Water Science & Technology

Water Science & Technology Vol 90 No 5, 1416 doi: 10.2166/wst.2024.293

Hybrid modelling of nitrogen removal by biofiltration using high-frequent operational data

Marcello Serrao[a,b,*], Vincent Jauzein[c], Ilan Juran[d], Bruno Tassin[a]
and Peter Vanrolleghem[b]

[a] Laboratoire eau environnement et systèmes urbaines (LEESU), Ecole des Ponts, Université Paris Est Créteil, Institut Polytechnique de Paris, Créteil F-94010, Marne-la-Vallée, France
[b] modelEAU, Université Laval, 1065 av de la Médecine, Québec, QC G1V 0A6, Canada
[c] SIAAP, Direction Innovation, 82 av Kléber, Colombes 92700 France
[d] W-SMART, 9 rue Victor Schoelcher, Paris 75014, France
*Corresponding author. E-mail: marcello.serrao@suez.com

MS, 0009-0001-9798-4223

ABSTRACT

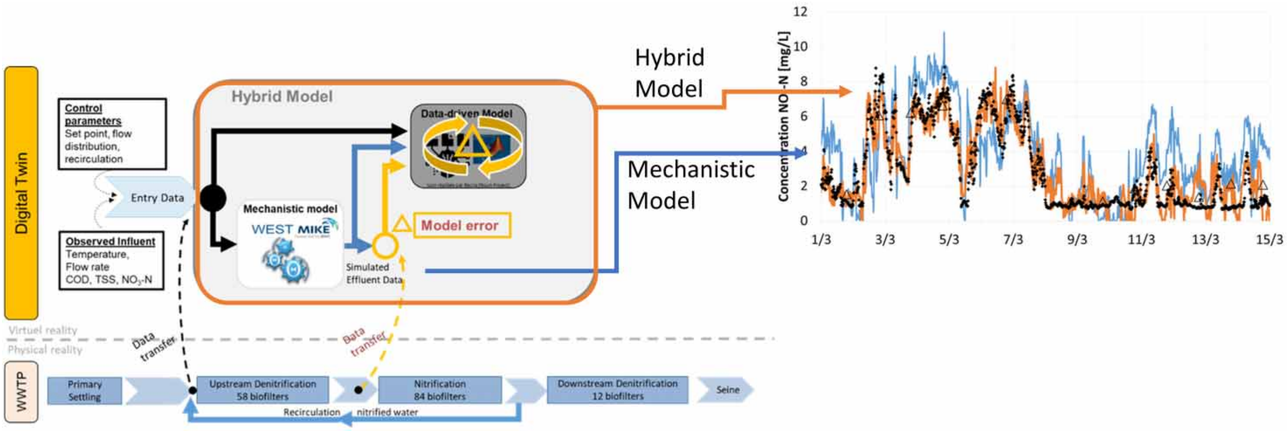
In this research, a parallel hybrid model is presented for the simulation of nitrogen removal by submerged biofiltration of a very large-size wastewater treatment plant. This hybrid model combines a mechanistic and a machine learning model to produce accurate predictions of water quality variables. The models are calibrated and validated using detailed and quality-controlled operational data collected over a period of 3.5 months in 2020. The mechanistic model is a modified activated sludge model that describes the biological, physical and chemical processes taking place in a biofilm reactor based on the domain knowledge of these processes. A three-layer feed-forward artificial neural network with a rectified linear activation function that aims to reduce the mechanistic model's residual error and then correct its output. The results show how the hybrid model outperforms and significantly reduces the size of the mechanistic model's prediction errors of the effluent nitrate concentration from a relative mean error of 12% (mechanistic model) to 2% (hybrid model) during training. The error on nitrate simulations increases to 8% during hybrid model testing, still significantly lower than the error of the mechanistic model. These results support future operational applications of hybrid biofilm models, such as in digital twins.

Key words: biofiltration, data-driven models, hybrid models, mechanistic models, water resource recovery facility

HIGHLIGHTS

- First hybrid model for a biofilm wastewater system for nitrogen removal.
- Feed-forward neural networks came out as the best of different data-based modelling approaches.
- Validation error (15 min interval) of the hybrid model is three times smaller than the error of the mechanistic-only model.
- The hybrid model is designed so that it can be used in digital twin applications.

GRAPHICAL ABSTRACT



ABBREVIATIONS

ANN	Artificial neural network
ASM	Activated sludge model
COD	Chemical oxygen demand
CODs	Soluble fraction of the chemical oxygen demand
CSTR	Continuous stirred tank reactor
DDM	Data-driven model
HM	Hybrid model
MAE	Mean absolute error
ME	Mean error
ML	Machine learning
MM	Mechanistic model
MSE	Mean squared error
NIT	Nitrification
PostDN	Postdenitrification
PreDN	Predenitrification
R^2	Coefficient of determination
RMSE	Root mean square error
SVM	Support vector machines
TKN	Total Kjeldahl nitrogen
TSS	Total suspended solids
WRRF	Water resource recovery facility

1. INTRODUCTION

Improving the performance and efficiency of water resource recovery facilities (WRRFs) involves strengthening the control of the processes that take place there. Objectives for the wastewater treatment sector nowadays include optimization of energy performance, reduction of the use of reagents, of greenhouse gas emissions and of the overall environmental impact of operations, improvement of the recovery of resources and the stabilization of effluent quality. An important field of research aims to study the potential of the application of artificial intelligence techniques to achieve this aspired improvement in the performance of WRRFs (Schneider *et al.* 2022).

Wastewater treatment includes several successive physical, biological and chemical processes, that are non-linear and non-stationary, which complicates the optimization of their management (Henze *et al.* 2008). To facilitate and improve treatment and ensure the achievement of performance in accordance with regulatory obligations, in recent decades, the instrumentation, control and automation of WRRFs have developed significantly (Ingildsen & Olsson 2016; Corominas *et al.* 2018) and so is the development of mathematical models for simulation, scenario analysis and support of process control (Garrido-Baserba *et al.* 2020).

Since the 1990s, mechanistic dynamic models describing the suspended growth process of biomass in activated sludge systems, such as the activated sludge model 1 (ASM1) and ASM2 models (Henze *et al.* 2000) have acquired a solid and valid basis within the scientific community. These models contain many parameters and may require significant calibration when used outside typical conditions (Rieger *et al.* 2012). In the case of treatment by biofiltration, in which a dense 'fixed-culture' biomass is exploited, the available models that also need to describe biofilm-specific processes (e.g., diffusion, competition for space among several types of biomasses and regular backwashing to prevent filter clogging) are considered more complex (Eberl *et al.* 2006).

A literature review of mechanistic modelling studies for nitrogen removal by biofiltration reveals that over the last 25 years, the models have indeed become increasingly complex to include more processes taking place in both the biofilm and bulk liquid compartment, the gaseous compartment, and the transfer between these compartments. Behrendt (1999) developed a two-step nitrification biofilter model; however, it failed to account for the evolution of biomass growth and decay or the transport of particulates (Samie *et al.* 2010). Falkentoft *et al.* (2000) developed a model based on ASM2d describing denitrification and phosphorus removal in a biofilter. It introduced the concept of an initial biofilm thickness to account for the impact of filter washing. The model of Viotti *et al.* (2002) simulates concentration profiles of the chemical oxygen demand

(COD) and NH_4-N inside the biofilm and along the filter bed. It also models the accumulation of biomass interfering with filtration efficiency and affecting head loss due to clogging. Hidaka & Tsuno (2004) added NO_3-N concentration profiles over the filter bed represented by five continuously stirred tank reactors (CSTRs). However, it does not account for diffusion, assuming complete substrate penetration into a uniform biofilm consisting of one layer.

Samie *et al.* (2010) adequately developed a biofiltration model with a three-layered biofilm for carbon and nitrogen (nitrification followed by denitrification) removal calibrated with daily laboratory samples from a large WRRF placing the modelled stages in series. Moreover, this model estimated greenhouse gas emissions (Samie *et al.* 2011). Vigne *et al.* (2010) successfully applied a model with six CSTRs, each containing a five-layered biofilm with a fixed-depth boundary layer and bulk liquid compartment to a nitrification biofilter with operational data from a large WRRF. During washing, partial mixing of the biofilter media along the filter bed is modelled, reducing the vertical gradient of biofilm properties.

Following that work, Bernier *et al.* (2014) modelled a biofilter with seven CSTRs but with a simplified two-layered biofilm. It computes head loss as a function of the thickness of the biofilm layers. The model was successfully calibrated with operational data from a large WRRF. Fiat *et al.* (2019) provided a model extension to include the main nitrous oxide (N_2O) biological pathways during nitrification and denitrification.

Finally, Zhu (2020) enhanced the biofiltration model of Bernier *et al.* (2014) by returning to a five-layer biofilm but making the boundary layer variable thickness depending on hydraulic conditions. It not only considers media mixing during washing but adds less intense mixing during normal operation. The biological production and removal of N_2O and a more detailed gas–liquid transfer process were added. Energy consumption could be calculated as well with this model. Placed in series, it is a biofilter model integrating three stages of nitrogen removal (predenitrification–nitrification–postdenitrification).

The increased complexity of these mechanistic biofilter models is making their parameterization, calibration, and validation much more laborious, requiring qualified experts and highly frequent observational data of good quality. This can become a bottleneck to the application of mechanistic models in an operational context (Vanrolleghem *et al.* 2005). In addition to the complexities mentioned in the review above, biofiltration models 'suffer' from a lack of widely accepted guidelines available to the modelling community (Rittmann *et al.* 2018). This increased complexity and inherent numerical challenges also require more computing power, which makes them less suitable for real-time applications where model updates are needed in real time with a short time horizon of action (Schneider *et al.* 2022).

