water

MDPI

Article

Data-Driven Drift Detection in Real Process Tanks: Bridging the Gap between Academia and Practice

Bolette D. Hansen [1,2], Thomas B. Hansen [1], Thomas B. Moeslund [2] and David G. Jensen [1,*]

[1] EnviDan A/S, Climate Adaption & Drainage Systems Department, 8600 Silkeborg, Denmark; bdha@create.aau.dk (B.D.H.); tbh@envidan.dk (T.B.H.)

[2] Visual Analysis and Perception Lab, Department of Architecture and Media Technology, Aalborg University, 9000 Aalborg, Denmark; tbm@create.aau.dk

* Correspondence: dgj@envidan.dk

Abstract: Sensor drift in Wastewater Treatment Plants (WWTPs) reduces the efficiency of the plants and needs to be handled. Several studies have investigated anomaly detection and fault detection in WWTPs. However, these solutions often remain as academic projects. In this study, the gap between academia and practice is investigated by applying suggested algorithms on real WWTP data. The results show that it is difficult to detect drift in the data to a sufficient level due to missing and imprecise logs, ad hoc changes in control settings, low data quality and the equality in the patterns of some fault types and optimal operation. The challenges related to data quality raise the question of whether the data-driven approach for drift detection is the best solution, as this requires a high-quality data set. Several recommendations are suggested for utilities that wish to bridge the gap between academia and practice regarding drift detection. These include storing data and select data parameters at resolutions which positively contribute to this purpose. Furthermore, the data should be accompanied by sufficient logging of factors affecting the patterns of the data, such as changes in control settings.

Keywords: wastewater; treatment; drift; anomaly; machine learning; data driven; detection; real data

check for updates

Citation: Hansen, B.D.; Hansen, T.B.; Moeslund, T.B.; Jensen, D.G. Data-Driven Drift Detection in Real Process Tanks: Bridging the Gap between Academia and Practice. *Water* **2022**, *14*, 926. https://doi.org/10.3390/w14060926

Academic Editors: Dimitrios E. Alexakis and Chi-Wang Li

Received: 9 February 2022
Accepted: 14 March 2022
Published: 16 March 2022

1. Introduction

With increased focus on the United Nations Sustainable Development Goals (SDGs) and increasing energy prices, there has been an increased interest in optimizing the performance of Wastewater Treatment Plants (WWTP) and Wastewater Recovery Facilities, prospectively referred to as WWTPs in this article. Optimizing the operation of WWTPs is the topic of several studies [1] where some of the more complex control systems include the energy usage and economics [2,3]. Several companies offer software for real time control of WWTP; however, few focus on the data quality involved [4]. This is problematic as sensor drift can induce decreased total N removal or over aeration, entailing a large increase in energy consumption. Bias in sensor data can easily counteract the energy reduction and cost savings obtained by advanced automatic control [5].

Drift in sensors is a commonly known problem. Due to fouling, optical Dissolved Oxygen (DO) sensors can easily be biased with one mg/L within a month [6]. In calibration data from two WWTPs, examples of drift with more than one mg/L can be found for both ammonia and potassium sensors. In plants, where the NH_4 level is typically below 3 mg/L in the outlet, 1 mg/L is a large deviation. Especially a positive drift in ammonia sensors is subject to increased costs at the utilities, and from an industrial perspective it should be of high priority to detect these faults.

Faulty sensor data is a problem in several different sectors. Teh et al. [7] reviewed 57 papers on sensor faults and methods used for detection and correcting faulty data. The faults included outliers, missing data, bias, drift, noise, constant values and the sensor being stuck at zero. The methods used for drift detection in the reviewed papers were

Principal Component Analysis (PCA), Artificial Neural Network (ANN), Ensemble Classifiers and Dempster-Shafer Theory and Mathematical Modelling (only one paper from 2008). Furthermore, PCA, calibration-based methods, PCA based methods and Kalman filter-based methods were used for drift detection and correction [7].

Several data-driven approaches for wastewater treatment operation have been developed and presented in the literature; however, in a review, Corominas et al. [4] report that only 16 percent of the developed solutions resulted in a commercial product, and seven percent were commercialized without full-scale testing. Within fault detection the most popular approach is PCA (27 papers) followed by ICA (9 papers) and clustering (7 papers) [4].

Within fault detection in process tanks, several studies have been made. Baklouti et al. [8] used univariate statistics to detect bias, drift and variating magnitudes in Dissolved Oxygen (DO) sensors. They introduced bias of two mg/l, drift with a slope of 0.005 and variating signal magnitude of three standard deviations. The calculation of 700 datapoints was made without information on the actual corresponding sampling frequency. Despite stating that it was in general problematic that models did not encounter seasonal changes, etc., the authors tested their model in simulated dry weather data [8].

In 2002 Thomann et al. [9] suggested using control charts to make it easier for WWTP staff to detect drift, outliers and shifts based on four months of collected data. Newhart et al. [10] stated that control charts are well suited for monitoring single variables which only contain a low degree of noise, having been measured on a daily to monthly basis.

Baggiani and Marsili-Libelli [11] used PCA combined with moving windows, T^2 and Q statistics, as well as threshold, to detect spikes and sensor faults in data from a real plant and obtained performances of 100% and 84% depending on the window size used [11]. It is worth noticing that the spikes and faults exemplified in the paper are very distinctive compared to the signal amplitude.

Alferes et al. [12] used PCA over six days and found two deviations in the PCA analysis. The first was explained by a high unusual discharge and the second was related to a turbidity sensor. It is worth noticing that when observing the turbidity data, another case is eye catching; however, it is found in the PCA analysis.

Cheng et al. [13] used kernel PCA (KPCA) and one-class support vector machine to detect anomalies in the inflow components of a real plant over seven years and obtained better results than when using linear PCA and K-nearest-neighbours.

Huang et al. [14] proposed a method for anomaly detection in a WWTP at a paper mill; however, only one case with faulty behaviour was available and the process was in a more closed and controlled environment than a normal treatment plant. This is indicated by specific time slots for different processes to take place.

In 2020 and 2021, several methods for fault detection in WWTPs were proposed in the literature. Ba-Alawi et al. [15] used stacked denoising autoencoders for detection of drift, bias, precision degradation and complete failure. The method was evaluated on simulated dry weather data and the authors state that the method was superior to existing methods and can reduce operating costs and improve the monitoring of the influent [15]. Kazemi et al. [16] showed that incremental PCA was able to distinguish between time varying events and faults in simulated data, while Kazemi et al. [17] investigated a number of technics including Support Vector Machine, Ensemble Neural Network and Extreme Learning and found that they performed better than a PCA based method after testing on simulated data. Luca et al. [18] applied PCA and statistic for fault detection in DO sensors in simulated data and stated that the method was successful in detecting the faults. Mali and Laskar [19] proposed an optimized Monte Carlo deep neural network and were able to detect faults of low magnitude in simulated data. Xu et al. [20] proposed a version of ICA called complex-valued ICA. The method was both evaluated for simulated data and for data from a real plant. In the real case, the authors had 213 samples of which 45 were from normal operation, and these were used for training; however, these samples were also included in the test set. The authors stated that this method could obtain more accurate, intuitive and efficient fault detection. Klanderman et al. [21] proposed a method based on auto correlation and Fused Lasso. The method was trained on an in-control data set and tested on a simulated data set with introduced faults and data from a real plant, which contained one fault that they were able to detect. Mamandipoor and Majd [22] possessed 11 months of data from 12 sensors in a real plant. The data were classified according to faulty NH_4 data by an expert and a Long Short-Term Memory Network was developed and outperformed PCA-SVM. Cecconi and Rosso [23] used ANN to predict the NH_4 concentration and used PCA along with Shewhart monitoring charts for detection of the variation between measured values and predicted values. This study was based on more than one year of data from a real plant. Six sensors were installed in the plant including two NH_4 sensors. The sensors were cleaned on a weekly basis and calibrated if there was a difference detected of more than 15% between the sensor and the reference. The faults considered in the study were sensor faults caused by wrong calibration, process anomaly and drift. For testing, three types of faults were introduced in real data. The suggested approach was able to detect the faults and the ANN prediction could be used for process control when a fault was detected [23]. Anter et al. [24] used fuzzy swarm intelligence and chaos theory to detect faults in a real data set from 1993 available at the UCI Machine Learning Repository [24]; however, details on the fault types detected are not described.

Except for Cecconi and Rosso [23] and Mamandipoor and Majd [22], none of the solutions proposed in 2020–2021 reflect contemporary conditions met at WWTPs, and while several papers acknowledge that there is a gap between the solutions in academia and in the real world [4], there is a lack of knowledge when it comes to implementing data-driven approaches in real WWTPs.

The aim of this paper is to bridge the gap between academia and practice by applying different approaches for machine learning to real-world data sets, and thereby identify challenges hindering implementation of data-driven drift detection at normal operating WWTPs. The main contribution of this work is identification of the shortcomings between academia and practice together with recommendations for future data usage and management obtained in collaboration between data scientists and water professionals. To ensure that the recommendations are as relevant as possible for both researchers and managers, this work is based on data available from operating WWTPs, and no extra data acquisition was made. This entails the data being of lower quality than if it is acquired with the specific purpose of developing algorithms for drift detection.

The remainder of this paper is structured as follows. Section 2 contains information on the data and approaches investigated in this study. Section 3 contains the results and description of how to interpret these. Section 4 is a discussion of the results and Section 5 contains perspectives on drift detection from both academic and practical perspectives. These perspectives are accompanied by recommendations for the future. The paper is concluded in Section 6.

