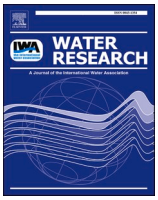
Water Research 268 (2025) 122754

Contents lists available at ScienceDirect

Water Research

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



# Forecasting nitrous oxide emissions from a full-scale wastewater treatment plant using LSTM-based deep learning models

Siddharth Seshan [a,b,*], Johann Poinapen [a], Marcel H. Zandvoort [c], Jules B. van Lier [b], Zoran Kapelan [b]

[a] KWR Water Research Institute, Nieuwegein, the Netherlands
[b] Section Sanitary Engineering, Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands
[c] Waternet, Amsterdam, the Netherlands

ARTICLE INFO

*Keywords:*
Nitrous oxide
Wastewater treatment
Artificial intelligence
Data-driven modelling
LSTM

ABSTRACT

Nitrous oxide ($N_2O$) emissions from wastewater treatment plants (WWTPs) exhibit significant seasonal variability, making accurate predictions with conventional biokinetic models difficult due to complex and poorly understood biochemical processes. This study addresses these challenges by exploring data-driven alternatives, using long short-term memory (LSTM) based encoder-decoder models as basis. The models were developed for future integration into a model predictive control framework, aiming to reduce $N_2O$ emissions by forecasting these over varying prediction horizons. The models were trained on 12 months and tested on 3 months of data from a full-scale WWTP in Amsterdam West, the Netherlands. The dataset encompasses seasonal peaks in $N_2O$ emissions typical for winter and spring months. The best performing model, featuring a 256–256 LSTM architecture, achieved the highest accuracy with test $R^2$ values up to 0.98 across prediction horizons spanning 0.5 to 6.0 h ahead. Feature importance analysis identified past $N_2O$ emissions, influent flowrate, $NH_4^+$, $NO_x$, and dissolved oxygen (DO) in the aerobic tank as most significant inputs. The observed decreasing influence of historical $N_2O$ emissions over extended prediction horizons highlights the importance and significance of process variables for the model's performance. The model's ability to accurately forecast short-term $N_2O$ emissions and capture immediate trends highlights its potential for operational use in controlling emissions in WWTPs. Further research incorporating diverse datasets and biochemical process inputs related to microbial activities in the $N_2O$ production pathways could improve the model's accuracy for longer forecasting horizons. These findings advocate for hybridising deep learning models with biokinetic and mechanistic insights to enhance prediction accuracy and interpretability.

## 1. Introduction

The urgency of climate change and global warming related challenges have led wastewater treatment plants (WWTPs) authorities to critically consider their greenhouse gas (GHG) emissions and carbon footprint. Nitrous oxide ($N_2O$) is one of the most potent GHG emitted from WWTPs, with a global warming potential 273 times greater than that of carbon dioxide ($CO_2$) on a 100-year time scale (Forster et al., 2021). In addition, its increasing atmospheric concentrations contribute to the depletion of the ozone layer in the stratosphere (Ravishankara et al., 2009). Hence, it is necessary to enhance the understanding of the underlying processes behind $N_2O$ production in WWTPs and develop effective mitigation strategies to reduce its emissions.

The production of $N_2O$ in WWTPs is associated with biological nitrogen removal (BNR) processes. During autotrophic nitrification conducted by the ammonia-oxidising bacteria (AOB), the incomplete oxidation of hydroxylamine ($NH_2OH$) to nitrite ($NO_2^-$) can cause $N_2O$ production (Pan et al., 2024). Additionally, under low dissolved oxygen (DO) conditions, $NO_2^-$ and NO accumulation can lead to $N_2O$ production, in a phenomenon known as nitrifier denitrification (Seshan et al., 2024). $N_2O$ is also an intermediate in heterotrophic denitrification, specifically in the final step, where $N_2O$ is reduced to $N_2$. Incompletion or inhibition of this step can lead to $N_2O$ accumulation (Massara et al., 2017a). Detailed monitoring of $N_2O$ and related parameters is necessary to better understand the actual production pathways and to develop relevant mitigation strategies. Long duration measurement campaigns have

* Corresponding author at: KWR Water Research Institute, Groningenhaven 7, 3430 BB, Nieuwegein, the Netherlands.
*E-mail address:* siddharth.seshan@kwrwater.nl (S. Seshan).

https://doi.org/10.1016/j.watres.2024.122754
Received 22 July 2024; Received in revised form 15 October 2024; Accepted 4 November 2024
Available online 5 November 2024

reported distinctive seasonal and diurnal variations in $N_2O$ emissions at various full-scale WWTPs (Daelman et al., 2013; Gruber et al., 2020; Kosonen et al., 2016). A significant seasonal peak has been recorded during the spring months (Gruber et al., 2021a), while in our previous work, we discussed a full-scale study reporting 54 % of the yearly $N_2O$ mass emitted during March and April (Seshan et al., 2024).

Hence, there is an increasing urgency to develop effective mitigation and operational control strategies that can be implemented in full-scale WWTPs to reduce $N_2O$ emissions. Models that can accurately predict $N_2O$ emissions, perform mitigation scenario analysis, and support advanced control strategies, offer a promising solution. The availability of long duration datasets is crucial for advancing the modelling of $N_2O$ production and emissions from WWTPs. Over the years, biokinetic models such as the activated sludge models (ASMs) have been extended to include $N_2O$ production pathways, to predict the emissions and simulate control strategies (Mampaey et al., 2013; Massara et al., 2017b; Ni et al., 2015; Guo and Vanrolleghem, 2013). Notably, in studies involving full-scale systems, the calibration process often used short to mid-term data, missing seasonal variations. Biokinetic models that were applied to long duration data revealed the active production pathways and plausible process conditions behind the seasonal peak. Nonetheless, these models struggled to predict the seasonal variations accurately (Seshan et al., 2024). This lack of accuracy shows that the current knowledge on the $N_2O$ production dynamics included in biokinetic models is limited in describing the changing kinetics that cause the strong seasonal variations in $N_2O$ emissions. Furthermore, biokinetic models, although comprehensive, are often over-parameterised, making calibration more complex (Domingo-Félez and Smets, 2016). Proper calibration requires monitoring intermediates of the production pathways such as $NH_2OH$, which can be costly and challenging (Khalil et al., 2023). These limitations raise questions about the suitability of biokinetic models for $N_2O$ emissions mitigation strategies, especially model-based control strategies.

The use of data-driven or machine learning (ML) models for $N_2O$ emissions from WWTPs has shown promise as an alternative to biokinetic models, given the increased availability of long-duration and high-resolution data. Various studies have utilised unsupervised learning methods such as principal component analysis (PCA) and clustering techniques to gain insights into the wastewater treatment operations and patterns associated with $N_2O$ production (Bellandi et al., 2020; Vasilaki et al., 2018). However, these methods cannot quantitatively predict $N_2O$ emissions. Song et al. (2020) used Random Forest (RF) to calculate feature importance, providing data-driven insights into the main contributors to $N_2O$ emissions. However, the study did not address the temporal forecasting of $N_2O$ emissions, a task for which RF models are generally not suited. More recently, Khalil et al. (2023) reported an ML modelling framework for developing $N_2O$ emissions soft sensors. Various decision tree-based models and a dense neural network (DNN) were trained to predict $N_2O$ emissions for the same time step as the model inputs. While these models provided accurate results, they are not capable of forecasting $N_2O$ emissions, which is crucial for implementing model-based or model-predictive control (MPC) frameworks that can steer the treatment processes to reduce $N_2O$ emissions. More specifically, ML models such as Artificial Neural Networks (ANNs) and Deep Learning (DL) have shown strong capabilities in time series forecasting. DL models can extract valuable knowledge from complex systems and identify patterns in the data (Zhang et al., 2018), offering the potential to utilise real-time online sensor data on the wastewater treatment process and operations to forecast $N_2O$ emissions. In this context, Xu et al. (2024) trained various Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models using 24 h of historical $N_2O$ emissions data to predict the $N_2O$ value one hour ahead. While the models achieved high accuracy ($R^2 > 0.90$), they lacked additional operation-based model inputs essential for implementing MPC, such as DO and internal recycle rates, and were restricted to one-time-step-ahead predictions which are not sufficient for MPC frameworks. Hwangbo et al. (2021) trained a DNN as a process model to predict $N_2O$ concentrations in the liquid phase and conducted preliminary investigations using an LSTM model to recursively forecast $N_2O$ concentrations. The LSTM model demonstrated high performance ($R^2 > 0.94$), but its evaluation was conducted on a single sample with a fixed prediction horizon, specifically corresponding to the last day of the dataset. Consequently, this assessment did not account for the model's forecasting capabilities across different seasonal variations and emissions peaks.

This study addressed key deficiencies in existing models by proposing new LSTM-based models to forecast $N_2O$ emissions using long-term data from a full-scale WWTP, including seasonal peaks. Unlike previous studies, we focussed on developing models that accurately predict $N_2O$ emissions over extended prediction horizons, capturing complex process dynamics and operational settings that trigger $N_2O$ emissions over time. We evaluated their performance to determine the feasibility of integrating these models into (near) real-time MPC strategies, ultimately enabling effective $N_2O$ emission mitigation in WWTPs. Additionally, we discussed the current capabilities and limitations of these ML models and offered new insights leading to potential future improvements.

## 2. Methodology

### 2.1. Overview

In this study, the investigated DL models utilised an LSTM-based encoder-decoder architecture. The methodology for DL model training and evaluation comprised five stages as explained in Fig. 1. First, model inputs were selected, measured using online sensor data, which are related to the nitrogen removal process and $N_2O$ production. Based on the assessment of WWTP operational data and understanding the system's response to varying control setpoints, an adequate length for the historical input data and prediction horizon length, which serves as the model outputs, was selected. In an MPC framework, the choice of the prediction horizon is also determined by the control variables and the control horizon applicable to the system. For reducing $N_2O$ production and emissions, optimised DO control is considered the most relevant variable for achieving (near) real-time reduction in $N_2O$ emissions in practice, given its role in the different $N_2O$ production pathways. The prediction horizon should be equal to, or longer than, the control horizon to account for process transients (Behera et al., 2015). Previous studies investigating MPC for DO control have utilised control horizons ranging from 1.25 – 2.50 h (Boruah and Roy, 2019; O'Brien et al., 2011; Shen et al., 2009), which can be considered an acceptable response time for the varying DO setpoint to influence the nitrification process. In this study, a longer prediction horizon of up to 6 h was considered to account for shorter control horizons and to test the feasibility of extending the prediction horizon for more extended forecasts.