Conversely, data-driven models that use artificial intelligence techniques to find patterns in data are computationally very fast, making them very interesting for real-time applications (Duarte *et al.* 2023). These models describe the system only based on information extracted from the provided data and have strong interpolation features, but they are less reliable for making predictions outside of training conditions (Newhart *et al.* 2019). Data-driven models typically require larger datasets than mechanistic models. Related to the inherent characteristics of wastewater treatment processes, which are nonlinear and show a large time dependency, frequent retraining of a data-driven model (DDM) remains necessary (Torfs *et al.* 2022).

Hybrid modelling combines the use of mechanistic models with data-based models and can take advantage of each approach (Von Stosch *et al.* 2014a): a mechanistic model that incorporates relevant knowledge about the processes, and a DDM that increases the predictive power of the model by including information on lesser-known sub-processes at reduced computing power. The training and testing of machine learning (ML) models is much faster than the calibration and validation of mechanistic models (Sundui *et al.* 2021). In recent literature, hybrid models (HMs) describing activated sludge systems are reported to improve model performance and support process monitoring and control (Lee *et al.* 2005; Quaghebeur *et al.* 2022; Sparks *et al.* 2022; Torfs *et al.* 2022). The main advantage of an HM is a higher benefit/cost ratio for solving complex problems, which is a key factor for process systems engineering (Schneider *et al.* 2022).

Serial and parallel structures are hybrid model architectures in which the mechanistic and DDM components work together. In a parallel structure, the outputs of both models are combined (usually by addition) to improve output. In a cooperative parallel structure, the DDM is trained to learn the mechanistic model residual error with respect to observational data (Lee *et al.* 2005) or to learn the residuals in the dynamics (Quaghebeur *et al.* 2022). With a competitive parallel hybrid model, the mechanistic and data-driven models independently make the predictions for the same variable, which are then weighted and combined into a final output (Dors *et al.* 1995; Peres *et al.* 2001; Galvanauskas *et al.* 2004; Ghosh *et al.* 2019).

The serial structure allows the output of a first model (commonly a DDM) to be input to the second (mechanistic) model with the objective of improving the second model's performance. This is useful when a specific subprocess is not sufficiently well described mechanistically. The data-driven component then represents these missed dynamics (Psichogios & Ungar

1992). With the inverse approach, the output of a mechanistic model is given to a DDM with the objective of supporting its features with domain knowledge (Tsen *et al.* 1996; Li *et al.* 2019; Hannaford *et al.* 2023).

The hybrid concept has been the subject of research for a decade in other industrial sectors; for example, in the biopharmaceutical industry (e.g., Von Stosch *et al.* 2014b) and the chemical industry (e.g., Schuppert & Mrziglod 2018). The wastewater sector is experiencing a lag in the development of HMs probably due to the complexity of the processes to be modelled, the high level of investments (that disagree with the low budgets available for wastewater infrastructures) and the long-life expectancy of the infrastructures which makes changes slow to be adopted. Additionally, wastewater treatment systems have peculiarities such as their nonlinear and nonstationary nature, making it a challenge to develop accurate mechanistic models (Corominas *et al.* 2018). On the other hand, great potential is expected in applying AI and its ML methodologies thanks to the increasing availability of 'big data' (Garrido-Baserba *et al.* 2020).

In this study, a hybrid model that is built to form an integral part of a 'digital twin' (e.g., for use in the HM predictive control application in Serrao *et al.* 2023) is developed by combining a mechanistic dynamic biofilm reactor model (MM) with a DDM capable of running in near real-time in a python environment controlling both the MM and the DDM. A key novel aspect of this study is that high-frequent operational data from a very large-size operational municipal water resource recovery facility is used for the training and testing of the HM instead of data from a pilot or otherwise controlled environment. The mechanistic biofilm reactor model and ML models are trained for longer periods (3.5 months) than in any other hybrid biofilm modelling study known to the authors. It is also the first hybrid model of a biofilter system.

A clear outcome of this study is that the HM improves the accuracy of the wastewater treatment model by reducing the size of prediction errors by a factor of 3 under challenging validation conditions. This increases the confidence in the model's performance and thus allows the application of hybrid biofiltration models as decision-making tools for managers, operators and process engineers in the wastewater industry and as an online ingredient for model-based predictive control (Serrao *et al.* 2023).

This modelling study culminated in the demonstration study of Serrao *et al.* (2023) in which the conventional feed-forward controller of methanol dosing in use today at the plant under study is compared to a model predictive controller of the methanol dosing to a biofilm reactor that takes advantage of the developed hybrid model in the digital twin central to this new controller.

2. MATERIALS AND METHODS

2.1. Case study site

The research project was carried out with observational data collected in the Seine-aval WRRF managed by SIAAP[1] located in the Greater Paris Region, France. This very large-size plant receives wastewater generated by nearly 6,000,000 population equivalents every day, i.e., a volumetric flow of approximately 1,250,000 m^3/day in dry weather (14.5 m^3/s) (SIAAP 2018).

The wastewater treatment chain consists of a physico-chemical treatment step (screening, grit-oil removal and primary settling enhanced by the addition of coagulant-flocculant reagents) followed by biological filtration carried out over three stages for the removal of organic matter and nitrogen in a configuration of predenitrification (preDN, 58 Biostyr®), nitrification (NIT, 84 Biostyr®) and postdenitrification (postDN, 12 Biofor®), as shown in Figure 1. The typical flow rate to the biofiltration train during the period studied (December 2019–October 2020) is 1,100,000 m^3/day.

The average upflow velocity through the Biostyr filters is 10 m/h for an average NO_3-N removal of 70% (preDN) and 7.2 m/h to remove 90% of the NH_4-N (NIT). The average upflow velocity of the Biofor filters (postDN) is 19 m/h with a NO_3-N reduction of 85% over the period studied for dry and rainy weather periods combined. More details can be found in Serrao *et al.* (2023).

The biological filters are backwashed daily on a fixed schedule by reversing the flow using treated water stored on site and injection of large air bubbles. During washing, which lasts 30 min per day per filter, the trapped particles and part of the biomass are removed in order to restore the filtration capacity. Additional short backwashing sessions of 10–15 min can be applied to a filter when the pressure loss reaches a determined value. More details can be found in Serrao (2023).

[1] Syndicat interdépartemental pour l'assainissement de l'agglomération parisienne

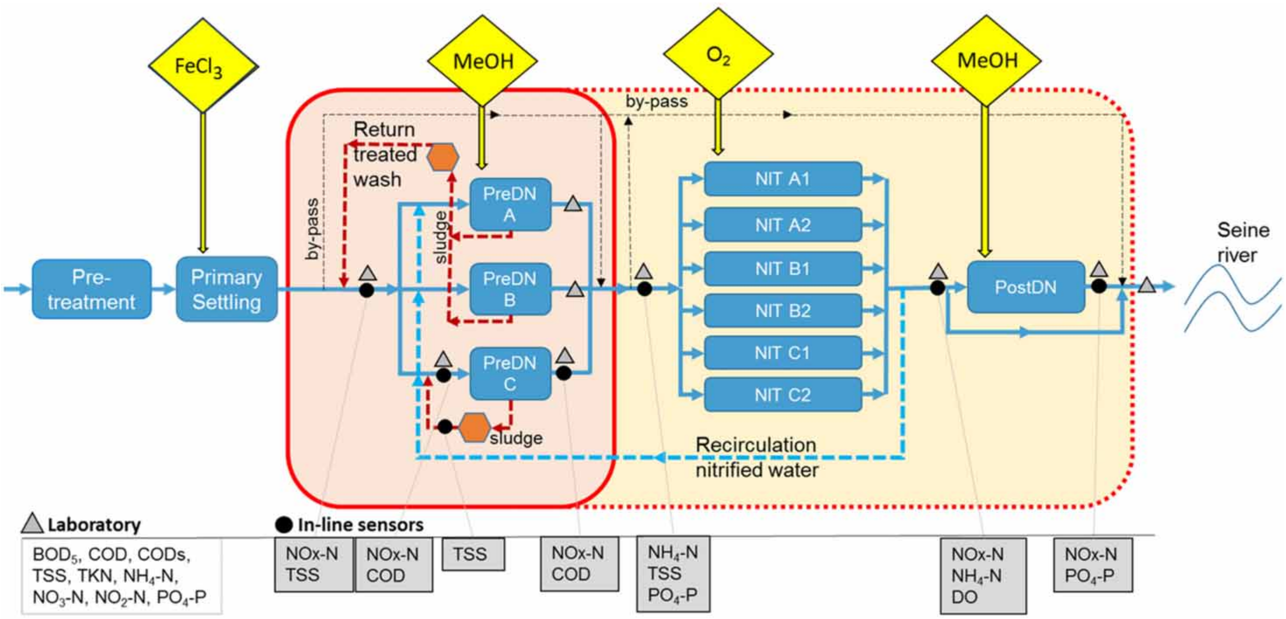


Figure 1 | Schematic diagram of the principal wastewater treatment lane at the Seine-aval WRRF (SIAAP) with indication of processes (blue boxes) and main dosing actions (yellow diamond shapes). Outlined in red frame are the secondary treatment predenitrification units (preDN), the nitrification units (NIT) and postdenitrification units (postDN). The focus of this study is on the preDN stage that includes methanol (MeOH) dosing. Also indicated are the sludge treatment units for preDN (brown hexagons) and location of measurements by inline sensors (black dots) and laboratory samples (grey triangles), that include biological oxygen demand over 5 days (BOD_5) and total Kjeldahl nitrogen (TKN).

