A critical review of the data pipeline: how wastewater system operation flows from data to intelligence

Jean-David Therrien IWA, Niels Nicolaï IWA and Peter A. Vanrolleghem IWA

ABSTRACT

Faced with an unprecedented amount of data coming from evermore ubiquitous sensors, the wastewater treatment community has been hard at work to develop new monitoring systems, models and controllers to bridge the gap between current practice and data-driven, smart water systems. For additional sensor data and models to have an appreciable impact, however, they must be relevant enough to be looked at by busy water professionals; be clear enough to be understood; be reliable enough to be believed and be convincing enough to be acted upon. Failure to attain any one of those aspects can be a fatal blow to the adoption of even the most promising new measurement technology. This review paper examines the state-of-the-art in the transformation of raw data into actionable insight, specifically for water resource recovery facility (WRRF) operation. Sources of difficulties found along the way are pinpointed, while also exploring possible paths towards improving the value of collected data for all stakeholders, i.e., all personnel that have a stake in the good and efficient operation of a WRRF.

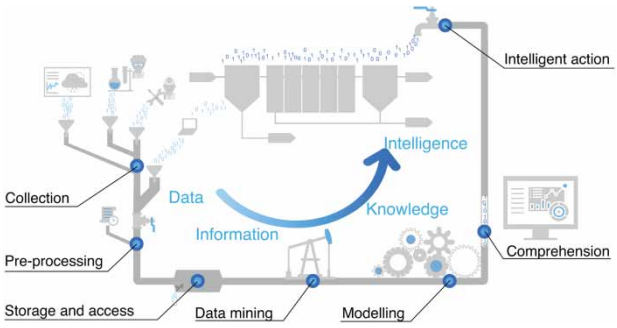
Key words | data treatment, digitalization, digital twin, metadata, wastewater modelling, water resource recovery facilities

Jean-David Therrien IWA (corresponding author)
Niels Nicolaï IWA
Peter A. Vanrolleghem IWA
modelEAU, Université Laval,
1065, Avenue de la Médecine, Québec,
Canada,
QC G1 V 0A6
E-mail: *jean-david.therrien.1@ulaval.ca*

HIGHLIGHTS

- Data can be abstract: with the data pipeline concept, issues are clarified.
- Digitalization of wastewater has begun, but it is lagging behind drinking water.
- Reliable digital twins need good quality data, but reaching that goal presents unique challenges.
- Collecting data without a proper strategy leads to data graveyards.
- More collaboration between data and water experts is critical to better use wastewater data.

GRAPHICAL ABSTRACT



doi: 10.2166/wst.2020.393

INTRODUCTION

Stringent limits on emissions, increasing water scarcity and rapid urbanization put a major strain on current wastewater systems. This means that the wastewater sector will have to adapt, and fast, if it is to fulfill its role as keeper of public and environmental health. In response to these stressors, WWTPs are increasingly being repurposed as water resource recovery facilities (WRRFs) (Water Environment Federation 2014). This value extraction is enabled by the deployment of a very diverse array of processes, from nutrient recovery to energy generation.

Moreover, the wastewater research community has been hard at work in recent years to develop new process models (Al-Omari *et al.* 2015; Mannina *et al.* 2016; Spérandio *et al.* 2016; Amaral *et al.* 2017; Rittmann *et al.* 2018; Varga *et al.* 2018), new process control strategies (Rieger *et al.* 2013; Jimenez *et al.* 2015; Regmi *et al.* 2015; Hernández-del-Olmo *et al.* 2016; Solon *et al.* 2017; Revollar *et al.* 2020) and new process monitoring systems (Alferes & Vanrolleghem 2016; Russo *et al.* 2019) to bridge the gap between current operation and the state-of-the-art control practices needed to efficiently and optimally run the multitude of processes involved in resource recovery given constrained capital and operational budgets.

These new developments share a need for data which has historically been difficult to collect. However, with sensor prices dropping, the increasing ubiquity of wireless communication and the proliferation of mobile devices able to ceaselessly gather information and perform sophisticated calculations, WWTPs have in recent years been exposed to an unprecedented amount of data. These trends span much wider than only the wastewater field, hence researchers from other disciplines have also dealt with this massive influx of data and have emerged with entirely new types of models leveraging machine learning and artificial intelligence. The interest in applying those new data-driven models to WRRF operation and control has recently become very strong. Some of the most discussed potential applications include the development of adaptive plant models, predictive maintenance and plant-wide control through the use of a digital twin, which have the potential to reduce costs, improve resource recovery, increase water quality and increase customer engagement (IWA and Xylem Inc. 2019). However, the pace of this digitalization has thus far been quicker in the drinking water distribution sector than in wastewater (Water Online and SWAN 2019). This could be attributed to the fact that wastewater treatment processes depend heavily on water quality sensors, which are known to be difficult to work reliably, whereas those of drinking water distribution processes rely mostly on well-established water quantity measurements.

For these newly deployed sensors and novel data-driven models to help close the digitalization gap and have a significant impact on the way WRRFs are run, they must be maintainable by the workers and professionals of the wastewater field. Simply put, the information yielded by those new techniques must be relevant enough to be looked at by these professionals; be clear enough to be understood; be convincing enough to be believed and be reliable enough to be acted upon. Regardless of the abundance of the data or the sophistication of the models, failure to attain any of those features can be a fatal blow to the adoption of even the most promising new technology. This review thus aims to provide an ensemble perspective on how data are handled in WRRFs and to point out pitfalls and possible paths for improvement to water professionals wishing to make better use of their data.

FROM DATA TO INTELLIGENCE

To have a meaningful discussion of data-driven technologies, some seemingly similar terms must be given distinct working definitions. Using the definitions put forward by Makropoulos & Savić (2019), one may categorize levels of knowledge along the following hierarchy:

- Data: Quantitative or qualitative measurement or recording of a phenomenon or process.
- Information: Fact or observation derived from the analysis of data.
- Knowledge: Insight into mechanisms that relate pieces of information together within a certain context.
- Intelligence: The ability to use knowledge of distinct aspects of a problem to develop new ideas and perspectives.

These definitions make clear the fact that different levels of understanding build on top of each other. They also show that in the pursuit of 'smart' systems, one has to attain the highest level – intelligence – by refining and analysing the available data. Makropoulos & Savić (2019) represent this process as beginning with a question about the world and

continuing with all the steps required to answer it. The collection stage requires one to gather data from different stores. Once the data of interest is assembled, it should be analysed to extract information. This information is integrated into a model to capture underlying mechanisms and gain new knowledge from the studied system. This knowledge, when general enough, can be built into new tools, which then saves one from having to repeat analyses. Finally, with enough of those knowledge-based tools, intelligent decisions can be made with confidence. Of course, data interpretation is an iterative process, so higher levels of understanding gathered over time feed back into the interpretation of the same data at later stages.

From the authors' perspective, this path from data to intelligence is common to any data-driven activity. A description of each of these steps is therefore attempted here in the hope of bringing clarity to the often arduous process of distillation required by smart water systems.

Laying down the data pipeline

Using the framework outlined by Makropoulos & Savić (2019), it becomes clear that data-driven systems must be created from the ground up: without data, there is no information; without information, we know nothing; and a system that embeds no knowledge can not be intelligent. Intelligence generation can, therefore, be imagined as a pipeline with data at one end and intelligence at the other, as is pictured in Figure 1. This pipeline begins and ends at a WRRF, indicating that the intelligence extracted from the pipeline can be leveraged at later iterations of every step of the pipeline. It must also be acknowledged that different data-driven activities require different levels of complexity of analysis. Corominas *et al.* (2018) classified data-driven analysis based on their complexity in the following way:

- Basic information extraction includes simple schemes such as univariate control charts and mass, energy and stoichiometric balances.
- Advanced information extraction includes multivariate data treatment such as dimensionality reduction, feature detection and supervised machine learning.
- Human-interpretable knowledge extraction includes tools such as generalized rules, fuzzy logic, environmental decision support systems or ontologies.

This break-down emphasizes the points that lower levels of analysis may be useful on their own, and that high levels of analysis depend on the lower ones to function, which fits nicely into the pipeline framework.

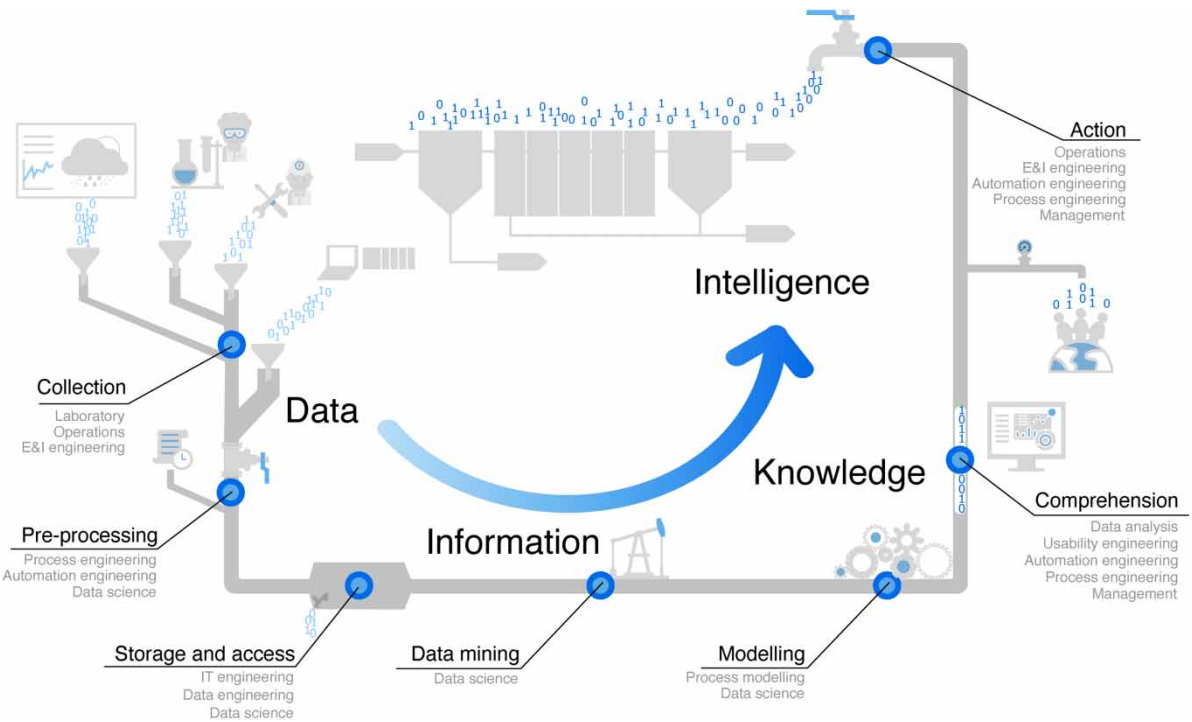


Figure 1 | From data to intelligence. For each step in the pipeline, the most essential professions are listed.

Some steps shown in the pipeline of Figure 1, data storage and data extraction, are not explicitly mentioned by Makropoulos & Savić (2019). They are, however, critical to creating usable data sets, as data collection and use may be separated by several months or years, and data collection may occur in a multitude of locations. Thus, the integration of the whole set requires a significant amount of effort and is worth discussing.

Figure 1 also lists the different actors involved in the manipulations and transformation of data along the pipeline. The variety of professions present underlines the fact that tending to the data produced by WRRFs is an inherently collaborative process and that several skill sets are required to achieve good results.

