Integrated Model for Understanding N₂O Emissions from Wastewater Treatment Plants: A Deep Learning Approach

Hwangbo, Soonho; Al, Resul; Chen, Xueming; Sin, Gürkan

Published in:
Environmental Science and Technology

Link to article, DOI:
[10.1021/acs.est.0c05231](https://doi.org/10.1021/acs.est.0c05231)

Publication date:
2021

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Hwangbo, S., Al, R., Chen, X., & Sin, G. (2021). Integrated Model for Understanding N₂O Emissions from Wastewater Treatment Plants: A Deep Learning Approach. *Environmental Science and Technology*, 55(3), 2143–2151. <https://doi.org/10.1021/acs.est.0c05231>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Integrated model for understanding N₂O emissions from wastewater treatment plants: A deep learning approach

Soonho Hwangbo,^{*,†,‡} Resul Al,[†] Xueming Chen,[†] and Gürkan Sin[†]

[†]Process and Systems Engineering Center (PROSYS), Department of Chemical and Biochemical
Engineering, Technical University of Denmark, Building 229, 2800 Kgs. Lyngby, Denmark

[‡]Department of Chemical Engineering, Gyeongsang National University, 501, Jinju-daero, Jinju-
si, Gyeongsangnam-do, South Korea

* Corresponding author

Email address: soohw@kt.dtu.dk

Abstract

This study aims to demonstrate the application of deep learning to describe quantitatively long-term full-scale data observed from wastewater treatment plants (WWTPs) from the perspectives of process modeling, process analysis, and forecasting modeling. Approximately 750,000 measurements incorporating influent flowrate, air flowrate, temperature, ammonium, nitrate, dissolved oxygen, and nitrous oxide (N₂O) during more than one year of Avedøre WWTP located in Denmark are utilized to develop deep neural network (DNN) through supervised learning for process modeling, and the optimal DNN ($R^2 > 0.90$) is selected for further evaluation. For process analysis, global sensitivity analysis based on variance decomposition is considered to identify key parameters contributing to high N₂O emission characteristics. For N₂O forecasting, the proposed DNN-based model is compared with long short-term memory (LSTM), showing that the LSTM-based forecasting model performs significantly better than the DNN-based model ($R^2 > 0.94$ and root mean squared error is reduced by 64 %). Results account for the feasibility of data-driven methods based on deep learning for quantitatively describing and understanding the rather complex N₂O dynamics in WWTPs. Research into hybrid modeling concepts integrating mechanistic models of WWTPs (e.g., ASMs) with deep learning would be suggested as a further direction for the monitoring of N₂O emissions from WWTPs.

Keywords: Deep neural network; Global sensitivity analysis; Long short-term memory; Nitrous oxide emission; Wastewater treatment plant

1. Introduction

Wastewater treatment plants (WWTPs) are an important contributor to the global greenhouse gas (GHG) emissions partly due to their significant production and emission of nitrous oxide (N_2O)^{1,2}, a potent GHG with a global warming potential 265 times stronger than carbon dioxide³. Full-scale investigations have reported dynamic N_2O emission factors, which could reach up to 14.6%⁴. Therefore, effective and efficient N_2O control for emission reduction is of imperative need in order to ensure the sustainable operation of WWTPs.

Mathematical models have been frequently employed to not only understand the N_2O production and emission mechanisms but also develop N_2O control strategies⁵⁻⁸. The development of mathematical models describing N_2O emissions based on the activated sludge model structure under dynamic conditions has been conducted⁹. Extended activated sludge model incorporating multiple N_2O production pathways and N_2O transfer in between liquid and gas phases has been developed to suggest N_2O mitigation strategies and calculate carbon footprint¹⁰. A case study of hundreds of domestic WWTPs in Greece has been applied to a mathematical model to roughly estimate the on-site and off-site GHGs including N_2O , and the future perspective on wastewater management has been discussed¹¹.

Mechanistic modeling, however, might be incapable of describing a large volume of the rather dynamic full-scale N_2O data quantitatively¹². This could be due to the incomplete consideration of the multiple N_2O producing pathways involved in biological nitrogen removal, which are typically regulated by a range of factors or the inconsistent representation of microbial parameters subject to changes in the microbial community¹³. While diverse data-driven methods to transform data into knowledge for WWTP operations have been consistently investigated^{14,15} and different variants of the application of artificial intelligence to WWTP have been examined¹⁶, a few