2. Materials and Methods

This section contains an overview of the available data and the applied methodology for anomaly and fault detection. With inspiration from the literature, several methods for anomaly and fault detection were initially considered; however, it became clear that many of the considered methods were not practically applicable. As the purpose of this paper is to bridge the gap between academia and research, descriptions of the unsuccessful methods have been included in this section, together with a description of why they were not successful in this case. Lastly, a description of how the detected anomalies are accessed is included.

2.1. Data

Data from three plants were available for this study. The resolution of the data was one sample per minute, and two of the WWTPs had a log accessible with calibration information. One of the WWTPs had one process tank (PCT) while the remaining two WWTPs had two PCTs. An overview of the PCTs can be seen in Table 1.

Table 1. Overview of the available data.

Process Tank	Data Period	Log *
WWTP 1 PCT 1	From 25 January 2021 To 14 September 2021	4 measurements 1 calibration
WWTP 2 PCT 1	From 13 June 2020 To 14 September 2021	Not available
WWTP 2 PCT 2	From 13 June 2020 To 14 September 2021	Not available
WWTP 3 PCT 1	From 1 February 2021 To 14 September 2021	3 measurements 3 calibrations
WWTP 3 PCT 2	From 1 February 2021 To 14 September 2021	3 measurements 2 calibrations

* Logs were available until 22 May 2021.

The control strategy for aeration of WWTP2 PCT1 was based on alternating operation where the air pumps turned on and off based on ammonia set points. The remaining PCTs were controlled by PID controllers. A PID controller tries to obtain a constant NH_4 level which is defined by a set point. The PID controller adjusted the amount of aeration based on the difference between the NH_4 concentration and the set point for the NH_4 concentration. How fast the PID adjust the aeration depends on three constants. This control strategy is beneficial as it allows for a more constant concentrations in the PCT.

Multiple parameters were available for the three plants including flow to the plant, NH_4, NO_3, DO, K and SS, while other parameters variated between the plants such as information on the aeration, if N_2O was measured, etc. The parameters flow, NO_3 and DO are highly related to the NH_4 level in the plant. Furthermore, plots of the data did not indicate that the remainder of the parameters, which were available for all the PCTs, should be included. Therefore, it was decided to focus on the parameters flow, NH_4, NO_3 and DO.

For two of the plants, lab measurements and calibration logs were kept for the NH_4 sensor and the NO_3 sensor. From the logs it could be seen that a drift of the NH_4 sensor of 0.5 mg/L was accepted, while a drift of 1 mg/L was accepted for the NO_3 sensor. In the log calibration, events were noted down; however, this was done manually. In some cases, it was stated that a sensor was adjusted, but it was not stated which sensor.

2.2. Machine Learning Approaches

As described in Section 1, several different data-driven approaches for drift and fault detection in WWTPs exist; however, these methods cannot be directly applied to the data available for this study.

A characteristic for almost all the methods presented in the literature is that they have been developed and tested on data sets where the faults are already known, either because the faults have been simulated or because the data come from well monitored WWTPs. Such labelled data sets are rarely available for normal WWTPs, which is also the case for the data available for this study. Therefore, it was sought to obtain a labelled data set for drift by manually labelling the data in the PCT with alternating operation, as this was the easiest PCT to assess. To do this, an interactive software tool for systematic labelling of each aeration cycle was made. Each aeration cycle could then be labelled as OK or as a fault type. However, during the labelling process it was observed that it was hard to label the data without introducing several faults. Reasons for this included the operators changing the control settings instead of calibrating the sensors and the utility accepting the NO_3 sensor to drift with up to 1 mg/L without considering it as an anomaly.

The lack of labelled data entails that it is not possible to use traditional supervised learning. Another approach initially tested was predicting each parameter based on one class learning. Thereby the variations between the prediction and the measurement would be the fault. However, this task was complicated by the fact that the immediate previous measurement could not be used as input for the predictive machine learning algorithm, as drift develops over time. Thereby most of the drift would also be present in the immediate previous measurement and consequently, the algorithm would predict the measured value and not the real value. Therefore, experimenters tried to train a Random Forest model, which is an ensemble method, on the first 80% percentage of the data and test it on the remaining 20% for each PCT. For this task, daily average values were used to neglect normal variations during a day such as increased flow in the rush hours, rainfall and when the aeration pump was activated. This approach showed low performance of the algorithm and the main bottleneck for obtaining better results was the large variations in control setting at the PCTs. Therefore, experimenters decided to use unsupervised learning.

2.3. Unsupervised Learning Algorithms

As described in Section 1, a commonly used unsupervised method for fault detection in WWTPs is PCA. Therefore, the data sets were normalized according to the standard deviation and examined through PCA. All combinations of Principal Components (PC) were then plotted per day and visually inspected. It was observed that the patterns changed over time, and especially changes in control settings caused the patterns to change. Changes in patterns when plotting principal components were also observed by Alferes et al. [12] who only looked at a few days of data. However, when considering several months of data this approach is not efficient, as the evaluation is based on visual inspection. Furthermore, it was found to be much simpler to interpret the data and changes by simply plotting all combinations of parameters per day. It was also investigated if using PCA on daily values could be used to detect anomalies. In this connection, it was tested if faults and anomalies could be removed by removing the least contributing PC; however, the anomalies were present in all principal components and this approach did not work.

More complex solutions such as deep auto-encoders were considered; however, based on the results with one class learning it was not expected that this approach would be efficient. Therefore, for the purpose of this study, it was found more relevant to use a simpler and more transparent approach.

The last approach considered was to use the Local Outlier Factor (LOF) [25] on daily values. Initial results showed that this method gave the most promising results, for which reason it was chosen to use LOF.

Local Outlier Factor

LOF is an unsupervised learning algorithm which measures the distance to a certain number of nearest neighbours and uses this distance as a measure of anomaly.

For the LOF it was decided to use daily data. This was done to neglect the large variations in inflow, wastewater composition and aeration periods during a day. After averaging the data to daily signals, the data were scaled according to the standard deviation. In the specific implementation of the LOF, the distance to the 20 nearest neighbours was used to calculate the LOF. To ensure that the method can be applied in real time, the LOF was implemented as a Moving LOF filter, where the LOF for a given day was based on the 99 previous days.

A threshold of two was applied to the Moving LOF, and all datapoints exceeding the threshold were considered as abnormal. All periods of abnormal behaviour were subsequently assessed.

2.4. Assessment of Anomalies

Several different types of anomalies were present in the data. For gaining an overview, the anomalies were categorized into five general groups, namely missing data, increased presence (referring to increased flow or increased presence of NH_4, NO_3 or DO), change in control settings, sensor drift or over aeration and other. In some cases, multiple anomalies were present, and in these cases it was evaluated, which was the primary reason for the detection. For instance, there could be a scenario where a sensor has drifted but nothing is detected until an increase in flow is present and after the next day, nothing is detected again. In such a case the anomaly is annotated as an "increased presence", even though the reason for the anomaly to be detected might be a combination of the drift and the increase in flow.

Plots were made for each PCT showing the anomaly category and relevant data examples were plotted. Additionally, examples of longer periods of anomalies not reaching the threshold were plotted.

3. Results

This section contains a description of the results of the anomaly detection. The results for each of the PCTs are presented in Sections 3.1–3.5. For each of the PCTs, examples of anomalies have been highlighted. The examples have been selected so that as many different scenarios as possible are shown, to give insight into as many scenarios as possible. Details on all observations are presented in Table 2. Furthermore, general observations are described in Section 3.6.

3.1. WWTP1 PCT1

The data available for WWTP1 PCT1, calibration and lab measurements, Moving LOF and the anomalies detected using the Moving LOF and thresholding can be seen in Figure 1. The detected anomalies are colour coded according to the anomaly type observed. In the figure, it is worth noticing that several different control settings have been used in the first period for which the data was available. Consequently, the algorithm does not consider this type of control setting as an anomaly if it is strongly present in the LOF window. This might be the reason that changes in control settings in the middle of May 2021 were not detected. As seen in the figure, most of the detected anomalies were related to increased presence of one or more of the parameters.

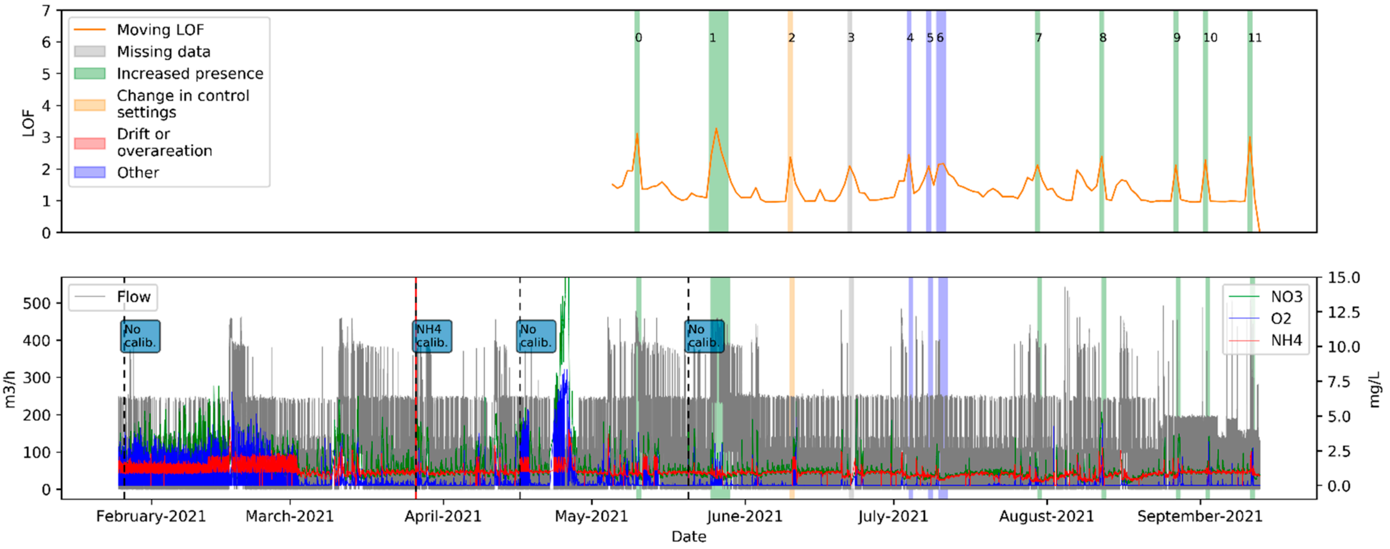


Figure 1. WWTP1 PCT1. In the upper graph, the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 11, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4 and the calibration data available from the plant.

