Water Research 157 (2019) 498–513

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Review

# Data-driven performance analyses of wastewater treatment plants: A review

Kathryn B. Newhart [a], Ryan W. Holloway [b], Amanda S. Hering [c, *], Tzahi Y. Cath [a, **]

[a] Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, CO, 80401, USA
[b] Kennedy/Jenks Consultants, Rancho Cordova, CA, 95670, USA
[c] Department of Statistical Science, Baylor University, Waco, TX, 76798, USA

ARTICLE INFO

*Article history:*
Received 9 September 2018
Received in revised form
12 March 2019
Accepted 16 March 2019
Available online 21 March 2019

*Keywords:*
Wastewater treatment
Big data
Statistical process control
Process optimization
Monitoring

ABSTRACT

Recent advancements in data-driven process control and performance analysis could provide the wastewater treatment industry with an opportunity to reduce costs and improve operations. However, big data in wastewater treatment plants (WWTP) is widely underutilized, due in part to a workforce that lacks background knowledge of data science required to fully analyze the unique characteristics of WWTP. Wastewater treatment processes exhibit nonlinear, nonstationary, autocorrelated, and co-correlated behavior that (i) is very difficult to model using first principals and (ii) must be considered when implementing data-driven methods. This review provides an overview of data-driven methods of achieving fault detection, variable prediction, and advanced control of WWTP. We present how big data has been used in the context of WWTP, and much of the discussion can also be applied to water treatment. Due to the assumptions inherent in different data-driven modeling approaches (e.g., control charts, statistical process control, model predictive control, neural networks, transfer functions, fuzzy logic), not all methods are appropriate for every goal or every dataset. Practical guidance is given for matching a desired goal with a particular methodology along with considerations regarding the assumed data structure. References for further reading are provided, and an overall analysis framework is presented.



**Contents**



* Corresponding author.
** Corresponding author.
*E-mail addresses:* Mandy_Hering@baylor.edu (A.S. Hering), tcath@mines.edu (T.Y. Cath).

https://doi.org/10.1016/j.watres.2019.03.030
0043-1354/

## 1. Introduction

Municipal wastewater treatment plants (WWTP) continuously monitor and collect data from unit processes, but the data are often underutilized. Due to the size and complexity of datasets currently generated by WWTP and the lack of data science background for WWTP professionals, it can be challenging to efficiently collect, manage, and analyze the data (Diebold, 2003; Kadiyala, 2018; Manyika et al., 2011; Regmi et al., 2018). Despite widespread interest in big data integration at WWTP, most raw data are stored in their original format for potential future performance analyses with little consideration to their structure or the organization of the data repository. To extract information from this "data lake," multiple factors need to be considered, including the unique characteristics of WWTP data and the goals of an individual WWTP. This review describes different data-driven methods and how they can be used to address problems specific to WWTP. While reviews exist for academic applications (Corominas et al., 2018; Hadjimichael et al., 2016; Olsson, 2012), this review is from an applied perspective; designed to demystify what methods should be used and under what circumstances. If operations data were analyzed in real-time with data-driven tools, WWTP could promptly detect and respond to process failures, inefficiencies, and abnormalities. Early correction of these WWTP faults could reduce (i) downtime, (ii) effluent discharge violation, and (iii) resource consumption such as energy, chemicals, and labor. Additional applications of big data to improve WWTP operation include data validation; online monitoring of difficult-to-measure variables; predictive maintenance (Golhar and Dallas, 2016); system and energy optimization; and tailored water reuse (i.e., producing water of distinct qualities for different reuse purposes).

Big data integration at WWTP will have the most substantial impact on process control. WWTP primarily use fixed upper and lower limits of process variables to monitor and control treatment processes. These limits are adjusted based on a WWTP operator's background knowledge of the specific system as well as online and offline water quality data, but rarely are more advanced methods of determining process limits (i.e., modeling) used. In part, this is due to the variability in the sensors that provide the data. Water quality is monitored in real-time by online digital sensors that transmit a voltage or current corresponding to an electrochemical reaction or physical change inside the sensors as they interact with the environment (e.g., constituent concentration, flowrate, pressure, level). To calibrate these sensors, measurements using analog devices or laboratory analyses are correlated to voltage or current changes from the sensor. However, solids deposition, biofilm formation, and precipitates can interfere with the sensor's voltage or current change and thus with the sensor's measurement accuracy. Offline analyses to calibrate sensors and monitor process performance is performed either on- or off-site, and the time required for each analysis can range from minutes to days, depending on the laboratory equipment and available staff. The resulting datasets often have missing values, contain outliers, and are sensitive to the interdependent, nonlinear, and nonstationary nature of WWTP data (Olsson et al., 2005; Rosen and Lennox, 2001), which makes WWTP difficult to model mathematically for the purpose of performing process control (Dürrenmatt and Gujer, 2012). Consequently, big data tools can provide an alternative approach (see Section 3.3.1).

To address the unique features of WWTP processes and the resulting data, WWTP need access to a labor pool of WWTP professionals with backgrounds in data science (Kadiyala, 2018; Sirkiä et al., 2017) and need more practical guidance on full-scale big data implementation (US EPA, 2014). This paper serves as a WWTP engineer's guide to understanding the advantages and limitations of applying different data-driven methods for process control and optimization. We present how big data has been used in the context of WWTP and review the academic literature that describes state-of-the-art methods of analyzing WWTP data for advanced control and process optimization, noting that state-of-the-art in WWTP does not reflect state-of-the-art in the data sciences. Many more advanced methodologies have been developed but not yet tested in the WWTP context. References for further reading for each broad category of methods are given along with suggestions for methods that have promise for the water industry. Additionally, much of this discussion can also be applied to water treatment. Section 2 provides the reader with an introduction to big data and data-driven analysis, WWTP, and the prominent data characteristics of WWTP processes that may impact the results of data analysis. Section 3 follows with analytical methods to improve process control using examples of real WWTP for the purpose of *fault detection*, *variable prediction*, and *advanced automated control* (Fig. 1). Some methods listed in Section 3 have multiple applications; thus, the full description is provided when the method is first presented. Section 4 concludes with lessons learned from this review of advanced data analysis at WWTP; outlines the challenges facing modern WWTP as they integrate data-driven solutions into their operations; and identifies some existing methodologies that have not yet been tested with WWTP data.

## 2. Background

### 2.1. Big data

The term "big data" encompasses the modern overabundance of data produced by online and offline analysis, and the innovative

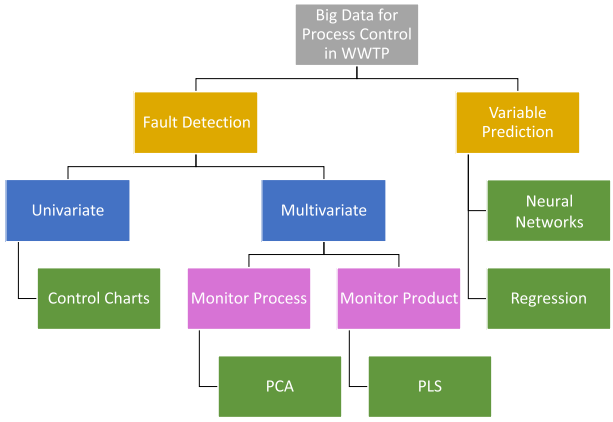



**Fig. 1.** The data-driven methods identified in green are examples of methods that have demonstrated good performance in WWTP for the purpose indicated by the tree diagram. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

tools used to analyze the data. Big data can be broadly characterized by the 5 Vs, which are volume, variety, velocity, veracity, and value (Golhar and Dallas, 2016; Laney, 2001; Slawecki et al., 2016). Volume is the physical storage size required to save collected data. WWTP rarely monitor the total size of the collected data, as storage is inexpensive relative to the operating budget of a facility. Variety refers to the different types of data collected, including file type and data structure. Maintenance notes are considered unstructured, but measured sensor values are structured because they have a quantifiable/measurable significance and are stored in a separate database. Velocity is the rate of data storage and analysis in real-time. The computer processing speed (i.e., velocity) to monitor a WWTP needs to be sufficiently fast; able to collect, sort, clean, analyze, and interpret data quickly and effectively. Veracity is the quality and trustworthiness of the data and can be considered a measure of uncertainty. One data source in WWTP with questionable veracity is sensing technology. Even with regular maintenance and calibration, sensor measurements drift over time, and the drift may differ between sensors of the same model exposed to the same environmental conditions (Haimi et al., 2013; Olsson, 2012; Vanrolleghem and Lee, 2003). Finally, value is a subjective characterization of data quality referring to (i) the cost of data collection and storage relative to the value it produces, and (ii) if analyses are performed to produce information from the data.

The two general steps for extracting information from industrial processes are data management and data analysis (Gandomi and Haider, 2015; Labrinidis and Jagadish, 2012). Broadly speaking, data management includes the acquisition, aggregation, and cleaning of raw data to prepare it for analysis. Analysis may include modeling or advanced statistics to make inferences about the process and can provide site-specific, actionable knowledge. In this paper, we primarily discuss the second step, approaches to data analysis, for addressing problems in modern WWTP.

To maximize value in data-driven analysis, engineers need to engage with statisticians, data scientists, and computers scientists to develop industry-specific tools. For WWTP, there are few data-driven tools that are commercially available, and most are designed as black-box, turnkey solutions with limited insight into computational details and causal factors. Given the nature of WWTP, in which an operator receives information from multiple sources and makes an educated decision, black-box systems are frequently not trusted by WWTP. Generic data-driven tools also exist, but the average WWTP engineer lacks the background in data science to apply these tools to a complex system like WWTP. In order to produce impactful and accurate results, big data analyses need to be implemented with WWTP-specific process knowledge and informed data characterization. In the next section, we provide a brief introduction to the types of processes in WWTP that are the focus of this paper.

### 2.2. Water & wastewater treatment

In the US, WWTP receive raw wastewater from sanitary sewer networks and use multiple unit processes to remove contaminants until the water meets standards for discharge or reuse as regulated under the Clean Water Act (Clean Water Act, 1977). Municipal wastewater treatment begins with physical treatment processes such as screening and grit trapping to remove large material and debris from the raw wastewater, followed by biological treatment. The most common method of biological wastewater treatment is the conventional activated sludge (CAS) process. Aeration and recirculation of biologically-active solids ("biosolids" or "solids") maintain diverse communities of microorganisms in CAS to degrade a wide range of organic compounds and nutrients. Clarification (gravity settling) separates treated water from the biosolids, followed by disinfection and discharge to the environment. Depending on the initial quality of the water, advanced treatment may also be required (e.g., diffusive membrane technology or advanced oxidation processes) to remove salts or contaminants of emerging concern (e.g., pharmaceuticals, personal care products, synthetic organic compounds).

The quantity and quality of water and solids are measured from the headworks of a facility, through the treatment train, to the final discharge point (Fig. 2). Some variables are a general measure of the health of a system, such as pH. Other variables are included in control loops with pumps, valves, and air blowers to optimize treatment, such as dissolved oxygen (DO), ammonia ($NH_4^+$), and nitrate ($NO_3^-$) concentrations. Additional variables indicate the operating state of a system, such as normal or peak operation in the event of unexpectedly high influent flow. These variables are categorical and can be assigned surrogate numerical values such as 0 = OFF and 1 = ON, depending on whether a piece of equipment is in operation. Unit processes can be designed to treat a continuous flow (e.g., disinfection plug-flow basin) or a batch (e.g., sequencing-batch reactor). Batch reactors have the additional variable of batch runtime, as contaminant transformation is time-dependent. A non-exhaustive list of WWTP variables that produce data of interest to process control are summarized in Table 1.

