ACTA UNIVERSITATIS UPSALIENSIS
Uppsala Dissertations from the Faculty of Science and Technology
145

Oscar Samuelsson

Sensor Fault Detection and Process Monitoring in Water Resource Recovery Facilities

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Zoom, Friday, 5 March 2021 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Elling W. Jacobsen.

Online defence: https://uu-se.zoom.us/j/69361047713

Abstract
Samuelsson, O. 2021. Sensor Fault Detection and Process Monitoring in Water Resource Recovery Facilities. *Uppsala Dissertations from the Faculty of Science and Technology* 145. 270 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1116-6.

Water resource recovery facilities (WRRFs) operate 24/7 to reduce the environmental impact from wastewater on receiving waters. Inaccurate measurements hinder the improvement of operations, limits the performance of automatic control, and deteriorate data quality for decision support and other purposes. This thesis studied how faults can be detected in sensors and impact the treatment process, including aeration diffusers. Simulation studies as well as three 6-18 months long pilot and full-scale experiments were conducted. Evidence was given for the commonplace problem with biofilm formation, and the consequence of biased measurements in two types of dissolved oxygen (DO) sensors. The condition of the energy critical aeration diffusers was monitored by combining process models and a tailored process disturbance, which subsequently improved the information content in existing measurements. The deliberate disturbance approach was also successful in predicting fouling and other faults in DO sensors, and further enabled separation of sensor faults from process variations. The practicability of several machine learning methods was studied for both sensor and process monitoring applications. Probabilistic one-class classification methods showed promising for automatically tuning the alarm threshold, although simple methods produced similarly good results in many situations. Lack of annotated data limited the applicability of the classification methods. For sensor fault detection, this was mitigated by using data from sensor maintenance routines. The need for overall good data quality to identify deviating measurements was underscored when data reconciliation was applied for process monitoring. Reaching a balance between theoretical and practical limitations was further pinpointed as a success factor for data reconciliation. Many previously unknown disturbances in the sensors and the treatment process were revealed during the experiments and resulted in improvement opportunities. A major negative impact from biased sensor signals on treatment efficiency was quantified and analysed in simulations, where the drift direction appeared to be vital. Knowledge gaps related to current sensor data quality were identified and studies were proposed to mitigate the identified shortcomings. Ultimately, the findings in this thesis underline the significance of analysing data using fault detection methods, which can enable a better overall system understanding and decision support.

Keywords: Aeration efficiency; anomaly detection; ammonium feedback control; change detection; covariance matrix; diffuser fouling; experimental design; Gaussian process regression; novelty detection; wastewater treatment

Oscar Samuelsson, Department of Information Technology, Division of Systems and Control, Box 337, Uppsala University, SE-75105 Uppsala, Sweden. Department of Information Technology, Automatic control, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

ISSN 1104-2516
ISBN 978-91-513-1116-6
urn:nbn:se:uu:diva-431488 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-431488)

To Lea and Thelma - part of the future

Contents

Populärvetenskaplig sammanfattning

Tillgång till rent vatten är livsviktigt och avgörande för att ett samhälle ska fungera. I takt med att förståelsen kring avloppsvattnets miljöpåverkan har ökat, så har även reningskraven och tekniken utvecklats. Idag är ett modernt reningsverk en processindustri med biotekniska, kemiska och mekaniska processer som begränsar utsläpp av näringsämnen (kväve och fosfor), substanser som kan leda till syrebrist i sjöar samt i vissa fall även läkemedelsrester. Ett reningsverk renar vatten dygnet runt och fel och störningar behöver därför snabbt upptäckas och åtgärdas.

Reglerteknik har varit en viktig komponent för att automatiskt kunna styra de alltmer avancerade reningsprocesserna på ett resurseffektivt sätt. Reglerteknik är dock känsligt för felaktiga mätningar som kan leda till oönskade konsekvenser. Felaktiga mätningar döljs inom en regulator och är tyvärr luriga att upptäcka.

Som exempel kan man ta en syrehaltsregulator i den biologiska reningen. Normalt så styr regulatorn mot ett målvärde på 2 mg/L i syrehalt. Regulatorn utgår från en sensor som mäter syrehalten i reningsverket och justerar lufttillförseln så att den önskade syrehalten, eller målvärdet, uppnås. Om nu syresensorn mäter 0,5 mg/L för lite (d.v.s. 1,5 mg/L när det i verkligheten är 2,0 mg/L) så kommer regulatorn ändå att styra mot (vad den tror är) 2,0 mg/L. Den sanna syrehalten kommer då bli 2,5 mg/L vilket kommer vara mindre energieffektivt än det önskade målvärdet. Det luriga är att den uppmätta syrehalten är 2,0 mg/L, vilket är det önskade, trots att mätningen egentligen är felaktig.

Hur fel i flera sensorer samverkar och påverkar reningsfunktionen och energiförbrukningen har studerats med hjälp av datorsimuleringar. Typen av fel (positivt eller negativt mätfel) och vilken sensor det var fel i visade sig ha betydelse för hur stora de negativa konsekvenserna blev. Detta visar på vikten av att förstå hur man kan undvika sensorfel samt prioritera de mest kritiska sensorerna.

En stor del av avhandlingens arbete har varit att utveckla metoder som kan detektera fel. Utöver sensorfel har detektion av processtörningar och fel på den tekniska reningsutrustningen studerats i experiment på Stockholms tre reningsverk i försök under 6–18 månader. Den vanligast orsaken till felaktiga sensormätningar är biofilmspåväxt. Detta studerades i detalj och visade att risken för mätfel varierar i takt med årstidsvariationerna. Biofilmspåväxt kan

även minska effektiviteten på luftningsutrustningen, vilket kunde övervakas med en av de framtagna metoderna.

Oavsett tillämpning har tillvägagångssättet varit att analysera data med hjälp av olika matematiska modeller. Modeller har använts för att filtrera fram den information som avslöjar om det finns ett fel eller inte i den studerade sensorn eller processen. En modell kan sägas sätta mätningarna i proportion till varandra, där modellen beskriver vår kunskap om mätningarna som matematiska samband. Detta kan till exempel användas för att förstå om en uppmätt syrehalt är rimlig, givet den aktuella lufttillförseln och syreförbrukningen. Flera typer av modeller har studerats och visat sig användbara för feldetektion. Dels de som enbart använder data och mönster i data för prediktioner (maskininlärning), samt de som även inkluderar kända fysikaliska samband som ekvationer (processmodeller).

Att bestämma larmgränsen för vad som bedöms som normalt eller felaktigt är ett kritiskt steg i all feldetektion. Resultaten i avhandlingen tydliggör det förväntade sambandet att god tillgång på information (data) om hur de normala och felaktiga tillstånden förenklar uppgiften att bestämma gränsen däremellan (d.v.s att välja larmgräns). Kunskap om både normal- och feltillstånd hos sensorer och processer är dock begränsad i praktiken, vilket därmed begränsar möjligheterna med maskininlärning. En enkel lösning förslås som åtgärd, vilken i korthet innebär att data om sensorunderhåll används för att identifiera sensorns normaltillstånd.

En annan metodik som studerades för att öka informationsinnehållet i data var att tillfälligt störa, eller excitera, processen. Genom denna metodik kan man undvika att variationer i mätningarna (mät- och processbrus) skymmer de förändringar man vill detektera. Metodiken visade sig användbar för att både detektera missvisande syresensorer, samt för att samla in mer informativa mätdata till en modell som i förlängningen används för att övervaka att luftningsutrustningen fungerar bra. Genom att enbart störa sensorn blir det möjligt att separera fel i sensorn från processvariationer, vilket vanligtvis är mycket svårt, såsom i exemplet med syreregulatorn. Liknande metoder har framgångsrikt använts i andra tillämpningar och resultaten i denna avhandling visade att det även är tillämpbart för feldetektion på reningsverk.

Resultaten i avhandlingen har bidragit till en bättre förståelse för hur sensorfel påverkar reningsprocessen, samt hur flera typer av fel kan detekteras genom att analysera data på olika sätt. Detta kan leda till mera robust och resurseffektiv drift av reningsverk. Flera feldetektionsmetoderna som använts i avhandling har dessutom utvärderats på verkliga processer och kan därför även tillämpas i praktiken.

Acknowledgements

A long journey has come to its end. Looking back, it has been challenging in many aspects, but also rewarding to see ideas spark during a cup of coffee, and develop into research projects. Many of those coffee cups were made empty in the company of my main supervisor Bengt Carlsson, who I had the privilege to learn from over the years. My deepest gratitude to you for all challenging discussions, and your patience with my more (or less) feasible ideas. I would also like to express my deepest gratitude to Anders Björk, my co-supervisor at IVL. You have been my main source of encouragement from day one, both for prioritizing important over urgent, and for providing the larger picture when experiments seemingly fail. I am deeply grateful for all discussions with Gustaf Olsson, my mentor for the last two years. You have been a great source of inspiration to me with your positive attitude and willingness to share both technical knowledge and personal experiences. I would also like to thank Thomas Schön at Uppsala University for raising my interest in probabilistic machine learning and your superb teaching effort.

My research effort has not been lonely, and owe a great deal of gratitude to all co-authors and peers. In addition to the mentioned supervisors, Jesús Zambrano, Erik Lindblom, Hong Quan Le, Stefan Erikstam, I have learned a lot from you during all our inspiring discussions. Kris Villez, you have challenged my mindset and together with the MetaCo task group, you have all been a great source of inspiration.

The staff at Stockholm Water, Käppala, and Syvab have been invaluable for executing the experiments. The same organizations also funded this project together with the foundation for Swedish Environmental Research Institute and the Swedish Water and Wastewater association (VA-kluster Mälardalen). I am must grateful for this funding and the opportunities they gave.

Thank you, all of my colleagues (former and present) at IVL for providing a dynamic and inspiring research environment. A special thanks to Max, for your support and made insights through our local PhD group. I would also like to thank Fredrik, Viktor, Andreas, Carl, Tatiana, and Rubén, for making me feel like a real PhD student during some of our joint courses, and my (occasional) weekly Uppsala visit.

Last, my largest gratitude goes to my family, and especially my beloved Marielle, Lea, and Thelma. Your support have been decisive to complete this work.

“However, not everything that can be counted counts, and not everything that counts can be counted”

William Bruce Cameron

1 Introduction

The goal with this thesis is to develop and apply methods that can detect deviations from the desired behaviour in components in water resource recovery facilities (WRRFs). In short, apply fault detection methods (FDMs) to improve wastewater treatment.

1.1 Research motivation

WRRFs operate 24/7 to limit the environmental impact from pathogens, nutrients, and recently also pharmaceutical residues in the used water (wastewater). Through the global urbanization trend, cities grow large and concentrate the anthropogenic impact on the locally receiving waters. For this reason, effluent wastewater permits are getting increasingly strict. This is certainly true for Stockholm in Sweden, with three large WRRFs neighbouring the nutrient-sensitive Baltic Sea. These represent about 15 percent of the total phosphorous load and 30 percent of Sweden's total nitrogen load from WRRFs, to the Eastern Baltic Sea (Käppalaförbundet 2018, Naturvårdsverket 2018, Stockholm vatten och avfall 2019). All three facilities are either currently re-built or are planned for retrofitted processes in order to compensate for the increased influent load and the stricter effluent permits. A benefit, however, with large centralized WRRFs is that advanced treatment technologies become feasible.

Automatic control has become an integral part for leveraging the benefits with advanced treatment processes. Automatic control has additionally allowed a balance between treatment efficiency and resource efficiency. For example, consumption of electrical energy and chemicals is a trade-off for emissions to the receiving water and air (Larsen 2015).

But automatic control is also vulnerable for failures. An automatic control loop can produce very poor performance, much worse than manual control, if fed with inaccurate information. In this respect, accurate on-line sensor measurements are essential. Additionally, many manually taken process operation decisions are based on data from on-line sensors.