Table 2. Overview of the anomalies detected in the five PCTs when using Moving LOF and a threshold of two.

WWTP 1 PCT 1	WWTP 2 PCT 1	WWTP 2 PCT 2	WWTP 3 PCT 1	WWTP 3 PCT 2
0. Increased flow 1. Increased flow 2. Change in PID 3. Missing data 4. Other 5. Other 6. Other 7. Increased flow 8. Increased NH_4, NO_3 and DO 9. Increased NO_3 and DO 10. Increased NO_3 and DO 11. Increased flow	0. Missing data 1. Increased flow + NH_4 drift (up) 2. Increased flow + NH_4 drift (up) 3. NH_4 drift (up) 4. NH_4 drift (up) 5. Increased flow 6. Increased flow 7. High concentrations of NH_4 and NO_3 8. High concentrations of NO_3 present or NO_3 sensor drifted (up) 9. Increased flow 10. Increased flow 11. Increased flow 12. NH_4 drift (up) 13. This anomaly starts as NH_4 drift (up). The second day data is missing for almost 13 h. Hereafter, the lower setpoint seems to be slightly increased with 0.1, which handles the problems with over-aeration. The last day of the anomaly is due to an increased flow. 14. All parameters are low except for the flow. Maybe this PCT has been out of operation or experiments had been performed. 15. Increased flow 16. Increased flow 17. Low parameters, see 14 18. Low parameters, see 14 19. Increased flow 20. High levels of NH_4 present day the first day, increased flow the second day 21. NH_4 drift (up) 22. NH_4 drift (up) 23. NH_4 drift (up) 24. Change in setpoint. In the period up to the detection of this anomaly the setpoints were increased multiple times. This also happened two days before this anomaly was detected. The day before this anomaly was detected, the setpoints were decreased inducing over aeration. The day after the setpoint was increased again, which was the case for the remainder of the anomaly. The LOF decreased over time as it learnt the new behaviour 25. Missing data 26. Increased flow and NH_4 concentration 27. Increased flow and NH_4 concentration	0. Increased flow combined with changed control settings the previous day 1. Increased flow 2. Increased flow 3. Increased flow 4. Increased flow and increased NO_3 unrelated to flow 5. Increased NO_3 6. Increased NO_3 7. Increased flow 8. Data shows low flow, very large amounts of DO and increasing NH_4. 9. Increased flow 10. NO_3 and NH_4 the first day, increased flow the second day 11. Increased concentrations of NH_4, NO_3 and DO. Possible because the other PCT at the WWTP was out of operation, see anomaly 14 for WWTP2 PCT1 12. Increased flow 13. Increased concentrations of NH_4, NO_3 and DO. One day with increased flow. Possible because the other PCT at the WWTP is out of operation, see anomaly 17–18 for WWTP2 PCT1. 14. Increased concentrations of NH_4, NO_3 and DO. One day with increased flow, see 13. 15. Increased flow 16. Increased flow 17. Increased flow 18. Missing data 19. NH_4 sensor drifted (up) 20. Increased flow 21. Increased flow 22. Increased flow 23. Increased flow	1. Change in PID. There is an increased flow starting the day before and continuing two days after the anomaly was detected. In the period of the anomaly, the pattern of the sensors changed, indicating change in control settings. 2. Increased NH_4 concentration due to missing aeration 3. Increased flow, inducing high NH_4, NO_3 and DO concentrations 4. Increased flow, inducing high NH_4 and NO_3 concentrations 5. Increased NH_4 concentrations inducing high NO_3 concentrations 6. Increased NH_4 concentrations inducing high NO_3 concentrations 7. Increased flow, inducing high NH_4 and NO_3 concentrations	0. Increased flow 1. Change in control settings 2. Increased flow 3. Increased flow 4. Increased NO_3, low DO 5. Increased flow, increased NO_3, low DO

Examples of an anomaly caused by increased flow and an anomaly caused by change in control settings are presented in Figure 2. Further details on the anomalies can be found in Table 2.

3.2. WWTP2 PCT1

Figure 3 shows the data, Moving LOF and detected anomalies for WWTP2 PCT1. The control strategy for WWTP2 PCT1 is based on alternating operation and the figure shows that several anomalies caused by sensor drifts or over aeration were found by the algorithm.

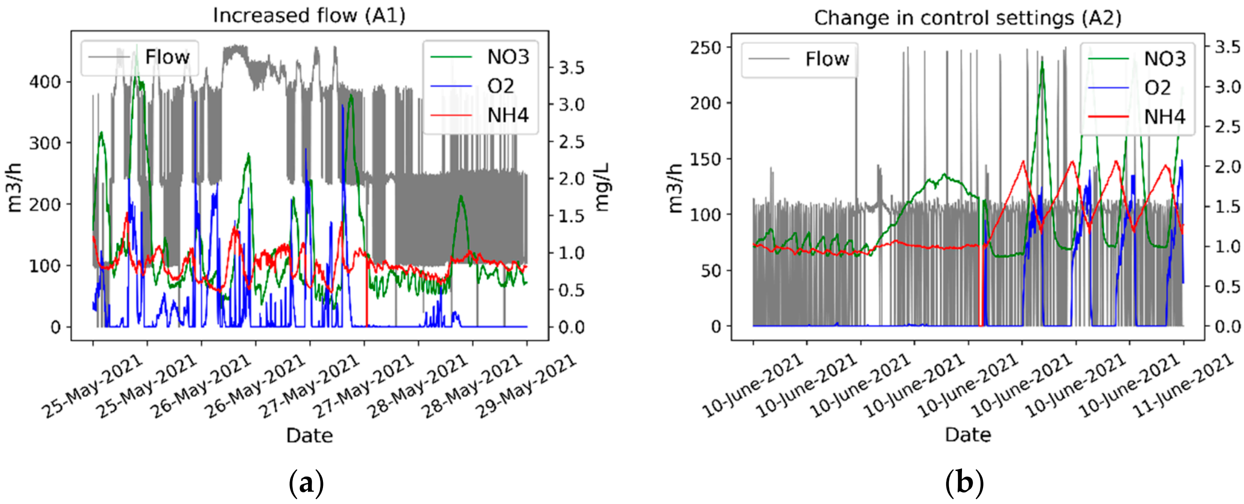


Figure 2. Examples of anomalies detected in WWTP1 PCT1. (**a**) increased flow, (**b**) change in control settings.

