PAPER • OPEN ACCESS

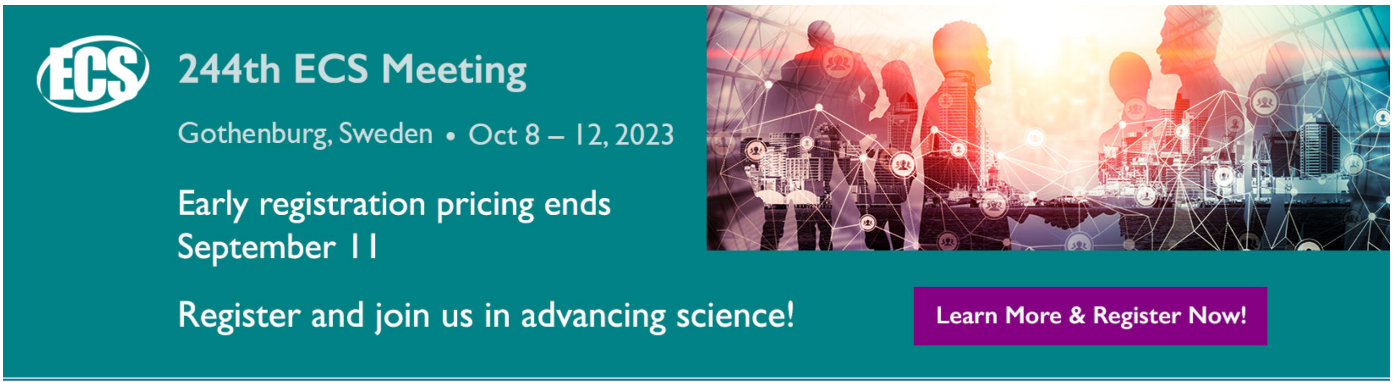
AI-based validation of wastewater treatment plant sensor data using an open data exchange architecture

To cite this article: Siddharth Seshan *et al* 2023 *IOP Conf. Ser.: Earth Environ. Sci.* **1136** 012055

View the article online for updates and enhancements.

You may also like

- The dynamics of discharge propagation and x-ray generation in nanosecond pulsed fast ionisation wave in 5 mbar nitrogen
  Bangdou Huang, Cheng Zhang, Jintao Qiu et al.
- Scaling of pulsed nanosecond capillary plasmas at different specific energy deposition
  Yifei Zhu, Svetlana M Starikovskaia, Natalia Yu Babaeva et al.
- Investigation of capillary nanosecond discharges in air at moderate pressure: comparison of experiments and 2D numerical modelling
  Andrei V Klochko, Svetlana M Starikovskaia, Zhongmin Xiong et al.



This content was downloaded from IP address 192.87.214.3 on 25/07/2023 at 09:18

IOP Conf. Series: Earth and Environmental Science 1136 (2023) 012055 doi:10.1088/1755-1315/1136/1/012055

AI-based validation of wastewater treatment plant sensor data using an open data exchange architecture

Siddharth Seshan[1], Dirk Vries[1], Maarten van Duren[2], Alex van der Helm[2] and Johann Poinapen[1]

[1] KWR Water Research Institute, Groningenhaven 7, 3433 PE, Nieuwegein, the Netherlands
[2] Waternet, PO Box 94370, 1090 GJ, Amsterdam, the Netherlands

siddharth.seshan@kwrwater.nl

Abstract. Typically, within the context of treatment plant-wide data, the quality of data can be impacted by sensor faults, sensor calibration issues, fouling of and obstruction to the sensors and connectivity problems between sensors, actuators and the data management system, therefore hampering advanced data driven monitoring and control of (critical) water operations. Here, a smart data validation scheme is proposed that validates sensor data from a wastewater treatment plant and is tightly integrated with the open-source data exchange system called FIWARE, an EU supported framework. The data validation application and FIWARE setup are integrated, tested and deployed at the water utility, Waternet. The validation scheme is based on an anomaly detector using (statistical) threshold techniques and a data reconciliation part that aggregates deep learning based autoencoder model predictions whenever an anomaly is detected. The autoencoder models proved to have a high accuracy and good reconciliation performance considering the variability of the signal. Furthermore, (near) real-time validated and raw data signals are relayed towards a dashboard. Finally, the validated data can be used as a screening for data ingested by another AI-based model that enables monitoring and smart control of the wastewater treatment plant in order to minimise greenhouse gas emissions and energy consumption while meeting effluent water quality standards.

