Journal of
Hydroinformatics

© 2024 The Authors

Journal of Hydroinformatics Vol 26 No 2, 441 doi: 10.2166/hydro.2024.167

LSTM-based autoencoder models for real-time quality control of wastewater treatment sensor data

Siddharth Seshan[a,*], Dirk Vries[a], Jasper Immink[a], Alex van der Helm[b] and Johann Poinapen[a]

[a] KWR Water Research Institute, Nieuwegein, The Netherlands
[b] Waternet, Amsterdam, The Netherlands
*Corresponding author. E-mail: siddharth.seshan@kwrwater.nl

SS, 0000-0001-8830-3601; DV, 0000-0003-2920-5926; JI, 0000-0003-1251-5442; AvdH, 0009-0005-2777-211X; JP, 0000-0002-1937-7292

ABSTRACT

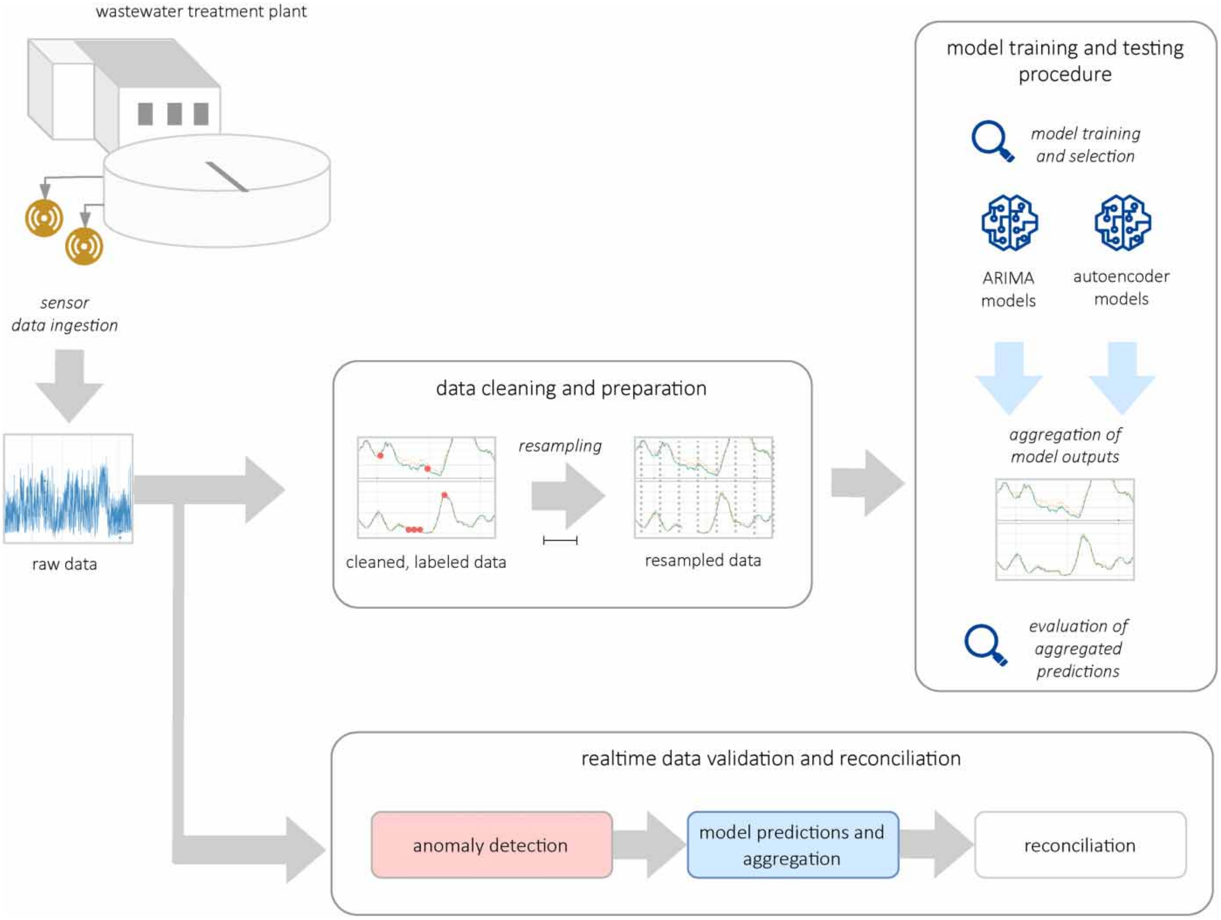
The operation of smart wastewater treatment plants (WWTPs) is increasingly paramount in improving effluent quality, facilitating resource recovery and reducing carbon emissions. To achieve these objectives, sensors, monitoring systems, and artificial intelligence (AI)-based models are increasingly being developed and utilised for decision support and advanced control. Key to the adoption of advanced data-driven control of WWTPs is real-time data validation and reconciliation (DVR), especially for sensor data. This research demonstrates and evaluates real-time AI-based data quality control methods, i.e. long short-term memory (LSTM) autoencoder (AE) models, to reconcile faulty sensor signals in WWTPs as compared to autoregressive integrated moving average (ARIMA) models. The DVR procedure is aimed at anomalies resulting from data acquisition issues and sensor faults. Anomaly detection precedes the reconciliation procedure using models that capture short-time dynamics (SD) and (relatively) long-time dynamics (LD). Real data from an operational WWTP are used to test the DVR procedure. To address the reconciliation of prolonged anomalies, the SD is aggregated with an LD model by exponential weighting. For reconciling single-point anomalies, both ARIMA and LSTM AEs showed high accuracy, while the accuracy of reconciliation regresses quickly with increasing forecasting horizons for prolonged anomalous events.

Key words: anomaly detection, ARIMA models, autoencoder models, data reconciliation, wastewater treatment plant data

HIGHLIGHTS

- A new methodology is proposed for the real-time validation of sensor data in wastewater treatment by the aid of anomaly detection and the subsequent reconciliation of sensor signals using a short timescale dynamic and a long timescale dynamic model.
- LSTM-based autoencoder models are used to reconcile anomalous data.
- Deep neural network-based models are compared with conventional time series modelling.

GRAPHICAL ABSTRACT


wastewater treatment plant
sensor data ingestion
raw data
data cleaning and preparation
cleaned, labeled data
resampling
resampled data
model training and testing procedure
model training and selection
ARIMA models
autoencoder models
aggregation of model outputs
evaluation of aggregated predictions
realtime data validation and reconciliation
anomaly detection
model predictions and aggregation
reconciliation

1. INTRODUCTION

The design and operation of wastewater treatment plants (WWTPs) are becoming increasingly complex due to the need to (i) improve effluent quality in order to comply with increasingly stringent standards such as the European Directive 91/271 (1991) (EEC) on urban wastewater, (ii) increase resource recovery activities and (iii) decrease the carbon footprint. To meet these objectives and enable advanced process monitoring and the control of sewage systems and treatment processes, data acquired from sensors are being increasingly used, and monitoring systems, software sensors, models and controllers are added or upgraded to bridge the gap between current practice and data-driven, smart water systems (Therrien *et al.* 2020).

In order to resolve the operational challenges related to meeting effluent quality criteria, while also reducing the carbon footprint and operational costs, the operation of WWTPs and, more generally, urban wastewater systems (UWS) is undergoing a transition from an industry 3.0 level where IT systems are set and process control is automated to an industry 4.0 level where the operation is getting smart and (semi) autonomous (Fernando *et al.* 2022). The industry 4.0 standard entails a data-driven approach, i.e. the use of artificial intelligence (AI) and machine learning (ML) models for decision support and advanced control. It also involves creating a strong connection between smart devices (such as Internet of Things (IoT)) and the processing of UWS including WWTP data, weather data and other external data sources to enable such a transition. However, the transition to a fully data-driven WWTP or UWS is not without hurdles: connectivity should be guaranteed as well as (cyber) security, management, availability and quality of data (Jagatheesaperumal *et al.* 2022), while UWS are inherently considered as complex, non-linear systems. More specifically, UWS are very dependent on environmental and operational conditions, exhibit seasonal dynamics (Daelman *et al.* 2015; Di Marcantonio *et al.* 2020), extreme weather events (Md Nor *et al.* 2020; Park *et al.* 2020) and often have cyclic behaviour due to recycling streams. Yet, it is recognised that the operation of WWTPs has evolved from sensor signal processing and using statistics and process identification in the seventies to knowledge-based systems where sensor data are increasingly fed into data mining techniques and predictive analytics in order to capture complexity. The aim is to support plant-wide control and decision-making that is optimized based on possibly multiple criteria (Poch *et al.* 2014). In this transition phase, perhaps one of the most important, first steps in the realisation of a smart, data-driven UWS is data validation and reconciliation (DVR) of critical sensor signals because the quality of data is frequently hampered by the intrinsically challenging measurement conditions in wastewater. Anomalous data stem from (Gaddam *et al.* 2020) either (1) intermittent sensor errors, such as communication and IT-related problems between sensors in the data acquisition system itself, leading to sensor data loss, and duplicate data entries, or (2) inherent sensor faults caused by fouling, aging or miscalibration, which can lead to offset, drifts, increased noise levels or freezing (flatlines) in the measurements. Additionally, (3) sensor events such as process faults or irregularities in the process behaviour can occur due to extreme weather events or maintenance activities of UWS assets. Manually curating data and ensuring its quality would be prohibitively time-consuming and expensive; hence, automated data quality control is highly needed.