Moving away from abstractions such as 'information', 'knowledge' and 'intelligence' for a moment, one may envision the data pipeline leading to concrete, actionable insight in the following way. Let us first imagine a WRRF equipped with online sensors characterizing the quantity (flow meters) and quality (TSS, COD, COD_s, NH_4-N and NO_3-N) of the influent of a nitrifying/denitrifying biological reactor, its internal DO and TSS concentrations, as well as its secondary clarifier's recirculation line and wastage line's TSS concentrations. After having validated, cleaned and gap-filled these data, one could build a physical model of the biological treatment using one of the ASM models and a settler model. This model could be fed data series similar to what the reactor typically receives with the help of an influent generator model. Combining these two models as sources of information, one could determine whether the bioreactor would still be able to denitrify an adequate amount of nitrate if the internal recirculation rate was reduced. Equipped with this new knowledge, the operator could choose to decrease the sludge return flow to optimize pumping energy consumption.

A more involved example could include a live feed of processed sensor data into the physical model. The pre-processing needed to clean the live data is made easier by metadata, which can be either captured by automatic systems or created by the plant operators while they observe the process and inspect the sensors. This live data, when fed to an Extended Kalman filter, could help the model adjust its parameters to better reflect the current state of the reactor. When combined with meteorological observations and predictions, the model and influent generator could predict how the effluent concentration of ammonia is likely to change in the near future. Equipped with this knowledge, an operator (or an automatic controller) could adjust the dissolved oxygen setpoints with confidence that regulatory limits will not be exceeded. As one can see from these examples, the value of the data and the computational tools put in place in a digitized WRRF take all their meaning when they finally lead to a concrete, intelligent action.

Good labels make good neighbours

Before jumping into the pipeline itself, however, it is important to acknowledge that data must always be interpreted with knowledge of its history and context. That is why, ideally, metadata is also produced and collected at every step of the data pipeline. Rieger & Vanrolleghem (2008) define metadata as data about data that enables the extraction of useful information out of a signal. In other words, metadata is what describes the context in which data is created, thus making the data more easily interpreted. For industrial processes, metadata typically relates to the instrumentation producing the data (e.g. measurement unit, sampling location, equipment model etc.) and its associated quality (e.g. normal value range, raw or filtered value, measurement accuracy etc.). The metadata may be very structured (e.g. timestamps), or completely unstructured (e.g. text). Unstructured metadata is especially useful when annotating sensor signals with rich information to indicate the occurrence of an event.

When they presented the FAIR guidelines, Wilkinson *et al.* (2016), argued for general principles of data and metadata stewardship that may apply to any research area. For them, the four foundational principles that should guide data producers and publishers are:

- **Findability**: The data should be indexed, uniquely identified and contain rich metadata.
- **Accessibility**: One should be able to browse metadata using standardized protocols.
- **Interoperability**: Data should be in a widely useable format.
- **Reusability**: Metadata should reflect the needs of the domain of inquiry in which the data has been produced. The data should also be released under a clear license.

Preparing data sets such that all best practices are followed is a difficult task – especially when these best practices are still ill-defined. The FAIR guidelines thus provide a very welcome framework to incrementally improve data set quality. That is, according to Wilkinson *et al.* (2016), any of the principles can be implemented independently from the other, thus causing less friction in the beginning stages of implementation.

FLOWING DOWN THE DATA PIPELINE

Based on Figure 1, one sees that data goes through different stages throughout its lifecycle (i.e. collection, pre-processing, storage and access, mining, modelling, comprehension and action) and that the flow from one step to the other can be thought of as following the flow through a pipeline. One implication of this is that more complex data-dependent WRRF processes are vulnerable to failures anywhere along that flow. Another is that errors in early steps will influence every step downstream, potentially leading to the proverbial 'garbage in; garbage out' conundrum. This section therefore also aims at describing some of the potential failure points that are encountered in WRRF data treatment, and possible paths for improvement.

Drawing from the data well – data collection

The data collected in WRRFs is integral to their instrumentation, control and automation (ICA) systems. According to (Olsson 2012) three main objectives originally motivated the introduction of ICA to WRRFs almost 50 years ago:

1. Keep the plant running safely, i.e. get the water from influent to effluent in a controlled and reliable way.
2. Maintain good effluent quality to fulfil the plant's mission of environmental protection.
3. Optimize the plant processes to fulfil its mission while consuming as few resources as possible.

These motivations hold to this day, and data collected within the plant are indispensable in trying to accomplish these goals. Indeed, since the introduction of ICA, practically exponential growth in the amount of data being automatically digitised and stored has been observed (Olsson 2012). Common examples of such data are univariate time series coming from on-line process sensors and at-line analysers, discrete signals denoting actuator states as well as process operational settings (Vanrolleghem & Lee 2003). Because these data are sourced from dynamic systems, time series are characterised by some interdependency between sequential observations (Box & Jenkins 1970). This is in contrast to data being generated as a result of ad hoc events that do not occur deterministically in time.

Though sensors and analysers provide discrete measurements, their sampling period is usually small enough (seconds or minutes) as compared to the characteristic time constants of the treatment processes (hours or days). As such, they are invaluable tools on the path towards smart wastewater utilities (Ingildsen & Olsson 2016).

Time series data in WRRFs are either collected manually or automatically. Examples of manual data collection are offline laboratory analysis of grab and composite samples (i.e. analytically determined data), as well as observations made during visual maintenance inspections of the process itself including its field and panel-mounted measurement indicators. Note that most regulatory agencies still consider laboratory analysis to be the gold standard for water quality measurements (Yuan *et al.* 2019). However, some might say that laboratory analysis also comes with multiple drawbacks. For example, data collection is labour intensive, the measurements are delayed and infrequent (i.e., non-equidistant in time) because of extensive sample analysis and manipulation, and the measurements are prone to gross errors stemming from human distraction and fatigue.

The most essential part of any measurement instrument is, of course, the sensing element involved in capturing a physical phenomenon from the analogue world. Wastewater, however, is a harsh environment where organic and inorganic pollutants commonly cause fouling and degrading of the sensing elements (Dürrenmatt & Gujer 2012). This is eloquently demonstrated by Vanrolleghem (2014) in Figure 2.

If not maintained correctly, or compensated for during subsequent data processing, errors in data collection will directly affect the decisions taken based on the information contained in the data, and thus decrease user confidence in the measurement system (Regmi *et al.* 2019). As such, it is of utmost importance that a well-defined strategy is available and put into practice to maintain sensors, as well as the data they generate, throughout their entire life cycle. This also means that already during the conceptual phase of sensor implementation, a trade-off should be made between the value of the information being generated and the full cost of ownership (Zegers *et al.* 2019).

Another important aspect, not to be forgotten when working with automatic data collection, is the intrinsic dynamics of the measurement system itself (Rieger *et al.* 2003). These must be faster than the dynamics of the process being monitored to ensure that measurements are useful in operation and decision-making. Moreover, other low-level components involved in data acquisition, such as signal transmitters, carriers, samplers, converters and networks, will also have an impact on the data collection process and thus the final data quality. Examples include noise due to electromagnetic interference, ground loop errors, power surges, jitter, signal aliasing, quantization errors and analogue filtering (Whitt 2012). Automated data collection also requires different types of process instrumentation,

Figure 2 | Sensors in their ideal condition versus their typical operating condition. Top-left: conductivity; bottom-left: NH_4-N, K and pH; right: pH (Vanrolleghem 2014).

equipment and software applications having to communicate with each other. Whereas in the past this was typically established using custom drivers and proprietary communication protocols, the current trend is towards interoperability based on standardised and open digital communication protocols (Korodi *et al.* 2018).

By now it should be clear that the efforts of process operators, lab technicians, electrical & instrumentation engineers and automation engineers are indispensable in the data-driven operation of smart WRRFs, as these workers operate at the front end of the data pipeline.

Besides data generated by on-site measurement instrumentation, other less conventional data sources can also be integrated. A common example is when rainfall data, coming from publicly accessible weather stations, together with weather forecasts, are integrated to anticipate future operation of wastewater treatment facilities (Hernández-del-Olmo *et al.* 2019; Vezzaro *et al.* 2020). In the context of energy consumption, energy costs and load management, Aymerich *et al.* (2015) showed that energy prices, as determined by the local tariffing structure, can also affect the choice of a WRRF's operational strategy. As such, it might be of interest to also collect data from energy markets. Steel production, for example, envisions process optimisation based on real-time electricity markets where decisions can be made as frequently as every 5 minutes (Shyamal & Swartz 2019).

Cutting through the noise – data pre-processing

The process of evaluating and augmenting the quality of data, so it can be used purposefully, is often referred to as data pre-processing. This step of the pipeline is necessary since, as a result of measurement noise, errors and failures, it is impossible to collect completely accurate data. The case of on-line sensor data is especially interesting because of their high sampling frequency. Figure 3 shows WRRF online sensor data series exhibiting common problems that can be alleviated with data pre-processing. As can be seen in (a), the presence of outliers muddies the trend of the shown turbidity signal, while in (b), a sensor fault made a part of the data series unusable. In the former case, the application of an exponentially-weighted moving average outlier detection algorithm and trend smoother removed the undue noise from the data series (Alferes *et al.* 2015), while in the latter case, data filling with an average daily dry-weather TSS profile provided a substitute for the lost data (Patry 2020).

Wishing to inventory the different possible sensor states, Rosén *et al.* (2008) enumerate 7 distinct states, while Newhart *et al.* (2019) add an eighth:

1. Operational: Sensor is working properly, with normal measurement noise.
2. Excessive drift: When a sensor outputs a value progressively further from the true value.
3. Shift: When the output of the sensor is a constant amount away from its true value.
4. Fixed value: When the sensor is stuck and keeps repeating the same value.
5. Complete failure: Similar to a fixed value fault, but the sensors either give off the maximum or minimum, value, zero or no value at all.
6. Wrong gain: When signals away from the calibration point are under- or over-amplified by the sensor.

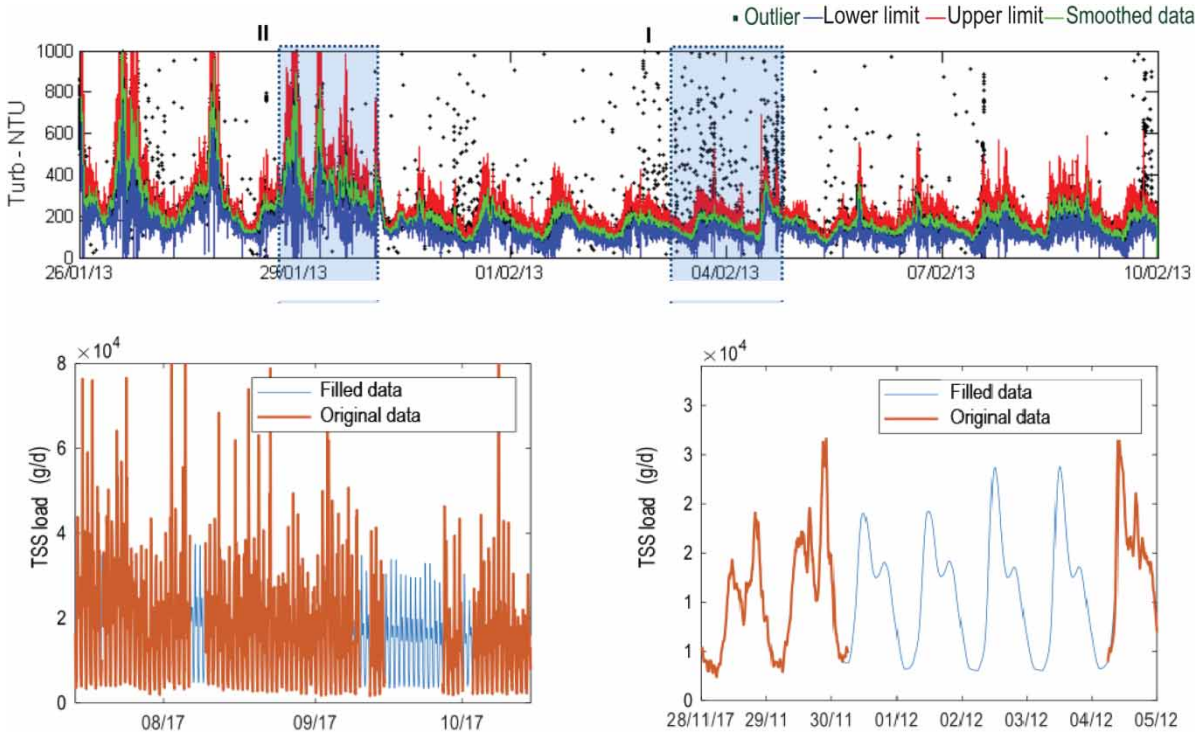


Figure 3 | (a) Raw turbidity sensor data (dots), along with the processed data (green) generated using an exponentially-weighted moving average (EWMA) filter (Alferes *et al.* 2015). (b) Raw total suspended solids (TSS) sensor signal containing gaps (orange), which are filled with average dry-weather daily profiles (blue) (Patry 2020). The full colour version of this figure is available in the online version of this paper, at http://dx.doi.org/10.2166/wst.2020.393.