1. Introduction

Wastewater treatment infrastructures get increasingly complex due to the need to meet stringent effluent quality standards such as the European Directive 91/271 on urban wastewater, but also due to resource recovery technologies and the increasing use of sensor data for advanced process monitoring and control of key variables. This leads to a stronger dependency on data. Therefore, ensuring high quality data has become even more relevant. For example, a water utility can be faced with (many) anomalous sensor data, i.e., missing values, outliers, irregular trend breaks in time series, bias due to connection errors, sensor fall-outs or wrong sensor calibrations, drifting due to sensor wear, fouling or obstruction. In addition, the multitude of sensors and their different characteristics and the availability of external data sources not only provide opportunities for intelligent data driven operation of wastewater treatment plants (WWTPs), but also a challenge in keeping the operability among different devices up-to-date. Analogously to the interconnection of smart devices as is the case in the Internet of Things (IoT), as IoT scales up, interoperability needs to be safeguarded in order to maintain operational performance and up-to-date information [1].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

In addition, WWTPs typically rely on multiple purification and sometimes resource recovery technologies which in turn are characterised by multigranular, interrelated nonlinear biological, physical and chemical processes. Furthermore, these systems are very dependent on environmental and operational conditions, and often have cyclic behaviour due to recycle streams. It is recognised that operation of WWTPs has evolved from sensor signal processing and using statistics and process identification in the seventies towards knowledge-based systems where sensor data is increasingly fed into data mining techniques and predictive analytics in order to capture complexity with the aim to support, possibly multi-criteria optimised, plant-wide control and decision support [2]. However, the quality of data is frequently hampered by the intrinsically challenging measurement conditions of the aquatic environment, which makes that on-line sensors can be affected by many faults [3].

Hence, the reliance on validated and clean data is increasingly becoming apparent and, moreover, open data management platforms are becoming attractive from (i) a software lifecycle assessment point of view and (ii) the prospect of managing heterogeneous data sources more efficiently.

Here, we present an automated data validation and reconciliation (DVR) framework which is embedded in an open data management system called FIWARE [4]. The DVR is one of the key steps in achieving intelligent, data-driven plant-wide control of a large WWTP called Amsterdam West, situated nearby Amsterdam, the Netherlands and operated by Waternet, the public water cycle utility of Amsterdam and surrounding areas. The Amsterdam West WWTP has a capacity of 1 million population equivalent and serves the city of Amsterdam. Currently, the control loops of Amsterdam West WWTP are for a large part locally distributed and dedicated to a single treatment process unit. To illustrate the context: the objective of the smart control application (not discussed in this work, see [5], [6] for more details) is to minimise energy consumption and N_2O emissions while meeting effluent quality criteria. N_2O has a large impact on greenhouse gas emissions and consequently, climate footprint.

In the following, we demonstrate the extension and coupling of FIWARE to the existing legacy data management system of Waternet in order to improve the utilisation of (near) real-time plant data for DVR of key sensor signals. For the Amsterdam West WWTP case, the sensor data of two key process parameters, nitrate (NO_3^-) and ammonium (NH_4^+), in the aerobic tank of the research lane's bioreactor unit were considered to demonstrate the prospect of using AI-based DVR and FIWARE as a data exchange platform.

2. Data management system

2.1. Distributed control system, information management and FIWARE4Water setup

Amsterdam West WWTP is real-time controlled by a distributed control system (DCS). The DCS stores and handles measured sensor data, e.g., online water quality and flow measurements, actuator data from e.g., valves and pumps, control-related data such as setpoints and local control settings. It also logs events, alarms and operator changes. Operators and water process engineers access the DCS via a graphical user interface.

Data is sent through the (legacy) Process Information Management System (PIMS) which acts as a historian and logs all process data, alarms and events. PIMS also hosts laboratory measurements, e.g., water quality and data signals manually validated by plant operators. All data and information are collected (near) real-time and can be accessed from office automation applications. Selected sensor data, i.e., NO_3^- and NH_4^+ concentrations, is forwarded to the FIWARE4Water setup which consists of FIWARE components, a dashboard and the DVR application, see Figure 1.

