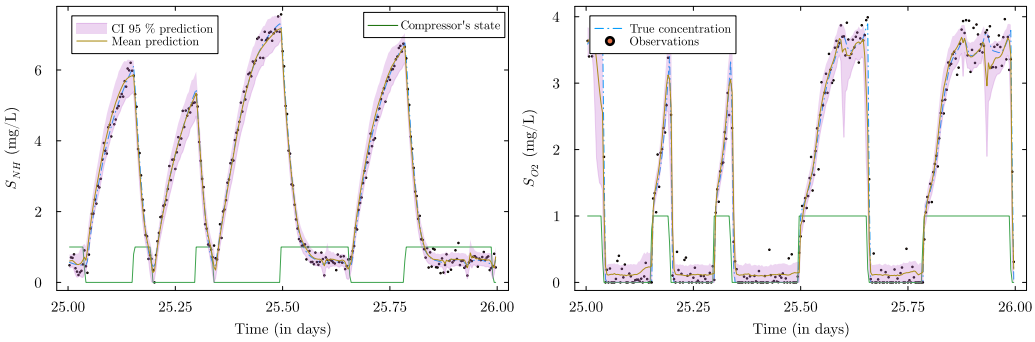

Figure 5: 24-hour prediction realizations for ammonium and oxygen. This figure compares the predicted and true concentrations of ammonium (left panel) and oxygen (right panel) for one realization of the non-parametric black-box model. The solid lines represent the mean predictions, the dotted line the true concentrations, and the shaded areas indicate the 95% confidence intervals.

The model successfully captures the dynamics of ammonium, though some errors that were not present in the single-species scenario become evident. The oxygen dynamics are also well-reconstructed; however, the as-

sumption of Gaussian noise for oxygen may not be optimal, suggesting that exploring alternative distributions could be beneficial in future studies. In conclusion, adding an additional sensor, such as oxygen, does not appear to improve model performance when the target species (ammonium) is already being observed. This finding suggests that, at least for the non-parametric black-box model, investing in additional sensors may not be justified if the primary species of interest is already monitored. However, it would be valuable to investigate whether observing other species could be beneficial when the target species is not directly observed—though such an exploration is beyond the scope of this paper.

3.2. Real system experiments

The non-parametric black-box model showed promising results on simulated data. However, to assess its performance in a real-world setting and ensure that the simulated data were not overly simplified, the model needed to be tested on real data. To achieve this, the non-parametric black-box model was applied to a real-world wastewater treatment plant located in Acigné, France, which serves a population equivalent (PE) of 14,000. Data collection spanned from April 12 to May 8, 2023, with the first five days used for training the model and the subsequent days for performance evaluation. Inflow and ammonium concentrations (NH_4^+) were recorded every minute. Since the inlet concentrations were not directly measured, it was excluded from the exogenous signals used as model inputs.

Figure 6 displays four 24-hour model predictions for the real dataset, each beginning at a different time of day. Although the real concentration data is unavailable, sensor readings were used for analysis. Compared to the simula-

