Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 19 No 3, 1016 doi: 10.2166/wpt.2024.045

Exploring data quality and seasonal variations of N_2O in wastewater treatment: a modeling perspective

Laura Debel Hansen [a,b,*], Peter Alexander Stentoft[b], Daniel Ortiz-Arroyo[a] and Petar Durdevic[a]

[a] Department of Energy, Aalborg University, Esbjerg, Denmark
[b] Krüger A/S, Veolia Water Technologies, Aalborg, Denmark
*Corresponding author. E-mail: ldh@energy.aau.dk

LDH, 0000-0003-2742-2310

ABSTRACT

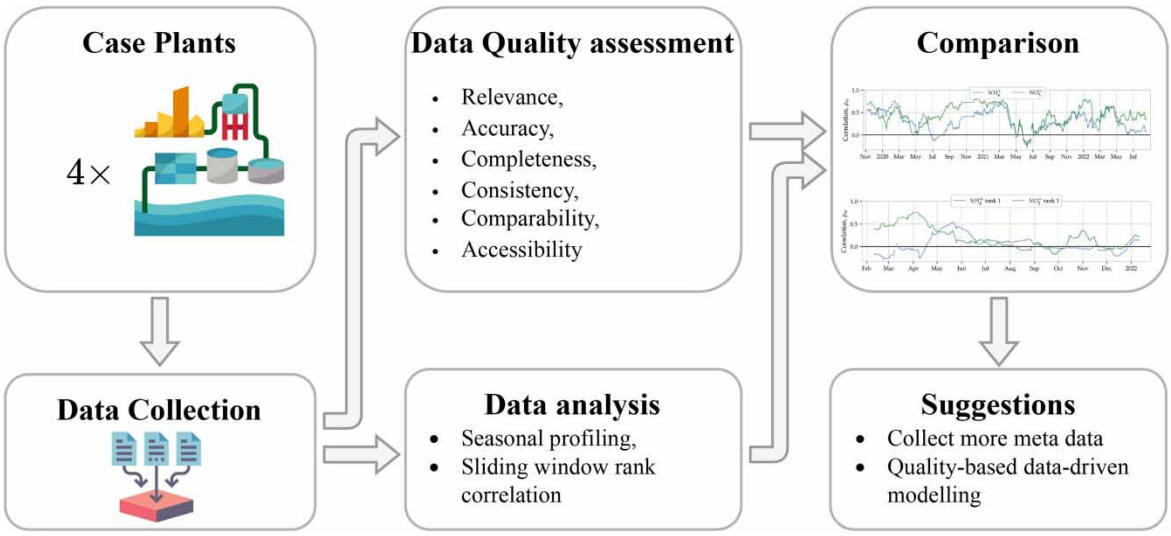
In this work, operational data collected from four Danish wastewater treatment plants (WWTP) are assessed for quality issues and analyzed to investigate the feasibility of data-driven modeling for control purposes. All plants have permanent N2O sensors installed in the biological reactors, and N_2O data are collected on the same terms as other operational data. We present and deploy a six-dimensional data quality assessment to the operational data evaluating (1) relevance, (2) accuracy, (3) completeness, (4) consistency, (5) comparability, and (6) accessibility. To increase the accuracy and completeness of the stored data, it is suggested that future initiatives are taken toward the collection and storing of metadata in WWTPs. Furthermore, seasonal variations and time-varying relationships between N_2O, nitrogenous variables, and oxygen are investigated and compared across various case plants and process designs. Results show that the quality of the operational data varies substantially between plants. The investigation of time-varying interrelation between N_2O and nitrogenous variables showed no clear pattern within or across different case plants. Furthermore, it is recommended that future research should consider adapting models so that more influence is linked to reliable measurements, contrary to assuming that all variables are of equal quality.

Key words: alternating activated sludge process, data quality assessment, exploratory data analysis, greenhouse gas emissions, heteroscedastic data, time series

HIGHLIGHTS

- The quality of operational data from full-scale WWTPs varies substantially.
- Metadata is required to ensure accurate data-driven modeling of N_2O dynamics.
- Nitrous oxide measurements show heteroscedasticity across different case plants.
- The relationship between N_2O and nitrogenous measurements is time varying.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Wastewater treatment plants (WWTPs) are known to produce substantial amounts of nitrous oxide (N_2O) (Law *et al.* 2012). With a global warming potential 265 times higher than CO_2 for a 100-year period (IPCC 2013), N_2O is a potent greenhouse gas (GHG) and one of the main contributors to climate footprint of WWTPs (Daelman *et al.* 2013; Delre & Scheutz 2019). N_2O has, therefore, gained focus in the wastewater industry and in the political debate; and the mitigation of N_2O has become a priority for WWTPs (Li *et al.* 2022).

N_2O has become the topic of several studies, where datasets of varying lengths have been collected from full-scale WWTPs and used for analysis and estimation of the N_2O emission and production (Vasilaki *et al.* 2018, 2019, 2020; Chen *et al.* 2019; Hwangbo *et al.* 2020, 2021; Myers *et al.* 2021; van Dijk *et al.* 2021; Li *et al.* 2022; Valk *et al.* 2022). Many studies apply data-driven modeling to model N_2O concentrations or emissions (Hwangbo *et al.* 2020, 2021; Vasilaki *et al.* 2020; Li *et al.* 2022); or exploratory data analysis of N_2O and other operational variables to investigate the relationship between those (Vasilaki *et al.* 2018; Chen *et al.* 2019; Myers *et al.* 2021; van Dijk *et al.* 2021). A monitoring campaign is usually conducted, where operational data are logged only within a predefined period. However, many WWTPs monitor the biochemical processes with several sensors and appertaining surveillance systems to ensure stable operation and minimize the risk of exceeding effluent standards and polluting surface waters. This practice generates tremendous amounts of operational data, which are stored and can be used for research purposes. Furthermore, as political initiatives are being notified (Danish Government 2020), many utilities (especially in Denmark) are installing N_2O sensors to quantify their climate footprint better and to prepare for N_2O production control. This enhances the prospects for data-driven modeling, which in various studies has shown great results for N_2O modeling (Vasilaki *et al.* 2020; Hwangbo *et al.* 2021; Li *et al.* 2022) and modeling of other variables in the activated sludge process (ASP) (Hansen *et al.* 2022). However, one major challenge regarding data-driven modeling is data quality. The quality of the data used to train and validate mathematical models can greatly affect the performance, since erroneous data can lead to inaccurate or unreliable models (Breck *et al.* 2019; Budach *et al.* 2022).

This study aims to build and support the foundation of data-driven model development for N_2O prediction and estimation. To this end, we investigate the feasibility of data-driven modeling using *only* operational data from four different treatment plants with various process designs. Consequently, the characteristics of the data collected align with that used to analyze or determine control actions for the plant. The acquired data are subject to a quality assessment where challenges associated with the operational data are identified, while examples from the real data are presented. Four quality issues are quantified, and the data quality across different case plants is compared. In addition, we also present a comparison of data quality across different sensor types used in the daily operation. The data from all four cases are profiled and investigated for seasonal variations, and the results for similar plant designs are compared. Furthermore, we investigate the relationship between N_2O and nitrogenous variables by means of a sliding window rank correlation. All statistical methods used are applied as sliding window operations, hence enabling a better understanding of the N_2O dynamics under changing conditions. The methods applied in this work differ from previously published literature, where statistical techniques were applied to entire datasets (Chen *et al.* 2019) or larger segments (Vasilaki *et al.* 2018). In addition, this study includes four case studies, where all biological treatment is done using the alternating ASP. As such, the main contributions of this study are the following:

(1) Using operational data from four Danish WWTPs, which combined provide over 8 years of N_2O measurements, we quantify the data quality and, thus, feasibility of data-driven methods for N_2O modeling in the future.
(2) We analyze and present the N_2O profile for the case plants and compare results across similar plant designs.
(3) To investigate the N_2O dynamics over time and over changing environmental conditions; we apply the sliding window rank correlation, which highlights the changing relationship between N_2O and nitrogenous variables. The outcomes of this analysis are evaluated within the context of the data quality investigation, considering potential significant interactions.