conventional machine learning techniques have been used for the analysis of N₂O emissions from monitoring campaign data. Multivariate statistical tools, including principal component analysis (PCA) and k-means clustering, have investigated dependencies between N₂O emissions and a multitude of influencing factors such as dissolved oxygen (DO), nitrogen concentrations, temperature and influent flowrate¹⁷. Research on applications of PCA coupled with clustering for N₂O production monitoring has been studied to provide an online N₂O mitigation strategy¹⁸. However, conventional machine learning techniques, especially PCA and k-means clustering techniques, suffer to some degree from non-stationary and nonlinear features in long-term data from WWTPs. Therefore, these studies provided limited ability to describe quantitatively long-term data dynamics. Recently, artificial neural networks for the simulation of N₂O emissions have been highlighted by demonstrating the feasibility of estimating N₂O emissions with data-driven models¹⁹. Especially, deep learning, which is one of the most powerful machine learning tools, and typically aims to learn hierarchical features and representations of big data by considering deep architectures, might play a dominant role in big data solutions to this problem²⁰. Despite state-of-the-art achievements in a broad range of deep learning applications, few studies report the feasibility of applications of deep learning for plant data related to N₂O emissions covering the entire operating period in a physical WWTP from the perspective of process systems engineering.

Therefore, the objective of this paper is to demonstrate the application of data-driven methods based on deep learning algorithms to process modeling, process analysis, and process forecasting of dynamic N₂O behaviors. First, process modeling using typical deep neural networks (DNNs) in collaboration with data smoothing and preprocessing of the long-term full-scale data from an existing WWTP (Avedøre WWTP located in Denmark) is performed to provide the suitable models of N₂O concentrations and N₂O emissions. Afterward, the DNN-based model is used to

perform a process analysis by means of global sensitivity analysis (GSA) to facilitate an understanding of N₂O emission characteristics. The GSA considered both the presence and absence of correlations among the input data to effectively reveal which process conditions are influential on N₂O emissions. Different deep learning approaches for forecasting N₂O concentrations using typical DNNs and long short-term memory (LSTM) recurrent neural networks are developed and compared with each other.

2. Materials and Methods

2.1. Description of raw data

A set of plant data used in this work originates from Avedøre WWTP located in Copenhagen, Denmark, with a daily capacity of treating around 70,000 m³ of wastewater under dry weather conditions. From the WWTP containing four parallel Carrousel reactors, each of which consists of two inter-connected compartments, dissolved oxygen (DO) [g/m³], ammonium (NH₄⁺-N) [g/m³], nitrate (NO₃⁻-N) [g/m³], temperature [°C], and N₂O concentration in the liquid phase [g/m³] are monitored online using sensors per 5-minute, while influent flowrate [m³/hour] is recorded online using a flow meter. Air flowrate [m³/hour] to each compartment is calculated through proportioning the online total airflow measurement based on the ratio of the valve opening measurement in each compartment to the sum of all valve opening measurements. Our previous study has reported specifically these measurements by carrying out the comparison of full-scale evaluation of control concepts²¹ and a more detailed explanation for the target WWTP is displayed in Section S1 in Supporting Information (SI).

Cumulative days of data collection are about 14 months (i.e., March 2018 to February 2019 and May 2019 to June 2019), and the total number of data points is 746,760. Collected data in the first

12 months are related to reactor operation by default DO setpoint (i.e., normal operating conditions), and the data in the last 2 months are produced by operating reactors with deliberately reduced DO setpoint by 0.5 mg/L for a week. Figure S1 illustrates the average, median, and maximum values of measurements per day. On the one hand, the daily pattern of temperature seems comprehensible; on the other hand, the rest of the data show diverse and irregular patterns, implying that data-driven modeling based on conventional machine learning tools would fail to reach for great model performance. For example, the configuration of air flowrate shows great differences between maximum values and other statistical values such as an average and a median since several values of air flowrate are equal to zero (i.e., the reactor in wastewater treatment has anoxic and aerobic phases). The anoxic phase corresponds to nonexistent air flowrate; on the other hand, the aerobic phase requires a large amount of air flowrate.). In the case of N₂O concentration in the liquid phase which is the target variable to be predicted in this research, there exist several hundred days consecutively incorporating average numbers of zero and small values near zero.

2.2. Deep neural network and long short-term memory recurrent neural network

Inputs consist of DO, NH₄⁺-N, NO₃⁻-N, temperature, influent flowrate, and air flowrate and the objective function aims to minimize the difference between real N₂O concentration in the liquid phase and a single value from the output layer, and it results in DNN-based model for process modeling of N₂O. The total of 6 hidden layers are considered, each of which has 16, 32, 16, 16, 4, and 4 hidden neurons respectively, and all the hidden neurons encompass the same activation function that is hyperbolic tangent sigmoid function. The main components of the DNN-based model are indicated in Table S5.