2.2. Hybrid model approach

The HM developed here aims at predicting effluent water quality variables. The MM specifies the basic dynamics of the relevant process variables and preserves process knowledge ('the First Principles'). The DDM learns unknown relationships from the data and helps improve simulation quality, especially for interpolation purposes. It has a parallel configuration, where the DDM and the MM model both produce an output that is subsequently combined, by the addition of the two outputs (Figure 2). The objective of the data-driven component is to calculate the residual error (Δ) of the knowledge-based model. It extracts relationships from the data inputs with which it can estimate the error of the MM-based simulation of

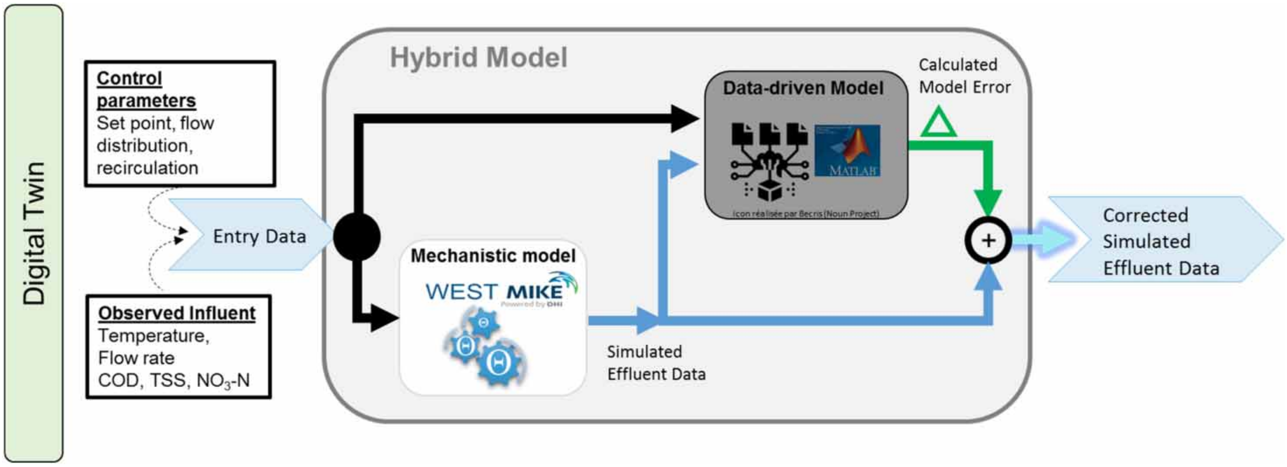


Figure 2 | Schematic representation of the developed hybrid model (grey box) where the simulations (blue arrow) of the knowledge-based mechanistic model (white box) and the estimated residual error (green arrow) by the data-driven model (black box) are combined to produce corrected values of effluent quality variables (for more details, see text).

the system. However, it does not learn explicitly the underlying process dynamics, which could form a problem when trying to make simulations in situations that fall outside the conditions with which the model was trained (Anderson *et al.* 2000). To minimize this limit in extrapolation power, the data-driven component needs to be exposed to a sufficiently large and representative dataset that captures the diversity of operational conditions (Quaghebeur *et al.* 2022). The parallel structure provides a significant advantage for conditions where the MM contains a large number of parameters necessary to describe complex processes, such as in biological WRRF applications (Lee *et al.* 2002, 2005).

In this configuration presented in Figure 2, the MM first simulates the effluent water quality variables, sludge concentrations, greenhouse gas emission and flow rates, as well as the intermediate state variables of the model. The selection of features for the DDM is based on data availability followed by an analysis of importance using an *F*-test method that allows reducing the dimensionality of the data. The selected simulated variables are then passed to the DDM (blue coloured arrow in Figure 2). Next to these MM simulation data, the data-driven component also receives measurement data of the influent quality and flow rates, as well as information on the principal operational settings, such as the effluent set points for nitrates, as specified in the MM (black arrow). After training, the DDM uses this information to calculate the residual errors of the mechanistic model simulations Δ (green arrow). Finally, the two results are added together to obtain corrected values of the effluent variable (light blue highlighted arrow). In this approach, a ML model is trained for each of the output variables, where for each variable, a corresponding subset of process input variables or features is used.

During the training phase of the DDM, observations of the water quality variables to be modelled are used to calculate the difference between observed and simulated variables. This residual error of the variables as simulated by the mechanistic model is then used by a Bayesian optimization algorithm to train the ML model so that it learns to calculate the proper errors. It serves as labelled responses in a supervised ML training environment.

2.2.1. Mechanistic model structure

The mechanistic model applied here is designed as a one-dimensional dynamic biofilm reactor model describing a 'fixed-culture' submerged upward flow biological activated filter. The model was previously developed (Zhu *et al.* 2018; Zhu 2020) and combines sub-models for the hydraulic representation of the filter media, the transfer and transport of soluble and particulate substrates to/from and within the biofilm, kinetics of bacterial growth and decay, and backwashing to avoid clogging of the filter. The model simulates the biological and physico-chemical conversion of carbon, nitrogen and phosphorus. In total, 22 biological conversions are accounted for by describing the evolution of substrates by mass balances.

The performance of biofilters is related to the filtration of suspended particles and the biological activity of purifying biomasses (Henze *et al.* 2008). Figure 3 shows the approximation of the vertical hydraulics of a filter into seven successive

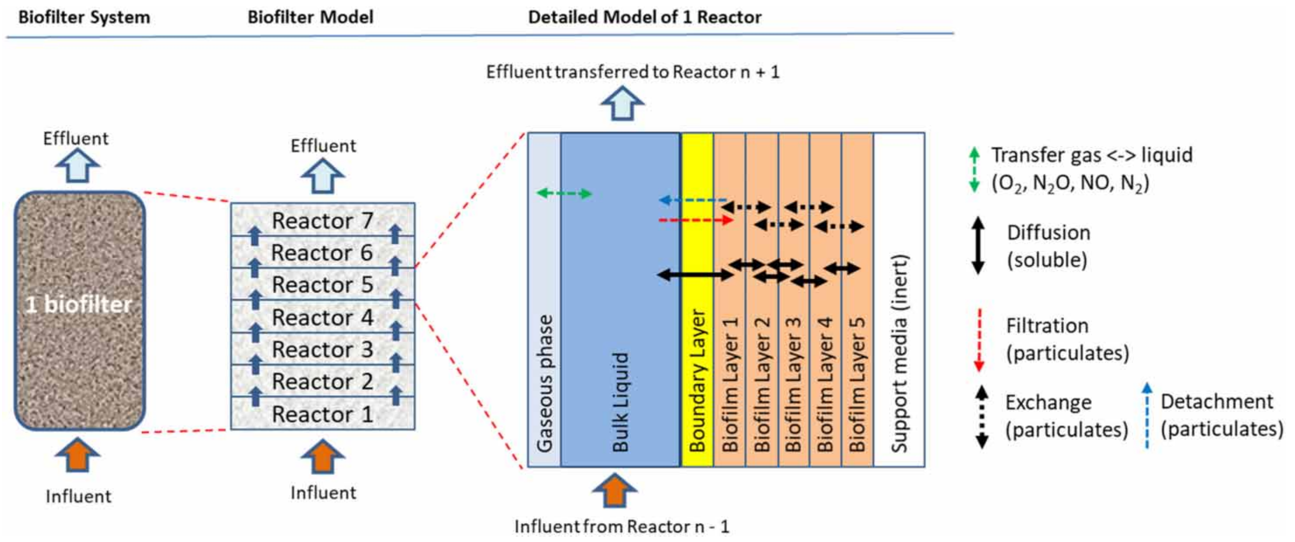


Figure 3 | General concept of the biofiltration model developed by Zhu (2020) that simulates the behaviour of one upflow biofilter, as a series of seven reactors. For each reactor there are compartments for gas, bulk liquid, the mass transfer boundary layer and the biofilm, divided into five layers. The horizontal arrows indicate the main physical phenomena modelled.

completely mixed reactors, a number that was selected to provide enough level of modelling detail for the gradient along the reactor and does not incur a too high computational cost (Zhu 2020). The volume of each mixed reactor is divided between the biofilm compartment (in orange), a bulk liquid phase (blue), a gaseous phase (grey) and the inert media (white). The mass transfer boundary layer (yellow) simulates the diffusion resistance of soluble components in the stagnant layer between the liquid phase and the biofilm (Boltz *et al.* 2011). A variable thickness of the boundary layer is modelled according to Ohasi *et al.* (1981), considering the media diameter and the soluble component's Sherwood number (Zhu 2020). The biofilm compartment is split into five homogeneous layers, a similar number applied previously by Vigne *et al.* (2010) that was found to not take too much computational time and still accurately simulate the concentration gradients of soluble substrates (Zhu 2020). The model assumes a dynamic biofilm volume with a variable thickness of each biofilm layer up to a predefined maximum allowable thickness, based on the mass of particulates present in each layer, the surface area of liquid to media and a constant biofilm density.

The physical processes considered in the model (the horizontal arrows in Figure 3) in each reactor are the transfer to and transport within the biofilm of soluble components by diffusion, the transfer and transport of particulate components by filtration, attachment and detachment, as well as a flux of four gaseous components (O_2, N_2, NO, N_2O), based on Pocquet *et al.* (2016). These gazes are considered as soluble components in the biofilm layers and are transported by diffusion (Zhu 2020). In addition, the model redistributes the particulate components between the successive mixed reactors during treatment and especially during backwashing by representing continuous media mixing at different intensities for normal operation and backwashing. Mixing occurs between the corresponding biofilm layers of adjacent bioreactors. The model allows for estimating energy consumption related to aeration and pumping, as well as greenhouse gas emissions (Zhu *et al.* 2018).

2.2.2. DDM structure

In the proposed hybrid model structure, different ML methods were tested: artificial neural networks (ANN), support vector machines (SVM), nonlinear regression and decision tree-based methods (Lee *et al.* 2002; Li & Vanrolleghem 2022; Shirkoohi *et al.* 2022). In this study, supervised learning was applied for training the ML models where input data are provided to the model along with the output. The selection of the optimal ML model in this study includes the exploration of data features, validation schemes and model performance.