### 2.2. Data pre-processing and preparation

Initially, the datasets underwent quality control to identify and address gross anomalies and sensor errors, such as values beyond sensor thresholds and sudden spikes. These anomalies were reviewed based on process-specific knowledge of the model inputs and replaced using linear interpolation between two known data points. The datasets were then resampled to a chosen data frequency to balance computational costs, while ensuring that the process dynamics and variations related to $N_2O$ production were still well represented. Additionally, the data were smoothed using a rolling mean with a window length of 5 data points. For training the LSTM-based DL models, the time series data were converted into a supervised learning problem, separating model inputs and providing labels of the outputs. The model inputs, which consisted of operational variables, such as ammonium ($NH_4^+$), mixed liquor suspended solids (MLSS) and DO in the aerobic tank, were transformed into sequences with a fixed temporal length of historical data ($n$). The model

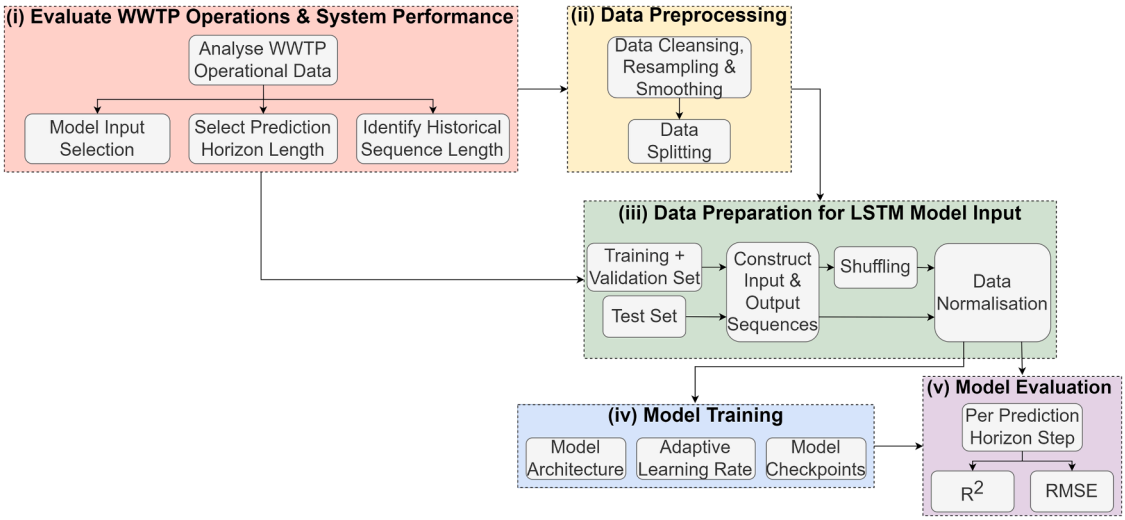



**Fig. 1.** Methodology overview employed in this study: (i) evaluate WWTP operations and system performance to determine key information for model inputs and outputs (described in Section 2.1); (ii) data pre-processing and (iii) data preparation for LSTM-based model input (both described in Section 2.2); (iv) model training, and (v) model evaluation (both described in Section 2.4).

output, which is the $N_2O$ concentration in the gaseous phase, matched the desired prediction horizon length ($h$), constituting the labels of observations following the input sequence. Each sample thus included sequences of both model inputs and outputs. Using a sliding window and moving 1 step at a time through the dataset, pairs of input and output sequences were created.

The entire dataset was split into training, validation and test sets. The training set was used to learn the underlying patterns and update the model weights, while the validation set was employed during training to tune hyperparameters, such as the learning rate, and save model checkpoints to prevent overfitting. The test set, an out-of-sample set completely unseen by the trained models, was used for evaluating and selecting the final model. Initially, the dataset was split with 80 % allocated to the combined model training and validation set, and 20 % reserved to the test set. The combined training and validation set was transformed into pairs of input and output sequences and was then randomly shuffled while retaining the elements of each sample pair. As illustrated in Fig. 2, for example, the model inputs and outputs sequences of *Sample 2 (purple colour boxes)* were retained, even though this sample was placed towards the end of training and validation set after shuffling. This shuffling is considered acceptable as predictions over the prediction horizon is a function of the model inputs of the last $n$ timesteps, and such a process can expedite the training process, leading to faster convergence (Kratzert et al., 2018). The shuffled dataset was further split into a training and validation set using an 85/15 ratio. The test set, which was separated prior to the shuffling, was kept unchanged to retain the temporal integrity of the original dataset. Finally, the training set, was standardised using Z-score normalisation by subtracting the training set's mean and dividing by the training set's standard deviation (Rahu et al., 2024). These same statistical parameters were then used to standardise the validation and test sets to prevent data leakage.

### 2.3. Deep learning models

In this study, LSTM units within an encoder-decoder architecture were used to forecast $N_2O$ emissions based on historical operational data. The model inputs consisted of sequences of historical data of length $n$, while the outputs were sequences of future $N_2O$ values over a prediction horizon length $h$. The following sections detail the structure and functionality of the LSTM units and the used encoder-decoder architecture.

#### 2.3.1. Long short-term memory units

Long Short-Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997), are a type of recurrent neural network (RNN) designed to learn long-term dependencies in sequential data, making them suitable for time series forecasting. Each LSTM unit comprises four gates: the forget gate ($f$), which regulates the retention of previously hidden and cell states' information; the input gate ($i$), which determines the amount of current information retained after updating the current cell state via the cell update gate ($c$); and the output gate ($o$), which controls the

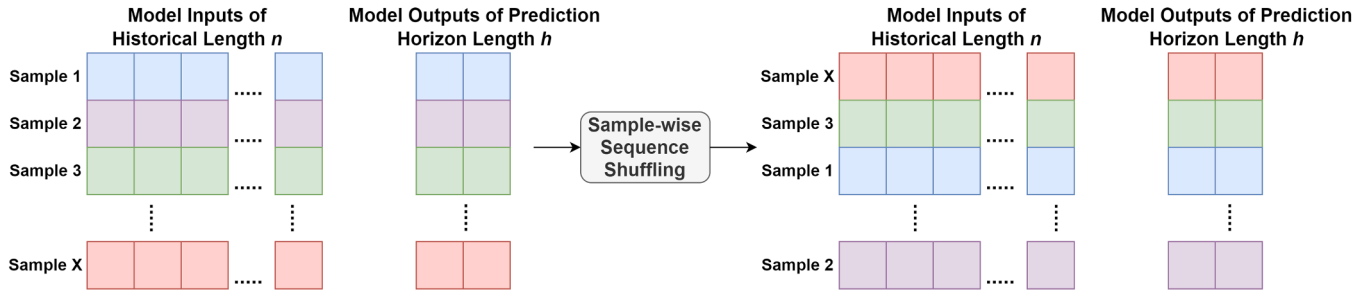



**Fig. 2.** Example representation of a sample-wise sequence shuffling of dataset used in model training prior to splitting into training and validation sets. For a dataset of X samples, each sample consist of a pair: model inputs that include historical data spanning back n timesteps, and model outputs that represent the forecasted values for the next h timesteps ahead, corresponding to the prediction horizon. The number of boxes shown is purely for illustrative purposes. After processing, each sample retains its pairs while the order of the samples is randomly shuffled.

information output based on the internal cell state. These gates are represented as linear transformations that consider the inputs, recurrent information from previous LSTM cell states, and trainable weights and biases. A sigmoid activation function is utilised for the *f, i* and *o* gates, specific to the retention of recurrent information, and a hyperbolic tangent activation function is used for updating the hidden and cell states (see Figure S1 for a visual representation).

### 2.3.2. Encoder-Decoder architecture-based DL model

A sequence-to-sequence (seq2seq) model architecture can be used to effectively process historical input sequences of length *n* to forecast $N_2O$ concentration over a prediction horizon of length *h*. Originally developed for natural language processing (Sutskever et al., 2014), seq2seq models have since been adapted for various applications, including text generation, conversational models (Ren et al., 2019) and time series prediction (Xu et al., 2021). Fig. 3 illustrates the generic seq2seq model architecture implemented in this study.

The model contained an encoder component that processed the input sequence, which includes the historical values of length *n* for each of the *k* number of model input variables, using an LSTM layer to generate the hidden and cell states. The hidden state served as the contextual and latent state representation of the input sequence, which is then repeated *h* times to match the output target sequence length. This state vector was then inputted to the decoder component, which contained another LSTM layer. The decoder's states were initialised with the hidden and cell state from the encoder's LSTM layer. The entire decoder output sequence was returned and then flattened to ensure a full connection with the output dense layer. This ensured that every recurrent output from the decoder was connected to the output layer. The model outputs $N_2O$ forecasted at once for the entire prediction horizon *h*.

## 2.4. DL model training and evaluation

Models using the LSTM-based encoder-decoder architecture were trained with the dedicated training set, containing input sequences of the chosen model inputs to forecast all *h* steps in the prediction horizon for $N_2O$ concentration in a one-shot manner, generating all steps simultaneously. A summary of the hyperparameter values and choices for the training is provided in Table 1. Models of different sizes and hence, with an increasing number of trainable weights, were trained to assess the necessary complexity needed for good forecasting performance. Specifically, models with varying LSTM units in the encoder and decoder components – 32–32; 64–64; 128–128; 256–256; 512–512 – were trained.

The model training utilised stochastic gradient descent optimisation with the AdamW optimiser (Loshchilov and Hutter, 2019), to minimise a mean squared error (MSE) loss function. For each iteration, the optimisation results were backpropagated, updating the model's trainable weights and biases. The batch size for each iteration of the gradient descent was set to 32. An adaptive learning rate procedure was employed, reducing the learning rate by 0.5 when the MSE loss on the validation set for a given epoch did not decrease by a threshold of 0.01. This ensured a smoother optimisation process and increased the likelihood of identifying the global minimum. The initial learning rate was set to 0.001, with a minimum learning rate of 0.00001. Furthermore, if the model's validation MSE loss decreased compared to the previous epoch, the model was saved, and a checkpoint was created to prevent overfitting in case of divergence during training. Each model was trained for 100 epochs. For each model network size, 10 different models with varying initialised weights and biases were trained to assess the reproducibility and consistency of the results. Finally, dedicated performance metrics were calculated for the training, validation and test set for each time step in the prediction horizon to assess and compare the forecasting performance. All model development and training activities were conducted using the Python software library of TensorFlow (Abadi et al., 2016).