137 Additionally, LSTM, which is one of the most commonly used recurrent neural networks in
138 predictive modeling ^{22, 23} and able to deal with the considerations above such as long-term
139 dependencies, is employed to design forecasting models. Figure S4 (B) shows that LSTM network
140 is comprised of diverse inner connections that are expressed as follows; (1) forget gate F
141 determines what the previous information will be removed away from the cell, (2) when cell update
142 gate C updates the cell state, input gate I decides what information will be saved in the cell, and
143 (3) output gate O controls what information will become an appropriate output based on the cell
144 state ²⁴, all of which include linear mathematical relationship between the input X , the recurrent
145 information H , the weight of gate W and the corresponding bias b . The sigmoid function σ
146 and the hyperbolic tangent function $\tanh$ play a role in the activation function of the LSTM gates.

147 The suggested models take advantage of parallel computing resources to save computational
148 cost by combining the High Performance Computing cluster by Technical University of Denmark
149 and Parallel Computing Toolbox by MATLAB[®] version R2019a. The step-by-step algorithm of
150 this study is provided in Section S2 in SI.

152 **2.3. Process analysis**

153 **2.3.1. Sobol method for global sensitivity analysis**

154 The Sobol method constitutes the benchmark in variance decomposition-based GSA methods.
155 It relies on an assumption of independent inputs, and the method provides first order (S_i) and total
156 order (S_{Ti}) sensitivity indices for each of the input variables in a given model. The theoretical
157 formulas and interpretations of these indices are as follows. The Sobol's first-order sensitivity
158 index quantifies the individual contribution of an input variable x_i to the total output variance

159 (unconditional) $\text{Var}[y]$ using its conditional variance $\text{Var}[E(y|x_i)]$, as reads in the formula
160 below.

$$S_i = \frac{\text{Var}[E(y|x_i)]}{\text{Var}[y]} \quad (1)$$

161 On the other hand, the total order sensitivity index quantifies the total contribution of an input
162 variable x_i to the total output variance, including also the contributions made due to the
163 interactions of that input variable with the other input variables. The formula for this index reads
164 as follows:

$$S_{Ti} = 1 - \frac{\text{Var}[E(y|x_{-i})]}{\text{Var}[y]} \quad (2)$$

165 where $\text{Var}[E(y|x_{-i})]$ is the conditional variance due to all variables except x_i . The interpretation
166 of the difference between the total and the first-order index, $(S_{Ti} - S_i)$, can be made as a direct
167 measure of the strength of interactions among input variables. When the analytical computation is
168 impractical, Monte Carlo simulations are commonly employed in order to estimate the conditional
169 and the unconditional variances in Eqs (1) and (2). Several different Monte Carlo-based numerical
170 estimators for the first and total indices exist, such as Jansen, Sobol, and Saltelli estimators, for
171 which a comparative study has been presented by Saltelli, et al. ²⁵. A MATLAB-based software
172 tool, easyGSA ²⁶, provides implementations of these estimators as well as generic workflows for
173 mechanistic and machine learning model-based calculation of the first and the total order Sobol
174 sensitivity indices.

175

176 2.3.2. Kucherenko indices method for global sensitivity analysis

In an effort to generalize the Sobol indices to the case of dependent input variables, Kucherenko, et al. ²⁷, proposed to divide the input space X into two complementary subsets, X_y and X_z , and to decompose the total output variance of a model $Y = f(X)$ as follows:

$$\text{Var}[Y] = \text{Var}[E[Y | X_y]] + E[\text{Var}[Y | X_y]] \quad (3)$$

After normalizing by the total output variance $\text{Var}[Y]$, Kucherenko et al. define the first order and the total order indices for the correlated inputs as follows:

$$S_i = \frac{\text{Var}[E_z(f(X_y, \overline{X_z}))]}{\text{Var}[Y]} \quad (4)$$

$$S_{Ti} = \frac{E_z[\text{Var}[Y | f(\overline{X_y}, X_z)]]}{\text{Var}[Y]} \quad (5)$$

where X_z and $\overline{X_z}$ represent random vectors generated from a priori known joint probability density function and a priori known conditional distribution function, respectively. For the uncorrelated case, the conditional distributions are reduced to the marginal distributions; hence the formulas become equivalent to the standard Sobol indices. Such a priori knowledge of joint and conditional distributions can be inferred from a given plant dataset using a copula-based approach, examples of which are also widely encountered in fields outside engineering, such as financial risk management ²⁸. The MATLAB-based software package UQLab provides a means to readily infer such copula information from datasets and also implements the Monte Carlo-based numerical estimators for the correlated Kucherenko indices ²⁹.