By contrast, early WRRF process monitoring was based on manual sampling and laboratory analyses. This was successively changed through the introduction of dissolved oxygen (DO) sensor in the 1970's, in combination with automatic process control (Olsson 2012). Since then, a multitude of measuring

techniques have been developed to provide high resolution data on time scales of minutes or seconds. Data availability has transitioned from being a bottleneck, to currently overwhelm operators. Examples of commonly measured parameters include dissolved oxygen, flow, suspended solids, nitrous, phosphorous and carbonaceous fractions measured in the water phase. In addition, gas sensors have become important to assess the contribution of greenhouse gases from the treatment processes. Today, on-line sensors are ubiquitous and produce ever-increasing data streams.

In parallel to the development of measurement technologies, the costs of computational power and data storage have drastically decreased since the 1970's, and this has made it economically feasible to record and save the large quantities of process data. Simultaneously, the exploitation of data has emerged outside process control to influence most of the WRRF's organization. Apart from process optimization and key process indicators, historic process data are now used for design of retrofitted processes, calibration of process models, and are recently expected to be part of the game-changer of data-driven support (artificial intelligence).

However, the increase in available data has led to a dual monitoring task, where operators need to monitor the process simultaneously as the status of the on-line sensors. High quality data do not come for free, despite affordable sensors. By contrast, the workload of sensor maintenance increases with the number of installed sensors. The maintenance typically requires manual cleaning, calibration, and replacement of disposable parts and this work is commonly routinely scheduled since there is currently no good method to support condition-based maintenance. The harsh measurement environment in WRRFs has often been underestimated in terms of the required sensor maintenance. A complicating aspect is that is difficult, or commonly impossible, to distinguish a sensor fault from a process disturbance.

These factors make data validation complicated, but nonetheless a central task. Several data quality validation approaches have been described (Irizar et al. 2008, Olsson et al. 2014, Rosén 2001), but there is no standard for automatic data validation in WRRFs (Olsson 2012). Ultimately, automatic data validation governed by sensor fault detection is needed for improving current operations and increasingly stricter effluent permits, automatically controlled advanced treatment technologies, and the partly unknown future data reuse purposes. In the end, data are expected to play an increasingly central role for how we treat our wastewater.

1.2 Research goal

The long-term goal with this research is to improve the data quality in the facilities. The initial hypothesis was that on-line sensor data quality will be

improved if faults can be quickly identified and subsequently corrected. Therefore, this thesis aims to

> "Develop and adapt fault detection methods with an emphasis on sensor faults, which are suitable for WRRF full-scale operations."

The need for improved data quality and fault detection was raised by stakeholders in the wastewater utilities, some of whom also funded part of this research project. An initial feasibility study was conducted in 2014 to identify the practical challenges related to data quality (Samuelsson 2014). A qualitative approach was used for the study with interviews of eight employees (process engineers, process managers, and instrument technicians), all working with data from different perspectives. The answers were considered as a good indication of the current needs for the three utilities during the initiation of this research project (2014). They are still relevant and have been a source of inspiration to target relevant problems.

This thesis was limited to consider problems related to FDMs, although the feasibility study illustrated a broader set of challenges. These included organizational- and measurements-technique-specific issues but were out of the scope for this thesis.

Two topics were considered in the feasibility study: 1) quality of on-line sensor data, and 2) FDMs applicable to such sensors. A few typical interview questions were:

Quality of on-line data
- Which issues exist today concerning data quality from on-line sensors?
- Which are the key factors that give rise to the aforementioned issues?

Fault detection methods
- What is the experience of fault detection and monitoring at wastewater treatment plants today?
- Which methods are used today to detect sensor faults and sensors disturbances?

The most frequent answers during the interviews are given in Table 1.1. They have a wide scope and would require extensive research (more than a PhD thesis at least) to be answered in full. Nevertheless, many problems will be considered throughout this thesis, as indicated in the first column in Table 1.1.

Table 1.1. Problems related to fault detection mentioned during interviews in the feasibility study. Table adapted from (Samuelsson 2014).

Considered in chapter	Problem	Importance
4	Can I trust this on-line value and more specifically, how much can I trust this on-line value?	
4,5,6,8	What is the normal condition in a plant?	
3,7	The constant state is a varying state. This makes it difficult to determine intervals for cleaning and calibration	Mentioned several times by people with different job assignments
2,3,7	The sensor is fouled, how can an interval for cleaning be decided before it is too late?	
-	Some processes are coupled in series and others in parallel, and this affects both sensors and their interaction with each other. How can this process knowledge be used to detect sensor faults	
3,5,7	How can slow drift in sensor values be detected in a continuous process?	
4,7,8	Is it possible to distinguish between sensor faults, process disturbances and dosage error (control issues)?	
4	The rapid increase in sensors and registered on-line data makes it hard to overview and use available information	Mentioned more than once
2	Which sensors are most important to maintain or fix right now? Is it possible to prioritize the importance of sensors, both by function and maintenance?	
3,7	Is it possible to detect sensor faults in sensors that show a reasonable process value?	

1.3 Fault detection

This section first gives a background to the definition of a fault (Section 1.3.1) and the specific conditions in WRRFs and how they impact the practical possibilities to detect faults (Section 1.3.2). Next, the historical development of fault detection (Section 1.3.3) is briefly outlined, and last, general theoretical aspects related to tuning and performance assessment of FDMs are described (Section 1.3.4).

1.3.1 A general definition of fault detection

A *fault* can be defined as

> "An unpermitted deviation of at least one characteristic property or parameter of the system from the acceptable/usual/standard condition" (Isermann and Ballé 1997)

and one definition of *fault detection* follows from the same article

> "Determination of the faults present in a system and the time of detection".

The natural step after fault detection is *fault diagnosis,* which includes isolation and identification as part objectives. The work described in this thesis has mainly been limited to fault detection.

The purpose of the fault definition was to achieve a consistency of terminology for a better understanding of various research efforts (Isermann and Ballé 1997) and this is the fault definition recommended by the International Federation for Automatic Control (IFAC). The definition nevertheless requires a subjective decision regarding the standard condition. In WRRFs, the standard or normal conditions vary considerably. To distinguish between normal variation and a fault is the key challenge and common sensor fault symptoms sensor signals are illustrated in *Figure 1.1*. The figure is adapted from and a summary of the symptoms described in (Olsson et al. 2005, Rosén 1998, Rosén et al. 2008, Yoo et al. 2008).

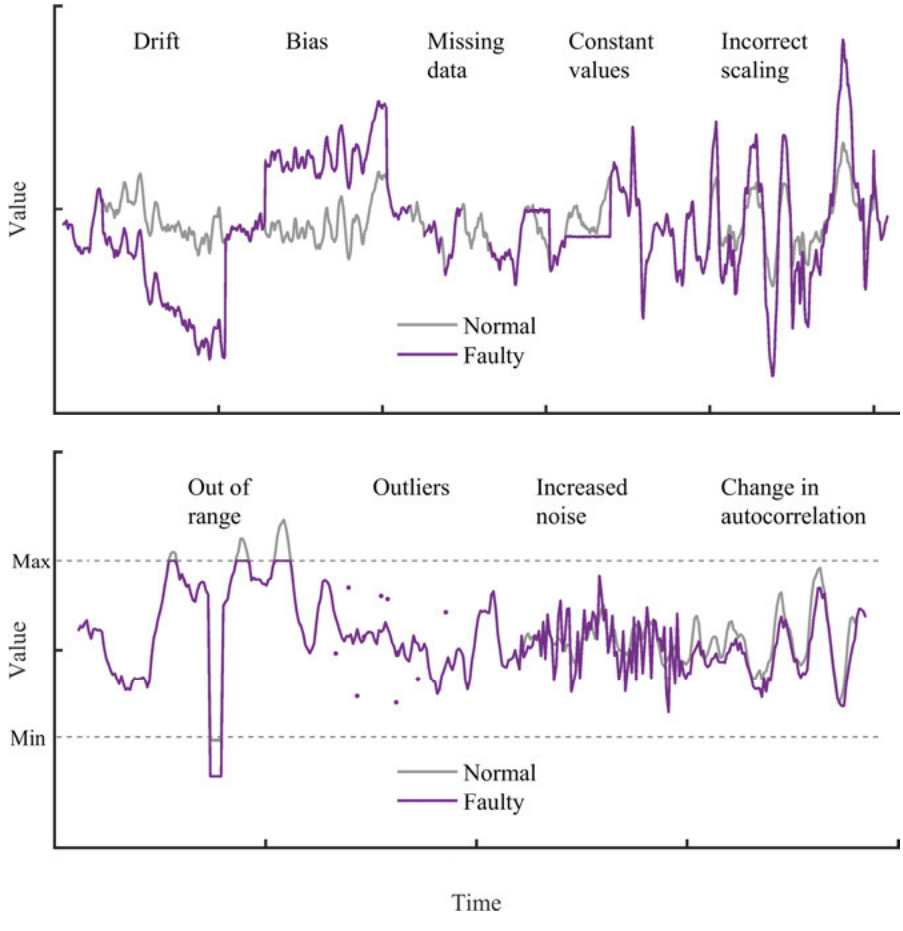


Figure 1.1. Common fault symptoms in sensor signals from WRRFs. The symptoms were simulated and added to data from a pre-processed dissolved oxygen measurements.

In practice, it is likely that a sensor fault would result in a combination of different symptoms. Unfortunately, little work has been done to characterize various faults and what symptoms they exhibit.

1.3.2 Wastewater data characteristics

The striking characteristic of WRRF data is that the signals vary considerably. The variations in data are a result of the highly varying conditions within the wastewater system, which are reflected in the influent. The measurement conditions for on-line sensors are also challenging and further contributes to the variations in data. The main sources of normal variations are introduced in this section since it is an integral task in fault detection to distinguish between normal variations and a change caused by a fault.

1.3.2.1 Influent variations

Wastewater collection and treatment naturally exhibit variations that are strongly linked to domestic water consumption patterns. Further, there is also a strong connection to surface or storm water sewers that are either combined or connected with the wastewater sewer at multiple overflow discharge outlets.

Although the relationship between water consumption and wastewater generation may seem straightforward, it is not possible to establish a clear relationship between the generating mechanisms and the expected wastewater patterns (Martin and Vanrolleghem 2014). Several efforts have been made to analyse (Farkas et al. 2020, Lindblom et al. 2019) and also generate artificial influent data for the purpose of modelling and simulating wastewater treatment processes (Martin and Vanrolleghem 2014, Saagi et al. 2017)

As an example, the influent generator described in (Gernaey et al. 2011) was developed to provide long-term influent data for simulations. The influent generator consists of six stochastic modules that illustrate the dominating factors leading to a variation in the influent. The six modules include: household wastewater generation, industrial discharge generation, soil infiltration, temperature effects, rain generation, and a sewer model that allowed for time delays and first flush effect in the transition from dry to storm weather. Rain (and snow melting) causes large deviations from the average flow.

The magnitude of the variations differs depending on the local conditions. Small wastewater collection systems (and resultingly small WRRFs) exhibit more pronounced variations than large systems since the flow variations are not dampened within the collection system (Metcalf and Eddy 2004). The magnitude (hydraulic residence time) of the collection system also leads to a time delay. The WRRFs in this thesis can be considered as large wastewater systems, treating wastewater from 300,000 up to 750,000 persons.

Characteristics such as daily, weekly and seasonal variation patterns are commonly evident, although the variations around an average pattern are substantial. In *Figure 1.2*, daily and seasonal variations for ammonium load are shown for 10 weeks in Bromma WRRF (300,000 persons).

The trend in decreasing ammonium concentration in *Figure 1.2* is a seasonal effect of the vacation period, initiated with the Swedish traditional midsummer celebration in the end of June. During the vacation period, the city population temporarily decreases with the effect of a decreasing load on the WRRF. The linkage between society and the influent composition variation has become widespread in the public through the pandemic in 2020 with extensive virus monitoring campaigns in the sewers, see e.g. (Randazzo et al. 2020).

