



Wastewater treatment plant operation optimization towards energy efficiency and reduced effluent loading based on optimized data-driven models

Norbert Botond Mihaly, Elisabeta Cristina Timis^{*}, Nicolae Cebotaru, Marius Adrian Brehar, Vasile Mircea Cristea

Babes-Bolyai University, Faculty of Chemistry and Chemical Engineering, Department of Chemical Engineering, 11 Arany Janos St, Cluj-Napoca, Cluj, 400028, Romania

ARTICLE INFO

Keywords:

WWTP
Data-driven modelling
ANN
WWTP energy efficiency
Reduced effluent loading

ABSTRACT

Traditional wastewater treatment plant operations often face challenges in balancing treatment performance with energy consumption, particularly under dynamic influent conditions. This study presents a novel approach to model wastewater treatment plants by leveraging statistical, data-driven and optimization methods to enhance both classical modelling and recent data-driven modelling approaches. Historical operational influent and effluent data from a full-scale plant was used to accurately predict effluent key performance indicators, such as total nitrogen and orthophosphate concentrations. The optimized data-driven model development methodology resulted in the most accurate models for all output features with model performance indicators of up to $R^2 = 0.99$ and mean absolute percentage error $< 1\%$. The best performing models were utilized to identify potential operation strategies of a one day real scenario. The multi-objective optimization aimed towards improved energy efficiency and reduced effluent loading by changing the aeration and external recycle flowrates. A potential energy consumption reduction of more than 20% was revealed by application of correction factors to the reference flowrate values coupled with marginal effluent loading reduction. The potential of intelligent modelling tools to facilitate sustainable and effective wastewater treatment plant operation is highlighted by this study.

1. Introduction

Within the framework of circular economy, wastewater management offers a huge potential in the face of several issues regarding environmental degradation, resource availability, climate change, and the rising need for sustainability (Elharbili et al., 2024). At the same time, pollutants released by commercial, industrial, residential, and agricultural operations can pose serious threats to human health and the environment (Batool et al., 2024). Despite sewage pollution being largely unrecognized (Elgarahy et al., 2021) pollutants might undesirably be released into natural water bodies when the effluent quality exceeds regulations under shocking loads, accidents or faulty operation at wastewater treatment plants (WWTPs), (Yin et al., 2025). To provide WWTP operators ample time to prevent such pollutant discharge breaches, water quality prediction tools can provide water quality

predictions and provide early warnings of standard overshooting (Yu et al., 2024). Moreover, water quality prediction tools support process optimization (Kumar and Goyal, 2025) and the wider evaluation of WWTPs environmental impact (Dikmen et al., 2025) and its minimization via water quality indicators, energy consumption (Hoz et al., 2026; S. Zhang et al., 2023) or greenhouse gas emissions (J. Chen et al., 2025; Z. Zhang et al., 2025), separately or combined (J. Chen et al., 2024; Du et al., 2025). In this context, energy use remains a major concern for wastewater management due to electricity consumption and to the indirect emissions related to conventional energy sources (Zakariazadeh et al., 2024) and has been intensively investigated with a wide range of methods, among which the artificial intelligence (AI) ones are very successful (J. Chen et al., 2025).

Among AI tools data-driven models (DDMs) gained appeal lately in such applications related to the urban environment (Majnoon and

This article is part of a special issue entitled: "SDEWES25" published in Cleaner Engineering and Technology.

^{*} Corresponding author. Babes-Bolyai University, Department of Chemical Engineering, 11 Arany Janos St., 400028, Cluj-Napoca, Cluj, Romania.

E-mail addresses: norbert.mihaly@ubbcluj.ro (N.B. Mihaly), elisabeta.timis@ubbcluj.ro (E.C. Timis), nicolae.cebotaru@stud.ubbcluj.ro (N. Cebotaru), marius.brehar@ubbcluj.ro (M.A. Brehar), mircea.cristea@ubbcluj.ro (V.M. Cristea).

<https://doi.org/10.1016/j.clet.2026.101207>

Received 30 December 2025; Received in revised form 7 March 2026; Accepted 5 April 2026

Available online 30 April 2026

2666-7908/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Saifoddin, 2025) due to their accuracy and because they do not require a thorough grasp of the underlying processes, otherwise necessary for the development of first-principle models (Dai et al., 2025; Picos-Benítez et al., 2020). WWTP effluent prediction models (De la Hoz-M et al., 2024), operation optimization tools (Kumar and Goyal, 2025), as well as more complex digital tools (Ma et al., 2025) were developed with the help of DDMs based on various techniques. Random forest has been widely used for effluent parameters prediction (Peerzade and Kamat, 2025), for energy consumption optimization (Du et al., 2025) and for energy consumption predictions with R^2 of 0.69 (Bagherzadeh et al., 2021) and with R^2 of 0.723 (Huang et al., 2023). Support vector machine (SVM) has been employed to predict a wide range of parameters (Du et al., 2025), among which the total nitrogen (TN) concentration in three WWTPs with R^2 values from 0.73 to 0.93 (Ly et al., 2022). A recent comparison between random forest, SVM and Decision Trees (DT) revealed random forest as the best predictor of TN (with R^2 of 0.70), biological oxygen demand (BOD, with R^2 of 0.56) and chemical oxygen demand (COD, with R^2 of 0.74), while DT was the best predictor for ammonia with R^2 of 0.36 (Peerzade and Kamat, 2025). Gaussian process (GP) was applied in hybrid structures with mathematical models for full-scale WWTP modelling (Hvala & Kocijan, 2020), and with artificial neural network (ANN) models for real-time effluent quality predictions (Shaban et al., 2024). Multiple types of ANN models were utilized for the modelling of various aspects of WWTPs (X. Chen et al., 2025) and for process optimization (Wei et al., 2023) providing significant advantages to leverage the plant performance improvement (Zamfir et al., 2025). These methods were found to be precise for steady state data, such as daily averages, however, their dynamic counterparts were better suited for higher frequency time series predictions (Alali et al., 2023). However, DDMs face challenges when presented with extensive datasets showing nonlinear and non-steady state characteristics (Ly et al., 2022), high-dimensionality, noise or unbalances (J. Chen et al., 2024).

To guarantee sustainable processes and regulatory compliance, such sophisticated modelling approaches that can precisely forecast WWTPs performance, optimize operating parameters and support environmental impact decrease are necessary (Pan et al., 2025). Robust AI and hybrid tools based on comprehensive process data and trusted techniques are being developed (Zamfir et al., 2025). Principal component analysis (PCA) is one of the multivariate statistical methods frequently used in the wastewater industry for improving model efficiency and predictive accuracy (Grace N et al., 2025). PCA has been successfully applied in hybrid model development involving ANNs, to predict water quality parameters (X. Chen et al., 2025), showing good results together with the nonlinear autoregressive networks with exogenous inputs (NARX) for the prediction of COD and TN (Yang et al., 2022). Optimization of DDM parameters was previously studied and proven to be an efficient method to find the most suitable NARX models (N.-B. Mihály et al., 2023) and to improve prediction accuracy and DDM stability (Xiong et al., 2025).

The strong potential of hybrid models evolved considerably during the last five years (X. Chen et al., 2025; Y. Chen et al., 2020; De la Hoz-M et al., 2024) and it has been demonstrated in a wide range of wastewater treatment directions (Xu et al., 2024), via robust tools (D. Zhang et al., 2026) with improved interpretability and predictive performance compared to conventional DDM (Wei et al., 2023), especially in complex operation scenarios (Yin et al., 2025). The majority of existing studies including DDM or hybrid models, such as PCA-ELM (extreme learning machine, ELM), (Grace N et al., 2025), PCA-NARX (Yang et al., 2022) or PCA-MLP (multi-layer perceptron, MLP) (Ramya et al., 2024) are based on simulated WWTP data (Kini et al., 2025), on laboratory data (Pan et al., 2025) or on limited measurements from a full-scale WWTP, in terms of measurements frequency (e.g., daily) (Dikmen et al., 2025; Jassim et al., 2025), and/or regarding the amount of data (Elharbili et al., 2024). Simulated data is widely employed, because the access to WWTP measurements may be too expensive (Wei et al., 2023), restricted, subjected to confidentiality policies or too expensive (Dai

et al., 2025). Though, significant discrepancies have been reported between predictions based on measurements and predictions based on simulated data (Du et al., 2025) and small datasets limit model robustness and applicability (Jassim et al., 2025). Therefore, there is a need of studies based on measurements reflecting real-time continuous process operation from full scale processes, over sufficient time. On the other hand, the majority of existing hybrid modelling studies validate each component independently (Elharbili et al., 2024; Kini et al., 2025), and it has been noted that limitations associated to feature selection (Dikmen et al., 2025) need to be attended.

Existing wastewater modelling approaches generally address only part of the complexity of full-scale WWTP operation. Dynamic PCA-NARX models based on daily data provide moderate accuracy but limited robustness and hyperparameter control (Yang et al., 2022). Architectures such as the probabilistic encoder-decoder long short-term memory (P-ED-LSTM) model improve resilience under shock-load conditions but remain restricted to single-pollutant (TN) prediction and lack dimensionality reduction or operational optimization components (Yin et al., 2025). Likewise, PCA-ELM frameworks deliver high static prediction accuracy but do not capture dynamic behaviour or provide optimization capabilities (Grace N et al., 2025), while ensemble-based PCA-XGBoost-NSGA-II methods focus mainly on COD and energy, these rely on small daily datasets and lack dynamic modelling (J. Chen et al., 2024). Newer deep spatio-temporal models such as the three-dimensional convolutional neural network – LSTM – Gaussian process regression (3D CNN–LSTM–GPR), though high-performing for TP prediction (R^2 of 0.92), are computationally intensive, single-pollutant, and do not support operation-level optimization (Shaban et al., 2024). These approaches offer valuable insights but remain limited in robustness, multi-pollutant capability, and operational applicability.

In contrast, this study introduces a unified, high-frequency, PCA-enhanced, dynamically optimized modelling framework for multi-pollutant prediction (TN and OP) that surpasses existing models in accuracy (R^2 up to 0.99; MAPE <1%), interpretability, and computational efficiency, while uniquely integrating real operational optimization for pollutant emission and energy use. The research focuses on two aspects: (1) Employing a real, new, high frequency dataset from the Cluj-Napoca WWTP for the integration of PCA-HO-DDM methodologies to enhance the classic DDM concerning the prediction of effluent loading, i.e., orthophosphates (OP) and TN, and comparison of their performance to the novel PCA-GA-NARX model; (2) To use the best performing model for WWTP operation optimization concerning the energy efficiency and effluent loading in the context of WWTP equipment limits and operator's acceptance. This comprehensive modelling-to-operations pipeline, absent from the aforementioned methods, constitutes a significant advancement toward energy-efficient and robust WWTP management, as models are coupled with real operational constraints to deliver actionable optimization of aeration and energy use. Moreover, our framework is less computationally demanding, more interpretable, supports multiple pollutants, and integrates operational optimization.

2. Methodology

The utilized dataset and modelling techniques will be presented in the following sections, where input and output parameters are referred to as features and the model parameters referred to as hyperparameters. The above presented methods have been implemented with the help of MATLAB software.

2.1. Description of data and principal component analysis

PCA is an unsupervised technique which extracts and analyses relevant information from large datasets. In the current study PCA was applied as a pre-processing step for the data-driven models to reduce the input feature space. The variance residual shows how important each

variable is, and it estimates the correlation structure of the process variables. As such, a smaller number of features can be selected which retain most of the original dataset's variance.

The employed dataset contains two months-worth of 22 sensor readings at 10-min frequency (i.e., 8352 points in time) from the Cluj-Napoca municipal WWTP (~400k population equivalents) which has an anaerobic-anoxic-oxic treatment system. The dataset contains the following initial (before PCA) input features: WWTP influent flowrate; influent temperature; influent total suspended solids (TSS); influent soluble COD; influent ammonia concentration; soluble COD after primary clarifier; flowrate from the primary clarifier to bioreactors; flowrate from anoxic section to oxic section; air flowrate for the oxic section; dissolved oxygen (DO) measured at two points of each of the four parallel oxic reactors; flowrate from bioreactors to secondary clarifier; external recycle flowrate. Another two sensor readings were utilized as output features for the models, i.e., effluent TN and OP. These 22 sensor readings were selected from more than 50 sensors across the WWTP due to these not having any missing data over the studied 2-month period. All sensors are periodically calibrated by certified Hach-Lange specialists under a premium maintenance and service contract, while calibration values are independently verified by the operator's accredited on-site analytical laboratory, ensuring traceability and metrological integrity. Raw time-series data are exported either directly from the Supervisory Control and Data Acquisition (SCADA) system or from dedicated analytical instruments, after which they undergo systematic pre-processing including the removal of physically impossible or implausible values, and detection and handling of statistical outliers. These procedures mitigate the influence of sensor drift, communication artefacts, and short-term measurement anomalies. Looking ahead, the operator is preparing a transition to secure cloud-based data storage fully compliant with Directive (EU) 2022/2555 (NIS2), which explicitly classifies water sector infrastructure, including SCADA, as essential. Collectively, these measures ensure a high-quality dataset that enhances the credibility and reproducibility of the modelling framework.

As such, the dataset resulting from the sensor readings were utilized close to their raw form. The only pre-processing step before the application of PCA was the reduction of the temporal frequency (down-sampling) of the dataset by moving average, as the original sensor readings occur at a 10-s frequency. In this manner, the data volume is greatly reduced while noise is also significantly diminished.

2.2. Data-driven modelling methods

This research includes a family of dimension-reduced, optimization-enhanced data-driven models, involving random tree regression (RTR), random forest regression (RFR), support vector regression (SVR), GPR and NARX (implemented also as deep neural network), all selected to align with the dynamic complexity of wastewater treatment processes while outperforming alternative static, deep, or ensemble models reported in recent literature. NARX networks have a well-documented advantage in environmental time-series modelling due to their ability to encode multi-lag nonlinear relationships and reproduce feedback-driven system states. It has been demonstrated that PCA-NARX significantly outperforms static ANNs (multilayer perceptron network (MLP) and radial basis function neural network (RBF)) for wastewater indicators due to its ability to embed temporal feedback into the modelling structure (Yang et al., 2022). Moreover, Yin et al. (2025) showed that deep recurrent architectures outperform memory-less models under shock-load events, highlighting the importance of dynamic modeling for real-world variability. On the other hand, the strong nonlinear patterns in WWTP data, with high potential of sensor noise, stochastic influent variability, and missing data, motivate the use of GPR and SVR, which outperform linear models, especially under fluctuating influent. GPR is ideal for small-to-medium datasets, offering probabilistic predictions and uncertainty quantification even for high-dimensional problems (Nadeem et al., 2025; Shaban et al., 2024). SVR handles nonlinearities

via kernels and is robust to noise, frequently achieving very good predictive accuracy for effluent quality indicators (Sahu et al., 2023) and energy consumption (Achite et al., 2024). RTs have been long used (Atanasova and Kompare, 2002), successfully, for modelling wastewater processes because it provide an interpretable baseline for nonlinearities due to the transparent rule-based structure and ability to capture hierarchical relationships among process variables. RFR builds on the strengths of RTR, providing enhanced predictive power, noise robustness, and generalization across variable influent conditions. RFR has repeatedly demonstrated state-of-the-art performance in wastewater modelling (Cheng et al., 2025) outperforming other regression algorithms (Sun et al., 2024) or neural networks (Zhou et al., 2019).