**Table 1**
Hyperparameter values and choices used during model training.

| Hyperparameter | Value/Choice |
|---|---|
| LSTM units (encoder-decoder) | 32–32, 64–64, 128–128, 256–256, 512–512 |
| Epochs | 100 |
| Optimiser | AdamW |
| Learning rate | Adapted during training |
| Loss function | Mean squared error |
| Activation function for LSTM layers | tanh |
| Batch size | 32 |

The performance of the trained models of varying network sizes was assessed by comparing their forecasting performance on the test set, that is, the unseen data not used during the training and validation procedure. For this, the root mean squared error (RMSE) was used:

$$RMSE = \left\{ \sqrt{\frac{\sum_i^N \left(y_{i,t} - \hat{y}_{i,t}\right)^2}{N}} \middle| t = 1,\ 2,\ \ldots\ldots,\ h \right\} \quad (1)$$

where $\hat{y}_{i,t}$ and $y_{i,t}$ are the predicted and measured values for a given sample $i$ and for a given time instance $t$ in the prediction horizon, respectively. $N$ is the total number of samples. This leads to a list of RMSE metrics of length $h$ representing the individual forecasting performance within the prediction horizon of length $h$. Furthermore, the performance metric coefficient of determination (CoD or $R^2$) was also used:

$$R^2 = \left\{ 1 - \frac{\sum_i^N \left(y_{i,t} - \hat{y}_{i,t}\right)^2}{\sum_i^N \left(y_{i,t} - \bar{y}_t\right)^2} \middle| t = 1,\ 2,\ \ldots\ldots,\ h \right\} \quad (2)$$

where $\bar{y}_t$ is the average value of the observed data for the given time instance $t$ in the prediction horizon, calculated over the $N$ data samples.

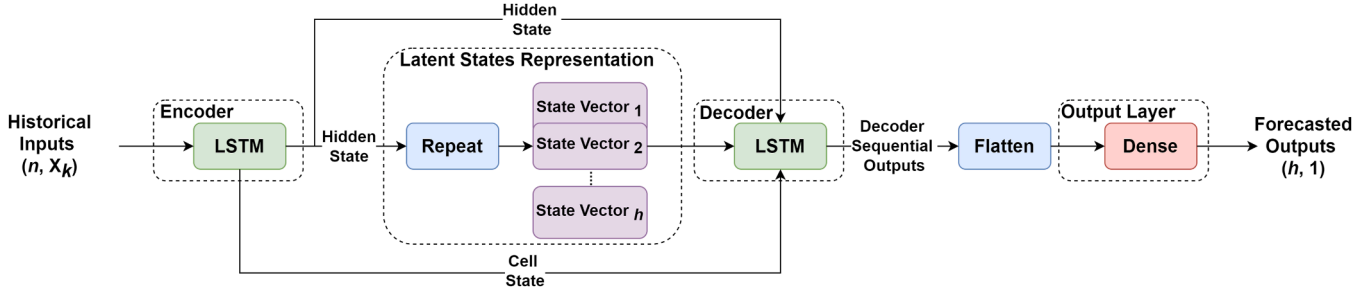



**Fig. 3.** A generic seq2seq model architecture adopted in this study, containing an encoder, latent state representation and decoder components. n denotes the length of the historical inputs, k denotes the number of inputs, and h denotes the length of the output and the prediction horizon utilised.

### 2.5. Permutation feature importance

Permutation feature importance evaluates the significance of individual model inputs (features) in predicting a given model output (target). This is achieved by permuting each input and assessing its impact on model predictions. Specifically, in this study, this method was used to assess the importance of the individual operational variables across each step of the prediction horizon, determining their relevance in forecasting $N_2O$ emissions within this timeframe. The influence of the inputs was evaluated for each step in the prediction horizon using the RMSE metric, and hence, a relative increase in the RMSE upon permutation, compared to the baseline, indicates the level of importance of the respective input. To ensure robustness, the procedure was iterated 10 times and an average RMSE value was computed.

## 3. Case study

### 3.1. Description of WWTP

The case study for this ML modelling investigation is the Amsterdam West WWTP with a capacity of 1.1 million population equivalent (average flow of 168 MLD), treating municipal wastewater. Details regarding the overall process configuration can be found in Seshan et al. (2024). Specifically, the WWTP contains seven treatment lanes that conduct the activated sludge (AS) process for biological nitrogen and phosphorous removal. The AS process configuration applied is the modified University of Cape Town (mUCT) process (Chen et al., 2023). Each treatment lane contains a bioreactor that has an anaerobic-anoxic-facultative-aerobic configuration in series. The facultative tank serves as a swing tank, providing additional denitrification or nitrification capacity based on the treatment requirements. The bioreactor units are covered, allowing for the capture of the off-gas emissions. Internal recycles are typically present for any continuous AS process configurations. In Amsterdam West, four internal recycles are operated; where one sludge recycle line returns active biomass from the secondary clarifier to the anoxic tank, another one returning active biomass from the anoxic to the anaerobic tank. The additional two internal recycles transfer the $NO_x$, produced during the nitrification process in the aerobic tank, to the anoxic and facultative tanks, when the latter is being operated as an anoxic zone. This modelling study was conducted only on one treatment lane due to the availability of online sensors measuring $N_2O$ and other crucial process variables .

### 3.2. WWTP data

Online sensor data from the treatment lane spanning 1 year and 3 months, from 11/2020 – 03/2022 were employed for training and evaluating the DL models. Process related variables were measured online up to a frequency of every 15 min. The $N_2O$ concentrations in the gaseous phase were measured in *ppm* every 15 min by sampling off-gases from the closed bioreactor, which were subsequently analysed by an infrared gas analyser (X-stream, Emerson, St. Louis, MO, US). The dataset comprised 13 model inputs, including the raw influent flowrate, $NH_4^+$, DO and $NO_x$ concentrations in the aerobic tank; $NO_x$ in the anoxic tank; MLSS and liquid temperature in the bioreactor; two internal recycles transferring the $NO_x$ from the aerobic tank to the anoxic tanks ($NO_x$ Recycle 1 and $NO_x$ Recycle 2), the position of three aeration valves (two in the aerobic tank and one in the facultative tank), where 0 % indicates fully closed and 100 % indicates fully open; and the $N_2O$ off-gas concentrations. The model output was the $N_2O$ off-gas concentrations.

Descriptive statistics for these variables are presented in Table 2, and their distributions over the dataset's duration are illustrated in Fig. 4. As detailed in Section 2.2, the dataset was resampled to a chosen frequency of 30 min. This choice was based on prior knowledge of the dynamics and rate of change typically observed within the wastewater treatment systems. This choice balanced data resolution with computational costs to allow for practical model training and evaluation. The datasets were prepared into sequences suitable for training LSTM-based models. A historical input length $n$ of 48 timesteps (equivalent to the last 24 h at 30 min-frequency) to capture diurnal patterns, including morning and evening peaks in total nitrogen (TN) as expected in the raw influent. As discussed in Section 2.1, the prediction horizon length $h$ was set to a maximum of 12 data points, corresponding to 6 h. The datasets were split into training, validation and test sets, as explain in Section 2.2. The period from 01/11/2020 to 25/11/2021 were allocated to the training and validation sets, encompassing all four seasons, and capturing a significant $N_2O$ emissions peak for model training. A dedicated test set covered the remaining dataset from 26/11/2021 to 28/02/2022, capturing another distinct $N_2O$ emissions peak. This partition ensured robust evaluation of the DL models' forecasting capabilities across varying seasonal and process conditions.

**Table 2**
Descriptive statistics of the model inputs and outputs used for model training and evaluation.

| Variable | Minimum | Mean | Median | Maximum | Standard Deviation |
|---|---|---|---|---|---|
| Influent [$m^3$/h] | 246.3 | 1003.3 | 984.4 | 3901.3 | 487.9 |
| $NH_4^+$- Aerobic [mg/L] | 0 | 1.6 | 1.1 | 20.0 | 2.2 |
| DO - Aerobic [mg/L] | 0.06 | 1.4 | 1.2 | 3.8 | 0.7 |
| $NO_x$ - Aerobic [mg/L] | 0.9 | 5.7 | 5.4 | 19.8 | 2.3 |
| $NO_x$ - Anoxic [mg/L] | 0.05 | 0.7 | 0.5 | 12.5 | 0.8 |
| MLSS [g/L] | 2.6 | 4.2 | 4.2 | 6.2 | 0.4 |
| Liquid Temp [Celsius] | 10.5 | 16.6 | 15.7 | 22.6 | 3.4 |
| Aeration Valve 1 – Aerobic [%] | 0 | 56.4 | 50.7 | 100 | 20.7 |
| Aeration Valve 2 – Aerobic [%] | 0 | 55.5 | 50.6 | 93.3 | 19.5 |
| Aeration Valve – Facultative [%] | 0 | 9 | 0.4 | 100 | 24.7 |
| $NO_x$ Recycle 1 ($m^3$/h) | 0 | 4609.6 | 5049.9 | 5050 | 975.9 |
| $NO_x$ Recycle 2 ($m^3$/h) | 0 | 5054.0 | 5960.8 | 6092.1 | 1305.7 |
| Gaseous $N_2O$ (ppm) | 0 | 52.5 | 21.9 | 522.5 | 74.4 |

## 4. Results

### 4.1. Forecasting performance of varying model network sizes

Each of the 5 LSTM-based encoder-decoder networks was trained 10 times, thereby yielding a distribution of performance results. All models performed well on the training and validation sets, with $R^2$ values for 6 h ahead prediction (i.e. $h = 12$) exceeding 0.96. The performances of the models on the training and validation sets are shown in Figure S2 and Figure S3, respectively. The RMSE performance on the training and validation sets indicated that the more complex models with 128–128, 256–256 and 512–512 LSTM units performed better than simpler models with 32–32 and 64–64 units, suggesting that model complexity enhanced learning accuracy. However, the 256–256 and 512–512 models showed similar training results, indicating that increasing model complexity beyond 256 LSTM units yielded diminishing returns in terms of prediction accuracy. Nonetheless, further analysis is needed to assess whether any additional benefits arise from increasing the number of LSTM units, particularly when considering the trade-off with higher computational costs.