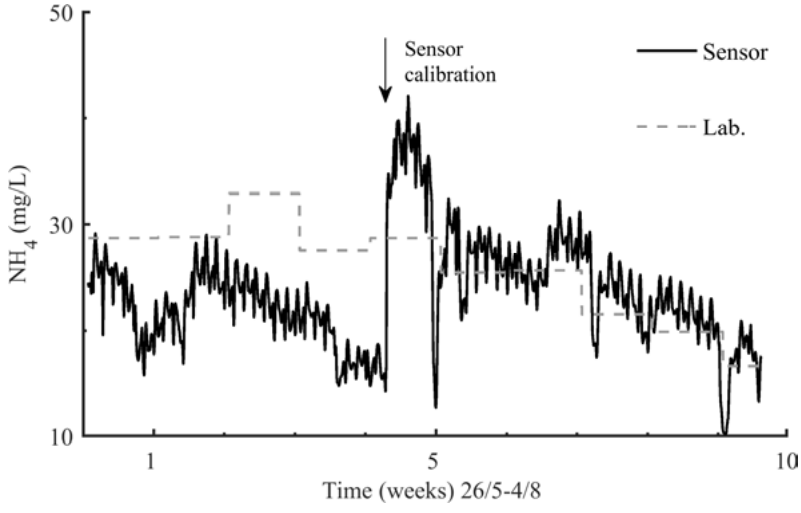


Figure 1.2. Daily and seasonal variations for ammonium load to the activated sludge process in Bromma WRRF. Weekly average laboratory samples (grey dashed lines) deviate from the on-line sensor values (black solid lines) before the sensor was calibrated. Similar drifts can be seen both before and after the sensor calibration.

Another illustration of variations in data is given in *Figure 1.3* where the influent flow during 3.5 years for Bromma WRRF is shown.

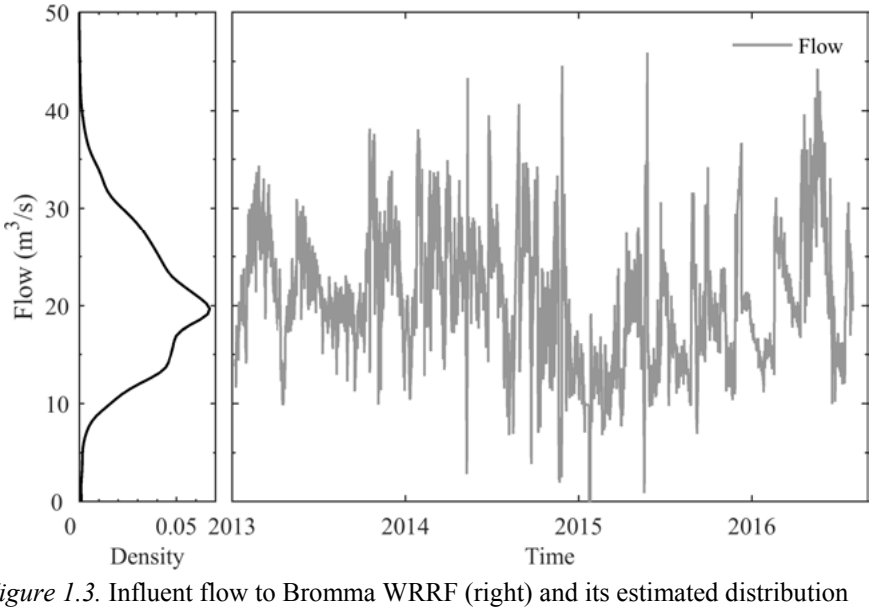


Figure 1.3. Influent flow to Bromma WRRF (right) and its estimated distribution (left).

In *Figure 1.3* it can be seen that the maximum (peak) flow is about twice as large as the average. Smaller WRRFs can exhibit much larger maximum to average ratios.

In summary, variations in influent data from wastewater systems exhibit regular key patterns (daily, weekly, and seasonal) that are frequently interrupted by rain and snow melting events.

1.3.2.2 Process variations

The influent variations also influence the biological treatment process since the treatment reaction rates are affected by both temperature and load concentrations (influent nitrogen, carbon, and phosphorous). The process dynamics for biological nutrient removal have been described in the well-established activate sludge model 1 (ASM 1) (Henze et al. 1987) using Monod equations. A simplified Monod model for the specific growth rate μ of nitrifying bacteria can be described by

$$\mu = \mu_{max} \frac{[NH_4]}{[NH_4] + K_{NH_4}} \tag{1.1}$$

where $[NH_4]$ is the ammonium (substrate) concentration in the reactor K_{NH_4} the half-saturation constant, and μ_{max} the maximum growth rate. The resulting ammonium removal rate r_{NH_4} in mass per volume and per unit time can be described by

$$r_{NH_4} = \frac{\mu}{Y_{max}} [X_B] \tag{1.2}$$

where $[X_B]$ are the nitrifying bacteria concentration in the reactor, Y_{max} is the maximum yield of nitrate given the ammonium load concentration. The maximum specific growth rate also depends on the temperature according to the van't Hoof equation (Henze et al. 2002)

$$\mu_{\max}(T) = \mu_{\max}(20°C) e^{\kappa(T-20)} \tag{1.3}$$

where κ is a temperature coefficient for the specific process. The example provided here is a simplification to illustrate that variations in the influent composition induce variations in the treatment process. In practice, additional factors such as dissolved oxygen concentration, pH, and possible inhibitors also affect the growth rate for nitrifying bacteria. Their effect can be described with coupled Monod expressions, similar to (1.1).

The maximum specific microbial growth rates varies from 0.6 – 0.8 d^{-1} (nitrifying bacteria) to 6.0 d^{-1} (heterotrophic bacteria)(Henze et al. 2002), indicating that changes in the microbiological composition are slow compared to influent flow rate and concentration dynamics. The microbial composition within a facility can have an important effect on the treatment rates, see for example the review on microbial ecology for denitrification processes (Lu et al. 2014). Recent development has excelled our knowledge about microbial ecosystem with genetic and species databases (Nierychlo et al. 2020). Still, understanding the function and the precise interplay between microbial composition and process conditions needs more research (Nierychlo et al. 2020, Saunders et al. 2016).

In addition to biological process variations, many process units are controlled by on-off actions including sludge pumping, intermittent aeration, and sludge scrapers. Such intermittent events can result in abrupt large variations or regular peaks in related variables.

1.3.2.3 Measurement related variations

The bulk data used for monitoring and control in WRRFs are obtained from on-line sensors, nutrient analysers, and laboratory analyses. In general, measurements contain variations that result in uncertainties about the true measured value. In control theory, measurement noise is commonly assumed to be white Gaussian noise and the dominant source of uncertainty. In practice, additional measurement uncertainties exist that complicate the assessment of variations from sensor data. To characterize sensor performance, a multitude of statistics have been described including: accuracy, precision, bias, trueness, repeatability, long-term stability, reproducibility, response time, calibration uncertainty, non-linearity, measurement noise, coefficient of variation, and limit of detection and quantification; to mention the most common. For on-line sensors at WRRFs, the most relevant performance standard is (ISO 15839:2003 *Water quality–On-line sensors/analysing equipment for water–Specifications and performance tests*.)

Although the ISO 15839:2003 standard was specifically developed for water quality sensors, few manufacturers follow it. Instead, they use other performance statistics or only part of the standard (Beaupré 2010). This makes it difficult to obtain a general view of the measurement uncertainties at WRRFs. For a comparison between the sensor performance statistics provided by sensor manufactures and ISO 15839:2003, see (Beaupré 2010). Although ISO 15839:2003 includes both recommendations for validation under laboratory conditions (without disturbances) and full-scale evaluation, the full-scale validation procedures are more of a general guideline, since they strongly depend on local measurement conditions (Beaupré 2010). Recent studies further suggest that pH-sensor drift depend on the individual sensor and should not be considered as a general and equal property (Ohmura et al. 2019).

In addition to the variations from the actual on-sensor or analyser, the location of the sensor in the WRRF also affects the measurement variations, mainly because of the heterogeneous conditions in the WRRF. The activated sludge process is heterogeneous in the sense that it is composed of flocs with different compositions (Wilén et al. 2003) and the solids-in-liquid and gases-in-liquid are unevenly distributed in the various process units (Samstag et al. 2016). The heterogeneity is difficult to quantify, although the current research in adopting computational fluid dynamics (CFD) for WRRFs indicates the importance of understanding the heterogeneous conditions, see (Karpinska and Bridgeman 2016, Samstag et al. 2016, Wicklein et al. 2016) for reviews on the topic.

Air bubbles (gas-to-liquid) are one major source of heterogeneity that affects on-line sensor measurements. For spectrometer and turbidity measurements, an increase in variance and off-set was noticed due to air bubbles (Beaupré 2010). Although the disturbance effect of bubbles has not been studied over a wide array of measurement technologies, it is general knowledge that on-line sensors have to be mounted in accordance with the manufacturer description, e.g. to avoid air bubbles stuck on the sensor's measurement cell. Methods for characterizing heterogeneous media have been suggested (Petersen and Esbensen 2005), and in the context of WRRF data (Rossi et al. 2011).

Incorrect sensor mounting should not be considered as normal, but it is normally difficult to obtain both an optimal sensor position (from a measurement perspective), and at the same time to allow easy sensor maintenance. This results in a trade-off for the selected sensor location where some measurement variations are normally an effect of the sensor location.

Ultimately, it is clear that WRRF data include measurement uncertainties, but the magnitude of those uncertainties are uncertain. It is an important objective for a FDM both to detect when measurement uncertainties exceed normal variation, but also to quantify the normal uncertainty magnitude in data.

1.3.3 The history and progression in fault detection

The field of fault detection and diagnosis is broad in the sense that most existing modelling methods can be adopted for fault detection purposes. As a result, a wide array of FDMs has been developed for various applications. The original driving force for developing fault diagnosis systems was safety critical applications starting with aerospace applications and electrocardiograms in the 1970's (Willsky 1976) followed by chemical processes (Himmelblau 1978), and submarines and space vehicles (Gertler 1988). More recently, fault detection has become increasingly important for cyber security (Showkatbakhsh et al. 2020).

As the field of fault detection evolved, numerous ways of classifying the FDMs have been suggested. In an early review by Gertler (1988), model- and

non-model based fault diagnosis system was sufficient to distinguish existing FDMs (Gertler 1988). Later, Isermann & Ballé (1997) reviewed five years development within the field and classified FDMs according to: observer based, parity space based, parameter estimation, frequency spectral analysis, and neural networks based fault detections methods. Change detection and fault classification methods were classified as: neural nets, fuzzy logic, Bayes classification, or hypothesis testing (Isermann and Ballé 1997). In the three-part extensive review by Venkatasubramanian *et al.* (2003), process FDMs were classified as either: quantitative model-based, qualitative model-based and process history-based were each group contained several sub-groups. For the application of nuclear power plants, fault detection and diagnosis methods were classified as: model-based, data-driven, and signal-based (Ma and Jiang 2011) were a similar classification was used in (Qin 2012) for process monitoring methods including the groups: model, signal, knowledge-based, and hybrid and active fault diagnosis methods. In a WRRF context, computer-based methods for data management were divided in order of complexity: Low-level data checking, basic or advanced information extraction, and Human-interpretable information extraction and knowledge management (Corominas et al. 2017). Grey-box modelling (Sohlberg and Jacobsen 2008) is one technique that has surprisingly received little attentions in WRRF applications.

Today, machine learning has entered the fault detection literature (Lei et al. 2020, Md Nor et al. 2020), which are grouped into supervised and unsupervised, classification or regression algorithms or deterministic and probabilistic methods. Related to fault detection with rare faults is the one-class classification, which include adaption of the supervised, and unsupervised methods based on the available data (Barnabé-Lortie et al. 2015, Sabokrou et al. 2018).