This work focuses on the development and assessment of AI-based anomaly detection techniques in combination with the reconciliation of faulty sensor data. A comprehensive review of advanced, AI-based data validation is given in Liu *et al.* (2023) where validation techniques are assessed from two points of view, i.e. the detection of process faults and instrumentation faults, which are related to the data collection (sensors) and acquisition systems (information and communication technologies (ICT)). With respect to the use of AI, or more specifically ML and deep learning techniques in data validation applications, quite some attention is being given to deep neural networks (DNNs), such as the long short-term memory (LSTM), convolutional neural networks and autoencoder (AE) models as promising techniques for the detection and reconstruction of faulty sensor signals. In Ba-Alawi *et al.* (2021), a denoising AE (D-AE) model is described that detects, identifies and reconciles faulty data based on real WWTP data in South Korea. The model is compared with a conventional principal component analysis (PCA) procedure, which is typically useful to detect process anomalies when using multiple sensor signals and is also known as being crucial in multivariate statistical process control theory. The same authors compared the performance of a deep residual network structure-based variational AE to impute missing sensor data with other neural network AE models and PCA for a WWTP data case (Ba-Alawi *et al.* 2022). In Pisa *et al.* (2020), an LSTM-based control strategy of a WWTP is coupled with three ML-based denoising techniques, including D-AEs, and model performance was compared to the performance of autoregressive integrated moving average (ARIMA) and PCA models using a benchmark WWTP model. Again, the best denoising was achieved with AE models.

The main aim of this work is to demonstrate and assess anomaly detection and reconciliation procedures using real plant data with a focus on anomalies from category 1 (faults stemming from data acquisition problems, e.g. missing data) and category 2, i.e. faults in sensor data streams. More specifically, the research objectives are to

- illustrate statistical anomaly detection techniques that will provide a basis for tagging the data, which is to be reconciled by data-driven models;
- demonstrate a novel reconciliation procedure, where a model that captures SD (hours) with high accuracy is combined with a model that captures LD (day); and
- compare and evaluate the performance of an LSTM-based AE with an ARIMA model for the reconciliation of the identified anomalous sensor data.

2. METHODOLOGY

2.1. Overview

In Figure 1, an overview of the methodology employed in this study is illustrated. Raw data of 1-min frequency acquired from online sensors from a full-scale WWTP were used. For the purpose of calibrating the ARIMA models and training the LSTM-based AEs, the datasets amounting to various months were pre-processed. Initially, the datasets were cleaned by addressing anomalies that were identified using the anomaly detection methods outlined in the proposed data validation methodology. Based on the knowledge of the processes represented in the parameters measured by the sensors, decisions were made on handling the identified anomalies. Erratically occurring and singular erroneous values were removed and filled using linear interpolation. For anomalous events of longer duration, for 12 h or more, the period was removed from the dataset altogether. This pre-processing step was conducted to ensure that the datasets were representative of process

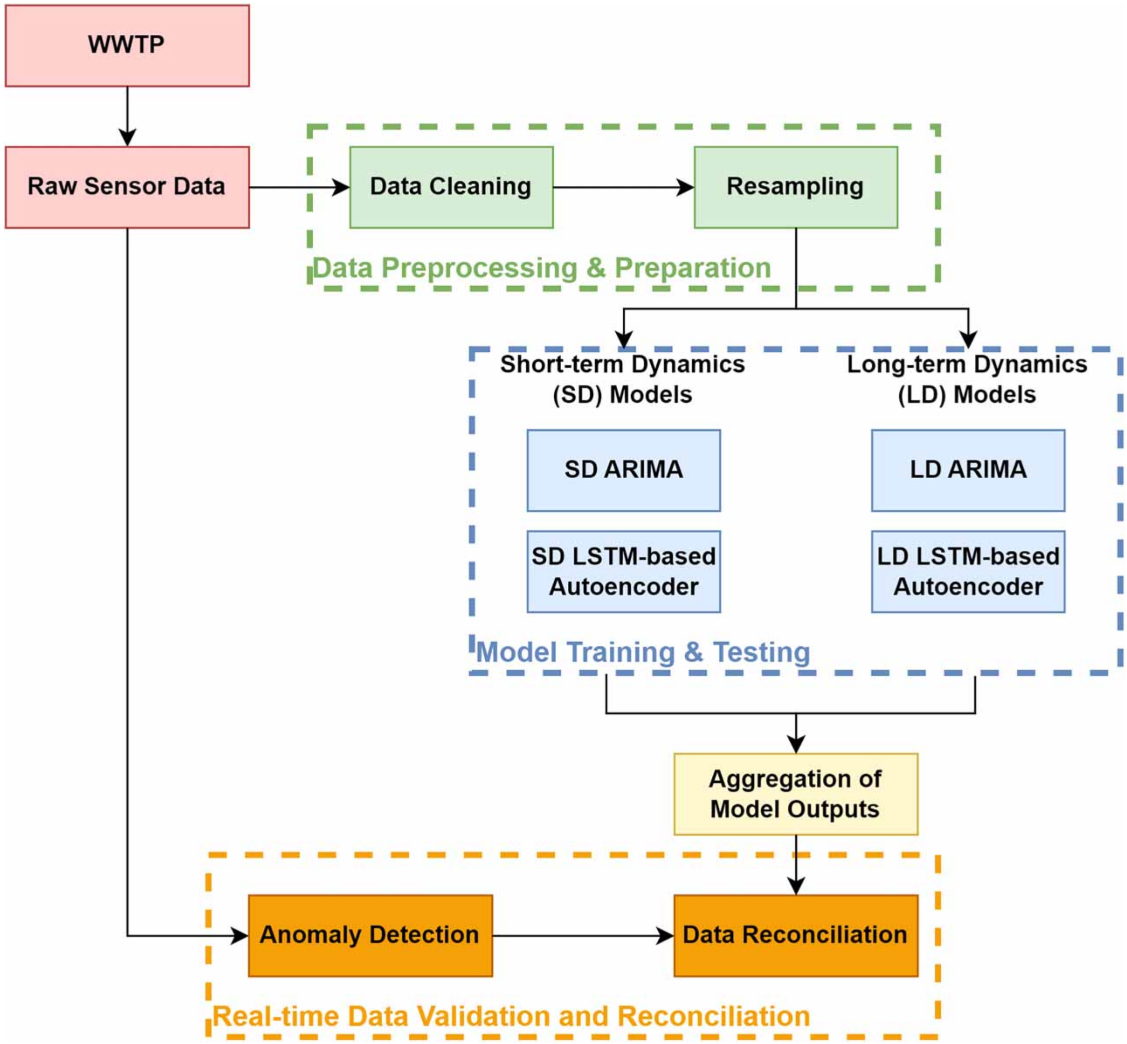

WWTP
Raw Sensor Data
Data Cleaning
Resampling
Data Preprocessing & Preparation
Short-term Dynamics (SD) Models
Long-term Dynamics (LD) Models
SD ARIMA
LD ARIMA
SD LSTM-based Autoencoder
LD LSTM-based Autoencoder
Model Training & Testing
Aggregation of Model Outputs
Anomaly Detection
Data Reconciliation
Real-time Data Validation and Reconciliation

Figure 1 | Overview of methodology employed in this study.

behaviour under normal operations and were the cleanest data available for model training, considering the models would subsequently be used within the data reconciliation process.

Furthermore, the datasets were then resampled into two granularities, which were determined based on the two most dominant time constants expected from the process. Models were then individually trained for the two resampled datasets. The computational time needed for model training was also considered. The dataset was then split into a training set and a test set, using an 80/20 ratio, where the models were trained on the larger set and evaluated on the smaller set, which represented the unseen data.

2.2. Anomaly detection

By definition, an anomaly is a data point that is likely to be erroneous and can be caused by the process or instrumentation irregularities. The detection of anomalies is considered as a classification problem, i.e. making the distinction between data that are likely faulty or correct. The techniques used in this application are statistical methods to flag gross anomalies within a dataset. Rule-based methods were developed that are generic and system-agnostic. The methods require the provision of certain metadata on the rules that are specific to the sensor signals. The anomaly detection methods can be subdivided into the detection of single-value anomalies and contextual anomalies. With regard to the single-value anomalies, faulty single measured values are detected, while for the contextual anomalies, anomaly regions of data are detected, in other words, consecutive faulty measured values over time. Through the execution of the anomaly detection methods, each sensor data point is provided with a flag or label providing information on whether the data point is considered an anomaly. All detection methods were implemented using the Python libraries pandas, Anomaly Detection Toolkit (ADTK) (Arundo Analytics Inc. 2020), and NumPy. Below is the description of the methods developed and incorporated within the data validation scheme:

- *NaN Value Detection:* The detection of any NaN (not a number) values that are present.
- *Zero Value Detection:* The detection of zero values for variables, where it is considered unfeasible to have zero values.
- *Negative Value Detection:* The detection of negative values in the dataset is considered unfeasible.
- *Threshold Detection:* The detection of values that are above or below given thresholds. The values of the thresholds are provided by the user (such as process operations and technologists) and are specific for a given variable and sensor.
- *Flatline Detection:* This method is to detect a form of contextual anomalies. It is a situation when consecutive observations have an equal value for an unfeasible duration of time. The detection of the flatlines is conducted using a user-defined minimum threshold value for the number of measurement points and a user-defined deviation. The number of data points that have the same value, or in other words, the length of a potential flatline, is compared with the threshold length to determine if a flatline is present.
- *Spike Detection:* Sudden spikes or drops with regard to previous data are detected by calculating a list of differences between all data points and their consecutive data points and calculating the mean μ and standard deviation σ for these differences. Should a difference value be further away than a threshold value ϵ_{sd} from μ, in units of σ, then the value is flagged as being anomalous. Spikes have been defined as a point that is preceded by an anomalous rise and followed by an anomalous fall, or vice versa.

