ICA2025OSLO
14th IWA Specialised Conference on
Instrumentation, Control and Automation

# Data-driven solutions transforming the wastewater sector: towards efficient and sustainable utilities

J Alferes*, F. Van Bauwel**, M. Van Loo***

*VITO, WaterKlimaatHub, Ostende Belgium, janelcy.alferescastano@vito.be

**VITO, WET Unit, Mol Belgium, frie.vanbauwel@vito.be

***VITO, WET Unit, Mol Belgium, maarten.vanloo@vito.be

**Abstract:** It is a reality that current and future challenges in the water sector cannot be faced with a traditional operation and considering water as a separate resource. Digitalization brings a huge potential to accelerate a successful transition towards more efficient, optimized and sustainable water and wastewater systems, with data as the key asset. In order to ensure data availability and reliability within data-driven solutions, real-time control, process optimization and decision making in practice, several methods are applied for data augmentation and data enrichment, with the overall goal of increasing the amount of data at the required frequency, and to improve its quality. Machine learning (ML) methods are applied to generate realistic synthetic data for the influent within water resource recovery facilities (WRRFs). Univariate and multivariate methods are also applied to verify the quality of sensor data and to provide validated and useful information for posterior use within the data-value chain. In both cases, emphasis is made on the applicability of the solutions at full-scale scenarios.

**Keywords**: data-driven solutions; data enrichment and augmentation; sustainable water system; wastewater treatment plants

The water and wastewater sector are experiencing a global transformation in view of facing current and future challenges. Traditional wastewater treatment plants (WWTPs) are on their way to be considered water resource recovery facilities (WRRF) as valuable source of clean water, nutrients and energy, and as a part of a greater ecosystem. In this transition, data-driven and Artificial Intelligence (AI)/ Machine Learning (ML) based solutions are becoming key to unlock the data value, to provide insights and to enhance decision-making towards optimization of resource recovery, autonomy and efficiency (Grievson et al., 2022, Zhang et al., 2023).

In the multilevel architecture presented in Figure 1, data is placed as the core element for a successful and economically viable transformation. Starting from the raw data collected at the physical systems, this innovative data-drive approach encompasses: (1) bottom layer with focus on data enrichment & augmentation, (2) top layer of digital twins and (3) layer intended for intelligent decision making, thus covering the whole data chain.

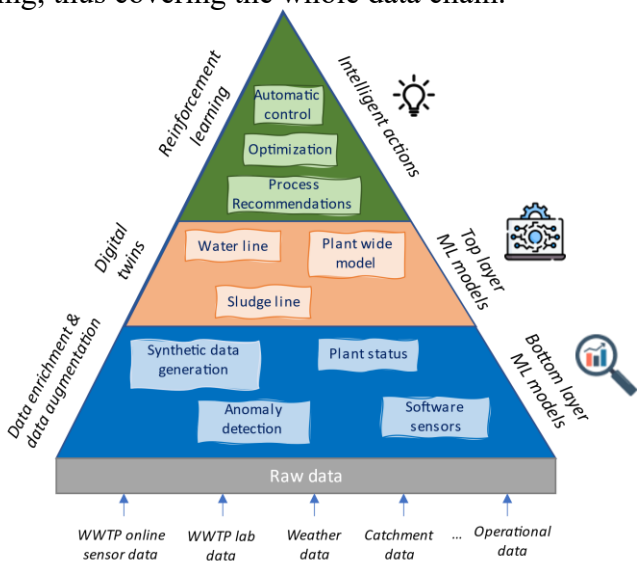


**Figure 1** Data-driven approach - unlocking the value of data for sustainable and efficient WRRFs

The availability of enough "good and rich" data becomes an enabler for the application of data-driven solutions for process optimization, control and decision making, which in practice can become a complex challenge (Therrien et al., 2020). First, large datasets of certain crucial parameters are not always available. Second, even if sensors may be providing large amounts of information, due to the harsh measurement environment present at WRRFs, sensors are prone to different type of faults affecting the reliability of the data (Samuelsson, 2017). Although several methods have been developed for synthetic data generation and data quality assessment purposes in different fields (Gaddam et al., 2020; Sit et al., 2020, Li et al., 2022), there is still an important gap in the water sector to bring these methods in practice, with a lack of comprehensive and systematic methodologies that take into account practical constraints in full-scale scenarios (Mamandipoor et al., 2020; Russo et al., 2022). In particular, parameter selection using a combination of domain knowledge and heuristics requires still significant manual intervention.