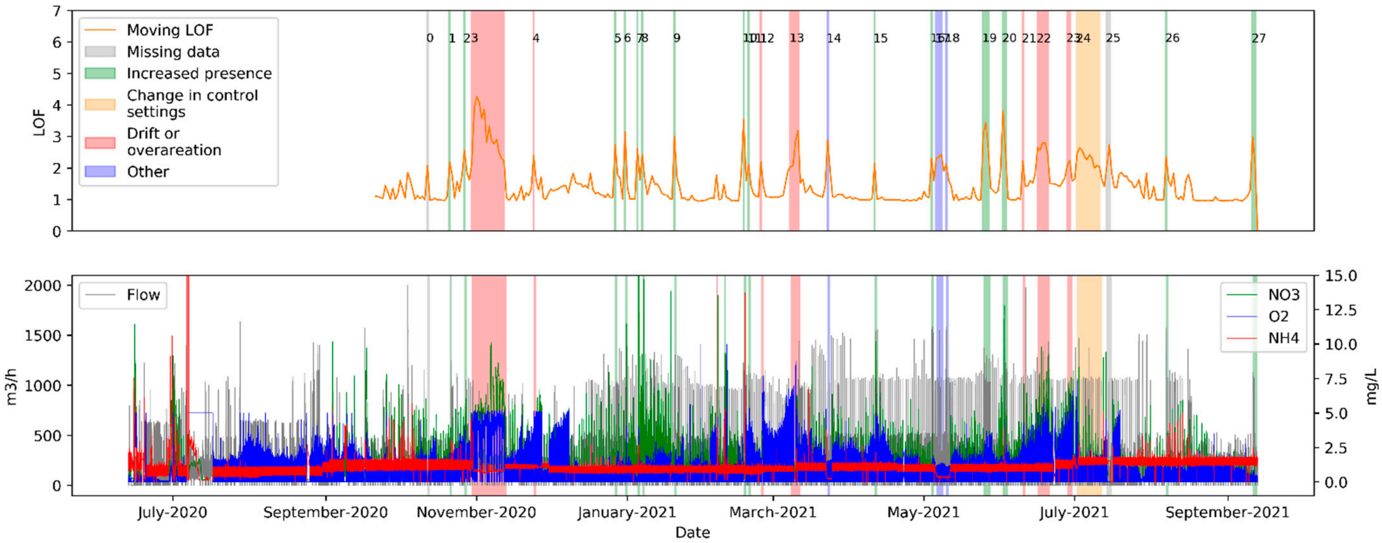
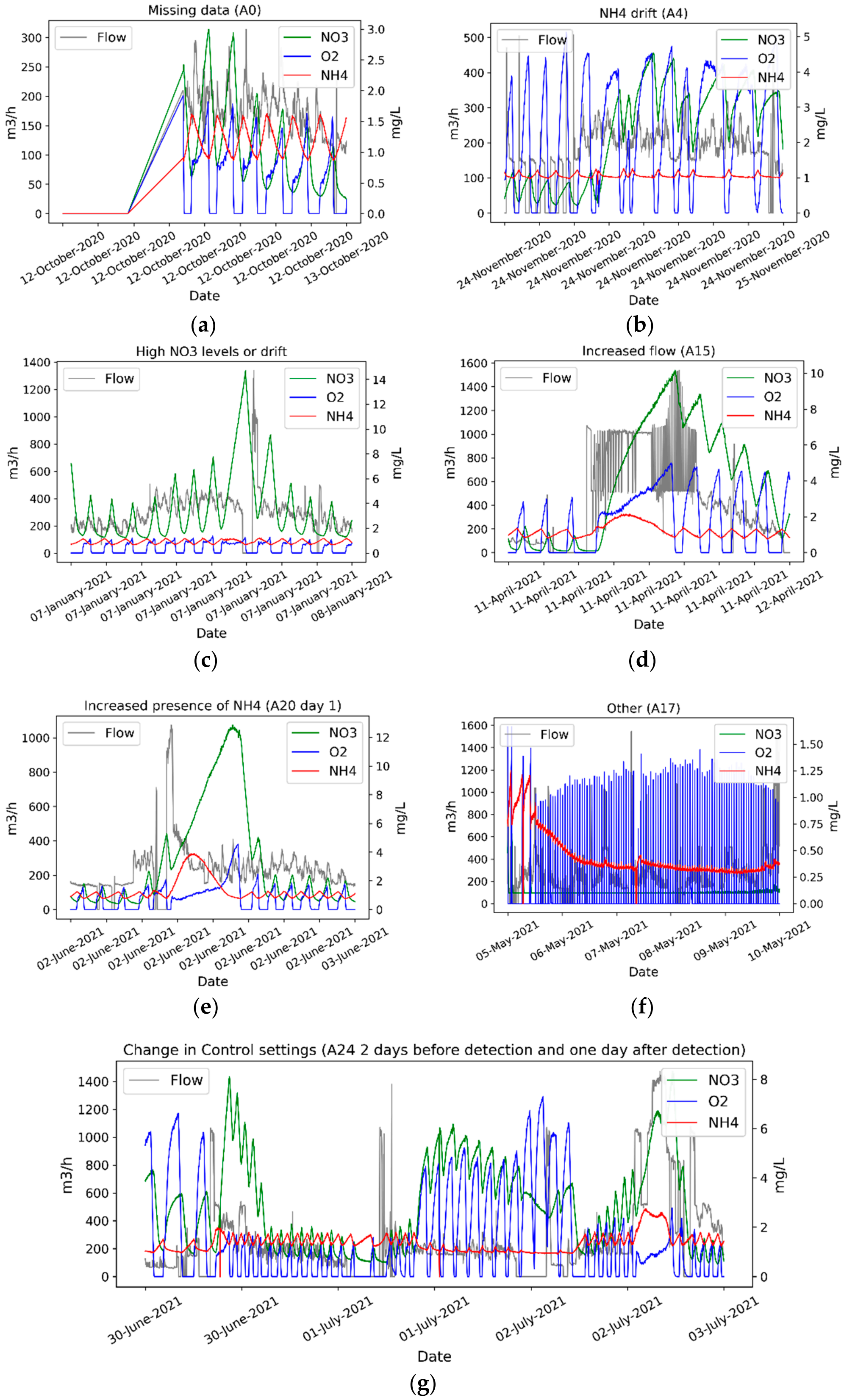


Figure 3. WWTP2 PCT1. In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 27, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4. No calibration data were available from the plant.

For WWTP2 PCT 1, in which alternating operation was used, 27 anomalies were detected. Of these, 15 were primarily detected due to increased presence of flow, NH_4, NO_3 or DO. However, in two of the cases the NH_4 sensor had already drifted but an increase in flow was the factor which made it exceed the threshold (Anomaly 1–2 in Figure 3).

Examples of missing data, NH_4 sensor drift, high NO_3 levels, increased flow, increased presence of NH_4, an anomaly categorized as other (which most likely is caused by the PCT being out of operation or experiments performed at the plant) and change in control settings are presented in Figure 4. Further details on all the anomalies detected in the PCT are presented in Table 2.



(**a**)

(**b**)

(**c**)

(**d**)

(**e**)

(**f**)

(**g**)

Figure 4. Examples of anomalies detected in WWTP2 PCT1. (**a**) missing data (W2P1A0), (**b**) NH_4 drift (W2P1A4), (**c**) High NO_3 levels or drift, (**d**) Increased flow (W2P1A15), (**e**) Increased presence of NH_4, (**f**) Other (W2P1A17), (**g**) Change in control settings (two days before onset of W2P1A24 until 1 day after detection).

In addition to the anomalies exceeding the threshold, some longer time periods with increased LOF were observed for the PCT with alternating operation. The increase in LOF was associated with a NH_4 sensor drift, which the operator compensated for by changing the set points. An example of this can be seen in Figure 5. It is worth noticing that the NO_3 and DO levels gradually increased in the period before a change in setpoints for NH_4 and suddenly decreased after the changes. This is especially clear in the period from 23 June 2021 to 1 July 2021.

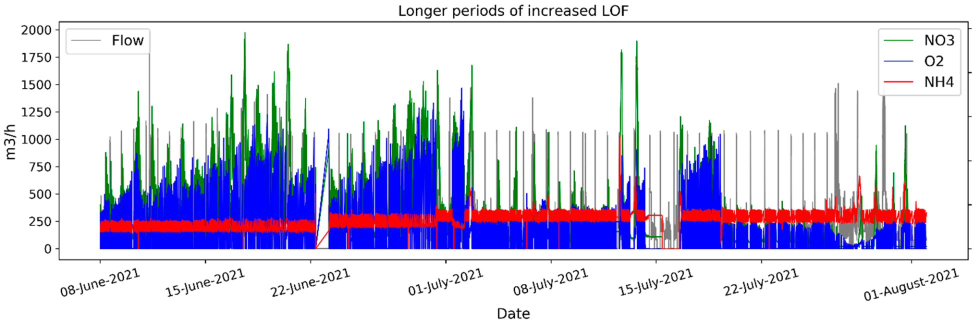


Figure 5. Long period of increased LOF in WWTP2 PCT1. The pattern in the data indicates that the NH_4 sensor had drifted and that the operator of the plant subsequently adjusted the setpoint instead of calibrating the sensor.

3.3. WWTP2 PCT2

Figure 6 shows the data, the Moving LOF and the detected anomalies for WWTP2 PCT2. As seen in the figure, increased presence of the different parameters is the most common reason for anomalies; however, increased presence of some of the parameters can be caused by other factors, such as change in the usage of the plant. For instance, anomalies 11, 13 and 14 coincide with anomalies 14, 17 and 18 in WWTP2 PCT1, which are most likely caused by PCT1 being out of operation and thereby cause an increased pressure on this PCT. Figure 6 also shows that several different control settings were used in the beginning of the data collection. However, as this was within the first 99 days of the data collection, the Moving LOF could not give the outlier score of the data for this period. When considering the anomalies detected by the Moving LOF, anomaly eight differs from previously elaborated anomalies. It has been classified as 'other', and the anomaly is most likely caused by a fault in the DO sensor as a constant increase in NH_4 concentration and low NO_3 concentration indicate a lack of DO in the PCT. A detailed plot of anomaly eight is shown in Figure 7. Further details on the anomalies detected in WWTP PCT2 can be found in Table 2.

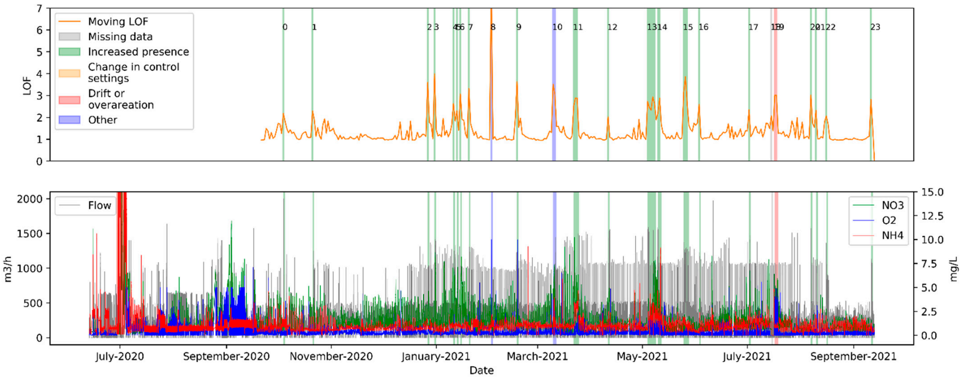


Figure 6. WWTP2 PCT2. In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 23, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4. No calibration data were available from the plant.