2. MATERIALS AND METHODS

2.1. Description of data

2.1.1. N_2O measurements

Dissolved N_2O concentrations were measured using the liquid-phase Clark-type electro-chemical N_2O sensor from Unisense Environment, Denmark (Unisense Environment A/S 2022). The sensor has a replaceable

sensor head for which the manufacturer guarantees a lifetime of 4 months from the date of receipt but expects a lifetime of +6 months. Before deploying the sensor, it is calibrated using a two-point calibration at the same temperature as the wastewater in which the sensor will be placed in. The measurements will then be valid with an uncertainty of ±5% within ±3°C of the calibration temperature; hence, it is recommended to perform a two-point calibration every 2 months to comply with seasonal variations in the wastewater temperature (Unisense Environment A/S 2022). In its default configuration, the sensor is designed to measure within the range of 0–1.5 mg N_2O-N/L with a resolution of 0.005 mg N_2O-N/L. Nevertheless, there is an option to calibrate the sensor to accommodate a working range tailored to the anticipated N_2O concentration at a particular plant. It is important to note that calibrating the sensor to a nonstandard range will necessitate a trade-off with resolution, leading to increased increments in sensor values. The response time of the sensor is 65 s for temperatures between 10 and 30°C. At lower temperatures, the sensor response will be slower. Information about the sensor is acquired from the sensor manual and through personal communication with the manufacturer.

Examples that demonstrate the hourly and daily variations are presented in Figure 1. The two examples are from different plants (Avedøre and Fredericia, respectively), and they clearly show how the dynamics of the N_2O production can change from one day to the other. Furthermore, Figure 1 shows the variation of cycle lengths and number of cycles per day. Here, a cycle refers to the period in which the concentration is near-constant close to zero before and after a period with varying concentration.

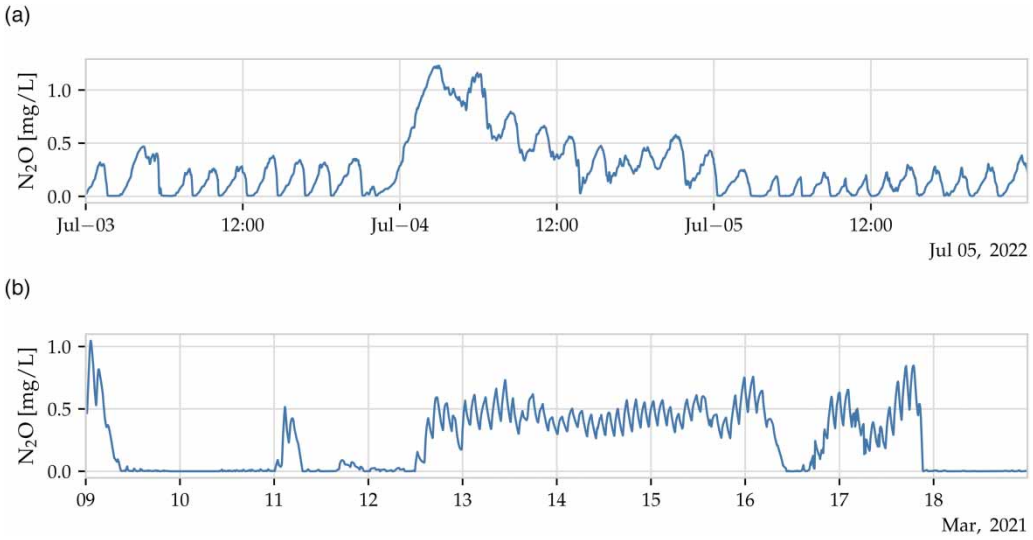


Figure 1 | Examples of variations in N_2O concentrations over arbitrary periods of (a) 3 days, Avedøre WWTP, and (b) 10 days, Fredericia WWTP.

2.1.2. Plant descriptions

This work is based on operational data collected from four Danish WWTPs from 2018 to early 2023. All plants involved in this study have permanent N_2O sensors installed in the biological reactors and collect N_2O data on the same terms as other measured operational variables (dissolved oxygen, nitrogen component concentrations, flow rates, temperature, etc.). Due to the warned political initiatives (Danish Government 2020), more Danish WWTPs are installing N_2O sensors for permanent use as part of the operation surveillance. However, few plants have collected data from 2022 or earlier, meaning that datasets of at least 1 year are difficult to obtain. The four datasets are of various sizes depending on the time of installation of the N_2O sensor(s) and the number of measured operational variables. Table 1 provides an overview of the individual dataset characteristics, where the following are specified: sample time (TS), temporal length of the dataset, the period in which the data were collected, number of installed N_2O sensors, plant design, and the population equivalent (PE) capacity.

As specified in Table 1, the plants differ in design and capacity. However, common for all the case plants included in this study is that they all utilize biological nutrient removal using the alternating ASP (Zhao *et al.* 1995), which is illustrated in Figure 2. As shown in the figure, first a primary settling is performed followed by a biological treatment before the secondary settling from where sludge is returned to the combined biological

Table 1 | Characteristics of the available data

	Dataset characteristics			Plant characteristics			
	Resolution	**Length**	**Period**	**Plant design**	**N_2O sensors**	**Lines**	**Capacity (PE)**
Avedøre	2 min	4 years	2018–2022	Biodenitro (Bundgaard *et al.* 1989)	4	6	350,000
Fredericia	5 min	3 years	2019–2022	Single tanks	1	4	420,000
Køge	2 min	10 months	2022	Biodenitro (Bundgaard *et al.* 1989)	2	3	100,000
Skanderborg	5 min	1 year	2021–2022	Single tanks	2	2	41,500

Note: All plants run the biological treatment using alternating ASP.

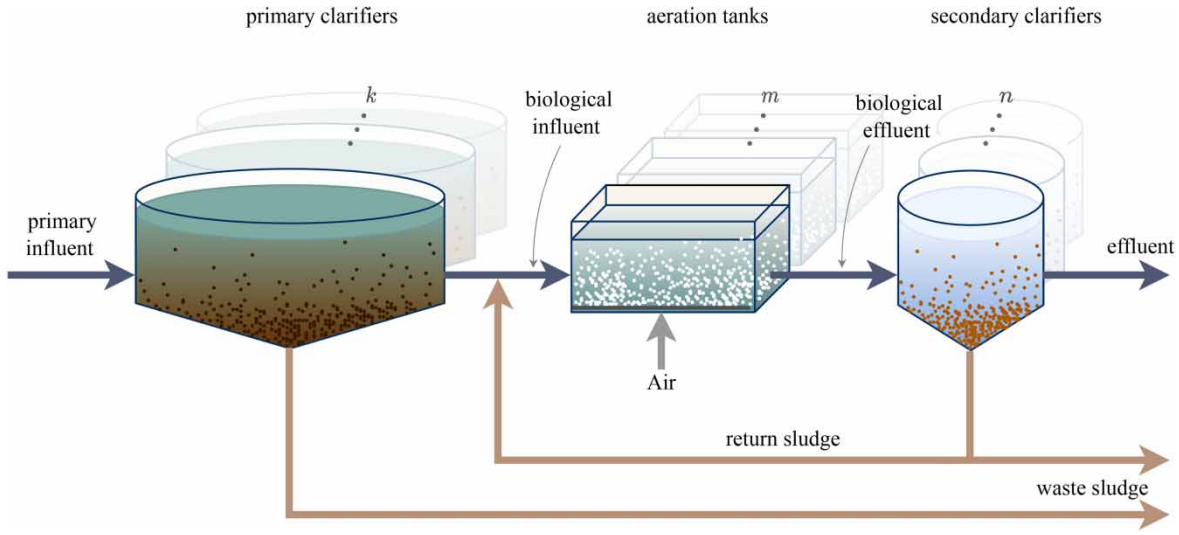


Figure 2 | Diagram of a general alternating ASP with single aeration tanks where *k*, *m*, and *n* vary from plant to plant depending on the wastewater influent characteristics.

influent. Depending on the capacity, the number of primary clarifiers, aeration tanks, and secondary clarifiers varies from plant to plant. Figure 2 shows an alternating ASP, which is used in several of the case plants. The other design presented in this work is the Biodenitro™ (Bundgaard *et al.* 1989). Further information describing the plant design and placement of sensors is presented in the Supplementary Material Section S1.