In turn, the FIWARE4Water setup consists of Docker™ containers running on a Windows VM in a Microsoft Azure™ cloud environment, each having dedicated predefined functionalities. The Docker containers contain the following components and are linked as follows:

- the IoT Agent JSON, ingests data via the DVR routines manager (see section 4.1. for more details) from the legacy system into the FIWARE setup and is coupled with the Orion-LD context broker;

- the Orion-LD Context Broker routes data between the IoT agent, the DVR application and notifies the Cygnus NGSI-LD component when new values are available
- the Cygnus NGSI-LD component writes all data that passes through Orion-LD to a PostgreSQL database;
- a PostgreSQL database stores data which has been ingested in FIWARE, as well as data generated by the DVR application;
- a Grafana dashboard queries the data from the PostgreSQL database which is visualised in real-time;
- a MongoDB database is used to store data models, FIWARE devices, FIWARE subscriptions, and other FIWARE configuration settings;
- a DVR docker containing the routines manager which handles FIWARE specific management tasks, e.g., creation and checks of data entities, subscriptions, etc. Furthermore, the routines manager serves as an application programming interface (API): it calls the relevant key components such as the anomaly detection methods, the trained autoencoder models to make a prediction and the reconciliation algorithm. As a final step, it outputs anomaly flags, predictions and reconciled data which is in turn passed through the IoT agent to the PostgreSQL database.