### 2.3. Data considerations

#### 2.3.1. Structure

Data-driven analytical methods are heavily dependent on the type of data collected. It is important to understand the unique structure and characteristics of the data used to determine how the data are organized and utilized (Cormen et al., 2009). Data at WWTP are acquired from a variety of sources: laboratory analysis, online sensor measurements, operations and maintenance management, and customer and technology manufacturer information. Each source produces data that are structured differently and can include numerical (sensor readings), categorical (ON or OFF), or unstructured (notes) variables. Differentiating between numerical or categorical variables is important for data-driven analysis. For example, an operator may determine the amount of time during a batch cycle that an air blower is ON or OFF, which dictates what a "normal" DO concentration in a reactor should be. If a controller is used, the speed of a blower can also be adjusted to meet a desired DO concentration. In this case, there is also a distinction between a controlled variable (air blower speed) and a variable that responds

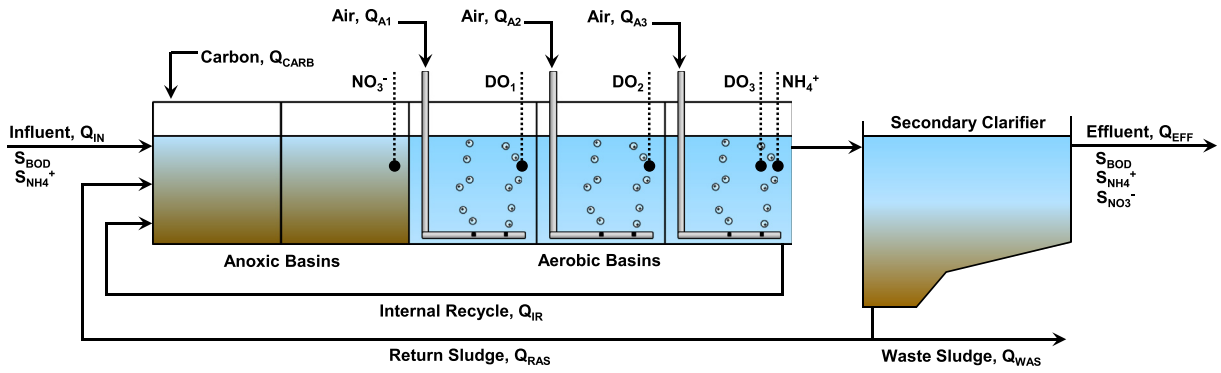



**Fig. 2.** A generic flow and sensor schematic of a CAS WWTP. $S_i$ denotes concentration of aqueous species *i* and is measured using offline laboratory analysis. $Q_i$ denotes the flowrate of either air, water, or supplemental carbon, which is measured using in-line flowmeters. DO, nitrate ($NO_3^-$), and ammonia ($NH_4^+$) are measured using sensors throughout the CAS process to evaluate treatment performance, and at some facilities nutrient measurements are used to control the rate of biosolid recycle.

**Table 1**
Examples of monitored features in WWTP and their associated data collection frequency and data structure.

| Feature | Frequency | Structure | Example |
|---|---|---|---|
| Water quality | Daily-Monthly-Quarterly | Numeric | Laboratory analysis: 5-day biochemical oxygen demand, alkalinity, nutrients |
| Water quality | Second-Minute | Numeric | Temperature, dissolved oxygen, pH, and nutrient concentrations from sensors |
| Equipment monitoring | Second-Minute | Categorical | Power state, valve position |
| Equipment monitoring | Second-Minute | Numeric | Operating speed, flowrate, pressure |
| Operating setpoints | Second-Minute | Categorical | Peak or normal operation for flow through, production or backwash for filters |
| Operating setpoints | Second-Minute | Numeric | Runtime for batch operations |

to the control (DO concentration). By measuring the effect of explanatory variables (i.e., control variables or other covariates) on response variables, data-driven methods can be developed to predict the outcome of a process. However, not all analyses differentiate between explanatory and response. Additionally, not all explanatory variables directly or measurably affect a process output, especially in a large process scheme like WWTP. Methods applied solely to explanatory variables are generally referred to as *unsupervised*, meaning that the goal is simply to identify patterns in the data without any advance knowledge of the relationships being sought. On the other hand, *supervised* learning occurs when the observations are "labelled" by their response values, and the goal is to characterize the link between the explanatory and response variables. When a distinction between variable types is required, it is mentioned in the first instance of the method in Section 3.

### 2.3.2. Frequency and temporal variability

WWTP data are collected at a variety of time intervals, from continuous online sensor measurements to quarterly laboratory results. For example, WWTP monitoring is considered "continuous" if data are collected at 15-min intervals or less (US EPA, 2015), but some effluent quality variables are measured only every few months, such as disinfection byproducts. Traditional data management segregates data by source, primarily due to the difficulty of merging data of different frequencies and formats. A common mathematical approach to handle different data frequencies is to scale data to a single time interval (Odom et al., 2018). However, datasets with a very large difference in frequencies cannot use this method because WWTP data are time-dependent, co-correlated (i.e., the relationships among variables are related to one another), and nonlinearly related, making downscaling challenging. Effluent quality variables may change either suddenly or gradually over time (Table 2), and they often change nonlinearly in relation to other process variables, which can make attributing the cause of change between sampling events difficult.

The monitoring frequency of the treatment process strongly depends on the goal of the analysis and the characteristics of the process (Venkatasubramanian, 1995). In process control, the data collection frequency should be sufficiently high to account for instrument noise and to track typical irregularities, but not so frequent that excessive computational power is required for full analysis. Short-term faults, like a clog in a pipe that can occur on the order of minutes to hours, require a different monitoring window of time than long-term faults like an increase in transmembrane pressure due to biological fouling of a membrane that occurs on the order of days to weeks (Table 2). Dürrenmatt and Gujer (2012) recommend a window width of at least three times the length of time over which the fault occurs to be detected.

**Table 2**
Examples of features with typically slow or rapid changes over time in a WWTP.

| Timescale | *Feature* |
|---|---|
| Slow (days-weeks) | Solids retention time (SRT)<br>Hydraulic retention time (HRT)<br>Transmembrane pressure (TMP) |
| Fast (seconds-minutes) | Dissolved oxygen (DO)<br>Nutrient concentrations<br>Turbidity<br>Conductivity<br>Flowrate |

### 2.3.3. Variable characteristics

Many WWTP process variables exhibit unique characteristics such as time-dependence and nonstationarity (e.g., the strong diurnal and seasonal swings of ambient temperature), but conventional control strategies rarely account for such relationships that need to be considered for data-driven fault detection, variable prediction, or automated control. Stationary variables have constant mean, variance, and covariances, making them predictable and more easily modeled. Conversely, the means and/or variances of nonstationary variables change over time. If a variable's measurements are correlated from one time step to the next, the

variable is said to be dependent over time. WWTP data exhibit these properties because of the dynamic nature of WWTP processes (Fig. 3); a constantly changing influent, batch as opposed to continuous processes, temperature, internal shifts in microbial ecology, and process control instability are a few causes of the nonstationarity and temporal dependence.

Many statistical methods assume data are normally distributed. A normal distribution is symmetric, unimodal, and bell-shaped and is characterized by two statistical parameters, its mean and variance. The multivariate case is additionally characterized by its covariances (i.e., the variance between each pair of variables). When the data are normally distributed, exact inferences can be made about the mean, variance, and covariances (e.g., confidence intervals, predictions, or hypothesis tests) because the distribution of the test statistics adhere to proven mathematical theories. When the assumption of normality is not met, it is more difficult to identify the distribution of the statistic. Without making assumptions about the data's distribution, the uncertainty in the estimate of interest cannot be accurately inferred.

The assumption of normality does not typically hold in WWTP, due to boundary limits of variables (i.e., sensor operating range), process variation, and outliers. In the event of a hardware malfunction, a contaminated lab sample, or a data entry error, observations may be missing or abnormal, compromising normality, analysis power, and reliability of results (Kwak and Kim, 2017). Each error can potentially bias features that are of interest to model, and the removal or correction of erroneous values (i.e., data cleaning) should be a high priority prior to data analysis to limit incorrect conclusions (Haimi et al., 2013; Kadlec et al., 2009).

Particular attention needs to be paid to how a data-driven methodology is implemented. In the event that nonlinear and nonstationary behavior is detected, there are two approaches for modeling nonstationary behavior: accounting for a known, or predictable, underlying trend or limiting the window of time over which a model is trained. Given the difficulty in modeling nonstationary behavior in WWTP (Fig. 3), relatively short windows of time (e.g., 3 to 10 days) may be the best option to achieve approximate stationary and normal behavior.

In addition to simple modifications of existing methods (e.g., using short training windows), distribution-free statistical methods, such as kernel density estimation (KDE) and bootstrapping, can be applied. KDE estimates a distribution using local smoothing, allowing practitioners to work around the normality assumption. However, KDE is very sensitive to the choice of tuning parameters (Izenman, 2013). Conversely, bootstrapping does not require any tuning parameters but is more computationally demanding. From a dataset, observations are randomly drawn with replacement; the statistic of interest is computed; and then these two steps are repeated many times to produce a distribution of the statistic (Efron and Tibshirani, 1994). James et al. (2013) provide a simple introduction to the bootstrap method.

### 2.4. Exploratory data analysis

Identifying the structure and characteristics of a dataset requires familiarity with the source of the data and the process itself. Plotting and visualizing data should be the first step in any analysis, but no one-size-fits-all approach exists. Observations recorded over time can be visualized in time series plots (Fig. 3); the strength of the temporal dependence can be assessed with autocorrelation function plots; potential outliers can be observed in boxplots; and the entire distribution can be plotted in a histogram. These plots work well for monitoring a single variable, but WWTP are often interested in the relationships among multiple variables. Pairwise scatterplots, multiple boxplots, functional boxplots, and cross-correlation function plots are just a few ways additional features can be assessed; examples of some of these plots can be found in Pfluger et al. (2018). There are many tests available to assess multivariate normality, including the Mardia test, Henze-Zirkler test, Royston test, Doornik-Hansen test, and the E-statistic (Korkmaz et al., 2014). Unsupervised learning methods for clustering or outlier detection are commonly used to identify structure in the data (James et al., 2013). Manual data inspection can be time-intensive, so the inclusion of more advanced statistical tools can

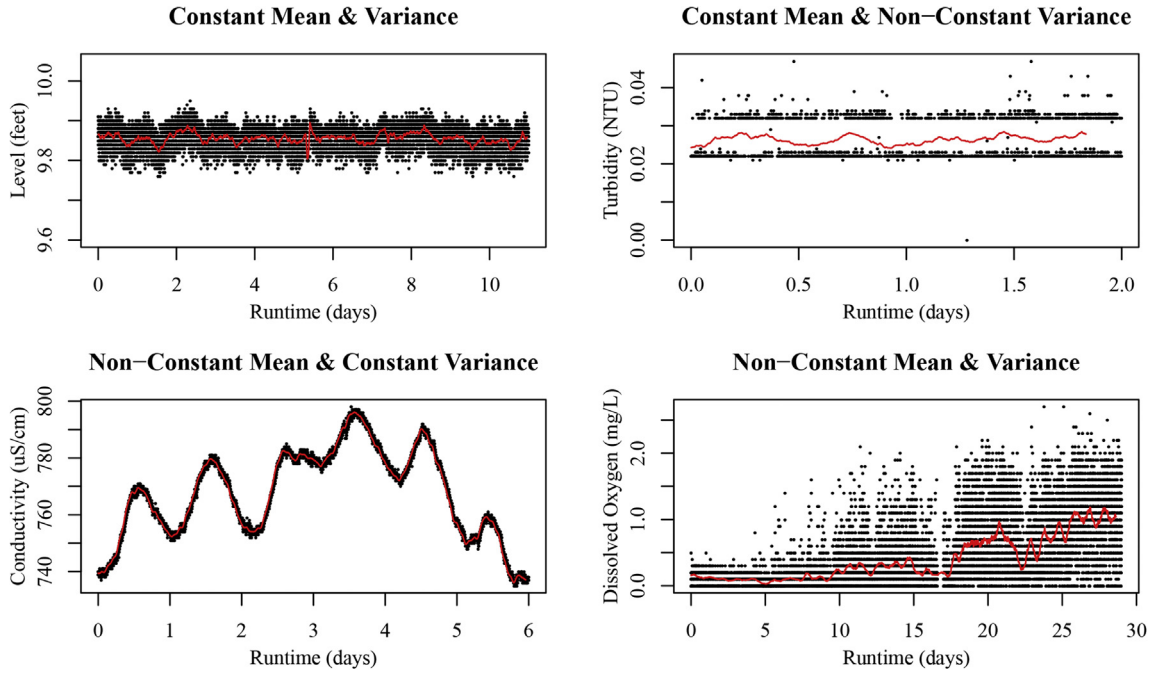



**Fig. 3.** Example of stationary and nonstationary process variables in WWTP. Membrane bioreactor (MBR) tank level (top left) is considered a stationary variable with constant mean and variance while permeate turbidity (top right), permeate conductivity (bottom left), and bioreactor (BR) DO concentration (bottom right) are nonstationary.

provide rapid insight into the data characteristics.

## 3. Methods & examples

In this section, we present the current use of process data in WWTP and a review of recent academic literature demonstrating the possibilities for advanced, data-driven process control in WWTP. The focus of this review is on control methods that have been tested on actual WWTP systems to provide practitioners with realistic examples. Simulation studies serve a valuable purpose but do not always represent the WWTP process realistically due to simplifying assumptions about data characteristics and the processes (Corominas et al., 2018). Primarily, data-driven analysis in WWTP can be used for fault detection, variable prediction, and automated control. Each requires a different and increasingly complex data processing, analysis, and control framework. The data-driven methods are therefore discussed in the context of the goal of each process control application: fault detection, variable prediction, or advanced control. We begin with a brief review of historical methods of process control in WWTP.