2.2. Data quality assessment and quantification

High-quality data are a precondition for analysis, development, and implementation of data-driven methods in any framework (Cai & Zhu 2015). Although several studies concern data-driven modeling of wastewater treatment processes (Ni *et al.* 2013; Chen *et al.* 2019; Hwangbo *et al.* 2020, 2021; Li *et al.* 2022; Valk *et al.* 2022), none address the quality of the collected data.

The definition of data quality depends on the perspective of, e.g., data users, producers, and custodians and the different use contexts (Fürber 2015). A very general definition could be *'data is of high quality if it is fit for use by data consumers'* (Wang 1996). Now the question arises, of whether the data are fit for use, and to answer that question, one may consider various different dimensions. Wang (1996) identified 15 important dimensions of data quality; however, several modern publications summarize the important aspects into fewer dimensions (Herzog *et al.* 2007; Cai & Zhu 2015; Fleckenstein & Fellows 2018), some with several elements in each dimension. In this study, the data quality was evaluated in six dimensions as inspired by several existing formulations (Wang 1996; Herzog *et al.* 2007; Cai & Zhu 2015; Fleckenstein & Fellows 2018), and those are the following:

- Relevance
- Accuracy
- Completeness
- Consistency, coherence, and clarity
- Comparability
- Accessibility

These dimensions were examined to evaluate the quality of operational data and its applicability in data-driven analysis and modeling. Challenges specific to WWTP data were identified and elucidated through examples drawn from operational data, with certain issues quantified. The evaluated variables encompass N_2O (nitrous oxide), NH_4 (ammonium), NO_3 (nitrate), DO (dissolved oxygen), temperature, influent flow rate, and airflow to the biological process. Note that certain challenges can be associated with multiple dimensions of data quality, highlighting the complexity of data quality assessment. In the following, we will define the six dimensions briefly. However, the full theoretical background is presented in Section S2 in the Supplementary Material along with visual examples of data quality issues from the four case plants.

The relevance of the data refers to which extent the data is applicable and useful for the intended task (Herzog *et al.* 2007). Given that we aim to determine the amount and quality of information possible to extract from operational data in this study, the relevance of the collected datasets is high. The relevance is a data quality dimension that is contextual; hence, a given dataset might be less relevant in different use-cases, for instance, due to missing variables.

Accuracy is a data quality dimension, which may partly rely on instinct (Wang 1996; Fürber 2015), but can also be assessed scientifically by comparing the data to a reference. However, when dealing with operational data, a reference measurement is rarely available since the operators are interested in low maintenance and operational costs, which avoids performing multiple measurements of the same variable when one is sufficient. Instead, a more accessible evaluation of data accuracy can be provided by the sensor manufacturers. Accuracy for the chemical sensors measuring, i.e., DO NH_4, NO_3, etc., is often provided with a percent-wise deviation from the true value. Other issues related to accuracy is the tendency of sensor drift or bias in the measurements. Both phenomena are visualized in Section S2 in the Supplementary material. Visual inspections of all chemical variables were performed using time series plots. Chemical variables that realistically cannot become negative involve N_2O, NH_4, NO_3, and DO. Hence, when encountering negative values for those variables, it indicates bias, drift, or another sensor fault. Similarly, the measurements for N_2O and DO are known to have cycles that drop to 0, meaning that drift can be clearly observed if the minimum of DO and N_2O never reaches 0. The number of variables that are identified to have drift or bias were identified. In addition, a more general quantification of negative values in the data was performed by calculating the number of negative values over the entire dataset and comparing it to the total number of measurements.

Completeness is tied to the SCADA system and operator philosophy. Incompleteness can arise from missing temporal data, excluded variables, or missing metadata (such as undocumented units). Incomplete data introduce uncertainties, often addressed by replacing missing values with the last available measurement. This practice, though, may create a misleading impression of stability in dynamic periods, emphasizing the need for comprehensive data practices and collection of metadata. An evaluation of completeness was performed by detecting missing and constant values in operational variables. Variables with missing or constant values for more than 30 min, were identified as having bad quality for the given period. Variables such as N_2O, DO, airflow, and flow of influent frequently measure 0 (mg/L or m^3/h), and periods longer than 30 min with a value of 0 were, hence, not identified as inaccurate for those specific variables. The accumulated duration of periods with incomplete data was compared to the total length of each dataset, providing a percentage of low-quality data in the dataset.

Consistency is a representational dimension of data quality, which refers to consistent representation and interpretation of data (Wang 1996; Fleckenstein & Fellows 2018). Changing, e.g., the sensor type, frequency of the measurements, sensor range, or unit of the measurement induces poor data quality in relation to consistency. Consistency can, despite its representational aspect, be difficult to quantify without metadata to support the questionable periods of data identified through process understanding and expert knowledge. More information and examples are available in Section S2 in the Supplementary material. Similar to bias and drift, the datasets were investigated visually for consistency issues. More specifically, the variables were inspected for changes in sensor ranges. Due to the lack of metadata (which is also related to completeness), the changes in sensor characteristics are difficult to quantify. Yet, one simple method is to examine visually whether the maximum sensor range changes over the dataset.

Comparability relates to the extent to which the dataset can be compared or perhaps merged with other data. Data used in this work show high comparability as all datasets are (1) from a real plant, (2) with a similar sampling rate (2 or 5 min), and (3) with analogous variables collected in the view of Danish standards. The

plant designs differ from each other, but for the purpose of data-driven modeling and exploratory analysis, this aspect increases the data quality as the aim is to cover several different plant designs and compare results.

Data availability involves challenges in collecting, storing, and retrieving data. High availability ensures quick access, but latency can impact data usefulness. Timeliness considers the appropriateness of data age, which, for offline methods (as done in this study), is usually not an issue unless processes change post-collection. Metadata is the most useful tool to determine the availability of the data; however, statistical and time series analysis may provide insight as well. Through personal communication with plant operators, the authors of this work have been informed that the data resolution is reduced in some systems. This is commonly done by down-sampling the time series or aggregation into periodic averages, and those methods are ways of lossy data compression. For data-driven modeling and system dynamics investigation, lossy compression may remove critical information, compromising data quality.

Quantification was performed on the dimensions of accuracy, completeness, and consistency, as these dimensions have aspects that are concrete, as opposed to the abstract and instinctive nature of for instance relevance.

2.3. Preliminary data processing

After quantifying some of the most important data quality issues, the data are pre-processed for further data analyses. The data retrieved from the SCADA system or online control platform have been already synchronized under the same time stamp, similar to the approach followed in other studies (Vasilaki *et al.* 2018; Chen *et al.* 2019). Due to daylight saving time (DST), starting in March and ending in October, the Danish local time is UTC+01 during the winter half-year and UTC+02 during the summer half-year. The data are, therefore, processed to be presented in UTC+01 throughout this work. The data are retrieved in the highest resolution possible, given the challenges related to availability in Subsection S2.6. To retain any information of system dynamics represented in the SCADA data, it is, hence, *not* averaged over a time period as is the case in other studies, where data were aggregated into 5-min averages (Chen *et al.* 2019; Myers *et al.* 2021) and hourly averages (Vasilaki *et al.* 2018). Missing data and samples that were given values out of the given sensor range were identified as erroneous and replaced with the last available value, hence inducing another data quality problem: completeness. Sensors subject to drift and bias are the nutrient-measuring sensors. Hence, these are investigated for signs of drift and bias and processed by manual visual inspection. By plotting the data like shown in Figure S2(a), the sensors with obvious drift or bias can be identified. As the alternating ASP often entails zero DO concentration and N_2O concentration, those two variables were straightforward to diagnose for drift or bias. If diagnosed to be inaccurate, the data X is divided into N smaller batches of fitting length and corrected using Equation (1),

$$X_{i,\text{corrected}} = X_i - \text{median}\,(\min^{\text{n}}\,(\text{X}_\text{i})), \quad \text{for i} = \{0, 1, \ldots, \text{N}\}, \tag{1}$$

in which X_i denotes the ith batch and $\min^n(X_i)$ is the n smallest values in the batch, with $n = 3$ in this work.

The DO concentration is effectively controlled by adjusting the airflow, and there are approximately 10–20 cycles per day. The N_2O concentration is not controlled; thus, the number of cycles per day varies in the range of 0.25–20 (as shown in Figure 1). The length of the batches in this study is 4 days to ensure there is at least one cycle per batch.