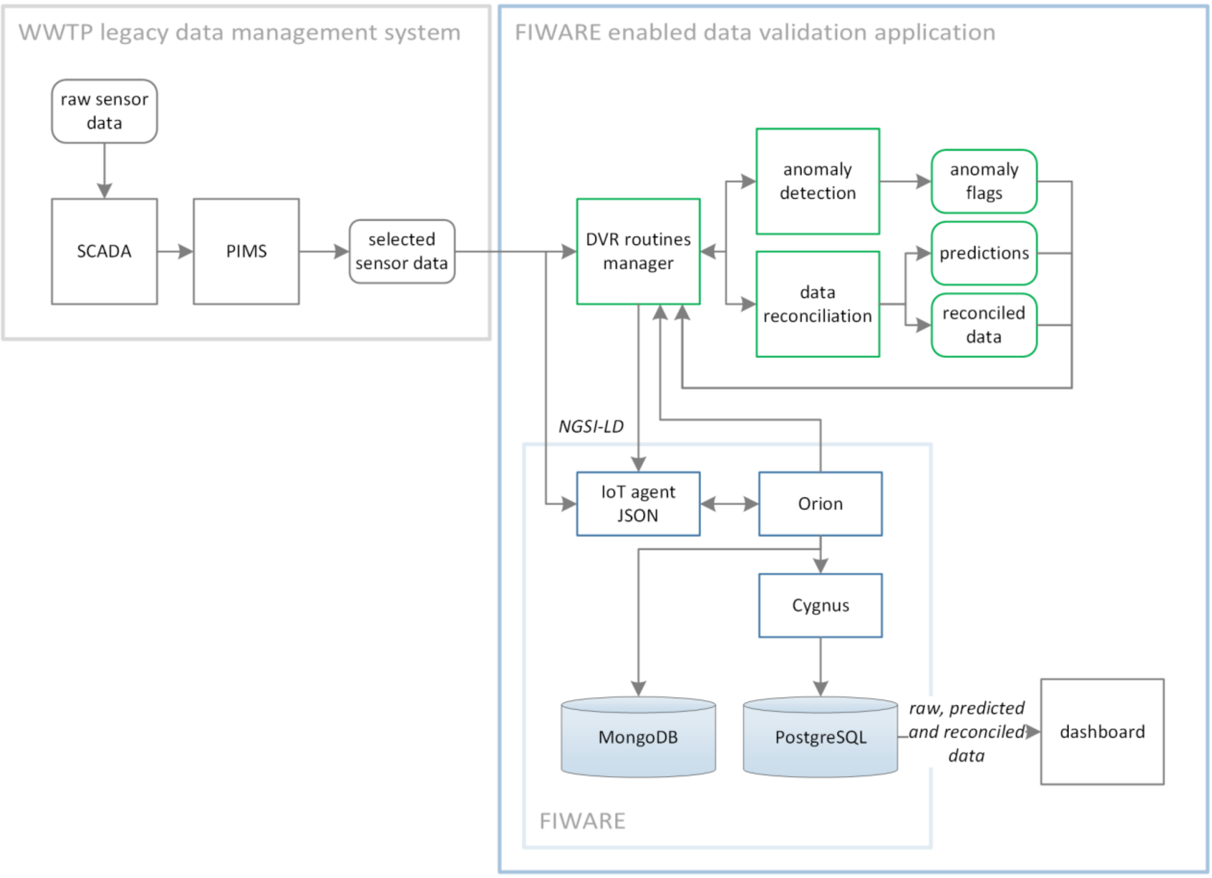


Figure 1: Schematic overview of the integration of FIWARE with the legacy control system.

Finally, raw, predictions from the models, and reconciled data from the DVR are visualised by the analytics and interactive visualisation web application Grafana by querying the PostgreSQL database.

2.2. FIWARE4Water data models

FIWARE relies on the NGSI-LD based data standard [7]. Smart data models for wastewater treatment have been specifically developed for this case and used in the current FIWARE setup, but have generic properties. Specifically, the following two NGSI-LD data models have been implemented for the DVR:

- a *WasteWaterTank*, an entity type used to describe the physical tanks in the bioreactor unit within the research lane. The properties of the entity are important process parameters, e.g., NO_3^- and NH_4^+, that are measured using online sensors.
- *WasteWaterSimulationResults*, a NGSI-LD entity that has been developed to describe the results provided from (data-driven) models. Note that this entity type represents digital-based solutions instead of wastewater treatment assets.

More details can be found in the smart data model repository available on GitHub [8]. In practice, the raw data as received from the legacy system is initially translated to NGSI-LD and then relayed to the IoT Agent JSON and the Orion-LD Context Broker. The data values specific to NGSI-LD entities are then accessed by the DVR routines manager. Similarly, once the DVR routines manager receives the output of the various DVR analytics, the data is once again transformed into the NGSI-LD data models, which are then sent back to the IoT Agent JSON.

3. Data Validation and Reconciliation

3.1. Procedure

The data validation consists of two components, first being the anomaly detection step where the raw data values are assessed as to being faulty or correct, and the second component includes the reconciliation of the anomalous values using predictions from trained, deep neural network models. The handling of data from and to the FIWARE4Water setup is taken care of by the FIWARE enabled DVR Routine Manager (FW-DVR). While streaming data is near real-time being ingested into FIWARE, the FW-DVR Routines Manager regulates when to trigger a given data validation routine, based on the availability of new data in FIWARE.

DVR follows a sequence of steps (Figure 2):

- The selected raw sensor data (NO_3^- and NH_4^+) is inputted initially into the anomaly detector to specifically flag faulty single values. The anomaly detector consists of two autoencoder models for each sensor signal, one representing short-term dynamics (5-minute) and one for long-term dynamics (30-minute). Furthermore, (historical) time series are inputted into the autoencoder models to conduct predictions that will be used in the reconciliation process. The autoencoder models are further explained in section 3.2.
- The single value anomaly detection is followed by flatline detection and flagging.
- The anomaly flags aid the data reconciliation process. Reconciliation is preceded by aggregation of predictions from the autoencoder models using exponential smoothening.
- Finally, three data outputs are provided by the data validation application, i.e., the raw sensor data, the autoencoder model predictions and reconciled data. The reconciled data constitutes raw sensor data during anomaly-free periods as well as aggregated, predicted values by autoencoder models during anomaly events.

3.2. Autoencoder models

Feedforward neural network models known as autoencoders were used for the purpose of data reconciliation. Autoencoders consists of two modules, an encoder and a decoder, where the encoder is trained for a given input of sensor data, to learn the underlying features and is represented in a reduced dimension. The decoder then reconstructs the input data, which is also the target variable [9], [10]. Conventionally, autoencoders are trained using only a single layer for the encoder and decoder each, however utilising deep autoencoders have known to yield better performance in the cases of non-linear and complex systems. Within wastewater treatment, autoencoder models are used for the purpose of reconstructing the input data which allows for such models to be used as soft sensors to predict the sludge volume index [11] and key effluent parameters such as biological oxygen demand [12]. Additionally, autoencoders have been used to denoise data from WWTPs, prior to its ingestion to data-driven based control strategies [13]. For the purpose of data validation, stacked denoising autoencoders have been used to detect faulty data and perform reconciliation for influent wastewater quality parameters [14] and additionally, deep autoencoders been used to reconcile missing or faulty values in wastewater treatment operations data [15]. In both cases, the prediction error between the reconstruction from the autoencoder model and the data evaluated was used for the fault detection.

