

METADATA: A MUST FOR THE DIGITAL TRANSITION OF WASTEWATER TREATMENT PLANTS

**Daniel Aguado García¹, Frank Blumensaat², Juan Antonio Baeza³, Kris Villez⁴,
María Victoria Ruano⁵, Oscar Samuelsson⁶, Queralt Plana⁷ and Janelcy Alferes⁸**

¹ CALAGUA – Unidad Mixta UV-UPV, Institut Universitari d'Investigació d'Enginyeria de l'Aigua i Medi Ambient – IIAMA, Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, Spain.

² ETH Zurich/Eawag, Institute of Environmental Engineering, Chair of Urban Water Management Systems, Zurich, Switzerland

³ Department of Chemical, Biological and Environmental Engineering, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain


⁴ Oak Ridge National Laboratory, Oak Ridge, TN, USA

⁵ Chemical Engineering Department, Universitat de València, Burjassot, Spain

⁶ IVL Swedish Environmental Research Institute, Sweden

⁷ modelEAU, Université Laval, Québec, Canada

⁸ VITO - Vision on technology, Boeretang 200, BE-2400 MOL, Belgium.

¹  daaggar@hma.upv.es, ² Frank.Blumensaat@eawag.ch, ³ JuanAntonio.Baeza@uab.cat,
⁴ villezk@ornl.gov, ⁵ m.victoria.ruano@uv.es, ⁶ oscar.samuelsson@ivl.se,
⁷ queralt.plana@gmail.com, ⁸ janelcy.alferes@gmail.com

Abstract

The increment in the number and diversity of available (and affordable) sensors together with the advances in information and communications technologies have made it possible to routinely measure and collect large amounts of data at wastewater treatment plants (WWTPs). This enormous amount of available data has boosted the interest in applying sound data-driven solutions to improve the current normal daily operation of these facilities. However, to have a real impact in current operation practices, useful information from the massive amount of data available should be extracted and turned into actionable knowledge.

Machine learning (ML) techniques can search into large amounts of data to reveal patterns that a priori are not evident. ML can be applied to develop high-performance algorithms useful for different tasks such as pattern recognition, anomaly detection, clustering, visualization, classification, and regression. These ML algorithms are very good for data interpolation, but its extrapolation capabilities are low. Hence, the data available for training these data-driven models require data covering the complete space for the independent variables. A significant amount of data is required for this purpose, but data of good quality.

To transform big data into smart data, giving value to the massive amount of data collected, it is of paramount importance to guarantee data quality to avoid “garbage in – garbage out”. The reliability of on-line measurements is a hard challenge in the wastewater sector. Wastewater is a harsh environment and poses a significant challenge to achieve sensor accuracy, precision, and responsiveness during long-term use. Despite the huge amount of data that are currently being recorded at WWTPs, in many cases nothing is yet being done with them (resulting in data graveyards). Moreover, the use of the data collected is indeed very limited due to the lack of documentation of the data generation process and the lack of data quality assessment.

Metadata is descriptive information of the collected data, such as the original purpose, the data-generating devices, the quality, and the context. Metadata is needed to clearly identify the data that should be used for the development of data-driven models. These data should be selected from the same category. If we include data that shouldn't be in the same data set because they were obtained under different operational conditions, this would lead to unreliable model predictions. ML algorithms learn from data, thus to be useful tools and to really improve the decision-making process in WWTP operation and control, representative, reliable, annotated and high-quality data are needed.

Effective digitalization requires the cultivation of good meta-data management practices. Unfortunately, there are no wastewater-specific guidelines available to the production, selection, prioritization, and management of meta-data. To address this challenge, the IWA Task Group on Meta-Data Collection and Organisation (MetaCO TG to which the authors of this paper belong) which has been supported by the International Water Association since 2020 will soon finish the scientific and technical report containing such guidelines specifically for WWTPs. This paper highlights why meta-data should be considered when collecting data as part of good digitalisation practices.

Keywords

Digitalization, metadata, wastewater treatment plant, water resource recovery facility.