7. Calibration: The sharp change in sensor output directly following a calibration.
8. Isolated fault: When a single point in a series shows an incorrect value.

Schraa *et al.* (2006) discuss several checks that may be performed on sensor signals to detect faults in the sensors or the process itself, such as whether user-defined bounds on the sensor measurement have been exceeded, or whether a measurement signal is outside the typical 4–20 mA range. However, by far the most common way to detect erroneous data used by analysts is by visual inspection of the time series (Alferes *et al.* 2015). However, given the sheer amount of data collected in any treatment plant, visual fault detection is simply impractical (not to mention, error-prone and requiring lots of expert knowledge). Thus, automatic detection of process and measurement faults is necessary. As it turns out, fault detection (determining whether a fault is present and at which time it occurred), fault isolation (determining which sensor caused the failure) and fault diagnosis (determining what kind of fault has occurred) (Isermann & Ballé 1997) are non-trivial tasks. As such, a lot of research has been carried out over the past decades specifically for faults present in WRRF.

Corominas *et al.* (2011) proposed a practical methodology to compare the accuracy of different univariate algorithms (e.g. exponentially weighted moving average (EWMA)) on artificial faults in time series generated by the Long-Term Control Benchmark Model (BSM_LT) proposed by Rosén *et al.* (2004). Meanwhile, Rosén & Olsson (1998), (Yoo *et al.* 2007) and Alferes *et al.* (2013) describe the use of multivariate techniques such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) to detect anomalies in time series. Russo *et al.* (2019) and Garneau & Vanrolleghem (2018) also demonstrate the use of autoencoder neural networks for the same purpose.

For their part, Ohmura *et al.* (2019) discuss two of the main assumptions made by most fault detection methods; that is, that different sensors used in multivariate fault detection fail at distinct times from each other, and that sensors work perfectly for some identifiable period. They demonstrate that these assumptions do not hold in every situation by simultaneously deploying 8 pH probes in wastewater and observing that all of them started drifting immediately. Consequently, the authors advocate for the development of new fault detection methods that are not based on these assumptions for use in such cases.

On the other hand, instead of focussing on fault detection as such, researchers such as Schneider *et al.* (2019) have investigated whether faulty data itself may be used with reasonable effectiveness despite their flaws. They concluded that unmaintained sensor signals were just as useful as maintained ones to detect specific process features (e.g. the 'knee' in an oxidation-reduction potential curve). Some types of sensors, however (e.g. dissolved oxygen) showed nonlinear disturbances that rendered their signal unusable when unmaintained.

Additionally, Rieger *et al.* (2010) recommend the use of mass balances inside plants to deduce missing information from a given sensor as well as to identify and correct for random and systematic gross errors in the data collection. Applying this procedure in process operation for data evaluation is referred to as data reconciliation. Important to the framework of data reconciliation are the concepts of measurement redundancy and variable observability. A methodology based on (bi)linear mass balances and specifically for WRRF data is proposed by Le *et al.* (2018) and Villez *et al.* 2020, setting up a measurement system to maximize the chance of detecting sensor faults and reconcile the faulty data.

Because of sensor faults and maintenance operations, it should be clear that no sensor boasts 100% uptime. Gap filling of data is a fact of life. Such gaps may be short or very long depending on the dynamics of the phenomenon being monitored and other factors. Different methods have been proposed to fill data gaps, the choice of which is influenced by the duration of the gap as well as the goal being pursued by the analyst. In their open-source wastewater data treatment toolkit, De Mulder *et al.* (2018) suggest five data imputation strategies specifically aimed at data generated by WRRF operation:

1. Interpolate.
2. Use a correlation with other available measurement signals.
3. Replace with a corresponding value in an average daily profile.
4. Repeat the values obtained on the preceding day.
5. Replace with the output of a model.

Though these gap filling methods may be used for reconstructing data series related to the plant itself, this is not their only use. Collecting influent water quality data, for example, is notoriously difficult and time consuming. This is unfortunate, as the state of the plant is directly related to the nature of the influent it is treating, both in terms of quantity and quality. Thus, water professionals may turn to influent generators to fill in the gaps in their influent data. These tools aim to generate complete time series from partial wastewater treatment plant influent data (Martin & Vanrolleghem 2014). As such, they replace expensive and time-consuming data collection campaigns; they allow modellers to infer concentrations of relevant wastewater components from other and correlated components, and they enable one to generate at will any number of instances of similar (and similarly plausible) time series. The generators can either be based on historical data from the influent of the plant (Devisscher *et al.* 2006) or characteristics of the upstream watershed and sewer network (Talebizadeh *et al.* 2016).

Say, in what folder was that already? – data storage and access

There exists a lag between the collection of data and its ultimate use. Hence, data storage and subsequent extraction are integral steps of data-driven systems, though they often go unmentioned in the WRRF literature.

The type of data being collected influences the choice of a storage system. In the case of offline laboratory data, manual record-keeping is still very often used, making data valorization a laborious task. However, laboratories nowadays have the option to use digital Laboratory Information Management Systems (LIMS) to organize their workflows and the results of their analyses in a central database (Skobelev *et al.* 2011). Similarly, data generated by sensors and other assets may be stored using a general-purpose database management system (DBMS) or an application-specific process data historian (Yee & Eren 2012). The difference between the two is that the latter is highly optimised in terms of writing, storing and extracting time series as a result of extensive data filtering, compression and caching. However, as data historians typically compress data to save storage space, their use may result in information loss. Indeed, compression reduces fidelity and increases the granularity of the data. Reconstruction of the original data without losing important statistical features is therefore not guaranteed (Thornhill *et al.* 2004).

In the pursuit of intelligent decision-making for WRRF operation, it is key that stakeholders have access to all relevant available data. Meaning that cross-subject and cross-facility data repositories such as spreadsheet logbooks, programmable logic controllers (PLC), supervisory control and data acquisition systems (SCADA), distributed control system (DCS), historians, lab information management system (LIMS), weather stations, computerized maintenance

management systems (CMMS), enterprise asset management systems (EAM), and so on, need to be integrated and centrally accessible. This is where the concept of a data warehouse comes in. Defined by Inmon (1998), a data warehouse is a logically centralized data repository where cleaned data originating from operational data stores are integrated and standardized to support business intelligence. Whereas the operational data stores are typically used to answer short-term questions in real-time, the objective of a data warehouse is to provide decision support for mid to long-term organisational strategies. For WRRFs, the obvious long-term goal stakeholders are trying to get a grip on is to increase operational efficiency, with existing resources, while meeting regulatory compliance. Although the concept of data warehouses has been around for several decades and applied in various contexts, their application in the water and wastewater utility industry remains limited or unreported. Indeed, isolated data stores, without open-data interfaces for enterprise-wide access, are still the norm for most water and wastewater utilities (Sirkiä *et al.* 2017). Recently, software providers are trying to counteract the existence of such data silos by upgrading process historians into full functional data warehouses (Matthews 2017).

Data warehouses are mostly based on the relational database model to store data in a structured format; that is, in a collection of different tables and records, each containing pre-defined fields with data and metadata. Various tables are then linked together by relations between fields, which allows for the cross-referencing of data from separate tables. The database schema prescribes the content of each field of each table and the links between tables. In the case of time series, relationships between the data itself are limited although metadata can provide structural relationships. An example related to the field of water and wastewater quality is the schema developed by Plana *et al.* 2019, which aims to store related values coming from sampling stations within a water network, a sewershed or a WRRF with the relevant metadata (see Figure 3). For their part, OGC Consortium (2011) have suggested a markup language to store relevant metadata along with water data according to a general structure, though without prescribing any specific database configuration for storing such data.

Note that metadata such as sampling location, sensor serial number, sensor manufacturer, measurement unit, etc., are commonly understood as being created and stored early in the process of data collection. However, any step within the data pipeline could potentially generate its own metadata. For example, the use of a fault detection algorithm can generate metadata recording the specific version of the algorithm used, the value of each of its parameters, the subset of data that was used for calibration, etc. Though metadata obtained during collection may provide important context to its associated data, examples of frameworks for automatic generation of metadata further on in the pipeline; that is, post-collection, are rarer for the wastewater operations field. One example, however, is found in De Mulder *et al.* (2018). Though the FAIR guidelines are not explicitly followed in this instance, the authors nonetheless provide a structured framework to also store metadata generated during data pre-processing. Such frameworks are crucial if one is to try to replicate outcomes from experiments or models using raw data that passed through data cleaning and gap-filling procedures. Given a long enough time series, virtually all wastewater-related time series are bound to undergo some degree of processing, hence the crucial necessity of expanding on these frameworks for the entire pipeline.

As the data sources related to the operation of a WRRF multiply, schema-based data storage becomes increasingly strained to its limits. For example, the relationships between the data being collected may become unclear or left undefined as the variety of data increases, or the speed of data creation may well exceed the speed at which the data can be parsed and processed by the database server. One then enters the realm of big data, for which specific large-scale data storage systems are used. Data may be called 'big' according to the degree to which it presents the properties summarized by the four V's (Farley *et al.* 2018):

- *Volume* of data being generated: counting every sensor, actuator, alarm and type of laboratory measurement taking place in WRRFs, typical plants have several thousands, if not tens of thousands of tags being logged. This, of course, generates a lot of digital data; for example, a single measurement with a sampling period of 10 seconds will produce 8,640 records a day, amounting up to 3,153,600 records a year.
- *Velocity* of data creation and collection: on-line sensors and instruments placed throughout a WRRF can collect data at frequencies up to tens of measurements a minute. Often the information content of this data is, however, very low since the dynamics of the process being measured are much slower than the sensor's sampling frequency (Olsson 2012).
- *Veracity* of the data being collected: the harsh environment WRRF sensors find themselves in mean that faults are omnipresent. Pre-processing is therefore critical for the data to be trustworthy.

- *Variety* of collected data: most of the sensor data collected in WRRFs are time-series data. However, a lot of information at WRRFs is nowadays being stored in other unstructured data formats such as photos, videos, spectral measurements, instrument data sheets, standard operating procedures, etc.

Data lakes are a big data alternative to data warehouses that allow for the storage of both structured and unstructured data, without concern for indexing or making sure that the metadata fits a specific schema (Nargesian *et al.* 2018). This proves much more flexible than data warehousing and is, therefore, easier to implement when the variety of data is high. However, data lakes provide no guarantee regarding the integrity of the data they hold, which might result in data swamps rather than lakes. Additionally, since there is no enforcement of metadata upon writing, much of the context in which the data was collected may be lost. Instead, the user reading the data will have the task of reconstructing context from available clues (Liu & Gawlick 2015). The risk of data being overlooked, misused or corrupted is thus increased compared to data warehouses. Though the use of data lakes is increasing in the business intelligence world, no published example of an implementation for a WRRF has been identified by the authors of this review.