The k-fold cross-validation method was selected for training so that the models could be compared with the same validation scheme. Cross-validation is a resampling procedure used to evaluate ML models that use smaller subsets of the original dataset to avoid an overfit or underfit of the model during training. This method makes efficient use of all the data and still gives a good estimate of the predictive accuracy of the trained model, although it can take a longer time to execute because the model is trained repeatedly on the subsets (Nti *et al.* 2021).

2.2.3. Observational data

The datasets used in this study cover December 2019 to October 2020 and consist of 15-min interval sensor data on flow rates, organic matter, total suspended solids (TSS) and nitrogen (NH_4-N and NO_x-N), supported by daily flow-proportional composite samples for laboratory analysis collected at key points in the system (see Figure 1). Details can be found in Serrao *et al.* (2023).

The calibration period covers December 2019 until 15 March 2020. The validation period runs from June to September 2020, a summer period characterized by low flows during the holiday month of August, conditions very different from the calibration period, making for an ambitious validation challenge.

Operational data collected in raw format were first cleaned to remove outliers and erroneous data using the method described by Alferes & Vanrolleghem (2016). Datasets for model input require gap filling of missing data, for which the moving median method was applied. Furthermore, data input for the ML models were normalized using the z-score, with a mean $\hat{y}$ and a standard deviation σ, and scaled in the range of 0–1 to provide uniform data across all features that still have the original shape properties. The normalized variables (features) included the type of day (weekday or weekend), type of weather (dry or rainy), temperature, as well as operational parameters such as the effluent NO_3-N set point and the methanol dosing rate.

2.2.4. Model performance criteria

The evaluation criteria used to evaluate the model performance are the mean error (ME) between observed and simulated data, the mean absolute error (MAE) and the root mean square error (RMSE), as well as their relative values. The ME indicates an uncertainty in a measurement or the difference between the measured value and modelled values. The MAE gives the average size of the error in a collection of predictions, without taking their direction into account. In an ideal case, the errors should be close to 0, but values up to 15% are considered acceptable. Model calibration for the mechanistic and HMs was based on the RMSE, which provides a quantitative measure of how well the model fits the data average, although it is more sensitive to outliers or large errors. The ME and MAE were supplementary indicators used for indication of a model's bias and of the error's size on average. The Janus coefficient, that evaluates the ratio of RMSE's of validation and calibration has an optimal value of 1 and indicates successful validation for values up to 2 (Hauduc *et al.* 2015).

2.2.5. Mechanistic model calibration

The mechanistic biofilter model developed by Zhu (2020) was first calibrated for the large-size municipal wastewater treatment plant (Seine-aval, SIAAP) using operational measurement data at the entry and exit of the predenitrification (preDN) stage. High-frequency data (interval of 15 min) were available for nitrate, nitrite, ammonia, total COD, soluble COD (CODs) and TSS concentrations, supported by daily composite laboratory results. The calibration/validation of this model is described in detail in Serrao *et al.* (2023).

The mechanistic modelling of the biofiltration system was performed in the WEST (DHI, Hørsholm, Denmark) environment.

2.2.6. DDM training

The selection of the optimal DDM type included training with linear regression models, regression decision tree models with a fine, medium and coarse structure; SVM and neural networks (NN) with one, two and three hidden layers. Since NN with one hidden layer performed equally good or better than models with multiple hidden layers that can be considered to be more 'complex'; those other models weren't selected for further training. To select each model's hyperparameters (number of layers, type of activation function, etc.), a five-fold cross-validation was applied that splits the data randomly into five subsets (folds) of equal size. One randomly chosen subset was first used for model validation, while the other four were used for training. This process was repeated five times so that each subset was used once for validation. The average error across all five folds is then reported as the average validation error, allowing for the selection of the optimal DDM structure. During the training phase of the ML models, the residual error of the mechanistic model simulations with respect to the measured data was transmitted to allow supervised training. This was done for each simulated univariate output.

Given the time-dependent behaviour of the system, the data cannot be divided randomly into a training and testing set but split chronologically where roughly the first 80% of the records in the time series are marked for training and the remainder 20% are set aside for testing.

The performance of each model was evaluated on the training set, in particular in terms of the RMSE criterion using a five-fold cross-validation scheme. The selected model was further tested with the testing dataset.

The raw data processing, the development and the training of the DDM were all performed in the MATLAB R2022a (MathWorks, Natick, MA, USA) environment using the *Machine Learning* and *Deep Network Designer* toolboxes.

2.2.7. Optimization of the selected DDM structure

Selection of the number of features and the model's hyperparameters was done to avoid under-fitting or over-fitting conditions over both training and testing periods. This was achieved by selecting the simplest model that still has a low RMSE score after removing features that are not contributing to the performance of the DDM (see Section 2.2.7.1) and by changing the values of the hyperparameters of the model, such as the number of neurons in the hidden layer (see Section 2.2.7.2).

2.2.7.1. NO_3-N DDM feature selection. Not all features contribute equally to the response variable; some can actually disturb the model output and decrease performance. Model improvement is achieved by removing features that have a low predictive power. The selection of which features to retain is based on a feature-ranking algorithm.

Here, the *F*-test was applied to identify the variables that allowed to best fit the data (Helsel & Hirsch 1992). The *F*-test uses a filter-type feature selection algorithm that measures feature importance based on characteristics, such as feature variance and feature relevance to the response. It examines the importance of each feature individually and allows defining the likelihood that an observed improvement of a fit to the data is worth the use of the feature. For each validation fold, the features that had a zero or near zero *F*-test score were removed, such that only the highest ranked features with a significant contribution (*F*-test above a value of 175, which corresponds to the half-value of the feature with the highest finite *F*-test score), were retained.

Figure 4 shows that 10 out of 12 features – selected by expert opinion out of the available measured and modelled variables – were identified to be influential to some degree, of which five have infinite importance on the model outcome. The most influential variables are water temperature, COD influent concentration (COD_IN), NO_3-N recirculation concentration (NOx_REC), TSS concentration in settled water (TSS_SW), NO_3-N concentration in effluent modelled by the mechanistic model (NOx_OUT_MOD), the flowrate of settled water (Q_SW) and the weather type – rain or dry (Weather). Of these features, all but one are measured variables that serve to characterize the influent water quality; the exception being NOx_OUT_MOD which is a simulated output of the mechanistic model. The features MeOH_Dose (methanol dosing rate) and NO3_SP (nitrate setpoint) appear to have no influence on the response variable but were selected nevertheless because of their operational importance in nitrogen removal during treatment.

The variables NOx_IN, which is the measured NOx-N concentration [g/m^3] in the preDN influent, and Q_REC (the measured flow rate of the recirculation of nitrified water [m^3/day]), are both identified as less influential. Their intrinsic information content is actually included in variables that have been classified as influential, namely the NOX_REC [g/m^3], which accounts for the significant source of nitrates and nitrites in the recycled water from the nitrification tank, after combination with the influent; and the Q_SW [m^3/day] (the settled water flow rate). Note that the flow rate of recirculated nitrified water Q_REC [m^3/day] is adjusted according to Q_SW to maintain a constant hydraulic loading of the biofilter; if Q_SW [m^3/day] increases, then the Q_REC [m^3/day] decreases.

2.2.7.2. NO_3-N DDM hyperparameters & model selection. Another important step in DDM model development is the identification of the DDM hyperparameters. A hyperparameter is a parameter that describes a configuration that is external to the model itself, like the learning rate, the number of estimators and the type of regularization. Hyperparameter tuning was performed using Bayesian optimization with MSE as a loss function. An acquisition function determines the next set of hyperparameter values to try.

The hyperparameters evaluated, are (1) the number of hidden layers with a range [1:3]; (2) the first layer size with a range [1:300]; (3) the second layer size with a range [1:300]; (4) the third layer size with range [1:300]; (5) the activation function, a search among the rectified linear unit, hyperbolic tangent, none, and sigmoid; (6) the regularization strength (Lambda). The optimization algorithm searches among log-scaled real values in the range [1e-5/n, 1e5/n], where n is the number of observations. For more information, refer to Serrao (2023).

Table 1 provides an overview of the 10 best scoring DDMs for the estimation of the residual errors between the mechanistic model simulation results and the observations of nitrate concentrations in the effluent of the preDN system. As can be seen,

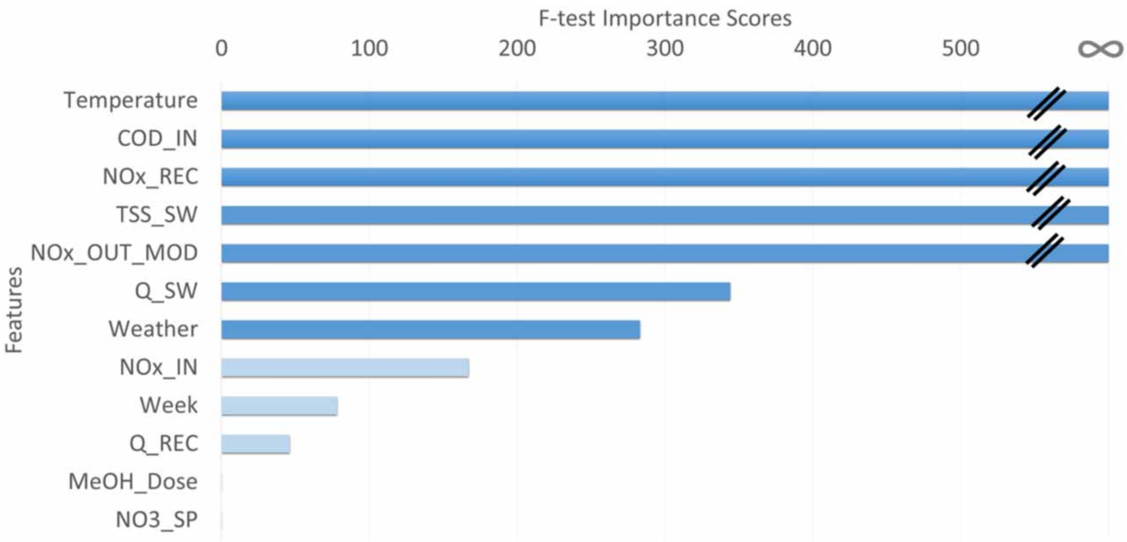


Figure 4 | Features used by the ML model and their ranking by *F*-test scores for the NO_3-N residual error.