2.3. Data reconciliation

The anomaly detection methods provide prior labelling of sensor data as being faulty or good. In the proposed data validation methodology, a data correction protocol is provided to handle the identified anomalies. This is conducted through the reconciliation of the sensor data with the outputs of the trained model. The values flagged as anomalies can be removed and replaced with the prediction values from the model. This leads to a reconciled signal, which is a combination of the original sensor data and the replaced values containing the predictions from a model. In this study, data reconciliation is achieved by training LSTM-based AEs. Furthermore, simpler statistical-based models called ARIMAs were used as benchmark models to compare with the more complex AE models for the purpose of conducting data reconciliation. In the following subsections, the ARIMA and LSTM-based AE models, along with the data reconciliation procedure, are described.

2.3.1. ARIMA as benchmark models

In this work, predictions of time-dependent water quality data measured at an operational full-scale plant are quantified and compared. To be able to adequately rate the predictions, a benchmark model is introduced to compare with the predictions of different models. For this purpose, ARIMA-style models are used (Box & Jenkins 1970). These are more generalised versions of the autoregressive moving average (ARMA) models, which are useful for predicting future values in a given time series (Mehdizadeh 2020; Moon *et al.* 2021). The ARIMA model is a three-part composite model: (i) an autoregressive (AR) part, (ii) an integrative (I) part, and (iii) a moving average (MA) part. The AR section is characterised by hyperparameter p, which defines the amount of data points in its immediate history that this section of the model considers. This AR model then determines, for each point i, which values in its immediate history have predictive power for target values, and also determines coefficients that quantify their importance for the final prediction. The AR model is described as:

$$\begin{aligned} Y_{t,\,\mathrm{AR}} &= \mu + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \epsilon \\ Y_{t,\,\mathrm{AR}} &= \mu + \epsilon + \sum_{i=1}^{p} \beta_i Y_{t-i} \end{aligned} \tag{1}$$

where $Y_{t,\,\mathrm{AR}}$ is the time series value at time t as predicted by the AR part of the model, μ is the estimated mean for this time series (intercept), β_i is the coefficient for lag time i that the AR model calculates and ϵ is a normally distributed error function that adds noise to the model. Model training comprises β_i- and μ-value determination.

The MA section is characterised by an input parameter q that also defines the number of data points in its immediate history that the ARIMA model considers. It is a linear combination of the errors of all q historical values:

$$Y_{t,\,\mathrm{MA}} = \mu + \beta_1' a_{t-1} + \beta_2' a_{t-2} + \ldots + \beta_q' a_{t-q} + \epsilon$$
$$Y_{t,\,\mathrm{MA}} = \mu + \epsilon + \sum_{i=1}^{q} \beta_i' a_{t-i} \tag{2}$$

where $Y_{t,\,\mathrm{MA}}$ is the time series value at time t as predicted by the MA part of the model, and β_i' is the coefficient that the MA model calculates. Furthermore, a_i is the lagged error from the AR model:

$$a_i = Y_{i,\,\mathrm{AR}} - Y_i \tag{3}$$

Model training comprises β_i'- and μ-value determination.

Finally, the integrative (I) part of the model is defined by the parameter d, which defines the degree of differentiation of the data before performing the AR and MA models. With time series data, first-order differentiation ($d = 1$) is defined as:

$$Y_t = y_t - y_{t-1} \tag{4}$$

with Y_t the differentiated data point at time t, and y_t the original data point at time t. In higher-order differentiation, this results in:

$$Y_t^{(d)} = \sum_{i=0}^{d-1} (-1)^i \binom{d-1}{i} y_{t-i} \tag{5}$$

with $Y_t^{(d)}$ the d-order differentiated data point at time t. The full ARIMA (p, d, q) model is a linear combination of the two AR (p) and MA (q) predictions, given that their input data were differentiated d times.

$$Y_{t,\,\mathrm{full}} = \mu + \epsilon + \sum_{i=1}^{p} \beta_i Y_{t-i}^{(d)} + \sum_{i=1}^{q} \beta_i' a_{t-i}^{(d)} \tag{6}$$

Here, $Y_i^{(d)}$ and $a_i^{(d)}$ are the time series values obtained by differentiation with order d.

2.3.2. LSTM-based AEs

2.3.2.1. Model description. An AE model is a special case of a feedforward neural network, consisting of two modules: an encoder and a decoder (Provotar *et al.* 2019). The encoder is trained for a given input of sensor data, learning the underlying features that are typically represented in a reduced dimension (Bengio 2009; Provotar *et al.* 2019). The decoder reconstructs the input data using the encoded outputs. Hence, the target of an AE is the input itself. Conventionally, AEs are trained using a single layer each for the encoder and the decoder. However, the use of deep AEs or recurrent neural network (RNN)-based layers can provide various benefits, such as yielding better compression of the data and, subsequently, better reconstruction of the input (Provotar *et al.* 2019).

The AE models trained in this study contained a combination of LSTM and dense layers. LSTM layers, introduced by Hochreiter & Schmidhuber (1997), are a subset of RNN layers that are efficient in learning long-term dependencies within data and are used frequently in predictive modelling. An LSTM unit is more complex than a conventional dense layer unit because it includes various gates to regulate the temporal information flow. A standard structure of an LSTM cell can be found in Figure S1 of the Supplementary Information. An LSTM cell typically consists of a forget gate (f) that determines how much information from the previous hidden and cell state will be removed, and an input gate (i) that decides how much of the current information will be kept after the current cell state is updated using the cell update gate (C), and finally an output gate (o) that controls the information that will be outputted based on the internal cell state. All the gates are defined as linear relationships while considering the inputs provided, recurrent information from previous cell states, weights of the gates and the corresponding bias terms. Typically, in an LSTM cell, a sigmoid activation function is used for recurrent related information, and the hyperbolic tangent function is used when updating the cell state.

2.3.2.2. Model training. To identify trained LSTM-based AE models providing accurate predictions, an orthogonal approach of fine-tuning hyperparameters was conducted as follows. After a training procedure is completed, first, the learning curve, which visualises the training and test loss, is inspected. This provided information on overfitting, the training dynamics over the epochs, and some indication of the convergence of the model to a stable state. An example learning curve for the LSTM-based AE models is provided in Figures S2 and S3. Subsequently, the predictions from the model on the training and test sets are visualised, and performance metrics are calculated to assess the model's performance and generalisation capabilities.

Initially, model trainings were done to identify the number of layers and an optimal model structure. While considering a trade-off between computational effort and accuracy, a combination of LSTM layers and dense layers was considered. This provided an increased number of trainable weights while reducing the number of memory-based units that could lead to overfitting issues. Additionally, LSTM layers required a 3D array of data as input. A key hyperparameter related to such an input is the dimension that represents the number of timesteps, or in other words, the amount of historical data inputted into the model to make predictions. This was determined and fixed based on the sensor signal behaviour and on prior knowledge of the wastewater treatment process dynamics that are prevalent. Further details on the choices made for this investigation are provided in Section 3.1.

To overcome the challenges of overfitting, based on the performance of the test dataset, a regularisation technique known as Dropout (Hinton *et al.* 2012) was implemented. In dropout regularisation, certain neurons in a layer are randomly ignored during training. This reduces the risk of the neurons of fully connected layers developing excess interdependencies among each other that lead to the model overfitting to the training data. The dropout hyperparameter, p, which represents the fraction of neurons within a layer that will be dropped at random, was fine-tuned. It must be noted that dropout was only incorporated in the LSTM layers, given that these layers contain a significant amount of trainable parameters.

All model trainings were conducted using the stochastic gradient descent optimisation while using the Adam optimiser. A fixed learning rate was incorporated with an objective to minimise the loss function. An activation function was selected for all hidden layers. The batch size for each iteration of the gradient descent was determined based on a sensitivity analysis conducted for each sensor signal used for model training. The models were trained for a fixed number of epochs, which were selected based on visual inspection of the learning curves generated and considering the available computational capacity. All model development and training were conducted using the Python software library of TensorFlow (Abadi *et al.* 2016).

2.3.3. Aggregation of model predictions using exponential weighting

Wastewater treatment processes are non-linear and dynamic, and environmental and operational factors can have influence over the processes at varying time constants. For example, the increase in dissolved oxygen concentration can lead to a delayed or lagged reduction of ammonium (NH_4^+) concentration (Rieger *et al.* 2006). Furthermore, operational changes can have delayed implications on the microorganism activity within the biomass (Rolfe *et al.* 2012), which in turn is reflected in some of the (online) measured parameters, such as nitrate (NO_3^-) concentration. Hence, to aid in monitoring and control of WWTPs, models that are developed must capture the process dynamics reflected in varying time constants. In this study, ARIMA and LSTM-based AE models were trained to capture short-term dynamics (SD) and long-term dynamics (LD) by a combination of resampling to lower sample frequencies while averaging. When dealing with anomalies that occur during an extended time period, multiple model predictions should be reliable for the period of the anomalous data. It was anticipated that models trained at a higher granularity would be sufficient to predict the real sensor signal during a short anomaly period, which is typically shorter than the history used to train the model. Since longer anomaly periods require recursive predictions, errors will accumulate when using a prediction as input. As a result, model predictions will quickly diverge from the real process parameter values. This accumulation issue was addressed by utilising predictions from the LD models, offsetting the limitations from the SD models. To combine the output of the SD and LD models, the data reconciliation protocol includes resampling and aggregation by exponential weighting, as follows.