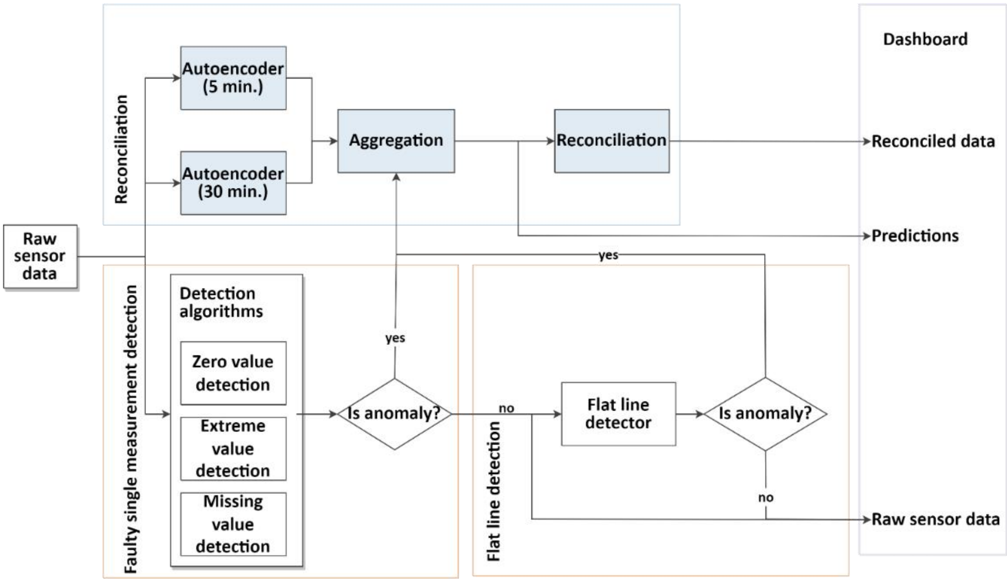


Figure 2: AI-based FIWARE enabled DVR Routines Manager.

In this study, over 70 different deep autoencoder models with varying architecture and hyperparameter combinations were investigated prior to model selection. The most relevant model structures (48 models) and the tuned hyperparameters are provided in [16]. The Python library TensorFlow was applied for development and training of the models. Initially, the model training was conducted to identify the type of layers that will be required within the model architecture, to be able to reconstruct the sensor behaviour with some degree of accuracy, where complexity was added with every iteration if deemed necessary. An architecture with only dense layers could not capture the non-linear effects present in the data. Hence, the model architecture was extended with layers that contain memorising units, i.e. more specifically, Long Short-Term Memory (LSTM) units. The use of LSTM layers resulted in an instant increase in prediction accuracy, which seems to confirm the autoregressive nature of the process. To mitigate the computational time involved when training LSTM models, the following was taken into account: (i) LSTM layers followed by dense layers in order to increase the depth of a model but keep the number of trainable parameters as small as possible, (ii) length of the input sequences, and (iii) use of dropout regularisation [17].

The resulting model structure settings are shown in Table 1 and training settings in Table 2. More details are provided in [16].

4. Real-time FIWARE-based data validation

4.1. Implementation of anomaly detection and reconciliation

For the purpose of recovering and reconciling anomalous NH_4^+ and NO_3^- signals, two granularities of 5-minute and 30-minute resolutions were chosen to perform the developed analytics. As a result, the

FW-DVR routines manager triggers the data validation methods every five or thirty minutes. Initially, the single value anomaly detection step is triggered, where the last available five or thirty data points are resampled and then assessed for being a faulty data point, thereby providing a flag for the current data point. Subsequentially, the flatline detection is triggered. For this use case, a minimum length of 3 consecutive data points has been used as a threshold (as available in the metadata for the sensor signal) to determine whether the points are part of a flat line. To perform this method, the manager therefore queries raw data points from the last 1.5 hours, that are resampled to the required resolutions separately, and inputted into the flat line detection. The method will provide a flag as an output which assesses the current data point to be part of a flat line, based on the values of the previous data points. The flags combined are used to judge whether the raw values should be reconciled using the predictions from the autoencoder models.