3. Results and Discussion

3.1. Process modeling of N₂O using DNN-based model

Figure 1 indicates the results of the evaluation of the optimal DNN-based model for process modeling of N₂O concentration through test data ($R^2 = 0.86$). The locally zoomed figure dealing with a couple of hundred test data shows that the developed model is considerably fit for both high low N₂O concentration behaviors. Data processing including a smoothing method plays a dominant role in providing a great model performance. When building a process model using plant data from sensors located in the reactors/plants, a data smoothing method capable of smoothing high-frequency noise and making the pattern more visible is needed to ensure the stability of the model performance. Given data in this study, the results from distributions of measurements in each data variable identified a number of outliers using commonly used statistical methods such as a boxplot (Table S1). However, process engineering knowledge from the wastewater treatment domain did not necessarily agree with some of the data deemed as outliers. For example, the statistical method identifies and removes peak/high values of N₂O from the dataset, which does not make sense physically to exclude when especially the purpose of the model is to describe the entire spectrum of observed N₂O emissions from the plant. Hence, it is crucial to embrace the outliers to some extent from the point of view of modeling purpose as well as process engineering knowledge, and the application of trailing moving average as a data smoothing to smoothen peak shapes that are explicitly related to the outliers is inevitably considered (Figure S3). Moreover, exploring different perspectives of using collected and processed plant data achieves to extract useful knowledge in the DNN-based supervised learning for process modeling. In the following, the results obtained from applying the suggested methods to several N₂O emission metrics are elaborated.

Figure 1. Predicted N₂O concentration from the proposed DNN-based model for process modeling over the entire time window of test data and for a locally zoomed illustration.

3.1.1. Comparison results of DNN models for N₂O concentration considering different sizes of plant data

The above results indicate that a smoothing processing for input variables incorporating DO, $\text{NH}_4^+\text{-N}$, $\text{NO}_3^-\text{-N}$, temperature, influent flowrate, and air flowrate is of importance to enhance model performance. Therefore, the strategy is moved to compare DNN-based models for process modeling as regards cumulative data that are consecutively accumulated over the entire duration of the given plant data. Three models with 4 months (March 2018 to June 2018), 10 months (March 2018 to December 2018), and 14 months (March 2018 to February 2019 and May 2019 to June 2019) are trained by 30 iterations, and test datasets are used to identify model performance to provide the model performance metrics relating to R^2 values (Table 1).

The reason why three different lengths of cumulative data have been examined is that the collected plant data in this study can be divided into three types of time series data from the perspectives of the target variable value and the operating condition as follows. The first time series datasets are related to a great quantity of N_2O concentration in the liquid phase. The second time series datasets encompass the first time series datasets and a low amount of N_2O concentration in the liquid phase. The last time series datasets include the second time series datasets and the remaining time series datasets that are generated from due to the DO setpoint change. As seen, the model performance of proposed DNN-based models using the processed inputs seems to be well balanced regardless of the abovementioned types of datasets. It is noted that the collaboration of the developed DNN-based model, a large amount of cumulative plant data, and a proper smoothing processing for input variables makes it possible to considerably facilitate a potential problem such as overfitting in an attempt for the model to continue to be trained by dissimilar data (Figure S7 and Table S6).

Table 1. Statistical results from the validation of DNN-based models using cumulative plant data.

Model	Data	R^2 (test data)			
		Min	Max	Mean	Standard deviation
Model S1	03.2018 – 06.2018	0.81	0.86	0.85	0.0164
Model S2	03.2018 – 12.2018	0.83	0.87	0.85	0.0171
Model S3	03.2018 – 06.2019	0.81	0.86	0.84	0.0133

Model S: DNNs based on the smoothed inputs from trailing moving average.

3.1.2. Target output selection for process modeling of N₂O emission rate

N₂O emission rate modeling is involved with a mathematical model using measurements from physical sensors and general biological constants (see Section S3 in SI). Two strategies to focus on process modeling of the N₂O emission rate are examined depending on how to extract the N₂O emission rate from the above mathematical model regarding the relationship of DNN-based models as follows. The first is where N₂O emission rates are estimated from mathematical equations based on N₂O concentrations predicted from the Model S3. The second is to design new DNN-based models in an attempt to replace the existing target output (i.e., N₂O concentration in the liquid phase) with the N₂O emission rate previously calculated by mathematical equations. It identifies that the timing of calculating N₂O emission rate per day from the measured N₂O concentration plays an important role in enhancing model performance in the DNN-based regression model.