### 3.1. Historical process control

Data-driven process control has historically been sparse in WWTP, with daily operational decisions considered more of an art than a science (Metcalf and Eddy, 2013; O'Day, 2004). As early as the 1920s, statistical tools like histograms and control charts were used for informal diagnostics. Control during this time relied on manual adjustments and observations, as digital control was not an option prior to the 1960s. The cost of computers and instrumentation was high; treatment dynamics were not well understood; facilities were not designed with additional flexibility; and control theory was not sufficiently developed (Olsson, 2012). In the 1980s, affordable computing power facilitated simple first principle models, although their complexity and lack of reliability made them poor advisory systems (Olsson et al., 1998). By the early 2000s the digital revolution reached WWTP, and most WWTP had integrated their own version of direct digital control into process monitoring in the form of programmable logic controllers (PLC) and supervisory control and data acquisition (SCADA) systems.

Despite the unique challenges posed by WWTP data, data-driven system automation and real-time control are integral to modern WWTP operation. The most common process control practice is to maintain a set-point (i.e., target value) using online sensor readings and feedback control. For example, DO concentrations can be controlled by adjusting air blower speed (i.e., aeration intensity). Rather than operating at a single blower speed, online measurements provide feedback to the SCADA system that determines whether blower speed should increase, decrease, or stay constant relative to a measured value, like DO concentration. Chemical dosing to enhance contaminant precipitation and additional carbon for biological processes are other examples of feedback control based on *in-situ* nutrient concentrations. This method of process control is commonly achieved by a controller with a variable frequency drive to change operating conditions in a continuous, smooth, and automated manner.

Establishing target values for process variables is one of the simplest methods of control and is a widespread practice in WWTP. This single-variable monitoring paradigm is the foundation for fault detection at most modern WWTP, in which a measured value is either within or outside of an operator-specified range. While this approach has a low false-alarm rate (e.g., if a flow rate measurement is below a set-point, a fault of unknown cause has certainly occurred somewhere in the system that affects flow rate), it can be very slow to detect faults, does not forecast future values, and does not account for correlations among variables. Plant operators must be available to respond quickly to a system fault to prevent equipment damage or system failure, putting additional stress on equipment and staff to reduce a fault's impact on effluent water quality. Proactive and comprehensive approaches to fault detection and forecasting are being developed (Capizzi and Masarotto, 2017; Jiang et al., 2012; Kazor et al., 2016; Odom et al., 2018; Wang and Jiang, 2009), which could help reduce cost and improve efficiency of WWTP systems. Some data-driven fault detection methods are currently being implemented, and the results are discussed in the next section.

### 3.2. Fault detection

A multitude of system faults or changes in conditions can cause process irregularities in WWTP. These include a change in influent quality (e.g., snowmelt, industrial discharge), an outbreak of microorganisms that inhibit treatment (e.g., filamentous bacteria, algae), irregularities or damage to treatment units (e.g., membranes, clarifiers), mechanical failures (e.g., pumps, air blowers), or sensor failure (e.g., drift, bias, electrical interference). Each type of fault can alter system performance differently, and it is important to consider the versatility of an analytical approach (i.e., which types of faults can be detected) when designing a fault detection program. For example, if a sensor failure occurs and the sensor's measurements are included in a control loop, many variables could be affected. In contrast, if the sensor's measurements are not included in a control loop, a sensor fault may only affect the measured sensor variable.

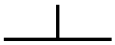
A “fault” is an unintentional deviation of a process characteristic that limits the process' ability to achieve its purpose (Isermann, 1984). Typical single-variable faults that occur in WWTP are easier to diagnose qualitatively and are illustrated in Table 3. However, multivariate faults can be much more difficult to discern visually. To detect faults in the dynamic, nonstationary, multivariate data found in WWTP, a quantitative approach, such as statistical process control (SPC), is needed. In SPC, a “fault” is identified when a consecutive series of observations are flagged as abnormal.

Olsson and Newell (1999) suggested data collection frequency be chosen to be at least one-fifth of the length of time over which the event of interest occurs. The distinction between normal (in-control or IC) and abnormal (out-of-control or OC) observations is determined by a statistical hypothesis test. Hypothesis tests are used to quantify the likelihood that an individual observation from a dataset is consistent with observations collected under IC conditions. IC observations are usually used to “train” an SPC model, which is a type of supervised learning because the initial data are known to be IC. Many SPC methods exist, but few have been implemented in WWTP. The following is a discussion of different SPC methods used in WWTP to determine if a significant change or fault has occurred.

#### 3.2.1. Control charts

Control charts are useful tools to determine, at a glance, if a process is IC. The most popular statistical control chart was outlined by Walter Shewhart of Bell Labs. The Shewhart control chart uses upper and lower control limits (UCL and LCL) for a process variable or statistic by adding or subtracting $k$ standard deviations from the variable's mean, with $k = 3$ being the industry standard (NIST/SEMATECH, 2003; Shewhart, 1926). If an observation is above the UCL (or conversely below the LCL), a statistically significant change has most likely occurred. Shewhart control charts employed at WWTP are typically constructed with a 3- or 5-day arithmetic moving average for variables designed to be stationary such as solids retention time (SRT) in a bioreactor or percent water

**Table 3**
Abnormal patterns in univariate WWTP data that could indicate a fault and potential causes of the fault pattern. Adapted from Capizzi and Masarotto (2017).

| Pattern | *Cause* |
|---|---|
| Isolated  | Power spike, air bubble on sensor, spike of contaminant in influent |
| Sustained  | Change in operational status, mechanical performance variation, sensor recalibration |
| Transient  | State change, sensor malfunction, cycle fluctuation |
| Drift  | Sensor or mechanical device degradation, biological shift, fluid flow restriction |

recovery of a membrane treatment unit. Additionally, control charts can be used in WWTP analytical labs for quality control (e.g., a sensor's measurements of a standard solution over time) (Rice et al., 2017) or other variables that change slowly (Table 2). However, the Shewhart method of calculating control limits is only valid for a variable that is normally distributed and whose observations are independent and stationary (Montgomery, 2009).

Updating the UCL and LCL to adapt to changing conditions using methods such as exponentially weighted moving average (EWMA) can account for some nonstationarity found in WWTP data (Wold, 1994). The Shewhart control chart assumes the process is stationary and weighs all past observations equally, ignoring trends (Montgomery, 2009). The EWMA gives more weight to the most recent observations, adapting to some process variation (NIST/SEMATECH, 2003) and is frequently used as a data smoothing technique (Berthouex and Box, 1996; Mina and Verde, 2007). The EWMA accounts for both the most recent observation and past behavior by multiplying the most recent observation by a forgetting factor ($0.05 \leq \lambda \leq 0.25$) and the geometric moving average by $1 - \lambda$ (Hunter, 1986; Montgomery, 2009; Roberts, 1959). However, the EWMA is not a good measure to distinguish between IC and OC for every WWTP process. Like many data-driven performance monitoring methods, the EWMA control limits are heavily impacted by outliers (Rosen et al., 2003). In both panels of Fig. 4, the assimilation of OC observations immediately widens the range of values that are considered IC. Thus, control chart limits should only be updated with IC observations, as explored by Corominas et al. (2011). Additionally, some sensors have a lower operating limit, which invalidates the standard EWMA LCL (Fig. 4a). In this case, a turbidity sensor outputs a current between 4 and 20 mA which is converted to turbidity units (NTUs) using a calibrated linear regression. Here, 4 mA correlates to 0.04 NTU. When the turbidity is below this threshold, the sensor continues to output 4 mA, which truncates the distribution of the turbidity data and invalidates the LCL. For small datasets (e.g., fewer than two variables in the case of flow and pressure of a water distribution system), univariate EWMA has shown to be better at detecting faults than multivariate EWMA (MEWMA) (Jung et al., 2013). However, most WWTP process variables violate the assumptions required for the Shewhart's or the EWMA control chart (Berthouex, 1989), resulting in a high percentage of false alarms, making them poor choices for fault detection. For larger datasets (e.g., monitoring more than 2 process variables), multivariate control charts can reduce a complicated dataset to a single measurement reflecting the "health" of the WWTP.

Monitoring multivariate processes (as opposed to individual variable monitoring) with a control chart method may provide WWTP operators with a better sense of the overall state of operating conditions (Schraa et al., 2006). Multivariate process statistics such as MEWMA (Lowry et al., 1992), multivariate cumulative sum (MCUSUM) (Crosier, 1988), and Hotelling's $T^2$ (Hotelling, 1947) can be used to examine the mean and dispersion of multiple variables but have rarely been implemented in industrial process monitoring due to the complex matrix algebra required (NIST/SEMATECH, 2003). MEWMA and MCUSUM have been shown to be good at detecting small changes in the mean, compared to Hotelling's $T^2$, but can have a high false-alarm rate (Alves et al., 2013). However, the assumption of multivariate normality is also required for these methods, and as mentioned previously, this is rarely observed in WWTP. A nonparametric approach, such as bootstrapping, may yield better results for a multivariate control chart in WWTP (Phaladiganon et al., 2011).

### 3.2.2. Principal component analysis

A widely used statistical method for monitoring multiple variables simultaneously is to capture the relationships among linear combinations of variables rather than the variables themselves by principal component analysis (PCA) (Jackson, 1991). PCA identifies independent, linear combinations of variables (principal components or PCs) by effectively calculating lines-of-best-fit through a dataset (Wise and Gallagher, 1996). PCs account for as much variation as possible (given the assumption of linearity) and can, therefore, reduce the number of model variables and eliminate noise and redundancy. For example, unsupervised PCA is frequently used to reduce the number of predictor (input) variables for multiple regression models (discussed further in Section 3.3.3) (Wallace et al., 2016; Wang et al., 2017) and can be used to identify "clusters" of related microbiological sample properties (Jałowiecki et al., 2016).

To use PCA for supervised, data-driven analysis, a "training" dataset that represents IC conditions is used to calculate the PC, then "testing" data are transformed into the model subspace (defined by the PC). If the overall distance from a new observation to the PCA-model is above a desired control limit (similar to the control chart methodology described above), then the new observation is considered abnormal and is a possible indication of a process fault. The benefit of performing PCA prior to calculating the control statistic (e.g., squared prediction error (*SPE*), Hotelling's $T^2$, etc.) is the reduction in false alarms due to the reduction in noise and removal of dependence among the features.

PCA has many applications in WWTP, from direct fault detection (King et al., 2006) to data reconstruction (Lee et al., 2006a; Schraa

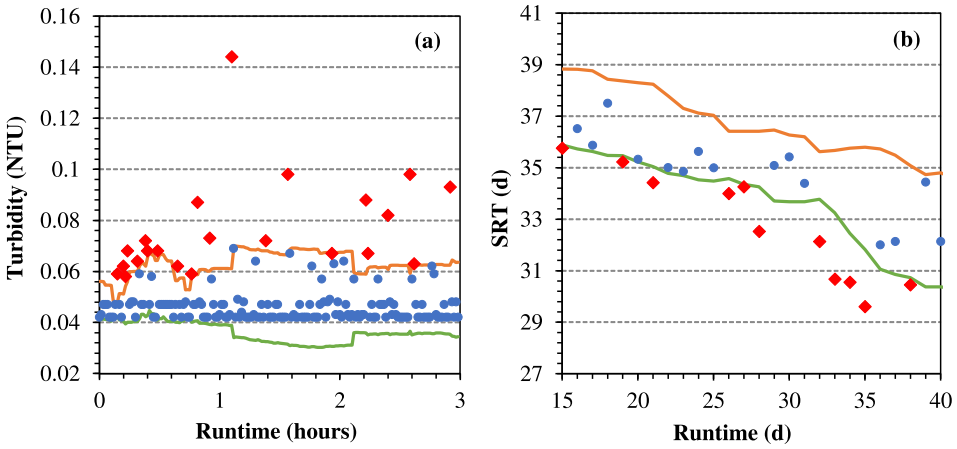



**Fig. 4.** EWMA control chart for (a) turbidity (clarity) sensor measurements of treated water and (b) SRT in an activated sldge wastewater treatment system. *Orange lines* are the EWMA UCL, *green lines* are the EWMA LCL, *blue dots* are measured values that fall within the control limits, and *red diamonds* are measured values that fall outside of the control limits. Both control limits were calculated using the previous 15 observations and a forgetting factor (λ) of 0.25. Sensor measurements for turbidity were recorded every minute, and SRT values are calculated daily. Both (a) and (b) demonstrate the power of outliers to dramatically change EWMA control limits. In (a), due to noise, natural process variation, and the range of the sensor (i.e., a sensor that communicates via a 4–20 mA current output has a lower (4 mA) and upper (20 mA) limit), many individual observations fall above the UCL, which, in this case, does not necessarily indicate a fault. In (b), the LCL detects the decline in SRT and responds to the change by widening the control range as the trend continues. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

et al., 2006). For dynamic WWTP data, variations of PCA are often used, with adaptive PCA being the most common (Baggiani and Marsili-Libelli, 2009; Kazor et al., 2016; Lee and Vanrolleghem, 2004; Rosen and Lennox, 2001). Adaptive PCA updates the model based on a "rolling window" of training observations. The training window is set to $n$ observations, and as time passes, the oldest observations are removed from the training dataset, and new observations are added to maintain a constant number of observations. The rolling training window can thereby account for temporal nonstationarity found in WWTP (i.e., conditions that change over time). However, if the training window is too large, faults could be ignored (Baggiani and Marsili-Libelli, 2009). If the training window is too small, normal observations could be flagged as faults (Rosen and Lennox, 2001). Given the type of process changes that a WWTP needs to detect, exploratory data analysis (Section 2.4) should be used to identify the shortest training window that achieves the desired true-detection rate. Some argue that the underlying correlation structure should not change with time, and therefore the rolling window concept defeats the purpose of PCA (Mina and Verde, 2007). This assumption may be valid for simulated WWTP data, but is unlikely for real WWTP data and is demonstrated by the improved performance of adaptive PCA as opposed to conventional PCA (Kazor et al., 2016).