The inconsistency in classifying FDMs is confusing in order to understand why various FDMs have been successful in different applications. A common approach for the reviews has been to describe the occurrence of FDMs and their applications. This is however of little use for this thesis, since it is mainly the interplay between characteristics of data and the method that is interesting for understanding the rationale of a FDM. From a mathematical perspective, it does not matter whether data originate from a pressure sensor in a WRRF or a nuclear power plant. What does matter is if the data characteristics widely differ between the applications in such way that the fault detection task becomes more or less complicated. Therefore, the illustration in *Figure 1.4* adapted from (Venkatasubramanian et al. 2003), is appealing to assess how and why certain FDMs have been successful in specific applications.

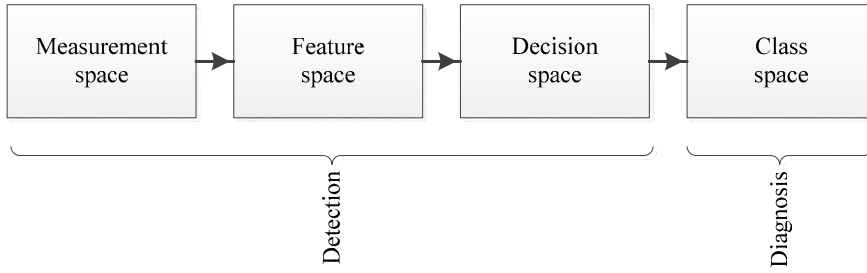


Figure 1.4. Transformation of measurement data to information suited for fault detection and diagnosis. Adapted from (Venkatasubramanian et al. 2003).

The main point in *Figure 1.4* is that it illustrates a FDM as a combination of two data transformation steps relating measurement space with *feature* and *decision space*. Transformation to feature space can be explained as raw data (measurement space) that are pre-processed in such a way that certain characteristics (features) become clear and better describe the difference between the normal and faulty mode, compared to measurement space. We thus use our a priori knowledge about the detection problem by considering relevant features. An example of a signal's feature is its variance. In the next step (decision space), a decision variable is defined in some test statistic. The last transformation step (class space) decide whether a fault is present or not (or for diagnosis, assigns the most likely fault hypothesis). Note that the decision space will be limited by the information provided from the feature space. At the same time, valued information from feature space will be of little use if a poor decision rule is applied. This interplay between feature and decision space is a hallmark of FDMs.

1.3.4 Tuning and performance assessment of fault detection methods

At the heart of fault detection lies the viability to distinguish normal from abnormal (faulty) observations. This relates to defining a suitable alarm threshold that balance false alarms and detection rate at a desired level. Due to the stochastic nature of most systems, probability theory has become the key tool for understanding how this threshold should be set.

The fundamental Neyman-Pearson theorem states an optimal hypothesis test when the probability distributions of the normal and faulty distributions are completely known (Kay 1998). In this test, which is denoted the likelihood ratio test (LRT), the test statistic $T_{LRT}(\boldsymbol{x})$, compares the conditional likelihood for the hypothesis of a faulty distribution, $p(\boldsymbol{x}|F)$, for the critical region of $\boldsymbol{x}$, with the equivalent likelihood for the normal hypothesis and distribution $p(\boldsymbol{x}|N)$.

The LRT is then given by

$$T_{LRT}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|F)}{p(\boldsymbol{x}|N)} > \gamma_{LRT} \tag{1.4}$$

where the threshold γ_{LRT}, identifies the maximum detection probability for the lowest, and user specified, false alarm probability α according to

$$\alpha = \int_{\phi} p(\boldsymbol{x}|N)d\boldsymbol{x} \tag{1.5}$$

where the integration is conducted over $\phi = \{\boldsymbol{x}|T_{LRT}(\boldsymbol{x}) > \gamma\}$. The optimally assigned threshold, in terms of a maximized detection probability for a given false alarm probability, is illustrated in *Figure 1.5*.

A decision threshold that is similar to (1.4) is obtained when instead the probability of error P_e (also known as the misclassification rate) is minimized, see (Bishop 2006, Kay 1998). The P_e is the sum of incorrect classifications, weighted by the probabilities for the normal $p(N)$ and faulty modes $p(F)$ as

$$P_e = p(F|N)p(N) + p(N|F)p(F). \tag{1.6}$$

Then, the P_e is minimized when an alarm is raised if

$$\frac{p(x|F)}{p(x|N)} > \frac{p(N)}{p(F)} = \gamma_{pe} \tag{1.7}$$

where γ_{pe} is the threshold (Kay 1998). The main difference between the LRT and P_e threshold is that the LRT emphasizes the false alarm rate and the P_e considers both the normal and faulty distribution. The P_e allows for a threshold adjustment by adding different costs for the misclassifications, see *Figure 1.5*. The cost adjusted misclassification rate, sometimes referred to as the Bayes risk raises an alarm if

$$\frac{p(x|F)}{p(x|N)} > \frac{(C_{fa} - C_n)\, p(N)}{(C_{md} - C_d)\, p(F)} = \gamma_{pe,} \tag{1.8}$$

where the costs for false alarm (C_{fa}) and missed detection (C_{md}) are user-defined values and the correct classifications true detections (C_d), and normal observations (C_n); are commonly assigned zero cost. As an example of the usefulness of the Bayes risk, consider an alarm that indicates broken diffusers in an activated sludge process. The mitigating counter action would involve a substantial maintenance effort. In contrast, the cost for running the WRRF one additional week to verify that the effort really is needed would be acceptable due to the large cost of false alarm.

The mentioned decision thresholds are optimal when the underling distributions are known. This is not true for most real applications such as the studied WRRF. For this reason, several variants of the LRT have been developed to approximate the true probability distributions.

Maximum likelihood estimation (MLE) is the most common approximation to estimate the distribution parameters for the faulty $\hat{\theta}_F$ and non-faulty (normal) distribution $\hat{\theta}_N$. This revised test statistic is known as the generalized likelihood ratio $T_{GLRT}(y)$, which is reliant on the MLE of assumed distributions (commonly Gaussian) in combination with data, y, as

$$T_{GLRT}(y) = \frac{p(y|\hat{\theta}_F)}{p(y|\hat{\theta}_N)} \tag{1.9}$$

The distribution parameters can also be marginalized out by assuming prior parameter distributions $p(\theta_N)$ and $p(\theta_F)$, which then gives the marginalized likelihood ratio (or Bayes factor) as

$$T_{MLRT}(y) = \frac{\int_{\theta_F} p(y|\theta_F)p(\theta_F)d\theta_F}{\int_{\theta_N} p(y|\theta_N)p(\theta_N)d\theta_N} \tag{1.10}$$

In the end, it should not be forgotten that the validity of any decision rule for real-world data are reliant on a good approximation of the underlying distributions and how well the generated data represent the true distributions. It should also be recalled that the original purpose for identifying an optimal decision threshold is to achieve good fault detection performance in a practical application.

The detection performance is commonly visualized by the Receiver operating characteristics (ROC). In a ROC-curve, the false alarm probability for varying thresholds of γ is plotted against the detection probability, resulting in a concave curve (*Figure 1.5*, bottom graph)

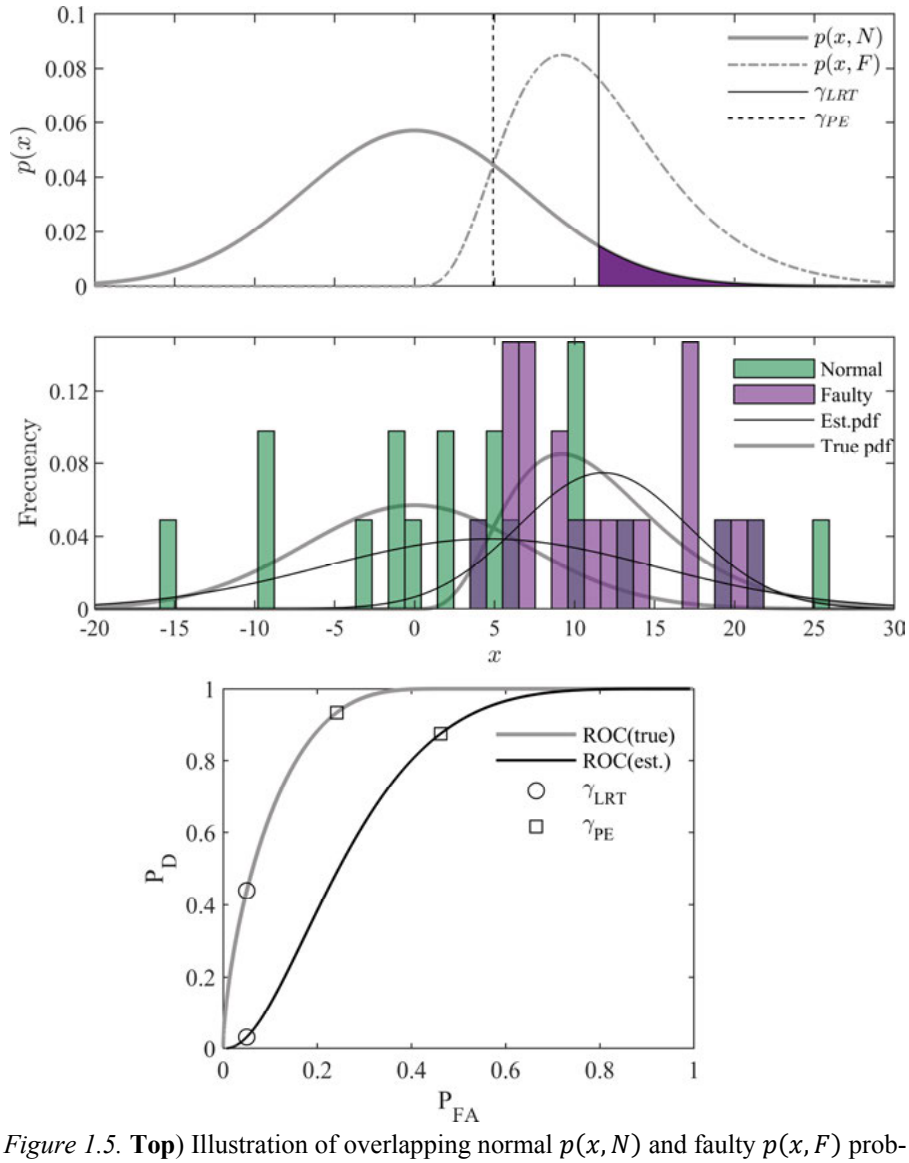


Figure 1.5. **Top**) Illustration of overlapping normal $p(x, N)$ and faulty $p(x, F)$ probability distributions. Shaded area indicates the critical region $\boldsymbol{x}$ and likewise the false alarm probability for the threshold γ_{LRT}. The minimized misclassification rate is indicated with γ_{PE}. **Middle**) Histograms based on 20 random samples from the true distributions (grey, thick lines), with the estimated distributions (black, thin lines). **Bottom**) The receiver operating characteristics for the true (grey, thick line) and estimated distributions (black, thin lines) in the middle graph. The circles and squares indicate the thresholds assigned by LRT and PE, respectively.

An optimal detector would give a complete detection probability with no false alarm resulting in a flat ROC-curve with detection probability one for all false alarm probabilities. Such a detector would indicate that the faulty distribution

is completely separated from the normal distribution. A poor detector that uses only random luck (similar to flipping a coin) would have equal detection and false alarm probabilities for all thresholds which would result in a straight line, splitting the ROC-graph into two equal halves and would give a lower limit for the performance of any (useful) fault detector. A scalar measure of the ROC is obtained by integrating the area under the ROC-curve, which is denoted the area under the curve (AUC). This gives an indication of the overall performance of the FDM, although only a part of the AUC may be informative about the relevant performance. For example, the partial AUC may be more relevant for comparing the performance of methods up to a given false alarm rate, than the complete AUC, see e.g. (Tian 2010). The power function (Casella and Berger 2001) is one additional probabilistic performance measures, which also considers the magnitude of the fault.