Figure 2 illustrates the results of the target output selection. Results of regression analysis for test datasets are presented in Figures 2 (A) and (B). Figure 2 (A) describes a comparison of

calculated N₂O emission rates using mathematical equations based on predicted N₂O concentrations from the Model S3 and real N₂O emission rates calculated using measured N₂O concentrations (R^2 is around 0.88). Directly predicted N₂O emission rates from the DNN model, which is trained by the calculated N₂O emission rates based on N₂O measurements, are provided in Figure 2 (B). The results demonstrate that calculated N₂O emission rates through the mathematical model using predicted N₂O concentrations from the DNN-based model can achieve relatively highly accurate model performance comparing to directly predicted N₂O emission rates based on the pre-calculated target variable. This modeling approach describes implicitly the feasibility of an integrated mathematical model and data-driven model, and it will be able to extend towards a hybrid modeling approach (e.g., N₂O emission factor also could be estimated using these results (see Section S4 in SI)).

Figure 2. Validation of model performance considering different approaches to N₂O emission rate as a target variable; Y_N2Oc represents the N₂O concentration in the liquid phase and Y_N2Oer represents the N₂O emission rate.

3.2. Understanding N₂O emission characteristic with global sensitivity analysis

In this section, we further investigate the N₂O emission characteristics of Avedøre WWTP, and thereby identify the factors that are contributing the most both to the N₂O concentration in the liquid and to the plant N₂O emission rate using global sensitivity analysis. To this end, two different GSA methods, namely Sobol indices and Kucherenko indices methods, are applied using the optimal DNN-based models found from the previously-described model selection step.

The sampling of the input space is achieved using highly space-efficient quasi-random Sobol sequences. Input factors are sampled using 50,000 samples both independently (no correlations among the input factors) and also dependently on the existing correlations inferred from the collected plant data, as shown in Figure S8. The data distribution profiles of separate factors are also extracted from the dataset and preserved during the sampling using kernel probability distributions. Figure 3 shows the results of both Sobol and Kucherenko indices for the N₂O concentration in the liquid.

The Sobol method relying on the independent sampling of the input factors identifies the temperature (T), NO₃⁻-N (NO₃) and NH₄⁺-N (NH₄) concentrations as well as influent flowrate as the most important factors contributing to the N₂O concentration in the liquid phase. As can be seen from Figure 3, while the magnitudes are slightly different, still the total Sobol indices (S_{Ti}) indicate that the same four input factors are driving the variances in their output. The most striking

outcome from Figure 3, on the other hand, is that the first-order indices (S_i) are significantly lower than the total indices (i.e., S_i is almost close to zero), and the sum of S_{Ti} is much higher than 1. This indicates that there are strong interaction effects of input parameters on the studied N_2O dynamics, which consists with the known process engineering knowledge. In other words, the joint interaction effects between temperature, NO_3^- -N, NH_4^+ -N, and influent flowrate are implied to drive the variability in N_2O dynamics in the studied plant, which is operated as alternating between aerobic and anoxic periods. This is, in fact, in agreement with our previous work ²¹, which has revealed the importance of elevated NO_3^- -N concentration at the end of aerobic cycles combined with insufficient carbon source where the following anoxic phase leads to incomplete denitrification of N_2O concentration in the liquid phase during the anoxic phase. Since N_2O is not completely denitrified during the anoxic phase, it is then stripped in the next aerobic phase, thereby contributing to the N_2O emissions from the plant. The importance of NH_4^+ -N can be explained through its central role in aerobic nitrification. In particular, NH_4^+ -N oxidizers play an important role in contributing to N_2O production during the aerobic phase through hydroxylamine oxidation and nitrifies (AOB) denitrification, both of which are influenced by NH_4^+ -N concentration in the aerobic phase ⁷. The effect of temperature can be attributed to its strong effect on both nitrifiers and denitrifiers activity; however, the exact mechanism remains to be elucidated. The effect of influent flowrate is thought to indicate not only the effect of influent loads such as nitrogen and COD but also the fact that the influent flowrate is one of the input signals used to decide the operation phase of the plant by the specific control scheme employed in the plant.

In the next step, we investigate the effect of taking into account the correlation among the inputs. The Kucherenko indices method additionally unveils the impacts made by the other process parameters through their contributions to the main process parameters abovementioned. The

results of the Kucherenko sensitivity analysis, as shown in Figure 3, present that air flowrate and DO setpoint are also among the important factors affecting the N_2O dynamics in the plant. This is to be expected from a process operation standpoint. The plant is operated using a complex control system (STAR control[®]), which determines the DO setpoint during the aerobic phases, the duration of the alternating aerobic and the anoxic phases as well as alternating influent flowrate to the compartments^{30, 31}. Therefore, the data of process variables, such as NO_3^- -N and NH_4^+ -N from such a plant, are naturally influenced by both DO setpoint during the aerobic phase as well as the duration of aerobic and anoxic phases. Moreover, air flowrate is correlated strongly not only with DO level in the aeration tanks but also with the stripping of liquid N_2O to the air, which subsequently affects the N_2O concentration in the liquid phase remaining in the tanks. Similarly to the Sobol indices, large differences are observed between the first and the total Kucherenko indices, implying that the interactional effects remain strong even when the correlations are taken into account as expected. Overall, both Sobol and Kucherenko methods successfully complement each other and confirm the importance of the key process parameters that affect the N_2O dynamics in the plant. While these findings are known from process engineering knowledge and operational experience^{12, 21}, the advantage of GSA is that it provides a quantification of the relationship. The latter is important when deciding on potential control strategies for the plant. In the case of the Avedøre plant, we have successfully tested the effect of DO setpoint in our previous study²¹; hence, the results from this study consistently and independently confirm the importance of DO setpoint on the N_2O emissions from the plant.