Dynamic PCA is another common modification to PCA for fault detection in WWTP (Lee et al., 2006a; Lee et al., 2006b; Mina and Verde, 2007). The dynamic extension accounts for autocorrelation among variables by lagging observations (i.e., shifting a dataset back by a given timestep and including the lagged values as new variables) (Kruger et al., 2004; Ku et al., 1995). For most WWTP applications, a lag of a single timestep is sufficient to account for how previous conditions affect current performance. However, if the process is cyclical (i.e., processes occur as a function of runtime, and the system returns to its initial state by the end of the cycle), then the lag should be the size of the cycle itself (Kazor et al., 2016).

Cyclical (i.e., batch) operations may also require a unique modification called multiway analysis (Smilde et al., 2005). Multiway PCA unfolds a dataset indexed in three-dimensions (e.g., cycle runtime, batch, monitored variables) to a long, two-dimensional array by combining two of the three-dimensions (e.g., cycle runtime and monitored variables) that can be analyzed with traditional SPC methods like PCA (Fig. 5) (Lee and Vanrolleghem, 2004; MacGregor et al., 1994; MacGregor and Kourti, 1995; Nomikos and MacGregor, 1994; Yoo et al., 2004). In WWTP, this approach is particularly useful for sequencing-batch reactors (SBR) (Villez, 2007). The new two-dimensional dataset can account for variability in the monitored variables across batches and variables measured at different temporal frequencies.

The major drawback of PCA, and many other SPC methods like partial least squares for WWTP, is the assumption that process variables are linearly related to each other. To account for the nonlinear components of WWTP, data can first be mapped into a higher-dimensional, nonlinear space where observations are more likely to be linear (Haykin, 1999). One such nonlinear PCA method is kernel PCA (KPCA). Kernel methods avoid computationally-intensive nonlinear optimization, and different nonlinearities can be captured using different kernel functions. Popular kernels are the polynomial, Gaussian, and sigmoid, but the most commonly used is the Gaussian because its associated parameter provides precise tuning of the model fit (Izenman, 2013; Nguyen and Golinval, 2010). KPCA has shown slightly better performance in simulated WWTP (Lee et al., 2004; Xiao et al., 2017); however, Kazor et al. (2016) found limited improvement in KPCA over PCA for fault detection in a decentralized WWTP; and Lee et al. (2006c) saw similar limited improvement for the performance of anaerobic filters.

### 3.2.3. Partial least squares

Similar to PCA, partial least squares (PLS) identifies independent linear combinations of the measured variables, and outliers can be identified with $T^2$ and *SPE* statistics (Chen et al., 2016). Unlike PCA, PLS differentiates between input variables (e.g., initial water quality, operational information) and output variables (e.g., effluent water quality) and performs dimension reduction on each set of variables separately (Hoskuldsson, 1996). PLS is an example of supervised dimension reduction; PLS only monitors output variables that are affected by the input variables, whereas PCA is used to monitor all variables in the process simultaneously. If an observation is abnormal but does not impact the final water quality, PLS will not flag the process as OC, but PCA will (Chiang et al., 2001; Nomikos and MacGregor, 1995). Hence, PLS is more frequently used

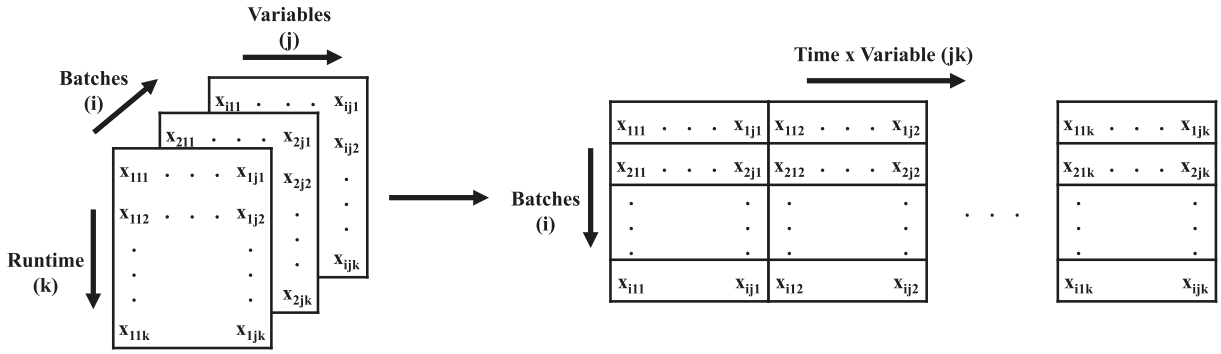



**Fig. 5.** Visual representation of how multiway PCA unfolds a three-dimensional array into a two-dimensional matrix that can be analyzed with PCA. This particular configuration is considered "batch-wise" because the length of the batch is fixed. Another multiway unfolding could be "time-wise" if runtime is the same for each batch, and batches and variables are merged.

in variable prediction than in fault detection.

In complex systems, fault detection may improve by dividing the process into units. Multiblock PLS (MB-PLS) subsets input variables into logical subsystems (e.g., primary sedimentation, aeration basin) prior to analysis (Wangen and Kowalski, 1989). Experimentally, MB-PLS does not improve prediction compared to the standard PLS; however, the results may be easier to interpret for fault diagnosis (Choi and Lee, 2005).

### 3.2.4. Neural networks

Conventional mechanistic models use complex formulas that are connected in mathematically simple ways (i.e., mass balance formulations to describe the sum of all unit processes). In contrast, neural networks (NN) use simple mathematical expressions that have complex relationships in which process inputs are nonlinearly linked to outputs without prior knowledge of an underlying mechanism (Dreyfus, 2005; Nielsen, 2015; Olsson and Newell, 1999). Process inputs and outputs are connected by "neurons" that are organized in layers (Fig. 6). A neuron in the input layer distributes an actual process input variable to neurons in the first hidden layer. Neurons in a hidden layer normalize and weigh multiple inputs, transform the value with an activation function, and produce a normalized output signal. NN can have one or multiple hidden layers, connected by input transformations and output signals. The output layer is a weighted sum of the final hidden layer's output signal.

In contrast to linear statistical models (e.g., multiple regression, PCA, PLS), NN model parameters do not have the same interpretability (i.e., no physical, chemical, or biological significance). To identify the parameters of each neuron in a NN model, learning algorithms are needed. Due to the extensive intricacies of the different learning algorithms for NN development, this paper will focus on the two types of NN training: supervised or unsupervised.

Supervised training requires data to be labelled in such a way that inputs and outputs are defined. There are many different learning algorithms for fitting a supervised NN, and each requires an iterative process in which parameters are estimated based on a large historical dataset (Dreyfus, 2005). One of the most common is the back-propagation learning algorithm that starts by randomly assigning parameter values, calculating an estimated output, and minimizing the error between the estimated and the actual output, node by node and layer by layer, starting with the hidden layer directly connected to the output. Hundreds of iterations must be performed to determine the best network for a particular dataset, requiring a substantial amount of computing power (Wei, 2013). Hence, it is valuable to minimize the number of variables used to construct the model.

Unsupervised training of a NN uses data that are unlabeled (i.e., the model is supplied with defined inputs but no outputs) and uses fundamentally different learning algorithms than the error-correction method in supervised training. Unsupervised NN act best as classifiers for pattern recognition. In fault detection, unsupervised NN can be trained to model a process by estimating the values of inputs and comparing the estimation to the actual values, also known as an auto-encoder NN (ANN). Xiao et al. (2017) used ANNs with "bottleneck" layers (i.e., the middle, hidden layer contains fewer nodes than the preceding or succeeding layers), which force the NN to effectively capture the principal components of the data to detect faults at a WWTP. *SPE* was calculated from the difference between actual and estimated values, and similar to PCA, an *SPE* threshold was calculated to determine if the process was IC or OC. Xiao et al. (2017) concluded that ANN-based fault detection was more sensitive to changes than conventional PCA. Additionally, Xiao et al. (2017) compared "deep" and "shallow" ANN. Until recently, training NN with many layers and nodes ("deep" NN) was not computationally efficient (Hinton et al., 2006). However, "shallow" NN are generally unable to capture highly-nonlinear systems. In this case, there was no conclusive evidence that the deep ANN performed better than the shallow ANN.

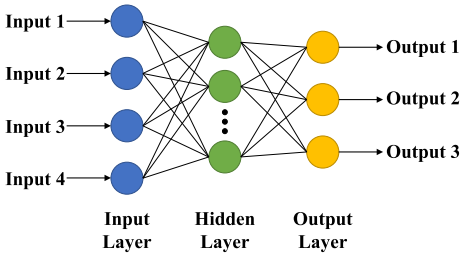



**Fig. 6.** A generic neural network structure. NN can contain multiple hidden layers with different numbers of neurons in each layer, and one hidden layer is shown here.

### 3.3. Variable prediction

SPC can be used to assess if a system is IC or OC, but this is a generic measure of product water quality and system health. To predict what a variable value should be under given conditions, model-based control could be used. Model predictive control (MPC) compares mechanistic model predictions to actual process measurements. Then, deviations from the model are identified as faults. The model can be derived from theory (i.e., fluid dynamics, microbial kinetics) or empirical trends (i.e., data-driven) and can be used to approximate additional process variables. In lieu of directly

monitoring the variable of interest, a link may be found among variables. In this way, a software sensor or "soft sensor" (also referred to as inferential sensors, virtual online analyzers, or observer-based sensors) can be developed for the online monitoring of variables that are too time-consuming or expensive to consistently monitor with lab-based analyses (Chéruy, 1997; Kadlec et al., 2009). Many different approaches to variable prediction have been proposed, and here we review the most commonly observed in literature.

#### 3.3.1. Activated sludge models

Some water quality variables can be adequately predicted with calibrated, first-principal models. The activated sludge model no. 1 (ASM1) is the most widely used deterministic model for biological carbon and nitrogen removal in WWTP (Henze et al., 1987). ASM1 was developed in 1985 by compiling novel research about the kinetic behavior and mechanisms of the CAS process. Subsequent CAS models (e.g., ASM2 and ASM2d) have additional parameters that account for fermentation, enhanced biological phosphorus removal, and chemical phosphorus removal (Gujer et al., 1999; Henze et al., 1999, 1995). Models based on first principals also exist for clarifiers/settlers, but due to a lack of a mathematical relationship between floc characteristics and settleability, they are still limited to process design and research purposes (Metcalf and Eddy, 2013; Olsson, 2012). Uncertainty in the model inputs and the simplified mathematical framework fundamentally limits the accuracy of the model. However, in many cases, pilot- or full-scale calibration can account for some of the error in the ASM (Metcalf and Eddy, 2013). Additional sources of error are model parameter estimates. Not all unit processes share a common set of state variables (i.e., variables that indicate the operating conditions of a process as opposed to variables that measure a constituent in the water), and linking models with variable estimates can lead to substantial error in the plant-wide model (Volcke et al., 2006). For example, ASM and secondary clarifier models are frequently coupled, but the models differ in how total suspended solids (TSS) concentration is calculated and incorporated.

Without calibration and only parameter estimates, the ASM may produce results that are only accurate within an order of magnitude, illustrating the variability of WWTP (Gujer, 2011). Olsson and Newell (1999) considered other sources of error for model predictions, including inaccurate or incorrect calibration, non-ideal process behavior, and lump-sum parameter assumptions. For these reasons, MPC with the ASM is rarely used in full-scale biological WWTP operations for variable prediction or fault detection.

To standardize control strategy testing at biological WWTP, a simulation benchmark was developed from the ASM1 (Henze et al., 1987) under EU COST Actions 682 and 624 (Alex et al., 1999; Jeppsson and Pons, 2004; Spanjers et al., 1998). The Benchmark Simulation Model No.1 (BSM1) includes a mathematical model of a five-reactor CAS treatment system followed by a clarifier (Fig. 2), ASM-specific parameters from literature, and simulated influent datasets for different weather events (Copp, 2002). The newest version, BSM2, incorporates extensions proposed in recent literature: a longer simulation study timeframe, inclusion of temporally dynamic parameters, and more realistic sensor behavior and failure (Jeppsson et al., 2007; Nopens et al., 2010; Rosen et al., 2004).