The N_2O and DO are measurements that repeatedly take values of zero; hence, the minimum concentration over a period that includes several cycles is always known. These measurements are, therefore, simple to correct for bias and drift. On the other hand, it is more difficult to correct variables such as NH_4, NO_3 and PO_4, as these measurements do not necessarily fall to zero over several cycles.

2.4. Data analyses

2.4.1. Seasonal variations

The first step after data preprocessing is to determine whether trend or seasonal patterns are present (Chatfield 2003). Understanding the profile and seasonal variations of N_2O within a specific plant is crucial for drawing meaningful conclusions and contextualizing results when conducting inter plant comparisons. Extrapolating findings related to N_2O from one plant to another may not be straightforward, particularly when the profiles exhibit significant differences. Furthermore, seasonal analysis may reveal issues such as data outliers or inconsistencies that can impact the overall quality of the time series data.

The seasonal variations of N_2O were investigated using a centered moving mean, μ_w, and standard deviation, σ_w, over a window w of 30 days. Comparison of the N_2O seasonal variations to the nitrogenous variables is performed using a scaled-centered moving mean and moving standard deviation of the compared variables. Scaling, using the min-max-scaling, enables comparison on a common axis as NH_4 and NO_3 endure up to 10 times higher concentrations (in mg/L) compared to N_2O.

2.4.2. Rank correlation

To explore the relationship between N_2O and the additional measurements, the sliding window rank correlation, ρ_w^i, as given in Equation (2) was investigated. Over a window of $w = 7$ days, the rank correlation coefficient ρ_w^i between N_2O (Y) and other nitrogenous measurements (X) was calculated.

$$\rho_w^i = \frac{E[(R(X_i) - \mu_{R(X_i)})(R(Y_i) - \mu_{R(Y_i)})]}{\sigma_{R(X_i)}\sigma_{R(Y_i)}}, \quad \text{for i} = \frac{\text{w}}{2}, \ldots, \text{N} - \frac{\text{w}}{2}, \tag{2}$$

in which X and Y are variables of the N-length dataset with subsets: $X_i = [x_{i-\frac{w}{2}}, \ldots, x_i, \ldots, x_{i+\frac{w}{2}}]$, $Y_i = [y_{i-\frac{w}{2}} \ldots y_i \ldots y_{i+\frac{w}{2}}]$. The operation $R()$ designates the ranking of variables, while μ and σ denote the mean and standard deviation of the subsets, respectively.

As opposed to the linear correlation coefficient, the rank correlation does not assume normally distributed variables, and the method is a measure of the monotonicity of the relationship between two variables. P-values lower than 0.01 were considered significant.

3. RESULTS AND DISCUSSION

3.1. Quality assessment of data

In accordance with Section 2.2, the quality of the data was assessed and quantified before the preliminary data processing and was used to evaluate the condition of wastewater data retrieved directly from SCADA systems or an online control platform. Quantification was performed on the dimensions of accuracy, completeness, and consistency, as these dimensions have aspects that can be quantified, as opposed to the abstract and instinctive nature of for instance relevance. Results are presented in Figure 3, Table 2, and further elaborated in the Supplementary material in Table S1 and S2, indicating that the quality of data varies across the four different plants and also across sensor types.

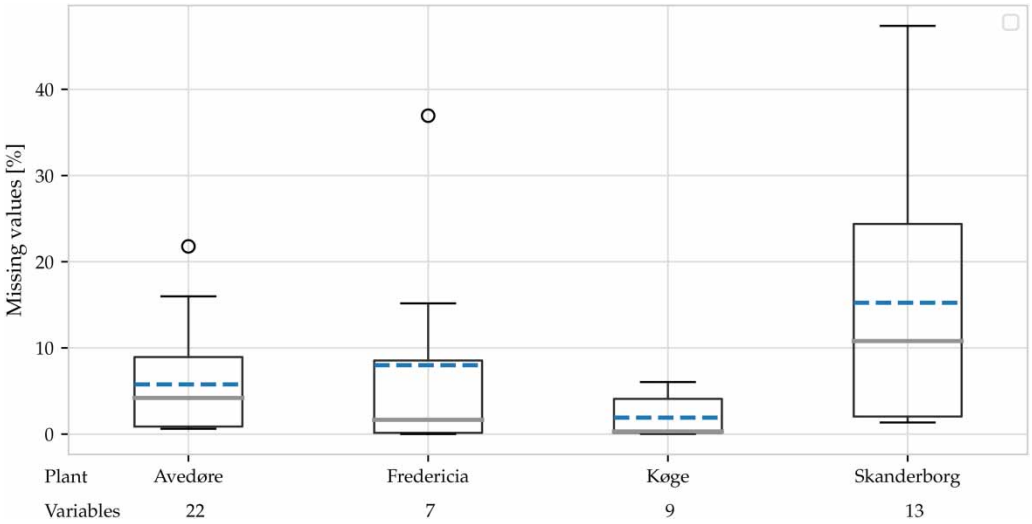


Figure 3 | Missing data grouped by plant. The gray solid line indicates the sample median while the blue dashed line indicate the sample mean.

In Table 2, the total number of investigated variables is shown alongside the number of alternating sensors (number of variables, where the sensor changes characteristics) and bias and drift in sensors grouped by N_2O and other sensors. Furthermore, the table also presents how many variables are affected by missing data and negative data, including the mean, minimum, and maximum percentage across all variables in every plant.

Table 2 | Quantification of quality issues across case plants

			Bias/drift[b]		Missing values				Negative values			
	Total number of sensors	Reconfigured sensors[a]	N_2O	Other	#[c]	Mean	Min	Max	#[d]	Mean	Min	Max
Avedøre	22	2	4	4	22	5.8	0.6	21.8	9	15.1	0.01	32.3
Fredericia	7	3	1	1	7	8.0	0.01	36.9	0	–	–	–
Køge	9	2	1	2	9	1.9	0.03	6.0	0	–	–	–
Skanderborg	13	–	2	–	13	15.2	1.4	47.4	2	40.2	33.2	47.2

Note: Mean, minimum, and maximum values are all given in percentages (%).
[a]Number of sensors subject to reconfiguration.
[b]Number of sensors subject to bias or drift grouped by N_2O and non-N_2O sensors.
[c]Number of sensors with missing values.
[d]Number of sensors with negative values.

The number of variables in each dataset varies between 7 (Fredericia) and 22 (Avedøre) variables. In spite of this, the number of sensors that have changing characteristics throughout the dataset is consistent over the datasets. The results suggest that this kind of quality issue does not depend on the size of the dataset. It was observed that the NO_3 sensor ranges often are changed to cover a larger span of concentrations as shown in Figure S5. Similarly, the N_2O sensor ranges have been observed to be adjusted after a period. This adjustment of both N_2O and NO_3 sensors suggests that the operators are not prepared for the high concentrations of nitrogenous substances.

The drift/bias investigation targeted nitrous oxide and dissolved oxygen sensors, revealing that 15 out of the 18 assessed variables exhibited signs of drift or bias. The findings suggest a correlation between operator attentiveness and sensor calibration frequency. Table 3 provides the minimum and maximum drift values for each sensor, expressed in mg/L. The outcomes underscore the presence of drift and/or bias in WWTP data, with a notable impact on 8 out of 15 affected sensors being N_2O sensors. This drift may arise from inadequate maintenance, marked by the non-replacement of sensor heads, or significant changes in wastewater temperature exceeding

Table 3 | Quantification of bias/drift across case plants

Plant	Sensor	Minimum (mg/L)	Maximum (mg/L)
Avedøre line 1	N_2O tank 1	−0.035	0.032
	N_2O tank 2	−0.035	0.013
	DO tank 1	−0.033	0.384
	DO tank 2	−0.031	0.03
Avedøre line 3	N_2O tank 1	−0.024	0.034
	N_2O tank 2	−0.024	0.054
	DO tank 1	−0.038	0.304
	DO tank 2	−0.039	0.319
Fredericia	N_2O	0	0.013
	DO	0	0.01
Køge	N_2O tank 1	0	0
	N_2O tank 2	0	0.01
	DO tank 1	0	0.066
	DO tank 2	0	0.069
Skanderborg	N_2O tank 1	−0.333	0.004
	N_2O tank 2	−0.066	0.038
	DO tank 1	0	0
	DO tank 2	0	0

Notes: Minimum and maximum values are given in mg/L. A visual representation of the analysis is provided in Figure S2(b).