The predictions of a model are written as $\hat{y}_n$, where n denotes the sampling period in minutes, and $\hat{y}_{\mathrm{LD},n_\mathrm{L}}$ and $\hat{y}_{\mathrm{SD},n_\mathrm{S}}$ denote LD predictions with an original sampling period n_L and SD predictions with a sampling period n_S, respectively. For the aggregation of both model outputs, $\hat{y}_{\mathrm{LD},\, n_\mathrm{L}}$ are up-sampled to a frequency $1/n_\mathrm{S}$, resulting in $\hat{y}_{\mathrm{LD},n_\mathrm{S}}$ using linear interpolation.

Consequently, predictions from $\hat{y}_{\mathrm{LD},n_{\mathrm{S}}}$ can be used for aggregation with $\hat{y}_{\mathrm{SD},n_{\mathrm{S}}}$. Furthermore, the total number of intermediate values within the LD sampling period is defined as k. Weights (w) are determined using an exponential function and a user-defined value termed as the ratio of importance (γ) defined at k and the time instance t_j where the index j is reset to zero if there is no anomaly:

$$w_j = e^{(j-1)\frac{\ln \gamma}{k-1}} \tag{7}$$

In this case, n_{L} is 30 min and n_{S} is 5 min, leading to a value $k = n_{\mathrm{L}}/n_{\mathrm{S}} = 6$. If we define the ratio of the importance of the 5-min model as 10% (i.e. $\gamma = 0.1$), then for this case, Equation (7) leads to γ being equal to the sixth weight factor (w_6).

Summarising, w determines the fractional value that should be given to $\hat{y}_{\mathrm{SD},n_{\mathrm{S}}}$ at time j and can be interpreted as how much $\hat{y}_{\mathrm{SD},n_{\mathrm{S}}}$ should contribute to the reconciled value consisting of the SD and LD predictions. In case there is a prolonged anomalous event, it is expected that the accumulation of prediction errors in the SD model will be substantially larger than the accumulated error in the LD model. Therefore, an exponential weighting is introduced where the importance of the SD model is reduced while increasingly relying on the LD model with increasing j as time progresses further from the last known non-anomalous value, i.e.

$$\tilde{y}_{n_{\mathrm{S}}}(j) = w_j \cdot \hat{y}_{\mathrm{SD},n_{\mathrm{S}}}(j) + (1 - w_j) \cdot \hat{y}_{\mathrm{LD},n_{\mathrm{S}}}(j) \tag{8}$$

where $\tilde{y}_{n_{\mathrm{S}}}$ is the reconciled signal with an n_{S} sampling frequency. The exponential weighting is solely dependent on the value of γ at some time window length k. An alternative to the approach is illustrated in this work, and γ can be set using a criterion for the prediction error (e.g. the mean squared error (MSE)) for specified horizon lengths within a test dataset and subsequently γ can be chosen based on the lowest MSE.

2.4. Model performance metrics

Hyperparameter values and model structures for both the benchmark ARIMA models and the LSTM-based AEs were selected by comparing their predictive performance on the test datasets, i.e. unseen data not used during the training process. To this aim, the MSE is used:

$$\mathrm{MSE} = \frac{\sum_{i}^{N} (y_i - \widehat{y}_i)^2}{N} \tag{9}$$

where $\widehat{y}_i$ and y_i are the predicted and measured values at time instant i, respectively, and N is the number of considered data points. Additionally, root mean squared error (RMSE), which is the root of the MSE, as shown in Equation (9), is used as a metric to provide further assessment of model performance while penalising large errors.

Furthermore, the performance metric coefficient of determination (CoD or R^2) was used:

$$R^2 = 1 - \frac{\sum_{i}^{N} (y_i - \widehat{y}_i)^2}{\sum_{i}^{N} (y_i - \bar{y})^2} \tag{10}$$

where $\bar{y}$ is the average value of observed data calculated over N data points. The R^2 for the training set is defined as R^2_{train}, and the R^2 for the test set is R^2_{test}, and we evaluate R^2_{test} to account for overfitting or underfitting. For the case of the AEs, R^2_{train} has been used as a metric to improve the model architecture in subsequent trainings.

3. CASE STUDY

3.1. Amsterdam West WWTP

The data validation application was developed and implemented for the Amsterdam West WWTP that is owned by the water authority Amstel, Gooi en Vecht, and is operated by Waternet, the water utility for the city of Amsterdam and surrounding areas. The WWTP has a capacity of 1.1 million population equivalent and seven treatment lanes. The control loops of the WWTP are largely locally distributed and are dedicated to a single wastewater treatment unit process. Currently, efforts are being made to upgrade the control process through the development and deployment of smart control applications that use real-time plant data and data-driven modelling to achieve more optimal plant-wide control. This is highly desired to achieve the goals of reducing the carbon footprint while meeting effluent quality criteria. Specifically, for the Amsterdam West WWTP, the objective is to minimise energy consumption and nitrous oxide (N_2O) emissions. N_2O emissions have a large impact on the carbon footprint of the treatment system, because N_2O has a global warming potential that is 273 times higher than carbon dioxide (CO_2) (Forster *et al.* 2021). Therefore, real-time and automated validation of sensor data is an integral part that must be implemented to ensure high data quality is ingested by data-driven forecasting and control tools to optimise such operational and environmentally based objectives.

3.2. WWTP data

In this data validation implementation, raw data from sensors measuring the NH_4^+ and NO_3^- concentrations in the aerobic tank of the bioreactor unit from one treatment lane were used. These process parameters are key indicators in the control and for the proper functioning of wastewater treatment, especially during the nitrification and denitrification process of activated sludge systems. Additionally, NH_4^+ and NO_3^- also provide information on the N_2O emissions, and their concentration levels in the system can provide indications on which production pathways are active for given process conditions. As a result, these sensor signals are important features when developing data-driven digital twins to model the wastewater treatment processes or key state variables within a control scheme with the objective to reduce N_2O emissions. The screening of the raw data from these signals to detect anomalies and conduct their reconciliation is therefore a crucial initial step prior to its ingestion.

For the anomaly detection routines, thresholds were assigned for the NH_4^+ and NO_3^- data signals, specifically for the threshold detection, flatline detection, and spike detection methods. This was conducted in consultation with process engineers of the WWTP who are familiar with the process signals and extensively work with the data. The thresholds can be found in Table S1 in the supplementary information. For model training and testing, a dataset comprising time series starting from September 2020 until April 2021 was used. Initially, the raw data were investigated in detail to remove anomalies, resulting in a clean and representative dataset, as detailed in Section 2.4. Furthermore, an additional dataset amounting to approximately 3 months of data, from May 2021 to July 2021, was used to assess the anomaly detection routines. In this dataset, no pre-processing step was done, which involved the prior identification and handling of erroneous values, as the primary goal in utilising this dataset was to test the anomaly detection routines. As mentioned in Section 2.3.3., two ARIMA and two LSTM-based AE models were trained for each sensor signal from the WWTP to capture the SD and LD. This was achieved by resampling the raw sensor data to a targeted granularity. Additionally, as required by both the ARIMA and LSTM-based AEs, a fixed amount of historical input must be decided, which would be used to make forecasts. A careful investigation and exploration of the raw data was undertaken to better understand the process dynamics behind the normal operations seen in this system. Through this analysis, supplemented with prior knowledge of wastewater treatment and while considering the trade-off with computation costs, it was decided that resampling the raw data to 5 min granularity would be sufficient to capture the SD and 30 min to capture the LD. Finally, the historical input that was used for the SD model was fixed to 3 h, and for the LD model, it was fixed to 24 h, as summarised in Table 1.

4. RESULTS AND DISCUSSION

4.1. Anomaly detection routines

The anomaly detection routines (Section 2.2.) were run on three different datasets. Dataset 1 consisted of 10^5 data points of non-resampled NO_3^- concentration data, of which 904 were manually identified to be anomalous (flatline-type anomalies). Dataset 2 was the same as dataset 1, but four lengths of 250 data points were replaced with artificial flatlines. Finally, dataset 3 was the same as dataset 2, but 1,000 points were randomly chosen and replaced with artificial single-point-like anomalies. Here, a single value was replaced with a randomly chosen value between the minimum and maximum in the dataset. The locations of artificial anomalies were chosen such that there were no adjacent or overlapping anomalies. The random generation of artificial anomalies and subsequent anomaly detection was repeated five times, and the results are averaged and noted in Table 2.

Table 1 | Amount of historical input to capture short-term and long-term dynamics for both NH_4^+ and NO_3^- sensor signals, as used by the ARIMA and LSTM-based autoencoder models

Model (ARIMA and LSTM-based AE)	Granularity (min)	Historical input (h)	Historical input (no. of timesteps)
Short-term dynamics	5	3	36
Long-term dynamics	30	24	48

Table 2 | Confusion matrices for three different datasets

	True positives	False positives	True negatives	False negatives
Uncleaned data	904	0	99,096	0
Uncleaned data, with four flatline anomalies	1,904	0	98,096	0
Uncleaned data, with four flatline anomalies and 1,000 spike anomalies	2,883	7	97,083	27

Note: The first row depicts a matrix for 105 data points of non-resampled NO_3^- data, the second row depicts the matrix for the raw data with four 250-point long flatline anomalies inserted, and the third row depicts the matrix for the raw data with four 250-point long flatline anomalies and 1,000 point-like anomalies added.

The anomaly detection routines, in particular the flatline detection routines, can detect all flatline anomalies in the data. This is independent of the flatline value. Detecting spike anomalies is, however, more difficult (see, e.g., Leigh *et al.* 2019).

4.2. Performance of ARIMA models

ARIMA models were trained to capture the SD and LD for the NO_3^- and NH_4^+ sensor signals. As highlighted in Section 3.2., the chosen granularity for SD and LD is 5 and 30 min, and the historical input used to make predictions is 3 h (36 steps) and 24 h (48 steps), respectively. As a result, the ARIMA-based hyperparameters p and q were set to the historical inputs assigned for a given granularity. Initially, a minor search was conducted to identify the optimal value for the hyperparameter, d. In all cases, it was found that $d = 0$ is optimal, indicating that the underlying data do not have a significant upward or downward trend.