1 INTRODUCTION

Water is a scarce natural resource, essential for life and for the exercise of the vast majority of economic activities. After its use, whether for human or industrial activities, its composition and quality notably degrade, and it becomes wastewater. Treatment of urban wastewater is essential to protect the aquatic environment as well as the human health [1]. The importance of access to clean water and sanitation as well as clean water at sea is embedded in goals 6 and 14 of the United Nations' Sustainable Development Goals (SDGs). Wastewater treatment is essential to reduce marine pollution (SDG 14); the development of energy-efficient treatment and control solutions for pollutants' removal (or resource recovery) and the production of renewable energy from the organic matter, can contribute to the affordable and clean energy goal (SDG 7) as well as to achieve more sustainable cities (SDG 11). Water scarcity, poor water quality and inadequate sanitation negatively impact food security.

Wastewater treatment is currently undergoing a paradigm shift as a result of the transition to the circular economy. This paradigm change consists in ceasing to consider wastewater as a mere waste, to consider it a source of resources and energy [2]. Wastewater Treatment Plants (WWTPs) can contribute to the circular economy in different ways, such as producing a water effluent that can be reused, resource recovery (e.g. phosphorus is a macro-nutrient essential for life and a non-renewable resource that can be recovered in WWTPs) and clean energy production through anaerobic digestion of the organic matter contained in the influent wastewater.

Traditionally, wastewater treatment has been centralized, due to the economy of scale associated with the construction of WWTPs. However, in less populated areas, other technologies have been implemented due to different reasons: less inversion is possible, the lack of suitable qualified personal to operate the WWTP, and the need to treat very low flow rates as well as to deal with strong fluctuations in influent flow and composition. As a consequence, many different treatment schemes and configurations for wastewater treatment can be found today, even in the same city. Moreover, the type and level of instrumentation, control and automation varies significantly from one WWTP to another. There are analysers whose installation and maintenance costs (require frequent maintenance by skilled staff) may not be covered by the energy cost savings on small or remotely located WWTPs. All these factors make that the number of analogical and digital signals registered vary widely from less than 500 to more than 30,000 from one WWTP to other [3].

Recent advances in information and communications technologies (ICT), online sensors, and autonomous energy supplies make ubiquitous sensing of WWTP viable today, even in remote locations. The monitoring capabilities offered by the ICT will greatly improve the decision-making process for design, operation, and control, but this will only be realized if the data produced by and sent to devices can be trusted with very high reliability [4].

The amount and diversity of available sensors in WWTPs and other data from the processes that take place in these facilities has increased massively in the last decades [5]. Thanks to increased efficiency in communication networks and extreme reductions in data storage costs, data

collection is extremely scalable which means today's WWTPs have entered the era of big data by covering the three V's introduced by Doug Laney in 2001 [6]. Volume (a high amount of data is gathered from the instrumentation deployed at WWTPs), velocity (data values are produced at high speed, with sensors recording data every second or less) and variety (data from WWTPs is heterogeneous: different types of sensors, data from laboratory analysis, spectrophotometry, omics, images...). More recently other V's have been added like veracity (data and sources must be trusted) and value (to refer that big data only is only of value when presents high veracity with low vulnerability).

Variety and veracity represent a global and big challenge for making the most of the available data in WWTPs. The heterogeneity of data types, including traditional univariate time series (e.g., dissolved oxygen, temperature...) structured multivariate data (e.g., from spectrophotometry) as well as unstructured data (e.g., omics, images), makes the creation of a data fusion processing pipeline to extract and synthesize valuable information from all available data sources challenging and extremely difficult to apply in a large-scale fashion. This, in turn, affects how data quality assessment and quality control is ensured in the WWTP sector. The combination of data variety together with the exposition of the instrumentation deployed in the facilities to the hostile environment that WWTP represent (wastewater, particles, sludge, corrosive gases...) means that ensuring fit-for-purpose data quality remains a tough challenge today. The hostile environment in which these sensors are installed make the instrumentation prone to malfunction (bias, drift, precision degradation, missing data....) leading to poor data quality and reduced sensor accuracy and reliability [7].