The chosen techniques have been successfully employed for the prediction of wastewater treatment (Nadeem et al., 2025) and represent an optimal compromise between dynamic fidelity, nonlinear modelling strength, generalization, interpretability, and computational efficiency. The mathematical description of the investigated models is provided in the Supplementary Materials, while other relevant details are discussed in the sub-sections below.

2.2.1. Nonlinear autoregressive network with exogenous inputs

The tapped delay line block between the ANN model time-series inputs and the first hidden layer is used in the design of the NARX type of feed-forward neural network to provide the ANN models a strong memory feature characteristic. This line enables the prediction of the target variable value at the proximate sampling moment of time using data from previous time moments in case of both the input and the target datasets. Due to this architecture, NARX models are ideal for time-series predictions. These NARX “open” structure that feeds the original target dataset to the ANN inputs standing for the past values of the targets. The NARX “closed” structure utilizes the ANN with self-simulated past target values to compute the target value at the subsequent moment of time, which was utilized for applications after the models were already developed.

The “open” structure of the NARX model was also implemented as a deep neural network (DNARX) which allows for different solvers to be utilized in the training process, such as adaptive moment estimation (*adam*), stochastic gradient descent with momentum (*sgdm*) and root mean square propagation (*rmsprop*). As a result, the graphics processing unit (GPU) may be utilized for the training process, a crucial aspect for intricate networks and large datasets. In this manner, the development process can become orders of magnitude faster than when performed on the central processing unit (CPU), depending on this hardware.

2.3. Development and optimization of data-driven models

The models input and output features were normalized prior to model development. The accuracy and generalization capability of DDMs strongly depend on finding the best model hyperparameters for the given problem. At the same time, overfitting the training model is to be avoided. This was ensured by a five-fold cross-validation method, except for the NARX models, for which the data was divided into subsets for training, validation and testing of which the validation subset was utilized to avoid overfitting. Bayesian optimization (BO) was utilized for all DDM hyperparameters except the NARX models in which case genetic algorithm (GA) was employed. The RT and RFR model hyperparameters that were subject to the optimization were: minimum leaf size, ensemble method, number of learners, learning rate and number of predictors to sample. In case of the SVR the optimized hyperparameters were as follows: kernel function, box constraint, kernel scale and ϵ . For the GPR the following hyperparameters were optimized: basis function, kernel function, kernel scale and σ .

In the case of PCA-GA-NARX and PCA-GA-DNARX models, the previously developed GA-NARX methodology (N.B. Mihály et al., 2023) was utilized and extended with further hyperparameters employing GA based optimization: training algorithm, number of hidden layers,

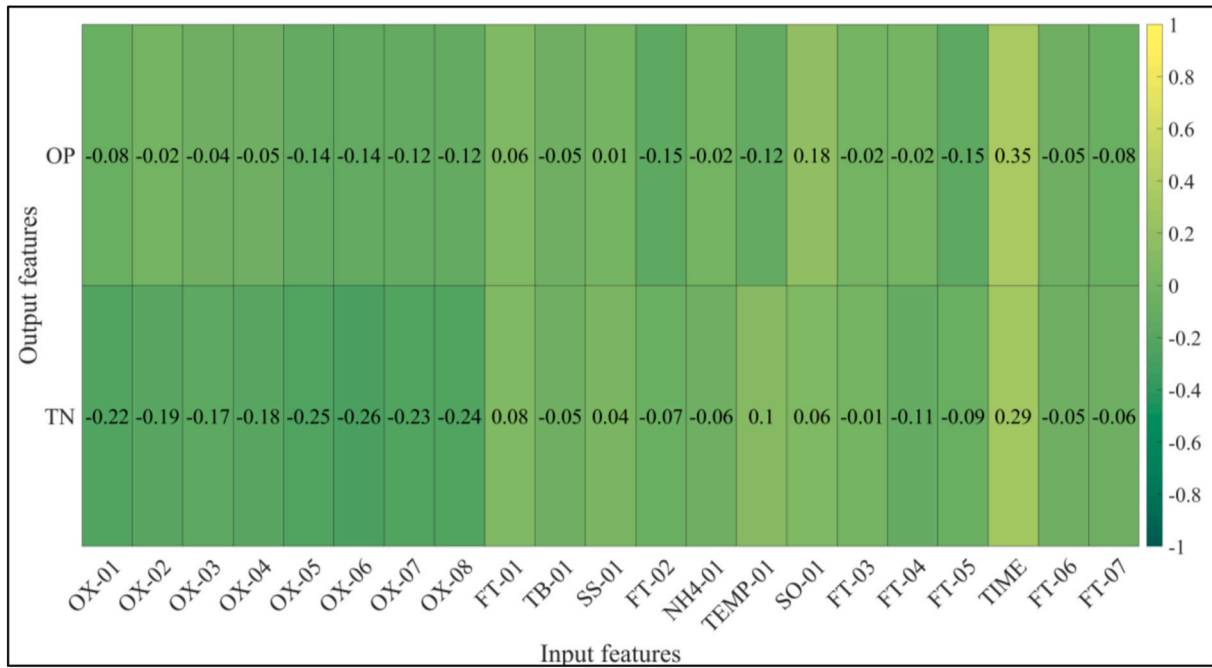


Fig. 1. Correlation coefficients for output and input features measured at the same point in time.

number of hidden neurons in each layer, transfer function of each hidden layer, percentage of data utilized for training and tapped delay line horizons.

All developed models have been further evaluated on a testing dataset (i.e., 10% of the original dataset) which was separate from the training subset based on three criteria, i.e., coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE), described in equations (1)–(4).

$$R^2 = 1 - \frac{\sum_{i=1}^N (t_i - s_i)^2}{\sum_{i=1}^N (\bar{t} - t_i)^2} \quad (\text{equation 1})$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (t_i - s_i)^2} \quad (\text{equation 2})$$

$$MAE = \frac{\sum_{i=1}^N |t_i - s_i|}{N} \quad (\text{equation 3})$$

$$MAPE = \frac{\sum_{i=1}^N \frac{|t_i - s_i|}{t_i}}{N} \cdot 100 \quad (\text{equation 4})$$

where t_i represents the desired value of the output feature at point i , $\bar{t}$ is the average value of the output feature and s_i is the value of the output calculated by the model at point i .

2.4. Optimization study energy efficiency and effluent loading

The best performing models have been employed to carry out optimization studies for the WWTP, considering input features related to the WWTP operation, i.e., air flowrate (AF) for the oxic section and external recycle flowrate (RF). AF and RF were chosen as decision variables as these can be manipulated at the WWTP and clearly influence both, the energy efficiency and the effluent loading. Aeration can account for up to 60% of energy consumption (Siatou et al., 2020), while the external recycle flowrate influences the mean cell retention time which in turn affects the treatment process (Cheng et al., 2024).

A 24-h period was left out of the DDM development process from the end of the original dataset, to be solely utilized for the optimization studies.

As both the energy efficiency and the effluent loading are of importance for a sustainable WWTP operation, a multi-objective optimization approach was implemented. A variation of the non-dominated sorting genetic algorithm II (NSGA-II) (Deb, 2001) was employed in the search of the Pareto-front. The two objective functions were both described as percentage of reduction compared to the reference case (i.e., the original 24-h dataset). In case of the first objective the effluent loading was expressed in TN equivalents as the OP concentration was multiplied by a factor of 16. This factor was chosen as the loading limit of TN is 8 mg/L, while that for total phosphorus (TP) is 0.5 mg/L (European Parliament and Council, 2024). In case of the second objective, both AF and RF were expressed in specific energy consumption utilizing the 0.05 kWh/m³ of air and 0.008 kWh/m³ (Alex et al., 2008) factors, respectively. Both objectives were calculated as the relative change compared to the reference case (i.e., the original 24-h scenario) and expressed in percentage as shown in the equations below.

$$f_{min} = \begin{cases} \frac{16 \cdot (\overline{OP} - \overline{OP}_{ref}) + (\overline{TN} - \overline{TN}_{ref})}{16 \cdot \overline{OP}_{ref} + \overline{TN}_{ref}} \cdot 100 \\ \frac{0.05 \cdot (\overline{AF} - \overline{AF}_{ref}) + 0.008 \cdot (\overline{RF} - \overline{RF}_{ref})}{0.05 \cdot \overline{AF}_{ref} + 0.008 \cdot \overline{RF}_{ref}} \cdot 100 \end{cases} \quad (\text{equation 5})$$

where, $\overline{OP}$, $\overline{TN}$, $\overline{AF}$ and $\overline{RF}$ represent the average concentrations and flowrates of the optimization solution, while $\overline{OP}_{ref}$, $\overline{TN}_{ref}$, $\overline{AF}_{ref}$ and $\overline{RF}_{ref}$ represent the average concentrations and flowrates of the reference case.

3. Results and discussion

3.1. Data description and principal component analysis

Firstly, the correlation of input and output features was studied. The Pearson correlation coefficients presented as a heatmap in Fig. 1 show the pronounced nonlinearity between the output and input features for data collected at the same point in time. The largest absolute values are around 0.3 in case of both output features for the time input. In case of

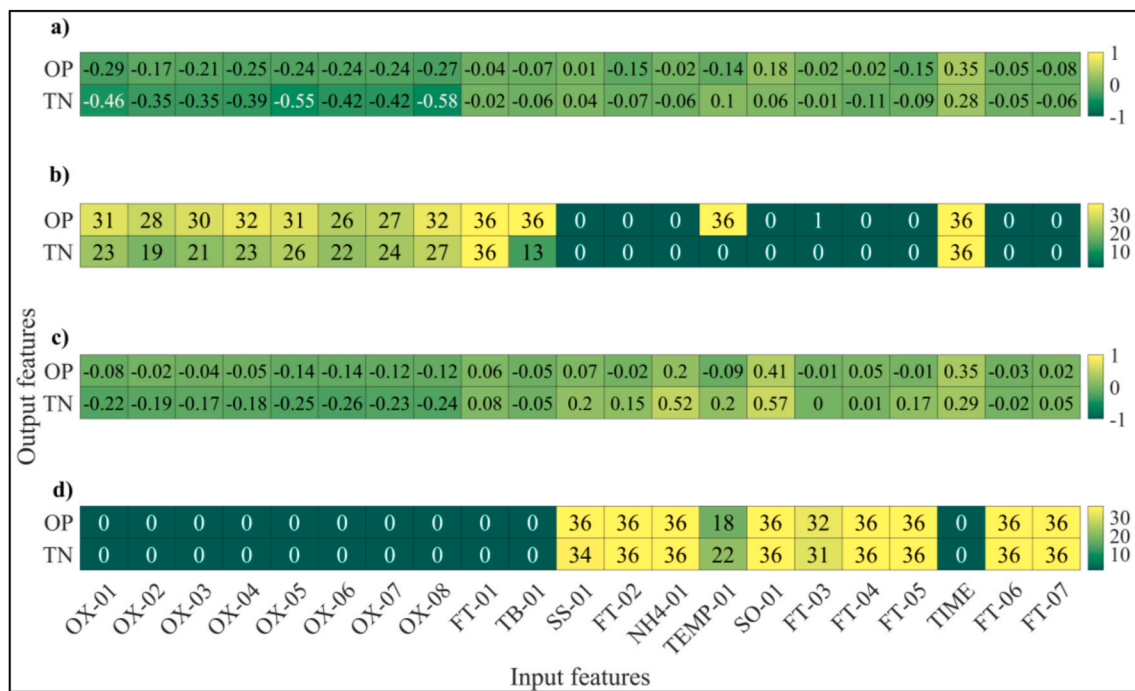


Fig. 2. a) Minimum correlation coefficient values over the investigated time delay; b) Time delay values at which the minimum correlation coefficients were obtained; c) Maximum correlation coefficient values over the investigated time delay; d) Time delay values at which the maximum correlation coefficients were obtained.

strictly sensor data, the largest absolute values are of 0.15–0.18 for OP output, and of 0.17–0.26 in case of TN output feature. The presented results showcase the high degree of nonlinearity of the process, i.e., absolute correlation values < 0.4 , when taking into account input and output features measured at the same time.

The influence of time delay on the correlation coefficients was also investigated, firstly up to a time delay of 36 moments which was also the upper boundary chosen in the development of the data-driven models. In this case, the minimum and maximum values of the correlation coefficients were recorded, as well as the time delay that they occurred at. These results are illustrated as heatmaps in Fig. 2. Compared to the correlations at the same point in time, some features exhibit larger linear correlation values. At the same time, the correlations even at delayed time steps cannot be considered of a high degree as the absolute values of the coefficients remain below 0.6. It is important to note that the aeration process displays a larger negative correlation for both output features and is most influential on the effluent TN and OP which will occur in 3 to 4.5 h and 4 to 5.5 h, respectively.

As the minimum and maximum correlation coefficients in a number of cases were limited by the time delay selected, these were also studied for an extended time period of around 2.5 days (i.e., time lag of 360). The results in this case are presented in the Supplementary Material, Fig. S1. A few sensors presented a medium degree of linearity with one or both output features due to the extension of the time delay, e.g., the negative correlation of the temperature with both outputs. However, in most cases the degree of linearity remained of the same degree with marginal changes in value, e.g., the ammonia ($\text{NH}_4\text{-N}$) and soluble COD (SO) sensors. The absolute values of the correlation coefficients also remained under 0.6. An interesting aspect to note is that although the negative correlation of the oxygen sensors remained the same, the positive correlation is of a much higher degree towards both output features with a delay of around 1.5 days compared to the greatest negative correlations. This can be explained by a possible over-aeration, in which case the anoxic zone drops in effectiveness due to DO carryover and the recycled nitrites and nitrates end up in the effluent 1.5 days later leading to an increase in TN. In case of OP, the high aeration can lead to

Table 1
PCA variability results.

Component	Variability %	Component	Variability %
1	50.90	11	0.76
2	11.48	12	0.65
3	10.41	13	0.59
4	8.74	14	0.52
5	6.93	15	0.42
6	2.92	16	0.29
7	1.64	17	0.23
8	1.30	18	0.17
9	0.97	19	0.15
10	0.78	20	0.12

the consumption of available carbon, resulting in a carbon-deficient system. As such, the PAOs cannot perform P-release in the anaerobic zone, which in turn affects the P-uptake of the aerobic zone.

For the PCA pre-processing step the singular vector decomposition algorithm was employed, and the first 13 principal components were selected for further utilization as these describe more than 98% of the variability as highlighted in Table 1.