For process technical reasons, the hyperparameters for the SD ARIMA model were set at $p, d, q = 36,\ 0,\ 36$; and the LD ARIMA models were set at $p, d, q = 48,\ 0,\ 48$. In Figure 2, predictions from the SD ARIMA models for (a) the NO_3^- and (b) NH_4^+ sensor signal are shown, respectively, on a section of the test set using the green dashed line. Both models provide highly accurate one-step-ahead predictions, with the models following the trends of the signal with minimal errors. This is also reflected in the high R^2 values, which are equal to 0.99, calculated for both the training and test sets for the SD ARIMA models, as summarised in Table 3. This can primarily be attributed to the fact that X_i is highly correlated with X_{i-1}, which is corroborated by the fact that lower R^2 values are found if the resampling rate is raised in the LD ARIMA models.

Figure 3 shows the predictions from the LD ARIMA models on a section of the test set along with the observed data. Here, and in Table 3, it can be seen that the accuracy of the one-step model predictions has reduced as compared to the SD ARIMA model predictions. The LD ARIMA model predictions are less accurate even though a higher amount of historical input has been provided.

4.3. LSTM-based AE models' performance

Based on the model training procedure detailed in Section 2.3.2., hyperparameters were fine-tuned to increase the prediction accuracy of the model for the given training and test datasets. In Table S2, detailed information is provided on the various model structures and hyperparameter combinations that were investigated. After the detailed search, a summary of the resulting choices of the hyperparameters used for training the best-performing model has been provided in Table 4. With respect to the dropout rate hyperparameter, it was seen that NH_4^+ required a slightly higher value (6%) as compared to NO_3^- (3%), given that the models were overfitting to the NH_4^+ data.

This can be explained by the fact that the NH_4^+ dataset contained various extreme peaks, where high concentration values are observed, particularly during rainfall events. However, during dry weather conditions and regular operating conditions, the NH_4^+ concentrations are relatively low and stable, followed by the incoming influent nitrogen trends. This proved challenging for the AEs to learn, as compared to the NO_3^- sensor data, which possess lesser extreme peaks and more gradual trends.

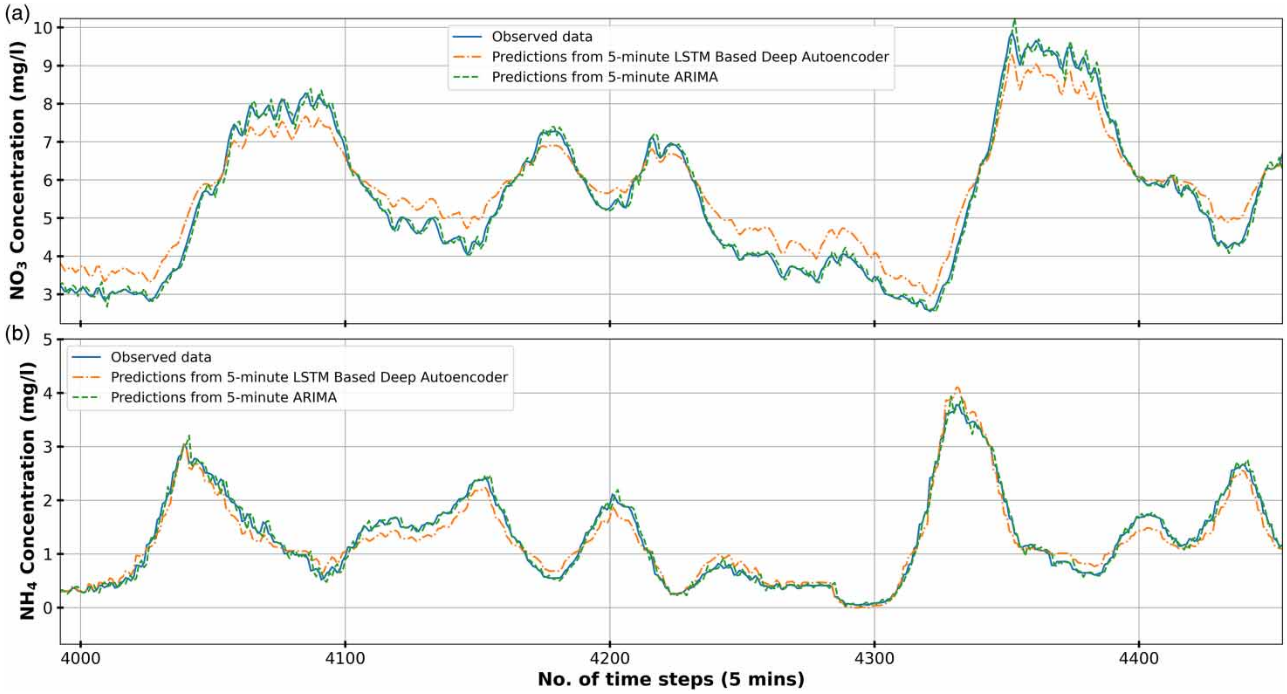


Figure 2 | Short-term dynamics (5-min) predictions from ARIMA and LSTM-based AE models for (a) NO_3^- and (b) NH_4^+ sensor data on the test set.

Table 3 | ARIMA and LSTM-based AE models MSE (mg/l), RMSE (mg/l), and R^2 values for NO_3^- and NH_4^+ data signal

	Short-term dynamics (5-min)			**Long-term dynamics (30-min)**		
Model	R^2	**MSE**	**RMSE**	R^2	**MSE**	**RMSE**
NO_3^-						
ARIMA						
Train	0.99	0.018	0.13	0.97	0.24	0.49
Test	0.99	0.015	0.12	0.95	0.25	0.50
LSTM-based AE **s**						
Train	0.97	0.24	0.49	0.99	0.048	0.22
Test	0.95	0.28	0.53	0.99	0.056	0.24
NH_4^+						
ARIMA						
Train	0.99	0.010	0.10	0.96	0.12	0.12
Test	0.99	0.015	0.12	0.95	0.098	0.31
LSTM-based AEs						
Train	0.90	0.30	0.55	0.98	0.05	0.22
Test	0.90	0.19	0.43	0.98	0.04	0.21

In Table 3, a summary of the SD and LD model performances from the LSTM-based AEs is provided. In Figure 2(a), the orange line represents the predictions obtained during a section of the test from the best-performing SD AE model to reconstruct the NO_3^- data signal. As shown, a fit was achieved with R^2 values of 0.97 (training set, as depicted in Figure S4 in the Supplementary Information) and 0.95 (testing set). Similarly, for the reconstructing of the NH_4^+ data signal, Figure S5 and the orange line in Figure 2(b) show the results obtained during the training and testing of the best-performing SD AE model,

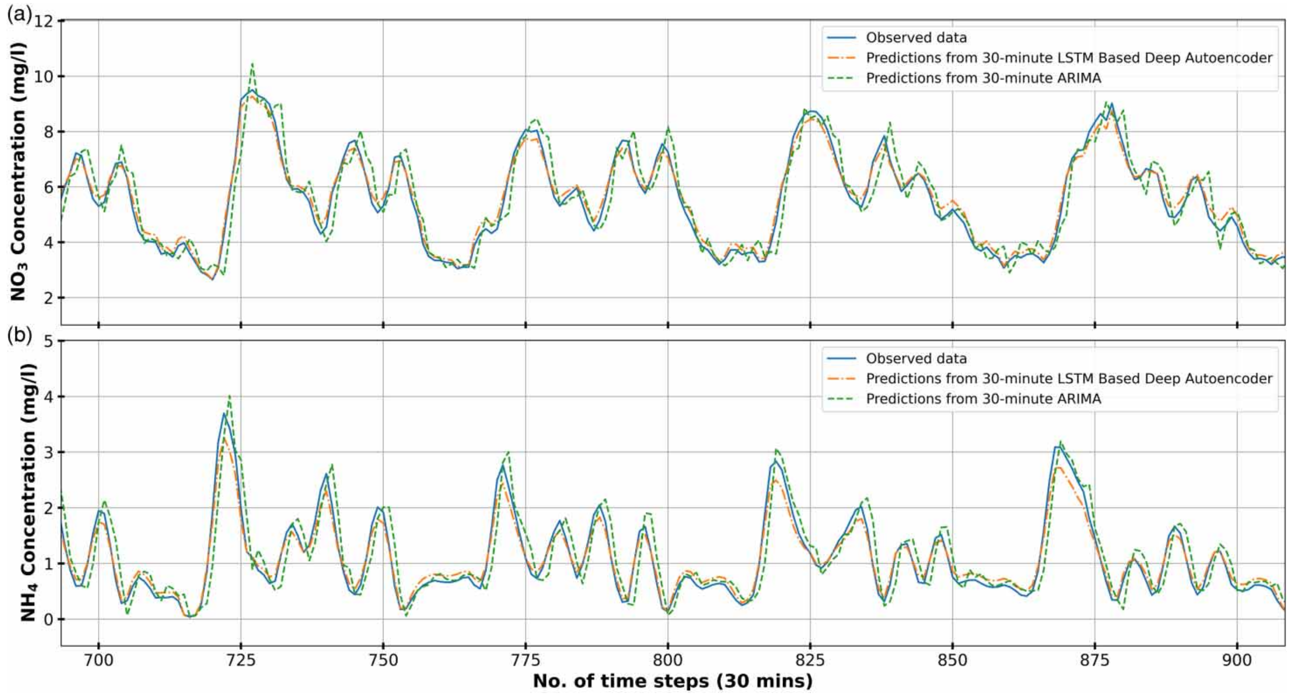


Figure 3 | Long-term dynamics (30-min) predictions from ARIMA and LSTM-based AE models for (a) NO_3^- and (b) NH_4^+ sensor data on the test set.