Fig. 5 presents the forecasting performance on the test (i.e. unseen) data set. Results showed that all models' performance deteriorated with

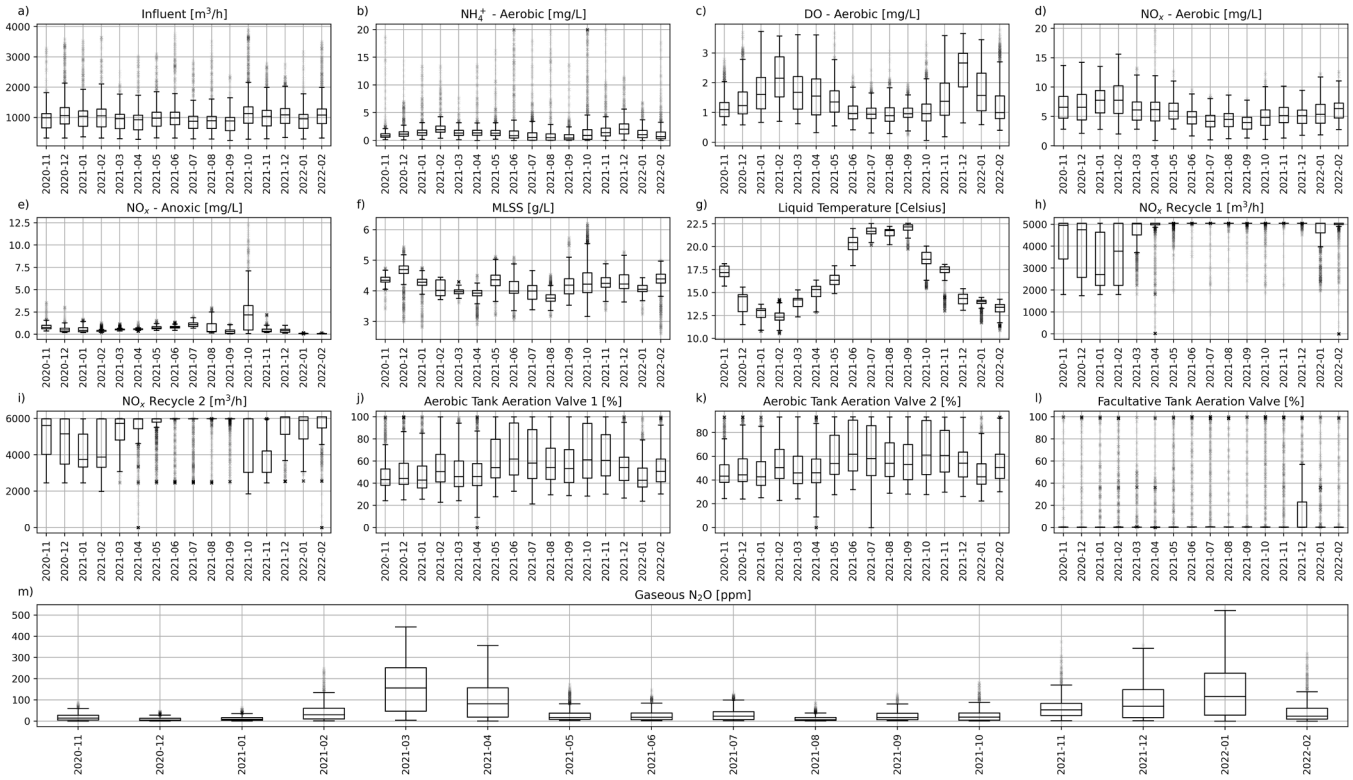



**Fig. 4.** Graphical representation of the distribution of the model input variables (a – m) and the model output variable (m) across the duration of the dataset used for model training and evaluation.

increasing prediction horizon steps. Simpler models, using 32–32 and 64–64 LSTM units, performed poorly, with RMSE values for the $h = 12$ step ranging from 64.7 to 74.5 ppm and 63.9 to 69.4 ppm, respectively, and $R^2$ values from 0.39 to 0.54 and 0.47 to 0.55. These models also showed larger variations in the evaluation results across the 10 trainings runs compared to the 128–128, 256–256 and 512–512 models. This suggested that more complex models containing higher number of LSTM units led to more robust and stable models with reproducible results. These findings highlighted the need for an adequate number of trainable weights in the model architecture to effectively learn the dynamics and process conditions from the model inputs, resulting in accurate forecasts of $N_2O$ emissions over the prediction horizon.

Consistent with the training and validation results, the 256–256 and 512–512 models exhibited better performances on the test sets, with RMSE values for 6 h ahead forecasts ($h = 12$) ranging from 61.1 to 67.5 ppm and 57.3 to 63.7 ppm, respectively and $R^2$ values 0.50 to 0.59 and 0.55 to 0.64, respectively. The best performing model with a 512–512 network was considered an outlier. This is illustrated in Fig. 5, as the model's performance did not represent the median model performance observed across all model training runs with the same network configuration. This can be attributed to the stochastic nature of the training process and the random initialisation of the trainable weights, which may result in isolated instances of high performance. This raises questions about the reproducibility of such model performance and it was excluded from further analysis. As a result, the best performing model chosen based on the results on the test set was a model with a 256–256 network (RMSE = 61.1 ppm and $R^2 = 0.59$ for $h = 12$), as detailed in Section 4.2.

### 4.2. Best performing model forecasts

The best model, based on performance for each step in the prediction horizon, was the 256–256 LSTM unit-architecture. A deterioration in forecasting accuracy was observed as the prediction horizon step increased, shown by an increase in RMSE and a decrease in $R^2$ testing values; see Table 3. The model achieved high accuracy for forecasts up to 2 h, with $R^2$ values above 0.86, and reduced performance for forecasts up to 4 h and the final prediction step of 6 h ahead, with the $R^2$ values being above 0.72 and 0.59, respectively. In Fig. 6, example windows are provided illustrating the one-shot forecasts made by the model, compared with the observed $N_2O$ concentrations in the gaseous phase. The observed $N_2O$ (x markers) leading up to the model forecasts signify the model inputs time period (see Section 3.2). Fig. 6a) and Fig. 6b), show that the model performed well in accurately forecasting the increase in $N_2O$ concentrations during the seasonal emissions peaks observed in January 2022. Examples of the model performing satisfactorily for low to mid $N_2O$ concentrations are provided in Fig. 6c) and Fig. 6d).

Fig. 7 illustrates time series forecasts of $N_2O$ concentrations for each prediction horizon step, compared to observed $N_2O$ concentrations, across different emission scenarios. The darkest red line represents forecasts of 0.5 h ahead and the lightest red line represents forecasts of 6.0 h ahead. The $N_2O$ forecasts for the entire test set period are provided in Figure S.4 and Figure S.5. As it can be seen from Fig. 7a), the model accurately captured the $N_2O$ peaks and responded to the fluctuation in $N_2O$ concentrations, including the diurnal peaks seen within a day. However, the model showed a sub-optimal fit for larger peaks when attempting to forecast across steps of the horizon, as illustrated by the peak observed on 21 January 2022 at 18:00, in Fig. 7a). This peak possibly represents an isolated incident, potentially caused by unstable operational conditions leading, to higher-than-normal $N_2O$ production during the nitrification or denitrification process. In addition, on 23 January 2022, an interesting observation appeared: the dynamics in $N_2O$ concentrations showed an unusual trend compared to typically seen in the dataset, notably with a diminished second diurnal peak. This shift could be attributed to a change in flow and load patterns, as the second diurnal peak was missing in the influent flowrate. However, the model failed to anticipate this trend shift and instead forecasted a peak in $N_2O$

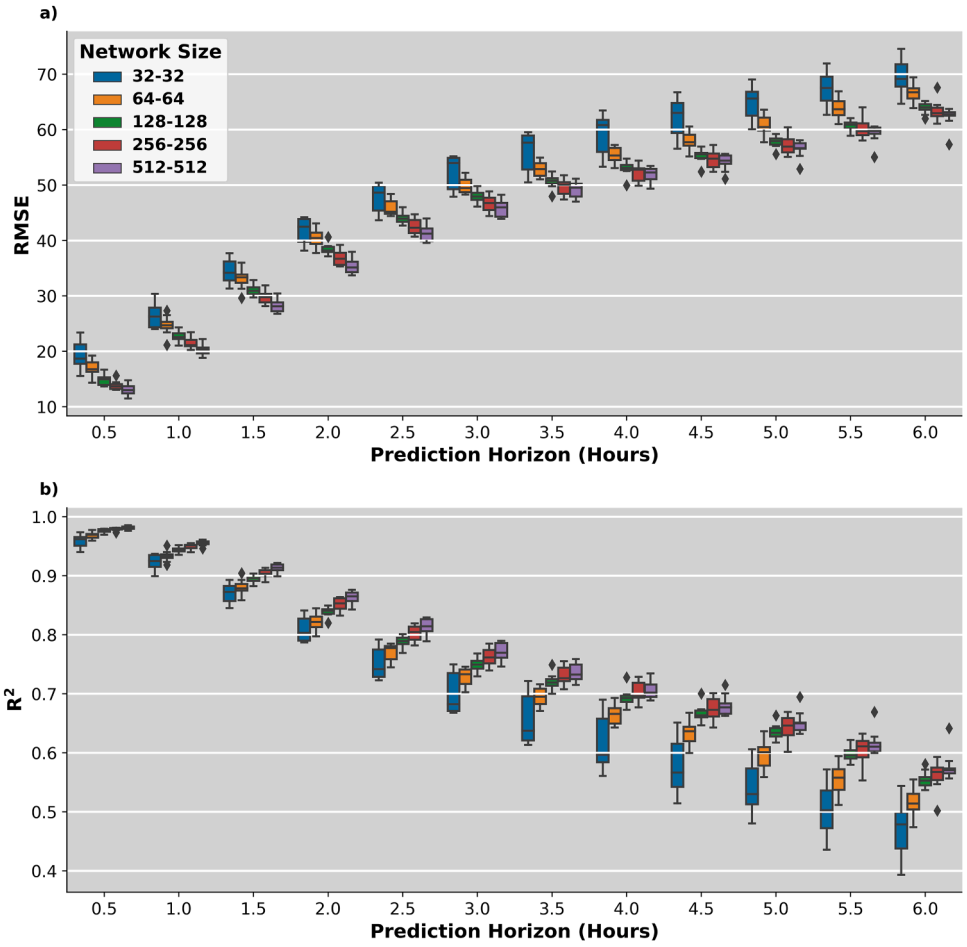



**Fig. 5.** Forecasting performance on the test set of 5 model network sizes and complexity over the prediction horizon using the a) RMSE (in ppm) and b) $R^2$ metrics.

**Table 3**
RMSE (ppm) and $R^2$ values by the best performing model with a 256–256 LSTM unit-architecture.

| Prediction horizon (hours) | RMSE (ppm) | $R^2$ |
|---|---|---|
| 0.5 | 13.6 | 0.98 |
| 1.0 | 21.1 | 0.95 |
| 1.5 | 29.0 | 0.91 |
| 2.0 | 35.9 | 0.86 |
| 2.5 | 41.4 | 0.81 |
| 3.0 | 45.5 | 0.77 |
| 3.5 | 48.4 | 0.74 |
| 4.0 | 50.7 | 0.72 |
| 4.5 | 52.8 | 0.69 |
| 5.0 | 55.2 | 0.67 |
| 5.5 | 58.0 | 0.63 |
| 6.0 | 61.1 | 0.59 |

concentrations. Therefore, it can be seen that the model is vulnerable to isolated incidents and sudden shift trends. These discrepancies suggested potential limitations in the training set, possibly due to insufficient examples of similar process conditions or inadequacies in the current set of model inputs that captured all factors influencing $N_2O$ production. The model performed well for more immediate forecasts for low $N_2O$ emissions but showed greater variability. This could be attributed to the ambiguous trends observed during low emission periods.