There are hundreds of proposed control strategies for the BSM, but simulation benchmarks do not yet exist for all common wastewater treatment technologies. While simulation studies are important to understand the potential behaviors of control strategies, the actual dynamics of a WWTP are nearly impossible to reproduce artificially. This is most evident when control strategies perform well on BSM but cannot be replicated with real WWTP data (Sin et al., 2006). Oppong et al. (2013) compared simulated datasets from the BSM models to real industrial WWTP data using anaerobic digestion in an attempt to develop a soft sensor. However, the variable of interest (volatile solids concentration) had substantially different co-correlation structure, both in magnitude and direction, among the simulated and actual process variables. The difference was attributed to infrequent sampling and a stable process with little change, but it is also possible that the BSM is inadequate for MPC in this case.

#### 3.3.2. Transfer function models

Transfer function models are a general class of models that describe the relationship between an input and output of a linear system using a mathematical function. When the system is not too complex (i.e., the number of output parameters is $\leq 2$), transfer functions can be a good approximation for dynamic systems (Box et al., 1994). Univariate autoregressive integrated moving average (ARIMA) models are a special case of transfer function models that do not depend on the input variables and are widely used for linear time series forecasting (Chen et al., 2007). The autoregressive (AR) portion predicts values that are mathematically related to the previous time-step(s) (i.e., lag). The moving average (MA) predicts values that are mathematically related to the error of the past time-steps. The integrated (I) portion of the ARIMA model indicates that the difference between observations (one or more) is modeled instead of the observation itself, and this step can remove some of the nonstationarity present in the data.

The primary application of ARIMA models in WWTP is to predict an effluent variable. Park and Koo (2015) showed that an ARIMA model can be used to predict effluent turbidity of a sedimentation basin. Berthouex and Box (1996) and West et al. (2002) successfully used an ARIMA model to predict effluent 5-day biochemical oxygen demand ($BOD_5$) of a WWTP. However, the ARIMA model's integration step is often insufficient to account for WWTP data non-stationarity in the long-term (West et al., 2002). As the prediction horizon increases, the accuracy of ARIMA models declines substantially; unable to account for nonlinear behavior in WWTP (Dellana and West, 2009).

#### 3.3.3. Multiple regression

Multiple regression is an extension of a simple, linear regression, using multiple independent variables ($X_i$, $i = 1, 2, \ldots, k$) to model a single dependent variable, $Y$, via $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + e$. The model parameters are commonly estimated by ordinary least squares, and more information about multiple regression model fitting can be found in Sheather (2009). When all independent variables are standardized (i.e., zero mean and unit variance), the strength of an input variable's impact on the output variable is directly proportional to the magnitude of $\beta_i$, giving tangible meaning to the model parameters. Categorical information can also be integrated by the use of dummy variables ($D_i$, $i = 1, 2, \ldots, k$), which take on binary values (e.g., if a blower is ON, then $D_i = 1$; if OFF, then $D_i = 0$). However, problems can arise when fitting a multiple regression model if the explanatory variables are exactly linearly related (multicollinearity), are highly variable, or are autocorrelated (Miah, 2016) as in WWTP.

Ebrahimi et al. (2017) used multiple regression models to predict various water quality variables like phosphorus, nitrogen, and TSS concentrations in a full-scale wastewater treatment plant with reasonable results ($r^2 = 0.71–0.87$, meaning the model explains 71–87% of total variation in the dependent variable), but they did not demonstrate sufficient accuracy for stand-alone fault detection. However, multiple regression techniques may have applications in soft sensor and empirical model development.

PLS regression is a combination of PLS and multiple regression and can be used to predict the output variables from the input

(Abdi, 2003). Because of this property, PLS is commonly used for industrial soft sensors and water quality variables in WWTP, such as chemical oxygen demand (COD), TSS, nitrate, and oil and grease concentrations (Langergraber et al., 2003; Qin et al., 2012). The use of nonlinear mapping with PLS (kernel PLS or KPLS) also shows promise for methane production from a full-scale anaerobic filter (Lee et al., 2006c) and COD, total nitrogen, and cyanide concentration in CAS (Woo et al., 2009). In these cases, KPLS performed better than conventional PLS, which demonstrates the importance of accounting for nonlinearities for some data-driven applications.

#### 3.3.4. Neural networks

Supervised NN have been used to predict raw wastewater flow from online rainfall data and historical influent data (Wei, 2013). Yang et al. (2008) reconstructed COD concentration from UV-254 and pH measurements using a back-propagation NN. Ömer Faruk (2009) used a hybrid NN-ARIMA to predict boron and DO concentrations and water temperature of a river over time. While the ARIMA model performed very poorly ($r^2 = 0.23–0.55$), the hybrid model performed only slightly better ($r^2 = 0.79–0.83$) than the NN model ($r^2 = 0.77–0.81$). NN-ARIMA hybridization has been a popular research topic because, in theory, both linear and nonlinear behavior could be described with the resulting model (Venkatasubramanian, 1995). However, few case studies exist that demonstrate a significant improvement of a hybrid model over a NN model (Chen et al., 2007).

A NN hybrid model was successfully used by Lee et al. (2008) to predict COD, total nitrogen, and total phosphorus concentrations in the effluent of small CAS WWTP from conductivity, temperature, pH, DO, oxidation-reduction potential (ORP), and turbidity. Input and output variables were lagged to account for dynamic process variation, combined with a transfer function model (auto-regressive representations with exogenous inputs or ARX). Lee et al.'s (2008) approach showed good results for variable prediction ($r^2 = 0.92–0.95$) and is promising for soft sensor development.

Lee et al. (2002) also used a hybrid NN structure for prediction of WWTP effluent quality. In principle, the hybridization of mechanistic and NN models bridges the gap between first principal and statistical approaches. The NN was placed in parallel and in series with the ASM1 model to estimate the model error or input parameters, respectively. The parallel hybrid NN model performed well at predicting effluent variables (e.g., cyanide $r^2 = 0.93–0.96$), but the series hybrid NN model did not perform better than the NN alone, indicating that there exists some error for which the ASM1 model itself does not account.

Models to determine sorption kinetics and the capacity of carbon to adsorb contaminants can also be mapped by NN. Vasanth Kumar et al. (2008) trained an NN model with batch experimental data under various conditions to predict equilibrium concentrations after the uptake of dye by powdered activated carbon (PAC). The resulting predictions were nearly perfect ($r^2 = 0.96$). However, hundreds of data points were needed to calibrate the model prior to the predictions; there was not significant variation among the input variables; only a single contaminant was used; and separate testing data was not used to verify results. Given the complexity of biological treatment modeling, generating a carbon adsorption isotherm for PAC treatment is very well understood, computationally straightforward, and accurate for design purposes. Unless NN can demonstrate the ability to account for large variations in initial water quality (of which the current isotherm paradigm cannot), use of the traditional adsorption isotherm models will continue to be used. A potential application that has not yet been explored is for the generation of isotherm models for micropollutants (e.g., per- and polyfluoroalkyl substances) in the presence of bulk organic carbon.

### 3.4. Advanced control

The goal of WWTP optimization is to achieve the desired effluent quality with fewer inputs (i.e., chemicals, energy, manpower). Future WWTP will also need to be able to adjust their effluent quality to meet new demands without risk of disturbance. As water resources dwindle and demand increases in urban centers, customizable water quality based on need and time of year (known as “tailored” water) has become an attractive option (Vuono et al., 2013). Using historical data and system knowledge, a function can be developed to minimize cost or energy while maintaining effluent quality in order to identify the best set of setpoints and control decisions. This is a fundamentally different approach than heuristically adjusting variable setpoints and observing the system's response. Various methods to achieve data-driven control (“advanced or automated control”) are discussed in this section. However, advanced control is still in its infancy for WWTP, and few full-scale demonstrations or installations exist.

#### 3.4.1. Model predictive control

MPC uses a mechanistic model of a process to predict a process variable accounting for the physical constraints of a system's actual process variable measurements (Richalet et al., 1978). The model predicts future process behavior over a time interval (known as the prediction horizon), and predictions are compared to online measurements to determine if a change has occurred (Fig. 7). MPC is less common in WWTP because most individual WWTP processes, especially biological processes, are too complex to develop sufficiently accurate first principal models (Section 3.3.1) for advanced control purposes due to their deviation from ideal, steady-state conditions (Patton et al., 2000). Furthermore, the computing power required to handle the nonlinearities has not been well documented.

MPC in WWTP takes on many forms, but all must address WWTP data's nonlinear behavior. Nonlinear models are computationally intensive to solve, and accounting for too many nonlinearities can substantially slow a controller's response. A less computationally intensive option is to use piecewise linear MPC in which multiple linear models approximate a nonlinear model (Ocampo-Martinez, 2010; Olsson and Newell, 1999). Another method is to update or adapt linear model parameters to fit current operating conditions (Zhang and Zhang, 2006). Adaptive MPC controllers have been shown to perform better than conventional PID controllers for nonlinear processes; however, strong nonlinearities are still better handled by alternative control approaches such as NN (Hermansson and Syafiie, 2015).

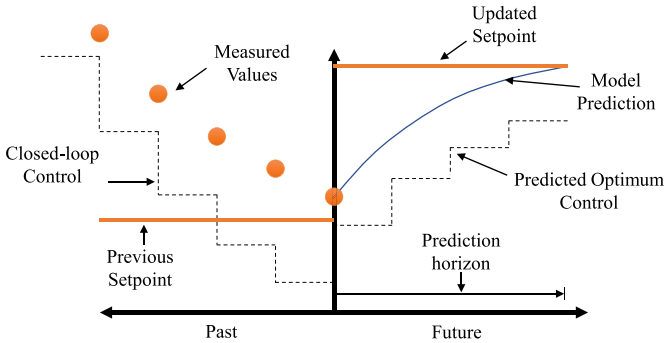



**Fig. 7.** Simplified example of MPC. At time = 0 (the intersection of the axes), the measured variable's setpoint is increased. The mechanistic model and a function describing the energy consumption of the control mechanism are used to identify the optimum control response to reach the new setpoint over a given time interval (i.e., prediction horizon).

MPC has been implemented for dynamically simplistic WWTP unit processes, such as membrane-based treatment, that can be controlled by a single variable. Membrane systems can be easily modeled using known relationships of fluid flow, mass transfer, and thermodynamics. Bartman et al. (2009) derived such a model to control a valve on a pilot-scale reverse osmosis (RO) membrane treatment system. The reject (concentrate) flowrate was controlled by the dynamic, nonlinear, lumped-parameter model and was validated with experimental data. In general, the system performed better when controlled by the dynamic model as opposed to a traditional controller.

Not all WWTP unit processes are fit for MPC simply because accurate analytical models do not yet exist, and the number of possible inputs makes real-time control computationally unreasonable. Attempts have been made in the literature to adapt MPC for WWTP, including CAS and membrane systems, but more research is needed to develop realistic and system-specific models before MPC can be implemented as a control strategy in full-scale WWTP. Alternatively, non-deterministic, nonlinear, data-driven models are an option for MPC of activated sludge systems (i.e., NN).

### 3.4.2. Neural networks

As discussed in Section 3.3.4, each parameter and layer in an NN model adds an additional degree of flexibility that can address the problem of nonlinear model fitting. However, a large number of NN model parameters can risk overfitting to noise in the data rather than the process itself and can unnecessarily increase computation. The computational power required to use an NN model for control applications is not well documented, and most studies in WWTP literature utilize only a few water quality variables to predict a single value (i.e., soft sensor development). The availability of reliable and plentiful online sensor data can also be a major constraint. More research is needed with constructing larger WWTP NN before the practicality of NN control strategies in WWTP can be assessed. To begin, WWTP NN model development should be performed incrementally, so that an unexpected and unmanageable amount of computational power is not required to achieve simple goals.

One proposed method of using NN in nonlinear dynamic process control is to adjust the NN structure (i.e., number of hidden neurons) and parameters (i.e., node weights) during the training phase, also referred to as an unsupervised, self-organizing NN. At each node, an optimization function determines if the node should be deleted, kept the same, or split into two, and the node parameters are adjusted accordingly. Post-training, self-organizing NN have been shown to perform better (i.e., lower computation time and testing error) than NN where the structure is fixed (i.e., number of nodes) (Han et al., 2010). Han and Qiao (2014) used a self-organizing NN to model aeration and recirculation (i.e., DO and nitrate concentration) and a multiple-objective controller to optimize control of a pilot-scale CAS system. However, the authors did not compare system performance to a conventional controller, making it difficult to justify implementation for the purpose of DO and nitrate concentration control, given computational requirements for real-time control of a larger system.

NN controllers are also being designed to detect nonlinear time-varying data features that indicate the end of a reaction, such as ORP in CAS (Fig. 8). Luccarini et al. (2010) used an NN program to control and optimize biological nitrogen removal for a pilot-scale SBR. The end of denitrification (i.e., a biological process to remove nitrate) could not be detected well due to a 50% historical completion rate at the pilot facility. The end of nitrification is difficult to detect because of noise and the small change in the rising ORP and DO. The lack of detection, in this case, demonstrates a common drawback of many data-driven systems—the desired performance must be demonstrated consistently.