Since external data is now less costly and ubiquitously available more data is still available at WWTPs. Consequently, information on important boundary conditions such as population estimates/ behaviour, weather ..., can be fed into supervisory control and data acquisition (SCADA) systems and wastewater services. Therefore, there is a realistic possibility to literally "drown" in data, especially if data management tools are not up to the task regarding quality checks and filtering out unreliable information. Many methods are available from data science [8], and the community seems to be familiar with them on a theoretical level. However, in practice, the developed capacities are not sufficient to harvest their full potential [4]. Transparent and standardized data treatment protocols may help to achieve this. Extracting information from the available data is not only important for the facilities themselves but also for other entities that can also be interested in WWTP data because they contain valuable information on societal behaviour (eg. spatial and temporal variations on drugs use as well as viruses tracking in wastewater-based epidemiology) or the expected quality of aquatic ecosystems.

Effective digitalization requires the cultivation of good meta-data management practices. Unfortunately, there are no wastewater-specific guidelines available to the production, selection, prioritization, and management of meta-data. To address this challenge, the IWA Task Group on Meta-Data Collection and Organisation (MetaCO TG to which the authors of this paper belong) which has been supported by the International Water Association since 2020 will soon finish the scientific and technical report containing such guidelines specifically for WWTPs. This paper highlights why meta-data should be considered when collecting data as part of good digitalisation practices.

2 META-DATA

Facilities usually feature many sensors that generate a large amount of data from the process (see Figure 1). It is important to highlight that to obtain accurate and useful data, every sensor:

- should be correctly installed (considering that it has also to be accessible for calibration, maintenance, and replacement when necessary),

- have a purpose (a reason to be there)
- and be adequately maintained by process operators according to manufacturer's instructions.

If only the numerical values of the variables are stored (which is the standard practice in most WWTPs), after a certain time, it will not be possible to know if the sensor was properly maintained, how many times it was calibrated, if it was replaced, the exact position where it was located or if it was moved temporarily to another position of the facility to gather data from another stream (e.g., influent, anoxic reactor, aerobic reactor, effluent of the treatment chain....) which could exhibit a completely different range of recorded values. All this information is of paramount importance to enable the current use of the data by an expert in data science as well as the future use of the data.

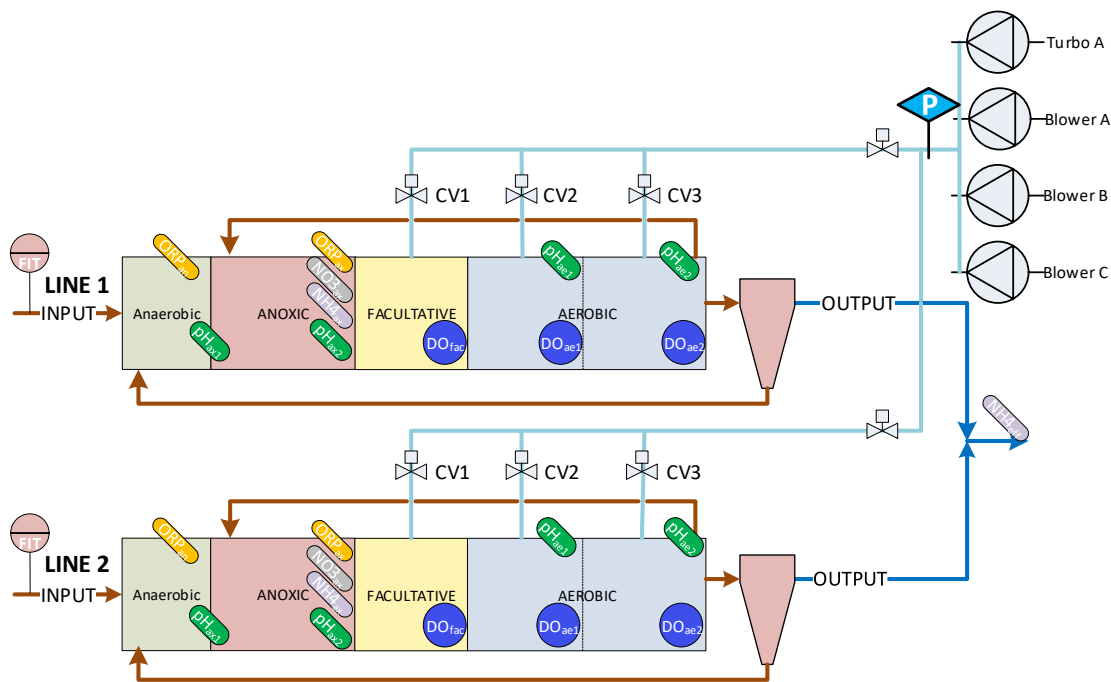


Figure 1. Layout of two parallel lines of a large WWTP featuring biological nutrient removal showing where each sensor is located within the biological reactor.