Data stored in databases can be accessed either through structured queries, as is the case for relational databases and data warehouses, or using metadata or pattern-recognition with data lakes and other unstructured data sources. To select data to be extracted and used, there needs to be a way for the data analysts to, as it were, know ahead of time what to search for. This is especially difficult to achieve when no schema exists to guide their search. There is thus a need for tools that enable one to preview, select and subset the available data. Query languages together with scripting languages are powerful interfaces, though they require programming literacy. Several closed and open-source software solutions have been developed (Demšar *et al.* 2013; Drucker & Fernandez 2015) over the years to enhance data exploration for a larger user base by creating graphical user interfaces that connect to structured and unstructured data sources alike.

Whereas data storage has traditionally been done on-site, there has been a shift in recent years to the adoption of off-site data storage using cloud computing. With virtually unlimited storage capacities and computing power, cloud solutions remove the burden of server acquisition, maintenance, and eventual upgrade away from their customers. However, such a system requires fast and reliable internet access, which is not available to WRFFs in remote locations. Specialized issues such as cyber security deserve special attention when working in the cloud, as the connection of WRRF data systems to the internet introduces the risk of unauthorized access and tampering (Blumensaat *et al.* 2019).

An aside on data mining

Though it is a separate process, data mining is linked with both data extraction and data analysis. Its increasing relevance, triggered by the ever-growing data sets of WRRFs, means that it warrants a brief discussion. Starting in the 1980s, access to more processing power has allowed for the adoption by businesses of data mining techniques, which are used to discover underlying relationships in data that were originally collected for another purpose. Lovell (1983) and Denton (1985) warned that this repurposing of data might lead researchers to find specious relationships in the mountains of data they analysed instead of detecting actual, meaningful relationships. They, therefore, underlined the necessity for researchers to remember that correlation does not imply causation when attempting data mining. Despite these reserves, data mining has flourished, and the water field has caught on as well.

Data mining should not be conflated with machine learning (ML) however – the former is the process of finding patterns in large amounts of data, while the latter is the process of using algorithms to find those patterns. Both activities are therefore unique but complementary, as both are required to automate the discovery of relevant trends and relationships in WRRF data.

Hadjimichael *et al.* (2016) have explored the potential of data mining for enhancing decision support systems (DSS) for urban water systems. They have found that the literature on data mining for DSS is very sparse compared to literature using data mining to create process models or for process optimization and that several obstacles stand in the way of the adoption of data mining in DSS systems. These challenges are just as applicable to the modelling field. They are that:

- Water professionals tend to lack the expertise in computer engineering required to develop adequate ML-derived models;
- Computer engineers do not possess the field-specific expertise needed to develop adequate ML-derived models for water systems on their own either;

- The tools and interfaces delivered to water professionals to explore their data are too difficult to use and lack adequate long-term support for them to be used in practice.

In light of these issues, it is clear that the wastewater field, though eager to leverage automated data analysis, generally lacks the required expertise to become proactive participants in its adoption and development for wastewater-related purposes.

This mix of eagerness and lack of expertise creates a perfect storm for the creation of data graveyards – large collections of data that never get used. These occur when data is collected without a clear motivation or purpose (Corominas *et al.* 2018). It is thus essential to treat the data pipeline as the critical part of the WRRF ecosystem that it is. Consequently, it must be just as carefully designed and maintained as the rest of the plant, with the collaboration of data governance experts.

A model is worth a thousand datasets – modelling

Modelling encodes, in a circumscribed form, all the knowledge and information relevant to a task. Knowledge may come from prior experience, whereas information must propagate from collected and processed data of the studied system. Though models are often formalized in mathematical language, they are not always. Sometimes, a schematic drawing or a written explanation of a phenomenon is all the modelling you need. However, such models can be considered more conceptual in nature. In the case of water resource recovery, however, mathematical modelling is extremely useful as it allows water professionals to perform virtual simulations of a WRRF system and its subprocesses. Simulations facilitate several tasks, such as the design of new plants, the optimization of existing ones, the prediction of the future behaviour of systems given an initial state, etc. To understand the current modelling trends, it is useful to define the types of models available to modellers developing operational support tools and tuning control systems. Models are typically categorized as follows:

1. White-box models are equivalent to mechanistic and phenomenological models in that their internal structure is legible and interpretable. Hence, they are deduced from first principles and are typically expressed as sets of differential algebraic equations to describe the steady-state or dynamic behaviour of a system (Gernaey *et al.* 2004). These models depend on data through their forcing input variables, but also for the calibration of their parameters, as well as for setting the initial and boundary conditions as required for the numeric calculations. White-box models can be further broken down into:
   a. Mechanistic models: these are based on physical laws; for example, the law of conservation of mass, or the laws of thermodynamics. They rigorously describe the behaviour of the system in an idealized form. A simple example of such a model is the mass transfer of oxygen (Garcia-Ochoa & Gomez 2009). Biofilm models, with their careful characterization of substrate transport through the biofilm (Pizarro *et al.* 2001) or the hydrodynamic flow of the bulk liquid in biofilm channels (Eberl *et al.* 2000) are other excellent examples of such models. Water resource recovery processes such as ammonia stripping and struvite extraction also use mechanistic models to characterize the chemical reactions underpinning these processes (Vaneeckhaute *et al.* 2018).
   b. Phenomenological models: these may be based on physical laws; however, they are not strictly beholden to them. Instead, these models include empirical relationships that describe the patterns of the observed phenomenon without having that description depend on the fundamental processes that generate the behaviour (Martin & Vanrolleghem 2014). For wastewater treatment, the most commonly used models belonging to this category surely are the Activated Sludge Model (ASM) family (Henze *et al.* 2000), which combines mechanistic mass balances of biochemical processes with heuristic relationships. Another widely used phenomenological model in the field of WRRF modelling is the settling model presented by Takács *et al.* (1991), which is based on mass balance equations, but relies on phenomenological descriptions of settling in activated sludge to model the storage and outflows of clarifiers.
2. Black-box models are models that map sets of inputs onto a certain output without any concern for embedding structured knowledge of the real processes that created these outputs. Because they contain no prior knowledge of the process they model, these models are completely dependent on the data being used to build them for their accuracy and applicability. Black-box models have proven to be very effective at finding unsuspected patterns in data generated by various application fields. The techniques employed for black-box modelling are varied, encompassing multivariate statistical models, time-series models (Box & Jenkins 1970), support vector machines (SVM) (Cortes & Vapnik 1995) and artificial neural networks (Werbos 1982). The basic choice of

each usually depends on whether the task at hand falls into the regression, classification or clustering categories. Haimi *et al.* (2013) provide an extensive overview of the various methods being applied for data-derived soft-sensors, specifically for biological wastewater treatment plants.

Because of recent advances in big data analytics and artificial intelligence, data-driven models will become progressively more widespread in the urban water field, hence strengthening the case for well-managed and readily available data (Garrido-Baserba *et al.* 2020). However, data-driven models have also been criticized for their lack of transparency, which could result in mistrust. A new line of explainability research has emerged with the aim of not only comprehending what a model did or might have done but also being able to question and audit the model (Gilpin *et al.* 2018). In this context, it is important to note that not all data-driven models are opaque to the user. For example, Principal Component Analysis (PCA) allows one to readily infer which variable of the input space contributed to the observed response in the output. The problem of explainability is, however, mostly encountered in the context of more sophisticated artificial intelligence (AI) methods based on neural networks.

3. Grey-box models, sometimes referred to as hybrid models, are a mixture of fundamental white-box and empirical black-box components. This category of models deserves special attention, as they combine the explanatory power inherent in first-principles models with the ability to detect subtle patterns in data. This hybrid formula is thus especially useful when white-box models contain parameters or state variables that are not readily evaluated experimentally (Psichogios & Ungar 1992). For example, Shiva Kumar & Venkateswarlu (2012) modelled a fixed bed biofilm reactor using a mechanistic model; however, they used an artificial neural network (ANN) to determine the kinetics of the growth rate, as it was unclear which mechanistic formulation was best suited to the behaviour of their biofilm. Also, Meirlaen (2002) trained a neural network to emulate the biokinetics of ASM2 while maintaining a mechanistic mass balance. Similarly, Lee *et al.* (2005) coupled an ASM1-derived model with different black-box models (namely, a Neural Network PLS scheme) to model the water quality of an industrial plant. They found that this approach not only yielded good performance, it also provided the authors with a readily interpretable signal to detect instances of unstable plant performance. It can thus be seen that the use of data-driven modelling schemes need not come at the cost of interpretability. Because of the growing interest of combining both domain-specific and data-derived knowledge, especially for complex nonlinear systems such as those encountered in WRRFs, new modelling frameworks are constantly being developed. The most notable hybrid modelling methods are currently coming from the emerging field of scientific machine learning (Baker *et al.* 2019), which lays the theoretical framework needed to embed mechanistic differential equations into deep learning models.

Putting on a friendly face – comprehension

As powerful as mathematics are at describing the physical world (possibly, even 'unreasonably effective' as Wigner (1960) famously remarked), the fact remains that very few humans are fluent in the language of maths. Our brains being what they are, most of us process information entirely differently than the machines running our models. There must, therefore, exist in the data pipeline interfaces that translate the information embodied in data, and the outputs of mathematical models, into a form that is more adapted to human cognition such as visual or linguistic models. Graphs are the most ubiquitous of these interfaces; however, all graphs are not created equal. Shah & Hoeffner (2002) explored the impact of graph formatting on the interpretation of data and found, among other things, that several parameters of graph design influence the ability of the user to make sense of the displayed information. More importantly, the authors also emphasize that these effects have different magnitudes based on the level of graph literacy of the user. This means that graphical tools used during model interpretation may need to be vastly different than those necessary to communicate those model results to wider audiences, namely WRRF operators, management and, why not, the general public. The translation of data from mathematical models to human actors must, therefore, be approached with care and with the user's needs in mind.

Knowing this, it must be noted that numerous efforts have been made towards concise but clear visual tools to present data and models (Figure 4). Amerlinck (2015) contributed a tool to quickly identify rate-limiting terms for processes of an ASM model using a colour scale. Similarly, Thürlimann *et al.* (2015) have developed a visual tool, based on colour bars and calendar-based views, for intuitive tracking of plant key performance indicators (KPIs) within a

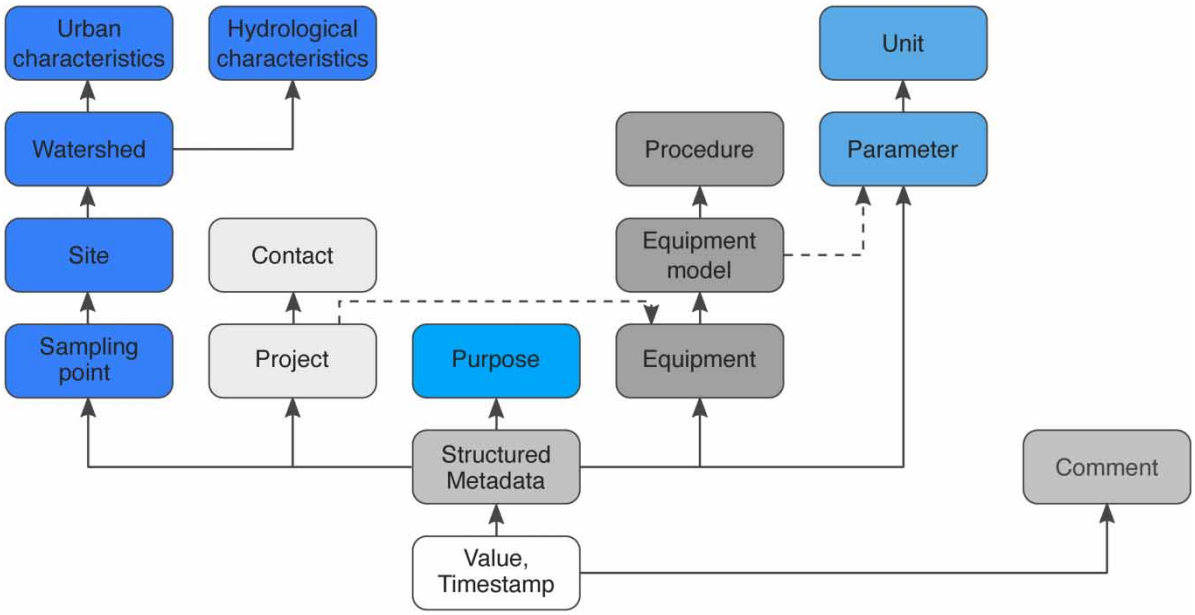
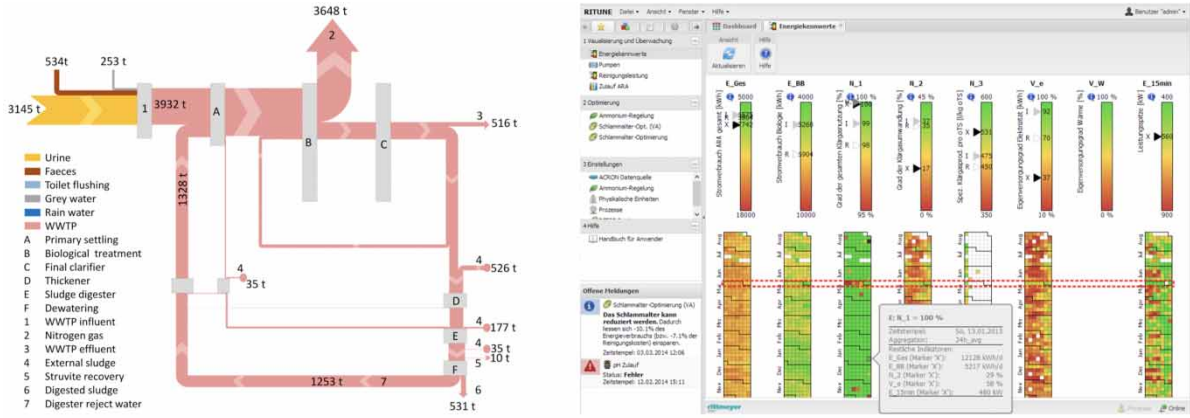