### 4.3. $N_2O$ forecasting for varying prediction horizon lengths

Among WWTPs, process dynamics and system response times can vary significantly in response to operational and control changes. Therefore, different prediction horizon lengths can be utilised in model development and training, to forecast $N_2O$ emissions. The proposed methodology can be applied for various wastewater treatment systems, where shorter prediction horizons allow adequate time for control adjustments, influencing the process. Fig. 8 illustrates the performance of various models, each having the 256–256 LSTM-unit architecture, on the test set. These models were trained with different prediction horizon lengths as outputs. When comparing the RMSE values (Fig. 8a), a trade-off can be seen in making immediate forecasts up to 1.5 h ahead. Models targeting shorter horizons showed higher performance compared to those with longer horizons. For example, a model trained to forecast 0.5 h ahead ($h = 1$) achieved an RMSE of 4.4 ppm, while models trained to forecast 5 or 6 h ahead ($h = 10$ or $h = 12$) achieved RMSEs of 13.7 and 13.6 ppm, respectively, for the 0.5-hour prediction horizon step. However, it became clear that choosing a longer prediction horizon length resulted in only a marginal decrease in model performance for forecasts spanning from 2.0 to 6.0 h ahead, where for example, RMSE (Fig. 8a) and $R^2$ (Fig. 8b) values for the 3.0-hours prediction horizon step, there was only an increase of 2.1 ppm in RMSE and a decrease in 0.02 in $R^2$. Such changes can be considered statistically insignificant and inconsequential in practice. As a result, selecting a longer prediction horizon length for the model does not significantly compromise the model's performance for intermediate forecasts.

Performance metrics such as RMSE and $R^2$ provide valuable indicators of how these models might perform when integrated into

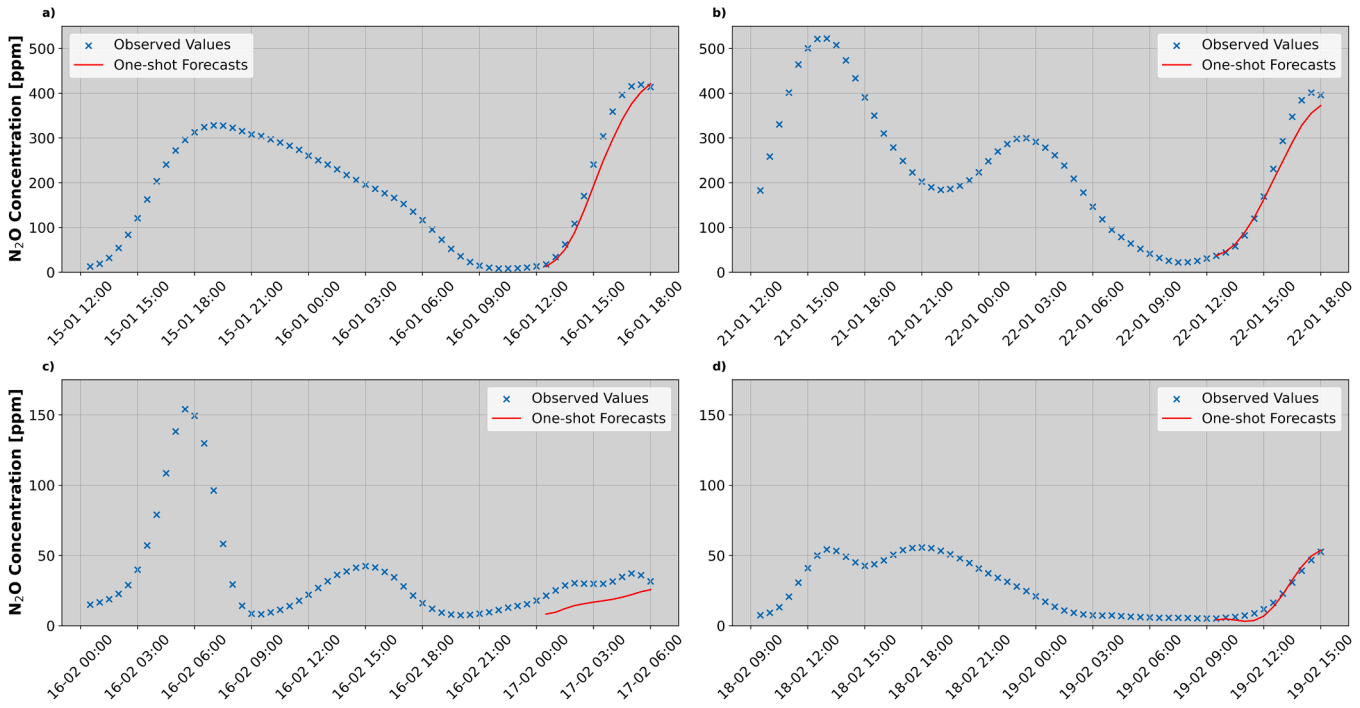



**Fig. 6.** Examples of comparing the $N_2O$ observed concentrations (x markers) and one-shot $N_2O$ emissions forecasts (red line) made 6 h ahead ($h = 12$) by the 256–256 LSTM-unit architecture, using data from the last 24 h ($n = 48$) with multiple model inputs. Panels a) and b) show high $N_2O$ emissions peaks, and c) and d) illustrate low to mid- $N_2O$ emissions peaks.

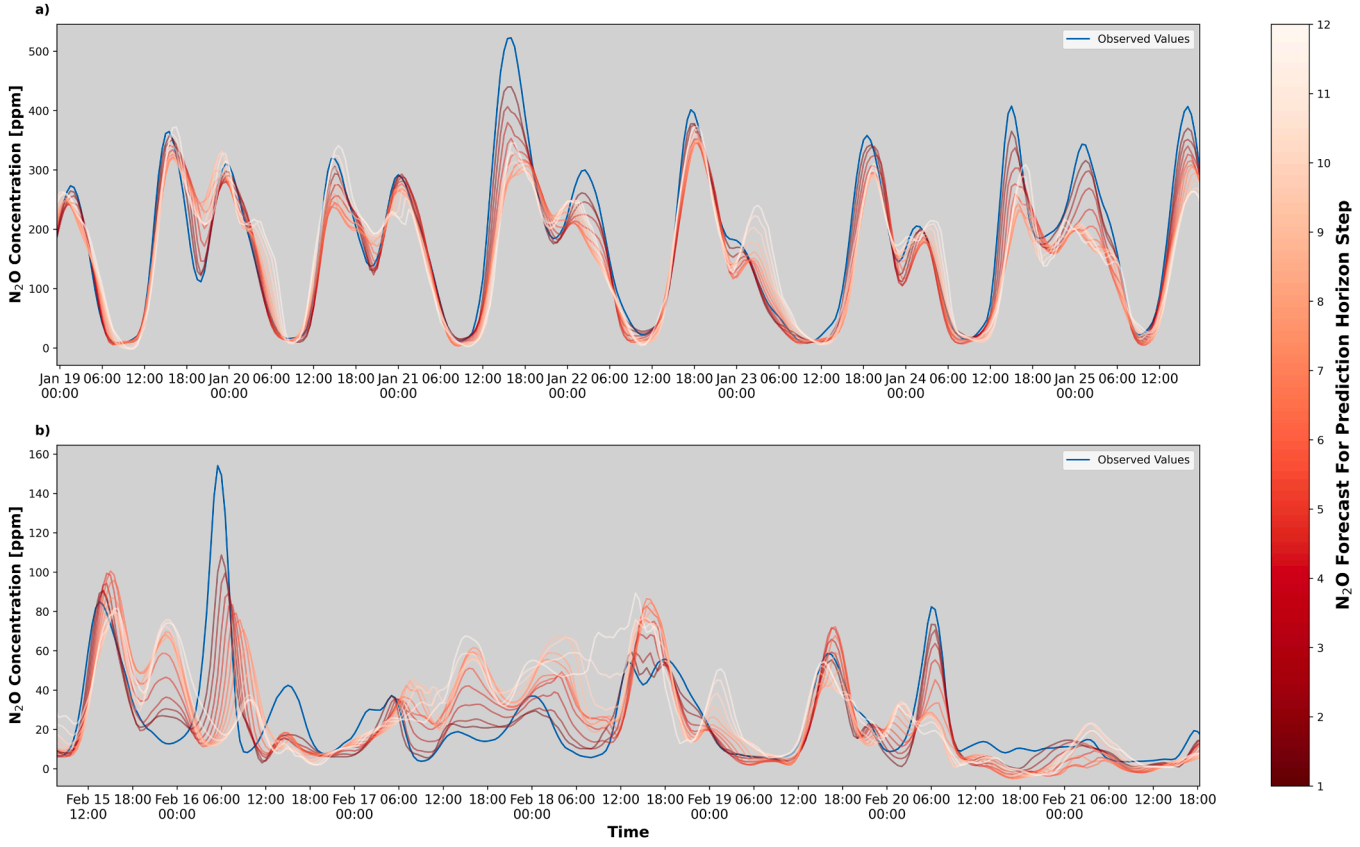



**Fig. 7.** Time series representation of $N_2O$ forecasts at each prediction horizon step (h) compared to observed $N_2O$ concentrations (blue line). Panel a) shows high $N_2O$ emissions, while panel b) shows mid to low $N_2O$ emissions.. The forecast for $h = 1$ is represented by the darkest red line and $h = 12$ by the lightest red line.