3°C since the last calibration. In Skanderborg, the N_2O sensor exhibits a substantial negative drift with a magnitude of 0.33 mg/L, likely attributed to a lack of calibration or sensor head replacement. In other plants, the observed drift/bias is comparatively more modest, reaching a maximum magnitude of 0.05 mg/L. While still significant, these deviations necessitate vigilant monitoring and potential corrective action.

All variables were investigated for missing data and incorrect negative values. For each dataset, the results in Table 2 show the number of variables affected along with the mean and min/max of percent-wise missing or negative values. Figure 3 illustrates the variation in data accuracy across plants, and the results support the finding that the general quality of data varies across plants. Note that for every case plant, all the signals/variables exhibit missing values, as the missing values (#) is the same as the total number of variables investigated.

Regarding accuracy, the mean percentage of missing data in all variables varies from 1.91 % for Køge to an alarming 15.24% for Skanderborg. In the datasets from Avedøre, Fredericia, and Skanderborg, we see high variations with differences between minimum and maximum above 20%. The missing data grouped by sensor type were likewise investigated, and the results are presented in Table S1. The highest percentage of missing values was for each of the plants – Avedøre, Fredericia, Køge, and Skanderborg – found in the variables N_2O, NH_4, and temperature.

Sensor maintenance and calibration are typical reasons for the incompleteness of data. Those two causes are often related to long-term monitoring where short monitoring campaigns of a few months do not have these issues to the same extent. Datasets based on monitoring campaigns only containing a few months of data may, hence, surpass the full-scale permanently monitored data on the dimension of completeness. Specifically for the N_2O sensor, there are several possible reasons for missing data, including (i) broken sensor membrane, (ii) water intrusion, and (iii) electrical fault (e.g., loose wiring) (Unisense Environment A/S 2022).

The four datasets include a total of 51 variables, for which 11 variables exhibited negative values, corresponding to 22%. The results presented in Table 2 show a variation across variables within the same plant, ranging from 0.01 to 32.3% for the Avedøre variables and from 33.2 to 47.2% for Skanderborg. Table S2 shows the distribution of erroneous negative values across different sensors. In the Skanderborg dataset, the percent-wise erroneous negative data are over a third of the entire dataset. This is alarming, especially because the two variables affected are N_2O measurements. Of the 11 variables experiencing incorrect negative values, the most common faulty variable was N_2O, representing 6 of those 11 affected variables.

These results shed light on an issue of high importance when data-driven modeling is applied to wastewater processes: that is, to which extend the operational data can be trusted, and thus following the accuracy of the model, when trained on similar operational data. Evidently, the findings presented here demonstrate that each case is unique, and model development must be adapted to the specific plant and appertaining dataset. Future research should consider adapting models so that more influence is linked to reliable measurements, rather than just assuming that all variables are of equal quality.

Five issues were identified with respect to data quality; however, there are more data quality issues that should be clarified to get the full extent and thorough quantification of data quality. The problem is that the datasets obtained from municipal WWTPs lack metadata. In the datasets presented in this study, none of the plants register or store metadata such as time of calibration, maintenance of sensors, and replacement of sensors. It has been noted through interviews with plant operators that some of the case plants involved in this study have had interruptions in the daily operation for various reasons. This is to be expected when operational data covering several years are collected, but the problem is the lack of metadata. To increase the accuracy and completeness of the stored data, it is thus suggested that future initiatives are taken toward the collection and storing of metadata in WWTPs.

3.2. Seasonal variations

After data cleaning, the seasonal variations were investigated across all case plants. To this end, the centered moving mean and standard deviation of the N_2O measurements were computed and illustrated in Figure 4. The centered window mean, μ_w, and standard deviation, σ_w, were calculated over a window, w, of 30 days.

From Figure 4, it is evident that the N_2O is a heteroscedastic variable in all the plants. For the Fredericia case presented in Figure 4(a), the process characteristics differ from year to year, and there is a pattern showing that N_2O production is highest in winter and spring around January to April. Fredericia and Skanderborg have similar plant designs with single circular tanks where nitrification and denitrification happen alternately. Nevertheless, when comparing the profiles of Skanderborg (Figure 4(b)) to Fredericia (Figure 4(a)),

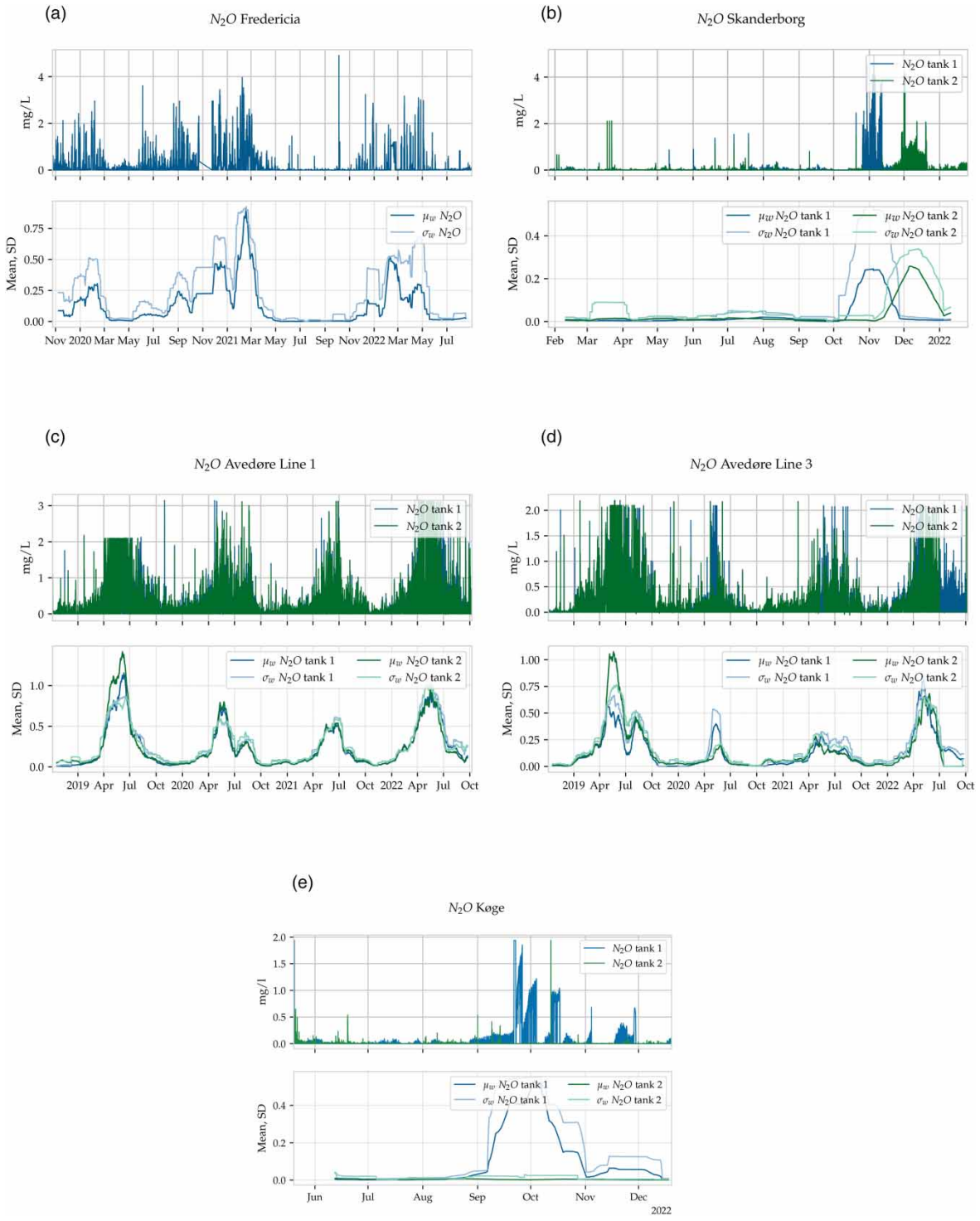


Figure 4 | Time series plot for comparison with moving mean and moving standard deviation over a window $w = 30$ days at (a) Fredericia, (b) Skanderborg WWTP, (c) Avedøre biological treatment line 1, (d) Avedøre biological treatment line 3, and (e) Køge.

we see different characteristics. The N_2O production in Skanderborg WWTP peaks in November and December as opposed to the Fredericia case. Through personal communication, the authors were informed that the Skanderborg plant underwent some challenging periods with limited resources for daily operation, resulting in decreased maintenance of sensors. This may be one explanation for missing data and drift in the N_2O sensor, but it may also explain why Skanderborg measures the highest concentrations of N_2O in the winter. The N_2O measurements at Skanderborg are only monitored for a year, but the data still provide valuable information as a frame of reference.