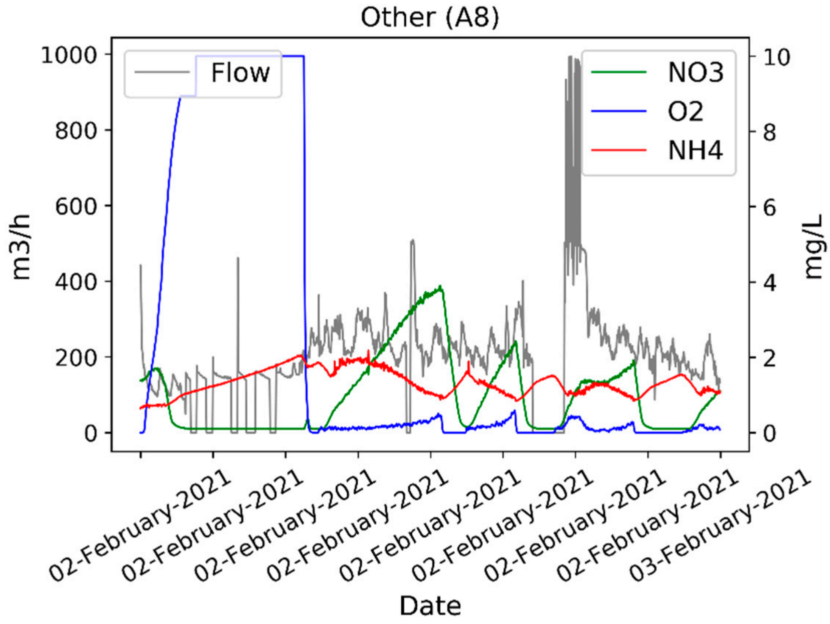


Figure 7. Example of an anomaly detected in WWTP2 PCT2. The anomaly is categorized as other. The anomaly is most likely caused by a fault in the DO sensor.

3.4. WWTP3 PCT1

The data, lab measurements and calibrations, Moving LOF and detected anomalies for WWTP3 PCT1 are shown in Figure 8. For this plant two anomalies distinguish themselves. These are the anomalies zero and one. Anomaly zero is observed during a longer period of increased flow. In the parallel PCT the full period of increased flow has been detected as an anomaly; however, for this PCT only one day during the increased flow was detected. During this day, changes in patterns indicated that the control settings were changed, possibly to deal with the increased flow. For anomaly one, it was observed that there were several hours with no DO, constantly increasing NH_4 levels and low NO_3 levels, indicating that the aeration pump had been out of operation. Detailed plots of anomaly zero and one can be found in Figure 9.

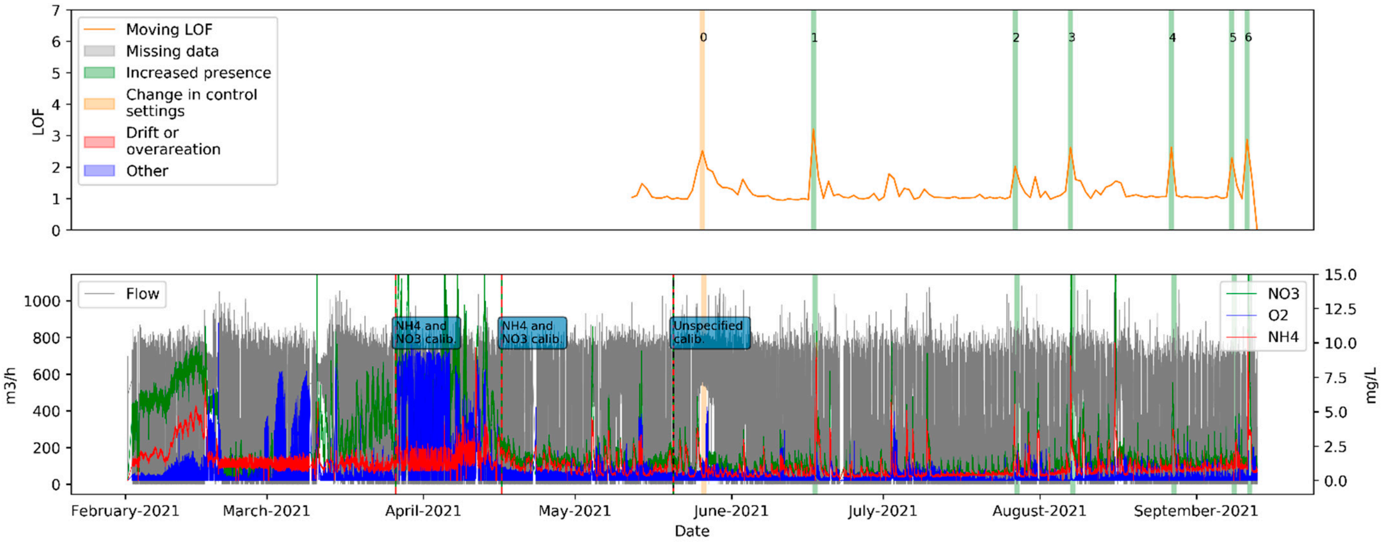


Figure 8. WWTP3 PCT1 In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 6, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4 and calibration data available from the plant.

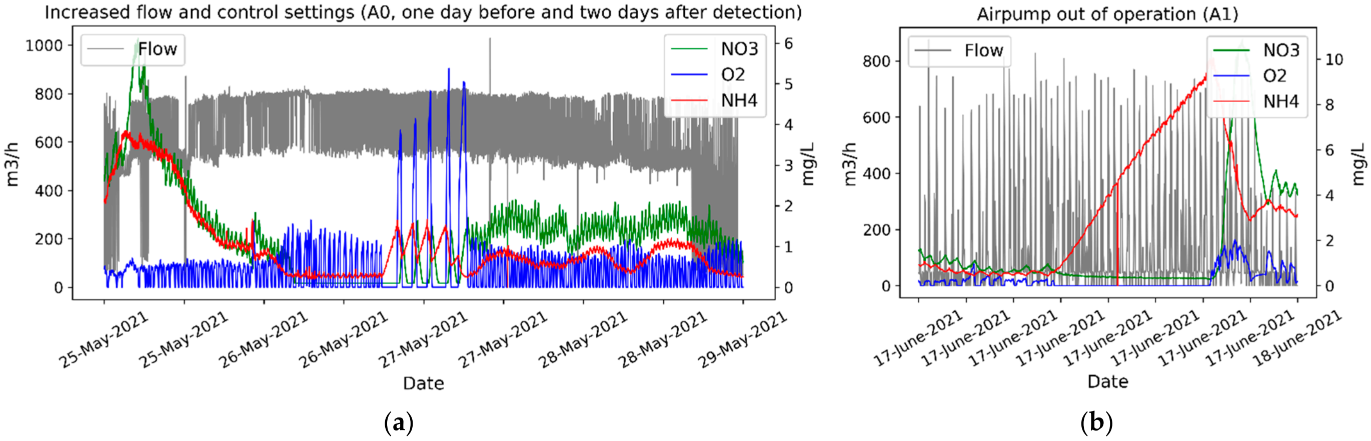


Figure 9. Examples of anomalies detected in WWTP3 PCT1. (**a**) Anomaly zero, increased flow prompting the operator to change the control settings. Only the 26 of May is detected as an anomaly, (**b**) Increased NH_4 levels due to lack of aeration.

3.5. WWTP3 PCT2

The data, measurements, and calibration as well as Moving LOF and detected anomalies for WWTP3 PCT2 are shown in Figure 10. In this PCT, anomaly one differs from previous observations. The pattern of the data indicates that the control settings were changed to reduce the NH_4 concentration in the outlet; however, for a while this entails that the air pump is constantly active as the NH_4 level does not decrease. Hereafter, a more normal pattern is observed again. A detailed plot of anomaly zero is presented in Figure 11. Another observation made for this PCT is a low concentration of DO, which is positive, as it indicates that all the DO has been used.

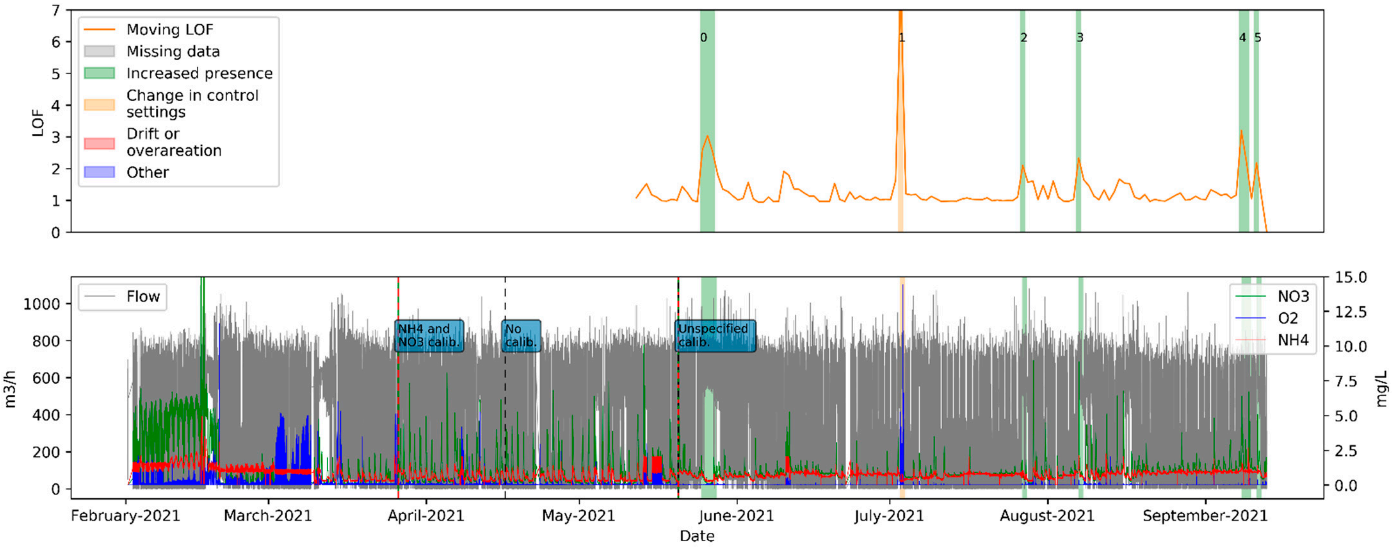


Figure 10. WWTP3 PCT1. In the upper graph the orange line shows the Moving LOF and the coloured areas, numbered from 0 to 5, show anomalies detected when the Moving LOF exceeded the threshold of two. The lower graph shows the flow, NO_3, DO and NH_4 and calibration data available from the plant.