Table 4 | Hyperparameters tuned during model training and choices for the final model trainings

Hyperparameter	Value/choice	Comment
No. of epochs	35	Decision made based on results obtained from a learning curve
Optimiser	Adam	
Learning rate	0.00001	–
Loss function	Mean squared error	–
Activation function	ReLU	Same activation function used for all hidden layers
Batch size	112 (5-min AE) and 14 (30-min AE)	Based on a sensitivity analysis, it was concluded that a batch size representing 2 weeks of data yielded the best results
Dropout rate (p)	NH_4^+ sensor signal – 0.06 NO_3^- sensor signal – 0.03	Dropout regularisation only applied to LSTM layers

respectively, where an R^2 value of 0.90 was achieved for both datasets. For the LD AE models to reconstruct the NO_3^- data signal, the results of the testing set are provided in the orange line in Figure 3. The results obtained in the training set are provided in Supplementary Material, Figure S6. A high-performance and prediction accuracy with an R^2 score of 0.99 was achieved for both during training and testing. Similarly, reconstruction NH_4^+ using the LD AE model, an R^2 score of 0.98 was achieved.

Additionally, the model architectures for the four AE models trained for NH_4^+ and NO_3^- are provided in Tables S3–S6 in the Supplementary Information. An interesting insight when comparing the SD AE model architectures (Tables S3 and S4) is the added complexity needed to capture the SD of NH_4^+. The need for an additional hidden layer to increase the prediction accuracy could potentially be attributed to the diurnal loads of NH_4^+ that the WWTP receives. These loads are based on consumption trends that influence the nitrification process in the aerobic tank. Additional observations can be deduced from the results regarding the LD AE models. Firstly, the predictive performance from the LD AE models was higher than the performance from the SD model predictions. This could be expected, considering that more historical data were used, where timesteps amounting to 24 h were entered for the LD AE models as compared to 3 h for the SD models. LSTM layers are known to adequately learn the long-term trends when provided with a long history. Secondly, the LD AE model for NO_3^- required more hidden layers in both the encoder and decoder components of the model, leading to the model becoming 'deeper' compared to the SD AE model for NO_3^- (as shown in Tables S3 and S5). This was not the case when reconstructing the NH_4^+ data signal while capturing the long-term dynamics, as shown in Table S6. This could be explained due to seasonal variations affecting the nitrification and denitrification processes. Should the nitrification process be affected (e.g., due to lower temperatures), the amount of NO_3^- in the aerobic tank could be influenced. Similarly, if the denitrification process is affected, an indirect influence of the NO_3^- in the aerobic tank could occur, given the Amsterdam West WWTP bioreactor is completely mixed with internal recycles. It is, therefore, hypothesised that the modelling of long-term dynamics of the NO_3^- data signal through the LD AE is affected more by seasonal effects, which therefore require more complex model architectures. Note that several authors, such as Do *et al.* (2022) and Yadav *et al.* (2023), have attempted to include seasonal effects of wastewater inflow behaviour with SARIMA, an extension of ARIMA models to capture seasonal effects.

4.4. Comparing reconciliation using recursive predictions and aggregation

During longer anomalous events, such as flatlines in raw data, the trained models used for reconciliation will be required to perform recursive predictions (inputting predictions as input) to increase the forecasted steps provided by the model. The recursive predictive power of the ARIMA and LSTM-based AE was compared for both the SD and LD models using the same fixed time horizon. This was conducted for two windows within the NO_3^- sensor test set with a forecasting horizon of 4 h. In Figure 4, the recursive performance within the forecasting horizon of the different models can be seen.

Furthermore, the combined MSE values calculated during the 4-h forecasting horizon are provided in Table 5. As can be seen, the SD models performed more accurate predictions compared to the LD models. With respect to the SD models (Figure 4(a) and 4(c)), the LSTM-based AE performed better than the ARIMA model, where the predictions from the former can be seen to follow the trends and have a reasonable fit to the observed data. The SD ARIMA model follows the general trend by having a poorer fit. This can be seen in the MSE values, as well as shown in Table 5. For the LD models (Figure 4(b) and 4(d)), both the ARIMA and LSTM-based AE models can be considered unsatisfactory in performing recursive predictions. The LD ARIMA model is seen to follow the general trend of the observed data but has a poor fit to the observed data. The LD LSTM-based AE model seems to struggle to follow the trend and fit the observed data. This highlights a limitation in modelling LD using the ARIMA and LSTM-AE. The use of lower granularity and increased historical input was found to be insufficient. Based on results obtained in these example cases, it can be deduced that the SD LSTM-AE model has the potential of accurately reconciling for an anomaly event in nitrate data that is up to 3 h long. However, this result is based on a limited analysis conducted on a fraction of samples. In-depth analysis by testing the models on various scenarios is warranted.

To mitigate the accumulation of errors when predicting future values, a methodology described in Section 2.4. uses an aggregation of the SD and LD models with exponential weighting. For the Amsterdam West WWTP use case, based on prior knowledge of the biological treatment process, certain choices were made for the parameter values used to calculate the exponential weights, as defined in Equation (8). The value for the parameter *ratio of importance* (γ) was set to be 0.1, and the value for k was set to 6. This means that the sixth prediction made by the 5-min SD model in a forecasting horizon will be weighted 10%, and therefore, the contribution from the 30-min model, which is up-sampled to 5 min, will be 90%. To test this methodology, two synthetic flatlines were created in the same windows in the NO_3^- test set that were demonstrated earlier for the recursive prediction results. In Figure 5, a comparison is shown between the aggregated ARIMA (maroon line) and AE (grey line) models' performance during synthetic flatlines (red line) generated. The aggregated results are also compared with the actual data observed (blue line) from the nitrate sensor signal. As expected, the aggregation of the LSTM-based AE models resulted in a poor fit to the observed data. This can be attributed to the lower-performing LD model with respect to the recursive predictions. The aggregated results from the ARIMA models performed marginally better than the AE aggregation, which can also be attributed to the better-performing recursive predictions by the LD ARIMA model, as can be seen in the performance metrics in Table 5.

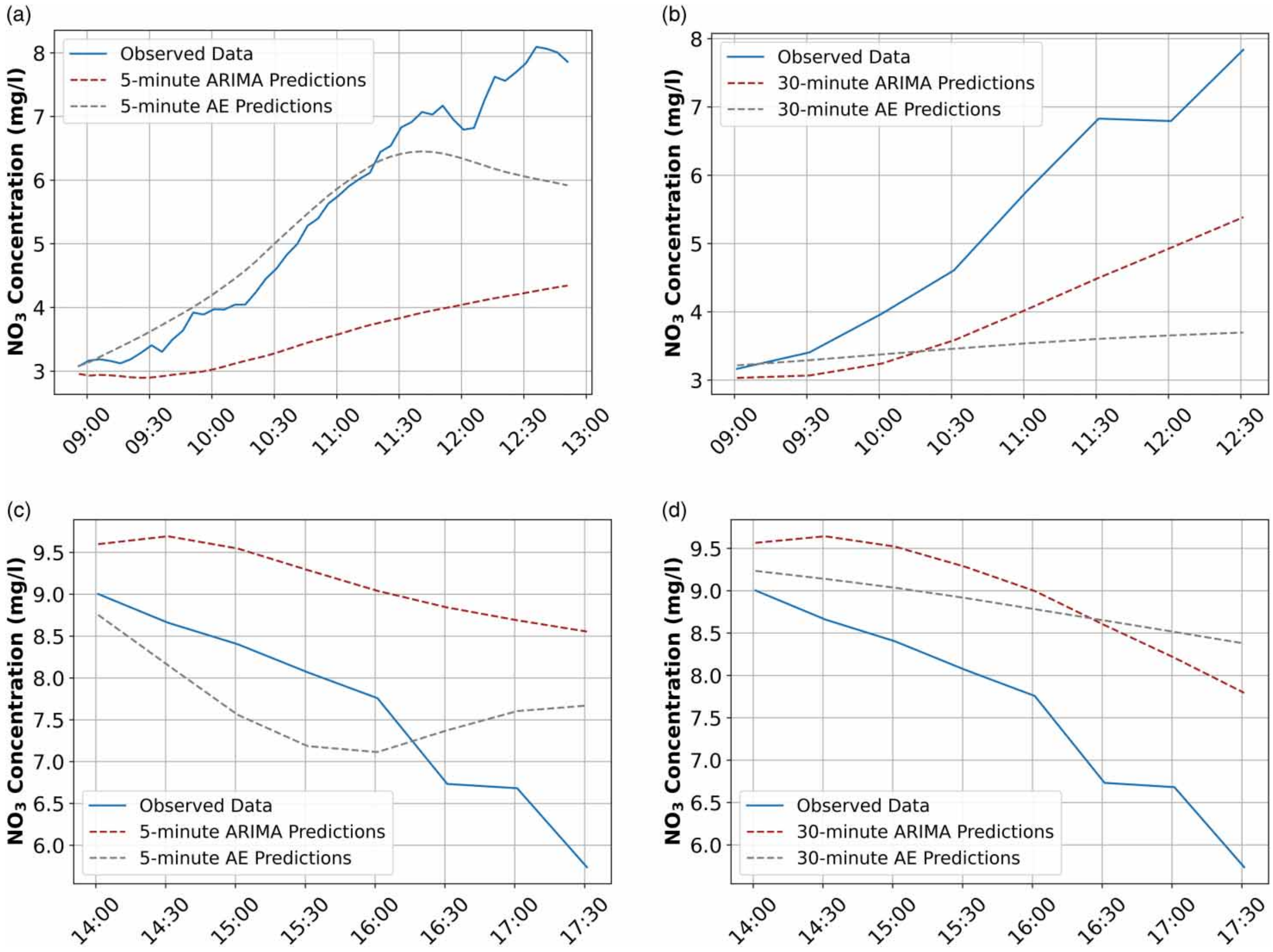


Figure 4 | Comparing recursive predictions made on two windows in the NO_3^- test set by the 5-min or SD models (a and c) and 30-min or LD models (b and d).