Figure 4 | Water-related metadata considered by Plana *et al.* (2019) (adapted from the original).

process optimisation software for WRRFs. Animation of simulation results may also prove useful. Stepping out of the WRRF context and into the catchment scale, Benedetti *et al.* (2006) proposed using diagrams with arrows of different widths (i.e. Sankey diagrams) to rapidly identify the sources and pathways of pollutants in a watershed. The same idea was also applied to the WRRF system by van der Hoek *et al.* (2018).

Besides presenting model outcomes, it is equally important to make the underlying modelling tools accessible to potential users. Dynamic process simulators for WRRFs, whether built in-house or commercially provided, are eminently fit to this task. By converting the complex mathematical equations of unit operations into drag-and-drop objects with user-friendly interfaces, the threshold to initiate a process modelling exercise is substantially decreased. As a result, simulators are nowadays used to design entire WRRFs, to support process operation, to help in the development of software sensors, and to train process operators. To improve model prediction accuracy, and consequently the perceived trustworthiness of the simulator, measurement data is used for calibration and validation. Although commercial WRRF simulators typically provide intuitive and automated model calibration tools, the implementation of data pre-processing and analysis tools remains somewhat limited. Including more powerful tools for data analysis, fault detection, gap filling, data-driven modelling and the like could be a big leap forward in the path towards even more successful WRRF simulators. With the advent of digital twins that include process simulators with a real-time data feed, the demand for efficient data treatment and hybrid process models will only rise in the predictable future.

Another important instrument used to disseminate information to plant operational staff is the human-machine interface (HMI). In the past, the design of such operator screens was entirely determined by the creativity of the automation engineers programming the system. The result was typically little more than copies of the P&IDs with confusing and distracting graphics. Such poor HMI practices prevent staff from operating plants near their most efficient point, and more importantly, they have been shown to contribute to major accidents in the process industries (HSE 1997). Nowadays, guidelines are available to avoid poor graphical principles during HMI design. Whereas data was previously scattered on the screen with a graphic of the process, the current trend is to add a high degree of context to the data. This way, continuous comparison is made possible, which simplifies the interpretation of complex process operation. Having knowledge directly embedded in the screens of such high-performance HMIs can drastically improve the situational awareness of process operators and thus decrease response time (Hollifield *et al.* 2008). In this context, Rieger & Olsson (2012) recognized the importance of clear visual communication of process control actions. They argue for embedding control actions and controller



AVERAGE	AN01	AN02	AN03	AN04	DT01	DT02	BT01	BT02	BT03	BT04	BT05	BT06	CA
Monod S_NH	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98
Monod S_A	0.91	0.91	0.92	0.92	0.63	0.52	0.08	0.06	0.04	0.04	0.03	0.03	0.03
Monod S_F	0.84	0.72	0.57	0.43	0.12	0.12	0.14	0.12	0.11	0.16	0.17	0.20	0.23
Monod S_Alk	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Monod S_PO	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.94	0.81	0.41	0.30	0.22	0.21
Monod S_NO	0.01	0.00	0.00	0.00	0.09	0.18	0.84	0.84	0.84	0.84	0.85	0.85	0.85
Inhibition S_O	1.00	1.00	1.00	1.00	0.99	0.96	0.12	0.22	0.49	0.05	0.05	0.05	0.03

Figure 5 | Top left: Sankey diagram of nitrogen flow through a WRRF) (van der Hoek *et al.* 2018); Top right: Decision support tool used to visualize WRRF key performance indicators based on energy and process data (Thürlimann *et al.* 2015); Bottom: Systems analysis tool that uses colour coding to identify growth limitations in different sections of an activated sludge tank (Amerlinck 2015).

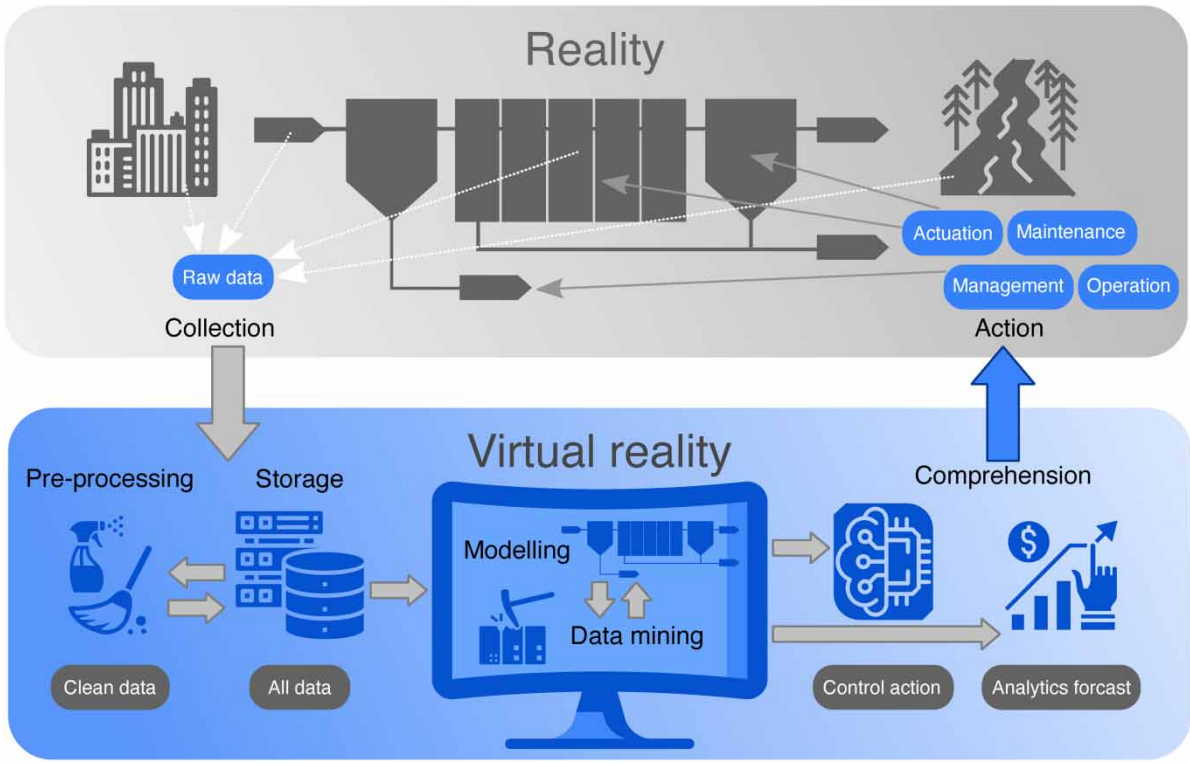


Figure 6 | Components of a digital twin.

settings directly in SCADA HMIs, along with the use of hierarchical parameter displays to guide users to the information most likely to help them accomplish their tasks. It can thus be concluded that, though seemingly simple, visual and intuitive components of the data pipeline are indispensable to ingest the massive amounts of information coming from data and models.

Is there a faucet on this thing? – action

For data to be truly useful, it must be put to work. One way of doing so is through the deployment of automatic process controllers and process models. The role of data in each of those is explored here.

Automation is the process of creating systems that can perform tasks without human intervention. The result is that process operators are alleviated from demanding repetitive tasks required to keep the plant running. In WRRFs, valve opening and shuttering, pump timings and sensor cleaning are among the candidates for automation. The first attempts at automated plant control were undertaken in the 1970s (Olsson 2012). Already at that time, it was clear to some researchers that massive gains in efficiency were possible by using automated controllers that encapsulated some knowledge about the plant. A lot of effort was applied to control the aeration process in particular, as aeration accounts for 45–75% of all wastewater treatment energy expenditures (Rosso *et al.* 2008). Early control strategies consisted of simply maintaining dissolved oxygen (DO) concentration in bioreactors at a fixed low value with the help of conventional feedback controllers. More complex schemes appeared over time, such as DO cascade control (Ingildsen *et al.* 2002), ammonium-based aeration control (Rieger *et al.* 2013) and multiple input multiple output control strategies, which modify the system setpoints to achieve the required oxidation capacity for shifting conditions within the plant (Åmand *et al.* 2013). This increase in complexity marks the shift from regulatory control – automatically manipulating an actuator to reach a setpoint value – to supervisory control – manipulating the set point itself to reach a higher-order objective (Ayesa *et al.* 2006).

As the complexity of control strategies grows, the likelihood that a component critical to that strategy will fail increases. It is therefore important to implement these strategies within a fault-tolerant framework, in which the process can either be automatically transferred to a fallback strategy or a safe parking-point until the fault is identified and corrected (Mhaskar *et al.* 2013). The use of fault-tolerant control is reviewed thoroughly by Blanke *et al.* (1997) and Zhang & Jiang (2008).

In addition to automatic control systems, one must not forget the people who oversee and support the automated systems. If successfully executed, the passage of data through the entire pipeline empowers those people to have a clear view of the state of, and power to act on, the plant through the automated system. The power to act hinges on the final transformation of knowledge into intelligence and intelligent action. Although this transformation is eased by visual interfaces, it ultimately occurs in the minds of water professionals. This means that the tools springing from the data pipeline must help workers synthesize their knowledge of the plant in ways that enable them to act intelligently on it. These tools come in many different shapes and forms. Managers can benefit from the data by having it embedded in a decision support system (Hadjimichael *et al.* 2016). Operators can benefit from dashboards indicating the state of sensors or key performance indicators (KPI) related to the processes they oversee (Thürlimann *et al.* 2015). Even the public can get involved with wastewater data if given the appropriate tools. For example, see the citizen science project of Damman *et al.* (2019), involving a community to sample and analyse the quality of their rivers downstream of their local wastewater treatment plant.

Of course, control for the sake of control achieves nothing: the control goal and whether that goal is reflective of the plant stakeholder's interests is what ultimately makes these systems effective. Weijers (2000) proposes a systematic methodology to derive appropriate control goals and the associated constraints for a given plant goal, as opposed to working from vague intentions such as 'minimizing costs while maximizing water quality'.