Figure 3. The sensitivity analysis results obtained from the optimal DNN models for the N_2O concentration in the liquid phase; the first order and the total order Sobol and Kucherenko indices are comparatively shown.

3.3. Optimal forecasting modeling of N_2O concentration based on LSTM network

A study to investigate whether the DNN-based model designed for process modeling is well conducted in terms of forecasting modeling or not has been implemented. The last one day of the entire time series data is fixed under the prediction horizon, and the accuracy of the forecasting model is examined according to an increase in the size of available historical data prior to the prediction horizon (see Section S5 in SI). Forecast N_2O concentrations from Model F3, which is determined as the optimal model among other DNN-based models (Table S2), in the suggested prediction horizon are depicted in Figure 4 (A). Despite errors between observed values and forecast values to a certain extent, the forecasting patterns are generally consistent with the patterns of real N_2O concentration. The major drawbacks to decide the optimal forecasting model based on

the suggested DNN are that 1) the process of model modification is inevitable when a prediction horizon moves back and forth and 2) the size of needed historical data for model training could be easily changed under the circumstance of manipulating prediction horizon length. To overcome the abovementioned problem of unstable performance of a general DNN for forecasting model development, LSTM network that is one of the most widely used recurrent neural networks in forecasting modeling is adapted to the given plant data. Since research on LSTM network analysis is beyond the scope of this study, the application of the LSTM network is focused on verifying model performance using the same range of a prediction horizon.

Forecast N_2O concentrations from the LSTM-based forecasting model in the prediction horizon are reported in Figure 4 (B), describing that specific peaks and flat areas, as well as holistic patterns, are forecast well. The reason is that the method used for utilizing available training datasets is systematically distinct from that used in the previous DNN-based forecasting model, and it can be explained as follows. The first is that the LSTM network requires time-series data to reflect time information during the design of a forecasting model. In doing so, meaningful historical data information can be efficiently stored and transmitted to enhance model accuracy. The second is the change of input datasets. When it comes to the aforementioned process modeling or DNN-based forecasting modeling, all measurements obtained from sensors in the reactor are delivered to a modeling processing. Otherwise, LSTM-based forecasting modeling depends on only the target measurement (i.e., N_2O concentration in this study). This plays the roles of both the inputs and the target output so that a new N_2O concentration is forecast by the current and/or previous N_2O concentrations due to the existence of time interval in between the input and the target output.

These differences state that the LSTM-based model outperforms the DNN-based model from the perspective of forecasting modeling. On the other hand, it entails that the LSTM-based model

is not suitable for process modeling that has to take into account every measurement variables. Consequently, depending on the purpose of the application of a deep learning model, the choice of a proper deep learning algorithm and appropriate data preparation/treatment thereof must be selected accordingly.

Figure 4. Forecast N₂O concentrations from (A) the DNN-based model and (B) the LSTM-based model over the fixed prediction horizon (one-day).

In addition to the results from the main applications, diverse perspectives such as model performance improvement using input combinations, comparison study to demonstrate a great

potential of the developed model in outperforming the existing models, and future outlooks of N₂O emissions are provided in Sections S6 – S8 in SI.

4. Conclusions

The objective of this study was to apply plant data-driven and deep learning-based models for process understanding of N₂O emissions in physical wastewater treatment (Avedøre WWTP in Denmark) from perspectives of explanatory and predictive modeling. Long-term full-scale plant data incorporating 746,760 measurements collected from sensors in the reactor are directly utilized in this work. The following conclusions are drawn.

- Process modeling of N₂O concentrations and emissions: Plant data-driven model coupled with DNNs for process modeling of N₂O concentrations and emission rates was developed, enabling to overcome limitations of the previously developed mechanistic models. The effects of the size of data samples, data smoothing, and data scaling on the model performance were examined. The optimal model ($R^2 > 0.9$) was able to be suggested by the application of the entire datasets based on the input combination incorporating original inputs and smoothed inputs followed by data scaling.
- Process understanding based on Monte Carlo sampling-based GSA methods: The Sobol indices results based on independent sampling revealed the temperature, NO₃⁻-N and NH₄⁺-N as the most important factors contributing to N₂O concentration and N₂O emission rate with strong interactions implied by the large differences observed between the first and total indices. Further light onto such interactions was brought by the Kucherenko indices results (based on dependent sampling), which also pinpointed air flowrate and DO setpoint as important factors, concurring with the earlier studies as well as the process operational know-how from the plant.