In Figure 4(c) and 4(d), the N_2O measurements from Avedøre biological treatment line 1 and 3 are presented. Similar to the data from Fredericia, the N_2O has seasonal patterns that repeat over the years. The N_2O measurements at Avedøre are highest in late spring and summer, peaking around April to June.

The N_2O profile of Køge WWTP is presented in Figure 4(e), where we see very different profiles for the two interconnected reactors. In tank 2, the N_2O concentration rarely exceeds 0.1 mg/L, whereas tank 1 generally has higher concentrations. Køge WWTP and Avedøre WWTP both utilize the Biodenitro design. Similar to the Skanderborg–Fredericia comparison, the two Biodenitro plants do not exhibit similar N_2O profiles. The N_2O at Avedøre seems to correlate across all lines and tanks, which is not the case at Køge WWTP. Furthermore, the concentrations at Køge WWTP peak around September to October, which does not match the profile of Avedøre WWTP.

The seasonal analysis of N_2O underlines the critical need for high-quality data. In line with the combined work of others (Kosonen *et al.* 2016; Vasilaki *et al.* 2018; Chen *et al.* 2019), the results identify substantial variations over time and over different plants, emphasizing that inaccurate data significantly influence the interpretation of observed plants. Considering the reported 5% uncertainty in the N_2O sensor and the substantial contribution of drift and missing data to the measurement characteristics, addressing these issues is imperative. This is crucial as it can impact the relationships identified between various process measurements. These results point to the need for more data, as seasonal variations are, evidently, not represented in datasets shorter than 1 year.

Another point that can be drawn from these results is that different plants require different models. This direct comparison, of periods longer than 8 months, emphasizes the need for case-specific models, developed solely to predict and estimate the processes that are represented in the dataset the model was trained on.

3.3. Time-varying rank correlation

The implementation of the sliding window rank correlation presented in Equation (2) is illustrated in Figures 5–8 with window $w = 30$ days. The method could not identify consistent correlations between N_2O and nitrogenous variables across case plants. Instead, it was shown that the similarity between N_2O and the variables NH_4 and NO_3 varies for all case plants without any clear seasonal pattern. As stated in Section 2.4.2, the established significance level is 0.01, indicating that only correlations deemed statistically significant are considered in this study. Consequently, the figures exclusively display statistically significant correlations, leading to interruptions in some of the plotted lines. Statistically insignificant correlations may be the result of bad data quality, especially if data are missing in the investigated variables.

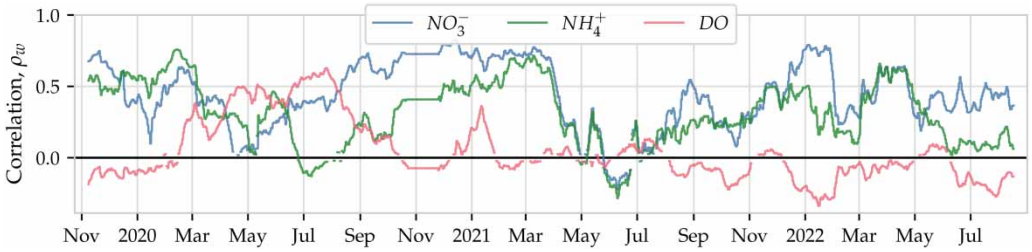


Figure 5 | Fredericia sliding window rank correlation.

Examining data from Fredericia in Figure 5 shows that the similarity between N_2O, NH_4, and DO varies throughout the dataset. The similarity between N_2O and NO_3 shows analogous behavior to that of N_2O and NH_4. Comparing Figure 5 to the seasonal variations shown in Figure 4(a), the mean and standard deviation of N_2O and correlation nitrogenous variables is simultaneously low. That is, in the periods from March to May 2020, May to November 2021, and May to September 2022, the N_2O mean and standard deviation are below 0.06 mg/L while the correlation to nitrogenous substances in the same periods is below 0.5. The correlation between N_2O and DO is, similarly, not constant over time.

Compared to each other, the two reactors at Skanderborg WWTP exhibit different characteristics for the correlation between NH_4, NO_3, and N_2O. In the peak periods, which are October to November for reactor 1 and December for reactor 2, Figure 6 shows a low correlation between nitrogenous variables and N_2O. The N_2O to DO correlation in this period is even lower and close to 0. The highest correlation is in March to June for

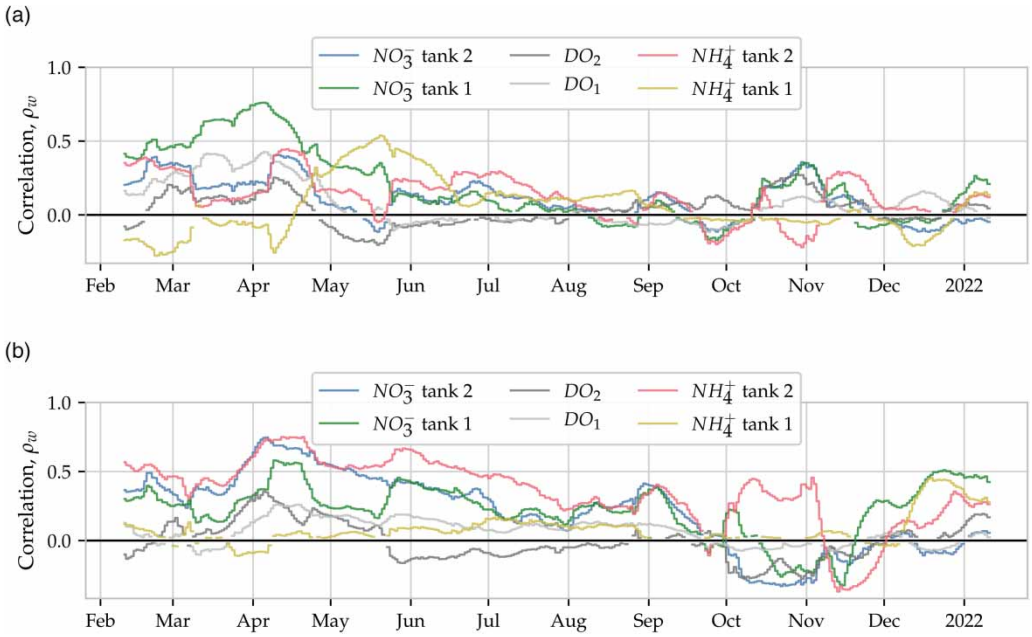


Figure 6 | Skanderborg sliding window rank correlation in relation to N_2O in (a) tank 1 and (b) tank 2.

tank 1 and April to July for tank 2, in which periods the N_2O concentrations are generally below 0.2 mg/L in the two reactors.

In Avedøre and Køge WWTPs, the two reactors in each line are connected. Hence, it is relevant to compare N_2O concentrations to operational conditions in both tanks to detect crossing relations. Results presented in Figure 7 show that the highest correlations are found between N_2O and NO_3 in the respective tanks. An interesting observation for the Avedøre WWTP is that in three out of four reactors, the second highest correlation is between N_2O and the adjacent reactor's NH_4 concentration. This may be explained by the way the reactors are controlled, where the two reactors often operate opposite each other. The lowest rank correlation is in all four tanks between N_2O and NH_4 in the same tank and DO in the adjacent tank. The correlation to NH_4 is consistent with existing studies, as nitrite occurs in the nitrification process where NH_4 is converted; and several studies find that N_2O production is linked with nitrite concentrations (Desloover *et al.* 2012; Massara *et al.* 2017; Vasilaki *et al.* 2019). Note that this study exclusively concentrates on operational data. Since none of the case plants analyzed in this study incorporate nitrite measurements into their daily operations, it has not been feasible to make comparisons between N_2O and nitrite trends.