The coefficients (also known as loadings) of the first five principal components (PC) that describe $>88\%$ of the input dataset's variability are presented in Fig. 3. It can be observed that the largest share of the variability (PC 1 and PC 4) can be attributed to the aeration process represented by the 8 oxygen sensors (i.e., OX-01 to OX-08), as these two PC have the largest absolute coefficient values for these sensors. Other important factors for the variability of the dataset are the two flow sensors describing the sludge production and external recycle flowrate (i.e., FT-03 and FT-06), as well as the time. The time input feature represents the date and time of the sensor readings as a numeric value counting from 0 at the start of the dataset. Time was utilized as an input feature to help in the description of daily and weekly tendencies. The PCA results show that time is indeed an important input feature as it almost solely describes PC 3.

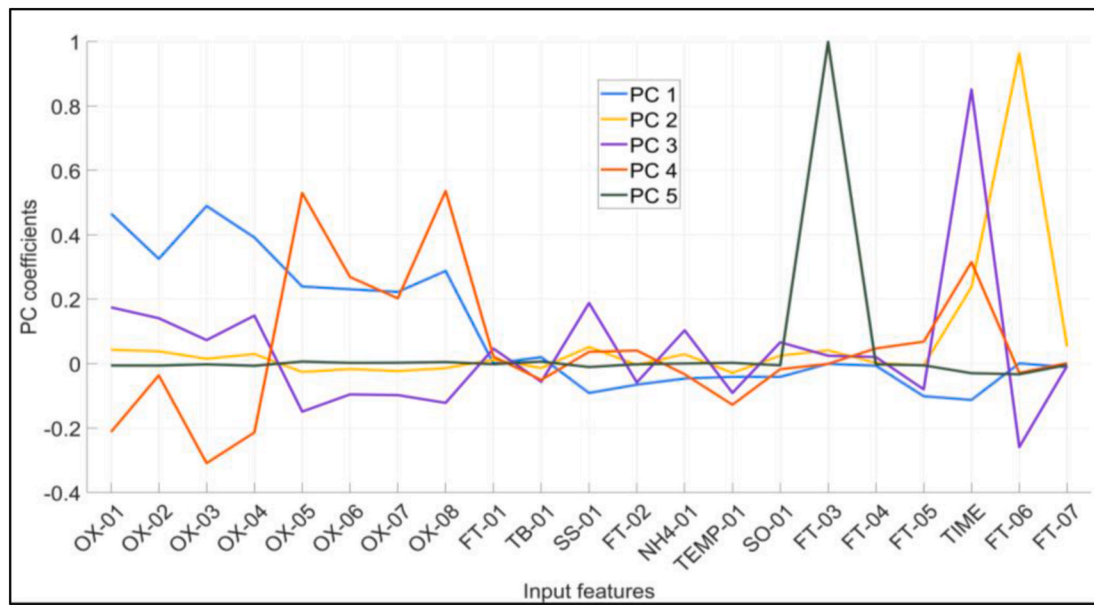


Fig. 3. Principal component (PC) coefficients in function of input features for the first five PC.

Table 2
Optimized data-driven model results.

Model	Output feature and performance indicators							
	TN				OP			
	R ²	RMSE	MAE	MAPE	R ²	RMSE	MAE	MAPE
PCA-BO-RTR	0.59	0.877	0.585	6.73%	0.63	1.936	1.002	157.33%
PCA-BO-RFR	0.79	0.630	0.445	5.13%	0.81	1.371	0.867	185.33%
PCA-BO-SVR	0.70	0.765	0.487	5.55%	0.84	1.234	0.702	158.78%
PCA-BO-GPR	0.85	0.524	0.345	3.97%	0.95	0.696	0.402	90.27%
PCA-GA-NARX	0.99	0.019	0.014	0.16%	0.99	0.141	0.075	26.98%
PCA-GA-DNARX	0.99	0.074	0.062	0.70%	0.99	0.128	0.086	23.30%

3.2. Data-driven models

All considered models were developed as multiple input single output structures, as these were reported to be superior to multiple input multiple output structures (N. B. Mihály et al., 2022). The results on the testing dataset for each of the three output features is presented in Table 2. It is important to note that although the models were developed utilizing normalized data, the prediction performance evaluation criteria were calculated utilizing the real values of the output features.

It is observed that PCA-BO-GPR outperformed the other non-ANN based methods in case of all performance criteria and for all output features. At the same time, the PCA-GA-NARX models showed better results than the PCA-BO-GPR in almost all cases. The correlation between expected and simulated output features was clearly the best for the developed PCA-GA-NARX and PCA-GA-DNARX models, obtaining R² values of 0.99. In terms of MAE, only the ANN models developed through the proposed methodology (i.e., PCA-GA-NARX and PCA-GA-DNARX) can be considered for further applications. This is due to the regulations regarding effluent TN and TP at values of 8 mg/L and 0.5 mg/L (European Parliament and Council, 2024), and only models with expected errors smaller than 0.1 were considered satisfactory. All developed models presented less ideal accuracy in terms of MAPE for the OP output feature which can be attributed to the dataset, and as such MAPE should not be a deciding factor regarding model selection for the OP output feature. This claim is also supported by the results shown in the Supplementary Material Table S1 that presents how the MAPE criteria changes in function of the OP values considered. The developed PCA-GA-DNARX is three times as accurate in terms of MAPE if only OP

Table 3
Optimized NARX type model architectures.

Model	Details
PCA-GA-NARX-TN	Levenberg-Marquardt training algorithm; tapped delay line of 12; 2 hidden layers with 13 and 11 neurons; tangent sigmoid transfer function at both hidden layers; 70-15-15 % of data to train-validation-test datasets
PCA-GA-NARX-OP	Levenberg-Marquardt training algorithm; tapped delay line of 15 and 15; 2 hidden layers with 14 and 11 neurons; logistic sigmoid and tangent sigmoid transfer functions; 80-10-10 % of data to train-validation-test datasets
PCA-GA-DNARX-TN	Rmsprop training algorithm; tapped delay line of 1; 2 hidden layers with 4 and 33 neurons; parametric rectified linear unit and rectified linear unit transfer functions; 87-6.5-6.5 % of data to train-validation-test datasets
PCA-GA-DNARX-OP	Rmsprop training algorithm; tapped delay line of 1; 2 hidden layers with 44 and 47 neurons; exponential linear unit and rectified linear unit transfer functions; 89-5.5-5.5 % of data to train-validation-test datasets

values above 0.25 mg/L are considered in the criteria calculation. When most of the data are at the lower end of the range the relative errors of the developed models tend to greatly increase. This tendency is especially evident when the targeted output feature values are closer to 0, which is the case of the PCA-GA-NARX and PCA-GA-DNARX models for OP with almost ideal R² value of 0.99, a small RMSE of 0.128 and 0.141, but MAPE between 23.3% and 27%.

It should also be noted that the deep networks (i.e., PCA-GA-DNARX) presented comparable results to those obtained in case of their shallow

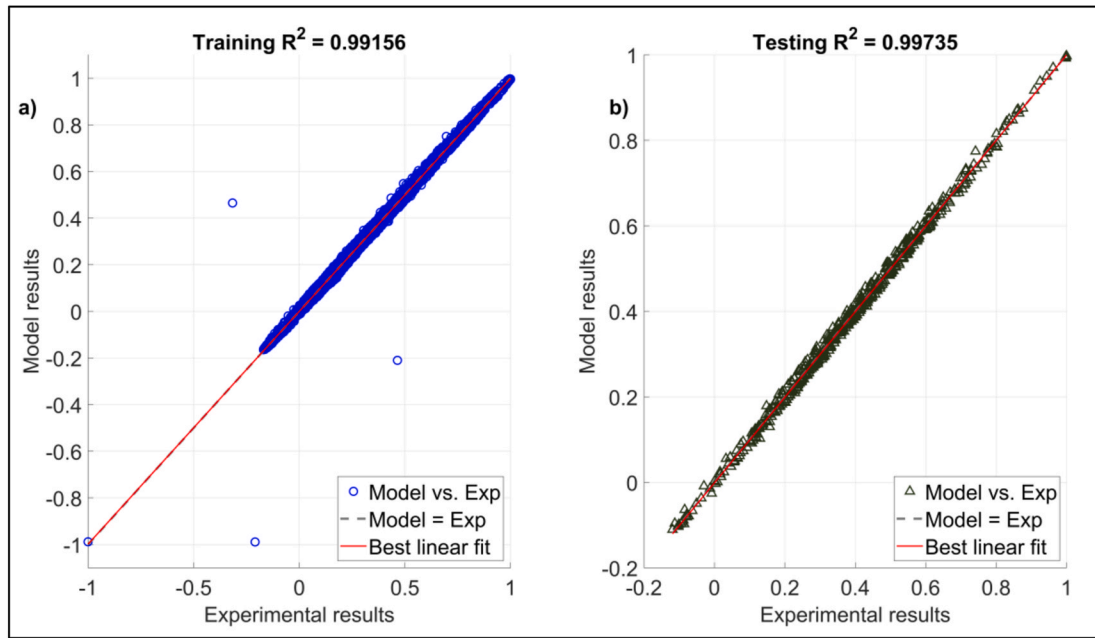


Fig. 4. PCA-GA-DNARX regression results for output feature TN in case of: a) training dataset and b) testing dataset.

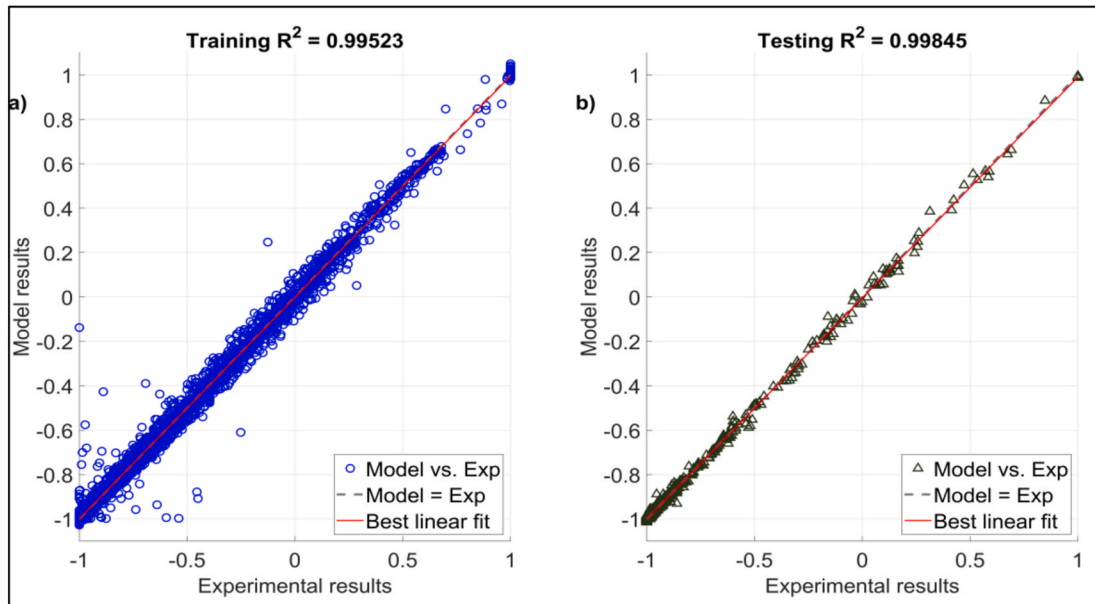


Fig. 5. PCA-GA-DNARX regression results for OP in case of: a) training dataset and b) testing dataset.

networks counterparts (i.e., PCA-GA-NARX). The architectures of the best developed networks are shown in Table 3.

The results show that for the NARX models the Levenberg-Marquardt training algorithm was best, whereas in case of the DNARX models training with the rmsprop algorithm yielded the best performing models. At the same time, the best DNARX models utilize less historical data (i.e., smaller tapped delay line) compared to the NARX models. In contrast, a larger portion of the available dataset was employed for the training of DNARX models. Another important aspect is the previously mentioned difference in computational power between the CPU and GPU. The same network architecture (i.e., number of hidden layers, number of hidden neurons and tapped delay line horizon) was trained in both manners to compare the processing times. Firstly, the NARX model was trained utilizing the Levenberg-Marquardt method on a 12th

generation Intel i5-12400F CPU. Secondly, the DNARX model was trained utilizing the rmsprop method on an NVIDIA RTX 4070 Super GPU. The development of the NARX model required 196 s, whereas the DNARX model was developed in less than 52 s, close to four times faster. The development on the CPU resulted in 12.5 to 13.5 s per epoch in the training process, while the development on the GPU showed 1.12 to 1.15 s per epoch. Consequently, the DNARX models were chosen as the best option taking into consideration the one order of magnitude smaller time requirements in terms of training epochs as well as the possibility of future models to be developed with an extended database. The accuracy of these two single-output models is illustrated in Figs. 4–6.

A good correlation can be observed in case of the PCA-GA-DNARX models for OP and TN output features which was expected from the previously presented R^2 values. Moreover, it can be observed that the

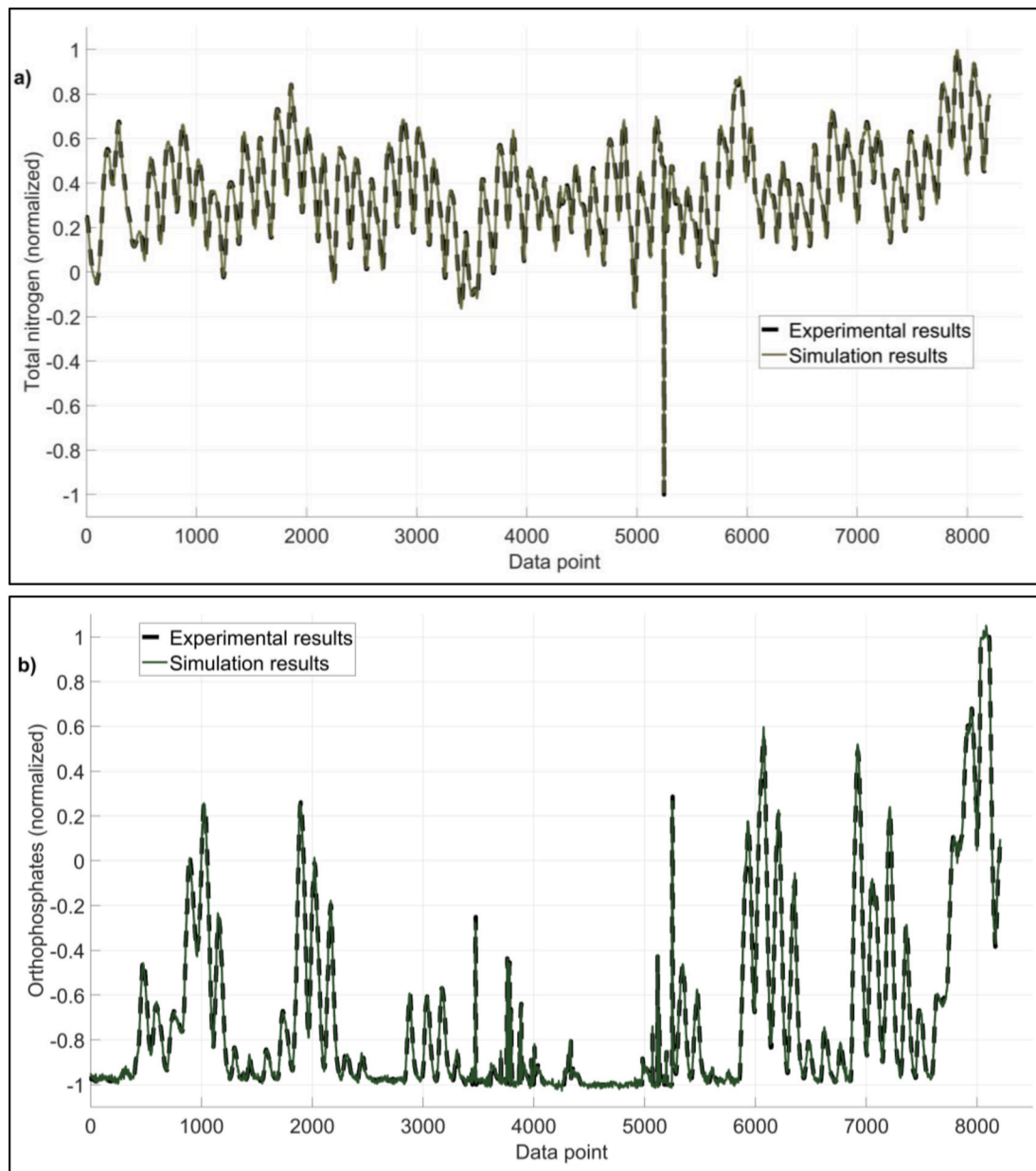


Fig. 6. PCA-GA-DNARX simulation results vs. experimental data for: a) TN and b) OP.