Table 1 | Overview of the 10 best scoring DDM models on the training and testing dataset for the estimation of the NO_3-N residual error

Model type	Hyperparameters	Selected features	RMSE training	RMSE testing
Neural network	**Neurons: 12:25:1**	**12**	**0.058**	**0.170**
Decision tree	Minimum leaf size: 12 Surrogate splits: off	9	0.063	0.191
Decision tree	Minimum leaf size: 4 Surrogate splits: off	9	0.059	0.193
Decision tree	Minimum leaf size: 4 Surrogate splits: off	12	0.059	0.194
Decision tree	Minimum leaf size: 4 Surrogate splits: off	10	0.059	0.194
SVM kernel	Regularization strength: auto	12	0.062	0.246
SVM kernel	Regularization strength: auto	10	0.062	0.274
SVM kernel	Regularization strength: auto	9	0.059	0.295
Neural network	Neurons: 9:25:1	9	0.059	0.297
Neural network	Neurons: 10:25:1	10	0.058	0.365

NNs provide good performance with the lowest RMSE score for both testing and training data of NO_3-N. The optimal network configuration was determined by simulations with the smallest configuration of an input, hidden and output layer and by varying the number of nodes in the hidden layer. The model with the lowest RMSE value during testing is a three-layer neural network with 12, 25 and 1 neurons, respectively, for the input, middle and output layers. The second to fifth positions are taken by decision tree models, indicating that models with a simpler structure could have a similar performance. However, considering the training RMSE values, where ANN models are listed at first, second and third places, it becomes clear that ANN models have a stronger performance. This is confirmed by the testing performance, where the lowest RMSE is found for a NN model. Overall, the neural network models consistently outperform other model types, such as Decision Trees and Supported Vector Machines.

As indicated in Table 1, during training the ANN model that applies 12 features has a score similar to the ANN model using 10 features; however, during testing the 12 features model has a lower RMSE of 0.17 compared to the 10 features model (RMSE: 0.37). It was therefore decided to select the three-layer feed-forward neural network with 12 neurons in the input layer, 25 neurons in the middle layer and one neuron in the output layer with a rectified linear unit activation function.

The selected ANN for NO_3-N removal was specifically chosen for use in the HM predictive control application presented in Serrao *et al.* (2023) as an integral part of a digital twin, where it is combined with the mechanistic dynamic biofilm reactor model described in paragraph 2.2.1. This digital twin runs in a Python environment using the WEST application programming interface, with timely provision of input data collected from the supervisory control and data acquisition (SCADA) system and allows controlling in near real-time the methanol dosing in a preDN biofilter.

An overview of the 10 best scoring data-driven models for the estimation of the total COD and TSS residual error is presented in Supplementary Information. For these two output variables too, the selected model is a three-layer feed-forward artificial neural network. For the total COD, the same 12:25:1 configuration was found. For TSS, the ANN model contains 10 neurons in the hidden middle layer, which makes it different from the NNs selected for the nitrate and total COD model. This relates to differences in the removal of carbon as compared to nitrogen. Both for the training and testing datasets, decision tree models take up prominent places in the RMSE ranking. These model types are easier to interpret and more transparent to explain output, but also prone to overfitting as indicated by the RMSE values for the test data that are higher than those of the selected ANNs (Bramer 2007).

3. RESULTS AND DISCUSSION

The configuration of the hybrid model implies an interaction between a mechanistic model and a DDM, with the mechanistic model output being used by the DDM to be trained using the calculated residual error. For that reason, this paragraph commences with a quick overview of the mechanistic model's performance.

3.1. Mechanistic model performance

Simulation results from the calibrated MM for effluent nitrate concentrations are presented in Figure 5(a), which shows that the observed trend is well simulated, in particular for the period starting from the end of January until mid-March. However, the model predictions are overestimated by 2–3 mgN/L during the period end of December until the end of January. The MM is able to reproduce the daily average performances of the denitrifying filter for NO_3-N, NO_2-N, NH_4-N, total COD, soluble COD and TSS concentrations, with relative mean errors between 10 and 30% of the observed mean. The model results for the validation period (Figure 5(b)) shows a significant overestimation of the sensor observations with an ME of −3.22 mgN/L, except for the short periods related to observed peak values, such as on the 25th of July, 16th of August and 12th of September. Similar overestimations are observed for the total COD and TSS simulations.

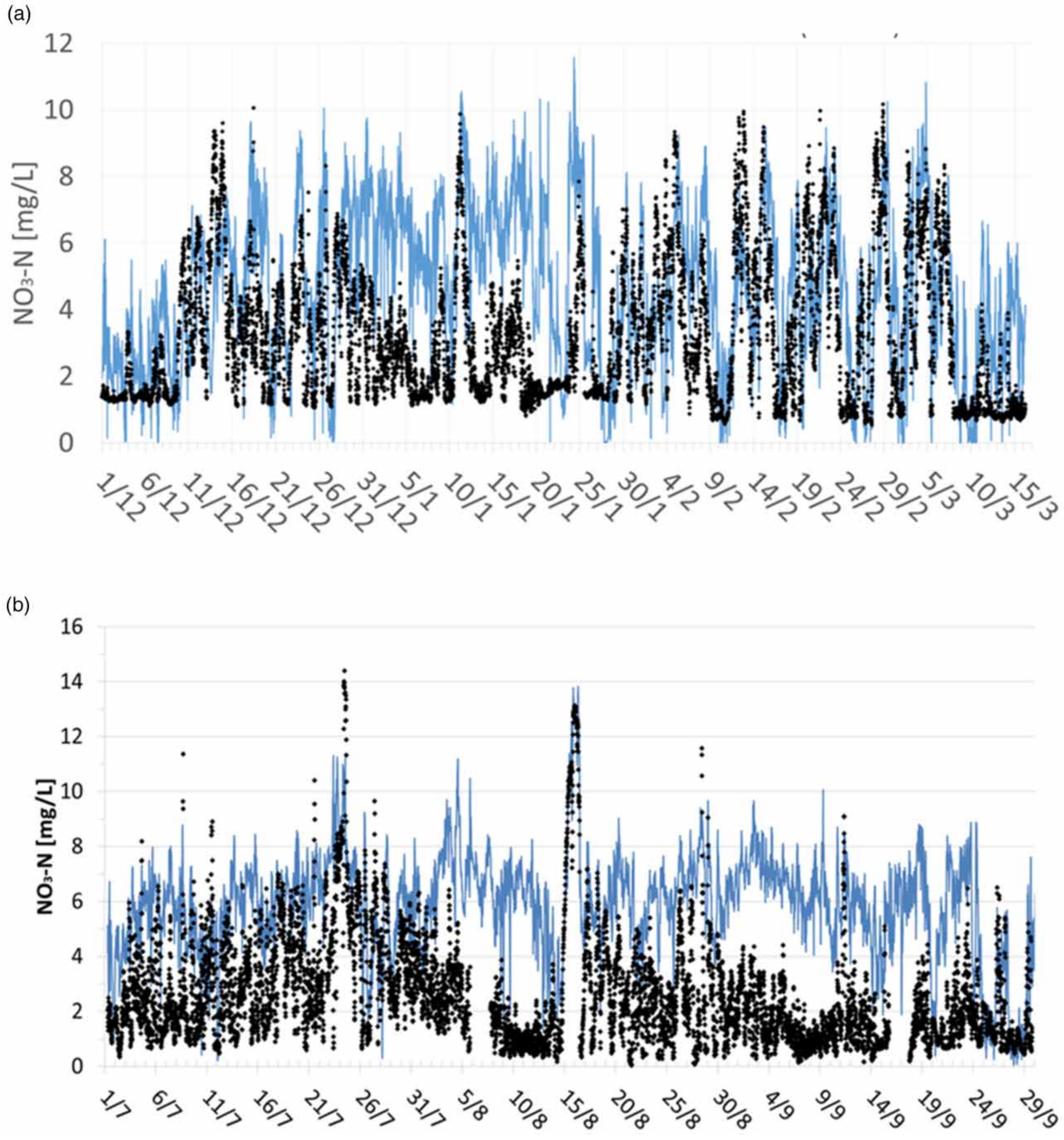


Figure 5 | Simulation results of the mechanistic model MM simulations for the NO_3-N concentration at the outlet of the preDN stage for (a) calibration, (b) validation (black points: sensor observations at 15 min interval, blue line: simulation results).

3.2. Training and testing of hybrid model

Detailed results with the training and testing data for effluent nitrate are now presented. Details on the TSS and total COD simulations with the HM are presented in the Supplementary Information.

3.2.1. Effluent NO_3-N with high-resolution data

Figure 6 compares the high-frequency (15-min interval) observations with the nitrate concentration simulations for the training (Figure 6(a)) and testing (Figure 6(b)) periods by the hybrid and mechanistic models, respectively. The outputs of the hybrid model are much closer to the observed values than to those of the mechanistic model for both of the two periods. By comparing the performance of the MM-only model with the HM, the improvement by the DDM is clearly noticeable.

Interesting to notice are the model simulated values lower than 1.0 mgNO$_3$-N/L (e.g., between January 20th and 30th). In fact, here the cleaned sensor data never dropped below 1 mg NO$_3$-N/L, a result of the lower precision for inline sensors at low concentrations. Conversely, the simulations seem to correspond better with the laboratory analysis results for composite samples that occasionally indicate daily averages below 1 mg NO$_3$-N/L.