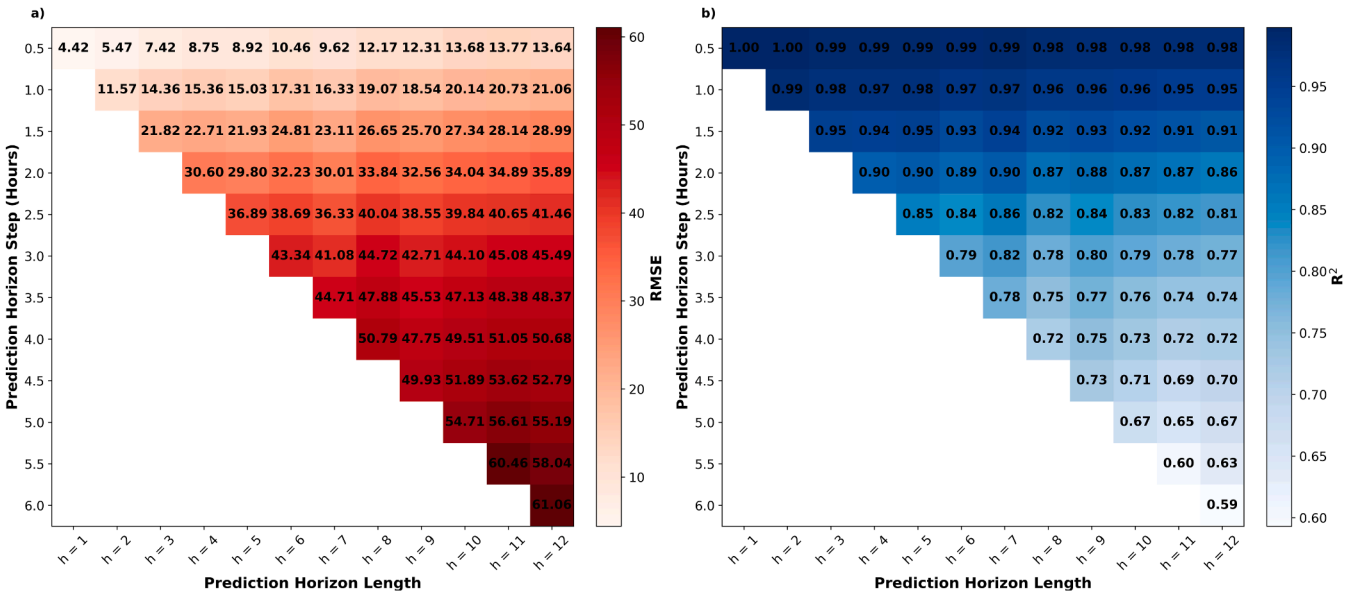



**Fig. 8.** Performance of models with 256–256 LSTM-unit architecture when trained with varying prediction horizon lengths and using the a) RMSE and b) $R^2$ metrics for each prediction horizon step.

control strategies. RMSE penalizes large errors more heavily and thus reflects the model's ability to forecast high $N_2O$ emission events, especially important for shorter prediction horizons where rapid responses are needed. In contrast, $R^2$ measures how well the model captures variability and predicts fluctuations in $N_2O$ emissions, providing insight into how accurately the model can forecast trends over the prediction horizon. Together, these metrics offer confidence in the model's capabilities and reveal its limitations in terms of accuracy and robustness when applied to operational control, particularly when considering different system response times to control changes over various prediction horizons.

### 4.4. *Feature importance for forecasting $N_2O$ emissions*

Fig. 9 illustrates the permutation feature (i.e. model input) importance results as a relative change in RMSE on the test set, compared to the baseline where no model inputs were permuted. The importance of each input is computed for every prediction horizon step, providing insights into the individual impact they have in making subsequent forecasts for $N_2O$ emissions. As expected, due to the strong autoregressive nature, the $N_2O$ concentrations in the gaseous phase had a very high impact on the model's immediate forecasts. However, there is a significant drop in this input's influence as the prediction horizon step increases, from 747 % to 29 % relative RMSE change, suggesting the

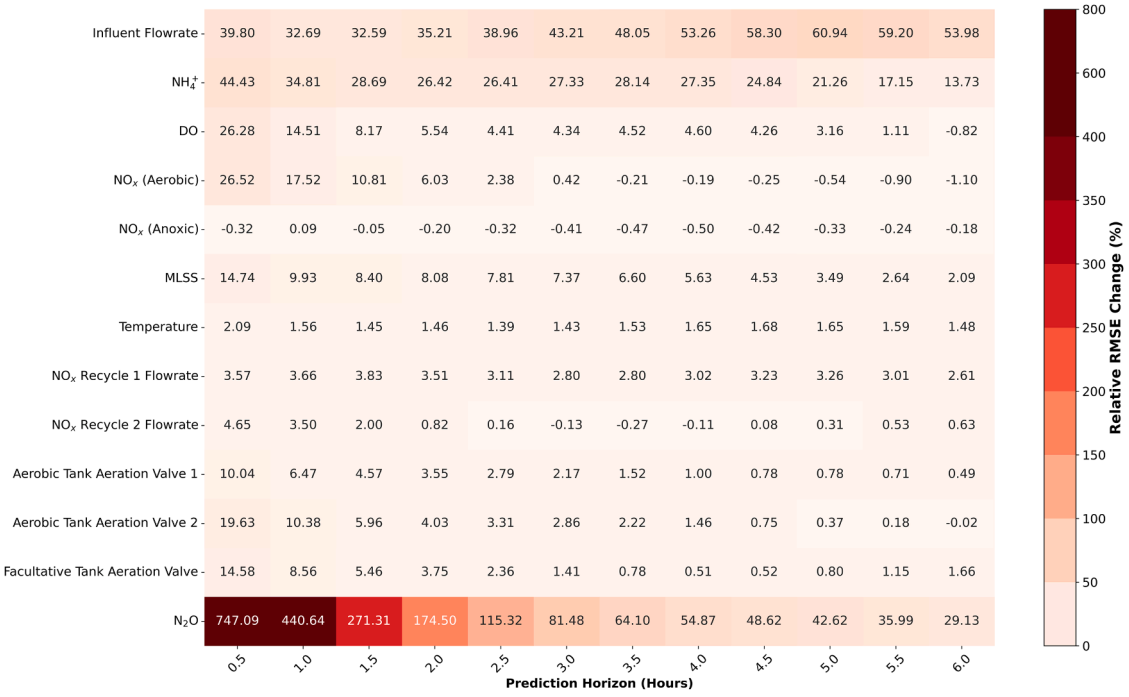



**Fig. 9.** Heatmap of permutation feature importance showing the relative change in RMSE for each prediction horizon step across all model inputs, compared to the baseline where no inputs are permuted. The colour intensity represents the magnitude of RMSE change.

necessity of including other inputs that represented related process dynamics to adequately explain the $N_2O$ emissions.

The $NH_4^+$ input in the aerobic tank showed an influence on immediate forecasts, impacting RMSE changes of 44 % and 35 %, for 0.5 and 1.0 h ahead, respectively. For the intermediate forecasts up to 4.5 h ahead, the $NH_4^+$ input maintained a constant influence, with RMSE impacts between 25 % and 29 %, followed by a drop in its influence for the remaining horizon steps. Similarly, the influent flowrate impacted immediate forecasts, with RMSE changes of 32 % and greater for forecasts up to 1.0 h ahead. Interestingly, the influent flowrate was observed to increase in importance with subsequent prediction horizon steps, eventually becoming more important than the autoregressive input of the $N_2O$ concentrations at the 4.5-hour mark. The influences of $NH_4^+$ and influent flowrate could be attributed to the fact that the $N_2O$ emissions were largely a response to the $NH_4^+$ load received by the WWTP and, subsequently, the bioreactor. The $NO_x$ and DO inputs in the aerobic tank showed limited influence on the $N_2O$ forecasts and a significant decrease in importance across the prediction horizon. This could be attributed to the fact that the $N_2O$ emissions primarily followed the $NO_x$ and DO trends. Consequently, historical $NO_x$ and DO inputs had limited impacts on forecasting $N_2O$ emissions in the chosen frequency, as they were more reflective of present conditions rather than predictive for future changes.

The MLSS input had a minor influence on the $N_2O$ forecasts, which could be due to this input representing slow dynamic processes such as the sludge retention time (SRT). The aeration valve inputs also showed minor influences on $N_2O$ forecasts, particularly for immediate forecasts, and had no influence on subsequent prediction horizon steps. The aeration valves indirectly represented the amount of airflow received by the aerobic and facultative tanks. Therefore, the minor influence seen for immediate forecasts could be partly attributed to the stripping of $N_2O$ concentrations from the liquid to the gaseous phase. When considering other model inputs, such as the $NO_x$ in the anoxic tank and temperature, they had negligible impact in making the $N_2O$ forecasts over the prediction horizon. The insignificance of the anoxic $NO_x$ concentrations suggests that intermediates $NO_3^-$ and $NO_2^-$ are being reduced to $N_2$ during heterotrophic denitrification. Apparently, any inhibition preventing the reduction of $N_2O$, leading to its accumulation during this process, was not represented in the inputs. The insignificance of temperature may be due to this input representing seasonal variations and hence, no variations are seen in these inputs within the time windows used for the model inputs and the prediction horizon.

The decrease in forecasting performance, combined with the lack of highly impactful inputs for the 4 to 6-hour prediction horizon steps, suggested missing inputs that could have improved the forecasting of $N_2O$ emissions for these timeframes. For example, at the 6-hour prediction horizon step, all model inputs contributed to less than a 60 % change in RMSE, suggesting the absence of critical inputs.

## 5. Discussion

The best performing LSTM-based encoder-decoder model of 256–256 architecture achieved relatively high accuracy in forecasting $N_2O$ concentrations across different time steps of the prediction horizon (see Table 3). Specifically, on a fully sequentially separated and unshuffled test set, the model attained $R^2$ values of 0.98, 0.95, 0.86, 0.72 and 0.59 for 0.5, 1.0, 2.0, 4.0 and 6.0 h ahead, respectively.

A key advantage of the LSTM-based encoder-decoder model presented in this study is its ability to seamlessly utilise historical sequential data to forecast all timesteps within the prediction horizon in a single model run. Therefore, this model effectively captured slower process dynamics and time lags within the system, considering the non-instantaneous triggers of $N_2O$ production and subsequent emissions. In contrast, other predictive methods, such as Random Forest or Support Vector Machines, often require extensive feature engineering to account for temporal dependencies and need the training of separate models for each prediction horizon step. This makes the LSTM-based encoder-decoder model more efficient and well-suited for real-time control applications in WWTPs.

In addition, the best performing LSTM-based encoder-decoder model reported in this work outperformed the existing models reported in the literature. For the one-step-ahead prediction (0.5 h ahead), the model's performance was slightly better than the soft sensor models based on Random Forest, XGBoost, DNN, and AdaBoost developed by Khalil et al. (2023), which achieved $R^2$ values of 0.91, 0.93, 0.94 and 0.95, respectively, on a randomly split test set from the same dataset, used to predict $N_2O$ emissions for the same time instance as the inputs. It is noteworthy that in that work, the soft sensors were not designed as time series models and did not consider any temporal dependencies. Hence, they lack the capabilities for $N_2O$ forecasting and are therefore not suitable for MPC of $N_2O$ emissions.