The main limitation with probabilistic measures is that they require detailed knowledge about the normal and faulty mode. How an inaccurately specified hypothesis (faulty or normal distribution) impact the performance assessment can be measured in terms of the Kullback-Liebler divergence between the true probability distribution $p(x)$ and the incorrect one $q(x)$

$$KL\big(p(x), q(x)\big) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \tag{1.11}$$

Equation (1.11) measures how much statistical power is lost due to an incorrectly specified hypothesis (Eguchi and Copas 2006). Note that (1.11) is the expectation of the logarithm of the likelihood ratio between the true probability distribution $p(x)$ and the incorrect one $q(x)$, which is zero when the assumed distribution is correct. The validity of the performance assessment derived from probability theory will therefore increase with an increasing number of representative samples.

An increasing number of samples will also improve the sensitivity in distinguishing a fault in the likelihood ratio in (1.4). This also relates to how the signal-to-noise ratio of normal and faulty distributions (the magnitude of the fault) impact the detection performance. The impact from signal-to-noise ratio can be realized from *Figure 1.5*, top graph, where a better separation between the normal and fault distribution is obtained for decreasing the noise variance and increasing the fault magnitude (bias). Similarly, for an increasing number of measurements, the uncertainty about the mode separation due to the variance in the distributions decrease, which can be understood as an increased (fault) signal-to-noise ratio.

The observation that the detection performance depend on both the fault magnitude and the number of measurements implies the inherent trade-off for FDMs - a fast detection time versus robust fault detection in terms of few false

alarms. Several time related measures have therefore been suggested, for example, the mean time to detection (MTD), the fault detection index (Carlsson and Zambrano 2016), and the average run length (ARL) between false alarms (Basseville and Nikiforov 1991, Gustafsson 2000). Hybrid measures that combine time to detection and the correctness of an alarm have also been proposed (Corominas et al. 2011).

The most prevalent performance assessment, however, only considers the frequency of correct fault detection rate (D), normal observations (N), incorrect classifications, false alarm rate (FA), and missed detection (MD). These frequencies are commonly summarized in a confusion matrix (*Figure 1.6*) and can be derived from the underlying normal and faulty distributions as previously illustrated in *Figure 1.5*.

		Predicted: Normal	Predicted: Faulty
Actual: Normal		N	FA
Actual: Faulty		MD	D

Figure 1.6. The binary confusion matrix.

Several meta-performance measures have been suggested, which are based on the confusion matrix content. Apart from the mentioned ROC and AUC, the precision-recall and F1-measure are prevalent in machine learning applications (Chen et al. 2020, Naseer et al. 2018).

So far, I have implicitly considered fault detection as a binary classification task. But fault detection can also include continuous variables, for example, a regression or a dynamic model simulation that is combined with a decision function in the decision space. Then, it can be relevant to assess the performance of the model, where conventional performance metrics such as the mean squared error and equivalent measures are applicable (Gustafsson 2000, Hauduc et al. 2015).

There is clearly a large toolbox available for performance assessment of fault detection methods. Their applicability depends on how detailed knowledge that is available about the studied system. For example, time-based measured require knowledge of the exact start and end time for the fault, i.e. complete knowledge about both the normal and faulty modes. This is available in simulations, but seldom in practice. In practice, such annotated data can be scarce, especially for rare faults. How such annotations can be facilitated by domain experts is a current research area (Lei et al. 2020, Russo et al. 2020). How these expert annotations can be conducted and obtained in an efficient

way is currently a bottleneck for performance assessment. Still, when annotated data are available, it is critical that they are representative and in sufficient amount to represent the true probability distribution, see *Figure 1.5*.

1.4 Contributions

This thesis is based on the following publications and manuscripts. The article content has been printed with permission from the publishers. Chapter 9 and 10 are my own conclusions from the results in Chapter 2-8. The Chapters 2-8 were, by contrast, a collaborative research effort by all co-authors. My contribution to these joint publications is described in the following section.

Peer-reviewed journal publications

Paper I: Samuelsson, O., Björk, A., Zambrano, J. and Carlsson, B. (2018) Fault signatures and bias progression in dissolved oxygen sensors. *Water Science and Technology,* 78(5), 1034-1044

Paper I provides evidence about biofilm progression on electrochemical- and fluorescent-based dissolved oxygen sensors in-situ, and how biofilm fouling impacts the measurement accuracy. Further, two annotated benchmark datasets were produced containing high-frequency (1 Hz) data for the dissolved oxygen impulse response due to automatic air-cleaning. Last, different symptoms of sensor faults were identified in the impulse responses. Paper I is contained in Chapter 3 with minor stylistic changes.

I was responsible for the conception, design, and execution of the experiments with subsequent data analysis and interpretation of the results with support from the co-authors. I drafted the original manuscript and revised it through collaboration with the co-authors.

I received great support from Simon Robertsson regarding execution of the experiments in Henriksdal WRRF and IVL Hammarby Sjöstadsverk R&D facility.

Paper II: Samuelsson, O., Zambrano, J., Björk, A. and Carlsson, B. (2017) Gaussian process regression for monitoring and fault detection of wastewater treatment processes. *Water Science and Technology,* 75(12), 2952-2963

In Paper II, the Gaussian process regression method is introduced in a wastewater context where the usefulness of sequential Monte Carlo hyperparameter approximation over the common maximum likelihood estimation for avoiding local optima is demonstrated. The study further emphasizes the logic to only consider advanced FDMs when required, as simple methods are

enough for simple problems. Paper II is contained in Chapter 5 with minor stylistic changes.

I was responsible for the conception, design, and execution of the experiments with subsequent data analysis and interpretation of the results with support from the co-authors. I drafted the original manuscript and revised it through collaboration with the co-authors.

Paper III: Zambrano, J., Samuelsson, O. and Carlsson, B. (2019) Machine learning techniques for monitoring the sludge profile in a secondary settler tank. *Applied Water Science,* 9(6), 146

In Paper III, the practical limitation with only normal training data is emphasized and illustrated with two probabilistic supervised machine learning methods on real data, which essentially relates to solving a one-class classification problem. The applicability of such methods in a realistic WRRF context was novel at the time for the study's initiation. Paper III is contained in Chapter 6 with minor stylistic changes.

I contributed to the conception, and interpretation of the results and suggested improvement on the manuscript drafted by the first author.

Paper IV: Samuelsson, O., Zambrano, J., Björk, A. and Carlsson, B. (2019) Automated active fault detection in fouled dissolved oxygen sensors. *Water Research* 166, 115029

Paper IV demonstrates the importance of gaining feasible (preferably automated) access to sufficient and representative normal data, both for fault detection tuning and performance assessment. The study introduces the active fault concept into a WRRF context by extending an existing conceptual model about data transformation steps related to fault detection. Paper IV is contained in Chapter 7 with minor stylistic changes.

I was responsible for the conception, design, and execution of the experiments with subsequent data analysis and interpretation of the results with support from the co-authors. I drafted the original manuscript and revised it through collaboration with the co-authors.

Paper V: Samuelsson, O., Olsson, G., Lindblom, E.U., Björk, A., and Carlsson, B. Sensor bias impact on efficient aeration control during diurnal load variations (*accepted for publication in Water Science and Technology*)

In Paper V, the impact from biased sensors used for automatic aeration control is assessed in a simulation study, which is facilitated by response surface methodology. Anticipated negative effects from sensor bias were verified and identified to be dependent on the sensor bias direction and magnitude. The study motivates sensor fault detection. Paper V is contained in Chapter 2.

I was responsible for the conception, design, and execution of the simulation experiments with subsequent data analysis and interpretation of the results with support from the co-authors. I drafted the original manuscript and revised it through collaboration with the co-authors.

Manuscripts submitted, or to be submitted, for journal publication

Paper VI: Samuelsson, O., Le, H. Q., Björk, A., Erikstam, S., Volcke, E.I.P., and Carlsson, B. Practicability of mass balance based data reconciliation for process rate monitoring (*manuscript, to be submitted for publication)*

Paper VI evaluates the trade-off between following theoretical assumptions and pragmatic ad-hoc settings related to the error covariance matrix and gross error detection in mass balance-based data reconciliation. A new approach for assigning the error matrix is further suggested and evaluated in a pilot plant experiment. Practical advice and future research directions are given to avoid pitfalls that have not been emphasized in the literature. Paper VI is contained in Chapter 4 with minor stylistic changes.

I and Hong Quan Le were jointly collaborating on the conception, and design of the experiments with subsequent data analysis and interpretation of the results, with support from the co-authors. I drafted the original manuscript and revised it through collaboration with the co-authors.

I received great support from the numerous people at Käppala WRRF regarding sensor installation, experimental execution, and conducting laboratory analyses.

Paper VII: Samuelsson, O., Björk, A., and Carlsson, B. Monitoring diffuser fouling with grey-box modelling (*submitted for publication)*

Paper VII demonstrates that the combination of simple process disturbances and mechanistic models is useful to assess the condition of aeration diffusers. Several unknown measurement disturbances were identified during the 18 months long full-scale evaluation of the proposed method. Paper VII is contained in Chapter 8.

I was responsible for the conception, design, and execution of the experiments with subsequent data analysis and interpretation of the results with support from the co-authors. I drafted the original manuscript and revised it through collaboration with the co-authors.

I received great support from Christer Persson (Bromma WRRF) who implemented the staircase method on the existing control system, and Anders Påhlsson (Bromma WRRF) who modified the existing sensor data collection routine.

Licentiate thesis
Samuelsson, O. (2017) Fault Detection in Water Resource Recovery Facilities. Licentiate Thesis, Uppsala University, Uppsala.

Part of the results have also been presented on the following conferences

Conferences publications
Samuelsson, O., Le, H. Q., Björk, A., Erikstam, S., Volcke, E.I.P., and Carlsson, B. (2019). Steady-state data reconciliation of a dynamic process. In: *WATERMATEX 2019, September 1–4, Copenhagen, Denmark*.

Samuelsson O., Olsson, G., Lindblom, E.U., Björk A. and Carlsson B. (2019). Sensor bias impact on controlling daily variations in wastewater loads. In: *WATERMATEX 2019, September 1–4, Copenhagen, Denmark*.

Samuelsson O., Zambrano J., Björk A. and Carlsson B. (2019). Automated active fault detection in fouled dissolved oxygen sensors. In: *Nordiwa 2019, September 23–25, Helsinki, Finland*.

Samuelsson O., Olsson, G., Lindblom, E.U., Björk A. and Carlsson B. (2019). Sensor bias impact on controlling daily variations in wastewater loads. In: *Nordiwa 2019, September 23–25, Helsinki, Finland*.

Samuelsson O., Björk A., Zambrano J. and Carlsson B. (2017). Monitoring of fouled DO-sensors with active fault detection. In: *12th IWA Specialized Conference on Instrumentation, Control and Automation (ICA), Quebec, Canada*.

Samuelsson O., Zambrano J., Björk A., Chistiakova T. and Carlsson B. (2015). Detecting anomalous air flow-ammonia load ratios, using Gaussian process regression. In: *WATERMATEX 2015, June 14–17, Gold Coast, Queensland, Australia*.

Zambrano J., Samuelsson O., Chistiakova T., Liu H. and Carlsson B. (2015). Gaussian process regression for monitoring a secondary settler. In: *2nd IWA Conference on New Developments in IT & Water, February 8–10, 2015, Rotterdam, The Netherlands*.

Zambrano J., Samuelsson O. and Carlsson B. (2016). Monitoring a secondary settler using Gaussian mixture models. In: *EUROSIM 2016, September 12–16, Oulu, Finland*, IEEE Computer Society.