Table 1: Amount of historical input used for training of the autoencoder models.

Autoencoder Model	Resolution (minutes)	Historical Input (hours)	Historical Input (no. of timesteps)
5 minutes	5	3	36
30 minutes	30	24	48

Table 2: Model training settings.

Hyperparameter	Value	Comment
# 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 layers
Batch size	112 (5 min. autoencoder), 14 (30 min. autoencoder)	Based on a sensitivity analysis, it was concluded that a batch size representing 2 weeks of data yielded the best results.
Dropout rate (p)	NO_3^- – 0.06 NH_4^+ – 0.03	-

Faulty raw data values are reconciled by the predictions of autoencoder models. To perform a prediction with the autoencoder models, a certain period of historical data is necessary as input, in order to make a one-step ahead prediction. For the case of the 5-minute autoencoder, the last 3 hours of data are needed (36 data points) and for the 30-minute autoencoder, the last 24 hours of data (48 data points). The FW-DVR Routines Manager queries the historical data which are needed for the model predictions. Model predictions are then pushed back into FIWARE via the IoT Agent, which are then subsequently stored in the database under the relevant table. The autoencoder models make use of recursive predictions in case an anomaly is detected to prevent recursive propagation of anomalous data. The procedure is as follows: raw data values flagged as non-faulty and model predictions which would replace faulty data points are stored in intermediate processed data streams and pushed into the IoT Agent. Each (5-minute and 30-minute) data stream is linked to its own intermediate processed data stream, and the data stream is used as an input to the autoencoder models. In the final step, predictions from both the autoencoder models are used in the reconciliation of the data stream per sensor signal, via exponential smoothing and resampling to a targeted resolution of 15 minutes. Finally, the calculated reconciled data points are pushed to the IoT Agent and stored in the PostgreSQL database. The various outputs from the data validation application, namely the predictions from the autoencoder models and the reconciled signals are (near) real-time visualised in the Grafana dashboard.

4.2. Performance of the anomaly detection method
The performance of the autoencoder models is discussed in detail in [16]. The accuracy of the methods, as well as the autoencoder prediction performance are summarised below.

The determination scores of the autoencoder models (Table 3) show that a very high prediction accuracy was achieved, indicating the sensor data signals can accurately be reconstructed. Such a performance subsequently provides adequate confidence to use the predictions as provided by the autoencoder models for the purpose of reconciling anomalous values and to therefore realise validated NO_3^- and NH_4^+ data signals.

The predictions of the 30-minute NH_4^+ autoencoder for the test set are shown in Figure 3 for illustration purposes.

Table 3: R^2 Score of autoencoder models for NO_3^- and NH_4^+ signals.

	NO_3^- autoencoder model		**NH_4^+** autoencoder model	
	Train	*Test*	*Train*	*Test*
5 Minutes	0.97	0.95	0.90	0.90
30 Minutes	0.99	0.99	0.98	0.98