The model presented here also outperformed a Support Vector Machine regression model predicting dissolved $N_2O$ concentrations in an aerobic tank, as reported in Vasilaki et al. (2020), which achieved an $R^2$ of 0.72. The best LSTM based encoder-decoder model developed in this study also outperformed the model trained by Xu et al. (2024), who reported a best performance of $R^2 = 0.915$ from their LSTM model forecasting $N_2O$ emissions one step (i.e. one hour) ahead. Similarly, Hwangbo et al. (2021) trained an LSTM model using only liquid-phase $N_2O$ concentrations to forecast one-step-ahead. The model was used recursively to predict subsequent steps, achieving an overall $R^2$ of 0.94. However, this evaluation was based on a limited sample, where a fixed one-day prediction horizon was used. In contrast, our study evaluated the model using several months of seasonally varying $N_2O$ emissions data, providing a more comprehensive assessment of forecasting capabilities.

The best model provided accurate forecasts by capturing operational and seasonal variations within the unseen test dataset (see Figure S4 and S5). Notably, as the test set included one of the two emission peaks observed in the full dataset, the model demonstrated its ability to predict this previously unseen peak. Furthermore, the best model of this study provided accurate forecasts for high $N_2O$ emissions, particularly during seasonal peaks, while exhibiting greater variations in forecasting low $N_2O$ emissions (see Fig. 7). This is desirable within an MPC framework, since the primary objective of optimising the control of wastewater treatment processes is to avoid the high $N_2O$ emissions. Additionally, measuring low $N_2O$ concentrations can be highly uncertain due to the detection limits of the analytical equipment and the small mass fraction of $N_2O$ compared to the overall nitrogen balance. Therefore, if only the model's performance in predicting high $N_2O$ emissions is considered, such a model shows promise for operational use within an MPC framework.

The best performing model reported in this work identified slightly different key model inputs than previously reported in the literature. The feature importance analysis shown in Section 4.4 identified the key model inputs as past $N_2O$ emissions followed by influent flowrate, $NH_4^+$, $NO_x$, and DO in the aerobic tank, indicating the influence of the fast and dynamic nitrification process and AOB-related $N_2O$ production pathways. The minor importance of $NO_x$ and the lack of influence of temperature in making short-term $N_2O$ forecasts differ from previous findings in the literature (Hwangbo et al., 2020; Khalil et al., 2023), where $NO_2^-$ and $NO_3^-$ were identified as critical and temperature was concluded to be the most influential input for making $N_2O$ predictions. These discrepancies can be attributed to differences in modelling objectives. Prior studies focussed on predicting $N_2O$ for the current situation, whereas our study focused on historical temporal dependencies and predicting future situations. Therefore, $NO_x$ can have limited influence in our forecasting model as it reflects immediate $N_2O$ levels and lacks predictive information for future concentrations. Similarly, the influence of temperature is minimal, since it represents slow and seasonal changes, making it unlikely to affect $N_2O$ emission forecasts for immediate timesteps. Thus, the primary objective of the DL model

(predict present/actual vs. future/forecast) influences the necessary model inputs for optimal performance. As the prediction horizon extends, the impact of the autoregressive input decreased, while the importance of inputs like influent flowrate increased, highlighting the need for process variables for accurate $N_2O$ forecasts. The significant drop in the autoregressive input's influence suggests that past $N_2O$ concentrations become less predictive of future concentrations over longer prediction horizons. This could be due to the changing dynamics in $N_2O$ production, driven by operational changes such as varying influent loads or adjustments to aeration, which can lead to the activation of different $N_2O$ production pathways. Thus, including process variables as inputs becomes critical for reliable forecasting of $N_2O$ emissions over longer horizons.

The best LSTM-based encoder-decoder model identified still has certain limitations. Notably, there is a significant decline in forecasting performance across the prediction horizon in the test set. This could stem from the absence of specific inputs representing the underlying biochemical processes driving $N_2O$ production. For example, the roles of free nitrous acid ($HNO_2$) during nitrifier denitrification and its inhibition of the last step of heterotrophic denitrification are not represented in the model inputs. Additionally, the model inputs lack direct information on the AOB and nitrite oxidation bacteria (NOB) activities, which could provide insights into the microbial dynamics and potential NOB washout that causes $NO_2^-$ accumulation (Gruber et al.,2021b). Other model inputs specific to the raw influent quality characteristics such as TKN, chemical oxygen demand (COD) and its fractions (slowly biodegradable COD [sbCOD] and readily biodegradable COD [rbCOD]), total suspended solids (TSS) and volatile suspended solids (VSS), could provide key information for the model to learn how different influent conditions affect the magnitude of $N_2O$ emissions. Incorporating these additional inputs would provide the DL models with more relevant data, enabling them to learn the underlying patterns and relationships between these inputs and $N_2O$ emissions. These inputs could represent potential causes of $N_2O$ emissions, allowing the model to capture complex interactions more effectively and forecast emissions with greater accuracy, ultimately improving overall model performance.

Moreover, the training dataset might still be inadequate in terms of length and data balance, lacking sufficient varied scenarios of operational conditions that influence $N_2O$ production and emission. Given the dynamic nature of factors contributing to $N_2O$ emission peaks, such as the activation of different production pathways in varying operational conditions, providing ample data samples for the DL model training is crucial for enhancing forecasting accuracy. Another limitation of the LSTM-based model proposed here is its "black-box" nature, offering limited insights into the active production pathways. This limitation could hinder the implementation of targeted control strategies or interventions aimed at mitigating specific production pathways. These challenges suggest the need for improving the DL models to effectively learn the underlying processes from data, and to achieve better model generalisation and transparency. Addressing these issues could facilitate broader application of DL models in full-scale WWTPs with $N_2O$ control strategies such as MPC.

Finally, while the DL models investigated in this study demonstrate potential as alternatives to biokinetic models for $N_2O$ forecasting, they could potentially benefit from integration with biokinetic models to form hybrid models. Some early work along these lines has already started. For example, Li et al. (2022) combined a first-principle ASM1 model with a teacher-forcing LSTM model in series, using the biokinetic model outputs as inputs to the DL model to predict $N_2O$. However, even though their results indicated improved performance of 22.5 % by the hybrid model compared to the standalone DL model, the dataset used was restricted to a rather short, 23-day monitoring campaign, i.e. the hybrid model was not tested on seasonal variations. In addition, this hybrid model has limitations in interpreting the specific active pathways responsible for $N_2O$ production. Mehrani et al. (2022) developed another hybrid model in which simulated data from a mechanistic model for a lab-scale sequencing batch reactor (SBR) was used to train various ML models. However, their study reported sub-optimal results with an ANN achieving an $R^2 = 0.67$. Therefore, there is plenty of scope and need for the development of better hybrid models for $N_2O$ emissions, possibly by investigating different hybridisation strategies (Khalil et al., 2024). Incorporating biokinetic and mechanistic information can enhance the accuracy of $N_2O$ emissions forecasts by including relevant attributes related to $N_2O$ production and emissions. Additionally, these novel models can promote greater generalisability and transparency, thereby increasing their robustness and applicability to various WWTP configurations. In this context, integrating domain knowledge biases (Cicirello et al., 2024), such as biokinetic constraints (inductive biases) and biochemical processes that fully or partially represent $N_2O$ production, into the DL model architecture and training, represents a promising avenue for future research.

## 6. Conclusions

This study developed LSTM-based deep learning (DL) models to forecast $N_2O$ emissions in the gaseous phase using data from a treatment lane of a full-scale WWTP in Amsterdam. The models were evaluated over a prediction horizon ranging from 0.5 to 6 h and were designed for integration within a model predictive control (MPC) framework to optimise key control variables, such as the dissolved oxygen (DO), with the aim to reduce $N_2O$ production and emissions. Based on the results obtained, the key findings are as follows:

- The LSTM-based encoder-decoder model architecture performed well, with the best model (256–256 LSTM units) achieving test set RMSE values of 13.6–61.1 ppm, and $R^2$ values of 0.98–0.59 for prediction horizon steps of 0.5-6.0 h ahead, respectively. Models with higher complexity outperformed the simpler ones, delivering more robust and reproducible results. Despite the promising results, the developed LSTM model exhibited reduced performance as the prediction horizon length increased.
- The best LSTM-based encoder-decoder model outperformed existing $N_2O$ emissions models published in the literature, particularly with regard to one-step-ahead predictions (0.5 h ahead).
- The feature importance analysis suggests that past $N_2O$ emissions, influent flowrate, $NH_4^+$, $NO_x$, and DO in the aerobic tank were the most significant model inputs. The importance of recent $N_2O$ emissions decreases over the prediction horizon, highlighting the increasing significance of process-related inputs.

To further improve $N_2O$ forecasting, future research should explore hybridising DL models with biokinetic models. This approach would leverage domain-specific knowledge of mechanistic and biochemical processes, alongside the strengths of data-driven modelling, to enhance both model robustness and interpretability, offering a promising avenue for more reliable $N_2O$ emissions predictions in wastewater treatment systems.

## CRediT authorship contribution statement

**Siddharth Seshan:** Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Johann Poinapen:** Writing – review & editing, Supervision, Methodology. **Marcel H. Zandvoort:** Writing – review & editing, Data curation. **Jules B. van Lier:** Writing – review & editing, Supervision, Methodology. **Zoran Kapelan:** Writing – review & editing, Supervision, Methodology.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

## Acknowledgements

We would like to thank the water authority Amstel, Gooi and Vecht, for making their plant WWTP Amsterdam West available for monitoring campaigns. Moreover, we are grateful to them for providing historical online sensor data, including the crucial $N_2O$ emission measurements, which were used for this modelling study. We would like to express our gratitude to Dr. Riccardo Taormina, of Department of Water Management, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands, for his ideas and discussions in the development of the ML model architecture. The authors acknowledge the use of computational resources of the DelftBlue supercomputer, provided by Delft High Performance Computing Centre (https://www.tudelft.nl/dhpc). This research was funded by the Fiware4Water project, which has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No. 821036.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2024.122754.

## Data availability

Data will be made available on request.

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X., 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. https://arxiv.org/abs/1603.04467v2.

Behera, C.R., Srinivasan, B., Chandran, K., Venkatasubramanian, V., 2015. Model based predictive control for energy efficient biological nitrification process with minimal nitrous oxide production. Chem. Eng. J. 268, 300–310. https://doi.org/10.1016/J.CEJ.2015.01.044.

Bellandi, G., Weijers, S., Gori, R., Nopens, I., 2020. Towards an online mitigation strategy for N2O emissions through principal components analysis and clustering techniques. J. Environ. Manage. 261, 110219. https://doi.org/10.1016/J.JENVMAN.2020.110219.