For their part, Rieger & Olsson (2012) remark that the stakeholders of wastewater treatment plants may vastly disagree on what constitutes a good control objective. In their view, this is because each of the stakeholders acts under contradictory incentives. Thus, the human aspect of control, including goal setting and the relationships between stakeholders and co-workers, is found to have a tremendous impact on the success of control strategies. Rieger & Olsson (2012) add:

> *'There has to be a qualified team of people who feel a deep sense of ownership of the system and the WWTP, and who are committed to its continuous improvement. It is important that all employees increase their competence through continual education.'*

It is thus crucial that the complex human interactions that form the backbone of wastewater treatment systems stay in view when designing technical solutions seeking to enhance these systems. Nevertheless, there is no denying that technology is a strong ally to the wastewater field in their quest to turn their data into intelligent actions.

DIGITAL TWINS

In recent years, a lot of excitement has been sparked by the perspective of creating so-called digital twins for WRRFs (IWA and Xylem Inc. 2019). In the context of water management, Kolditz *et al.* (2019) define digital twins as virtual systems that 'contain all important characteristics and features of the real system, depending on the specific purpose for an application'. This vague definition does not explain in detail what a digital twin is. It does, however, suggest the following key features:

1. Digital twins are virtual systems that aim to embody and simulate the physical components of a real system. Therefore, not physically existing as such, but made by software to appear to do so.
2. A digital twin sources its data from measurements performed on the physical system itself or from its environment.
3. The digital twin constantly mirrors the current state of the physical plant. Consequently, some of the data the digital twin requires needs to be provided in real-time.
4. The digital twin, having access to the current state of the plant, can make predictions on the future state of the physical system.
5. The predictions that are produced can be fed back into the real physical system in the form of intelligent actions. Depending on the intrinsic characteristics of the system, these actions can be made automatically or manually.

It is easy to see that the concept of a digital twin is reliant on every part of the data pipeline. Indeed, it cannot exist without extensive, fault-free and continuous data from the plant and its environment; it embodies knowledge of the plant through a constantly updated model; it provides insight into the plant via interfaces, visualization and analytics, and it allows for action on the plant via automatic control and the insight it generates for water professionals.

Moreover, the digital twin concept for process industries is often thought of as a close cousin of a model-based control system used to automatically optimise plant performance. To the authors, this point of view limits itself to the continuous operation of the process, meaning that it does not take into account the potential value generated by a digital twin in other aspects of a plant, namely plant maintenance and asset management. Consensus on which components must be present in a digital system for it to be called a digital twin is therefore much needed so as not to misuse this powerful concept. Note that this effort has already been initiated in other industries and other branches of the water industry, including the field of water and wastewater networks (Water Online and SWAN 2019).

LIGHT AT THE END OF THE PIPE?

This review attempted to inventory the steps required to turn raw data into intelligent action for the operation of a WRRF. As these steps were discussed, possible problems were pointed out, as well as possible ways to cope with them. The main sticking points discussed in this review are gathered in Table 1.

For some of these issues (e.g. fault detection or data gaps), there seems to exist convincing technological and scientific tools to alleviate the problems. For others, however, the way forward is simply the continued dedication of water professionals (e.g. sensor maintenance), or their willingness to collaborate with or be trained by experts from other fields (e.g. development of a data warehousing strategy or adapted user interfaces).

The nature of the issues encountered along the pipeline and their potential fixes may be diverse, but they certainly will all require a significant amount of effort and dedication to effectively tackle. This has implications for research, of course, but also for the practical applicability of smart wastewater applications such as the digital twin to WRRFs around the world. Indeed, since the success of such a system requires the creation and the maintenance of an integrated data pipeline, as well as considerable modelling efforts to implement relevant control strategies within the twin, only the most sophisticated WRRFs may be able to attempt to create such a system for many years to come. Nonetheless, progress in the development of robust data pipelines may very well have benefits for smaller WRRFs as well. Indeed, one can easily imagine remote WRRFs benefitting from better fault detection for remote monitoring and from gap-filling to reconstruct faulty time series between maintenance operations, for example. Thus, as WRRFs brace themselves for a smart future, hopefully water professionals will consider patching leaks in the data pipeline to be much more than a chore, but rather see it as participating in the development of a sophisticated system that is just as complex and

Table 1 | Summary of data-related issues encountered in WRRFs

Step of the pipeline	Issue encountered	Paths for improvement
Collection	Harsh WRRF environment negatively affects online sensors	• Careful, regular, sensor maintenance • Fault detection procedures
	Lab data is often recorded manually. Maintenance of sensors can be resource intensive.	• Use of LIMS software • Define a strategy for the entire lifecycle of the sensor before installing.
	Enthusiasm for collection without purpose leads to data graveyards.	• Treat data collection as its own process; let it be engineered by experts, carefully planned and rigorously executed. • Develop a strategy ahead of time for data use.
Pre-processing	Manual fault detection is too time consuming.	• Univariate automatic fault detection algorithms • Multivariate automatic fault detection
	Automatically captured and treated data contain gaps.	• Time-series generators • Model-based gap filling • Interpolation
	Difficult to know which data has been changed by pre-processing algorithms, and what algorithm was used.	• Keep thorough metadata accounting for pre-processing • Data versioning • Version control of pre-processing algorithms
	Fault detection algorithms often assume we can pinpoint a period of 'good data', but sensors may begin to drift immediately after commissioning.	• Sensor redundancy • Mass balances for data reconciliation • Use signal features that are identifiable even if the sensor is unmaintained/faulty
Storage and access	Data historians use destructive compression	• Consider signal properties when choosing a signal resolution
	Simple databases are ill-suited to handle WRRF 'big data'.	• Data warehouses • Data lakes
	Data is often spread in multiple storage sites and is thus difficult to access.	• Improved collaboration between data experts and water professionals.
	IT infrastructure may be difficult to maintain on-premise because of lack of technical know-how.	• Increased use of cloud computing • Training
	Extracting data requires prior knowledge of what data is available, and water professionals do not have access to adequate interfaces to explore their data.	• Access to data exploration software • Collaboration between data workers and water professionals
Data mining	Mining data for relationships increases the odds of finding spurious correlations.	• Stronger statistical tests to determine significance. • Extensive use of domain-specific knowledge.
Modelling	Water professionals don't have the required skills to build data-driven models, while data scientists don't have the WRRF-specific knowledge required to develop adequate wastewater treatment models.	• Strong collaboration between water professionals and data scientists. • Extensive use of domain-specific knowledge. • Training
	Simulation software is mostly aimed at mechanistic modelling and doesn't provide tools to create data-driven models	• Provide an integrated toolchain that supports more modelling methodologies.

(*continued*)

Table 1 | continued

Step of the pipeline	Issue encountered	Paths for improvement
Comprehension	Computers think with math; humans don't.	• Synthesis of information via interactive dashboards, visualizations and reports. • Use of colour, shape, spatial placement. • Provide adapted interfaces for each task.
	Human-machine interfaces are sometimes confusing and distracting	• Make use of high-performance HMI guidelines. • Collaborate with UI designers to create powerful interfaces.
Action	Complex control strategies are prone to fail when data quality drops.	• Online automatic fault detection. • Controller reconfiguration – automatic switch to fallback strategy.
	Control objectives are often ill-defined among different plant stakeholders.	• Systematic objective definition. • Continuous communication and collaboration between all stakeholders.

fascinating as wastewater treatment and resource recovery themselves.

ACKNOWLEDGEMENTS

This work has been carried out at Université Laval, supported financially by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the award of an NSERC Discovery Grant awarded to Peter Vanrolleghem (grant number RGPIN-2016-06522) and an NSERC ES D PhD scholarship for Jean-David Therrien (grant number PGSD3-519336-2018). Peter Vanrolleghem holds the Canada Research Chair on Water Quality Modelling.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Alferes, J. & Vanrolleghem, P. A. 2016 Efficient automated quality assessment: dealing with faulty on-line water quality sensors. *AI Communications* **29** (6), 701–709. doi:10.3233/AIC-160713.

Alferes, J., Tik, S., Copp, J. B. & Vanrolleghem, P. A. 2013 Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection. *Water Science and Technology* **68** (5), 1022–1030. doi:10.2166/wst.2013.302.

Alferes, J., Copp, J. B., Weijers, S. & Vanrolleghem, P. A. 2015 Validating data quality for water quality monitoring: Objective comparison of three data quality assessment approaches. In: *Proceedings of the New Developments in IT & Water Conference*, February 8–10, 2015, Rotterdam, The Netherlands.

Al-Omari, A., Wett, B., Nopens, I., De Clippeleir, H., Han, M., Regmi, P., Bott, C. & Murthy, S. 2015 Model-based evaluation of mechanisms and benefits of mainstream shortcut nitrogen removal processes. *Water Science and Technology* **71** (6), 840–847. doi:10.2166/wst.2015.022.

Åmand, L., Olsson, G. & Carlsson, B. 2013 Aeration control – a review. *Water Science and Technology* **67** (11), 2374–2398. doi:10.2166/wst.2013.139.

Amaral, A., Schraa, O., Rieger, L., Gillot, S., Fayolle, Y., Bellandi, G., Amerlinck, Y., Gori, R., Neves, R. & Nopens, I. 2017 Towards advanced aeration modelling: from blower to bubbles to bulk. *Water Science and Technology* **75** (3), 507–517. doi:10.2166/wst.2016.365.

Amerlinck, Y. 2015 *Model Refinements In View of Wastewater Treatment Plant Optimization: Improving The Balance In Sub-Model Detail*. PhD Thesis, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium, pp. 289.

Ayesa, E., De la Sota, A., Grau, P., Sagarna, J. M., Salterain, A. & Suescun, J. 2006 Supervisory control strategies for the new WWTP of Galindo-Bilbao: the long run from the conceptual design to the full-scale experimental validation. *Water Science and Technology* **53** (4–5), 193–201. doi:10.2166/wst.2006.124.

Aymerich, I., Rieger, L., Sobhani, R., Rosso, D. & Corominas, L. 2015 The difference between energy consumption and energy

cost: modelling energy tariff structures for water resource recovery facilities. *Water Research* **81** (2015), 113–123. doi:10.1016/j.watres.2015.04.033.

Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., Wilcox, K., Karen, L. & Lee, S. 2019 *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*. Richland, WA, USA, p. 96.

Benedetti, L., Dirckx, G., Bixio, D., Thoeye, C. & Vanrolleghem, P. A. 2006 Substance flow analysis of the wastewater collection and treatment system. *Urban Water Journal* **3** (1), 33–42. doi:10.1080/15730620600578694.

Blanke, M., Izadi-Zamanabadi, R., Bøgh, S. A. & Lunau, C. P. 1997 Fault-tolerant control systems – a holistic view. *Control Engineering Practice* **5** (5), 693–702. doi:10.1016/S0967-0661(97)00051-8.

Blumensaat, F., Leitão, J. P., Ort, C., Rieckermann, J., Scheidegger, A., Vanrolleghem, P. A. & Villez, K. 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 and Technology* **53** (15), 8488–8498. doi:10.1021/acs.est.8b06481.

Box, G. E. P. & Jenkins, G. M. 1970 *Time Series Analysis, Forecasting, and Control*, 2nd edn. Holden-Day Publishing, San Francisco, CA, USA, pp. 553.

Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C. & Vanrolleghem, P. A. 2011 Performance evaluation of fault detection methods for wastewater treatment processes. *Biotechnology and Bioengineering* **108** (2), 333–344. doi:10.1002/bit.22953.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U. & Poch, M. 2018 Transforming data into knowledge for improved wastewater treatment operation: a critical review of techniques. *Environmental Modelling and Software* **106** (August), 89–103. doi:10.1016/j.envsoft.2017.11.023.

Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Machine Learning* **20**, 273–297. doi:10.1007/BF00994018.