This paper will dive in the mentioned innovative data-driven approach, with emphasis on the bottom layer tools aimed at providing reliable and insightful data in WRRFs. Several data-driven modules have been developed and applied as shown in Figure 2 for data augmentation and enrichment, specifically with two objectives:

1. Increase data availability: use of ML algorithms for generation of synthetic data that mimics the dynamics of relevant wastewater components and flow, with focus on the influent to the WRRF. Deep learning models are used to identify patterns on the available data (e.g. online sensors and laboratory information), and to complete or to generate new profiles of similar characteristics at the required data frequency. With this aim, several models have been developed as shown in Table 1. Since the influent flow is dependent on specific operation rules of the pumping screws influenced by a level sensor, the choice was made to estimate the level in first place, and then convert the level to flow measurements. Figure 3 illustrates as examples the satisfactory results from the influent level prediction and the influent COD laboratory values estimations.
2. Generate data with verified quality from online sensors: several methods are applied to automatically detect anomalies, or abnormal data, in sensors that may be considered critical for the operation and control of key processes within the WWRFs. The anomalies considered are associated with intrinsic sensor errors (faulty sensors) or with events that may affect the sensor readings. A combination of several univariate statistic and simple tests, each of them aimed to detect a particular type of anomaly, is applied to generate an overall estimation of the data quality per each data point and the corresponding quality label. A multivariate detection of anomalies is also applied, based on a principal component analysis (PCA), in order to exploit the existing correlation between the measured variables (Miletic et al., 2019, Newhart et al., 2019). Two practical statistics (hotelling T2 and squared prediction error Q) are used to detect faults and to measure model adequacy.

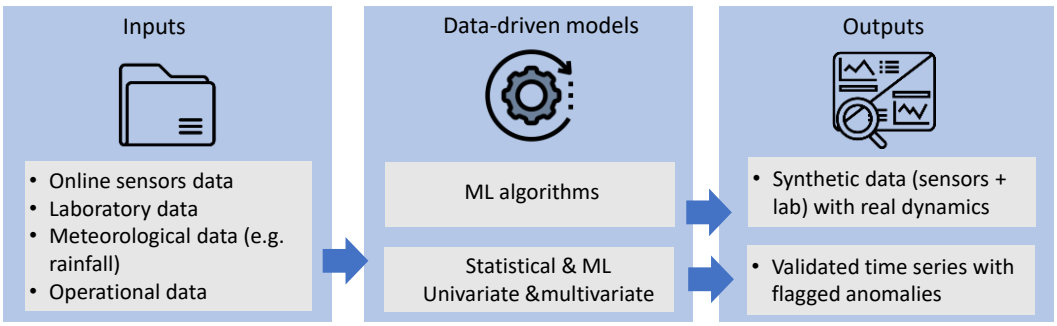


**Figure 1** Data-driven modules for data enrichment & data augmentation

**Table 1** Data-driven models to increase data availability

| | Models | | |
|---|---|---|---|
| | Influent flow | Influent concentrations | Influent lab data |
| Goal | Prediction of influent flow | Prediction water quality parameters (e.g. COD, NH4, TSS…) | (1) Fill in data gaps (lab samples dataset)<br>(2) Prediction lab values for the upcoming day |
| Model type | LSTM (level prediction), ANN (conversion from level to flow) | LSTM/ANN | Ensemble model |
| Input data | Inflow/Level, precipitation, temperature, dry weather flow | Precipitation, temperature, available sensor data at the influent, time information (day, week, holidays) | Precipitation, temperature, dry weather flow, lab data |
| Output data | ○ Inflow 12hr ahead (frequency =30 min)<br>○ Inflow 24hr ahead (frequency = hourly min) | Concentrations 24hr ahead (frequency =hourly) | Daily values for relevant parameters (e.g. COD, BOD, PO4, TSS, TOC, Ptot, Ntot) |