Boruah, N., Roy, B.K., 2019. Event triggered nonlinear model predictive control for a wastewater treatment plant. J. Water. Process. Eng. 32. https://doi.org/10.1016/j.jwpe.2019.100887.

Chen, G., van Loosdrecht, M.C.M., Ekama, G.A., Brdjanovic, D., 2023. Biological Wastewater Treatment: Principles, Modelling and Design. IWA Publishing, p. 269. https://doi.org/10.2166/9781789060362.

Cicirello, A., 2024. Physics-Enhanced Machine Learning: a position paper for dynamical systems investigations. https://arxiv.org/abs/2405.05987.

Daelman, M.R.J., De Baets, B., van Loosdrecht, M.C.M., Volcke, E.I.P., 2013. Influence of sampling strategies on the estimated nitrous oxide emission from wastewater treatment plants. Water. Res. 47 (9), 3120–3130. https://doi.org/10.1016/J.WATRES.2013.03.016.

Domingo-Félez, C., Smets, B.F., 2016. A consilience model to describe N2O production during biological N removal. Environ. Sci.: Water Res. Technol. 2 (6), 923–930. https://doi.org/10.1039/c6ew00179c.

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D.J., Mauritsen, T., Palmer, M.D., Watanabe, M., Wild, M., Zhang, H., 2021. The Earth's energy budget, climate feedbacks, and climate sensitivity. Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekçi, O., Yu, R., Zhou, B.. In: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S. (Eds.), Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, pp. 923–1054. https://doi.org/10.1017/9781009157896.009.

Gruber, W., Villez, K., Kipf, M., Wunderlin, P., Siegrist, H., Vogt, L., Joss, A., 2020. N2O emission in full-scale wastewater treatment: proposing a refined monitoring strategy. Sci. Total Environ. 699, 134157. https://doi.org/10.1016/J.SCITOTENV.2019.134157.

Gruber, W., von Känel, L., Vogt, L., Luck, M., Biolley, L., Feller, K., Moosmann, A., Krähenbühl, N., Kipf, M., Loosli, R., Vogel, M., Morgenroth, E., Braun, D., Joss, A., 2021a. Estimation of countrywide N2O emissions from wastewater treatment in Switzerland using long-term monitoring data. Water Res. X 13, 100122. https://doi.org/10.1016/J.WROA.2021.100122.

Gruber, W., von Känel, L., Vogt, L., Luck, M., Biolley, L., Feller, K., Moosmann, A., Krähenbühl, N., Kipf, M., Loosli, R., Vogel, M., Morgenroth, E., Braun, D., Joss, A., 2021b. Estimation of countrywide N2O emissions from wastewater treatment in Switzerland using long-term monitoring data. Water. Res. X. 13, 100122. https://doi.org/10.1016/J.WROA.2021.100122.

Guo, L., Vanrolleghem, P.A., 2013. Calibration and validation of an activated sludge model for greenhouse gases no. 1 (ASMG1): prediction of temperature-dependent N2O emission dynamics. Bioprocess. Biosyst. Eng. https://doi.org/10.1007/s00449-013-0978-3.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1739–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hwangbo, S., Al, R., Sin, G., 2020. An integrated framework for plant data-driven process modeling using deep-learning with Monte-Carlo simulations. Comput. Chem. Eng. 143, 107071. https://doi.org/10.1016/J.COMPCHEMENG.2020.107071.

Hwangbo, S., Al, R., Chen, X., Rkan Sin, G., 2021. Integrated model for understanding N2O emissions from wastewater treatment plants: a deep learning approach. Environ. Sci. Technol. 55. https://doi.org/10.1021/acs.est.0c05231.

Khalil, M., AlSayed, A., Liu, Y., Vanrolleghem, P.A., 2023. Machine learning for modeling N2O emissions from wastewater treatment plants: aligning model performance, complexity, and interpretability. Water Res. 245, 120667. https://doi.org/10.1016/J.WATRES.2023.120667.

Khalil, M., AlSayed, A., Elsayed, A., Sherif Zaghloul, M., Bell, K.Y., Al-Omari, A., Laqa Kakar, F., Houweling, D., Santoro, D., Porro, J., Elbeshbishy, E., 2024. Advances in GHG emissions modelling for WRRFs: from State-of-the-Art methods to Full-Scale applications. Chem. Eng. J. 494, 153053. https://doi.org/10.1016/J.CEJ.2024.153053.

Kosonen, H., Heinonen, M., Mikola, A., Haimi, H., Mulas, M., Corona, F., Vahala, R., 2016. Nitrous oxide production at a fully covered wastewater treatment plant: results of a long-term online monitoring campaign. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.5b04466.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth. Syst. Sci. 22 (11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.

Li, K., Duan, H., Liu, L., Qiu, R., van den Akker, B., Ni, B.J., Chen, T., Yin, H., Yuan, Z., Ye, L., 2022. An integrated first principal and deep learning approach for modeling nitrous oxide emissions from wastewater treatment plants. Environ. Sci. Technol. 56 (4), 2816–2826. https://doi.org/10.1021/acs.est.1c05020.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. ICLR 2019. https://arxiv.org/abs/1711.05101.

Mampaey, K.E., Beuckels, B., Kampschreur, M.J., Kleerebezem, R., Van Loosdrecht, M.C.M., Volcke, E.I.P., 2013. Modelling nitrous and nitric oxide emissions by autotrophic ammonia-oxidizing bacteria. Environ. Technol. 34 (12), 1555–1566. https://doi.org/10.1080/09593330.2012.758666.

Massara, T.M., Malamis, S., Guisasola, A., Antonio Baeza, J., Noutsopoulos, C., Katsou, E., 2017a. A review on nitrous oxide (N2O) emissions during biological nutrient removal from municipal wastewater and sludge reject water. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2017.03.191.

Massara, T.M., Solís, B., Guisasola, A., Katsou, E., Baeza, J.A., 2017b. Development of an ASM2d-N2O model to describe nitrous oxide emissions in municipal WWTPs under dynamic conditions. Chem. Eng. J. 335, 185–196. https://doi.org/10.1016/J.CEJ.2017.10.119.

Mehrani, M.J., Bagherzadeh, F., Zheng, M., Kowal, P., Sobotka, D., Mąkinia, J., 2022. Application of a hybrid mechanistic/machine learning model for prediction of nitrous oxide (N2O) production in a nitrifying sequencing batch reactor. Process Saf. Environ. Protect. 162, 1015–1024. https://doi.org/10.1016/j.psep.2022.04.058.

Ni, B.-J., Pan, Y., Van Den Akker, B., Ye, L., Yuan, Z., 2015. Full-scale modeling explaining large spatial variations of nitrous oxide fluxes in a step-feed plug-flow wastewater treatment reactor. Environ. Sci. Technol. 49, 42. https://doi.org/10.1021/acs.est.5b02038.

O'Brien, M., Mack, J., Lennox, B., Lovett, D., Wall, A., 2011. Model predictive control of an activated sludge process: a case study. Control Eng. Pract. 19 (1), 54–61. https://doi.org/10.1016/J.CONENGPRAC.2010.09.001.

Pan, K., Guo, T., Liao, H., Huang, Z., Li, J., 2024. Nitrous oxide emissions from aerobic granular sludge: a review. J. Clean. Prod. 434, 139990. https://doi.org/10.1016/J.JCLEPRO.2023.139990.

Rahu, M.A., Shaikh, M.M., Karim, S., Soomro, S.A., Hussain, D., Ali, S.M., 2024. Water quality monitoring and assessment for efficient water resource management through internet of things and machine learning approaches for agricultural irrigation. Water Resour. Manag. https://doi.org/10.1007/s11269-024-03899-5.

Ravishankara, A.R., Daniel, J.S., Portmann, R.W., 2009. Nitrous oxide (N2O): the dominant ozone-depleting substance emitted in the 21st century. Science (1979) 326 (5949), 123–125. https://doi.org/10.1126/science.1176985.

Ren, D., Cai, Y., Lei, X., Xu, J., Li, Q., Leung, H.fung, 2019. A multi-encoder neural conversation model. Neurocomputing. 358, 344–354. https://doi.org/10.1016/J.NEUCOM.2019.05.071.

Seshan, S., Poinapen, J., Zandvoort, M.H., van Lier, J.B., Kapelan, Z., 2024. Limitations of a biokinetic model to predict the seasonal variations of nitrous oxide emissions from a full-scale wastewater treatment plant. Sci. Total Environ. 917, 170370. https://doi.org/10.1016/J.SCITOTENV.2024.170370.

Shen, W., Chen, X., Pons, M.N., Corriou, J.P., 2009. Model predictive control for wastewater treatment process with feedforward compensation. Chem. Eng. J. 155 (1–2), 161–174. https://doi.org/10.1016/J.CEJ.2009.07.039.

Song, M.J., Choi, S., Bae, W.bin, Lee, J., Han, H., Kim, D.D., Kwon, M., Myung, J., Kim, Y. M., Yoon, S., 2020. Identification of primary effecters of N2O emissions from full-scale biological nitrogen removal systems using random forest approach. Water Res. 184, 116144. https://doi.org/10.1016/J.WATRES.2020.116144.

Sutskever, I., Vinyals, O., & Le, Q.v., 2014. Sequence to sequence learning with neural networks. https://arxiv.org/abs/1409.3215.

Vasilaki, V., Conca, V., Frison, N., Eusebi, A.L., Fatone, F., Katsou, E., 2020. A knowledge discovery framework to predict the N2O emissions in the wastewater sector. Water Res. 178, 115799. https://doi.org/10.1016/J.WATRES.2020.115799.

Vasilaki, V., Volcke, E.I.P., Nandi, A.K., van Loosdrecht, M.C.M., Katsou, E., 2018. Relating N2O emissions during biological nitrogen removal with operating conditions using multivariate statistical techniques. Water Res. 140, 387–402. https://doi.org/10.1016/J.WATRES.2018.04.052.

Xu, J., Wang, K., Lin, C., Xiao, L., Huang, X., Zhang, Y., 2021. FM-GRU: a time series prediction method for water quality based on Seq2seq framework. Water (Switzerland) 13 (8). https://doi.org/10.3390/w13081031.

Xu, X., Wei, A., Tang, S., Liu, Q., Shi, H., Sun, W., 2024. Prediction of nitrous oxide emission of a municipal wastewater treatment plant using LSTM-based deep learning models. Environ. Sci. Pollut. Res. Int. 31 (2), 2167–2186. https://doi.org/10.1007/s11356-023-31250-9.

Zhang, Q., Yang, L.T., Chen, Z., Li, P., 2018. A survey on deep learning for big data. Inf. Fusion 42, 146–157. https://doi.org/10.1016/J.INFFUS.2017.10.006.