As can be seen in Figure 1, the same type of sensor has been installed in different locations along the water-line of the facility. For instance, in this case, six dissolved oxygen (DO) sensors have been located in different parts of the biological reactor, eight pH sensors, and so on. The sensors have been installed in different positions because variations in the recorded values of the variable are expected. The locations have been selected to gather information from the biological processes taking place as well as for control purposes. Note that if the location of each sensor would not be stored together with the recorded data, how would a data scientist be able to extract valuable information from the recorded data? In contrast, sometimes for variables that are critical for the process as well as for cheap and ubiquitous sensors such as pressure and temperature, the sensor is duplicated to obtain a reference measurement. Thus, in these cases two sensors measuring the same variable are located in the same reactor close to each other (and similar recorded values from both sensors are expected). The reason is to achieve hardware redundancy for data quality assessment. Thus, to extract valuable information from the data and to enable the use and reuse and future use of the gathered data for multiple purposes, additional information describing the context of the data should be also stored. This additional descriptive information is what is known as meta-data.

Meta-data is any descriptive information that is required and/or useful to interpret the data from a sensor or from the laboratory. There are several pieces of basic information that should be collected as much as possible, like the following aspects (full-details and extended information will be available soon in the scientific and technical report specifically targeted for WWTPs):

- Unit of measurement
- Measurement range
- Sensor location
- Sensor ID
- Temporal resolution
- If any change of units is applied
- If any type of missing data imputation is applied
- Operational state (operational, calibration, validation, maintenance)
- Purpose of the sensor
- Operational condition of the plant (normal conditions, toxic spill, dry-weather conditions, type of mechanical failure,...)

The first and foremost evident piece of information required to interpret the numerical value of a variable is the measuring unit which allows the correct interpretation of the values recorded.

The measuring range describes the interval along which the value of the measured variable must be situated so that the technical specifications of the sensor are met and thus they are the possible values that can be generated by a properly functioning sensor. Knowing the lowest and highest value that can be produced by a sensor that is functioning properly, can be useful to detect a malfunctioning sensor.

Maintenance actions often require a change in the exact location of the sensor, and occasionally a sensor can be temporary moved to another position within the facility for a given period of time to measure the variable of interest in another stream. These changes in position can lead to a complete different range of values (e.g., an expensive ammonia sensor moved from the aerobic reactor to the anoxic reactor, or to the influent or effluent of the facility), making the recorded data not always representative of the original measurement location. Therefore, logging the exact location of sensors is important being ideal that they would incorporate a position measurement system that enable automatic logging. The location is highly important because it can provide the context for the values interpretation, in addition to being also useful for the detection, isolation and diagnosis of anomalous values.

Knowing the purpose of a sensor can assist on to taking decisions on the prioritization of maintenance actions and it can also help a data scientist in deciding whether some signals could be useful to develop a data-driven model.

The operational condition of the plant (normal conditions, toxic spill, dry-weather conditions, type of mechanical failure, jammed valve...) can be considered contextual information that can be added through annotation. Thus, it is normally a manual process executed by a human expert. The information provided is of great help for the later use of the data. Normally this information is provided in a freeform text since the type and number of events that can occur is very large. However, free text allows different terms used for the same event, for instance if a different person is annotating. This can be avoided by providing a set of labels with an accepted description of most of the possible events and let a final option for including a free text in case the expert considers that none of the labels describe what it is wanted to reflect on the observed data.

It should be noted that the more unstructured would be the meta-data it would also be more difficult to managed and analysed in a conventional way.