generalization capability of both models is satisfactory as the results for the testing datasets exhibit the same degree of accuracy as which can be seen for the training dataset. Although the determination coefficients present values close to ideal and the RMSE values also close to zero, the MAPE might be concerning in case of the OP predictions. Results in Fig. 5 confirm the analysis of results in Table 2 regarding the slightly decreased model performance for several samples of OP from the lower range, which cause the increased value of MAPE, around 27%, despite R^2 values above 0.99. Though, very good model performance in depicting the process dynamic is revealed, as in Fig. 6 the predicted series of data overlap the experimental data equally well for the OP and TN output features.

The previously discussed correlation coefficients between input and output features both at the same point in time as well as considering time delay showed that there is no high correlation or temporal autocorrelation of input features with the outputs. The training and testing results illustrate the generalization capability of the developed models. The learning curves provided in the Supplementary Materials (i.e. Fig. S2 to

Fig. S5) also attest to the avoidance of overfitting through the utilization of validation in the training process. Consequently, the obtained model performances can be attributed to the proposed PCA-HO-DDM development methodology.

3.3. Optimization studies

A 24-h period of monitoring data was utilized, and two input features, the aeration flowrate (AF) and the external recycle flowrate (RF), were considered as decision variables in the investigation. In order to utilize the previously presented PCA-GA-DNARX models, these had to be converted to a “closed” structure, their predictions to be utilized in the forecast of further values. In this manner, the models would be utilized for a larger prediction horizon compared to the ones these models were trained for (i.e., prediction horizon of one). As such, the models were first tested on this 24-h scenario to establish a prediction horizon that still resulted in accurate predictions. Thus, the 24-h scenario would be divided into smaller, continuous scenarios which in turn results in a

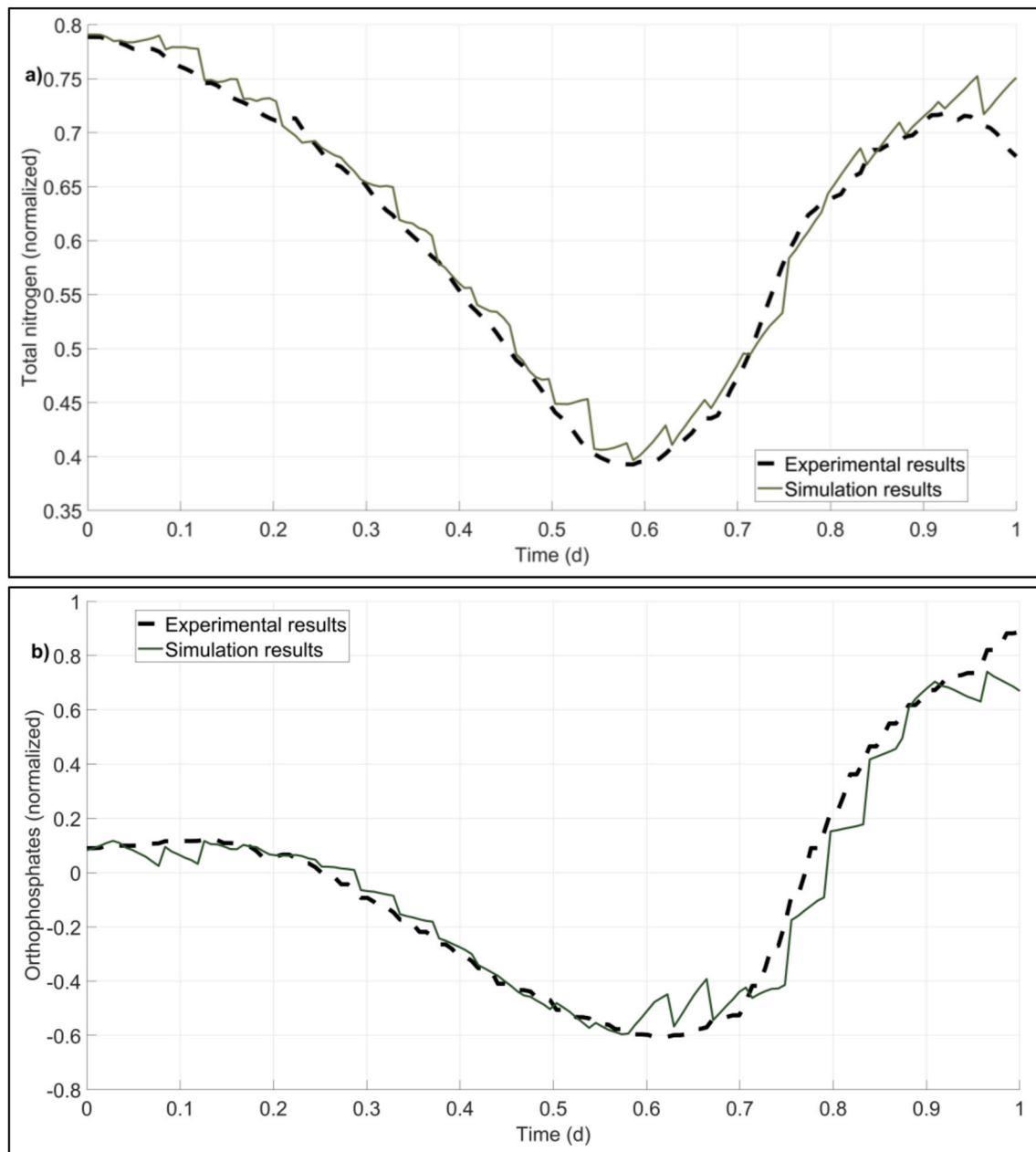


Fig. 7. PCA-GA-DNARX predictions of the 24-h scenario for: a) TN and b) OP.

larger number of decision variables and greater computational requirements. Both models showed satisfactory prediction accuracy of TN and OP with a prediction horizon of six (i.e., 1-h prediction scenarios) as illustrated in Fig. 7. In this fashion, a total of 48 decision variables are employed in the optimization studies representing the 24 hourly values for AF and RF, respectively.

Two cases were defined in how the optimal values for the decision variables were searched for. In the first case, the decision variables were considered as weights to be applied to the reference actions concerning the AF and RF of the 24-h scenarios with values searched for in the 0.5 to 1.5 region. In the second case, the decision variables were the hourly, normalized values of the AF and RF parameters with values of -1 to 1 . Results of both cases are analysed below.

The solutions found in the first case range from -40% to 7% in terms of energy consumption compared to the reference case, whereas a lesser change can be obtained in this case concerning the effluent loading, i.e., -8% to 0% . It is noted that maximizing the reduction of effluent loading

is only possible by increasing the energy consumption of the treatment process, as shown in Fig. 8a). At the same time, it is possible to operate the WWTP in a manner which results in both a diminished energy consumption as well as lower effluent loading in terms of TN and OP. Solution 7 of the Pareto-front in Fig. 8a) was chosen to showcase how this can be achieved, this solution representing approximately the middle point of the trade-off between energy consumption and effluent loading. Fig. 9 illustrates the output feature (i.e., TN and OP) predictions when the AF and RF would be set to the values shown in Fig. 10 compared to the reference represented by the original 24-h scenario.

The illustrated solution results in relative energy savings of $\sim 22\%$ and close to 5.5% reduction in effluent loading. The lower effluent loading can be attributed to lower OP values throughout the 24-h scenario, whereas the relative energy savings arise from the AF values, which in average are lower than the reference. It can be observed that the AF is, in several cases, above the reference value which shows that a more targeted approach is viable and leads to lower average energy

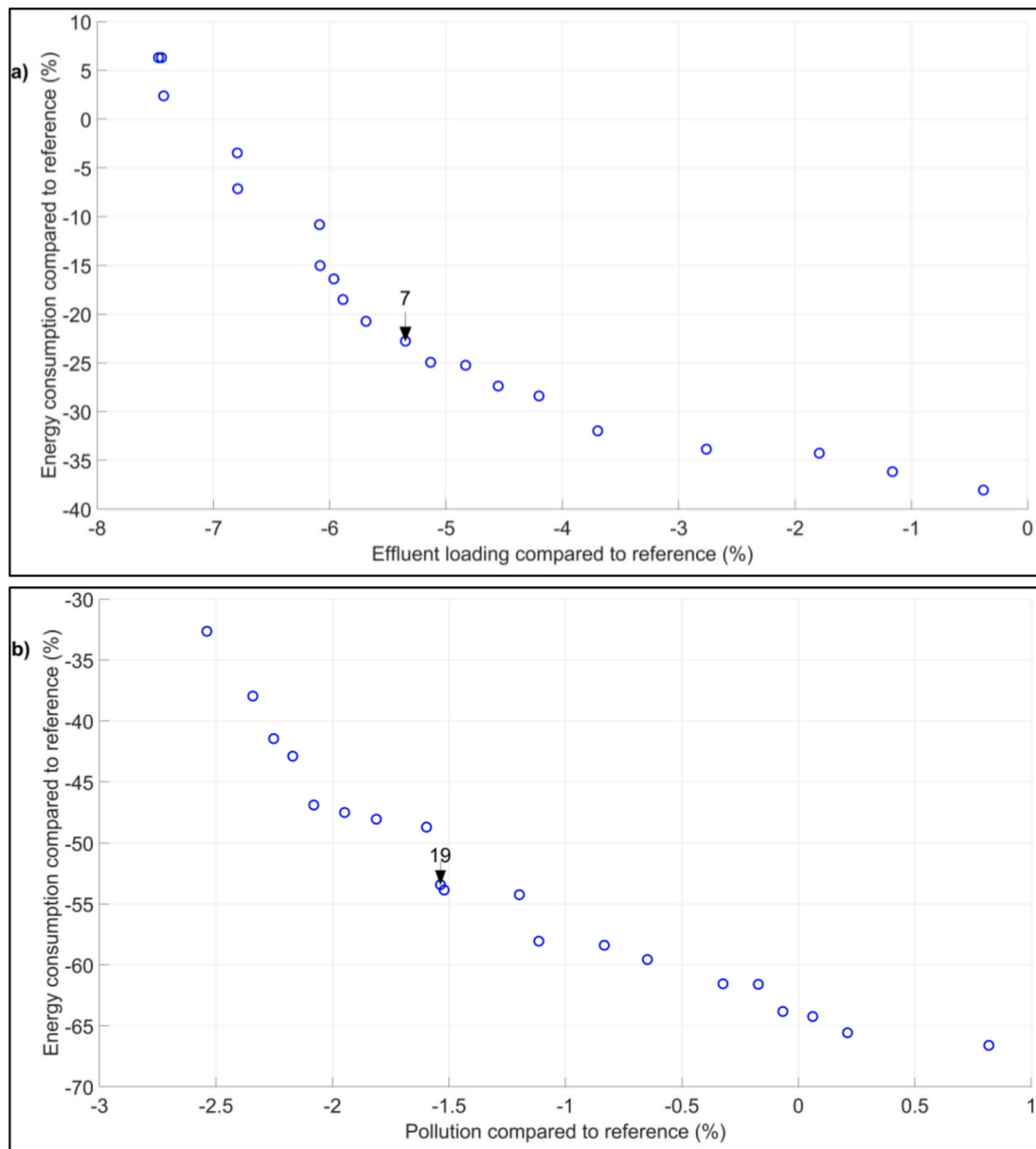


Fig. 8. Pareto-fronts of multi-objective optimizations: a) first case and b) second case.

consumption throughout the studied day. However, the implementation of this approach is only possible through model predictions and optimization, for which fast and accurate models, such as the PCA-GA-DNARX developed, are required. In case of the RF, the values periodically go to a predetermined minimum and are kept there for less than an hour which might be related to the sludge blanket (i.e., maintaining a minimum level). In this first optimization case, solution 7, both AF and RF parameters have values larger than 1 throughout the 24-h optimized scenario. In turn, the predictions of the utilized models might be less accurate as these values are out of the training range. On the other hand, these optimal values might not be possible to achieve due to equipment limitations.

The second optimization case addresses both these issues by maintaining the possible optimal solutions in the -1 to 1 range that represents the experimental range of values and the training range. As such, values in this range were already obtained with the available equipment and the models were also developed with input features in this range. Fig. 8b) presents the Pareto-front for the second multi-objective optimization case, the energy consumption in this case reached values from

-33% to -67% compared to the reference, whereas the changes in effluent loading were between -2.6% and 0.9% . The Pareto-front solution 19, which roughly represents the middle position of the trade-off between energy use and effluent loading, was selected to demonstrate how these savings may be accomplished. In comparison to the reference provided by the initial 24-h scenario Fig. 11 shows the output feature (i.e., TN and OP) predictions when the AF and RF take the values displayed in Fig. 12.

The above-mentioned solution 19 results in average relative energy savings of 53% coupled with a very slight reduction in effluent loading regarding TN and OP of 1.5% . Similarly to the first optimization case, the energy savings may be related to the reduction of AF throughout the studied 24 h, while the decision variables have a greater influence on the OP in relation to the effluent loading parameters. It can be observed that although the optimal values of RF were searched for in the -1 to 1 region, the periodical minimum flowrate was kept as in the reference case. In case of the AF, it can be observed that the values are lower throughout the whole 24-h scenario compared to the reference data which significantly cuts the energy consumption of the studied treatment process.