### 3.4.3. Transfer function models

Transfer function models can be used for MPC and optimization in addition to variable prediction discussed in Section 3.3.2. O'Brian et al. (2011) demonstrated the ability of a first-order, six variable ARX MPC to optimize energy consumption by reducing aeration by 25% at a full-scale CAS WWTP, compared to the facility's original PLC-based control strategy. However, in this case, providing

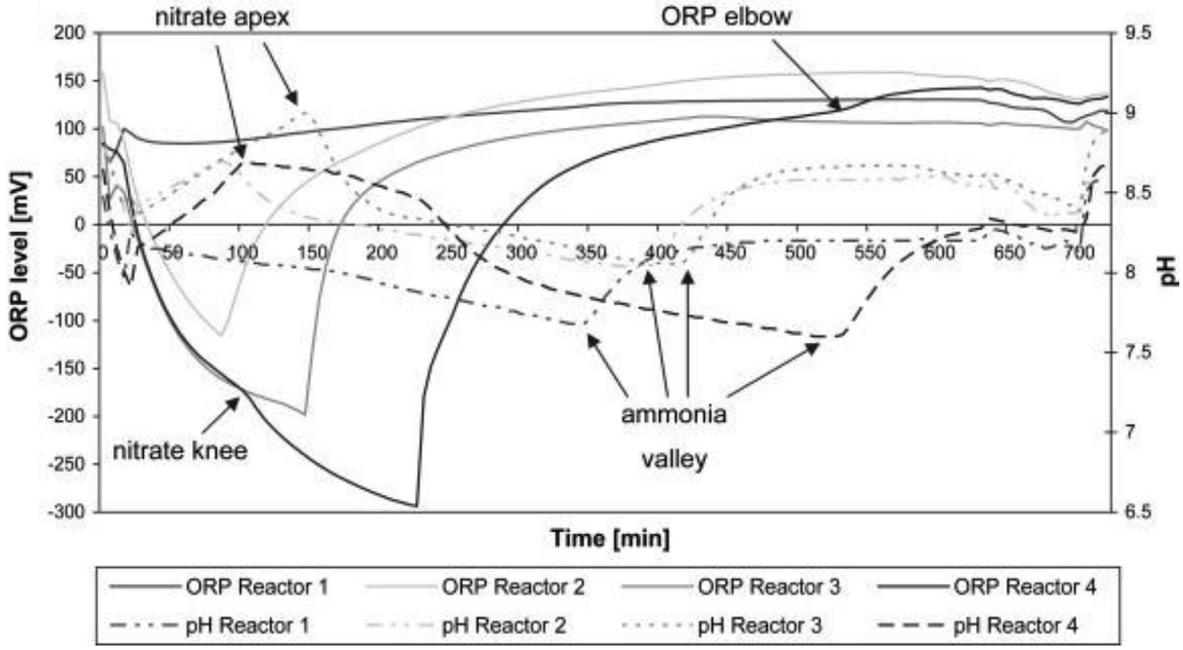



**Fig. 8.** Illustration of repeating patterns of ORP and pH data in parallel reactors of a batch activated sludge system. The ORP elbow, nitrate apex and knee, and ammonia valley indicate the completion of different stages in biological nitrogen removal. Adopted from Dubber and Gray (2011).

aeration based on influent organic loading is not a novel concept, and much of the improvement can be attributed to a poorly calibrated or performing controller.

*3.4.4. Fuzzy logic*

In diagnosing process upsets, multiple WWTP operators may logically reach different conclusions regarding the cause of a problem. Unlike computers, human decision-making is not always logical, and choices are not always binary (i.e., true or false). Fuzzy logic mimics the attributes of human reasoning by "blurring" the inputs and rules to allow for "partial" truth. To achieve this, fuzzy logic uses linguistic variables in place of numerical variables, defines relationships among variables ("clustering") with IF-THEN statements that allow for different degrees of truth, and characterizes the relationships by fuzzy algorithms. The seminal paper describing fuzzy logic by Zadeh (1973) is recommended for readers interested in more details on fuzzy model structure.

The classic rule development approach for fuzzy models is to write IF-THEN relationships explicitly, which is time-consuming for both computer scientists and WWTP operators. This process can be simplified by using an NN to map operator observations into fuzzy rules. Enbutsu et al. (1993) re-structured the traditional NN model with fuzzy neurons in the input and output layers to model PAC dosing and established rules that were more accurate than those derived from interviews with water treatment operators.

Taw-Hwan et al. (1997) also used a hybrid fuzzy model-NN approach, but calculated PAC-dosing rate with a fuzzy model under normal conditions and used an NN model when abnormalities were detected (i.e., turbidity >30 NTU). Jar-tests were used to collect data to build both the fuzzy model and NN model, which predicted PAC-dosing rate very accurately during a one-year field test at a full-scale water treatment plant.

Yoo et al. (2003) used PCA combined with fuzzy clustering and a fuzzy model to predict COD removal from a full-scale industrial WWTP. PCA was used to reduce the complexity of the fuzzy model as well as to reduce co-linearity. Results were able to generally predict COD removal but could not be used for direct control, demonstrating that not all combinations of data-driven solutions are always effective.

Bello et al. (2014) used fuzzy clustering to define rules for a multi-input, multi-output coagulant-dosing system in a water treatment plant. The pH predictions were calculated from previous pH values and flowrates of three coagulants and coagulant-aids. In this case, fuzzy MPC performed slightly better than a nonlinear model approximated by linearization. If conditions were to change substantially over time (i.e., if the model parameters needed to adapt over time), fuzzy models may be a practical alternative to nonlinear MPC.

## 4. Conclusions and recommendations

The future of data-driven and big data analytics in WWTP (and water treatment) is in improving process control to reduce energy demand, ensure effluent water quality, and prevent system failure. To achieve this, WWTP need to incorporate data-driven fault detection, variable prediction, and automation into their current process control paradigms. Despite a large body of literature on many data-driven process control methods in WWTP, there is no consensus on a singular "best" approach. For fault detection and diagnosis, WWTP need to understand past and present behavior in terms of IC or OC." Control charts are good for monitoring single variables that are measured daily to monthly and do not contain a lot of noise (e.g., laboratory analysis, SRT). To evaluate multiple variables for fault detection and diagnosis, PCA is good for use with composite samples because it does not distinguish between input and output variables like PLS. To use big data in wastewater treatment to predict future performance, the monitored variable(s) must have a high sample frequency and number of historical observations, but do not need to be linearly related or parametrically distributed.

A small, decentralized facility may experience so much operational variability that MPC is not effective; thus, SPC may be implemented to detect faults and reduce the number or length of time on-site operations staff must be present. For a large, centralized facility with the buffering capacity to operate at quasi-steady-state (compared to the decentralized case), MPC may be useful for reducing chemical inputs and energy optimization. An additional consideration is the development of tools to optimize operations at the unit process scale (e.g., aeration basin with recycling, membrane bioreactor) in addition to the plant-wide scale. Different tasks will employ different problem-solving methodologies. Models that reveal more mechanistic information to assist with diagnosis tend to have poor fault detection accuracy in highly nonlinear systems. Hence, a hybrid method combining model and statistical process control may be a superior problem-solving approach. Each approach will come at a computational cost, which is rarely reported in the literature; with the major limiting factor being the quality and quantity of data generated by WWTP.

To develop high-quality and accurate big data tools for wastewater treatment industry data scientists, computer scientists, and engineers must continue to collaborate to maximize data's potential. The effectiveness of many state-of-the-art data science tools have not yet been tested in WWTP. Random forests, support vector machines, and reinforcement learning have the potential to accommodate many of the features of WWTP data, but they still require large training datasets to fit and must produce reliable results (Kusiak et al., 2013; Verma et al., 2013). With WWTP operations, transparency in methodology is one of the keys to adoption, so some advanced methodologies may continue to be eschewed in favor of simpler but interpretable methodologies.

In summary, WWTP looking to integrate data-driven control should:

1. Define the scope of the problem and desired goals.
2. Identify which variables are currently being monitored (or should be monitored) to effectively capture the scope of the problem.
3. Use plotting tools to investigate the characteristics of each variable as well as the relationships between variables.
4. Based on the features observed in the data and analysis goals, identify the appropriate method to implement. Recommendations for further reading on each broad category of methods are given throughout the text.
5. Fit the models and assess their validity. Visualize results to ensure that the conclusions are logical and realistic.
6. Share the results with other WWTP via industry-specific publications and conferences to develop mainstream, data-driven process control tools for WWTP.

## Acknowledgments

Support for this study was provided by the National Science Foundation Partnership for Innovation: Building Innovation Capacity project 1632227 and by the National Science Foundation Engineering Research Center program under cooperative agreement EEC-1028968 (ReNUWIt). The authors would also like to thank Dr. Tanja Rauch-Williams and Dr. Eric Dickenson for their comments on earlier drafts. The anonymous comments from three reviewers also greatly contributed to framing the content of this review.

## Acronyms

| | |
|---|---|
| ANN | Auto-Encoder NN |
| AR | Autoregressive |
| ARIMA | Autoregressive integrated moving average |
| ARX | Auto-regressive representations with exogenous inputs |
| ASM | Activated sludge model |
| $BOD_5$ | 5-day biochemical oxygen demand |
| BSM | Benchmark simulation model |
| CAS | Conventional activated sludge |
| COD | Chemical oxygen demand |
| DO | Dissolved oxygen |
| EWMA | Exponentially weighted moving average |
| IC | In-control |
| KDE | Kernel Density Estimation |
| KPCA | Kernel PCA |
| LCL | Lower control limits |
| MCUSUM | Multivariate Cumulative Sum |
| MA | Moving average |
| MB-PLS | Multiblock PLS |
| MEWMA | Multivariate EWMA |
| MPC | Model predictive control |
| $NH_4^+$ | Ammonia |
| NN | Neural networks |
| $NO_3^-$ | Nitrate |
| NTU | Nephelometric Turbidity Unit |
| OC | Out-of-control |
| ORP | Oxidation-reduction potential |
| PAC | Powder activated carbon |
| PCA | Principal component analysis |
| PLC | Programmable logic controllers |
| PLS | Partial least squares |
| SBR | Sequencing batch reactors |
| SCADA | Supervisory control and data acquisition |
| SPC | Statistical process control |
| *SPE* | Squared prediction error |
| SRT | Solids retention time |
| TSS | Total suspended solids |
| UCL | Upper control limits |
| WWTP | Wastewater treatment plant |

## References


Abdi, H., 2003. Partial least squares (PLS) regression. In: Lewis-Beck, M., Brayman, A., Liao, T.F. (Eds.), Encyclopedia of Social Sciences Research Methods. Sage, Thousand Oaks, CA.

Alex, J., Beteau, J.F., Copp, J.B., Hellinga, C., Jeppsson, U., Marsili-Libelli, S., Pons, M.N., Spanjers, H., Vanhooren, H., 1999. Benchmark for Evaluating Control Strategies in Wastewater Treatment Plants. European Union Control Association.

Alves, C., Henning, E., Samohyl, R., Cristina Konrath, A., Formigoni Carvalho Walter, O., 2013. Application of multivariate control charts for monitoring an industrial process. Tecno-Lóg. 17, 101–107. https://doi.org/10.17058/tecnolog.v17i2.3636.

Baggiani, F., Marsili-Libelli, S., 2009. Real-time fault detection and isolation in biological wastewater treatment plants. Water Sci. Technol. 60. https://doi.org/10.2166/wst.2009.723.

Bartman, A.R., Christofides, P.D., Cohen, Y., 2009. Nonlinear model-based control of an experimental reverse-osmosis water desalination system. Ind. Eng. Chem. Res. 48, 6126–6136. https://doi.org/10.1021/ie900322x.

Bello, O., Hamam, Y., Djouani, K., 2014. Fuzzy dynamic modelling and predictive control of a coagulation chemical dosing unit for water treatment plants. J. Electr. Syst. Inf. Technol. 1, 129–143. https://doi.org/10.1016/j.jesit.2014.08.001.

Berthouex, P.M., 1989. Constructing control charts for wastewater treatment plant operation. J. Water Pollut. Control Fed. 61, 1534–1551.

Berthouex, P.M., Box, G.E., 1996. Time series models for forecasting wastewater treatment plant performance. Water Res. 30, 1865–1875. https://doi.org/10.1016/0043-1354(96)00063-2.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. Time Series Analysis - Forecasting and Control, fourth ed. Wiley.

Capizzi, G., Masarotto, G., 2017. Phase I distribution-free analysis of multivariate data. Technometrics 59, 484–495. https://doi.org/10.1080/00401706.2016.1272494.

Chen, A., Zhou, H., An, Y., Sun, W., 2016. PCA and PLS monitoring approaches for fault detection of wastewater treatment process. In: Presented at the 2016 IEEE 25th International Symposium on Industrial Electronics (ISIE). https://doi.org/10.1109/ISIE.2016.7745032.