The figures below illustrate some examples of the relevant results from implementation of the different methods for a wastewater treatment plant in the Netherlands (Tilburg – 35000 p.e.). Concerning data availability, Figure 3a shows the estimated inflow level 12 hours ahead, with a satisfactory follow up of daily and seasonal dynamics. The influent flow is calculated in a second step, as it is dependent on specific operation rules of the screws influenced by the level measurements. On other hand, Figure 3b displays the suitable results for the laboratory gap filling for the chemical oxygen demand (COD) from the available lab dataset. The "data gaps" in this case are days for which no lab samples were taken or available. The same methodology has been applied to provide complete time series of laboratory data for several relevant parameters.

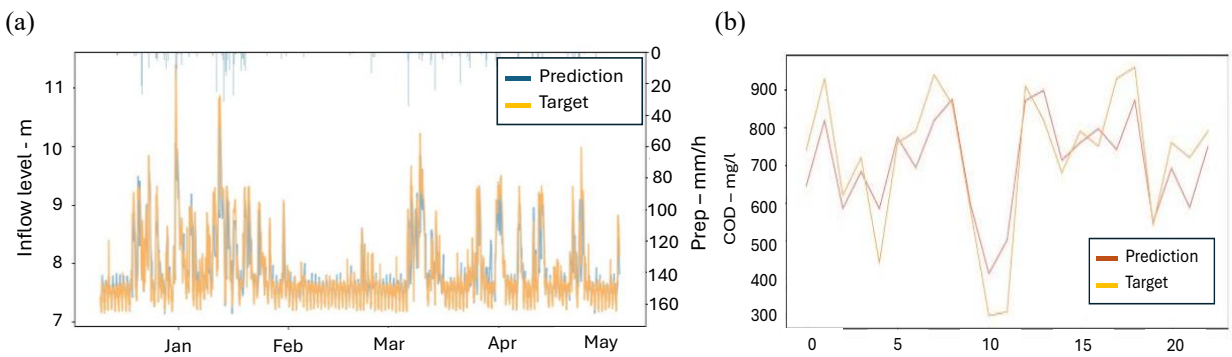


**Figure 3** Synthetic data generation. a) Influent flow. b) COD laboratory data gap filling

Concerning data validation, Figure 4a illustrates the obtained scores (data transformed in the PCA model describing around 90% of the data variability), for a historical dataset (one year) that includes measurements from nine on-line sensors (e.g. ammonia, nitrate, dissolved oxygen, temperature, phosphorus, nitrate) located at one of the biological lines, and influent flow. Each point in the plot represents a data observation vector. The vectors represent each variable and its contribution to the two first principal components (PC1, PC2). The multivariate statistics allowed for the detection of fault situations that affected several sensors (example shown in Figure 4b), adding to univariate detection of anomalies per sensor (results not shown). These statistics are used in a later stage to reveal which process variables are related to the fault.