2 Sensor bias impact on controlling daily variations in wastewater loads

This chapter highlights the need to increase our understand of the interplay between sensor drift and the performance of the automatic control system. The impact from biased sensors on the automatic control systems is rarely considered when different control strategies are assessed in water resource recovery facilities. Still, the harsh measurement environment with negative effects on sensor data quality is widely acknowledged. Simulations were used to show how sensor bias in an ammonium cascade feedback controller impacts aeration energy efficiency and total nitrogen removal in an activated sludge process. Response surface methodology was used to reduce the required number of simulations, and to consider the combined effect of two simultaneously biased sensors. The effects from flow variations, negatively biased ammonium (-1 mg/L) and suspended solids sensors (-500 mg/L) reduced the nitrification aeration energy efficiency between 7 and 25 percent. Less impact was seen on total nitrogen removal. There were no added non-linear effects from two simultaneously biased sensors, apart from an interaction between a biased ammonium sensor and dissolved oxygen sensor located in the last aerated zone. Negative effects from sensor bias can partly be limited if the expected bias direction is considered when the controller setpoint-limits are defined.

2.1 Introduction

Automatic control has developed to be an essential tool for balancing consistent treatment and energy efficiency (Olsson 2012). Substantial efforts have been devoted to develop different aeration control strategies (Åmand et al. 2013). Most control system studies, however, assume an ideal situation with accurate sensor measurements (Santín et al. 2016). In practice, on-line measurements are far from ideal. The water resource recovery facility (WRRF) constitutes a harsh measurement environment. This is generally recognized, and the commonly accepted standpoint is that biased measurements are widespread and have a negative impact on the desired control target. To our knowledge, this assumption has not been verified in studies although tools have been developed for that purpose (Rosén et al. 2008).

We consider accurate measurements and adequate automatic control to be increasingly important for four reasons. First, stricter effluent permits reduce the time when control can be out of specifications without violating the regulations. Second, retrofitted advanced treatment processes that need to operate together with existing processes makes control more elaborate and sensitive for consequential errors. Third, control is essential for a resource efficient treatment process. Last, automatic control can attenuate negative effects from increasing influent variations that are expected due to global warming. Therefore, we need to identify the most critical sensors and prioritize sensor maintenance to minimize bias impact as the number of sensors steadily increases. Ultimately, these aspects in combination can aggravate the operation and make it important to understand the impact from biased sensors on control.

Different methodologies can be used for studying the impact of biased sensors. Both full-scale and simulation experiments have been applied to assess the impact of biased DO sensors (Carlsson and Zambrano 2016). Full-scale studies are, however, time-consuming and impractical for assessing the effects of the combination of many biased sensors. It is also difficult (commonly impossible) to control the influent load, which would be needed to repeatedly assess whether the impact of biased sensors depends on different load conditions. Thus, simulation studies are preferred. Simulations also enable a precise interpretation (as interpreted within the model's predictive accuracy), without noise that can mask small effects on a full-scale plant. The widely used benchmark simulation model platform (BSM) (Jeppsson et al. 2006) is well suited for sensor bias evaluation.

Even a suitable model simulation will be time-consuming and difficult to evaluate when a vast number of simulation results are to be compared. As an example, the total number of combinations for ten biased sensors with three bias magnitudes (consider e.g. bias of -1, 0, +1) is $3^{10} = 59{,}049$. Thus, it is clearly a challenge to assess interaction effects between several biased sensors at different load scenarios also in simulation studies.

In this study, we adopt the response surface methodology (Myers et al. 2004), which limits the required number of simulations, but still enables to identify the key effects. The method is reliant on a representative set of simulations (Box and Behnken 1960) that are interpreted via linear regression coefficients.

The goal of this study is to assess the impact of sensor bias on ammonium cascade feedback control at different influent variations. The energy efficiency of applying this type of controller has been demonstrated in practice (Ingildsen et al. 2002, Rieger et al. 2014, Åmand 2014), but it is possible that biased sensors can reduce its advantage. Here, we study how different degrees of diurnal variations in the influent (flow and concentration variations) combined with sensor bias impact aeration energy efficiency and total nitrogen

removal. Bias in dissolved oxygen (DO), ammonium (NH) and suspended solids (SS) sensors are studied. The results show that the bias direction is critical and can be both beneficial and detrimental depending on the control target.

2.2 Material and Methods

This section describes the methodology (Section 2.2.1) and how the simulated system with related influent scenarios and sensors bias were defined (Section 2.2.2). The applied response surface methodology is described in Section 2.2.3.

2.2.1 Methodology

A dynamic process model was simulated with different sensor bias magnitudes and evaluated with respect to their impact on two process performance indicators, the energy efficiency of nitrification (NIT_{eff}), and total nitrogen removal (N_{rem}). Bias in five sensors (three dissolved oxygen, one ammonium and one suspended soldids sensor) and variations in influent flow rate and concentrations were studied at three different magnitudes. The combinatorial complexity, and likewise the required number of simulations, was reduced by applying a reduced factorial design (Box and Behnken 1960) and evaluated using response surface methodology (RSM) (Bezerra et al. 2008). The findings indicated by regression coefficients produced by RSM were further analysed for causal explanations by evaluating the simulation results in detail.

2.2.2 System description

The studied system was a dynamic model of a continuous activated sludge process (CAS) representing parts of the Henriksdal WRRF in Stockholm, Sweden (750,000 p.e.). The CAS consists of pre-denitrification followed by three aerated zones for nitrification and a final deaeration zone (*Figure 2.1*). The model is further described in (Lindblom et al. 2019)

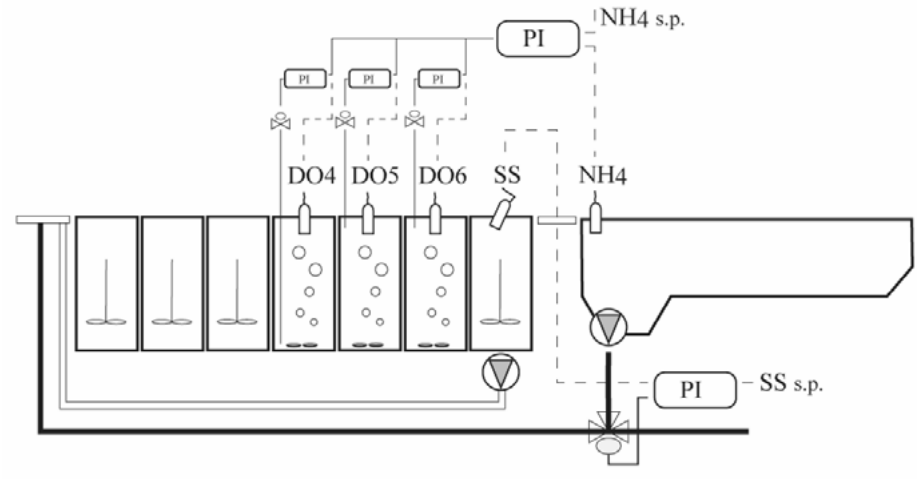


Figure 2.1. Process configuration for the simulated Henriksdal WRRF model with ammonium cascade feedback and suspended solids controller.

2.2.2.1 Controller configuration

The CAS air supply was controlled by an ammonium cascade feedback controller with DO PI-controllers operating as slave controllers under the master ammonium PI-controller (*Figure 2.1*). Each aerated zone had a separate slave DO-controller with equal setpoints provided by the master ammonium controller. This setting was developed by Åmand (2014) and is currently in use at Henriksdal WRRF. In practice, there is also an airflow rate slave controller for each DO-controller, which was excluded in this study.

The solids retention time (SRT) was controlled by adjusting the wastage of active sludge (WAS) with a PI-controller to obtain an average SS concentration of 2,500 mg/L. In practice, and in the studied model, this resulted in a variable solids retention time (SRT) of 16±3 days. The large SRT variability was due to the studied influence of the bias in the SS-sensor, which is described in Section 2.2.2.3. The SRT was sufficient to achieve near complete nitrification for the studied ammonium load and temperature of 14°C for the whole range of studied scenarios.

All PI-controllers were tuned with the Lambda method (Åström and Hägglund 1995). A relatively fast disturbance rejection rate was chosen to assure that the NH controller reacted on diurnal variations. The controller parameters obtained from the tuning are given in Table 2.1 with DO setpoint-limits.

Table 2.1. *Controller parameter settings used in the study.*

	K (-)	Ti (h)	Setpoint (mg/L)
DO4-DO6	1000	0.12	1.0/3.8 (min/max)
NH_4	-4	2.4	1.0
SS	0.67	360	2500

2.2.2.2 Process performance indicators

The two process indicators NIT_{eff} and N_{rem} were defined as

$$NIT_{eff} = \frac{mNHred}{\sum_{z=4}^{6} k_L a(z)} \tag{2.1}$$

$$N_{rem} = \frac{mNred}{mNtot,in} \tag{2.2}$$

where the prefix m refers to the mass of: removed total nitrogen ($Nred$), total nitrogen in the influent ($Ntot,in$), nitrified ammonium nitrogen ($NHred$) and $k_L a(z)$ refers to the total $k_L a$ (used as proxy for the mass of oxygen transferred from air to water) in zone z. In practice, the $k_L a$ depend on the current airflow rate and α factor. Both these factors are site specific and difficult to assess in practice. The results should therefore be interpreted in terms of an efficient aeration system (constant $k_L a$ per airflow rate, regardless of load situation, and new diffusers) and an influent free from surfactants that may reduce $k_L a$ and α.

2.2.2.3 Sensor bias magnitudes

The studied sensor bias magnitudes are given in Table 2.2, which are suggested to represent a reasonable bias due to, for example, a fouled sensor or inaccurate calibration. To clarify the notation: a negative bias means that the sensor value is lower than the true concentration, and vice versa. Drift in pH and DO sensors have been studied in (Ohmura et al. 2019, Samuelsson et al. 2018), but to our knowledge, bias in SS and NH sensors have not yet been estimated. It is important to assume bias magnitudes that may appear in practice. At the same time, it should be recognized that sensor bias also varies with site specific conditions and sensor maintenance.

Table 2.2. *The three sensor bias magnitudes (levels) considered in the study.*

Sensor	Bias (mg/L)
NH_4	-1/0/+1
DO4-DO6	-1/0/+1
SS	-500/0/+500

The assumed accuracy in an SS sensor is reflected by the accuracy for laboratory samples that are used during calibration. Here, the lab samples had an ± 20% analytical uncertainty at a 95% confidence interval. This would correspond to a bias of ± 500 mg/L at 2,500 mg/L concentration.

NH measurements can be conducted with ion-selective probes, gas sensitive and spectrophotometric analysers. The two latter are expected to have a higher accuracy than the former. The expected bias should be small in absolute terms if following the same reasoning about analytical uncertainty during calibration, as for the SS sensor. This is because the effluent NH concentration is expected to be low (0-3 mg/L) in nitrifying WRRFs. Our experience is, however, that ion-selective sensors can show substantial drift (Samuelsson et al. 2017), although the NH sensor drift was not explicitly assessed in that study.

In the end, we limited the NH bias magnitude to -1 mg/L. This is the largest possible bias before the sensor would measure a negative concentration. Similarly, the positive NH bias was limited to +1 mg/L to use the same magnitude in absolute terms, which facilitates interpretation of the results described in Section 2.2.3.

It was assumed a DO sensor bias of ± 1 mg/L. An approximate 1 mg/L negative bias was obtained after about one month of biofilm growth without manual cleaning on an electrochemical sensor. By contrast, an optical DO sensor showed a correspondingly large positive bias after 14 days of biofilm growth without manual cleaning and only automatic air cleaning (Samuelsson et al. 2018).

2.2.2.4 Influent variations

Three levels of variations were defined for both influent flow and concentrations: constant, normal and high influent variations (*Figure 2.2*). The purpose was to assess if bias in sensors is more critical at large variations.

The ‘normal diurnal influent’ and reference scenario was produced by combining measurements from the influent monitoring program with diurnal influent flow and load patterns to Henriksdal WRRF. The model by Gernaey et al. (2011) was used to estimate COD fractions required in the simulation model. Both influent concentration variations (nitrogen and carbon) and influent flow show similar diurnal patterns (*Figure 2.2*).

The constant variation scenarios were produced by either setting the influent flow or the influent concentration to be constant. The diurnal mass influents of both NH and COD were kept identical to the normal scenario by adjusting the influent concentration or flow mean value correspondingly.