Damman, S., Helness, H., Grindvoll, I. L. T. & Sun, C. 2019 Citizen science to enhance evaluation of local wastewater treatment – a case study from Oslo. *Water Science and Technology* **79** (10), 1887–1896. doi:10.2166/wst.2019.180.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M. & Zupan, B. 2013 Orange: data mining toolbox in Python. *Journal of Machine Learning Research* **14**, 2349–2353.

De Mulder, C., Flameling, T., Weijers, S., Amerlinck, Y. & Nopens, I. 2018 An open software package for data reconciliation and gap filling in preparation of water and resource recovery facility modeling. *Environmental Modelling and Software* **107** (September), 186–198. doi:10.1016/j.envsoft.2018.05.015.

Denton, F. T. 1985 Data mining as an industry. *The Review of Economics and Statistics* **67** (1), 124–127. doi:10.2307/1928442.

Devisscher, M., Clacci, G., Fé, L., Benedetti, L., Bixio, D., Thoeye, C., De Gueldre, G., Marsili-Libelli, S. & Vanrolleghem, P. A. 2006 Estimating costs and benefits of advanced control for wastewater treatment plants – the MAgIC methodology. *Water Science and Technology* **53** (4–5), 215–223. doi:10.2166/wst.2006.126.

Drucker, S. M. & Fernandez, R. 2015 *A Unifying Framework for Animated and Interactive Unit Visualizations. Technical Report. Microsoft Research*. p. 9. Available from: http://research.microsoft.com/pubs/252104/sanddance.pdf (accessed 26 May 2020).

Dürrenmatt, D. J. & Gujer, W. 2012 Data-driven modeling approaches to support wastewater treatment plant operation. *Environmental Modelling and Software* **30** (April), 47–56. doi:10.1016/j.envsoft.2011.11.007.

Eberl, H. J., Picioreanu, C., Heijnen, J. J. & Van Loosdrecht, M. C. M. 2000 Three-dimensional numerical study on the correlation of spatial structure, hydrodynamic conditions, and mass transfer and conversion in biofilms. *Chemical Engineering Science* **55** (24), 6209–6222. doi:10.1016/S0009-2509(00)00169-X.

Farley, S. S., Dawson, A., Goring, S. J. & Williams, J. W. 2018 Situating ecology as a big-data science: current advances, challenges, and solutions. *BioScience* **68** (8), 563–576. doi:10.1093/biosci/biy068.

Garcia-Ochoa, F. & Gomez, E. 2009 Bioreactor scale-up and oxygen transfer rate in microbial processes: an overview. *Biotechnology Advances* **27** (2), 153–176. doi:10.1016/j.biotechadv.2008.10.006.

Garneau, C. & Vanrolleghem, P. A. 2018 Neural network for tuning-friendly automatic outlier detection in water quality time series. In *Proceedings of the 9th International Congress on Environmental Modelling and Software, Fort Collins, CO, USA*.

Garrido-Baserba, M., Corominas, L., Cortés, U., Rosso, D. & Poch, M. 2020 The fourth-revolution in the water sector encounters the digital revolution. *Environmental Science and Technology* **54** (8), 4698–4705. doi:10.1021/acs.est.9b04251.

Gernaey, K. V., Van Loosdrecht, M. C. M., Henze, M., Lind, M. & Jørgensen, S. B. 2004 Activated sludge wastewater treatment plant modelling and simulation: state of the art. *Environmental Modelling and Software* **19** (9), 763–783. doi:10.1016/j.envsoft.2003.03.005.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. & Kagal, L. 2018 Explaining explanations: an overview of interpretability of machine learning. In: *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, October 1–4 2018, *Turin, Italy*, pp. 80–89.

Hadjimichael, A., Comas, J. & Corominas, L. 2016 Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector. *AI Communications* **29** (6), 747–756. doi:10.3233/AIC-160714.

Haimi, H., Mulas, M., Corona, F. & Vahala, R. 2013 Data-derived soft-sensors for biological wastewater treatment plants: an overview. *Environmental Modelling and Software* **47**, 88–107. doi:10.1016/j.envsoft.2013.05.009.

Henze, M., Gujer, W., Mino, T. & van Loosdrecht, M. C. M. 2000 *Activated Sludge Models. Reprint. Technical Report.*

Scientific and Technical Report Series. IWA Publishing, London, UK, p. 130.

Hernández-del-Olmo, F., Gaudioso, E., Dormido, R. & Duro, N. 2016 Energy and environmental efficiency for the N-ammonia removal process in wastewater treatment plants by means of reinforcement learning. *Energies* **9** (9), 755. doi:10.3390/en9090755.

Hernández-del-Olmo, F., Gaudioso, E., Duro, N. & Dormido, R. 2019 Machine learning weather soft-sensor for advanced control of wastewater treatment plants. *Sensors (Switzerland)* **19** (14), 1–12. doi:10.3390/s19143139.

Hollifield, B. R., Oliver, D., Nimmo, I. & Habibi, E. 2008 *The High Performance HMI Handbook: A Comprehensive Guide to Designing, Implementing and Maintaining Effective HMIs for Industrial Plant Operations*, 1st edn. PAS, Inc., Houston, TX, USA, p. 206.

HSE GB 1997 *The Explosion and Fires at the Texaco Refinery, Milford Haven, 24 July 1994: A Report of the Investigation by the Health and Safety Executive Into the Explosion and Fires on the Pembroke Cracking Company Plant at the Texaco Refinery, Milford Haven on 24 J. Incident Report Series*. HSE Books, London, UK, p. 66.

Ingildsen, P. & Olsson, G. 2016 *Smart Water Utilities: Complexity Made Simple*. IWA Publishing, London, UK, p. 304.

Ingildsen, P., Jeppsson, U. & Olsson, G. 2002 Dissolved oxygen controller based on on-line measurements of ammonium combining feed-forward and feedback. *Water Science and Technology* **45** (4–5), 453–460. doi:10.2166/wst.2002.0649.

Inmon, W. H. 1998 *Building the Data Warehouse*, 3rd edn. Wiley Computer Publishing, New York, NY, USA, p. 412.

Isermann, R. & Ballé, P. 1997 Trends in the application of model-based fault detection and diagnosis of technical processes. *Control Engineering Practice* **5** (5), 709–719. doi:10.1016/S0967-0661(97)00053-1.

IWA and Xylem Inc. 2019 *Digital Water Report. Technical Report*. London, UK, p. 44.

Jimenez, J., Latrille, E., Harmand, J., Robles, A., Ferrer, J., Gaida, D., Wolf, C., Mairet, F., Bernard, O., Alcaraz-Gonzalez, V., Mendez-Acosta, H., Zitomer, D., Totzke, D., Spanjers, H., Jacobi, F., Guwy, A., Dinsdale, R., Premier, G., Mazhegrane, S., Ruiz-Filippi, G., Seco, A., Ribeiro, T., Pauss, A. & Steyer, J. P. 2015 Instrumentation and control of anaerobic digestion processes: a review and some research challenges. *Reviews in Environmental Science and Biotechnology* **14** (4), 615–648. doi:10.1007/s11157-015-9382-6.

Kolditz, O., Rink, K., Nixdorf, E., Fischer, T., Bilke, L., Naumov, D., Liao, Z. & Yue, T. 2019 Environmental information systems: paving the path for digitally facilitated water management (Water 4.0). *Engineering* **5** (5), 828–832. doi:10.1016/j.eng.2019.08.002.

Korodi, A., Radu, M. A. & Crisan, R. 2018 Non-invasive control solution inside higher-level OPC UA based wrapper for optimizing groups of wastewater systems. In: *Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, September 4–7 2018, *Turin, Italy*, pp. 597–604.

Le, Q. H., Verheijen, P. J. T., van Loosdrecht, M. C. M. & Volcke, E. I. P. 2018 Experimental design for evaluating WWTP data by linear mass balances. *Water Research* **142**, 415–425. doi:10.1016/j.watres.2018.05.026.

Lee, D. S., Vanrolleghem, P. A. & Jong, M. P. 2005 Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. *Journal of Biotechnology* **115** (3), 317–328. doi:10.1016/j.jbiotec.2004.09.001.

Liu, Z. H. & Gawlick, D. 2015 Management of flexible schema data in RDBMSs – opportunities and limitations for NoSQL-. In: *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR)*, January 4–7 2015, *Asilomar, CA, USA*.

Lovell, M. C. 1983 Data mining. *The Review of Economics and Statistics* **65** (1), 1–12. doi:10.2307/1924403.

Makropoulos and Savić 2019 Urban hydroinformatics: past, present and future. *Water (Switzerland)* **11** (10), 1959. doi:10.3390/w11101959.

Mannina, G., Ekama, G., Caniani, D., Cosenza, A., Esposito, G., Gori, R., Garrido-Baserba, M., Rosso, D. & Olsson, G. 2016 Greenhouse gases from wastewater treatment – a review of modelling tools. *Science of the Total Environment* **551–552** (1 May 2016), 254–270. doi:10.1016/j.scitotenv.2016.01.163.

Martin, C. & Vanrolleghem, P. A. 2014 Analysing, completing, and generating influent data for WWTP modelling: a critical review. *Environmental Modelling & Software* **60** (October 2014), 188–201. doi:10.1016/j.envsoft.2014.05.008.

Matthews, W. 2017 *New Roles for Process Historians*. *InTech Magazine* Nov–Dec, https://ww2.isa.org/intech/20171202/.

Meirlaen, J. 2002 *Immission Based Real-Time Control of the Integrated Urban Wastewater System*. PhD Thesis, Department of Mathematical Modelling, Biometrics and Process Control, Gent University, Ghent, Belgium, p. 246.

Mhaskar, P., Liu, J. & Christofides, P. D. 2013 *Fault-Tolerant Process Control: Methods and Applications*. Springer London, London, UK, p. 261.

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q. & Arocena, P. C. 2018 Data lake management: challenges and opportunities. *Proceedings of the VLDB Endowment* **12** (12), 1986–1989. doi:10.14778/3352063.3352116.

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

OGC Consortium 2011 OGC WaterML 2.0: Part 1- Timeseries. Measurement (1–151). http://www.opengeospatial.org/.

Ohmura, K., Thürlimann, C. M., Kipf, M., Carbajal, J. P. & Villez, K. 2019 Characterizing long-term wear and tear of ion-selective pH sensors. *Water Science and Technology* **80** (3), 541–550. doi:10.2166/wst.2019.301.

Olsson, G. 2012 ICA and me – a subjective review. *Water Research* **46** (6), 1585–1624. doi:10.1016/j.watres.2011.12.054.

Patry, B. 2020 *Suivi, Compréhension et Modélisation d'une Technologie à Biofilm Pour l'augmentation de La Capacité Des Étangs Aérés*. PhD Thesis, Département de génie civil et de génie des eaux, Université Laval, Québec, Canada, p. 190.

Pizarro, G., Griffeath, D. & Noguera, D. R. 2001 Quantitative cellular automaton model for biofilms. *Journal of Environmental Engineering* **127** (9), 782–789. doi:10.1061/(ASCE)0733-9372(2001)127:9(782).

Plana, Q., Alferes, J., Fuks, K., Kraft, T., Maruéjouls, T., Torfs, E. & Vanrolleghem, P. A. 2019 Towards a water quality database for raw and validated data with emphasis on structured metadata. *Water Quality Research Journal* **54** (1), 1–9. doi:10.2166/wqrj.2018.013.

Psichogios, D. C. & Ungar, L. H. 1992 A hybrid neural network-first principles approach to process modeling. *AIChE Journal* **38** (10), 1499–1511. doi:10.1002/aic.690381003.

Regmi, P., Holgate, B., Fredericks, D., Miller, M. W., Wett, B., Murthy, S. & Bott, C. B. 2015 Optimization of a mainstream nitritation-denitritation process and anammox polishing. *Water Science and Technology* **72** (4), 632–642. doi:10.2166/wst.2015.261.