Chen, S.-H., Wang, P.P., Kuo, T.-W., 2007. Computational Intelligence in Economics and Finance. Springer.

Chéruy, A., 1997. Software sensors in bioprocess engineering. J. Biotechnol. 52, 193–199. https://doi.org/10.1016/S0168-1656(96)01644-6.

Chiang, L.H., Russell, E.L., Braatz, R.D., 2001. Fault Detection and Diagnosis in Industrial Systems.

Choi, S.W., Lee, I.-B., 2005. Multiblock PLS-based localized process diagnosis. J. Process Control 15, 295–306. In: https://doi.org/10.1016/j.jprocont.2004.06.010.

Clean Water Act, 1977.

Copp, J.B., 2002. The COST Simulation Benchmark: Description and Simulator Manual. European Cooperation in the field of Scientific and Technical Research (COST).

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to Algorithms, third ed. The MIT Press.

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. Environ. Model. Softw., Special Issue on Environmental Data Science. Applications to Air quality and Water cycle 106, 89–103. https://doi.org/10.1016/j.envsoft.2017.11.023.

Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C., Vanrolleghem, P.A., 2011. Performance evaluation of fault detection methods for wastewater treatment processes. Biotechnol. Bioeng. 108, 333–344.

Crosier, R.B., 1988. Multivariate generalizations of cumulative sum quality-control schemes. Technometrics 30, 291–303. https://doi.org/10.1080/00401706.1988.10488402.

Dellana, S.A., West, D., 2009. Predictive modeling for wastewater applications: linear and nonlinear approaches. Environ. Model. Softw 24, 96–106. https://doi.org/10.1016/j.envsoft.2008.06.002.

Diebold, F.X., 2003. Big data" dynamic factor models for macroeconomic measurement and forecasting. In: Dewatripont, M., Hansen, L.P., Turnovsky, S. (Eds.), Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society. Cambridge University Press, Cambridge, pp. 115–122.

Dreyfus, G., 2005. Neural networks: an overview. In: Neural Netw. SE - 1 1–83. https://doi.org/10.1007/3-540-28847-3_1.

Dubber, D., Gray, N.F., 2011. The effect of anoxia and anaerobia on ciliate community in biological nutrient removal systems using laboratory-scale sequencing batch reactors (SBRs). Water Res. 45, 2213–2226. https://doi.org/10.1016/j.watres.2011.01.015.

Dürrenmatt, D.J., Gujer, W., 2012. Data-driven modeling approaches to support wastewater treatment plant operation. Environ. Model. Softw 30, 47–56. https://doi.org/10.1016/j.envsoft.2011.11.007.

Ebrahimi, M., Gerber, E.L., Rockaway, T.D., 2017. Temporal performance assessment of wastewater treatment plants by using multivariate statistical analysis. J. Environ. Manag. 193, 234–246. https://doi.org/10.1016/j.jenvman.2017.02.027.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC Press.

Enbutsu, I., Baba, K., Hara, N., Waseda, K., Nogitat, S., Hitachi-shi Ibaraki, O., Japan, ken, Japan, T., 1993. Integration of multi AI paradigms for intelligent operation support systems – fuzzy rule extraction from a neural network. Water Sci. Technol. 28, 333–340.

Gandomi, A., Haider, M., 2015. Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manag. 35, 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Golhar, S., Dallas, C., 2016. Big Plants Produce Big Data that They Do Not Know of…. Presented at the WEFTEC. Water Environment Federation, pp. 2065–2077.

Gujer, W., 2011. Is modeling of biological wastewater treatment a mature technology? Water Sci. Technol. 63, 1739–1743. https://doi.org/10.2166/wst.2011.323.

Gujer, W., Henze, M., Mino, T., Van Loosdrecht, M.C.M., 1999. Activated sludge model No. 3. Water Sci. Technol. 39, 183–193.

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 Commun 29, 747–756. https://doi.org/10.3233/AIC-160714.

Haimi, H., Mulas, M., Corona, F., Vahala, R., 2013. Data-derived soft-sensors for biological wastewater treatment plants: an overview. Environ. Model. Softw 47, 88–107. https://doi.org/10.1016/J.ENVSOFT.2013.05.009.

Han, H., Qiao, J., 2014. Nonlinear model-predictive control for industrial processes: an application to wastewater treatment process. IEEE Trans. Ind. Electron. 61, 1970–1982. https://doi.org/10.1109/TIE.2013.2266086.

Han, H.G., Chen, Q.L., Qiao, J.F., 2010. Research on an online self-organizing radial basis function neural network. Neural Comput. Appl. 19, 667–676. https://doi.org/10.1007/s00521-009-0323-6.

Haykin, 1999. Neural Networks. Prentice-Hill, Englewood Cliffs.

Henze, M., Grady, C.P.L.J., Gujer, W., Marais, G. v. R., Matsuo, T., 1987. Activated Sludge Model No. 1 (No. 1900222248). IAWPRC, London.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M.C., Marais, G. v. R., 1995. Activated Sludge Model No. 2. IAWQ, London.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M.C., Marais, G. v. R., Van Loosdrecht, M.C.M., 1999. Activated sludge model No. 2d. Water Sci. Technol. 39, 165–182.
Hermansson, A.W., Syafiie, S., 2015. Model predictive control of pH neutralization processes: a review. Contr. Eng. Pract. 45, 98–109. https://doi.org/10.1016/j.conengprac.2015.09.005.
Hinton, G.E., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.
Hoskuldsson, A., 1996. Prediction Methods in Science and Technology. Thor Publishing, Copenhagen.
Hotelling, H., 1947. Multivariate quality control. In: Eisenhart, C., Hastay, M.W., Wills, W.A. (Eds.), Techniques of Statistical Analysis. McGraw-Hill, New York, NY, pp. 111–184.
Hunter, J.S., 1986. The exponentially weighted moving average. J. Qual. Technol. 18, 203–210.
Isermann, R., 1984. Process fault detection based on modeling and estimation methods—a survey. Automatica 20, 387–404.
Izenman, A.J., 2013. Kernel density estimation. In: Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer Texts in Statistics. Springer, New York, pp. 88–100.
Jackson, J.E., 1991. A User's Guide to Principal Components. John Wiley & Sons, Inc.
Jałowiecki, Ł., Chojniak, J.M., Dorgeloh, E., Hegedusova, B., Ejhed, H., Magnér, J., Płaza, G.A., 2016. Microbial community profiles in wastewaters from onsite wastewater treatment systems technology. PLoS One 11, e0147725. https://doi.org/10.1371/journal.pone.0147725.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7.
Jeppsson, U., Pons, M.N., 2004. The COST benchmark simulation model—current state and future perspective. Contr. Eng. Pract. 12, 299–304. https://doi.org/10.1016/j.conengprac.2003.07.001.
Jeppsson, U., Pons, M.N., Nopens, I., Alex, J., Copp, J.B., Gernaey, K.V., Rosen, C., Steyer, J.P., Vanrolleghem, P.A., 2007. Benchmark simulation model no 2: general protocol and exploratory case studies. Water Sci. Technol. 56, 67–78. https://doi.org/10.2166/wst.2007.604.
Jiang, W., Wang, K.B., Tsung, F., 2012. A variable-selection-based multivariate EWMA chart for process monitoring and diagnosis. J. Qual. Technol. 44, 209–230.
Jung, D., Kang, D., Liu, J., Lansey, K., 2013. Improving resilience of water distribution system through burst detection. In: World Environmental and Water Resources Congress 2013. Presented at the World Environmental and Water Resources Congress 2013. American Society of Civil Engineers, Cincinnati, Ohio, pp. 768–776. https://doi.org/10.1061/9780784412947.073.
Kadiyala, R., 2018. Leveraging Other Industries - Big Data Management. Water Environment Federation Water Research Foundation.
Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven soft sensors in the process industry. Comput. Chem. Eng. 33, 795–814. https://doi.org/10.1016/j.compchemeng.2008.12.012.
Kazor, K., Holloway, R.W., Cath, T.Y., Hering, A.S., 2016. Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility. Stoch. Environ. Res. Risk Assess. 30, 1527–1544.
King, K.L., Wang, Z., Kroll, D.J., 2006. Classification of Deviations in a Processes, 3723429343.
Korkmaz, S., Goksuluk, D., Zararsiz, G., 2014. MVN: an R package for assessing multivariate normality. Rom. Jahrb. 6, 151–162.
Kruger, U., Zhou, Y., Irwin, G.W., 2004. Improved principal component monitoring of large-scale processes. J. Process Control 14, 879–888. In: https://doi.org/10.1016/j.jprocont.2004.02.002.
Ku, W., Storer, R.H., Georgakis, C., 1995. Disturbance detection and isolation by dynamic principal component analysis. Chemometr. Intell. Lab. Syst. 30, 179–196.
Kusiak, A., Zeng, Y., Zhang, Z., 2013. Modeling and analysis of pumps in a wastewater treatment plant: a data-mining approach. Eng. Appl. Artif. Intell. 26, 1643–1651. https://doi.org/10.1016/j.engappai.2013.04.001.
Kwak, S.K., Kim, J.H., 2017. Statistical data preparation: management of missing values and outliers. Korean J. Anesthesiol. 70, 407–411. https://doi.org/10.4097/kjae.2017.70.4.407.
Labrinidis, A., Jagadish, H.V., 2012. Challenges and opportunities with big data. Proc. VLDB Endow. 5, 2032–2033.
Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety (No. 0950–5849), Application Delivery Strategies. META Group.
Langergraber, G., Fleischmann, N., Hofstädter, F., 2003. A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. Water Sci. Technol. 47, 63–71.
Lee, C., Choi, S.W., Lee, I.B., 2006. Sensor fault diagnosis in a wastewater treatment process. Water Sci. Technol. 53, 251–257. https://doi.org/10.2166/wst.2006.027.
Lee, D.S., Jeon, C.O., Park, J.M., Chang, K.S., 2002. Hybrid neural network modeling of a full-scale industrial wastewater treatment process. Biotechnol. Bioeng. 78, 670–682. https://doi.org/10.1002/bit.10247.
Lee, Dae Sung, Lee, M.W., Woo, S.H., Kim, Y.-J., Park, J.M., 2006. Nonlinear dynamic partial least squares modeling of a full-scale biological wastewater treatment plant. Process Biochem. 41, 2050–2057. In: https://doi.org/10.1016/j.procbio.2006.05.006.
Lee, D.S., Lee, M.W., Woo, S.H., Kim, Y.-J., Park, J.M., 2006. Multivariate online monitoring of a full-scale biological anaerobic filter process using kernel-based algorithms. Ind. Eng. Chem. Res. 45, 4335–4344. https://doi.org/10.1021/ie050916k.
Lee, D.S., Vanrolleghem, P.A., 2004. Adaptive consensus principal component analysis for on-line batch process monitoring. Environ. Monit. Assess. 92, 119–135.
Lee, J.-M., Yoo, C., Choi, S.W., Vanrolleghem, P.A., Lee, I.-B., 2004. Nonlinear process monitoring using kernel principal component analysis. Chem. Eng. Sci. 59, 223–234. https://doi.org/10.1016/j.ces.2003.09.012.
Lee, M.W., Hong, S.H., Choi, H., Kim, J.-H., Lee, D.S., Park, J.M., 2008. Real-time remote monitoring of small-scaled biological wastewater treatment plants by a multivariate statistical process control and neural network-based software sensors. Process Biochem. 43, 1107–1113. In: https://doi.org/10.1016/j.procbio.2008.06.002.
Lowry, C.A., Woodall, W.H., Champ, C.W., Rigdon, S.E., 1992. A multivariate exponentially weighted moving average control chart. Technometrics 34, 46–53. https://doi.org/10.1080/00401706.1992.10485232.
Luccarini, L., Bragadin, G.L., Colombini, G., Mancini, M., Mello, P., Montali, M., Sottara, D., 2010. Formal verification of wastewater treatment processes using events detected from continuous signals by means of artificial neural networks. Case study: SBR plant. Environ. Model. Softw 25, 648–660. https://doi.org/10.1016/j.envsoft.2009.05.013.
MacGregor, J.F., Kourti, T., 1995. Statistical process control of multivariate processes. Contr. Eng. Pract. 3, 403–414.
MacGregor, J.F., Nomikos, P., Kourti, T., 1994. Multivariate statistical process control of batch processes using PCA and PLS. IFAC Proc 27, 523–528. https://doi.org/10.1016/S1474-6670(17)48203-6.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011. Big Data: the Next Frontier for Innovation, Competition, and Productivity. No. 0983179697, 978–0983179696.
Metcalf, Eddy, 2013. Wastewater Engineering: Treatment and Resource Recovery, fifth ed. McGraw-Hill Education.
Miah, A.Q., 2016. Multiple regression. In: Applied Statistics for Social and Management Sciences. Springer Singapore, Singapore, pp. 233–244.
Mina, J., Verde, C., 2007. fault detection for large scale systems using dynamic principal components analysis with adaptation. Int. J. Comput. Commun. Control II 185–194.
Montgomery, D.C., 2009. Introduction to Statistical Quality Control. John Wiley & Sons, New York.
Nguyen, V.H., Golinval, J.-C., 2010. Fault detection based on kernel principal component analysis. Eng. Struct. 32, 3683–3691.
Nielsen, M.A., 2015. Neural Networks and Deep Learning. Determination Press.
NIST/SEMATECH, 2003. Univariate and multivariate control charts. In: E-handbook of Statistical Methods. U.S. Department of Commerce.
Nomikos, P., MacGregor, J.F., 1995. Multi-way partial least squares in monitoring batch processes. Chemometr. Intell. Lab. Syst. 30, 97–108. https://doi.org/10.1016/0169-7439(95)00043-7.
Nomikos, P., MacGregor, J.F., 1994. Monitoring batch processes using multiway principal component analysis. AIChE J. 40, 1361–1375. https://doi.org/10.1002/aic.690400809.
Nopens, I., Benedetti, L., Jeppsson, U., Pons, M.N., Alex, J., Copp, J.B., Gernaey, K.V., Rosen, C., Steyer, J.P., Vanrolleghem, P.A., 2010. Benchmark Simulation Model No 2: finalisation of plant layout and default control strategy. Water Sci. Technol. 62, 1967–1974. https://doi.org/10.2166/wst.2010.044.
O'Brien, M., Mack, J., Lennox, B., Lovett, D., Wall, A., 2011. Model predictive control of an activated sludge process: a case study. Contr. Eng. Pract. 19, 54–61. https://doi.org/10.1016/j.conengprac.2010.09.001.
Ocampo-Martinez, C., 2010. Suboptimal MPC approach based on piecewise linear functions. In: Grimble, M.J., Johnson, M.A. (Eds.), Model Predictive Control of Wastewater Systems, Advances in Industrial Control. Springer.
O'Day, D.K., 2004. Statistical Process Control Can Help Treatment Plants. Water World 20.
Odom, G.J., Newhart, K.B., Cath, T.Y., Hering, A.S., 2018. Multistate multivariate statistical process control. Appl. Stoch Model Bus. Ind. 34, 880–892. https://doi.org/10.1002/asmb.2333.
Olsson, G., 2012. ICA and me – a subjective review. Water Res. 46, 1585–1624. https://doi.org/10.1016/j.watres.2011.12.054.
Olsson, G., Aspegren, H., Nielsen, M.K., 1998. Operation and control of wastewater treatment—a Scandinavian perspective over 20 years. Water Sci. Technol. 37, 1–13.
Olsson, G., Newell, B., 1999. Wastewater Treatment Systems: Modelling, Diagnosis and Control. IWA Publishing, London, UK.
Olsson, G., Nielsen, M., Yuan, Z., Lynggaard-Jensen, A., Steyer, J.-P., 2005. Instrumentation, Control and Automation in Wastewater Systems. IWA publishing.
Ömer Faruk, D., 2009. A hybrid neural network and ARIMA model for water quality time series prediction. Eng. Appl. Artif. Intell. 23, 586–594. https://doi.org/10.1016/j.engappai.2009.09.015.
Oppong, G., Montague, G.A., O'Brien, M., McEwan, M., Martin OBE FREng, E.B., 2013. Towards advanced control for anaerobic digesters: volatile solids inferential sensor. Water Pract. Technol. 8. https://doi.org/10.2166/wpt.2013.002.
Park, S.-H., Koo, J., 2015. Application of transfer function ARIMA modeling for the sedimentation process on water treatment plant. Int. J. Control Autom. 8, 135–144. https://doi.org/10.14257/ijca.2015.8.10.13.
Patton, R.J., Frank, P.M., Clark, R.N., 2000. Issues of Fault Diagnosis for Dynamic Systems. Springer-Verlag.
Pfluger, A.R., Hahn, M.J., Hering, A.S., Munakata-Marr, J., Figueroa, L., 2018. Statistical