The high variation scenarios were produced by stretching the normal variation by multiplying the dynamic diurnal profile with two, but still compensating its mean value so that an identical mass influent as the reference scenario was obtained (*Figure 2.2*).

Variations in flow and concentrations were treated as separate factors to assess if either of them would give a larger impact in combination with biased sensors.

The study was purposefully limited to consider only diurnal influent variations, neglecting other variations such as temperature. The reason was to learn about the sensor bias impact during the most common disturbances. For the Henriksdal WRRF, the diurnal variations represent the typical disturbance pattern for about 80% of the time. It is expected that impact from weekly and seasonal variations are similar to diurnal load variations, but with a changed mass load. Here, we assess interactions between variations in flow/concentration and sensor bias at a *fixed* diurnal mass load, to allow a fair comparison. This would be slightly different if seasonal and weekly variations were to be included.

The impact from rain and drainage water can be substantial, both in terms of flow and impact on the wastewater temperature. How such stormwater events impact the WRRF will, however, be very site specific, and therefore difficult to generalize. We also expect a large (possibly the largest) negative effect on the settler operation. These conditions are difficult to model and would introduce uncertainty to the result interpretation. The results here assume a biological treatment process with good settling properties. For these reasons, we limited this study to daily normal variations that still represent the main time of operation.

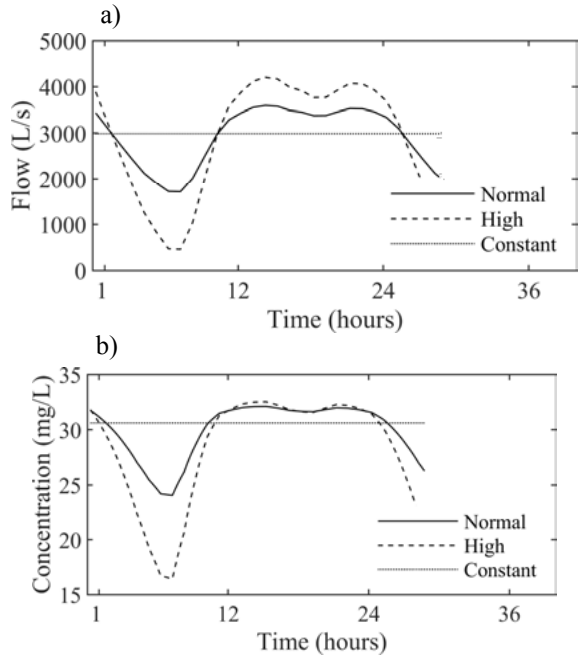


Figure 2.2. Influent patterns for flow (**a**) and ammonium (**b**). Different line types indicate variation levels: normal (solid), high (dashed) and constant (dotted).

2.2.3 Response surface methodology

RSM originally gained interest as a tool for industrial product and process optimization, and is iterative in its nature (Myers et al. 2004). The RSM applied here followed four steps

1. Design experiment
   a. Define factors, their levels and output variables
   b. Define which combination of factors and levels to evaluate
2. Execute experiment
   a. Execute simulations as specified in 1b
3. Construct a valid regression model
   a. Assess model quality
   b. Remove insignificant model terms
   c. Iterate a-b) until the model only contain significant factors
4. Interpret the results
   a. Interpret significant model terms (regression coefficients)
   b. Analyse causal reasons for the results indicated in 4a

Especially steps 3 and 4 are commonly iterated to obtain a final model. Conclusions obtained in step 4 often induce additional experiments, reinitiating the four-step procedure all over again.

The experimental design and evaluation were conducted in MODDE 12.1 (Sartorius), a software for experimental design. The model simulations and evaluations were conducted in MATLAB/SIMULINK version R2020a (MathWorks).

2.2.3.1 Step 1. Design of experiment

Three levels were considered for the seven quantitative factors: sensor bias (factor 1-5, Table 2.2) and influent variations (factor 6-7, constant/normal/high flow and concentration variations). Three levels are required to identify any quadratic effects, which would here require $3^7 = 2187$ combinations for a full factorial design. The reduced design suggested in (Box and Behnken 1960) was used, which required only 57 combinations. The Box-Behnken design was preferred over the more common central composite design since extreme points are excluded (Bezerra et al. 2008). Extreme points refer to combinations where all factors are at their minimum or maximum levels. Here, these extreme points would correspond to a situation when all sensors have, for example, a positive bias simultaneously as the influent flow and concentration variations are high. This is not expected to be common in practice, and the exclusion of those combinations was therefore not expected to influence the applicability of the results. The dependent variables (responses) NIT_{eff} and N_{rem} were defined in (2.1) and (2.2), respectively.

2.2.3.2 Step 2. Simulation procedure

The model was simulated with the different combinations of factor settings described in Step 1. For each factor setting, the model was first simulated to reach steady state with a constant load for 60 days. Next, the same factor settings were simulated for an additional 120 days with the dynamic influent, which allowed the SS to reach steady state. The last seven days were then used for evaluation.

2.2.3.3 Step 3. Construct a valid model

An initial linear regression model was defined to contain all linear, quadratic, and two-factor interaction terms. Nonsignificant model terms measured by the corresponding confidence intervals were removed. This procedure was repeated until the final model with only significant model terms remained.

The final model was assessed by verifying a large (>0.8) Q^2, which measures the regression model's predictive ability, in contrast to the common R^2, which only measures the explained variation in the output. Both measures are used to assess the model validity. Golub et al. (1979) defined Q^2 as

$$Q^2 = 1 - \frac{\sum_{n=1}^{N} \left(\frac{y_n - \hat{y}_n}{1 - x_n (X^T X)^{-1} x_n^T} \right)^2}{\sum_{n=1}^{N} (y_n - \bar{y})^2} \quad (2.3)$$

where x_n are the inputs that consists of seven factors (sensor bias and influent variation) and y_n are the observed outputs NIT_{eff} and N_{rem} for the n dynamic model simulations. X is a n-by-7 matrix that contain the factor settings for the n dynamic model simulations. $\bar{y}$ is the mean of all observed outputs (dynamic model simulations) and $\hat{y}_n$ is the regression model's predicted output for the corresponding simulation n. In total, N dynamic model simulations were performed with the bias and influent settings obtained from the Box-Behnken design. The regression was conducted by regressing the observed values (dynamic model simulations) on the Y-axis for the predicted values on the X-axis (regression model predictions) as suggested by Piñeiro et al. (2008).

2.2.3.4 Step 4. Interpret the results

The regression coefficients from the final model were analysed with causal analysis from the dynamic simulation results.

2.3 Results

First, an overview of the results is given, expressed as the obtained regression models where the validity of the regression models is also analysed. Then, the impact of biased sensor measurements on the process indicators are interpreted. Last, the effects from changes in influent variations and interactions are analysed.

2.3.1 The obtained regression model and its interpretation

The simulation results resulted in two different regression models with six and ten significant model terms for NIT_{eff} and N_{rem}, see *Figure 2.3(a)* and *Figure 2.3(b),* respectively. Changes in the influent flow (F) were significant in contrast to influent concentration variations that did not have a significant effect on any of the process indicators. In general, the impact on NIT_{eff} was larger than the impact on N_{rem} (*Figure 2.3*, note the different scales). The largest effects were seen from biased NH and SS sensors that included both significant linear and quadratic model terms. Interaction effects with other factors were seen for the ammonium sensor and are further analysed in the next section.

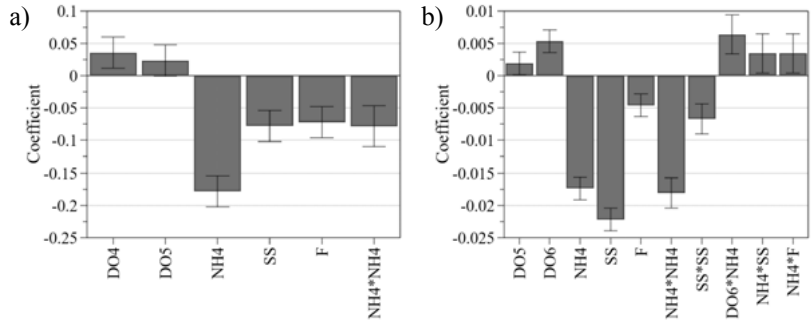


Figure 2.3. Coefficients of the final regression models for NIT_{eff} (**a**) and N_{rem} (**b**), including 95% confidence intervals. Bias in ammonium sensor (NH), suspended solids sensor (SS), and dissolved oxygen sensors (DO) located in zone 4-6 (denoted with suffix 4,5,6).

The regression models were obtained by removing non-significant model terms as described in Section 2.2.3. Both models showed a good predictive fit (Q^2_{NITeff} =0.82; Q^2_{Nrem}=0.94) and were assumed valid for further analysis. The agreement between predicted and simulated values is shown in *Figure 2.4*. It is worth noting that the reference scenario showed a larger value for NIT_{eff} than the model prediction (square, *Figure 2.4*). This will not violate the conclusions made from the regression model. However, it highlights the fact that certain combinations of levels and factors will produce results that deviate from the general conclusions, which ought to influence the interpretation of the results. We therefore complement the regression model analysis with a causal model evaluation of the mass flows and time series in the underlying simulation results where non-obvious explanations can be made.

To avoid confusion about how to interpret the bias direction (positive or negative) and the signs and magnitudes of the regression coefficients we here exemplify how to read *Figure 2.3*.

A positive regression coefficient in *Figure 2.3* indicates an increase in the performance indicator when the corresponding model term has a positive value. For example, a positive bias in DO5 will result in an increased NIT_{eff} and N_{rem} since the regression coefficients for DO5 are positive (yet small) in *Figure 2.3*. Note that a *positive bias* in DO5 will result in a *lower* DO concentration than the desired setpoint. Similarly, negative coefficients for a biased NH sensor indicate the opposite relationship. That is, a negative bias in the ammonium sensor (and likewise an increased effluent ammonium) will increase NIT_{eff} and N_{rem} as indicated in *Figure 2.3*.

The magnitude of a regression coefficient for a main effect should be interpreted as - the impact on the process indicator for a unit change in the factor related to the coefficient, while keeping remaining factors constant. Here, a unit change is defined as a 1 mg/L bias for DO and ammonium sensors, and 500 mg/L bias in the SS sensor (Section 2.2.2). As an example, the regression

coefficient for the NH sensor main linear effect is -0.17 (*Figure 2.3(b)*). This indicates that a 1 mg/L ammonium sensor bias is expected to reduce NIT_{eff} with about 17 percent on average from the linear effect.

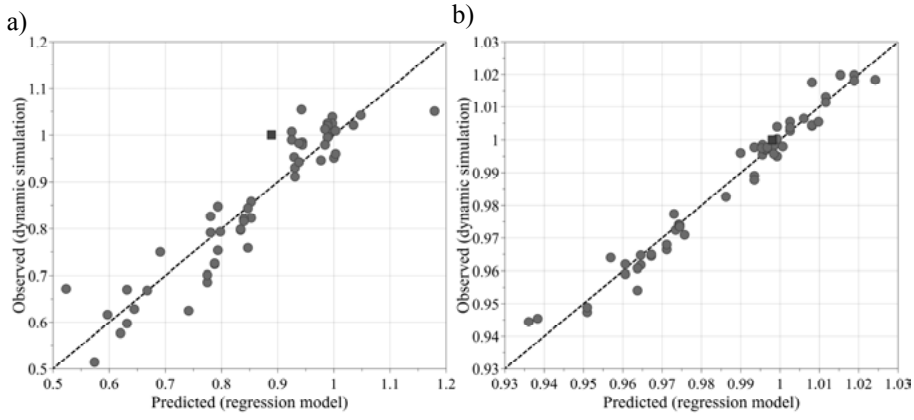


Figure 2.4. Observed (simulated) and predicted values for NIT_{eff} (**a**) and N_{rem} (**b**) from the RSM model (note that 'Predicted' refers to the regression model prediction and 'Observed' are the computed values from the dynamic model simulations). Each circle represents one simulation with a combination sensor bias magnitudes and influent variations given by the Box-Behnken experimental design. Interaction effects between several biased sensors can be assessed since the biased sensors are not evaluated one-at-the-time. Squares are the reference simulation with normal influent variations and no sensor bias.