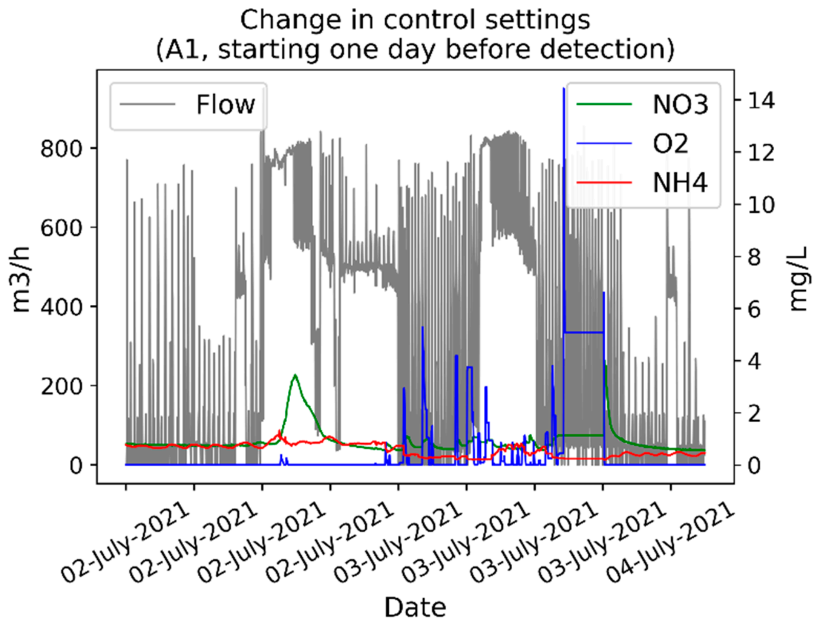


Figure 11. Example of an anomaly detected in WWTP3 PCT2. The anomaly is most likely caused by change in control settings.

3.6. General Observations

An overview of the observations and detailed descriptions for each of the PCTs is presented in Table 2.

Generally, it is worth noticing that the easiest drift to detect was the NH_4 sensor measuring too high values during alternating operation. In these cases, indications of drift could be visually observed in the data before the threshold was exceeded. However, reducing the threshold would also introduce more anomalies due to increased flow, which can be considered as false positives.

In several cases, faults such as drift were included in the data window used by the Moving LOF. Thereby faults would not be as abnormal for the Moving LOF as if a clean data set was available.

In some cases, increased amounts of NO_3 were detected indicating that the NO_3 sensor measured too high values. However, it was not possible to evaluate if the sensor measured within the accepted range of ± 1 mg/L.

In general, the algorithm did not detect drifts when the sensors measured too low of values. However, when reviewing the data manually there were indicators of the NO_3 sensor measuring too low values. Generally, drift towards low concentrations is harder to detect than drift towards high concentrations, as there is a natural limit in how much a drift towards zero can be distinguished from normal behaviours. Furthermore, reaching a low number of particles in the outlet of the plant is also an indicator of optimal operation of the plant.

Several cases of changes in control settings were detected as anomalies. Changes in control settings are not faults; however, they change the basis for any type of data-driven algorithm significantly.

Regarding missing data, it is worth noticing that this type of anomaly can easily be detected using rule-based methods. This type of anomaly was present several times but was not removed before applying the LOF algorithms, as daily values were based on average values for a given date.

Other Observations

The problem with NH_4 sensors measuring too high values is that this can entail plants over-aerating, which is expensive. An increase in multiple data parameters was observed when the NH_4 sensor measured too high values in alternating operation. Thereby it is possible that an increase in the daily average of concentrations could indicate NH_4 drift and that NH_4 sensor drift hereby could be detected by utilization of a simple rule-based algorithm, such as alarming, if a threshold is exceeded for a longer period. An overview of the average values per day for WWTP2 PCT1 can be seen in Figure 12.

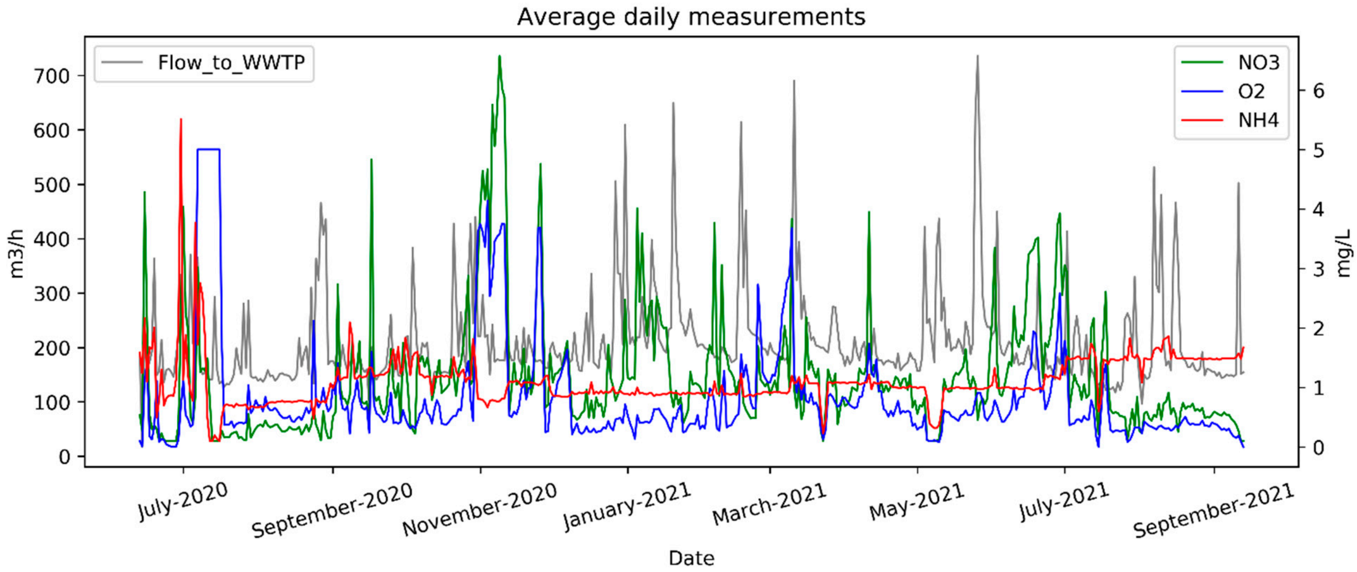


Figure 12. Average daily values of Flow, NH_4, NO_3 and DO in WWTP2 PCT1.

4. Discussion

This section contains a discussion of the results.

Based on initial tests described in Section 2, it was chosen to use LOF combined with a threshold. LOF was chosen despite several more complex methods having been previously presented in the literature, as it was not possible to apply the more complex methods to the data, due to low data quality in real WWTPs. The results showed that it was possible to detect anomalies; however, the most detected anomalies were related to increased presence of flow or substances. Increased presences of flow and substances are not faults. It is problematic that the most detected fault is increased in different parameters as previously published papers primarily focus on dry weather data, simulated data or in-control data, because this means that the developed methods do not encounter the challenges met at real plants. It is important to be aware of this shortcoming as it entails that automatically detecting outliers as faults most likely will entail that valid datapoints are considered as faults whereas actual faults are overlooked.

The results presented in this paper are highly dependent on the used threshold and the window length. If one lowers the threshold more anomalies would be detected; however, it would also entail a larger number of anomalies caused by increased flow or substances. For plants with more than one PCT, the number of anomalies caused by increased flow or increased presence of substances could be reduced by removing detections which are present in both PCT simultaneously, as can be seen in Tables 3 and 4. It is important to mention that increased presence of substance can both be caused by external factors, such as industrial discharges, which are anomalies, and internal factors such as sensor drifts, which are faults. This makes it complicated to distinguish between anomalies and faults.

Table 3 shows that if anomalies present in both PCTs are removed for WWTP2, the anomalies caused by increased flow and increased presence of other substances will change from 0.48% of the detected anomalies to 27% of the anomalies for PCT1 and from 79% to 63% for PCT2. For PCT1 sensor drift increases from 21% to 56% and for PCT 2 the percentage of anomalies caused by drift would increase from 4% to 17%. Due to the low number of anomalies for WWTP3, it would be misleading to make similar calculations for this plant.

Table 3. Overview of anomaly types detected in WWTP2 for the two PCTs. The table shows how many of the different types of anomalies are detected in total for each of the PCTs. Furthermore, it shows how many anomalies are detected if anomalies present in both PCTs are removed. The numbers in parentheses are the number of anomalies which have some overlap with the other PCT, but the period of the detections is not similar.

Anomaly Type	PCT1, All Detections	PCT1 Overlapping Detections Removed	PCT2, All Detections	PCT2 Overlapping Detections Removed
Increased flow	8	1 (1)	13	2 (4)
Increased presence incl. substances	5	2 (2)	6	3 (3)
Missing data	2	1 (1)	1	0 (1)
Drift	6	6	1	1
Change in control settings	1	(1)	0	0
Combinations of multiple types	3	1 (1)	1	1
Other	3	(3)	2	1 (1)

Table 4. Overview of anomaly types detected in WWTP3 for the two PCTs. The table shows how many of the different types of anomalies are detected in total for each of the PCTs. Furthermore, it shows how many anomalies are detected if anomalies present in both PCTs are removed. The numbers in parentheses are the number of anomalies which have some overlap with the other PCT, but the period of the detections is not similar.