exposé of a multiple-compartment anaerobic reactor treating domestic wastewater. Water Environ. Res. 90, 530–542. https://doi.org/10.2175/106143017X15131012153068.
Phaladiganon, P., Kim, S.B., Chen, V.C.P., Baek, J.-G., Park, S.-K., 2011. Bootstrap-based T2 multivariate control charts. Commun. Stat. Simulat. Comput. 40, 645–662. https://doi.org/10.1080/03610918.2010.549989.
Qin, X., Gao, F., Chen, G., 2012. Wastewater quality monitoring system using sensor fusion and machine learning techniques. Water Res. 46, 1133–1144. https://doi.org/10.1016/j.watres.2011.12.005.
Regmi, P., Stewart, H., Amerlinck, Y., Arnell, M., Garcia, P.J., Johnson, B., Maere, T., Miletic, I., Miller, M., Rieger, L., Samstag, R., Santoro, D., Schraa, O., Snowling, S., Takacs, 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., 2018. The future of WRRF modelling – outlook and challenges. Water Sci. Technol. (in press) https://doi.org/10.2166/wst.2018.498.
Rice, E.W., Baird, R.B., Eaton, A.D., 2017. Standard Methods for the Examination of Water and Wastewater, 23rd ed. American Water Works Association/American Public Works Association/Water Environment Federation.
Richalet, J., Rault, A., Testud, J.L., Papon, J., 1978. Model predictive heuristic control: applications to industrial processes. Automatica 14, 413–428. https://doi.org/10.1016/0005-1098(78)90001-8.
Roberts, S.W., 1959. Control chart tests based on geometric moving averages. Technometrics 1, 239–250.
Rosen, C., Jeppsson, U., Vanrolleghem, P. a, 2004. Towards a common benchmark for long-term process control and monitoring performance evaluation. Water Sci. Technol. 50, 41–49.
Rosen, C., Lennox, J.A., 2001. Multivariate and multiscale monitoring of wastewater treatment operation. Water Res. 35, 3402–3410. https://doi.org/10.1016/S0043-1354(01)00069-0.
Rosen, C., Röttorp, J., Jeppsson, U., 2003. Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation. Water Sci. Technol. 47, 171–179.
Schraa, O., Tole, B., Copp, J.B., 2006. Fault detection for control of wastewater treatment plants. Water Sci. Technol. 53, 375–382. https://doi.org/10.2166/wst.2006.143.
Sheather, S., 2009. A Modern Approach to Regression with R. Springer Science & Business Media.
Shewhart, W.A., 1926. Quality control charts. Bell Syst. Tech. J. 5, 593–603. https://doi.org/10.1002/j.1538-7305.1926.tb00125.x.
Sin, G., Villez, K., Vanrolleghem, P.A., 2006. Application of a model-based optimisation methodology for nutrient removing SBRs leads to falsification of the model. Water Sci. Technol. 53, 95–103. https://doi.org/10.2166/wst.2006.114.
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. Util. Pol. 45, 69–75. https://doi.org/10.1016/j.jup.2017.02.002.
Slawecki, T., McMaster, C., Lodhi, M., Hornback, C., Nakamura, B., 2016. Big Data: preparing for the future of "smart water" and "big data. Water Environ. Technol. 88–91.
Smilde, A., Bro, R., Geladi, P., 2005. Multi-way Analysis: Applications in the Chemical Sciences. John Wiley & Sons.
Spanjers, H., Vanrolleghem, P., Nguyen, K., Vanhooren, H., Patry, G.G., 1998. Towards a simulation-benchmark for evaluating respirometry-based control strategies. Water Sci. Technol. 37, 219–226. https://doi.org/10.1016/S0273-1223(98)00373-4.
Tae-Hwan, H., Eui-Suck, N., Kwang-Bang, W., Kim, C.J., Jeong-Woong, R., 1997. Optimization of coagulant dosing process in water purification system. In: Presented at the Proceedings of the 36th SICE Annual Conference. International Session Papers, pp. 1105–1109. https://doi.org/10.1109/SICE.1997.624942.
US EPA, 2015. Continuous Monitoring Data Sharing Strategy. EP-C-12-052 Task Order No. 0005.
US EPA, 2014. Configuring Online Monitoring Event Detection Systems. 600/R-14/254.
Vanrolleghem, P. a, Lee, D.S., 2003. On-line monitoring equipment for wastewater treatment processes: state of the art. Water Sci. Technol. 47, 1–34.
Vasanth Kumar, K., Porkodi, K., Avila Rondon, R.L., Rocha, F., 2008. Neural network modeling and simulation of the solid/liquid activated carbon adsorption process. Ind. Eng. Chem. Res. 47, 486–490. https://doi.org/10.1021/ie071134p.
Venkatasubramanian, V., 1995. Towards integrated process supervision: current status and future directions. In: Arzen, K.-E. (Ed.), Computer Software Structures Integrating AI/KBS Systems in Process Control. Pergamon, Lund, Sweden.
Verma, A., Wei, X., Kusiak, A., 2013. Predicting the total suspended solids in wastewater: a data-mining approach. Eng. Appl. Artif. Intell. 26, 1366–1372. https://doi.org/10.1016/j.engappai.2012.08.015.
Villez, K., 2007. Multivariate and Qualitative Data-Analysis for Monitoring, Diagnosis and Control of Sequencing Batch Reactors for Wastewater Treatment. Ph.D. thesis. Ghent University, Ghent, Belgium.
Volcke, E.I.P., van Loosdrecht, M.C.M., Vanrolleghem, P.A., 2006. Continuity-based model interfacing for plant-wide simulation: a general approach. Water Res. 40, 2817–2828. https://doi.org/10.1016/j.watres.2006.05.011.
Vuono, D., Henkel, J., Benecke, J., Cath, T.Y., Reid, T., Johnson, L., Drewes, J.E., 2013. Flexible hybrid membrane treatment systems for tailored nutrient management: a new paradigm in urban wastewater treatment. J. Membr. Sci. 446, 34–41. https://doi.org/10.1016/j.memsci.2013.06.021.
Wallace, J., Champagne, P., Hall, G., 2016. Multivariate statistical analysis of water chemistry conditions in three wastewater stabilization ponds with algae blooms and pH fluctuations. Water Res. 96, 155–165.
Wang, K.B., Jiang, W., 2009. High-dimensional process monitoring and fault isolation via variable selection. J. Qual. Technol. 41, 247–258.
Wang, X., Ratnaweera, H., Holm, J.A., Olsbu, V., 2017. Statistical monitoring and dynamic simulation of a wastewater treatment plant: a combined approach to achieve model predictive control. J. Environ. Manag. 193, 1–7. https://doi.org/10.1016/j.jenvman.2017.01.079.
Wangen, L.E., Kowalski, B.R., 1989. A multiblock partial least squares algorithm for investigating complex chemical systems. J. Chemom. 3, 3–20. https://doi.org/10.1002/cem.1180030104.
Wei, X., 2013. Modeling and Optimization of Wastewater Treatment Process with a Data-Driven Approach. University of Iowa.
West, D., Dellana, S., Jarrett, J., 2002. Transfer function modeling of processes with dynamic inputs. J. Qual. Technol. 34, 315–326.
Wise, B.M., Gallagher, N.B., 1996. The process chemometrics approach to process monitoring and fault detection. J. Process Control 6, 329–348. https://doi.org/10.1016/0959-1524(96)00009-1.
Wold, S., 1994. Exponentially weighted moving principal components analysis and projections to latent structures. Chemometr. Intell. Lab. Syst. 23, 149–161.
Woo, S.H., Jeon, C.O., Yun, Y.-S., Choi, H., Lee, C.-S., Lee, D.S., 2009. On-line estimation of key process variables based on kernel partial least squares in an industrial cokes wastewater treatment plant. J. Hazard Mater. 161, 538–544. https://doi.org/10.1016/j.jhazmat.2008.04.004.
Xiao, H., Huang, D., Pan, Y., Liu, Y., Song, K., 2017. Fault diagnosis and prognosis of wastewater processes with incomplete data by the auto-associative neural networks and ARMA model. Chemometr. Intell. Lab. Syst. 161, 96–107. https://doi.org/10.1016/j.chemolab.2016.12.009.
Yang, W., Nan, J., Sun, D., 2008. An online water quality monitoring and management system developed for the Liming River basin in Daqing, China. J. Environ. Manag. 88, 318–325. https://doi.org/10.1016/j.jenvman.2007.03.010.
Yoo, C.K., Lee, D.S., Vanrolleghem, P.A., 2004. Application of multiway ICA for on-line process monitoring of a sequencing batch reactor. Water Res. 38, 1715–1732. https://doi.org/10.1016/j.watres.2004.01.006.
Yoo, C.K., Vanrolleghem, P.A., Lee, I.-B., 2003. Nonlinear modeling and adaptive monitoring with fuzzy and multivariate statistical methods in biological wastewater treatment plants. J. Biotechnol. 105, 82–54. https://doi.org/10.1016/S0168-1656(03)00168-8.
Zadeh, L.A., 1973. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybern. 1100, 38–45.
Zhang, B., Zhang, W., 2006. Adaptive predictive functional control of a class of nonlinear systems. ISA Trans. 45, 175–183.