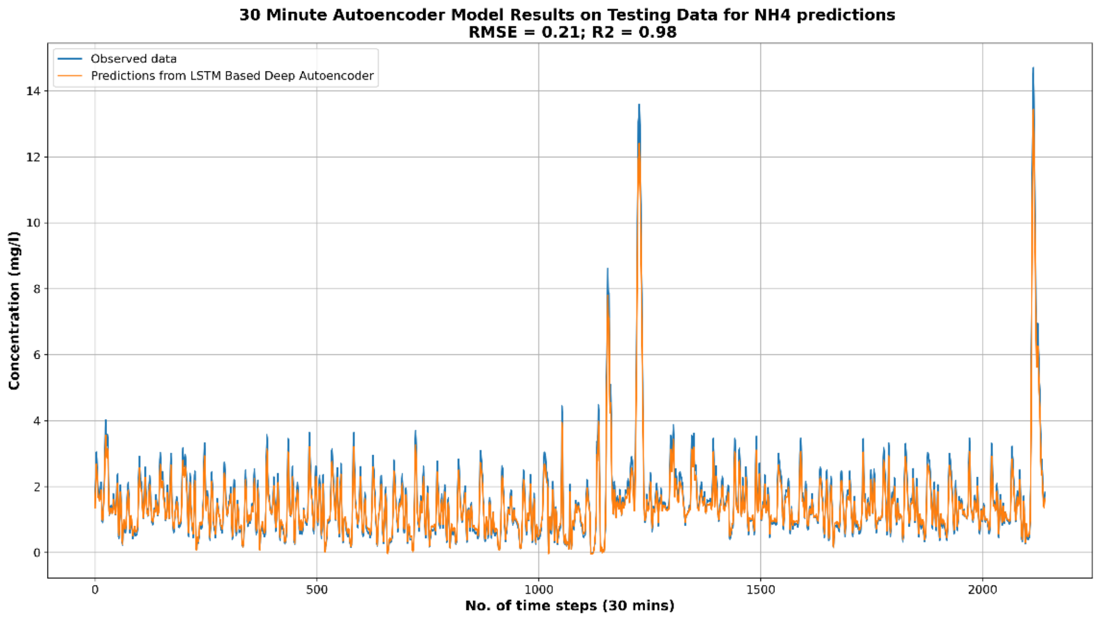


Figure 3: 30-minute autoencoder model predictions for the NH_4^+ signal (orange) and measured data (blue).

4.3. Visualisation
The DVR application has been running successfully near real-time. Yet, further tests are needed with synthetically generated anomalies and by running the DVR for an extended period of time to further test the reliability of the method. Early results indicate that reconciliation of long duration (more than 2 hours) anomalous events fail because of error propagation. Possible improvements are (i) extending the DVR with another autoencoder specifically trained on slow dynamics by using a larger than 30-minute time resolution, (ii) and/or including other sensor data signals. Future directions in development are to extend the DVR with a technique for classification of anomalies, and to deploy the DVR as a screening layer for soft sensors of key wastewater process variables [5] and the control agent for minimising the climate footprint while meeting regulatory effluent standards [6].

Measurement data of NO_3^- and NH_4^+, as well as predictions and reconciled values, are shown in **Figure 4** where the DVR is running in production in real-time at Waternet.

Figure 4: Grafana dashboard of the DVR of $NO3^-$ (left) and $NH4^+$ (right). Upper panels: raw signal, second row: 5-minute autoencoder predictions, third row: 30-minute autoencoder predictions, bottom row: reconciled data.

5. Concluding remarks

In this work, crucial sensor data for monitoring the performance of a WWTP, i.e., ammonium and nitrate, are checked for errors and anomalous data values are reconciled by model predictions using an advanced data validation and reconciliation (DVR) method which makes use of recurrent neural network (RNN) models and exponential smoothing. The DVR application is deployed in a cloud environment where sensor data is retrieved from the (legacy) WWTP's process information management system and relayed towards a setup using the open data exchange platform FIWARE. The FIWARE setup is extended to allow wastewater treatment specific data and data is successfully communicated, stored and processed near real-time to a DVR application.

The following can be concluded:

- The RNN autoencoder models are able to detect extreme or unexpected sensor values, as well as flatlines with differing duration;

- Reconciliation is successful for short to medium (i.e. approximately less than 2 hours) horizon lengths, but performance is dependent on the amplitude of the signal;
- The DVR provides a robust and accurate screening and correction layer for further use of sensor data in the monitoring and control applications – especially for anomaly events with a short duration;
- The DVR procedure can be used for other sensor signals using standardised data models.

In order to improve the performance and usability of the DVR, it is proposed to: (i) improve the accuracy of predictions, thereby extending the duration of reconciliation possible, by extending the DVR with another autoencoder specifically trained on slow dynamics, and/or by inclusion of other sensor data signals, (ii) extend the DVR with a technique for classification of anomalies

Acknowledgments
The presented research and development is part of the Fiware4Water project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 821036.