Regmi, P., Stewart, H., Amerlinck, Y., Arnell, M., García, P. J., Johnson, B., Maere, T., Miletić, I., Miller, M., Rieger, L., Samstag, R., Santoro, D., Schraa, O., Snowling, S., Takács, I., Torfs, E., van Loosdrecht, M. C. M., Vanrolleghem, P. A., Villez, K., Volcke, E. I. P., Weijers, S., Grau, P., Jimenez, J. & Rosso, D. 2019 The future of WRRF modelling – outlook and challenges. *Water Science and Technology* **79** (1), 3–14. doi:10.2166/wst.2018.498.

Revollar, S., Vilanova, R., Vega, P., Francisco, M. & Meneses, M. 2020 Wastewater treatment plant operation: simple control schemes with a holistic perspective. *Sustainability (Switzerland)* **12** (3), 1–28. doi:10.3390/su12030768.

Rieger, L. & Olsson, G. 2012 Why many control systems fail. *Water Environment and Technology* **2012** (June), 42–45. doi:10.2175/193864711802764779.

Rieger, L. & Vanrolleghem, P. A. 2008 monEAU: a platform for water quality monitoring networks. *Water Science and Technology* **57** (7), 1079–1086. doi:10.2166/wst.2008.135.

Rieger, L., Alex, J., Winkler, S., Boehler, M., Thomann, M. & Siegrist, H. 2003 Progress in sensor technology – progress in process control? Part I: sensor property investigation and classification. *Water Science and Technology* **47** (2), 103–112. doi:10.2166/wst.2003.0096.

Rieger, L., Takács, I., Villez, K., Siegrist, H., Lessard, P., Vanrolleghem, P. A. & Comeau, Y. 2010 Data reconciliation for wastewater treatment plant simulation studies – planning for high-quality data and typical sources of errors. *Water Environment Research* **82** (5), 426–433. doi:10.2175/106143009×12529484815511.

Rieger, L., Jones, R. M., Dold, P. L. & Bott, C. B. 2013 Ammonia-based feedforward and feedback aeration control in activated sludge processes. *Water Environment Research* **86** (1), 63–73. doi:10.2175/106143013×13596524516987.

Rittmann, B. E., Boltz, J. P., Brockmann, D., Daigger, G. T., Morgenroth, E., Sørensen, K. H., Takács, I., van Loosdrecht, M. C. M. & Vanrolleghem, P. A. 2018 A framework for good biofilm reactor modeling practice (GBRMP). *Water Science and Technology* **77** (5), 1149–1164. doi:10.2166/wst.2018.021.

Rosén, C. & Olsson, G. 1998 Disturbance detection in wastewater treatment plants. *Water Science and Technology* **37** (12), 197–205. doi:10.2166/wst.1998.0542.

Rosén, C., Jeppsson, U. & Vanrolleghem, P. A. 2004 Towards a common benchmark for long-term process control and monitoring performance evaluation. *Water Science and Technology* **50** (11), 41–49. doi:10.2166/wst.2004.0669.

Rosén, C., Rieger, L., Jeppsson, U. & Vanrolleghem, P. A. 2008 Adding realism to simulated sensors and actuators. *Water Science and Technology* **57** (3), 337–344. doi:10.2166/wst.2008.130.

Rosso, D., Larson, L. E. & Stenstrom, M. K. 2008 Aeration of large-scale municipal wastewater treatment plants: state of the art. *Water Science and Technology* **57** (7), 973–978. doi:10.2166/wst.2008.218.

Russo, S., Disch, A., Blumensaat, F. & Villez, K. 2019 Anomaly detection using deep autoencoders for in-situ wastewater systems monitoring data. In: *Proceedings of the 10th IWA Symposium on Systems Analysis and Integrated Assessment (Watermatex2019)*, September 1–4 2019, Copenhagen, Denmark.

Schneider, M. Y., Carbajal, J. P., Furrer, V., Sterkele, B., Maurer, M. & Villez, K. 2019 Beyond signal quality: the value of unmaintained pH, dissolved oxygen, and oxidation-reduction potential sensors for remote performance monitoring of on-site sequencing batch reactors. *Water Research* **161**, 639–651. doi:10.1016/j.watres.2019.06.007.

Schraa, O., Tole, B. & Copp, J. B. 2006 Fault detection for control of wastewater treatment plants. *Water Science and Technology* **53** (4–5), 375–382. doi:10.2166/wst.2006.143.

Shah, P. & Hoeffner, J. 2002 Review of graph comprehension research: implications for instruction. *Educational Psychology Review* **14** (1), 47–69. doi:10.1023/a:1013180410169.

Shiva Kumar, B. & Venkateswarlu, C. 2012 Estimating biofilm reaction kinetics using hybrid mechanistic-neural network rate function model. *Bioresource Technology* **103** (1), 300–308. doi:10.1016/j.biortech.2011.10.006.

Shyamal, S. & Swartz, C. L. E. 2019 Real-time energy management for electric arc furnace operation. *Journal of Process Control* **74** (February 2019), 50–62. doi:10.1016/j.jprocont.2018.03.002.

Sirkiä, J., Laakso, T., Ahopelto, S., Ylijoki, O., Porras, J. & Vahala, R. 2017 Data utilization at Finnish water and wastewater utilities: current practices vs. state of the art. *Utilities Policy* **45** (April 2017), 69–75. doi:10.1016/j.jup.2017.02.002.

Skobelev, D. O., Zaytseva, T. M., Kozlov, A. D., Perepelitsa, V. L. & Makarova, A. S. 2011 Laboratory information management systems in the work of the analytic laboratory. *Measurement Techniques* **53** (10), 1182–1189. doi:10.1007/s11018-011-9638-7.

Solon, K., Flores-Alsina, X., Kazadi Mbamba, C., Ikumi, D., Volcke, E. I. P., Vaneeckhaute, C., Ekama, G., Vanrolleghem, P. A., Batstone, D. J., Gernaey K, V. & Jeppsson, U. 2017 Plant-wide modelling of phosphorus transformations in wastewater treatment systems: impacts of control and operational strategies. *Water Research* **113**, 97–110. doi:10.1016/j.watres.2017.02.007.

Spérandio, M., Pocquet, M., Guo, L., Ni, B. J., Vanrolleghem, P. A. & Yuan, Z. 2016 Evaluation of different nitrous oxide production models with four continuous long-term wastewater treatment process data series. *Bioprocess and Biosystems Engineering* **39** (3), 493–510. doi:10.1007/s00449-015-1532-2.

Takács, I., Patry, G. G. & Nolasco, D. 1991 A dynamic model of the clarification-thickening process. *Water Research* **25** (10), 1263–1271. https://doi.org/10.1016/0043-1354(91)90066-Y.

Talebizadeh, M., Belia, E. & Vanrolleghem, P. A. 2016 Influent generator for probabilistic modeling of nutrient removal wastewater treatment plants. *Environmental Modelling and Software* **77**, 32–49. doi:10.1016/j.envsoft.2015.11.005.

Thornhill, N. F., Shoukat Choudhury, M. A. A. & Shah, S. L. 2004 The impact of compression on data-driven process analyses. *Journal of Process Control* **14** (4), 389–398. doi:10.1016/j.jprocont.2003.06.003.

Thürlimann, C. M., Dürrenmatt, D. J. & Villez, K. 2015 Energy and process data processing and visualisation for optimising wastewater treatment plants. *Water Practice and Technology* **10** (1), 10–18. doi:10.2166/wpt.2015.002.

van der Hoek, J., Duijff, R. & Reinstra, O. 2018 Nitrogen recovery from wastewater: possibilities, competition with other resources, and adaptation pathways. *Sustainability (Switzerland)* **10** (12), 4605. doi:10.3390/su10124605.

Vaneeckhaute, C., Claeys, F., Tack, F., Meers, E., Belia, E. & Vanrolleghem, P. A. 2018 Development, implementation, and validation of a generic nutrient recovery model (NRM) library. *Environmental Modelling & Software* **99**, 170–209.

Vanrolleghem, P. A. 2014 Bits and bytes and bugs – On monitoring and control in WRRF's (aka WWTPs). In *Smart Wastewater Virginia WEA Seminar*, April 30-May 1 2014, Richmond, VA, USA (Invited Keynote Lecture).

Vanrolleghem, P. A. & Lee, D. S. 2003 On-line monitoring equipment for wastewater treatment processes: state of the art. *Water Science and Technology* **47** (2), 1–34. doi:10.2166/wst.2003.0074.

Varga, E., Hauduc, H., Barnard, J., Dunlap, P., Jimenez, J., Menniti, A., Schauer, P., Lopez Vazquez, C. M., Gu, A. Z., Sperandio, M. & Takács, I. 2018 Recent advances in bio-P modelling – a new approach verified by full-scale observations. *Water Science and Technology* **78** (10), 2119–2130. doi:10.2166/wst.2018.490.

Vezzaro, L., Pedersen, J. W., Larsen, L. H., Thirsing, C., Duus, L. B. & Mikkelsen, P. S. 2020 Evaluating the performance of a simple phenomenological model for online forecasting of ammonium concentrations at WWTP inlets. *Water Science and Technology* **81** (1), 109–120. doi:10.2166/wst.2020.085.

Villez, K., Vanrolleghem, P. A. & Corominas, L. 2020 A general-purpose method for pareto optimal placement of flow rate and concentration sensors in networked systems – with application to wastewater treatment plants. *Computers and Chemical Engineering* **139**, 106880. doi:10.1016/j.compchemeng.2020.106880.

Water Environment Federation 2014 *Moving Toward Resource Recovery Facilities. Technical Report*. Alexandria, VA, USA, pp. 16.

Water Online and SWAN 2019 *Smart Water Report: Navigating The Smart Water Journey: From Leadership to Results*. Horsham, PA, USA. Water Online, p. 58.

Weijers, S. 2000 *Modelling, Identification and Control of Activated Sludge Plants for Nitrogen Removal*. PhD Thesis, Eindhoven University of Technology, Eindoven, The Netherlands, p. 235.

Werbos, P. J. 1982 Applications of advances in nonlinear sensitivity analysis. In: *System Modeling and Optimization* (R. F. Drenick & F. Kozin, eds). Springer-Verlag, Berlin/Heidelberg, Germany, pp. 762–770.

Whitt, M. D. 2012 *Successful Instrumentation and Control Systems Design*, 2nd edn. International Society of Automation, Durham, NC, USA, p. 531.

Wigner, E. P. 1960 The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics* **13** (1), 1–14. doi:10.1002/cpa.3160130102.

Wilkinson, M. D., Dumontier, M., IjJ, A., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. 2016 The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* **3** (1), 160018. doi:10.1038/sdata.2016.18.

Yee, I. & Eren, H. 2012 'Data historian.' In: *Instrument Engineers' Handbook, Volume Three: Process Software and Digital Networks* (B. G. Liptak & H. Eren, eds). CRC Press, Boca Raton, FL, USA, pp. 465–470.

Yoo, C. K., Villez, K., Lee, I. B., Rosén, C. & Vanrolleghem, P. A. 2007 Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. *Biotechnology and Bioengineering* **96** (4), 687–701. doi:10.1002/bit.21220.

Yuan, Z., Olsson, G., Cardell-Oliver, R., van Schagen, K., Marchi, A., Deletic, A., Urich, C., Rauch, W., Liu, Y. & Jiang, G. 2019 Sweating the assets – the role of instrumentation, control and automation in urban water systems. *Water Research* **155**, 381–402. doi:10.1016/j.watres.2019.02.034.

Zegers, E., Atlin, S., Nyman, K. & Berenyi, A. 2019 Toronto's new data management platform and how we plan to manage the quality of data that feeds it. In *Proceedings of WEFTEC 2019*, September 21–25 2019, Chicago, IL, USA.

Zhang, Y. & Jiang, J. 2008 Bibliographical review on reconfigurable fault-tolerant control systems. *Annual Reviews in Control* **32** (2), 229–252. doi:10.1016/j.arcontrol.2008.03.008.

First received 29 May 2020; accepted in revised form 5 August 2020. Available online 18 August 2020