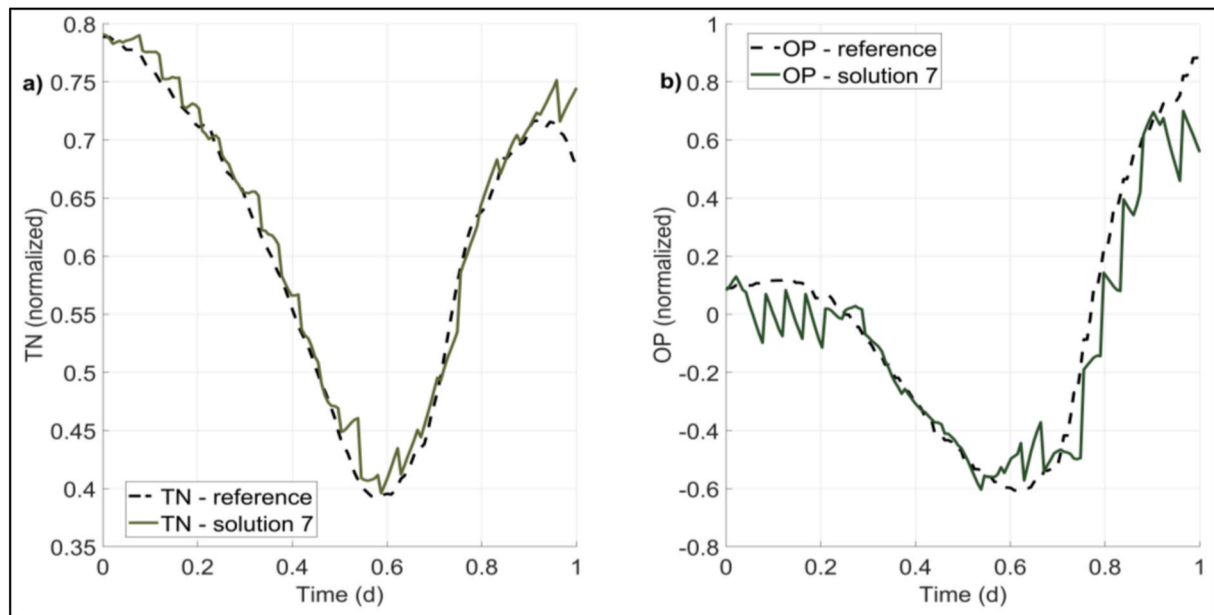


Fig. 9. Multi-objective optimization case 1, solution 7, output predictions: a) TN and b) OP.

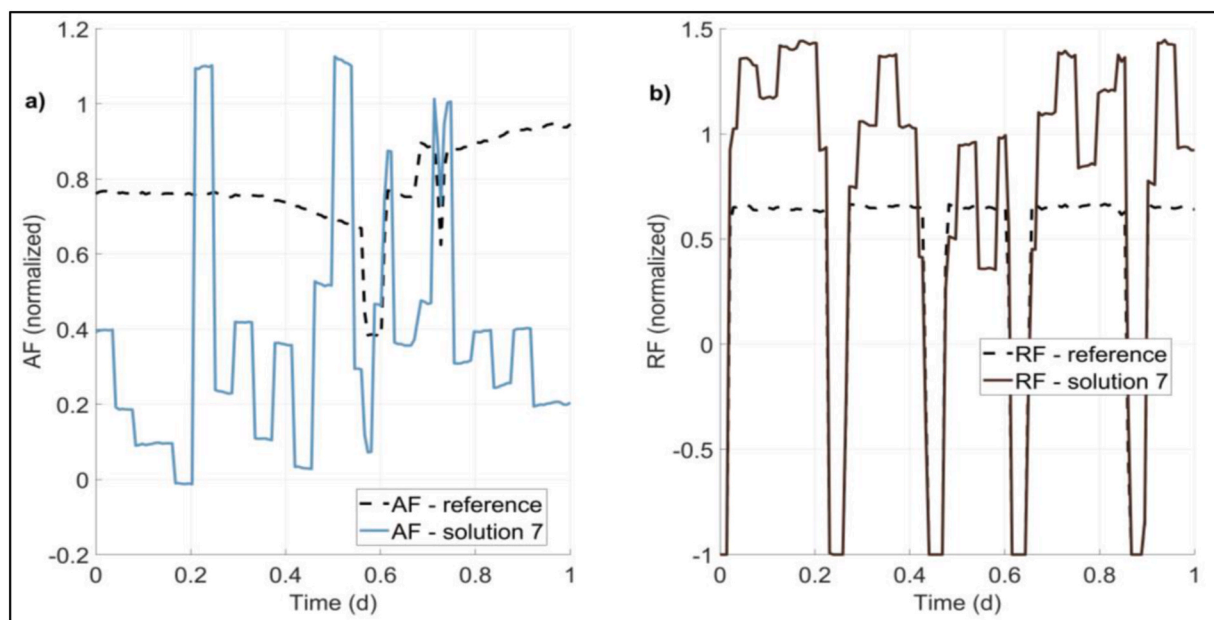


Fig. 10. Multi-objective optimization case 1, solution 7, WWTP operation variables: a) AF and b) RF.

The applicability of resulted optimal solutions is of critical importance. In both optimization cases, the AF and RF values were limited to their real bounds based on the available equipment at the studied WWTP. As such, all of the optimal results can be implemented in terms of real actuator bounds. The current dataset was obtained at the start of the year (i.e., January and February). In this period the influent flowrate was smaller compared to the highest loading of the WWTP, that usually occurs in the spring and autumn. As such, the real actuator limits of both AF and RF are of approximately 1.5 times greater than the maximum values present in the dataset. In case of AF, e.g., during these 2 months at most three out of the five identical blowers were utilized to aerate the bioreactors. In this sense, the solutions presented previously are considered as applicable for RF as well, where the single major difference were the flowrate values compared to the reference case. In terms of the AF, the optimal scenario significantly differs from the reference

case in both optimization cases. The sudden changes of the AF can be performed as a response time of 4 min is considered for the aeration system (Alex et al., 2008), while the ramping changes of the optimal solution occur over a period of 10 min. The understanding and acceptance by the operator of the optimal solutions is crucial. For optimization case 2 the lower AF values can be attributed solely to the relative energy reduction (objective 2), in which case the effluent pollutants are barely reduced or even increased compared to the reference case (Fig. 8 b)). In the first optimization case however, some of the proposed actions regarding AF do not relate to this second objective, i.e., the three time periods where the optimal AF is above the normalized value of 1. The further investigation of the optimal solution revealed that the increases of the AF are alike predictive actions as all of these time periods are followed by increases in effluent loading, as illustrated in Fig. 13. These results provide a better understanding of the optimal solution and help

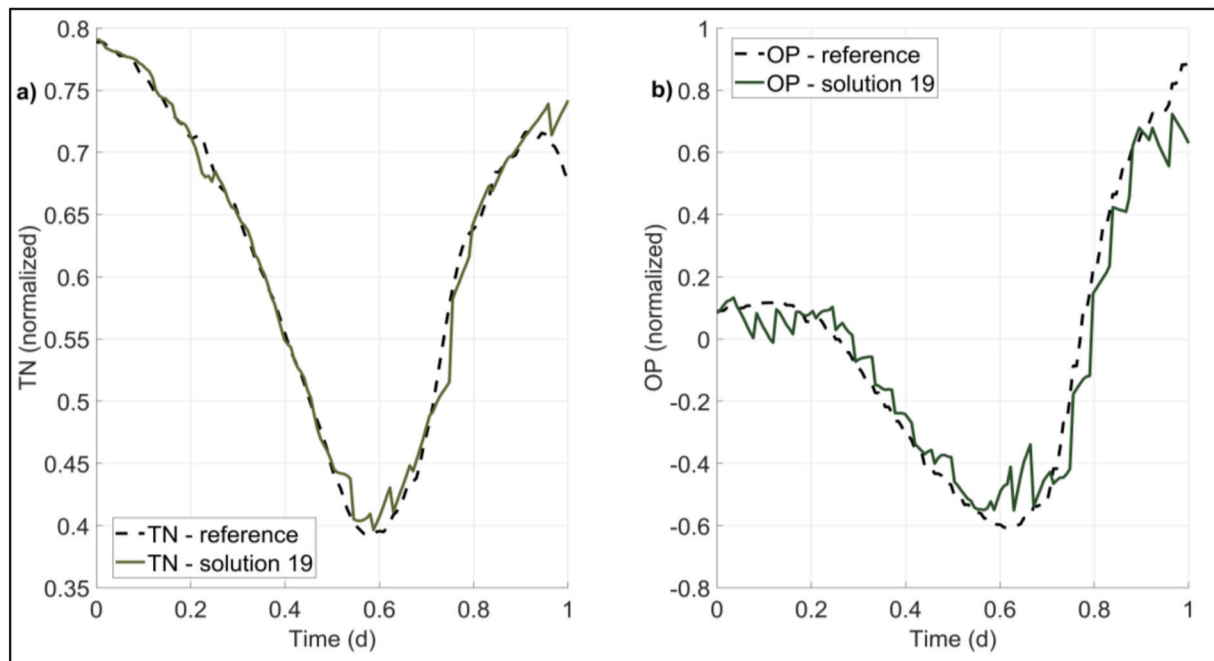


Fig. 11. Multi-objective optimization case 2, solution 19, output predictions: a) TN and b) OP.

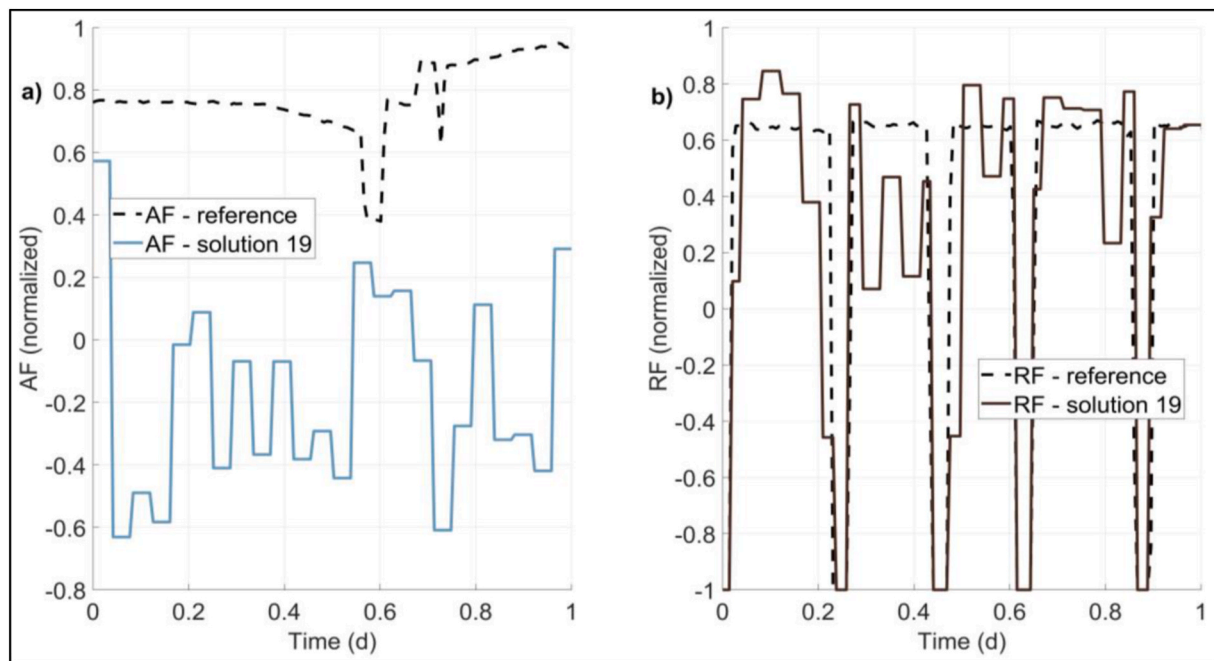


Fig. 12. Multi-objective optimization case 2, solution 19, WWTP operation variables: a) AF and b) RF.

in operator acceptance of the proposed AF values. This also reflects the findings from the analysis of the dataset that the dissolved oxygen level of the aerated bioreactors that is closely tied to AF is negatively correlated with the modelled output features (i.e., TN and OP) in the following few hours. However, constant high AF values are to be avoided as the dissolved oxygen levels slowly become positively correlated with the output features.

3.4. Comparative positioning within the literature and study limitations

The comparative analysis summarized in Table 4 highlights how the

proposed PCA-HO-DDM modelling framework advances beyond the dominant families of wastewater treatment models previously reported in the literature. The added value of this study relies on one hand on the use of comprehensive measured field data at high frequency (compared to the daily, hourly and bi-hourly data frequency used by prior studies), capturing the real-time operation of a full scale WWTP, and on the other hand on its methodological developments. The coupling of PCA with a HO loop, which classic DDM workflows typically lack, facilitates the dimension reduction and the systematic hyperparameter search, which further improve the model generalization ability and stability under operational variability. The developed PCA-HO-DDM models offer

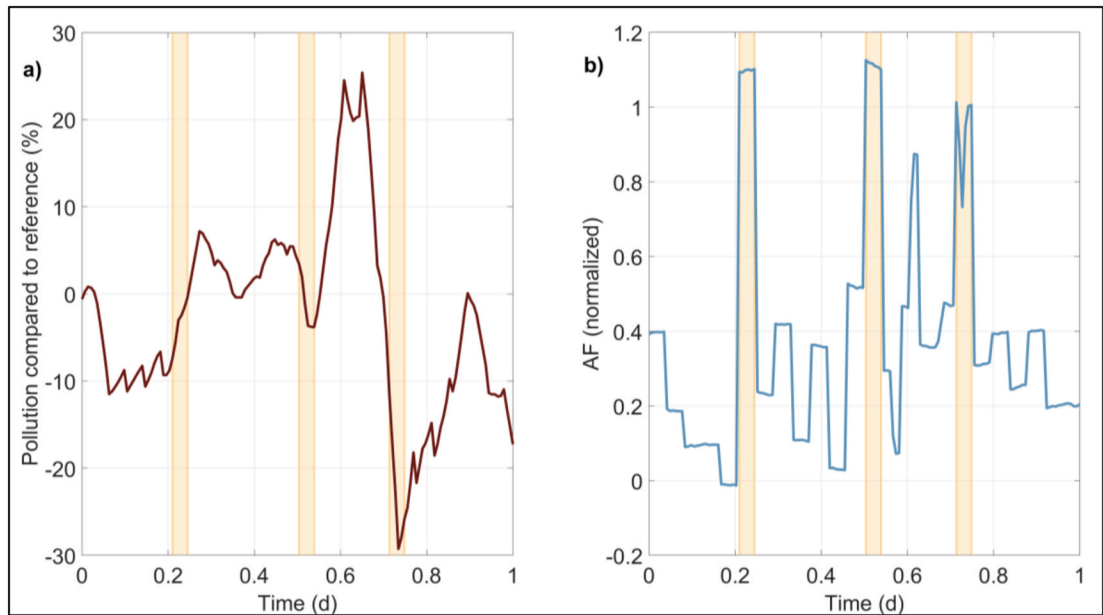


Fig. 13. Multi-objective optimization case 1, solution 7 results: a) Objective 1, pollution compared to reference case; and b) AF. Highlighted areas show the same timeframes on both figure a) and b).