References
[1] Ahlgren B, Hidell M and Ngai ECH 2016 Internet of Things for Smart Cities: Interoperability and open data *IEEE Internet Computing* **20**(6) 52–56 doi: 10.1109/MIC.2016.124
[2] Poch M., J. Comas, J. Porro, M. Garrido-Baserba, L. Corominas and M. Pijuan 2014 Where are we in wastewater treatment plants data management? A review and a proposal *Int. Congress on Environmental Modelling and Software* pp 1450–1455 Accessed: May 01, 2022 [Online] Available: https://scholarsarchive.byu.edu/iemssconference
[3] Branisavljević N, D. Prodanović and D. Pavlović 2010 Automatic, semi-automatic and manual validation of urban drainage data *Water Science and Technology* **62**(5) 1013–1021 doi: 10.2166/WST.2010.350
[4] Cirillo F, G. Solmaz, E.L. Berz, M. Bauer, B. Cheng and E. Kovacs 2020 A standard-based open source IoT platform: FIWARE, *IEEE Internet of Things Magazine*, **2**(3) 12–18, doi: 10.1109/IOTM.0001.1800022
[5] van der Helm AWC, M.R. Haas, D. Amoiridis, M.H. Zandvoort, D. Vries and L. Echeverria 2021 Artificial intelligence for wastewater treatment – The development of a digital twin used to train a control optimization agent. Accessed: Feb. 02, 2022. [Online]. Available: https://aiww2021.com/event/smart-water-tools-in-waste-water-treatment/
[6] Echeverria L, N. Nievas, A.W.C. van der Helm, M.H. Zandvoort, M. Ribalta, X. Domingo and C. Fernandez 2021 Artificial intelligence for wastewater treatment – AI-agent-based service for optimal control *Proc. Amsterdam International Water Week*
[7] Privat G, A. Abid, F. le Gall, G. Tropea and J. A. Martinez 2021 Guidelines for modelling with NGSI-LD LD - Guidelines for modelling with NGSI 2 Contributors LD - Guidelines for modelling with NGSI. *Sophia Antipolis, France: ETSI*. Accessed: May 01, 2022. [Online]. Available: www.etsi.org
[8] Abella A and S. Seshan 2022 Data models for wastewater treatment. https://github.com/smart-data-models/dataModel.WasteWater
[9] Bengio Y, Ducharme R and Vincent P 2000 A neural probabilistic language model https://proceedings.neurips.cc/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html (accessed May 06, 2022)
[10] Provotar OI, Y. M. Linder and M.M. Veres 2019 Unsupervised anomaly detection in time series using LSTM-based autoencoders *IEEE Int. Conf. Advanced Trends in Information Theory, ATIT 2019 - Proceedings*, pp. 513–517. doi: 10.1109/ATIT49449.2019.9030505
[11] Safder U, J. Loy-Benitez, H.T. Nguyen and C.K. Yoo 2022 A hybrid extreme learning machine and deep belief network framework for sludge bulking monitoring in a dynamic wastewater

treatment process *Journal of Water Process Engineering* **46** 102580 doi: 10.1016/J.JWPE.2022.102580

[12] Babiker Y, M. Osman and W. Li 2020 Soft sensor modeling of key effluent parameters in wastewater treatment process based on SAE-NN doi: 10.1155/2020/6347625

[13] Pisa I, A. Morell, J.L. Vicario and R. Vilanova 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 doi: 10.3390/S20133743

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

[15] Ba-Alawi AH, J. Loy-Benitez, S. Kim and C. Yoo 2021 Missing data imputation and sensor self-validation towards a sustainable operation of wastewater treatment plants via deep variational residual autoencoders *Chemosphere* **288** 132647 1–15 doi: 10.1016/j.chemosphere.2021.132647

[16] Seshan S, D. Vries, A. van der Helm, D. Amoidiris, N. Nievas and L. Echeverria 2022 *D3.3 FIWARE-Enabled Applications for Wastewater Treatment* (Fiware4Water consortium) https://www.fiware4water.eu/sites/default/files/delivrables/F4W-D3.3-WEBSITE-FIWARE-AppWastewaterTreatment_final.pdf

[17] Hinton GE, N. Srivastava, A. Krizhevsky, I. Sutskever and R.R. Salakhutdinov 2012 *Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors* doi: 10.48550/arxiv.1207.0580 Available: https://arxiv.org/pdf/1207.0580.pdf