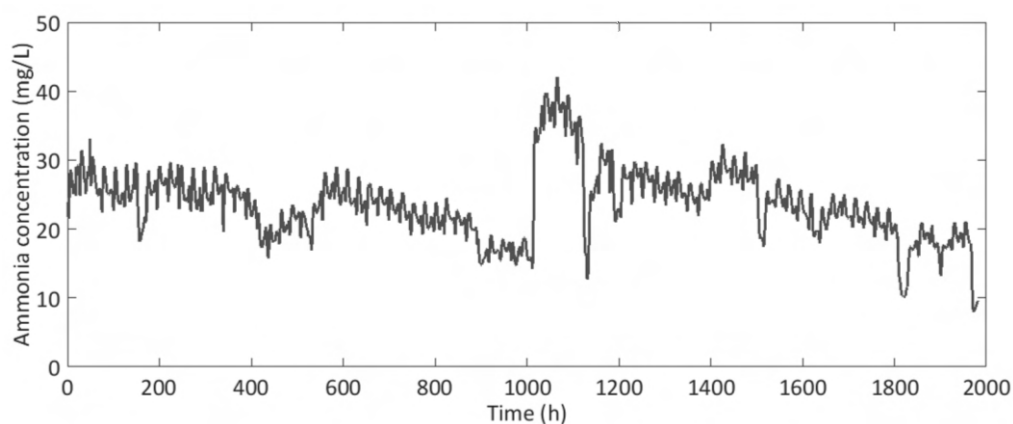
3 ILLUSTRATIVE EXAMPLE

Figure 2a shows the temporal evolution of ammonium concentration at the influent of the biological process of a large WWTP along nearly three months of operation. In this figure it can be seen noticeable variations in the ammonia concentration. Apart from a daily pattern, a temporal trend can be seen together with several relative abrupt changes in the recorded values of ammonia concentration.

In Figure 2b, the same variable and period is depicted, together with information of the observed changes by the process expert. As can be seen, Figure 2b is much more informative and useful, from the point of view of a later and future use of the data, for instance by a Data Scientist (with little knowledge on wastewater treatment processes). It should be highlighted how these pieces of information provided as meta-data can assist in an automation of the data triage to a high degree, and how the need for subjective assessment of the recorded time series has been significantly reduced, fact that will boost the trust in the data-driven predictions and decision-making.

The supervised approach in machine learning applications need a labelled dataset with the real anomalies and the normal-operating conditions data (i.e., clean data) for training and testing the data-driven models used for classification purposes as well as those used prediction applications. Thus, meta-data will play a central role to implement supervised machine learning models and intelligent tools in practice.

The most recent developments in data mining, machine learning, and artificial intelligence promise to extract the maximum of information from historical data. However, not understanding the process and the conditions under which data is collected makes even the best of algorithms fail. The use of WWTP data after they has been collected is typically limited due to a lack of documentation on the data generating process and a lack of data quality assessment. The scalability of the AI methods is expected to enable an effective use of data from multiple, uncoordinated sources, including data sources outside of the wastewater resource recovery plant's fence (e.g., weather forecasting data) to provide answers to questions through understanding new relationships that are ever-larger in scope. Unfortunately, this will remain a daydream unless the raw data becomes traceable.



(a)

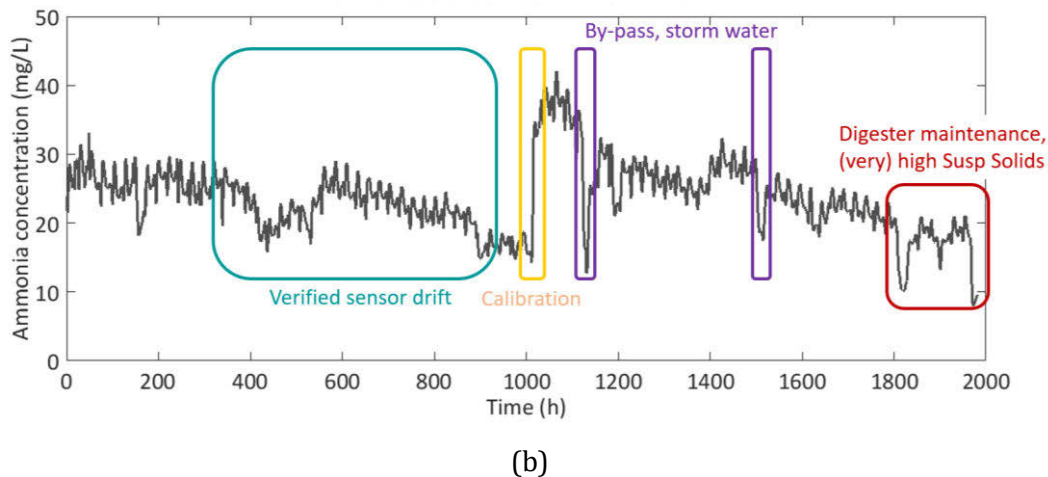


Figure 2. Time evolution of influent ammonium concentration along almost three months of operation: (a) only recorded values of the sensor including the units of measurement (b) adding process expert annotation.