Table 5 | MSE values (mg/l) calculated for the two windows used within the NO_3^- test set to assess the recursive predictive performance

Model type	5-min/SD model	30-min/LD model
ARIMA	3.86	2.19
LSTM-based AE	0.73	3.77

4.5. Limitations and future perspectives

In this study, a methodology to validate dynamic sensor data from a full-scale WWTP was demonstrated. The reconciliation of the data signals requires high-performance data-driven models where the predictions can be used to replace the anomalous values observed in the raw dataset. In the case of one-step-ahead predictions, it was seen that traditional time series regression models such as the ARIMA as well as more complex AI-based neural networks provide high [$\geq$0.95 for nitrate and $\geq$ 0.90 for ammonium] R^2 values and would be sufficient in reconciling single-value-based anomalies. In more challenging tasks of reconciling long anomalous events such as flatlines, highly accurate model predictions are needed to mitigate the accumulation of residual errors. However, certain limitations in the models identified to perform this task were seen. Although LSTM-based AE models have the potential to show superior performance as compared to ARIMA models (Siami-Namini *et al.* 2019), the results in this study tell a different story. Firstly, AEs are designed to provide a

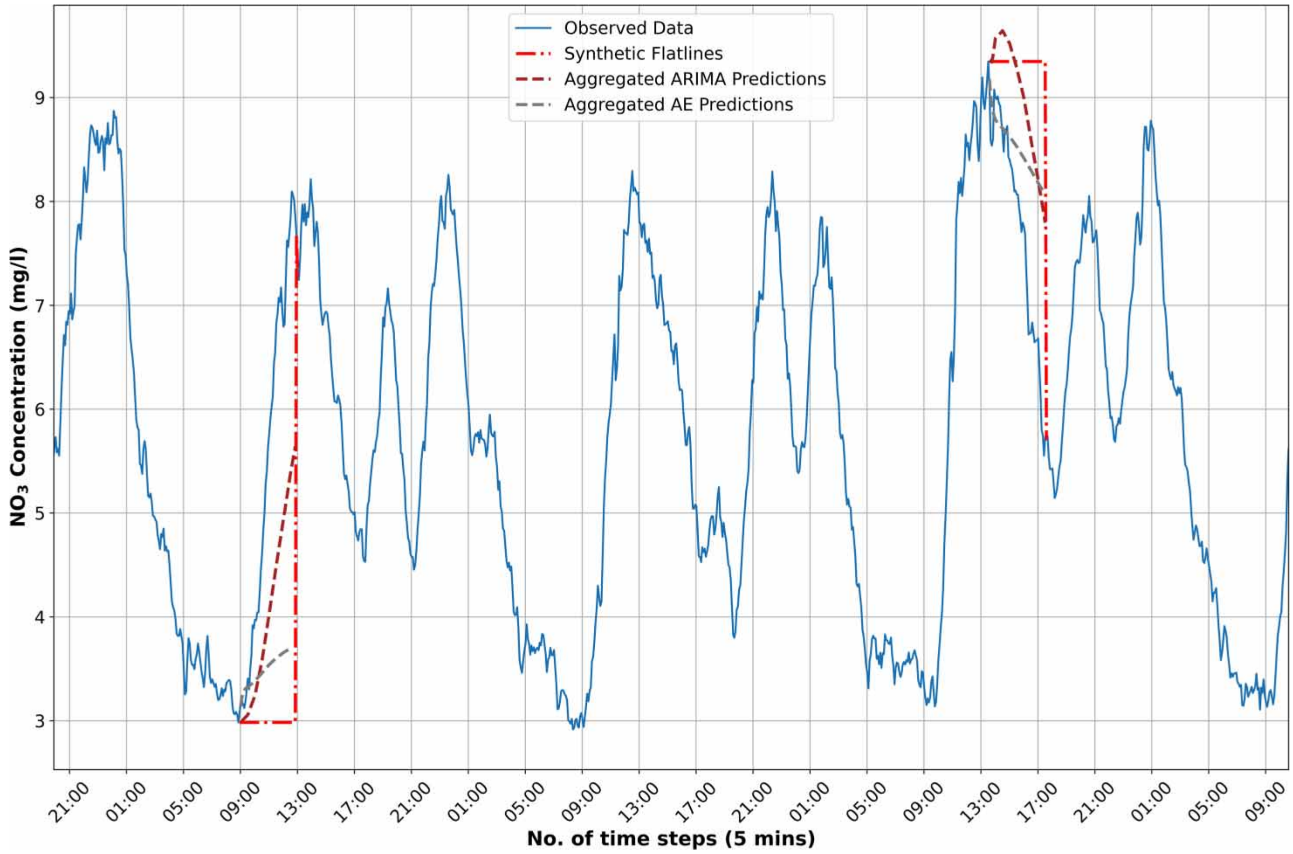


Figure 5 | Comparing the aggregated ARIMA (maroon dashed line) and AE models (grey dashed line) for two examples where synthetic flatlines (red dashed-dot line) were added. The aggregated results were compared with observed NO_3^- concentrations (blue line).

reconstruction of a dataset, and the approach used for training and selection here might not be optimal for training models with good forecasting ability. The training process can be set up in such a way as to steer the model to learn to improve the recursive prediction accuracy by employing a multi-step-ahead loss function, which measures the accumulated error for consecutive time steps (Bentivoglio *et al.* 2023). Alternatively, more conventional neural networks tuned at forecasting performance is a direction worth investigating, considering the decreased amount of the complexity of the model structures compared to LSTM-AE. Furthermore, only univariate models were considered in this study, where predictions for a sensor signal were made while only considering its own history. More data or information, such as inputting other correlated sensor signals that influence the nitrification and denitrification processes, can help in improving the recursive predictive, i.e. forecasting performance as well as capturing seasonal effects by, e.g., SARIMA models (Do *et al.* 2022; Yadav *et al.* 2023). Finally, the calculation of a prediction interval of (autoregressive) models can be an interesting addition (Hill & Minsker 2010). A prediction interval can provide valuable information on the increasing level of uncertainty that a user can expect when making predictions over a future horizon. Therefore, future research will be targeted at improving the reliability of the forecasting performance by assessing models of intermediate complexity, e.g. SARIMA and neural network models.

The aggregation methodology introduced in this study is a straightforward method to utilise two different models. In this case, models have differently trained parameters and sample time durations. The methodology can further be refined by minimising (for example) the MSE with the aid of the weighting parameter γ. Furthermore, calibration of the reconciliation procedure can be improved further by minimising (for example) the MSE with the aid of the weighting parameter γ and the fixed parameter k. Finally, the methods are being tested in a real production environment of the water utility Waternet, while connections to the legacy system of the water company are preserved using a similar procedure as described in Seshan *et al.* (2023).

5. CONCLUSIONS

In this paper, a detection and reconciliation procedure has been described where the reconciliation accuracy of data tagged as anomalous has been successfully assessed. For reconciliation, the performance of LSTM-AE models that were identified after an extensive hyperparameter search was compared with the performance of ARIMA models using a training procedure that captures SD and LD timescale dynamics. The following conclusions are drawn:

- For the detection of sensor signal faults, the performance of the statistical and heuristic methods is flawless for single-point anomalies, and only 2.7% of the points were incorrectly classified due to spike anomalies.
- One-step-ahead predictions from the SD and LD models are used for reconciling single-value-based anomalies, which are provided highly accurate predictions for both NO_3^- (SD: $R^2 \geq 0.95$, RMSE ≤ 0.53 mg/l; LD: $R^2 \geq 0.95$, RMSE ≤ 0.50 mg/l) and NH_4^+ (SD: $R^2 \geq 0.90$, RMSE ≤ 0.55 mg/l; LD: $R^2 \geq 0.95$, RMSE ≤ 0.31 mg/l) sensor data. When performing recursive predictions where residual errors accumulate, the accuracy of the SD AE model was significantly higher than the performance of the SD ARIMA model, whereas the performance metrics were in the same order of magnitude for the LD models and the ARIMA model proved to be slightly more accurate. However, the structure of ARIMA models is less complex, and the model selection is relatively straightforward compared to that of LSTM AEs.
- The proposed aggregation method of the SD and LD model predictions allows the user to tune the accuracy of the reconciled signal by aggregating the outputs of both models with time-weighted importance.

Hence, the DVR procedure proposed in this work shows that it is possible to detect a large number of single-point anomalies and correct these with high accuracy, where the accuracy of AE models surpasses the accuracy obtained by ARIMA models. For contextual anomalies, especially when dealing with environmental sensor data subject to (sudden) environmental events, the current univariate modelling approach reaches its performance limits in reconciling the real sensor signal with increasing forecasting horizon due to, most probably, 'unknown' influent dynamics. As a next step, a multi-step-ahead loss function during the training process should be able to improve the prediction accuracy for long(er) time horizons. Finally, it is hypothesised that inputting other correlated sensor signals and including the calculation of prediction intervals will benefit the usability of the DVR approach by increasing its accuracy for longer contextual anomaly events and by showing the uncertainty involved in forecasting, respectively.