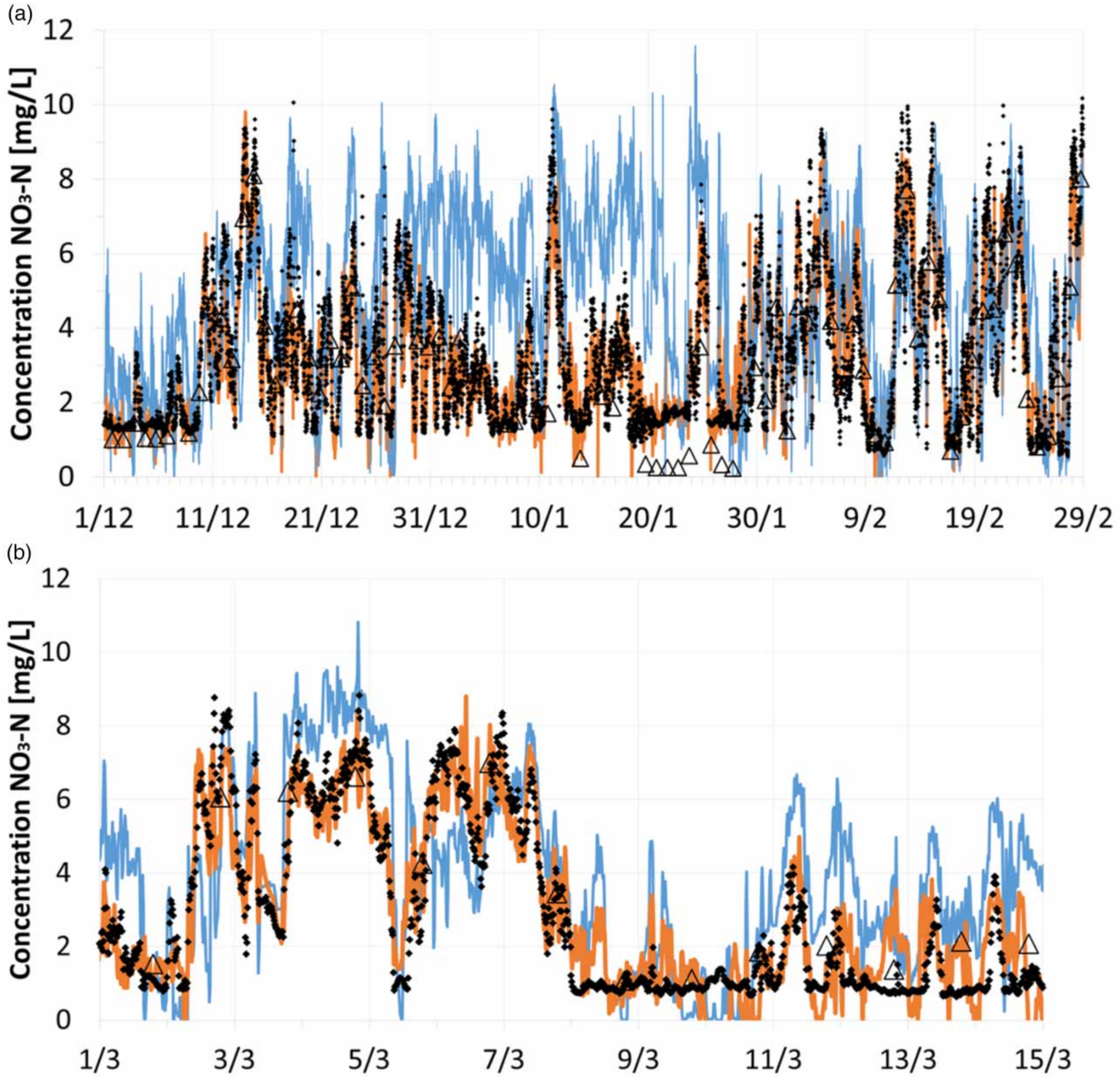


Figure 6 | Results of the training (a) and testing (b) of the hybrid model (orange line) for the NO$_3$-N concentration at the outlet of the preDN. Blue line: mechanistic model simulation; orange line: hybrid model results; black points: high-resolution observations; black-outlined triangles: laboratory composite observations.

Figure 6(b) shows that the hybrid model simulates well the trends and variability of the nitrate concentration over the testing period and performs much better than the MM. Towards the end of the testing period, however, the relative improvement of the HM seems to be lower. Starting from 8 March, a period occurs in which the HM has difficulties to compensate for the residual error of the MM simulations, this is especially in those cases when the MM predicts low NO_3-N concentrations of 1 mgN/L or less. The daily composite laboratory samples however show average values between 1.0 and 2.0 mgNO_3-N/L, which conceal the high variability in the 15-min values that can temporarily reach extremely high and low values.

This reduced performance could be due to different reasons, such as a lack of good quality measurements (due to the lower precision of the nitrate sensor below 1 mg NO_3-N/L) or less reliable simulations by the MM model in the lower range of nitrate and nitrite concentrations. A lower performance with the test data could indicate that the data from the training period does not contain a sufficiently wide range of information to train the DDM. Still, the Janus coefficient of 1.0 has the optimal value and does not indicate an error of over- or under-fitting.

The statistical performance indicators are listed in Table 2 for the simulations of the effluent variables at the preDN outlet with the hybrid model and are compared to the MM. In this table, the performance of the models is relative to the observed data from the calibration/training and validation/testing periods; MM-calibration (1 December 2019 to 15 March 2020); MM-validation (01 July 2020 30 September 2020); HM-training (1 December 2019 to 29 February 2020) and HM-testing (01 March 2020 to 15 March 2020). The MM model validation period runs from July to September 2020, a summer period characterized by low flows during the holiday month of August, conditions very different from the calibration period, making for an ambitious validation challenge. For this reason, the DDM was not trained or tested with the summer dataset.

For the effluent nitrate, the HM outperforms the MM with relative mean errors that are much smaller during training and testing for the HM compared to the calibration and validation of the MM, indicating a much smaller bias for the HM. The other performance criteria (MAE and RMSE) demonstrate a better approximation of the dynamics by the HM. The Janus coefficient, with an optimal value of 1.0, indicates that the RMSE values are nearly identical for training and testing, such that the model can be considered valid for nitrate simulations. Overall, even though the relative MAE and relative RMSE range from 24 to 32%, the scores remain globally acceptable given the highly dynamic conditions of the process and considering the sensor measurement limitations.

3.2.2. Effluent NO_3-N with daily values

In contrast to high-resolution data with a time interval of 15 min, which can capture short-term fluctuations in wastewater conditions and are therefore more appropriate for real-time control operations, daily average values may be more appropriate for process analysis purposes because they provide a noise-reduction and provide a better indication of overall performance instead of being determined by short-term extreme values.

Figure 7 plots the daily flow-proportional averages of nitrate effluent concentrations calculated from the simulations by the mechanistic and the hybrid model as compared to the daily flow-proportional averaged sensor data for the training period. Plotted as well are the daily composite flow-proportional laboratory samples. It is noticed that the HM simulations are much closer to the daily averaged sensor data, with an R^2 value of 0.95, as compared to the daily averages of the MM with an R^2 of only 0.21. The MM tends to overestimate the nitrate effluent concentrations, especially in the period between 24 December 2019 and 25 January 2020. The HM is capable of correcting for this and reducing the overall relative ME to 2% under training conditions (see Table 2).

The daily laboratory analysis results follow in general the daily average values of the sensor data, with a few notable exceptions (e.g., a short period from 19 to 22 of January, 1, 12 and 28 February). The HM is here not able to compensate since the

Table 2 | Statistical performance indicators for the MM and HM models during calibration/training and validation/testing with high-frequency data for the effluent NO_3-N concentrations simulated at the outlet of the preDN process

			Calibration (MM) or training (HM)					Validation (MM) or testing (HM)					
Variable	Model	Unit	*N*	Mean obs ($\bar{y}$)	ME/$\bar{y}$	MAE/$\bar{y}$	RMSE/$\bar{y}$	*N*	Mean obs ($\bar{y}$)	ME/$\bar{y}$	MAE/$\bar{y}$	RMSE/$\bar{y}$	Janus coef.
NO_3-N	MM	gN/m^3	10,656	2.89	−47%	67%	86%	8,015	2.62	−123%	129%	146%	1.5
NO_3-N	HM	gN/m^3	8,736	2.94	−11%	25%	34%	1,440	3.23	8%	24%	32%	1.0

Shown are the relative errors compared to the observed mean. *N*, number of observations; $\bar{y}$ is the observed mean.