• N₂O forecasting modeling considering two deep learning models: Two different forecasting models, each of which is based on the DNN (i.e., having the same structure in process modeling part) and based on the LSTM network (i.e., the most commonly used recurrent neural network in the field of deep learning), were investigated and compared. The model accuracies from forecasting models indicated that the LSTM-based model could outperform the DNN-based model by decreasing 64 % of root mean square error and presenting 0.94 of R² value over the prediction horizon since it explicitly considers interactions of time series data such as long-term dependencies.

• As a future perspective, we believe there is a promising potential for exploring synergies from integrating mechanistic ASM-based plant models with deep learning in a so-called hybrid concept. In such a concept, deep learning is used for predicting related substances usually at small concentrations (such as N₂O) in the plant while the ASM model provides a mechanistic framework and mass balance assurances to describe process variables of large magnitudes such as NH₄⁺-N, NO₃⁻-N, biomass, DO, COD fractions, among others.

Acknowledgements

Soonho Hwangbo would like to thank the Danish Council for Independent Research (DFF) for financing the project under the grant ID: DFF-6111600077B. Resul Al acknowledges the funding from the EU Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 675251. Xueming Chen acknowledges the financial support by the European Union's Horizon 2020 research and innovation programme through Marie Skłodowska-Curie Individual Fellowship under grant agreement No. 790231.

- 433 1. Czepiel, P.; Crill, P.; Harriss, R., Nitrous Oxide Emissions from Municipal Wastewater
434 Treatment. *Environmental Science & Technology* **1995**, 29, (9), 2352-2356.
- 435 2. Hanaki, K.; Hong, Z.; Matsuo, T., Production of nitrous oxide gas during denitrification of
436 wastewater. *Water Science and Technology* **1992**, 26, (5-6), 1027-1036.
- 437 3. IPCC *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and*
438 *III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC,
439 Geneva, Switzerland, 2014; p 151.
- 440 4. Daelman, M. R.; van Voorthuizen, E. M.; van Dongen, U. G.; Volcke, E. I.; van
441 Loosdrecht, M. C., Seasonal and diurnal variability of N₂O emissions from a full-scale municipal
442 wastewater treatment plant. *Science of the Total Environment* **2015**, 536, 1-11.
- 443 5. Mannina, G.; Ekama, G.; Caniani, D.; Cosenza, A.; Esposito, G.; Gori, R.; Garrido-
444 Baserba, M.; Rosso, D.; Olsson, G., Greenhouse gases from wastewater treatment—A review of
445 modelling tools. *Science of the Total Environment* **2016**, 551, 254-270.
- 446 6. Ni, B. J.; Yuan, Z. G., Recent advances in mathematical modeling of nitrous oxides
447 emissions from wastewater treatment processes. *Water Research* **2015**, 87, 336-346.
- 448 7. Boiocchi, R.; Gernaey, K. V.; Sin, G., Understanding N₂O formation mechanisms through
449 sensitivity analyses using a plant-wide benchmark simulation model. *Chemical Engineering*
450 *Journal* **2017**, 317, 935-951.
- 451 8. Boiocchi, R.; Gernaey, K. V.; Sin, G., A novel fuzzy-logic control strategy minimizing
452 N₂O emissions. *Water research* **2017**, 123, 479-494.
- 453 9. Massara, T. M.; Solís, B.; Guisasola, A.; Katsou, E.; Baeza, J. A., Development of an
454 ASM2d-N₂O model to describe nitrous oxide emissions in municipal WWTPs under dynamic
455 conditions. *Chemical Engineering Journal* **2018**, 335, 185-196.
- 456 10. Zaborowska, E.; Lu, X.; Makinia, J., Strategies for mitigating nitrous oxide production and
457 decreasing the carbon footprint of a full-scale combined nitrogen and phosphorus removal
458 activated sludge system. *Water research* **2019**, 162, 53-63.
- 459 11. Koutsou, O. P.; Gatidou, G.; Stasinakis, A. S., Domestic wastewater management in
460 Greece: Greenhouse gas emissions estimation at country scale. *Journal of Cleaner Production*
461 **2018**, 188, 851-859.
- 462 12. Mampaey, K. E.; Spérandio, M.; van Loosdrecht, M. C.; Volcke, E. I., Dynamic simulation
463 of N₂O emissions from a full-scale partial nitrification reactor. *Biochemical Engineering Journal*
464 **2019**, 152, 107356.
- 465 13. Schreiber, F.; Wunderlin, P.; Udert, K. M.; Wells, G. F., Nitric oxide and nitrous oxide
466 turnover in natural and engineered microbial communities: biological pathways, chemical
467 reactions, and novel technologies. *Frontiers in microbiology* **2012**, 3, 372.
- 468 14. Corominas, L.; Garrido-Baserba, M.; Villez, K.; Olsson, G.; Cortés, U.; Poch, M.,
469 Transforming data into knowledge for improved wastewater treatment operation: A critical review
470 of techniques. *Environmental modelling & software* **2018**, 106, 89-103.
- 471 15. Dürrenmatt, D. J.; Gujer, W., Data-driven modeling approaches to support wastewater
472 treatment plant operation. *Environmental Modelling & Software* **2012**, 30, 47-56.
- 473 16. Zhao, L.; Dai, T.; Qiao, Z.; Sun, P.; Hao, J.; Yang, Y., Application of artificial intelligence
474 to wastewater treatment: A bibliometric analysis and systematic review of technology, economy,
475 management, and wastewater reuse. *Process Safety and Environmental Protection* **2019**.