Similar patterns are not found in the sliding window correlation of Køge data, which is shown in Figure 8. It should, however, be emphasized that the period only covers 8 months and there are only NH_4 and NO_3 sensors in the second tank. Figure 8(a) shows low correlations between N_2O and NH_4, NO_3, which does not agree with the observations of the Avedøre data. Comparing nitrogenous measurements in tank 2 to N_2O production in tank 2, shows low correlation coefficients, with only $\rho_w > 0.5$ in November 2022. There is an occasionally high correlation between DO and N_2O in reactor 2, which is not observed in reactor 1. In general, the correlations in reactor 1 are very small, which may be explained by the lack of NH_4 and NO_3 sensors in the tank.

In addition to the observed seasonal variations of N_2O in all the case plants, the relationship between N_2O and NH_4, NO_3 varies as well. Vasilaki *et al.* (2018) concluded that there was no significant correlation between N_2O and operational variables in a plug-flow carousel reactor; and Kosonen *et al.* (2016) also identified different relationships over two periods of monitoring N_2O. Our results are in line with those findings, as the sliding window rank correlation shows period-wise high correlation but also periods with no significant relationship between N_2O and nitrogenous variables. The results show clear variations within the case plants, where there were no recognizable patterns. A likely cause of this may be threefold. First, the plants differ in size and capacity. Second, some of the plants have different process designs and primary treatment facilities. In this study, we only address the secondary treatment, and specifically the biological nutrient removal, but future research may include analyses involving data from the primary treatment. Third, for seemingly similar process designs, the processes

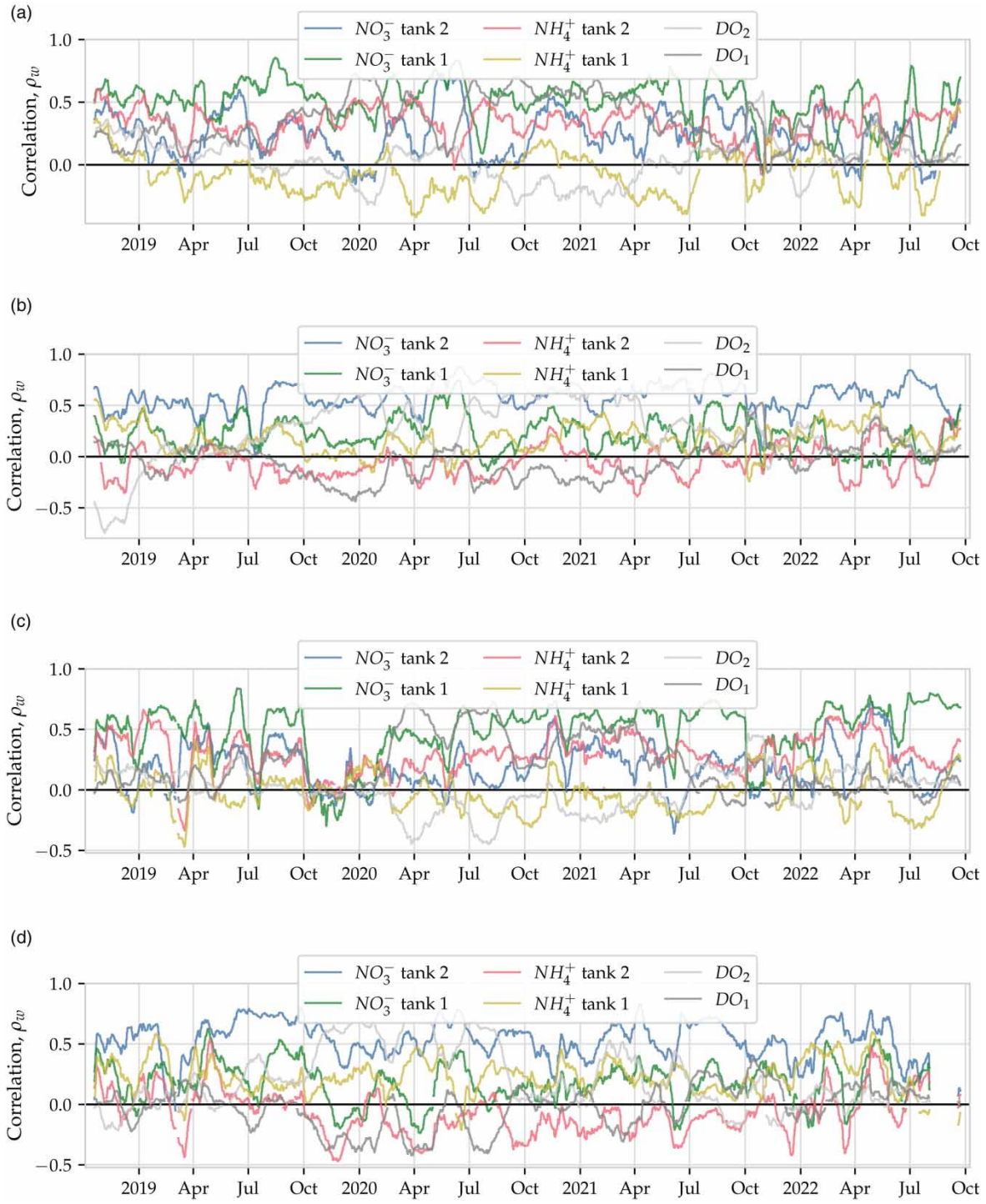


Figure 7 | Avedøre sliding window rank correlation with respect to N_2O in (a) line 1 tank 1, (b) line 1 tank 2, (c) line 3 tank 1, and (d) line 3 tank 2.

are monitored differently with different sensors, sensor locations, sensor ranges, etc. This fact alone highly supports and suggests that mechanistic modeling/first principal modeling will be difficult to implement, and also a challenge to achieve great results due to the inherent differences between all plants and the characteristics of the influent wastewater.

Taken together, these findings provide support for the belief that, to model N_2O dynamics across different WWTPs, there is a need for several years of high-quality operational data and an application of data-driven modeling.

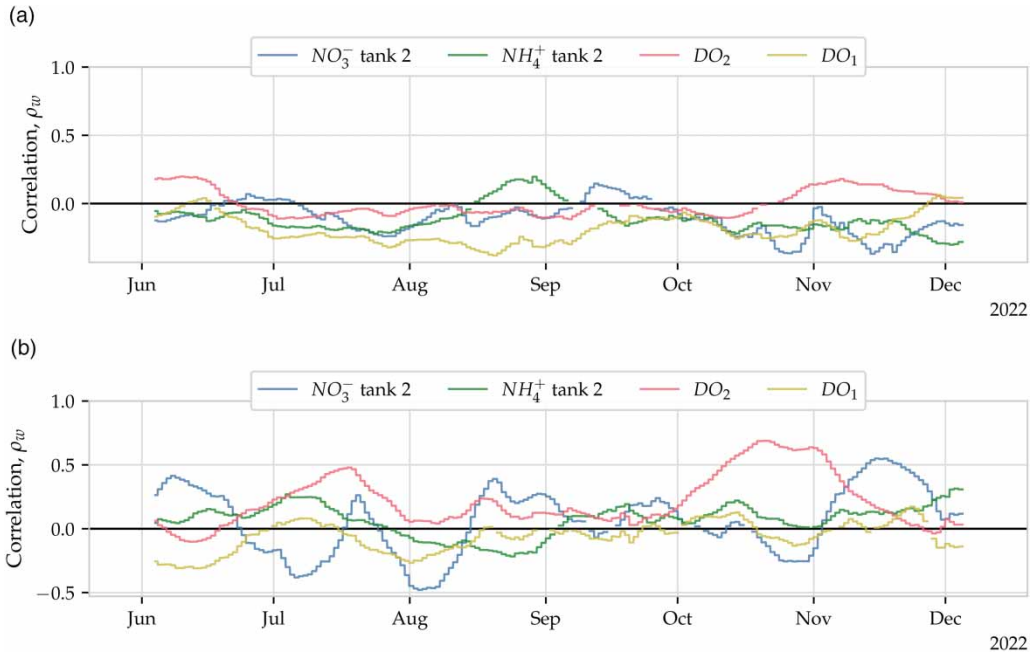


Figure 8 | Køge sliding window rank correlation with respect to N_2O in (a) tank 1 and (b) tank 2. Note that NH_4 and NO_3 are only measured in tank 2, whereas N_2O is measured in tank 1 and tank 2.