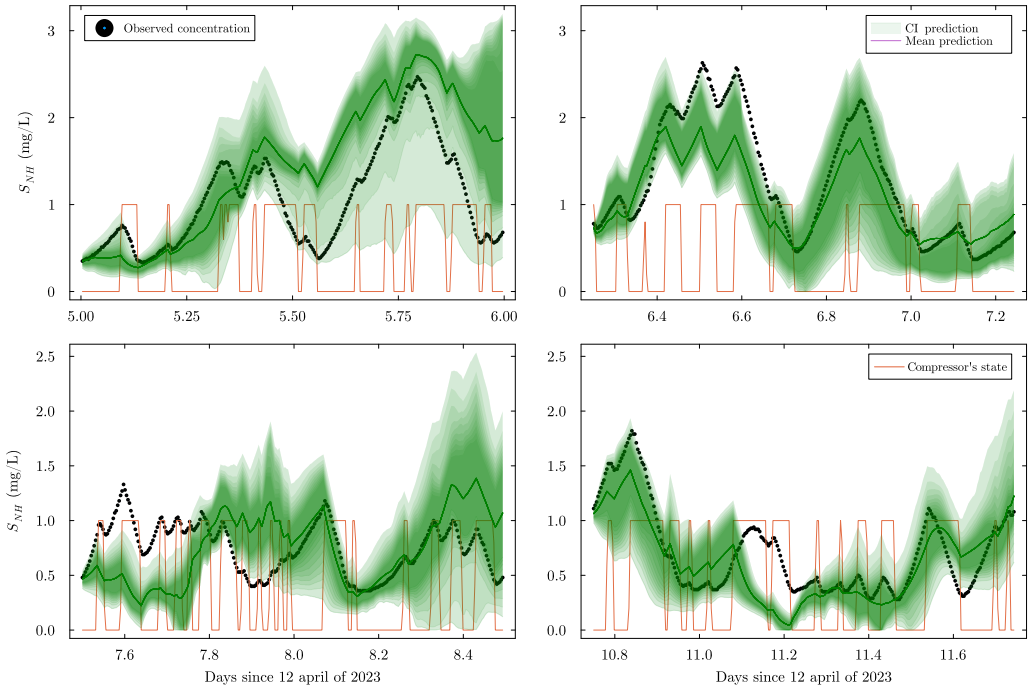

Figure 6: 24-hour model predictions on real system data, initiated at different times of the day. The black dots represent sensor readings of ammonium concentration (s_{NH}) in mg/L, while the green-shaded regions represent different percentages of confidence interval (CI) for the predicted values. Darker shading corresponds to lower percentages of values contained within the interval.

tion results, the real-world dynamics are less accurately captured. The model shows higher uncertainty, as reflected by the wider confidence intervals, which still manage to encapsulate the majority of the observations. Notably, the model's deviations are more pronounced in amplitude rather than in trend evolution, likely due to the assumption—borrowed from the simulation—that the inlet concentrations remain constant. This assumption does not hold true for the real system, where significant daily variations in inlet concentrations are observed (Gernaey et al., 2014). As the model lacks this critical

information, it must generalize across varying conditions, leading to higher confidence intervals despite generally capturing the underlying dynamics. To provide further insights into the predicted distribution, confidence intervals of varying percentages were included, highlighting the asymmetry in the intervals, as the distribution is non-Gaussian. For example, in the top left plot, although the 95% confidence interval is wide near the end of the prediction, the 90% interval is significantly narrower, demonstrating a key strength of the method used in this study. This behavior indicates that the model can still maintain tight predictions even when broader uncertainty exists for the outer confidence bands. In the bottom prediction, however, certain sections appear too confident, with overly narrow confidence intervals. This might indicate that the model is underestimating uncertainty during these periods. Nonetheless, it is a localized issue and doesn't detract from the overall performance. For instance, in the bottom right prediction, between days 11.0 and 11.2, the model predicts a decrease in ammonium concentration during aeration, whereas the observed values show an increase, likely caused by an unusual spike in inlet pollution (ammonium). As this is an infrequent, abnormal event, it is understandable that the model failed to capture it.

To further explore the impact of inlet variation, Figure 7 compares the scaled mean derivative of the non-aeration periods across different times of the day. Figure 7 shows that under simulated conditions with a fixed inlet, the derivative remains nearly constant, as expected. However, in the real data, the derivative fluctuates throughout the day, increasing during daylight hours and decreasing at night. This diurnal variation mirrors patterns observed in other studies measuring real wastewater treatment plant opera-