Table 4
Study achievements in the context of existing literature.

Feature	PCA-HO-DDM	PCA-NARX	P-ED-LSTM	PCA-ELM	PCA-RF/LightGBM/ XGBoost	3D CNN-LSTM-GPR
Reference	Present study	Yang et al. (2022)	Yin et al. (2025)	Grace N et al., 2025	Chen et al. (2024)	Shaban et al. (2024)
Data frequency	Minutes	Daily	Bi-hourly	Daily	Daily	Hourly
Dynamics	Yes	Yes	Yes	No	No	No
PCA used	Yes	Yes	No	Yes	Yes	No
Hyperparameter optimization (HO)	Full HO loop	Limited	Limited	None	Random search	Limited to grid search
Optimization	Operational optimization	None	None	None	NSGA-II	None
Interpretation	High (PCA and sensitivity)	Medium	Medium (SHAP)	Low	High (SHAP)	Medium (Feature Importance Ranking)
Shock-load robustness	Medium-High	Low	High	Low	Low	Low
Predicts TN	Yes	Yes	Yes	Yes	No	No
Predicts OP	Yes	No	No	Yes	No	No
Model accuracy	TN and OP: R ² up to 0.99 MAPE <1%	TN and COD: R ² up to 0.85 MAPE 6.4-13.9%	TN: R ² of 0.60–0.80	TN and TP: R ² up to 0.90	COD: R ² of 0.87, Energy R ² of 0.876	TP: R ² of 0.92

empirical validation allowing for superior interpretability, and enhanced predictive performance under real-time fluctuating influent conditions.

Unlike existing PCA-NARX, PCA-ELM, ensemble-learning, or deep-learning models, which rely on low-frequency datasets, limited hyperparameter optimization, or single-architecture assumptions, the present approach integrates systematic model selection, dimensionality reduction, and dynamic predictive structures to achieve state-of-the-art accuracy (MAPE <1%, R² up to 0.99). At the same time, its limitations ought to be mentioned. As the models worked with a prediction horizon of 6 steps (i.e., 1 h) the DNARX models are set to the reference values after each hour. Thus, the identified optimal solutions only exhibit their influence over the 1-h scenario they are a part of. Consequently, the presented optimization studies should be considered as the combination of 24 distinct hourly optimization scenarios. Another limitation is the assumption that, in real-time applications, 1 h of “future” data is available. This may consist of predicted values which are later updated when real measurements become available. An alternative approach is to

develop models that rely on past data, using information from at least 1 h before the prediction time. In this manner, the required 1 h input window is available entirely from real data. Another important aspect is that the energy usage was calculated utilizing constant specific consumption factors. As such, the energy savings obtained are not absolute, but relative energy savings (as mentioned throughout the study) which do not take into consideration nonlinear blower efficiency, head losses, and pump performance in function of its load. The utilized dataset describes only two months worth of operation from a particular WWTP, thus the generalization capability of the developed models cannot extend beyond the events captured by the utilized dataset. In consequence, the presented models are yet to be broadly applicable due to seasonal variability, significant temperature shifts or operational regime changes not being described by the dataset. However, the proposed methodology can be applied in the future to a more comprehensive dataset, that also captures the mentioned events.

4. Conclusion

The proposed integrated PCA-HO-DDM methodology applied to an extensive nonlinear dataset from a real full scale WWTP resulted in the enhancement of classical approaches available in the literature. The PCA-BO-GPR outperformed the other classical methods, reaching $R^2 = 0.95$ and $MAPE < 5\%$. The proposed PCA-GA-NARX was found to be the best approach for all applications, with performance values of $R^2 = 0.99$ and $MAPE < 1\%$. This was further utilized for the identification of operation strategies during a 24-h scenario to reduce both energy consumption and effluent loading. The optimal solutions showed that relative energy savings can be achieved through a better management of the aeration flowrate. The simple application of a correction factor to the standard operational values can lead to relative energy savings of more than 20%, while the effluent loading is also slightly reduced. The developed tools can provide engineers, operators, and decision-makers with powerful, data-driven insights for a more informed, efficient, and sustainable management of the WWTP.

As future perspectives, the PCA-GA-NARX method is to be enhanced for the development of more performant models for TSS and other relevant effluent predictions. At the same time, the development of models with larger tapped delay line suitable for real-time optimization tasks will be studied.

CRedit authorship contribution statement

Norbert Botond Mihaly: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Elisabeta Cristina Timis:** Funding acquisition, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Nicolae Cebotaru:** Data curation. **Marius Adrian Brehar:** Resources, Validation, Visualization. **Vasile Mircea Cristea:** Software, Visualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Elisabeta Cristina Timis reports financial support was provided by Executive Unit for Financing Higher Education Research Development and Innovation. Norbert Botond Mihaly reports financial support was provided by Executive Unit for Financing Higher Education Research Development and Innovation. Marius Adrian Brehar reports financial support was provided by Executive Unit for Financing Higher Education Research Development and Innovation. If there are other authors, they 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 work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, Romania, project number PN-IV-P2-2.1-TE-2023-0666, within PNCDI IV.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.clet.2026.101207>.

Nomenclature

Symbols

n	tapped delay line horizon for NARX model
N	number of data points

s	model output feature values
σ	error standard deviation
t	target output feature values
t_i	timestep i
x	input features
y	output features

Subscripts/superscripts

i	counter for data points
j	counter for data points
T	transposed matrix

Acronyms

AI	artificial intelligence
AF	aeration flowrate
ANN	artificial neural network
BO	Bayesian optimization
BOD	biological oxygen demand
COD	chemical oxygen demand
CPU	central processing unit
DDM	data-driven model
DNARX	deep nonlinear autoregressive network with exogenous inputs
DO	dissolved oxygen
DT	decision tree
ELM	extreme learning machine
GA	genetic algorithm
GP	Gaussian process
GPR	Gaussian process regression
GPU	graphics processing unit
HO	hyperparameter optimization
MAE	mean absolute error
MAPE	mean absolute percentage error
MLP	multi-layer perceptron
MSE	mean squared errors
NARX	nonlinear autoregressive network with exogenous inputs
NSGA II	non-dominated sorted genetic algorithm II
OP	orthophosphates
PCA	principal component analysis
RF	external recycle flowrate
RFR	random forest regression
RMSE	root mean squared error
RT	random tree
RTR	random tree regression
SVM	support vector machine
SVR	support vector regression
TN	total nitrogen
TP	total phosphorus
TSS	total suspended solids
WWTP	wastewater treatment plant

Data availability

The data that has been used is confidential.

References

- Achite, M., Samadianfard, S., Elshaboury, N., Toubal, K.A., Abdelkader, E.M., Sharafi, M., 2024. A combined support vector regression with a firefly algorithm for prediction of energy consumption in wastewater treatment plants. *Water Sci. Technol.* 90, 2747–2763. <https://doi.org/10.2166/wst.2024.375>.
- Alali, Y., Harrou, F., Sun, Y., 2023. Unlocking the potential of wastewater treatment: machine learning based energy consumption prediction. *Water* 15 (13), 2349. <https://doi.org/10.3390/w15132349>.
- Alex, J., Benedetti, L., Copp, J., Gernaey, K.V., Jeppsson, U., Nopens, I., Pons, M., Rieger, L., Rosen, C., Steyer, J.P., Vanrolleghem, P., Winkler, S., 2008. Benchmark simulation model no. 1 (BSM1). In: *Control*, 1. Issue 1.
- Atanasova, N.B., Kompare, B., 2002. Modelling of wastewater treatment plant with regression trees. In: Zanasi, A., Brebbia, C.A., Ebecken, N.F.F., Melli, P. (Eds.), *Data Mining III*. WIT Press, Southampton, UK.

- Bagherzadeh, F., Nouri, A.S., Mehrani, M.-J., Thennadil, S., 2021. Prediction of energy consumption and evaluation of affecting factors in a full-scale WWTP using a machine learning approach. *Process Saf. Environ. Prot.* 154, 458–466. <https://doi.org/10.1016/j.psep.2021.08.040>.
- Batool, R., Sahrish, Ahad, S.A., Maqsood, Q., Ali, S.W., Abbas, S.M., 2024. Redefining sustainability: next-gen wastewater treatment breakthroughs. *Clean. Water* 1, 100018. <https://doi.org/10.1016/j.clwat.2024.100018>.
- Chen, J., Li, J., He, Y., Chen, S., Wu, L., He, Y., Lu, Y., Ling, H., Su, C., 2025. Research status and prospects of using machine learning for promoting carbon emission reduction in wastewater treatment processes: a review. *J. Water Proc. Eng.* 76, 108224. <https://doi.org/10.1016/j.jwpe.2025.108224>.
- Chen, J., Wan, J., Ye, G., Wang, Y., 2024. Prediction and optimization of wastewater treatment process effluent chemical oxygen demand and energy consumption based on typical ensemble learning models. *Bioresour. Technol.* 411, 131362. <https://doi.org/10.1016/j.biortech.2024.131362>.
- Chen, X., Lei, Z., Chang, J.S., Lee, D.J., 2025. Navigating future wastewater treatment plants with artificial intelligence: applications, challenges, and innovations. *J. Clean. Prod.* 504, 145467. <https://doi.org/10.1016/j.jclepro.2025.145467>.
- Chen, Y., Song, L., Liu, Y., Yang, L., Li, D., 2020. A review of the artificial neural network models for water quality prediction. *Appl. Sci.* 10 (17), 5776. <https://doi.org/10.3390/AP10175776>.
- Cheng, H.-H., Huang, P.-W., Whang, L.-M., 2024. Optimization of biological nitrogen removal in full-scale municipal WWTPs using activated sludge model simulation. *Chemosphere* 362, 142939. <https://doi.org/10.1016/j.chemosphere.2024.142939>.
- Cheng, Q., Houtao, L., Jin, Z., Fangqian, H., Jing, J., Li, Q., Wang, M., 2025. Enhancing sustainability in a sequential batch reactor for wastewater treatment: soft sensor-based ammonia prediction with PSO-LSTM and random forest model. *Environ. Sci. Water Res. Technol.* 11 (7), 1691–1703. <https://doi.org/10.1039/D5EW00063G>.
- Dai, W., Pang, J.W., Ding, J., Wang, J. hui, Xu, C., Zhang, L.Y., Ren, N.Q., Yang, S.S., 2025. Integrated real-time intelligent control for wastewater treatment plants: Data-driven modeling for enhanced prediction and regulatory strategies. *Water Res.* 274, 123099. <https://doi.org/10.1016/j.watres.2025.123099>.
- De la Hoz-M, J., Ariza-Echeverri, E.A., Vergara, D., 2024. Exploring the role of artificial intelligence in wastewater treatment: a dynamic analysis of emerging research trends. *Resour. 13* (12), 171. <https://doi.org/10.3390/resources13120171>.
- Deb, K., 2001. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons.
- Dikmen, F., Demir, A., Özkaya, B., Raza, M.O., Rasheed, J., Asuroglu, T., Alsubai, S., 2025. AI-driven wastewater management through comparative analysis of feature selection techniques and predictive models. *Sci. Rep.* 15 (1), 25347. <https://doi.org/10.1038/S41598-025-07124-0>.
- Du, Y., Lei, T., Jin, J., Chen, B., Wu, X., Yang, N., Ge, Y., Wang, Y., Tan, Q., Fan, D., Lei, Y., Yao, L., 2025. Harnessing machine learning for energy optimization and intelligent process control in wastewater treatment. *J. Water Proc. Eng.* 80, 109210. <https://doi.org/10.1016/j.jwpe.2025.109210>.
- Elgaray, A.M., Elwakeel, K.Z., Mohammad, S.H., Elshoubaky, G.A., 2021. A critical review of biosorption of dyes, heavy metals and metalloids from wastewater as an efficient and green process. *Clean. Eng. Technol.* 4, 100209. <https://doi.org/10.1016/j.clet.2021.100209>.
- Elharbil, R., El Moussaoui, T., El Ass, K., Belloulid, M.O., El Alaoui El Fels, A., Samiri, M. Y., 2024. Hybrid data driven approach based on ANNs-PCA for wastewater treatment plant performance assessment. *Clean. Water* 2, 100058. <https://doi.org/10.1016/j.clwat.2024.100058>.
- European Parliament and Council, 2024. *Concerning urban wastewater treatment (recast)*. In: OJ L, 2024/3019, 12.12.2024, p. 3019.
- Grace N, V.A., Aldehim, G., Alruwais, N., T, N.P., 2025. Novel PCA-driven extreme machine learning for comprehensive modelling of metropolitan wastewater treatment systems. *Desalination Water Treat.* 321, 101037. <https://doi.org/10.1016/j.dwt.2025.101037>.
- Hoz, J.L., Bappy, M.M., Islam, M.S., Marcantel, M., Hayes, M.P., 2026. Interpretable forecasting of dissolved oxygen leveraging foundation model for proactive aeration in rural wastewater treatment systems. *Water Res.* 289, 124931. <https://doi.org/10.1016/j.watres.2025.124931>.
- Huang, R., Yu, C., Wang, H., Zhang, S., Wang, L., Li, H., Zhang, Z., Zhou, Z., 2023. Spatial and temporal modeling on energy consumption of wastewater treatment based on machine learning algorithms. *ACS ES&T Water* 4 (3), 1119–1130. <https://doi.org/10.1021/ACSESTWATER.3C00430>.
- Hvala, N., Kocijan, J., 2020. Design of a hybrid mechanistic/Gaussian process model to predict full-scale wastewater treatment plant effluent. *Comput. Chem. Eng.* 140, 106934. <https://doi.org/10.1016/j.compchemeng.2020.106934>.
- Jassim, M.S., Hammad, S., Coskuner, G., 2025. AI-Driven optimization models for prediction of effluent quality parameters in a wastewater treatment plant. *Process Integr. Optim. Sustain.* 2025, 1–17. <https://doi.org/10.1007/S41660-025-00576-X>.
- Kini, K.R., Harrou, F., Madakary, M., Sun, Y., 2025. Enhancing fault detection in wastewater treatment plants: a multi-scale principal component analysis approach with the Kantorovich distance. *Environ. Sci. Water Res. Technol.* 11 (5), 1215–1232. <https://doi.org/10.1039/d4ew00831f>.
- Kumar, R., Goyal, M.K., 2025. Role of AI&ML in modernizing water and wastewater treatment processes. *Water, Air, Soil Pollut.* 236 (1), 31. <https://doi.org/10.1007/s11270-024-07618-z>.
- Ly, Q.V., Truong, V.H., Ji, B., Nguyen, X.C., Cho, K.H., Ngo, H.H., Zhang, Z., 2022. Exploring potential machine learning application based on big data for prediction of wastewater quality from different full-scale wastewater treatment plants. *Sci. Total Environ.* 832, 154930. <https://doi.org/10.1016/j.scitotenv.2022.154930>.
- Ma, Z., Zhu, Y., Chen, C., Li, T., Li, Y., Li, X., Wang, Y., Waite, T.D., Guan, J., 2025. Towards the digitalization of water treatment facilities: a case study on machine learning-enabled digital twins. *J. Water Proc. Eng.* 77, 108316. <https://doi.org/10.1016/j.jwpe.2025.108316>.
- Majnoon, A., Saifoddin, A., 2025. AI-driven energy optimization enhancing efficiency in urban environments with hybrid machine learning models. *Clean. Eng. Technol.* 28, 101072. <https://doi.org/10.1016/J.CLET.2025.101072>.
- Mihály, N.-B., Luca, A.-V., Simon-Várhelyi, M., Cristea, V.M., 2023. Improvement of air flowrate distribution in the nitrification reactor of the waste water treatment plant by effluent quality, energy and greenhouse gas emissions optimization via artificial neural networks models. *J. Water Proc. Eng.* 54, 103935. <https://doi.org/10.1016/j.jwpe.2023.103935>.
- Mihály, N.B., Simon-Várhelyi, M., Cristea, V.M., 2022. Data-driven modelling based on artificial neural networks for predicting energy and effluent quality indices and wastewater treatment plant optimization. *Optim. Eng.* 23, 2235–2259. <https://doi.org/10.1007/s11081-022-09724-5>.
- Nadeem, N., Bentaleb, A., Qadir, M.B., Ahmad, F., Shahzad, M.W., Jumah, A. B., 2025. A machine learning approach to wastewater treatment: gaussian process regression and Monte Carlo analysis. *Nanoscale Adv.* 7, 4436–4449. <https://doi.org/10.1039/d4na01064g>.
- Pan, C., Huang, Y., Lu, Y., Bu, Y., Ma, B., 2025. Enhancing predictive accuracy of wastewater treatment process: an approach via optimizing data collection and increasing operating state diversity. *J. Clean. Prod.* 487, 144621. <https://doi.org/10.1016/j.jclepro.2024.144621>.
- Peerzade, S., Kamat, P., 2025. Enhancing water quality prediction: a machine learning approach across diverse water environments. *Water Qual. Res. J.* 60 (1), 298–317. <https://doi.org/10.2166/wqrj.2025.083>.
- Picos-Benítez, A.R., Martínez-Vargas, B.L., Duron-Torres, S.M., Brillas, E., Peralta-Hernández, J.M., 2020. The use of artificial intelligence models in the prediction of optimum operational conditions for the treatment of dye wastewaters with similar structural characteristics. *Process Saf. Environ. Prot.* 143, 36–44. <https://doi.org/10.1016/j.psep.2020.06.020>.
- Ramya, S., Srinath, S., Tuppad, P., 2024. Hybrid modeling techniques for predicting chemical oxygen demand in wastewater treatment: a stacking ensemble learning approach with neural networks. *Environ. Monit. Assess.* 196 (12), 1259. <https://doi.org/10.1007/S10661-024-13390-8>.
- Sahu, P., Londhe, S.N., Kulkarni, P.S., 2023. Prediction of water quality parameters using support vector regression. *Innov. Infrastruct. Solut.* 8. <https://doi.org/10.1007/s41062-023-01195-6>.
- Shaban, W.M., Xie, D., Elbaz, K., Shen, S.-L., 2024. Real-time water quality prediction of wastewater treatment plants using advanced deep learning networks. *J. Water Proc. Eng.* 65, 105775. <https://doi.org/10.1016/j.jwpe.2024.105775>.
- Siatou, A., Manali, A., Gikas, P., 2020. Energy consumption and internal distribution in activated sludge wastewater treatment plants of Greece. *Water* 12 (4), 1204. <https://doi.org/10.3390/W12041204>.
- Sun, J., Guan, X., Sun, X., Cao, X., Tan, Y., Liao, J., 2024. Water quality prediction and carbon reduction mechanisms in wastewater treatment in Northwest cities using Random Forest Regression model. *Sci. Rep.* 14, 31525. <https://doi.org/10.1038/s41598-024-83277-8>.
- Wei, Z., Wu, N., Zou, Q., Zou, H., Zhu, L., Wei, J., Huang, H., 2023. Data modeling of sewage treatment plant based on long short-term memory with Multilayer Perceptron network. *Water* 15 (8), 1472. <https://doi.org/10.3390/W15081472>.
- Xiong, Z., Liu, X., Igou, T., Li, Z., Chen, Y., 2025. Using hybrid machine learning to predict wastewater effluent quality and ensure treatment plant stability. *Water* 17 (13), 1851. <https://doi.org/10.3390/W17131851>.
- Xu, W.L., Wang, Y.J., Wang, Y.T., Li, J.G., Zeng, Y.N., Guo, H.W., Liu, H., Dong, K.L., Zhang, L.Y., 2024. Application and innovation of artificial intelligence models in wastewater treatment. *J. Contam. Hydrol.* 267, 104426. <https://doi.org/10.1016/J.JCONHYD.2024.104426>.
- Yang, Y., Kim, K.R., Kou, R., Li, Y., Fu, J., Zhao, L., Liu, H., 2022. Prediction of effluent quality in a wastewater treatment plant by dynamic neural network modeling. *Process Saf. Environ. Prot.* 158, 515–524. <https://doi.org/10.1016/J.PSEP.2021.12.034>.
- Yin, H., Chen, Y., Zhou, J., Xie, Y., Wei, Q., Xu, Z., 2025. A probabilistic deep learning approach to enhance the prediction of wastewater treatment plant effluent quality under shocking load events. *Water Res.* X 26, 100291. <https://doi.org/10.1016/j.wroa.2024.100291>.
- Yu, Y., Chen, Y., Huang, S., Wang, R., Wu, Y., Zhou, H., Li, X., Tan, Z., 2024. Enhancing the effluent prediction accuracy with insufficient data based on transfer learning and LSTM algorithm in WWTPs. *J. Water Proc. Eng.* 62, 105267. <https://doi.org/10.1016/j.jwpe.2024.105267>.
- Zakariazadeh, A., Ahshan, R., Al Abri, R., Al-Abri, M., 2024. Renewable energy integration in sustainable water systems: a review. *Clean. Eng. Technol.* 18, 100722. <https://doi.org/10.1016/J.CLET.2024.100722>.
- Zamfir, F.S., Carbureanu, M., Mihalache, S.F., 2025. Application of machine learning models in optimizing wastewater treatment processes: a review. *Appl. Sci.* 15 (15), 8360. <https://doi.org/10.3390/app15158360>.
- Zhang, D., Liu, J., Wang, H., Wang, Y., 2026. Advancing carbon-neutral wastewater treatment: artificial intelligence-driven strategies for emission mitigation and process optimization. *Environ. Res.* 290, 123449. <https://doi.org/10.1016/J.ENVRES.2025.123449>.