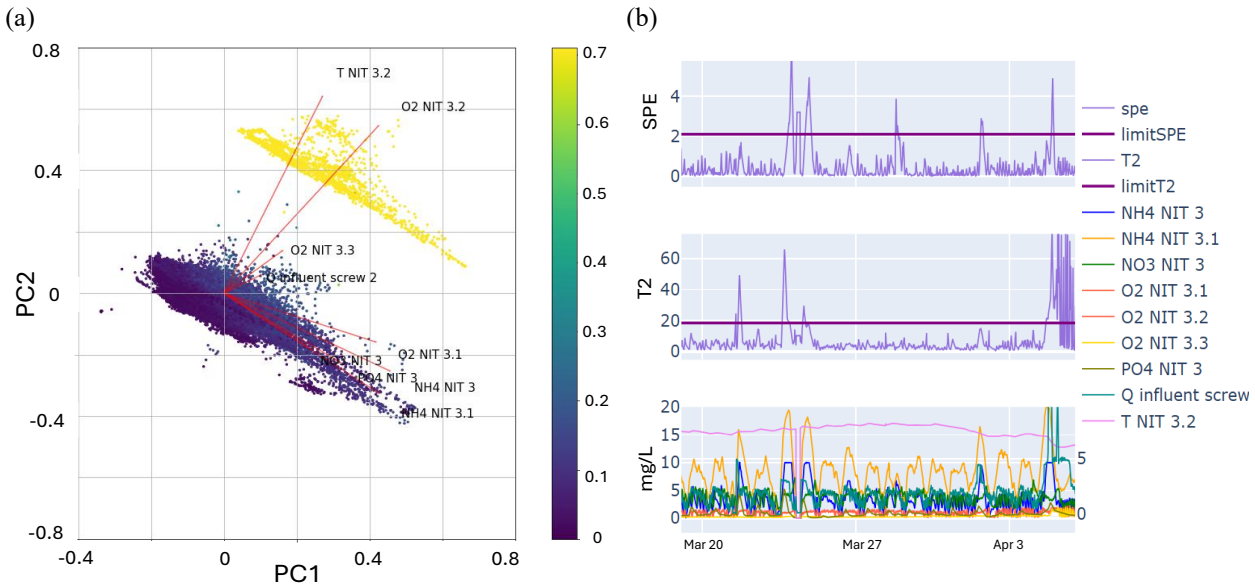


**Figure 4** Anomaly detection. a) scores for PCA implementation. b) multivariate statistics

Generated influent data (both for high frequency sensor data and laboratory values) and validated data are used afterwards to feed mechanistic and reinforcement learning models used within process modelling and process optimization strategies (top layers models in Figure 1), seeking to obtain the most value from the data and leveraging resource recovery and energy efficiency in WRRFs. All data-driven models are being integrated in a scalable environment to facilitate their portability, replication and scalability to any generic case study, in function of its digitalization level. Further details on the applied methodology an practical implementation of the different models will be provided in the full paper.

## REFERENCES

Grievson O., Holloway T. and Johnson B. (2022) A Strategic Digital Transformation for the Water Industry. IWA Publishing, 2022. https://doi.org/10.2166/9781789063400

Li F. and Vanrolleghem P.A. (2022) An influent generator for WRRF design and operation based on a recurrent neural network with multi-objective optimization using a genetic algorithm. Water Sci Technol, 85 (5): 1444–1453. doi: https://doi.org/10.2166/wst.2022.048

Gaddam A, Wilkin T , Angelova M and Gaddam J. (2020) Detecting Sensor Faults, Anomalies and Outliers in the Internet of Things (9), 511.

Mamandipoor B., Mahshid Majd M., Sheikhalishahi S., Modena C., and Osmani V. (2020) Monitoring and Detecting Faults in Wastewat er Treatment Plants using Deep Learning. Environmental Monitoring and Assessment (192).

Miletic I., Rieger L., Rosenthal A. and Schraa O. (2019) Get More Out of Your Data – Advanced Methods for Automated Data Integration, Cleaning and Augmentation. WEFTEC 2019 Proceedings, 3266–3271.

Newhart K. B., Holloway R. W., Hering A. S. and Cath T. Y. (2019) Data-driven performance analyses of wastewater treatment plants: a review. Water Research, 157, 498–513, https://doi.org/10.1016/j.watres.2019.03.030

Russo S., Moritz L., Hao W., Matthews B. and Villez K. (2020) Active learning for anomaly detection in environmental data. Environmental Modelling and Software 134

Samuelsson Oscar. (2017) PhD thesis on Sensor Fault Detection and Process Monitoring in Water Resource Recovery Facilities. Uppsala University

Sit M., Demiray B.Z., Xiang Z., Ewing G., Sermet Y. and Demir I. (2020) A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources." Water science and technology : a journal of the International Association on Water Pollution Research 82 12 (2020): 2635-2670 .

Therrien J.D., Nicolaï N. and Vanrolleghem P.A. (2020) A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. Water Sci Technol 82 (12): 2613–2634.

Zhang S., Jin Y., Chen W., Wang J., Wang Y. and Ren H.. (2023) Artificial intelligence in wastewater treatment: A data-driven analysis of status and trends. Chemosphere (336). https://doi.org/10.1016/j.chemosphere.2023.139163.