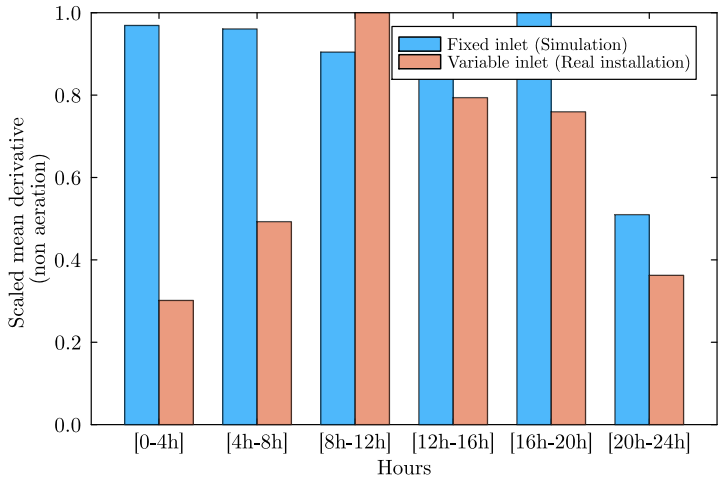

Figure 7: Scaled mean derivative of non-aeration periods comparing simulated and real installation data across different time windows. The blue bars represent fixed inlet simulations, showing nearly constant derivative values. The orange bars represent real installation data, exhibiting significant diurnal variation, with increased derivatives during the day and decreased values at night, reflecting the inlet variability observed in real wastewater treatment plants.

tions. These findings suggest that the model's diminished performance in the real case may stem from the unaccounted variability in inlet concentrations. To improve accuracy, additional covariates such as time of day, rainfall, and other environmental factors should be incorporated into the model. These improvements will be addressed in future research.

4. Conclusion

This study compared different types of models, from white-box to black-box models, to predict ammonium concentration in wastewater treatment plants (WWTP) with a focus on integrating data assimilation techniques.

The objective was to identify the most effective approach for model-based control, taking into account real-world constraints such as limited data availability and sensor noise. Data assimilation proved to be crucial for simultaneously estimating model parameters, latent variables (such as unmeasured species concentrations), and quantifying prediction uncertainty which was rarely done before in the water research literature. Different data assimilation methods, including the Kalman filter, particle filter, and conditional particle filter, were combined with parameter estimation algorithms such as EM and MCEM to achieve robust model calibration.

The simulation results showed that the non-parametric black-box model outperformed all other models in terms of predictive accuracy and uncertainty estimation. This finding highlights the effectiveness of machine learning when integrated with data assimilation techniques to capture information from the training set, even in the presence of limited data. The linear model, although quick to train, failed to capture the complex dynamics of the reactor. The white-box model, while potentially powerful, proved difficult to train, time-consuming, and highly dependent on good initial model parameters, making it less generalizable. The grey-box model offered a good balance between accuracy and computational efficiency, though certain simplifications may have hindered its performance.

Interestingly, the addition of an extra sensor, such as an oxygen sensor, did not improve the model's performance when the target species (ammonium) was already being observed. This finding suggests that, at least for the non-parametric black-box model, investing in additional sensors may not be justified if the primary species of interest is already being monitored.

Experiments on the real system showed that the non-parametric black-box model was able to capture the general dynamics of ammonium concentration in a real wastewater treatment plant. However, the model's performance was somewhat reduced compared to the simulation results, likely due to the variability in the input concentration, which was not accounted for in the model.

In conclusion, this study highlights the importance of data assimilation for robust WWTP modeling, especially when data are limited. It also demonstrates the potential of non-parametric "black-box" models for prediction and uncertainty quantification in this context. The study also highlighted the importance of accounting for input data variability, such as inlet concentrations, to further improve model accuracy. Future research should focus on incorporating additional covariates, such as time of day, precipitation, and other environmental factors, to enhance model accuracy under real-world conditions. Investigating the impact of noise levels, noise distribution, and observation time steps could also provide valuable insights for model refinement. Further improvements in confidence interval estimation could be achieved by incorporating heteroscedastic noise, allowing for more accurate representation of uncertainty in model predictions (Cheng et al., 2023). Additionally, future research could investigate the applicability of the proposed approach to WWTPs with more complex configurations, such as plants with parallel or serial tanks, or those receiving highly concentrated influents (e.g., reject water from sludge treatment), to assess its robustness under varying operational conditions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This material is based upon work supported by the ANRT (Association nationale de la recherche et de la technologie) with a CIFRE fellowship [grant numbers 2022/1582] in collaboration between the IRMAR (Institut de Recherche Mathématique de Rennes), the SATIE (Systèmes et applications des technologies de l'information et de l'énergie) and the Purecontrol Company.