- Zhang, S., Jin, Y., Chen, W., Wang, J., Wang, Y., Ren, H., 2023. Artificial intelligence in wastewater treatment: a data-driven analysis of status and trends. *Chemosphere* 336, 139163. <https://doi.org/10.1016/J.CHEMOSPHERE.2023.139163>.
- Zhang, Z., Qi, F., Liu, Y., Asif, M.B., Ikhlaq, A., Wang, Z., Chen, C., Li, C., Chang, J., Li, Q., Li, Y., Li, Y., Jia, Y., Liu, Y., Xu, B., Sun, D., 2025. Comprehensive assessment, intelligent prediction, and precise mitigation strategies for greenhouse gas emissions in full-scale wastewater treatment plants. *Environ. Res.* 270, 121052. <https://doi.org/10.1016/j.envres.2025.121052>.
- Zhou, P., Li, Z., Snowling, S., Baetz, B.W., Na, D., Boyd, G., 2019. A random forest model for inflow prediction at wastewater treatment plants. *Stoch. Environ. Res. Risk Assess.* 33, 1781–1792. <https://doi.org/10.1007/s00477-019-01732-9>.

Wastewater Treatment Plant Operation Optimization Towards Energy Efficiency and Reduced Effluent Loading Based on Optimized Data-Driven Models

Norbert Botond Mihaly^a, Elisabeta Cristina Timis^{a,*}, Nicolae Cebotaru^a, Marius Adrian Brehar^a,
Vasile Mircea Cristea^a

^aBabes-Bolyai University, Faculty of Chemistry and Chemical Engineering, Department of Chemical Engineering, 11 Arany Janos St., 400028, Cluj-Napoca, Cluj, Romania

The aim of the present document is to offer supplementary information related to the mathematical description of the data-driven models, results on the dependence of MAPE on the values utilized in its calculation, time lagged correlation coefficients between input and output features, and learning curves from the development of the neural network models.

Random tree and random forest regression

With core nodes representing decisions based on input feature values, branches describing the decisions' results, and leaf nodes holding the predicted continuous values, a random tree regression (RTR) model results in a tree-like structure. By choosing features and matching split points that minimize a predetermined loss function (i.e., the mean squared errors (MSE) in the case of regression as in the current study) the model recursively divides the input space (Breiman, 2001). The random elements of these models are due to the random subset of features at each node that can be considered for splitting, rather than all available input features, as well as the bootstrap sampling of the training data for the construction of the different trees in a random forest regression (RFR) model. For regression studies, the features and split points are selected in such manner that the MSE of the resulting regions are reduced, as described by equation 1.

$$MSE = \sum_{i \in R_1} (y_i - \bar{y}_1)^2 + \sum_{i \in R_2} (y_i - \bar{y}_2)^2 \quad (\text{equation 1})$$

where, R_1 and R_2 are the regions resulting from a split, while $\bar{y}_1$ and $\bar{y}_2$ are the mean values of the output feature in each of these regions.

* Corresponding author: elisabeta.timis@ubbcluj.ro; Babes-Bolyai University, Department of Chemical Engineering, 11 Arany Janos St., 400028, Cluj-Napoca, Cluj, Romania
Email addresses: norbert.mihaly@ubbcluj.ro (N.B. Mihaly), elisabeta.timis@ubbcluj.ro (E.C. Timis), nicolae.cebotaru@stud.ubbcluj.ro (N. Cebotaru), marius.brehar@ubbcluj.ro (M.A. Brehar), mircea.cristea@ubbcluj.ro (V.M. Cristea)

Support vector regression

Support vector regression (SVR) is the term used to describe the employment of a SVM model to a regression task. SVR is a supervised learning method that predicts continuous output values instead of classes. It is based on the same concepts as SVM for classification, aiming to find a function $f(x)$ to describe the output feature based on model inputs with at most ε deviation. At the same time, the resulting function should be as flat as possible which in turn influences the complexity of the model. A linear SVR model can be defined as shown in equation 2, with the minimization goal of equation 3, subjected to the constraints in equation 4.

$$f(x) = x^T \alpha + b \quad (\text{equation 2})$$

$$J(\alpha) = \frac{1}{2} \alpha^T \alpha \quad (\text{equation 3})$$

$$|y_i - (x_i^T \alpha + b)| \leq \varepsilon \quad (\text{equation 4})$$

where x_n is a multivariate set of input features with N observations that correspond to y_n output feature values; and $\alpha^T \alpha$ is the norm value of $f(x)$.

For a complex modelling problem, however, there is a chance that no function that satisfies the constraints in all cases can be found. As such, the previously presented model is extended with slack variables that allow regression deviations to occur and still satisfy the conditions resulting in the primal formula presented in equation 5 (Vapnik, 2000).

$$J(\alpha) = \frac{1}{2} \alpha^T \alpha + C \sum_{i=1}^N (\beta_i + \beta_i^*) \quad (\text{equation 5})$$

where β_i and β_i^* are the slack variables, while C is the box constraint.

The resulting optimization task may be computed with less effort in its Lagrange dual formulation that introduces nonnegative multipliers for each observation x_i as described in equation 6. The SVR then can be modified to a nonlinear model by simply utilizing a different nonlinear (e.g., gaussian or polynomial) kernel function instead of the dot product as presented in equation 7.

$$L(\gamma) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\gamma_i - \gamma_i^*) (\gamma_j - \gamma_j^*) x_i^T x_j + \varepsilon \sum_{i=1}^N (\gamma_i + \gamma_i^*) + \sum_{i=1}^N \gamma_i (\gamma_i^* - \gamma_i) \quad (\text{equation 6})$$

$$L(\gamma) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\gamma_i - \gamma_i^*) (\gamma_j - \gamma_j^*) G(x_i, x_j) + \varepsilon \sum_{i=1}^N (\gamma_i + \gamma_i^*) + \sum_{i=1}^N \gamma_i (\gamma_i^* - \gamma_i) \quad (\text{equation 7})$$

where γ_i and γ_i^* are the nonnegative multipliers, $x_i^T x_j$ is the mentioned dot product, while $G(x_i, x_j)$ is the kernel function.

Gaussian process regression

Any finite number of random variables with a joint multivariate Gaussian distribution is called a GP. In case of regression, a GP establishes a prior distribution over feasible functions that fit the data which can be utilized together with the input data to construct a posterior distribution utilized for predictions, known as gaussian process regression (GPR). A GPR model is described in equation 8 (Rasmussen, 2004).

$$h(x)^T \alpha + f(x) \quad (\text{equation 8})$$

where $f(x) \sim GP(m(x), k(x, x^T))$, i.e., $f(x)$ are from a $m(x)$ mean function GP with $k(x, x^T)$ as the covariance function; coefficients α are estimated based on the data; and $h(x)$ represents a set of basis functions that transform the input feature vector in x into a new vector of features $h(x)$ in R^p (i.e., a p -dimensional feature space). The response (y) of such a model considering a zero mean GP can be described as in equation 9.

$$P(y_i|f(x_i), x_i) \sim N(y_i|h(x_i)^T \alpha + f(x_i), \sigma^2) \quad (\text{equation 9})$$

where σ^2 is the variance of errors.

It can be observed that GPR is a probabilistic model, and the latent variable $f(x_i)$ for each observation (x_i) makes it a nonparametric model.

Nonlinear autoregressive network with exogenous inputs

A broad description of how to compute the output for a NARX model is provided by equation 10.

$$y(ti) = f(y(ti-1), y(ti-2), \dots, y(ti-n_y), x(ti-1), x(ti-2), \dots, x(ti-n_x)) \quad (\text{equation 10})$$

where y is the output signal, x represents the input signal, while n_y and n_x are tapped delay line horizons for the two signals.

Fig. S1 describes the minimum and maximum values of correlation coefficients obtained in the study of input and output features considering a time delay limit of 360 moments in time which translates to roughly 2.5 days.

Table S1 illustrates how the developed model's accuracy considering the MAPE criteria is dependent on the values of the OP output feature. It is apparent how OP values close to zero greatly increase the percentual errors, as the model is twice as performant for values above 0.1 mg/L, while for values above 0.25mg/L it is more than three times as accurate.