Anomaly Type	PCT1, All Detections	PCT1 Overlapping Detections Removed	PCT2, All Detections	PCT2 Overlapping Detections Removed
Increased flow	0		3	(1)
Increased presence incl. substances	6	2 (1)	2	(1)
Change in control settings	1	(1)	1	1

It is important to mention that drifts primarily were detected in the PCT with alternating operation and for this plant several of the cases with drift could be detected earlier by visual inspection of the data, if good visualization tools were provided.

5. Perspectives and Recommendations

Drift in sensors, especially in NH_4 sensors, causes non-optimal operation at WWTPs, which can induce inefficient N removal, increased resource usage and extra economic costs. Sensor drift is common in most WWTPs and this needs to be handled for resource optimization. Several data-driven algorithms for drift detection have been developed in academia; however, they often remain as academic projects entailing a gap between academia and the real world. This gap was investigated by applying different data-driven solutions on real WWTPs. The results showed a number of significant challenges in real data, which have not been handled in the current academic solutions.

This study showed that it was not possible to obtain valid, consistent and precise labelling of sensor drift in the data after its collection. Only extreme sensor drifts and sensor drifts induing over aeration could be identified in PCTs with alternating operation. The study also showed that NH_4 drift, to some extent, can be identified using unsupervised learning. It also shows that more anomalies were detected in the WWTP with alternating operation than in the plants with PID control systems. However, the performance does not meet the needs at the WWTPs.

The challenges described in the Method and the Result sections clearly illustrate that action needs to be taken if optimal operation at the plants should be widespread among WWTPs. In the following, the challenges met at the plants are described and discussed in Sections 5.1–5.5. Section 5.6 contains a discussion of whether it is feasible to acquire a sufficient data set for data-driven drift detection. Section 5.7 contains perspectives on other ways of handling drift while Section 5.8 contains a discussion of why well considered data acquisition from the plants is still important. Section 5.9 contains perspectives on how the data available today can still create value.

5.1. Variations between Plants

There is a large variance between plants. For instance, there is a large variation in the design of WWTPs, the sensors installed at the plants, the composition of raw wastewater, the control strategies and the data stored from each plant. It is worth noticing that factors such as the composition of the wastewater can induce the NH_4 sensor to drift earlier in one plant than in another. The control methods and settings vary largely between different plants, due to factors such as variation in the discharge requirements of the plants.

From the perspective of a data scientist, it would be plausible to gain more knowledge on changes and abnormalities in the PCTs by comparing two PCTs in the same plant. This approach is not feasible from the perspective of water professionals as the tanks often have different control settings. However, it might be beneficial to compare the results of the anomaly detection. If an anomaly is detected in both PCTs simultaneously, the anomaly is most likely caused by surrounding factors and not faults in the sensors at the plant.

Like in other fields, such as maritime image recognition [26], large variation in the environment prevents formulation of specific general requirements; however, it is possible to discuss the main factors which need to be considered.

5.2. Control Settings

In situ changes in control settings at the WWTPs were largely observed in the data. Some of the changes were detected as anomalies but it was not always the case. In some cases, the changes were a consequence of drift in the NH_4 sensor. This is a practical solution at the plant and solves the present problem; however, it also introduces a bias in the data and makes faulty data normal. Furthermore, it was observed that sometimes, when the conditions at a plant using a PID-controller changed, the control settings were changed. This could, for instance, be due to increased flow. Some of the more extreme cases were visible for a human observer while it is uncertain if changes of less extreme character were present in the data. Change in control settings largely affects the patterns in the data, complicating development of data-driven solutions and in cases where it is found necessary to change the settings, it is essential that the changes are logged.

5.3. Logging Strategies

Missing and insufficient logging was a large challenge met in this study. In the cases where a log was available it solely contained information on measurement and calibration, and in some cases, it was not clear which sensors were calibrated due to unprecise documentations. Changes in control settings were never mentioned, despite being essential for the patterns in the data. The lag of logs at the plants is not solely a problem from a data science perspective, but it also makes it hard for newcomers to understand the plant, as they cannot see what has been done previously.

From a data science perspective, all lab tests, calibrations and change in control settings should be documented in a software system with constrained input parameters selected either from dropdown menus or check boxes, leaving solely numbers for manual entering. However, from the operator's point of view this can easily be considered as unnecessary bureaucracy. Therefore, the logging software should be as simple as possible while still providing sufficient data, and the operators should be included in the design and implementation processes and be able to see a benefit.

5.4. Data Quality

Multiple definitions of data quality can be found; however, the key element is that data is of high quality if it is 'fit for use' for the given purpose. Thereby data can have a high quality in one perspective while being of low quality from another perspective. Data need to contain a certain level of completeness, consistency, validity and timeliness, which all depend on the particular purpose [27,28].

This study shows that the information in the data available was insufficient for comprehensive drift detection in multiple sensors. Furthermore, due to a combination of missing logs and low resolution in the data, four out of seven data sets available for this study were not used.

Insufficient data quality is a problem in multiple other industrial cases. Despite companies collecting data with the purpose of using it, there is a high amount of data, which are collected without being actionable [28].

Prospectively, the authors suggest that data owners at utilities and municipalities consider what they wish to gain from their data and, based on this, select which data to store and what the resolution should be for the data to contain sufficient information. It is possible that other factors not directly connected to the content of the plant such as energy usage and cost at a given time could be relevant factors for benchmarking the performance of the plant. Generally, it is important that the pattern in the data is relevant. Changes in patterns can occur by change in control settings, sensor drift, change in the catchment area etc. From a data quality perspective, the changes in patterns should be minimized, and when they occur, they should be well documented. In case a lot of information needs manual entry, it could be considered to use a well-defined user interface, to reduce faults in the manual documentation and increase the precision of the data. Generally, data should be easily accessible and interpretable [27]. In this connection it is important to ensure coherent naming of parameters, etc. For more information on data quality, please refer to Mahanti [27].

5.5. Learning Algorithms

Development of data-driven drift detection in treatment plants is complicated since the control systems are based on feedback loops. Thereby the system is automatically adjusted to the drift, minimizing the changes in faulty data compared to correct data. Furthermore, constant concentration levels in the outlet, where the NH_4 and NO_3 sensors are placed, are considered optimal; however, a constant value further decreases the level of information in the data. As there is a large uncertainty in the composition of the wastewater arriving at the plant, it can be difficult to distinguish between natural variations and sensor drift from a data perspective. A solution to this could be sensors located at the inflow. This would give the possibility of performance evaluation, etc.; however, it would also result in more sensors to maintain.

Due to the costs of sensors, it is often not feasible to implement additional sensors. Therefore, when selecting sensors in WWTPs and deciding which parameters to store, it is important to consider the indirect information in sensors and potential use cases. For instance, Thürlimann et al. [29] suggested a soft sensor using the pH in the inlet and the outlet to detect NH_4 peak load events. Another parameter worth considering in the future is the airflow. The correlation between the airflow and the DO most likely contains usable information of the processes in the plant.

5.6. Data-Driven Drift Detection–Is It Worth It?

Due to large variations between plants, it is necessary to acquire a high-quality data set for each plant and subsequently adjust the model to the plant. Acquiring the data set entails that the operators systematically measure and calibrate sensors. Furthermore, the control settings should not be changed and if they need change due to external factors, this should be documented. With such a high-quality data set it is possible to detect faults in the plant [23]. If the catchment area of the plant is changed or if the control system needs updates, for instance due to better algorithms, the data acquisition needs to be remade. This means that the operators need to be systematic in the operation of the plant for several months, or preferably a year, every time a change is made. Lab measurements are easy to perform, and the biggest obstacle is to obtain a culture among the operators where lab measurements are performed instead of ad hoc adjustment of the control settings. A utility that can acquire the needed data set might already have obtained a culture of high-quality sensor maintenance, making data-driven drift detection redundant.

5.7. Other Ways of Handling Drift

The above statement yields a need for higher quality in sensor data at wastewater plants. This is especially relevant for the sensors which record data that are used by the plant's control system. Data quality can be obtained by regular monitoring, calibration and cleaning. Other approaches include self-calibrating sensors and soft sensors. It could also be argued that in some cases, multiple sensors of the same type could be used for drift detection; however, as the sensors would be in the same environment, they would also be affected by the same environmental factors such as fouling or drift after heavy rain or high NH_4 levels. Contemporary, ion-selective sensors are widely used as they are cheap to operate. Another solution could be to use sensors based on gas chromatography for quality control. This sensor uses chemicals and measure once an hour; however, this would be an expensive solution.

A different approach to manage drift could be to include more rules in the control strategy, for instance by stopping aeration if the NO_3 level does not increase or by finding the actual NH_4 level by aerating until the NH_4 level does not decrease more during night-time.

5.8. Why Well Considered Data Acquisition from Plants Is Still Necessary

Increased focus on the SDGs emphasizes that the utilities optimize the operations at the plants by reducing energy usage and lowering greenhouse gas emissions while ensuring a high degree of N removal. However, to benchmark performance of experiments performed to optimize the performance, the general performance of the plant needs to be known. Newhart et al. [10] stated that it is essential to a define the problem scope and desired goals when integrating data-driven control at WWTPs. This can be generalized to other tasks involving data-driven solutions.

5.9. Can Low Quality Data Still Create Value?

Data quality is a relative concept, and it is related to the purpose of the data [27]. Therefore, the data can be of high quality if used for other purposes. For instance, comparing the available parameters for a given day with average values of the previous days, days with similar flow or similar weekdays can give information to water professionals and help them evaluate the operation of the plant. If available, the energy usage can give information on the effectiveness of the operation of the plant. Furthermore, comparing the average price of the energy used at the plant a given day to the average energy price the same day can give information on how sustainable the energy usage is, as low energy prices are often related to a surplus production of green energy. This is relevant as it can help operators evaluate and optimize the control strategy of the plant and thereby contribute to a more holistic cross sectorial optimization, which is essential to obtain smart cities.