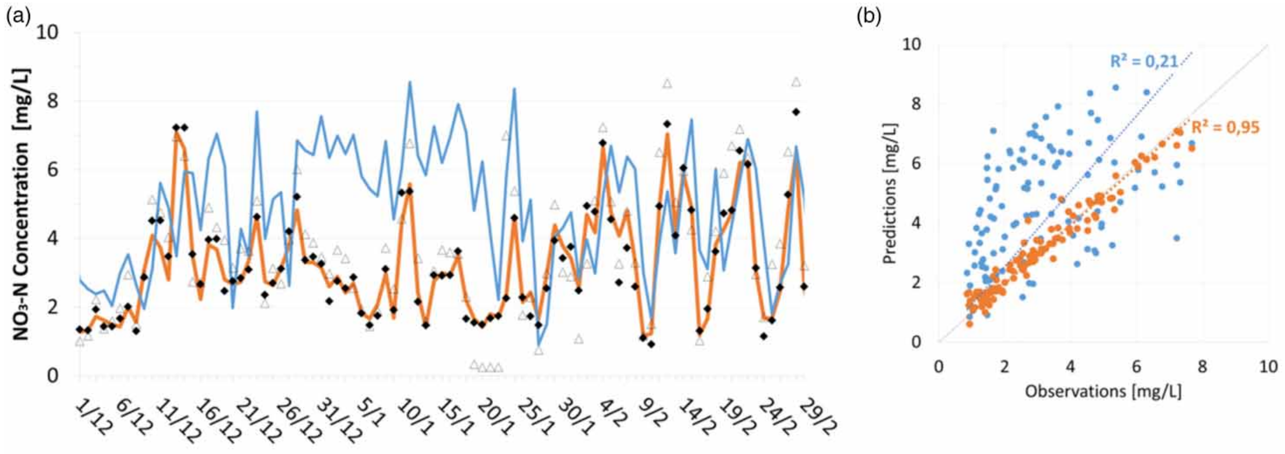


Figure 7 | (a) Performance of hybrid model (orange) and mechanistic model (blue) evaluated for NO_3-N effluent daily average sensor data for the training of the HM at 15-min interval; orange line: connected points of the daily average of the HM simulations; black diamonds: daily average of the inline observations; black-outlined triangles: daily laboratory samples. (b) Parity plot of the simulated data of the MM and HM compared to the observed data using daily averaged sensor values.

DDM was trained on high-resolution measurements and is thus unaware of the discrepancy between laboratory and sensor observations.

Concerning the results of the testing period with the daily averaged values (Figure 8), the overall high performance of the HM was maintained when compared to the daily averages of the sensor data with a R^2 of 0.97 (see Figure 8(b)). This could be explained by the fact that the daily averages are to a much lower degree affected by the high-frequency dynamics around the 1 mgNO_3-N/L that falls in the lower precision range of the nitrate sensor, as is the case with the high-resolution data with a time interval of 15 min.

The daily averages of the simulations with the mechanistic model follow the trend of the operational data, but beginning on 8 March they start to show an overestimation. This is confirmed in the parity plot in Figure 8(b), where the mechanistic model only has an R^2 value of 0.21 for the same data period.

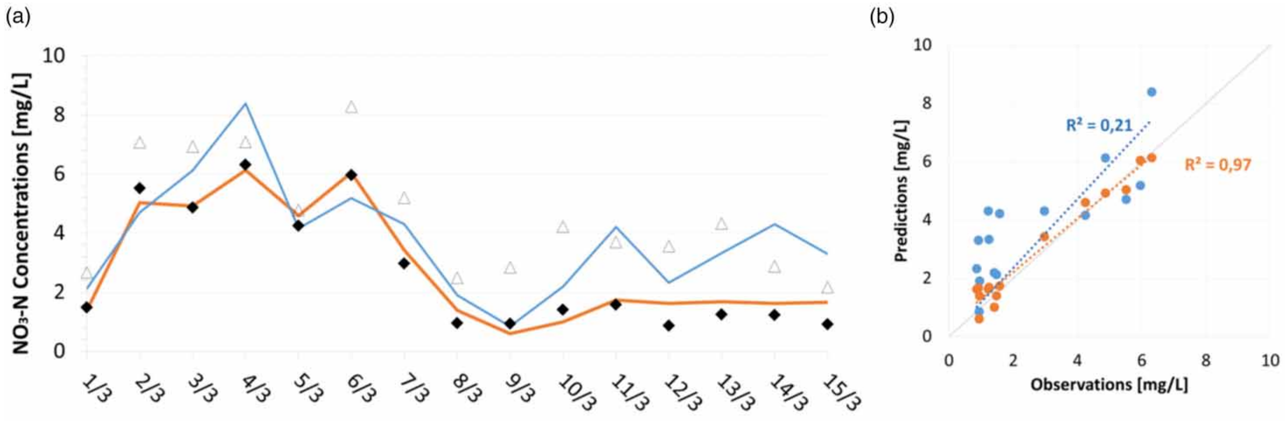


Figure 8 | (a) Results of the test of the MM (blue) and HM (orange) models for the NO_3-N concentration at the outlet of the preDN using daily averaged values of the 15 min simulation results. Blue line: connected points of daily average of the mechanistic model simulations; orange line: connected points of daily average of the hybrid model simulation results; black diamonds: daily average of the online observations; black-outlined triangles: daily laboratory samples. (b) Parity plot of the simulated data of the HM and MM models compared to the observed data using daily averaged values.

4. CONCLUSIONS

A parallel hybrid model was developed in which a calibrated/validated mechanistic biofilm model of a submerged biofilter for nitrogen removal is corrected by a data-driven neural network model to improve the accuracy of water quality simulations. For the first time, a mechanistic model and a DDM were calibrated and trained on high-frequent operational data collected at a very large municipal WRRF by inline sensors and supported by daily laboratory sampling analyses over a long period of 3.5 months.

The simulations obtained over the training and testing periods well establish the validity of the hybrid model built by the addition of a feed-forward neural network: the trained DDM can pick-up 'hidden' dynamic information that corrects the mechanistic model's residual error. The validation error (15 min interval) of the hybrid model is three times smaller than the error of the mechanistic-only model. This excellent performance is confirmed by the strong Janus coefficients, which evaluate the capability of the model to perform under unseen conditions.

Considering the highly dynamic conditions observed in the treatment plant being studied, the considerable measurement errors observed in the monitoring data, and the sensitivity of the model calculations to outliers and faults in the input data, it is fair to conclude that the modelling results could further benefit from more intense (and automated) data preparation and reconciliation processes, as well as more data from longer training and testing periods.

Altogether, the results of this study with the developed hybrid model indicate that the data-driven component captures sufficient residual information to compensate for the inaccuracy of the mechanistic model. This shows that mechanistic models, in particular those developed for biofiltration, require assumptions to be made that fail to include all dynamics, some of which can be 'recovered' by the hybrid model. The hybrid model reduces the size of residual errors and makes predictions of future states more reliable, thus improving the confidence in the model's output.

ACKNOWLEDGEMENTS

The authors would like to thank the operators of the Seine-aval WRRF (SIAAP) for access to the facility and collection of the data.

FUNDING

This work was supported by SIAAP (Paris, France) as part of the Mocopée Research Program (SIAAP 2018) that includes partnerships with model *EAU* – Université Laval, LEESU and W-SMART (innEAUvation.fr). The Natural Sciences and Engineering Research Council of Canada awarded a Discovery grant to model *EAU* entitled: 'Towards digital twin based control of water resource recovery facilities – Methods supporting the use of adaptive hybrid digital twins'.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Alferes, J. & Vanrolleghem, P. A. 2016 Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. *AI Communications* **29** (6), 701–709.

Anderson, J. S., McAvoy, T. J. & Hao, O. J. 2000 Use of hybrid models in wastewater systems. *Industrial & Engineering Chemistry Research* **39** (6), 1694–1704.

Behrendt, J. 1999 Modeling of aerated upflow fixed bed reactors for nitrification. *Water Science and Technology* **39** (4), 85–92.

Bernier, J., Rocher, V., Guerin, S. & Lessard, P. 2014 Modelling the nitrification in a full-scale tertiary biological aerated filter unit. *Bioprocess and Biosystems Engineering* **37** (2), 289–300.

Boltz, J. P., Morgenroth, E., Brockmann, D., Bott, C., Gellner, W. J. & Vanrolleghem, P. A. 2011 Systematic evaluation of biofilm models for engineering practice: Components and critical assumptions. *Water Science and Technology* **64** (4), 930–944.

Bramer, M. 2007 Avoiding overfitting of decision trees. *Principles of Data Mining*. Springer Verlag, Heidelberg, Germany, pp. 119–134.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Ces, U. & Poch, M. 2018 Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling & Software* **106**, 89–103.

Dors, M., Simutis, R. & Lübbert, A. 1995 *Hybrid Process Modeling for Advanced Process State Estimation, Prediction, and Control Exemplified in A Production-Scale Mammalian Cell Culture*. In: Rogers, K., Mulchandani, A. & Zhou, W. (eds) ACS Symposium Series (613): Biosensor and Chemical Sensor Technology. Process Monitoring and Control, ACS, Washington, DC, USA, pp. 144–145.

Duarte, M. S., Martins, G., Oliveira, P., Fernandes, B., Ferreira, E. C., Alves, M. M., Lopes, F., Pereira, M. A. & Novais, P. 2023 A review of computational modeling in wastewater treatment processes. *ACS ES&T Water* **4** (3), 784–804.

Eberl, H., Morgenroth, E., Noguera, D. R., Picioreanu, C., Rittmann, B., Van Loosdrecht, M. & Wanner, O. 2006 *Mathematical Modelling of Biofilms (IWA Task Group on Biofilm Modelling, ed) Scientific and Technical Report No. 18*. IWA Publishing, London, UK.

Falkentoft, C. M., Harremoes, P., Mosbæk, H. & Wilderer, P. 2000 Combined denitrification and phosphorus removal in a biofilter. *Water Science and Technology* **41** (4–5), 493–501.

Fiat, J., Filali, A., Fayolle, Y., Bernier, J., Rocher, V., Sperandio, M. & Gillot, S. 2019 Considering the plug-flow behavior of the gas phase in nitrifying BAF models significantly improves the prediction of N2O emissions. *Water Research* **156**, 337–346.

Galvanauskas, V., Simutis, R. & Lübbert, A. 2004 Hybrid process models for process optimisation, monitoring and control. *Bioprocess and Biosystems Engineering* **26**, 393–400.

Garrido-Baserba, M., Corominas, L., Ces, U., Rosso, D. & Poch, M. 2020 The fourth revolution in the water sector encounters the digital revolution. *Environmental Science & Technology* **54** (8), 4698–4705.

Ghosh, D., Hermonat, E., Mhaskar, P., Snowling, S. & Goel, R. 2019 Hybrid modeling approach integrating first-principles models with subspace identification. *Industrial & Engineering Chemistry Research* **58**, 13533–13543.