Table S1

PCA-GA-DNARX results for MAPE criteria in function of OP values

OP values utilized in MAPE calculation	MAPE results
all OP	23.30%
OP > 0.10	11.65%
OP > 0.15	11.41%
OP > 0.20	8.97%
OP > 0.25	7.54%

The two learning curves of the PCA-GA-NARX models are presented in Fig. S2 and Fig. S3 which were saved at the time of development of the mentioned models. The learning curves for the PCA-GA-DNARX models were not saved during their development, however, these training processes were performed with the identified optimal hyperparameters. The learning curves resulting from these trainings are presented in Fig. S4 and Fig. S5.

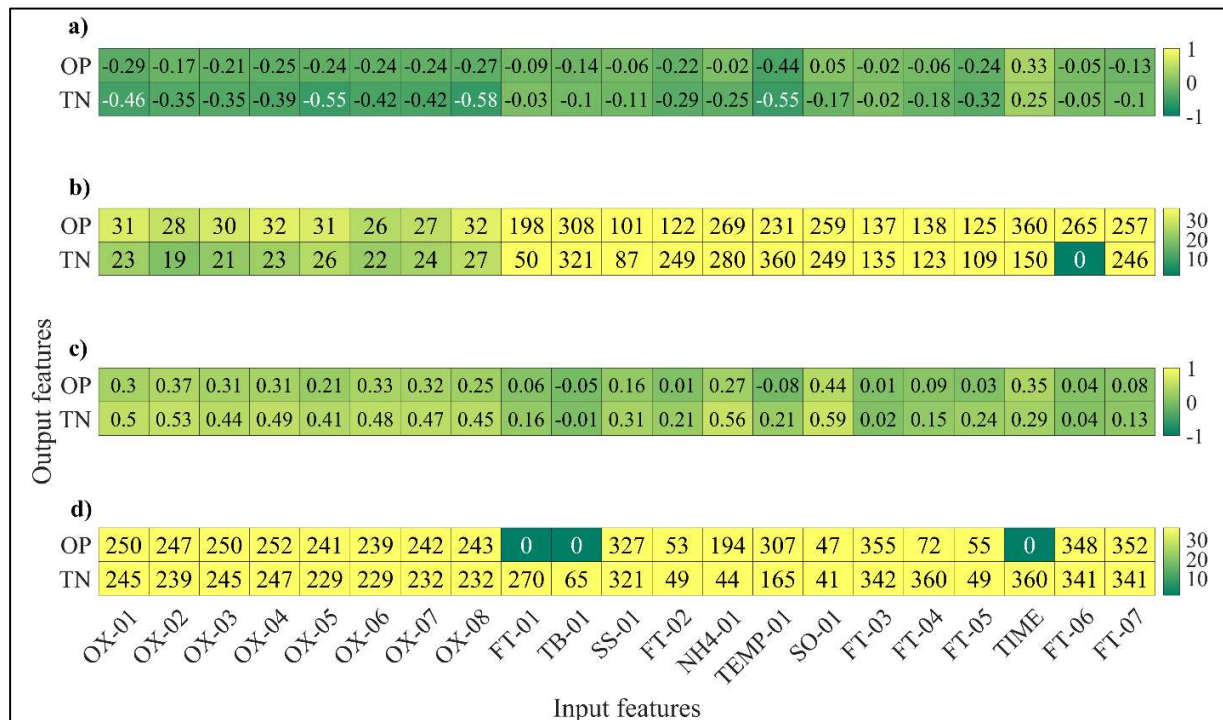


Fig. S1. a) Minimum correlation coefficient values over the investigated time delay; b) Time delay values at which the minimum correlation coefficients were obtained; c) Maximum correlation coefficient values over the investigated time delay; d) Time delay values at which the maximum correlation coefficients were obtained

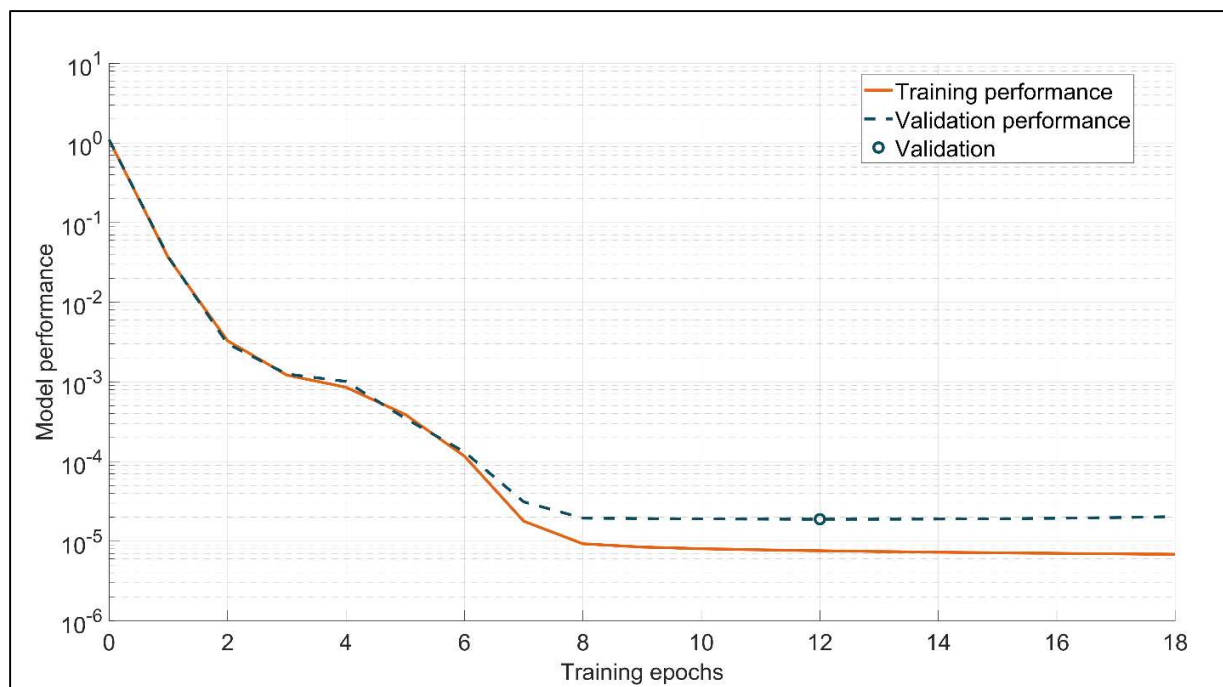


Fig. S2. Learning curve for PCA-GA-NARX model concerning the TN output feature

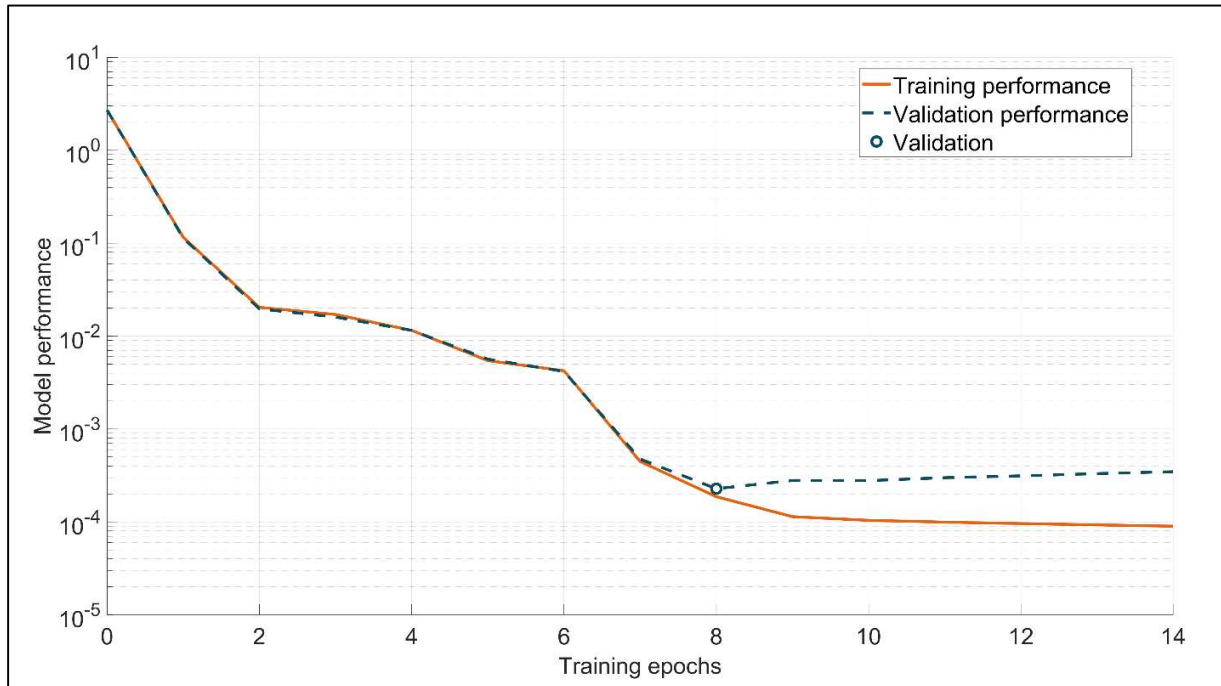


Fig. S3. Learning curve for PCA-GA-NARX model in case of the OP output feature

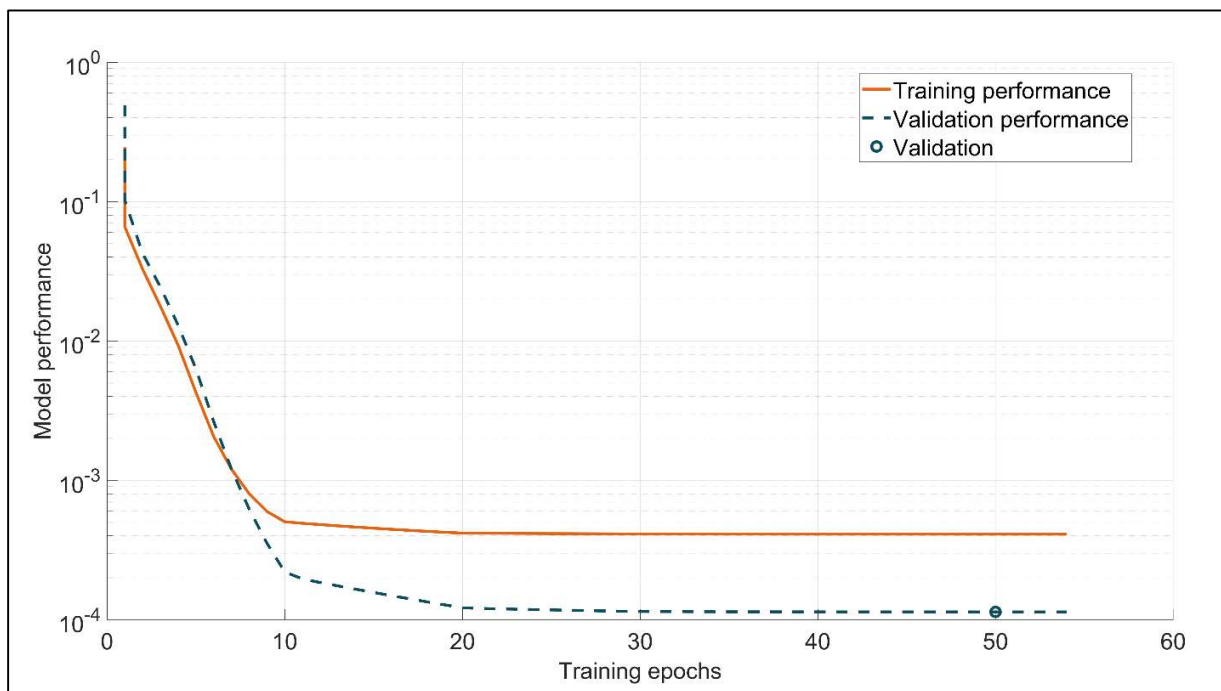


Fig. S4. Learning curve for PCA-GA-DNARX model concerning the TN output feature

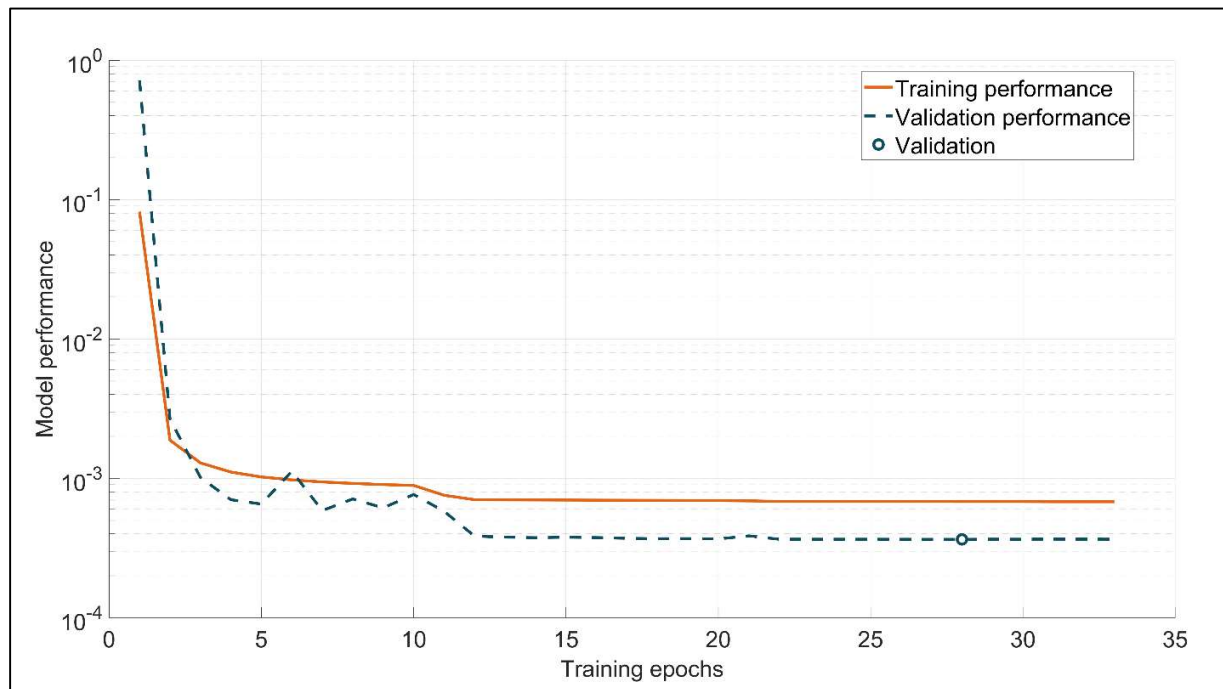


Fig. S5. Learning curve for PCA-GA-NARX model in case of the OP output feature

References

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Rasmussen, C.E. (2004). Gaussian Processes in Machine Learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds) *Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science*, vol 3176. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_4
- Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer New York. <https://doi.org/10.1007/978-1-4757-3264-1>