ACKNOWLEDGEMENTS

The authors would like to thank the water authority Amstel, Gooi and Vecht for making the data from the Amsterdam West WWTP available for this study. Part of the research was funded by the Fiware4Water project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 821036 and the Joint Research Programme of the Dutch Water Companies (BTO).

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

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

Arundo Analytics, Inc. 2020 *ADTK: Anomaly Detection Toolkit (Version 0.6.2) [Software]. PyPi*. Available from: https://adtk.readthedocs.io/en/stable/index.html.

Ba-Alawi, A. H., Vilela, P., Loy-Benitez, J., Heo, S. K. & Yoo, C. K. 2021 Intelligent sensor validation for sustainable influent quality monitoring in wastewater treatment plants using stacked denoising autoencoders. *Journal of Water Process Engineering* **43**, 102206. https://doi.org/10.1016/J.JWPE.2021.102206.

Ba-Alawi, A. H., Loy-Benitez, J., Kim, S. & Yoo, C. 2022 Missing data imputation and sensor self-validation towards a sustainable operation of wastewater treatment plants via deep variational residual autoencoders. *Chemosphere* **288** (132647), 132647. https://doi.org/10.1016/j.chemosphere.2021.132647.

Bengio, Y. 2009 Learning deep architectures for AI. In: *Foundations and Trends in Machine Learning*, Vol. 2, Issue 1. https://doi.org/10.1561/2200000006.

Bentivoglio, R., Isufi, E., Jonkman, S. N. & Taormina, R. 2023 *Rapid Spatio-Temporal Flood Modelling via Hydraulics-Based Graph Neural Networks*. Available from: https://egusphere.copernicus.org/preprints/2023/egusphere-2023-284/.

Box, G. E. P. & Jenkins, G. M. 1970 *Time Series Analysis: Forecasting and Control*. Holden-Day. Available from: https://books.google.com/books/about/Time_Series_Analysis.html?hl=nl&id=5BVfnXaq03oC.

Council Directive (91/271/EEC) of 21 May 1991 – European Environment Agency n.d. Available from: https://www.eea.europa.eu/policy-documents/council-directive-91-271-eec (accessed 22 September 2023).

Daelman, M. R. J., van Voorthuizen, E. M., van Dongen, U. G. J. M., Volcke, E. I. P. & van Loosdrecht, M. C. M. 2015 Seasonal and diurnal variability of N_2O emissions from a full-scale municipal wastewater treatment plant. *Science of the Total Environment* **536**, 1–11. https://doi.org/10.1016/J.SCITOTENV.2015.06.122.

Di Marcantonio, C., Chiavola, A., Dossi, S., Cecchini, G., Leoni, S., Frugis, A., Spizzirri, M. & Boni, M. R. 2020 Occurrence, seasonal variations and removal of organic micropollutants in 76 wastewater treatment plants. *Process Safety and Environmental Protection* **141**, 61–72. https://doi.org/10.1016/J.PSEP.2020.05.032.

Do, P., Chow, C. W. K., Rameezdeen, R. & Gorjian, N. 2022 Wastewater inflow time series forecasting at low temporal resolution using SARIMA model: A case study in South Australia. *Environmental Science and Pollution Research* **29** (47), 70984–70999. https://doi.org/10.1007/S11356-022-20777-Y/FIGURES/11.

Fernando, W. A. M., Khadaroo, S. N. B. A. & Poh, P. E. 2022 Artificial intelligence in wastewater treatment systems in the era of industry 4.0: A holistic review. In: *Artificial Intelligence and Environmental Sustainability: Challenges and Solutions in the Era of Industry 4.0*, pp. 45–85. https://doi.org/10.1007/978-981-19-1434-8_3.

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

Gaddam, A., Wilkin, T., Angelova, M. & Gaddam, J. 2020 Detecting sensor faults, anomalies and outliers in the internet of things: A survey on the challenges and solutions. *Electronics* **9** (3), 511. https://doi.org/10.3390/ELECTRONICS9030511.

Hill, D. J. & Minsker, B. S. 2010 Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software* **25** (9), 1014–1022. https://doi.org/10.1016/J.ENVSOFT.2009.08.010.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. 2012 *Improving Neural Networks by Preventing co-Adaptation of Feature Detectors*. arXiv. http://arxiv.org/abs/1207.0580.

Hochreiter, S. & Schmidhuber, J. 1997 Long short-term memory. *Neural Computation* **9** (8), 1739–1780.

Jagatheesaperumal, S. K., Rahouti, M., Ahmad, K., Al-Fuqaha, A. & Guizani, M. 2022 The duo of artificial intelligence and big data for industry 4.0: Applications, techniques, challenges, and future research directions. *IEEE Internet of Things Journal* **9** (15), 12861–12885. https://doi.org/10.1109/JIOT.2021.3139827.

Leigh, C., Alsibai, O., Hyndman, R. J., Kandanaarachchi, S., King, O. C., McGree, J. M., Neelamraju, C., Strauss, J., Talagala, P. D., Turner, R. D. R., Mengersen, K. & Peterson, E. E. 2019 A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. *Science of the Total Environment* **664**, 885–898. https://doi.org/10.1016/J.SCITOTENV.2019.02.085.

Liu, Y., Ramin, P., Flores-Alsina, X. & Gernaey, K. V. 2023 Transforming data into actionable knowledge for fault detection, diagnosis and prognosis in urban wastewater systems with AI techniques: A mini-review. *Process Safety and Environmental Protection* **172**, 501–512. https://doi.org/10.1016/J.PSEP.2023.02.043.

Md Nor, N., Che Hassan, C. R. & Hussain, M. A. 2020 A review of data-driven fault detection and diagnosis methods: Applications in chemical process systems. *Reviews in Chemical Engineering* **36** (4), 513–553. https://doi.org/10.1515/REVCE-2017-0069/ASSET/GRAPHIC/J_REVCE-2017-0069_CV_003.JPG.

Mehdizadeh, S. 2020 Using AR, MA, and ARMA time series models to improve the performance of MARS and KNN approaches in monthly precipitation modeling under limited climatic data. *Water Resources Management* **34** (1), 263–282. https://doi.org/10.1007/S11269-019-02442-1/METRICS.

Moon, J., Hossain, M. B. & Chon, K. H. 2021 AR and ARMA model order selection for time-series modeling with ImageNet classification. *Signal Processing* **183**, 108026. https://doi.org/10.1016/J.SIGPRO.2021.108026.

Park, Y. J., Fan, S. K. S. & Hsu, C. Y. 2020 A review on fault detection and process diagnostics in industrial processes. *Processes* **8** (9), 1123. https://doi.org/10.3390/PR8091123.

Pisa, I., Morell, A., Vicario, J. L. & Vilanova, R. 2020 Denoising autoencoders and LSTM-based artificial neural networks data processing for its application to internal model control in industrial environments – The wastewater treatment plant control case. *Sensors* **20** (13), 3743. https://doi.org/10.3390/S20133743.

Poch, M., Comas, J., Porro, J., Garrido-Baserba, M., Corominas, L. & Pijuan, M. 2014 Where are we in wastewater treatment plants data management? A review and a proposal. In *International Congress on Environmental Modelling and Software*, pp. 1450–1455. https://scholarsarchive.byu.edu/iemssconference.

Provotar, O. I., Linder, Y. M. & Veres, M. M. 2019 Unsupervised anomaly detection in time series using LSTM-based autoencoders. In *2019 IEEE International Conference on Advanced Trends in Information Theory, ATIT 2019 – Proceedings*, pp. 513–517. https://doi.org/10.1109/ATIT49449.2019.9030505.

Rieger, L., Alex, J., Gujer, W. & Siegrist, H. 2006 Modelling of aeration systems at wastewater treatment plants. *Water Science and Technology* **53** (4–5), 439–447. https://doi.org/10.2166/wst.2006.100.

Rolfe, M. D., Rice, C. J., Lucchini, S., Pin, C., Thompson, A., Cameron, A. D. S., Alston, M., Stringer, M. F., Betts, R. P., Baranyi, J., Peck, M. W. & Hinton, J. C. D. 2012 Lag phase is a distinct growth phase that prepares bacteria for exponential growth and involves transient metal accumulation. *Journal of Bacteriology* **194** (3), 686–701. https://doi.org/10.1128/JB.06112-11.

Seshan, S., Vries, D., Duren, M. v., Helm, A. v. d. & Poinapen, J. 2023 AI-based validation of wastewater treatment plant sensor data using an open data exchange architecture. *IOP Conference Series: Earth and Environmental Science* **1136** (1), 012055. https://doi.org/10.1088/1755-1315/1136/1/012055.

Siami-Namini, S., Tavakoli, N. & Siami Namin, A. 2019 A comparison of ARIMA and LSTM in forecasting time series. In *Proceedings – 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pp. 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227.

Therrien, J. D., Nicolaï, N. & Vanrolleghem, P. A. 2020 A critical review of the data pipeline: How wastewater system operation flows from data to intelligence. *Water Science and Technology* **82** (12), 2613–2634. https://doi.org/10.2166/WST.2020.393.

Yadav, P., Chandra, M., Fatima, N., Sarwar, S., Chaudhary, A., Saurabh, K. & Yadav, B. S. 2023 Predicting influent and effluent quality parameters for a UASB-based wastewater treatment plant in Asia covering data variations during COVID-19: A machine learning approach. *Water* **15** (4), 710. https://doi.org/10.3390/W15040710.

First received 3 July 2023; accepted in revised form 29 December 2023. Available online 17 January 2024