Hannaford, N. E., Heaps, S. E., Nye, T. M. W., Curtis, T. P., Allen, B., Golightly, A. & Wilkinson, D. J. 2023 A sparse Bayesian hierarchical vector autoregressive model for microbial dynamics in a wastewater treatment plant. *Computational Statistics & Data Analysis* **179**, 107659.

Hauduc, H., Neumann, M. B., Muschalla, D., Gamerith, V., Gillot, S. & Vanrolleghem, P. A. 2015 Efficiency criteria for environmental model quality assessment: A review and its application to wastewater treatment. *Environmental Modelling & Software* **68**, 196–204.

Helsel, D. R. & Hirsch, R. M. 1992 *Statistical Methods in Water Resources*, Vol. 49. Elsevier, Amsterdam, The Netherlands.

Henze, M., Gujer, W., Mino, T. & Van Loosdrecht, M. 2000 *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA Publishing, London, UK.

Henze, M., Van Loosdrecht, M., Ekama, G. & Brdjanovic, D. 2008 *Biological Wastewater Treatment*. IWA Publishing, London, UK.

Hidaka, T. & Tsuno, H. 2004 Development of a biological filtration model applied for advanced treatment of sewage. *Water Research* **38** (2), 335–346.

Ingildsen, P. & Olsson, G. 2016 *Smart Water Utilities. Complexity Made Simple*. IWA Publishing, London, UK.

Lee, D. S., Jeon, C. O., Park, J. M. & Chang, K. S. 2002 Hybrid neural network modeling of a full-scale industrial wastewater treatment process. *Biotechnology and Bioengineering* **78** (6), 670–682.

Lee, D., Vanrolleghem, P. A. & Park, J. 2005 Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. *Journal of Biotechnology* **115**, 317–328.

Li, F. & Vanrolleghem, P. A. 2022 An influential generator for WRRF design and operation based on a recurrent neural network with multi-objective optimization using a genetic algorithm. *Water Science and Technology* **85** (5), 1444–1453.

Li, B., Taniguchi, D., Gedara, J. P., Gogulancea, V., Gonzalez-Cabaleiro, R., Chen, J., McGough, A. S., Ofiteru, I. D., Curtis, T. P. & Zuliani, P. 2019 NUFEB: A massively parallel simulator for individual-based modelling of microbial communities. *PLOS Computational Biology* **15** (12), e1007125.

Newhart, K. B., Holloway, R. W., Hering, A. S. & Cath, T. Y. 2019 Data-driven performance analyses of wastewater treatment plants: A review. *Water Research* **157**, 498–513.

Nti, I. K., Nyarko-Boateng, O. & Aning, J. 2021 Performance of machine learning algorithms with different K values in K-fold cross-validation. *International Journal of Information Technology & Computer Science* **6**, 61–71.

Ohasi, H., Sugawara, T., Kikuchi, K. & Konno, H. 1981 Correlation of liquid-side mass-transfer coefficient for single particles and fixed-beds. *Journal of Chemical Engineering of Japan* **14** (6), 433–438.

Peres, J., Oliveira, R. & De Azevedo, S. F. 2001 Knowledge based modular networks for process modelling and control. *Computers & Chemical Engineering* **25** (4–6), 783–791.

Pocquet, M., Wu, Z., Queinnec, I. & Spérandio, M. 2016 A two pathway model for N_2O emissions by ammonium oxidizing bacteria supported by the NO/N_2O variation. *Water Research* **88**, 948–959.

Psichogios, D. C. & Ungar, L. H. 1992 A hybrid neural network-first principles approach to process modeling. *AIChE Journal* **38**, 1499–1511.

Quaghebeur, W., Torfs, E., De Baets, B. & Nopens, I. 2022 Hybrid differential equations: Integrating mechanistic and data-driven techniques for modelling of water systems. *Water Research* **213**, 118–166.

Rieger, L., Gillot, S., Langergraber, G., Ohtsuki, T., Shaw, A., Takacs, I. & Winkler, S. 2012 *Guidelines for Using Activated Sludge Models. Scientific and Technical Report No. 22*. IWA Publishing, London, UK.

Rittmann, B. E., Boltz, J. P., Brockmann, D., Daigger, G. T., Morgenroth, E., Sørensen, K. H., Takacs, I., Van Loosdrecht, M. & Vanrolleghem, P. A. 2018 A framework for good biofilm reactor modelling practice (GBRMP). *Water Science and Technology* **77** (5), 1149–1164.

Samie, G., Lessard, P. & Rocher, V. 2010 Simulation du comportement d'unités de biofiltration des eaux usées. *Techniques, Sciences, Méthodes* **11**, 85.

Samie, G., Bernier, J., Rocher, V. & Lessard, P. 2011 Modeling nitrogen removal for a denitrification biofilter. *Bioprocess and Biosystems Engineering* **34**, 747–755.

Schneider, M. Y., Quaghbeur, W., Borzoei, S., Froemelt, A., Li, F., Saagi, R. & Torfs, E. 2022 Hybrid modelling of water resource recovery facilities: Status and opportunities. *Water Science and Technology* **85** (9), 2503–2524.

Schuppert, A. & Mrziglod, T. 2018 Hybrid model identification and discrimination with practical examples from the chemical industry. In: Glassey, J. & von Stosch, M. (eds) *Hybrid Modeling in Process Industries*. CRC Press, London, UK, pp. 63–88.

Serrao, M. 2023 *Towards Intelligent Process Control of Municipal Wastewater Treatment: The Development of A Hybrid Model That Aims to Improve Simulation Performance and Process Optimization*. PhD Thesis, École National des Ponts et Chaussées – ParisTech, France.

Serrao, M., Jauzein, V., Azimi, S., Rocher, V., Tassin, B. & Vanrolleghem, P. A. 2023 Hybridizing a first-principles biofilm model with a data-based model to improve model accuracy for model predictive control of a 6 million PE WRRF. In *Proceedings WEF/IWA Innovations in Process Engineering*, 6–9 June 2023, Portland, OR, USA.

Shirkoohi, M. G., Tyagi, R. D., Vanrolleghem, P. A. & Drogui, P. 2022 Modelling and optimization of psychoactive pharmaceutical caffeine removal by electrochemical oxidation process: A comparative study between response surface methodology (RSM) and adaptive neuro fuzzy inference system (ANFIS). *Separation and Purification Technology* **290**, 120902.

SIAAP 2018 Innovate in the monitoring and operating practices of wastewater treatment plants. Scientific and technical lessons learned from phase I (2014–2017) of the Mocopée program. Report, SIAAP, Paris, France.

Sparks, J., Vanrolleghem, P. A., Bott, C. & Wadhawan, T. 2022 It's OK to be a Control Freak: Deploying machine learning algorithms and model-based controllers for WRRF optimization. In: *Proceedings: WEF Innovations in Process Engineering Conference (WEF/IPE)*, 20–23 June 2022, Miami, FL, USA.

Sundui, B., Ramirez Calderon, O. A., Abdeldayem, O. M., Lázaro-Gil, J., Rene, E. R. & Sambuu, U. 2021 Applications of machine learning algorithms for biological wastewater treatment: Updates and perspectives. *Clean Technologies and Environmental Policy* **23**, 127–143.

Torfs, E., Nicolai, N., Daneshgar, S., Copp, J. B., Haimi, H., Ikumi, D., Johnson, B., Plosz, B. B., Snowling, S., Townley, L. R., Valverde-Perez, B., Vanrolleghem, P. A., Vezzaro, L. & Nopens, I. 2022 The transition from WRRF models to digital twin applications. *Water Science and Technology* **85** (10), 2840–2853.

Tsen, A. Y.-D., Jang, S. S., Wong, D. S. H. & Joseph, B. 1996 Predictive control of quality in batch polymerization using hybrid ANN models. *AIChE Journal* **42**, 455–465.

Vanrolleghem, P. A., Benedetti, L. & Meirlaen, J. 2005 Modelling and real-time control of the integrated urban wastewater system. *Environmental Modelling & Software* **20** (4), 427–442.

Vigne, E., Choubert, J. M., Canler, J. P., Héduit, A., Sorensen, K. & Lessard, P. 2010 A biofiltration model for tertiary nitrification of municipal wastewaters. *Water Research* **44** (15), 4399–4410.

Viotti, P., Eramo, B., Boni, M. R., Carucci, A., Leccese, M. & Sbaffoni, S. 2002 Development and calibration of a mathematical model for the simulation of the biofiltration process. *Advances in Environmental Research* **7** (1), 11–33.

Von Stosch, M., Oliveira, R., Peres, J. & De Azevedo, S. F. 2014a Hybrid semi-parametric modelling in process systems engineering: Past, present and future. *Computers & Chemical Engineering* **60**, 86–101.

Von Stosch, M., Davy, S., Francois, K., Galvanauskas, V., Hamelink, J. M., Luebbert, A., Mayer, M., Oliveira, R., O'Kennedy, R., Rice, P. & Glassey, J. 2014b Hybrid modeling for quality by design and PAT-benefits and challenges of applications in biopharmaceutical industry. *Biotechnology Journal* **9** (6), 719–726.

Zhu, J. 2020 *Detailed Modelling of the Functioning of the Complete Biofiltration Sector of the Seine-Aval Wastewater Treatment Plant*. PhD Thesis, University of Technology of Compiègne, Sorbonne University, Paris, France.

Zhu, J., Bernier, J., Pauss, A., Vanrolleghem, P. A. & Rocher, V. 2018 Modelling of the Seine-aval station – Towards real-time optimization of operating and environmental costs. In *Proceedings Water Information Days*, 11 October 2018, Poitiers, France.

First received 7 February 2024; accepted in revised form 9 August 2024. Available online 27 August 2024