6. Conclusions

Sensor drifts are widely present in WWTPs and can result in less efficient operation at the plants. Several approaches for solving this problem have recently been proposed and documented in academia; however, the studies rarely reflect the conditions at real treatment plants and thereby remain as academic projects. The aim of this study was to investigate this gap between academia and practice by applying algorithms suggested in academia on data from real WWTPs. The results showed that obtaining a robust and valid model for fault detection is challenged by several factors such as low data quality, missing logging and in situ changes of control settings. The most often detected anomalies were related to increased flow or increased concentrations, which can be hard to distinguish from sensor drift. It is the author's interpretation that better algorithms and results could be obtained by increased focus on the data quality by including well-considered data management, logging strategies and consistency in the control settings of the WWTP. However, if a utility can obtain such a data set, the problems with drift might already have been solved. Other solutions to handle sensor drift include implementation of improved sensors for quality control, self-calibrating sensors and soft sensors based on informative parameters.

While the data quality might not be sufficient for automatic drift detection, the quality might be sufficient for statistical purposes, which can contribute to information for water professionals and help them evaluate the performance of the plant.

Author Contributions: Conceptualization B.D.H., T.B.M. and D.G.J.; methodology, B.D.H., T.B.M. and D.G.J.; software, B.D.H.; validation, T.B.H., T.B.M. and D.G.J.; formal analysis, B.D.H.; investigation, B.D.H.; resources, B.D.H.; data curation, B.D.H.; writing—original draft preparation, B.D.H.; writing—review and editing, B.D.H., T.B.H., T.B.M., and D.G.J.; visualization, B.D.H.; supervision, T.B.H., T.B.M. and D.G.J.; project administration, B.D.H., and D.G.J.; funding acquisition, B.D.H., T.B.M. and D.G.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Innovation Fund Denmark.

Acknowledgments: The authors wish to thank Lars Lading from EnviDan A/S for his assistance with accessing data and supervision.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, W.; Tooker, N.B.; Mueller, A.V. Enabling Wastewater Treatment Process Automation: Leveraging Innovations in Real-Time Sensing, Data Analysis, and Online Controls. *Environ. Sci. Water Res. Technol.* **2020**, *6*, 2973–2992. [CrossRef]
2. Santín, I.; Pedret, C.; Vilanova, R.; Meneses, M. Advanced Decision Control System for Effluent Violations Removal in Wastewater Treatment Plants. *Control Eng. Pract.* **2016**, *49*, 60–75. [CrossRef]
3. Stentoft, P.A.; Vezzaro, L.; Mikkelsen, P.S.; Grum, M.; Munk-Nielsen, T.; Tychsen, P.; Madsen, H.; Halvgaard, R. Integrated Model Predictive Control of Water Resource Recovery Facilities and Sewer Systems in a Smart Grid: Example of Full-Scale Implementation in Kolding. *Water Sci. Technol.* **2020**, *81*, 1766–1777. [CrossRef] [PubMed]
4. Corominas, L.; Garrido-Baserba, M.; Villez, K.; Olsson, G.; Cortés, U.; Poch, M. Transforming Data into Knowledge for Improved Wastewater Treatment Operation: A Critical Review of Techniques. *Environ. Model. Softw.* **2018**, *106*, 89–103. [CrossRef]
5. Samuelsson, O.; Olsson, G.; Lindblom, E.; Björk, A.; Carlsson, B. Sensor Bias Impact on Efficient Aeration Control during Diurnal Load Variations. *Water Sci. Technol.* **2021**, *83*, 1335–1346. [CrossRef] [PubMed]
6. Samuelsson, O.; Björk, A.; Zambrano, J.; Carlsson, B. Fault Signatures and Bias Progression in Dissolved Oxygen Sensors. *Water Sci. Technol.* **2018**, *78*, 1034–1044. [CrossRef] [PubMed]
7. Teh, H.Y.; Kempa-Liehr, A.W.; Wang, K.I.-K. Sensor Data Quality: A Systematic Review. *J. Big Data* **2020**, *7*, 11. [CrossRef]
8. Baklouti, I.; Mansouri, M.; Hamida, A.B.; Nounou, H.; Nounou, M. Monitoring of Wastewater Treatment Plants Using Improved Univariate Statistical Technique. *Process Saf. Environ. Prot.* **2018**, *116*, 287–300. [CrossRef]
9. Thomann, M.; Rieger, L.; Frommhold, S.; Siegrist, H.; Gujer, W. An Efficient Monitoring Concept with Control Charts for On-Line Sensors. *Water Sci. Technol.* **2002**, *46*, 107–116. [CrossRef] [PubMed]
10. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, *157*, 498–513. [CrossRef] [PubMed]
11. Baggiani, F.; Marsili-Libelli, S. Real-Time Fault Detection and Isolation in Biological Wastewater Treatment Plants. *Water Sci. Technol.* **2009**, *60*, 2949–2961. [CrossRef] [PubMed]
12. Alferes, J.; Tik, S.; Copp, J.; Vanrolleghem, P.A. Advanced Monitoring of Water Systems Using in Situ Measurement Stations: Data Validation and Fault Detection. *Water Sci. Technol.* **2013**, *68*, 1022–1030. [CrossRef] [PubMed]
13. Cheng, T.; Dairi, A.; Harrou, F.; Sun, Y.; Leiknes, T. Monitoring Influent Conditions of Wastewater Treatment Plants by Nonlinear Data-Based Techniques. *IEEE Access* **2019**, *7*, 108827–108837. [CrossRef]
14. Huang, F.; Shen, W.; Liu, Z. Applications of Sub–Period Division Strategies on the Fault Diagnosis with MPCA for the Biological Wastewater Treatment Process of Paper Mill. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; IEEE: Guangzhou, China, 2019; pp. 5138–5143.
15. Ba-Alawi, A.H.; Vilela, P.; Loy-Benitez, J.; Heo, S.; Yoo, C. Intelligent Sensor Validation for Sustainable Influent Quality Monitoring in Wastewater Treatment Plants Using Stacked Denoising Autoencoders. *J. Water Process Eng.* **2021**, *43*, 102206. [CrossRef]

16. Kazemi, P.; Giralt, J.; Bengoa, C.; Masoumian, A.; Steyer, J.-P. Fault Detection and Diagnosis in Water Resource Recovery Facilities Using Incremental PCA. *Water Sci. Technol.* **2020**, *82*, 2711–2724. [CrossRef] [PubMed]
17. Kazemi, P.; Bengoa, C.; Steyer, J.-P.; Giralt, J. Data-Driven Techniques for Fault Detection in Anaerobic Digestion Process. *Process Saf. Environ. Prot.* **2021**, *146*, 905–915. [CrossRef]
18. Luca, A.-V.; Simon-Várhelyi, M.; Mihály, N.-B.; Cristea, V.-M. Data Driven Detection of Different Dissolved Oxygen Sensor Faults for Improving Operation of the WWTP Control System. *Processes* **2021**, *9*, 1633. [CrossRef]
19. Mali, B.; Laskar, S.H. Incipient Fault Detection of Sensors Used in Wastewater Treatment Plants Based on Deep Dropout Neural Network. *SN Appl. Sci.* **2020**, *2*, 2121. [CrossRef]
20. Xu, C.; Huang, D.; Li, D.; Liu, Y. Novel Process Monitoring Approach Enhanced by a Complex Independent Component Analysis Algorithm with Applications for Wastewater Treatment. *Ind. Eng. Chem. Res.* **2021**, *60*, 13914–13926. [CrossRef]
21. Klanderman, M.C.; Newhart, K.B.; Cath, T.Y.; Hering, A.S. Fault Isolation for a Complex Decentralized Waste Water Treatment Facility. *J. R. Stat. Soc. C* **2020**, *69*, 931–951. [CrossRef]
22. Mamandipoor, B.; Majd, M.; Sheikhalishahi, S.; Modena, C.; Osmani, V. Monitoring and Detecting Faults in Wastewater Treatment Plants Using Deep Learning. *Environ. Monit. Assess.* **2020**, *192*, 148. [CrossRef] [PubMed]
23. Cecconi, F.; Rosso, D. Soft Sensing for On-Line Fault Detection of Ammonium Sensors in Water Resource Recovery Facilities. *Environ. Sci. Technol.* **2021**, *55*, 10067–10076. [CrossRef] [PubMed]
24. Anter, A.M.; Gupta, D.; Castillo, O. A Novel Parameter Estimation in Dynamic Model via Fuzzy Swarm Intelligence and Chaos Theory for Faults in Wastewater Treatment Plant. *Soft Comput.* **2020**, *24*, 111–129. [CrossRef]
25. Breunig, M.M.; Kriegel, H.-P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.* **2000**, *29*, 93–104. [CrossRef]
26. Pedersen, M.; Madsen, N.; Moeslund, T.B. No Machine Learning without Data: Critical Factors to Consider When Collecting Video Data in Marine Environments. *J. Ocean Technol.* **2021**, *16*, 21–30.
27. Mahanti, R. *Data Quality: Dimensions, Measurement, Strategy, Management, and Governance*; ASQ Quality Press: Milwaukee, WI, USA, 2018; ISBN 9780873899772.
28. Scarisbrick-Hauser, A.; Rouse, C. The Whole Truth and Nothing but the Truth? The Role of Data Quality Today. *Direct Mark. Int. J.* **2007**, *1*, 161–171. [CrossRef]
29. Thürlimann, C.M.; Dürrenmatt, D.J.; Villez, K. Soft-Sensing with Qualitative Trend Analysis for Wastewater Treatment Plant Control. *Control Eng. Pract.* **2018**, *70*, 121–133. [CrossRef]