17. Vasilaki, V.; Volcke, E.; Nandi, A.; Van Loosdrecht, M.; Katsou, E., Relating N₂O emissions during biological nitrogen removal with operating conditions using multivariate statistical techniques. *Water research* **2018**, *140*, 387-402.
18. Bellandi, G.; Weijers, S.; Gori, R.; Nopens, I., Towards an online mitigation strategy for N₂O emissions through principal components analysis and clustering techniques. *Journal of Environmental Management* **2020**, *261*, 110219.
19. Vasilaki, V.; Massara, T.; Stanchev, P.; Fatone, F.; Katsou, E., A decade of nitrous oxide (N₂O) monitoring in full-scale wastewater treatment processes: A critical review. *Water research* **2019**.
20. Zhang, Q.; Yang, L. T.; Chen, Z.; Li, P., A survey on deep learning for big data. *Information Fusion* **2018**, *42*, 146-157.
21. Chen, X.; Mielczarek, A. T.; Habicht, K.; Andersen, M. H.; Thornberg, D.; Sin, G., Assessment of Full-Scale N₂O Emission Characteristics and Testing of Control Concepts in an Activated Sludge Wastewater Treatment Plant with Alternating Aerobic and Anoxic Phases. *Environmental science & technology* **2019**, *53*, (21), 12485-12494.
22. Hwangbo, S.; Nam, K.; Heo, S.; Yoo, C., Hydrogen-based self-sustaining integrated renewable electricity network (HySIREN) using a supply-demand forecasting model and deep-learning algorithms. *Energy Conversion and Management* **2019**, *185*, 353-367.
23. Nam, K.; Hwangbo, S.; Yoo, C., A deep learning-based forecasting model for renewable energy scenarios to guide sustainable energy policy: A case study of Korea. *Renewable and Sustainable Energy Reviews* **2020**, *122*, 109725.
24. Yu, Y.; Si, X.; Hu, C.; Zhang, J., A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* **2019**, *31*, (7), 1235-1270.
25. Saltelli, A.; Annoni, P.; Azzini, I.; Campolongo, F.; Ratto, M.; Tarantola, S., Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications* **2010**, *181*, (2), 259-270.
26. Al, R.; Behera, C. R.; Zubov, A.; Gernaey, K. V.; Sin, G., Meta-modeling based efficient global sensitivity analysis for wastewater treatment plants—An application to the BSM2 model. *Computers & Chemical Engineering* **2019**, *127*, 233-246.
27. Kucherenko, S.; Tarantola, S.; Annoni, P., Estimation of global sensitivity indices for models with dependent variables. *Computer physics communications* **2012**, *183*, (4), 937-946.
28. Cherubini, U.; Mulinacci, S.; Gobbi, F.; Romagnoli, S., *Dynamic copula methods in finance*. John Wiley & Sons: 2011.
29. Marelli, S.; Sudret, B., UQLab: A framework for uncertainty quantification in Matlab. In *Vulnerability, uncertainty, and risk: quantification, mitigation, and management*, 2014; pp 2554-2563.
30. Nielsen, M. K.; Önnérth, T. B., Improvement of a recirculating plant by introducing STAR control. *Water Science and Technology* **1995**, *31*, (2), 171-180.
31. Nielsen, M.; Bechmann, H.; Henze, M., Modelling and test of aeration tank settling (ATS). *Water science and technology* **2000**, *41*, (9), 179-184.