References

RTE, Futurs énergétiques 2050, 2021.

D. J. Reardon, Turning down the power, Civil Engineering 65 (1995) 54.

B. Chachuat, N. Roche, M. A. Latifi, Long-term optimal aeration strategies for small-size alternating activated sludge treatment plants, Chemical Engineering and Processing: Process Intensification 44 (2005) 591–604.

S. Balku, Comparison between alternating aerobic–anoxic and conventional activated sludge systems, Water research 41 (2007) 2220–2228.

L. J. S. Lukasse, K. J. Keesman, G. van Straten, A recursively identified model for short-term predictions of nh4/no3 – concentrations in alternating activated sludge processes, Journal of Process Control 9 (1999) 87–100.

G. Fu, Y. Jin, S. Sun, Z. Yuan, D. Butler, The role of deep learning in urban water management: A critical review, Water Research 223 (2022) 118973.

R. S. Sutton, A. G. Barto, Reinforcement Learning, Adaptive Computation and Machine Learning series, 2 ed., Bradford Books, Cambridge, MA, 2018.

G. Dulac-Arnold, D. J. Mankowitz, T. Hester, Challenges of real-world reinforcement learning, CoRR abs/1904.12901 (2019).

M. Andreides, P. Dolejš, J. Bartáček, The prediction of wwtp influent characteristics: Good practices and challenges, Journal of Water Process Engineering 49 (2022) 103009.

E. E. Holmes, Derivation of an em algorithm for constrained and unconstrained multivariate autoregressive state-space (marss) models, 2013.

N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, N. Chopin, On particle methods for parameter estimation in state-space models, Statistical Science 30 (2015).

M. Henze, W. Gujer, T. Mino, M. van Loosdrecht, Activated sludge models ASM1, ASM2, ASM2d and ASM3, IWA Publishing, 2006.

B. Chachuat, Methodologie d'Optimisation Dynamique et de Commande Op-

timale des Petites Stations d'Epuration a Boues Activees, Ph.D. thesis, Institut National Polytechnique de Lorraine (INPL), 2001.

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

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

A. Sánchez, M. Katebi, M. Johnson, Optimal control of an alternating aerobic-anoxic wastewater treatment plant, IFAC Proceedings Volumes 35 (2002) 411–416.

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

M. Alvi, T. French, R. Cardell-Oliver, P. Keymer, A. Ward, Cost effective soft sensing for wastewater treatment facilities, IEEE Access 10 (2022) 55694–55708.

M. Bocquet, Surrogate modeling for the climate sciences dynamics with machine learning and data assimilation, Frontiers in Applied Mathematics and Statistics 9 (2023).

M. Bonavita, On some limitations of current machine learning weather prediction models, Geophysical Research Letters 51 (2024) e2023GL107377.

R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, P. Battaglia, Learning skillful medium-range global weather forecasting, Science 382 (2023) 1416–1421.

R. Lguensat, Les nouveaux modèles de prévision météorologique basés sur l'intelligence artificielle: opportunité ou menace ?, La Météorologie (2023) 011.

L. Verhaeghe, J. Verwaeren, G. Kirim, S. Daneshgar, P. A. Vanrolleghem, E. Torfs, Towards good modelling practice for parallel hybrid models for wastewater treatment processes, Water Science and Technology 89 (2024) 2971–2990.

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

C.-S. Gómez-Quintero, I. Queinnec, State and disturbance estimation for an alternating activated sludge process, IFAC Proceedings Volumes 35 (2002) 447–452. 15th IFAC World Congress.

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

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

C. Yang, P. Seiler, E. Belia, G. T. Daigger, An adaptive real-time grey-box model for advanced control and operations in wrrfs, Water Science and Technology 84 (2021) 2353–2365.

R. Fablet, B. Chapron, L. Drumetz, E. Mémin, O. Pannekoucke, F. Rousseau, Learning variational data assimilation models and solvers, Journal of Advances in Modeling Earth Systems 13 (2021) e2021MS002572.

M. A. Sacco, M. Pulido, J. J. Ruiz, P. Tandeo, On-line machine-learning forecast uncertainty estimation for sequential data assimilation, Quarterly Journal of the Royal Meteorological Society n/a (2024).