4 CONCLUSIONS

This paper has illustrated the importance of meta-data to make it possible to transform raw data into intelligent actions for the operation of WWTPs and WRRFs, and it has been highlighted why meta-data should be considered when collecting data as part of good digitalisation practices. The main take-home messages are the following:

- Nowadays, a large amount of data is collected in WWTPs and WRRFs, but the almost complete absence of meta-data is compromising their potential use as well as their future use of the data, allowing their loss as data graveyards.
- There are no wastewater-specific guidelines available to the production, selection, prioritization, and management of meta-data.
- Meta-data greatly facilitates the interpretation of the collected data, helping to avoid the saying “data rich but information poor” that characterises many wastewater treatment data sets around the world.
- Meta-data plays a central role to implement useful supervised machine learning models and intelligent tools in practice.
- Systematic meta-data management will make it possible the future leverage of sensor and laboratory data, increasing its value.
- To really improve the decision-making process in the operation and real-time control of WWTPs and WRRFs, based the on-line collected data, there is need of annotated, representative, reliable and high-quality data.
- The IWA Task Group on Meta-Data Collection and Organisation (MetaCO TG to which the authors of this paper belong) will soon finish the scientific and technical report containing such guidelines specifically for WWTPs.

5 COPYRIGHT NOTICE

This research is sponsored by the US Department of Energy (DOE), Office of Energy Efficiency and Renewable Energy, Advanced Manufacturing Office, under contract DE-AC05-00OR22725 with UT-Battelle LLC. This manuscript has been authored by UTBattelle LLC under contract DE-AC05-00OR22725 with DOE. The US government retains—and the publisher, by accepting the article

for publication, acknowledges that the US government retains—a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

6 REFERENCES

- [1] Metcalf & Eddy, G. Tchobanoglous, H.D. Stensel, R. Ryujiro Tsuchihashi, F. Burton (2013). *Wastewater Engineering: Treatment and Resource Recovery*. McGraw-Hill Education. ISBN: 978-0073401188.
- [2] E. Neczaj, A. Grosser (2018). Circular Economy in Wastewater Treatment Plant—Challenges and Barriers. *Proceedings*; 2(11):614. <https://doi.org/10.3390/proceedings2110614>.
- [3] Ll. Corominas, M. Garrido-Baserba, K. Villez, G. Olsson, U. Cortés, M. Poch (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques, *Environmental Modelling & Software*, 106, 89-103, doi: 10.1016/j.envsoft.2017.11.023
- [4] F. Blumensaat, J.P. Leitão, C. Ort, J. Rieckermann, A. Scheidegger, P.A. Vanrolleghem, K. Villez (2019). How Urban Storm- and Wastewater Management Prepares for Emerging Opportunities and Threats: Digital Transformation, Ubiquitous Sensing, New Data Sources, and Beyond - A Horizon Scan. *Environmental Science & Technology* 2019 53 (15), 8488-8498 doi:10.1021/acs.est.8b06481.
- [5] G. Olsson (2012). ICA and me — A subjective review. *Water Research*, 46 (6), pp. 1585-1624, 10.1016/j.watres.2011.12.054
- [6] Doug Laney, 3d data management: Controlling data volume, velocity and variety, META Group Research Note 6 (2001).
- [7] O. Samuelsson, A. Björk, J. Zambrano, B. Carlsson (2018). Fault signatures and bias progression in dissolved oxygen sensors. *Water Science & Technology* 78 (5), 1034-1044. doi: 10.2166/wst.2018.3501979.
- [8] K.B. Newhart, R. W. Holloway, A.S. Hering, T.Y. Cath (2019). Data-driven performance analyses of wastewater treatment plants: A review. *Water Research*, 157, 498-513. doi: 10.1016/j.watres.2019.03.030.