4. CONCLUSIONS

The quality of WWTP operational data varies across plants. Within the plants, the quality also varies across sensor types. The findings demonstrate that each case is unique, and model development must be adapted to the specific plant and appertaining dataset. It is, furthermore, suggested that future research should consider adapting models so that more influence is linked to reliable measurements identified by evaluation of data quality. This discovery underscores the complexity of legislating and regulating N_2O emissions. Crafting effective legislation on the subject extends beyond merely setting limits; it necessitates inclusion of requirements for data quality and measurement frequency. Regarding N_2O modeling, these findings emphasize the importance of metadata. Ensuring accurate models requires meticulous attention to metadata to prevent development and fitting to incorrect data. This study suggests that taking tangible action to curb GHG emissions demands substantial resource investment (time, sensors, maintenance) and commitment from plants and operators. Analyzing, quantifying, and acting upon actual GHG emissions require significant effort.

A recurrent seasonal N_2O concentration pattern was found for Avedøre and Fredericia WWTPs. For the remaining plants, the datasets were not long enough to confirm a recurrent pattern. The N_2O concentration is a heteroscedastic variable across all four case plants, hence confirming that the full N_2O dynamics are not represented in datasets shorter than 1 year.

The investigation of time-varying interrelation between N_2O and nitrogenous variables showed no clear pattern within or across different case plants. The findings establish that there is a need for several years of high-quality operational data and an application of data-driven modeling. To increase the accuracy and completeness of the stored data, it is also suggested that future initiatives are taken toward the collection and storing of metadata in WWTPs.

ACKNOWLEDGEMENTS

The authors thank Krüger-Veolia Colleagues for their technical and supervisory support. This project was supported by the Danish Innovation Foundation (grant number: 1044-00031B).

DATA AVAILABILITY STATEMENT

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

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Breck, E., Polyzotis, N., Roy, S., Whang, S. & Zinkevich, M. 2019 Data validation for machine learning. In: *Proceedings of Machine Learning and Systems*, pp. 334–347.

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Harmouch, H. & Naumann, F. 2022 The effects of data quality on machine learning performance. arXiv preprint arXiv:2207.14529, pp. 1–40.

Bundgaard, E., Andersen, K. L. & Petersen, G. 1989 Bio-denitro and bio-denipho systems – Experiences and advanced model development: The Danish systems for biological N and P removal. *Water Science and Technology* **21**(12), 1727–1730.

Cai, L. & Zhu, Y. 2015 The challenges of data quality and data quality assessment in the big data era. *Data Science Journal* **14**, 1–10.

Chatfield, C. 2003 *The Analysis of Time Series: An Introduction*, 6th edn. Chapman and Hall/CRC, New York.

Chen, X., Mielczarek, A. T., Habicht, K., Andersen, M. H., Thornberg, D. & Sin, G. 2019 Assessment of full-scale N_2O emission characteristics and testing of control concepts in an activated sludge wastewater treatment plant with alternating aerobic and anoxic phases. *Environmental Science and Technology* **53**(21), 12485–12494.

Daelman, M. R. J., Van Voorthuizen, E. M., Van Dongen, L. G. J. M., Volcke, E. I. P., Van Loosdrecht, M. C. M. & Haskoning, R. 2013 Methane and nitrous oxide emissions from municipal wastewater treatment – Results from a long-term study. *Water Science and Technology* **67**(10), 2350–2355.

Danish Government 2020 Klimaplan for en gron affaldssektor og cirkular okonomi (Climate plan for a green waste sector and circular economy). https://www.regeringen.dk/media/9591/aftaletekst.pdf.

Delre, A. & Scheutz, C. 2019 Site-specific carbon footprints of Scandinavian wastewater treatment plants, using the life cycle assessment approach. *Journal of Cleaner Production* **211**, 1001–1014.

Desloover, J., Vlaeminck, S. E., Clauwaert, P., Verstraete, W. & Boon, N. 2012 Strategies to mitigate N_2O emissions from biological nitrogen removal systems. *Current Opinion in Biotechnology* **23**(3), 474–482.

Fleckenstein, M. & Fellows, L. 2018 *Modern Data Strategy*, 1st edn. Springer, Cham.

Fürber, C. 2015 *Data Quality Management with Semantic Technologies*, 1st edn. Springer Gabler, Wiesbaden.

Hansen, L. D., Stokholm-Bjerregaard, M. & Durdevic, P. 2022 Modeling phosphorous dynamics in a wastewater treatment process using Bayesian optimized LSTM. *Computers & Chemical Engineering* **160**, 107738.

Herzog, T. N., Scheuren, F. J. & Winkler, W. E. 2007 *Data Quality and Record Linkage Techniques*, 1st edn, Vol. 1. Springer, New York. https://doi.org/https://doi.org/10.1007/0-387-69505-2.

Hwangbo, S., Al, R., Chen, X. & Sin, G. 2021 Integrated model for understanding N_2O emissions from wastewater treatment plants: A deep learning approach. *Environmental Science and Technology* **55**(3), 2143–2151.

Hwangbo, S., Al, R. & Sin, G. 2020 An integrated framework for plant data-driven process modeling using deep-learning with Monte-Carlo simulations. *Computers and Chemical Engineering* **143**, 107071.

IPCC 2013 *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Technical Report, Cambridge University Press, Cambridge, United Kingdom and New York, USA.

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. *Environmental Science and Technology* **50**(11), 5547–5554.

Law, Y., Ye, L., Pan, Y. & Yuan, Z. 2012 Nitrous oxide emissions from wastewater treatment processes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**(1593), 1265–1277.

Li, K., Duan, H., Liu, L., Qiu, R., 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. *Environmental Science & Technology* **56**(4), 2816–2826.

Massara, T. M., Malamis, S., Guisasola, A., Baeza, J. A., Noutsopoulos, C. & Katsou, E. 2017 A review on nitrous oxide (N_2O) emissions during biological nutrient removal from municipal wastewater and sludge reject water. *Science of the Total Environment* **596–597**, 106–123.

Myers, S., Mikola, A., Blomberg, K., Kuokkanen, A. & Rosso, D. 2021 Comparison of methods for nitrous oxide emission estimation in full-scale activated sludge. *Water Science and Technology* **83**(3), 641–651.

Ni, B. J., Ye, L., Law, Y., Byers, C. & Yuan, Z. 2013 Mathematical modeling of nitrous oxide (N_2O) emissions from full-scale wastewater treatment plants. *Environmental Science and Technology* **47**(14), 7795–7803.

Unisense Environment A/S 2022 N2O Wastewater system. https://unisense-environment.com/wp-content/uploads/2022/11/2022.11-N2O-WW-System_english-2.pdf.

Valk, L. C., Peces, M., Singleton, C. M., Laursen, M. D., Andersen, M. H., Mielczarek, A. T. & Nielsen, P. H. 2022 Exploring the microbial influence on seasonal nitrous oxide concentration in a full-scale wastewater treatment plant using metagenome assembled genomes. *Water Research* **219**(February), 118563.

van Dijk, E. J., van Loosdrecht, M. C. & Pronk, M. 2021 Nitrous oxide emission from full-scale municipal aerobic granular sludge. *Water Research* **198**, 117159.

Vasilaki, V., Conca, V., Frison, N., Eusebi, A. L., Fatone, F. & Katsou, E. 2020 A knowledge discovery framework to predict the N_2O emissions in the wastewater sector. *Water Research* **178**, 115799.

Vasilaki, V., Massara, T. M., Stanchev, P., Fatone, F. & Katsou, E. 2019 A decade of nitrous oxide (N_2O) monitoring in full-scale wastewater treatment processes: A critical review. *Water Research* **161**, 392–412.

Vasilaki, V., Volcke, E. I., Nandi, A. K., van Loosdrecht, M. C. & Katsou, E. 2018 Relating N_2O emissions during biological nitrogen removal with operating conditions using multivariate statistical techniques. *Water Research* **140**, 387–402.

Wang, R. Y. 1996 Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* **12**(4), 5–34.

Zhao, H., Isaacs, S. H., Søeberg, H. & Kümmel, M. 1995 An analysis of nitrogen removal and control strategies in an alternating activated sludge process. *Water Research* **29**(2), 535–544.

First received 21 November 2023; accepted in revised form 10 February 2024. Available online 28 February 2024