J. Durbin, S. J. Koopman, Time series analysis by state space methods, Oxford statistical science series, 2nd ed ed., Oxford University Press, Oxford, 2012.

S. Cheng, C. Quilodrán-Casas, S. Ouala, A. Farchi, C. Liu, P. Tandeo, R. Fablet, D. Lucor, B. Iooss, J. Brajard, et al., Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review, IEEE/CAA Journal of Automatica Sinica 10 (2023) 1361–1387.

M. Henze, M. C. M. V. Loosdrecht, G. A. Ekama, D. Brdjanovic, Biological Wastewater Treatment, IWA Publishing, 2008.

Metcalf & Eddy Inc., G. Tchobanoglous, F. L. Burton, R. Tsuchihashi, H. D.

Stensel, Wastewater engineering: Treatment and resource recovery, 5 ed., McGraw-Hill Professional, 2013.

V. Bertret, R. L. G. Latimier, V. Monbet, A stochastic expectation maximization algorithm for the estimation of wastewater treatment plant ammonium concentration, in: 2024 European Control Conference (ECC), 2024, pp. 3527–3532.

R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, R. Fablet, The analog data assimilation, Monthly Weather Review 145 (2017) 4093 – 4107.

F. Stulp, O. Sigaud, Many regression algorithms, one unified model: A review, Neural Networks 69 (2015) 60–79.

D. A. Otchere, T. O. Arbi Ganat, R. Gholami, S. Ridha, Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ann and svm models, Journal of Petroleum Science and Engineering 200 (2021) 108182.

A. Carrassi, M. Bocquet, L. Bertino, G. Evensen, Data assimilation in the geosciences: An overview of methods, issues, and perspectives, WIREs Climate Change 9 (2018) e535.

Y. Sasaki, Some basic formalisms in numerical variational analysis, Monthly Weather Review 98 (1970) 875 – 883.

B. D. Anderson, J. B. Moore, Optimal filtering, Courier Corporation, 2012.

G. Evensen, P. J. v. Leeuwen, An ensemble kalman smoother for nonlinear dynamics, Monthly Weather Review 128 (2000) 1852–1867.

A. Doucet, A. M. Johansen, A tutorial on particle filtering and smoothing : fiteen years later, in: D. Crisan, B. Rozovskii (Eds.), The Oxford handbook of nonlinear filtering, Oxford handbooks in mathematics, Oxford University Press, Oxford ; N.Y., 2011, pp. 656–705.

D. Dreano, P. Tandeo, M. Pulido, B. Ait-El-Fquih, T. Chonavel, I. Hoteit, Estimating model-error covariances in nonlinear state-space models using kalman smoothing and the expectation–maximization algorithm, Quarterly Journal of the Royal Meteorological Society 143 (2017) 1877–1885.

A. Doucet, N. De Freitas, N. J. Gordon, et al., Sequential Monte Carlo methods in practice, volume 1, Springer, 2001.

F. Lindsten, Particle filters and Markov chains for learning of dynamical systems, Ph.D. thesis, Linköping University Electronic Press, 2013.

F. Lindsten, M. I. Jordan, T. B. Schon, Particle gibbs with ancestor sampling, Journal of Machine Learning Research 15 (2014) 2145–2184.

F. Lindsten, An efficient stochastic approximation em algorithm using conditional particle filters, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6274–6278.

A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1977) 1–38.

G. Celeux, D. Chauveau, J. Diebolt, On Stochastic Versions of the EM Algorithm, Research Report RR-2514, INRIA, 1995.

L. Ljung, System Identification : Theory for the User, 2nd ed., Pearson Education, 1998.

E. Paul, S. Plisson-Saune, M. Mauret, J. Cantet, Process state evaluation of alternating oxic-anoxic activated sludge using orp, ph and do, Water Science and Technology 38 (1998) 299–306.

K. V. Gernaey, U. Jeppsson, P. A. Vanrolleghem, J. B. Copp, Benchmarking of Control Strategies for Wastewater Treatment Plants, IWA Publishing, 2014.