2.3.2 Ammonium sensor bias

A bias in the ammonium sensor had the largest influence on both NIT_{eff} and N_{rem}, as remarked in Section 2.3.1. When both quadratic and linear effects are added (as a straightforward summation), a bias in the NH sensor showed the largest impact on NIT_{eff} of all factors. The reduction in NIT_{eff} was about 25 percent for a 1 mg/L bias as predicted by the regression model (*Figure 2.3*). A large effect from bias in the NH sensor was expected, since it is the key information used in the NH-controller. This reduction in energy efficiency is larger than what is expected to be gained from the ammonium cascade feedback controller in the first place (compared to DO-controllers with fixed setpoints), see (Åmand 2014). This emphasizes the importance of unbiased measurements to achieve the desired benefits from automatic control.

The reason for the reduced NIT_{eff} was that a positive NH sensor bias will result in a lowered effluent NH concentration that requires more aeration energy per mass nitrified nitrogen. In effect, the 'true' NH-setpoint is 0 mg/L at 1 mg/L bias and 1 mg/L NH-setpoint. To reach such low NH effluents, extensive aeration is required. The NH-controller solved this by assigning high DO-setpoints close to, or at, the maximum setpoint-limit during peak loads, with

a poor NIT_{eff} as consequence. The opposite reasoning can be applied to a negative NH bias.

A positive bias in the NH sensor did not only have a negative effect on NIT_{eff}, but also on N_{rem} (*Figure 2.3(b)*). Again, high DO-setpoints were the reason for the reduced nitrogen removal. High DO-setpoints cause a higher DO concentration in the last unaerated zone and the oxygen is recirculated to the pre-denitrification (*Figure 2*) resulting in a decreased pre-denitrification rate and N_{rem}. The same effect on N_{rem} was seen for a negative bias in DO6 (higher DO than desired). The effect was even more pronounced when NH and DO6 simultaneously had a negative bias, which was amplified through their interaction effect (*Figure 2.5*).

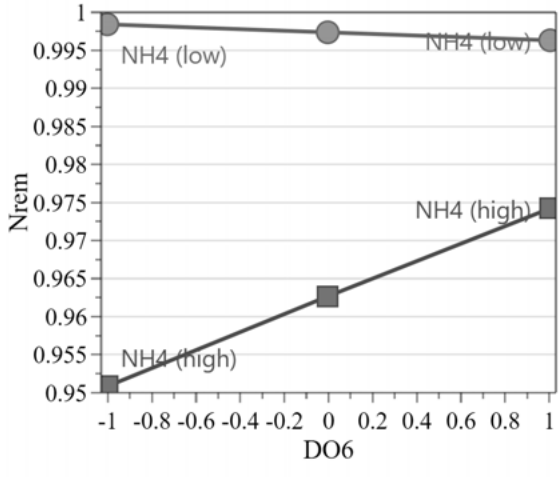


Figure 2.5. Interaction between NH and DO6 on N_{rem}. When DO6 have a negative bias (- 1 mg/L) simultaneously as NH sensor has positive bias (NH4 (high)) there is an additional decrease in N_{rem}.

A negative bias in the NH sensor (lowered DO-setpoints) *also* led to a decrease in N_{rem}. The reason for this was that less nitrification also lowered the total nitrogen removal rate as less nitrate became available for denitrification. Note that both NH bias directions had a negative effect on the total nitrogen removal. This raises the question whether the NH-setpoint was optimal with respect to the N_{rem} requirement. The possibility of an optimum is further supported by the presence of quadratic NH model term (*Figure 2.3(b)*). The consequences of a biased NH sensor for NIT_{eff} and N_{rem} are demonstrated in *Figure 2.6*. In *Figure 2.6(b)*, the modelled main effect (linear and quadratic) for NH is shown for different bias magnitudes, which indeed shows an optimum at -0.5 mg/L NH bias. The explanation for the optimum is that for this specific WRRF configuration and load, a certain amount of nitrification is needed to obtain a low (< 0.5 mg/L) DO in the recirculation stream. This occurs at a NH

effluent concentration of 1.5 mg/L, i.e. the optimum in *Figure 2.6(b)*. If the NH effluent is larger than 1.5 mg/L, the minimum DO-setpoint at 1 mg/L instead increases the recirculated DO concentration.

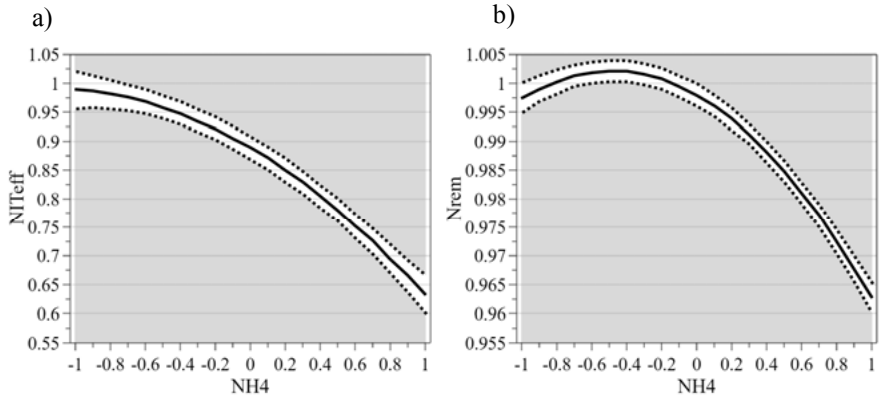


Figure 2.6. Main effects of a biased NH sensor (± 1 mg/l) including both linear and quadratic model terms for NIT_{eff} (**a**) and N_{rem} (**b**). 95% confidence intervals are indicated with dashed lines.

2.3.3 DO sensor bias in the ammonium loop

Bias in the DO sensor in cascade with the ammonium controller cannot be neglected. As noted in Section 2.3.2, increased DO concentrations reduced both NIT_{eff} and N_{rem}. Similarly, a negative bias in any DO sensor (higher DO than measured) also resulted in reduced NIT_{eff} and N_{rem} (*Figure 2.3*). It was initially expected that the NH master controller would compensate for any bias inside the DO slave controller. After all, a positively biased DO sensor causing an increase in NH effluent should be possible to compensate for with an increased DO-setpoint. The mistake in the previous reasoning is that the setpoint-limits of the slave controllers were not considered. An increase in both the minimum and maximum setpoint-limits due to negative bias in DO4 will be unfavourable as shown in *Figure 2.7*. For example, during low loads at night, the minimum true DO concentration is 2 mg/L instead of the desired 1 mg/L minimum setpoint.

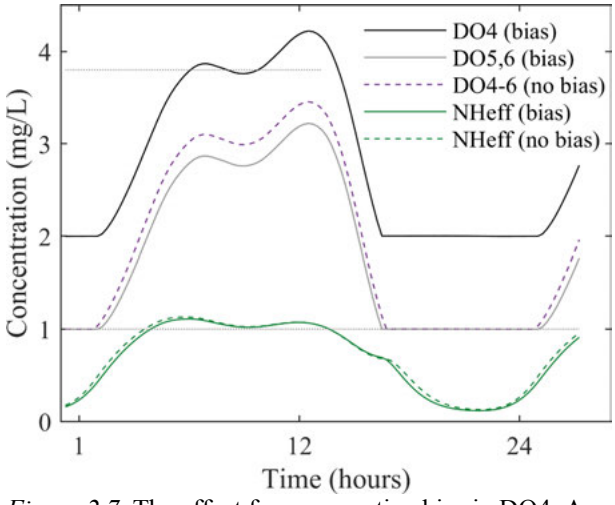


Figure 2.7. The effect from a negative bias in DO4. Ammonia effluent (green solid and dashed lines) and DO setpoints (black, grey purple dashed lines) during a diurnal cycle. Dotted grey lines indicate DO setpoint-limits at 1 mg/L and 3.8 mg/L respectively.

It should be noted that the location of the biased DO sensor matters, since the impact direction was different for NIT_{eff} compared to N_{rem} (*Figure 2.3*). Bias in the DO4 sensor had the largest effect on NIT_{eff}, which contrasts N_{rem} and the results obtained with a biased DO6. This is in agreement with previous studies of the optimal DO-profile, where a lower DO in first zone was shown to be the most energy efficient (Åmand and Carlsson 2013).

2.3.4 Suspended solids sensor bias

A negative 500 mg/L bias in the SS sensor resulted in a moderate increase in NIT_{eff} (about 6 percent) and N_{rem} (about 2 percent), see *Figure 2.3*. This indicates that there would only be positive effects from an increased SS concentration. This is logical as the biomass is the limiting factor for reducing ammonium peaks (Rieger et al. 2014). We acknowledge that the key limitation for reaching a high SS is the sedimentation capacity, which is not fully described in the applied model. Especially operating issues related to the bacterial sludge composition are not considered. Still, the results highlight the importance to strive for the maximum *practically feasible* SS-concentration. We further expect that the impact from a biased SS sensor, on N_{rem}, to increase if the activated sludge process is operating close to its maximum capacity where occasional ammonium break-through is expected. This would be especially critical during cold wastewater temperatures where a higher SRT is needed to maintain nitrification. Here, the NH load was within the modelled plant nitrification capacity even for the reduced SS concentration.

2.3.5 Interactions between sensors and influent flow variations

As expected, a decrease in influent flow variations also led to an increase in NIT_{eff}. It was less expected that the effluent NH concentration decreased with increasing influent variations. At a first glance, this is non-intuitive but is reasonable since the opposite applied for aeration energy – the aeration energy increased with increased influent variations. The reason is that the NH-controller increases the DO-setpoint during peak loads, resulting in an increased nitrification at the cost of increased energy consumption. The effluent ammonium concentration was below the setpoint of 1 mg/L, apart from simulations with a constant influent mass flow. This was not a consequence of poor tuning of the controller but caused by the DO-setpoint-limits. During low loads at night, the ammonium effluent concentration approaches zero, which should produce a zero setpoint value to the DO-controller. However, as there was a minimum DO-setpoint value of 1 mg/L, the ammonium concentration remains below the setpoint value until the load increases and justifies a DO-setpoint of 1 mg/L.

Note that in contrast to the effect from influent *flow* variations, there were no significant effects from influent *concentration* variations (*Figure 2.3*). A possible reason is that concentration changes are within the NH-control authority and that the nitrification rate can temporarily be increased with increased aeration. Such action, however, is not enough to compensate for short hydraulic retention times due to flow variations. This needs to be verified by further studies.

There were no significant interaction terms between influent variations and sensor bias (apart from a minor interaction term between NH and F, *Figure 2.3(b)*). This indicates that avoiding biased sensor does not become more important when influent variations increase, compared to when the influent flow is constant. The results, however, may be different during storm water conditions, or when a sudden load increase temporary exceeds the plant treatment capacity.

Apart from the interaction between bias in NH and DO6 mentioned in Section 2.3.2, there were no large interactions between a pair of two biased sensors. Thus, we should expect that the combination of several biased sensor will not be more problematic, compared to the problems caused by bias in the individual sensors, one at the time. The lack of non-linear interaction effects opposes the common assumption of the non-linear nature of wastewater treatment models. Instead, the result indicated fairly linear changes for the studied scenarios, which can simplify and reduce the needed scope for future studies.

2.4 Discussion

The consequences of sensor bias are discussed in wider context, including relation to costs in Section 2.4.1. The impact from bias direction and the interplay with controller setpoint-limits are discussed in Section 2.4.2 and Section 2.4.3, respectively. The impact on effluent permits is analysed in Section 2.4.4. The applicability of the RMS method is evaluated in Section 2.4.5. Finally, possible mitigating actions are considered in Section 2.4.6.

2.4.1 Sensor bias impact on costs

A negative bias in the NH sensor will cause an increase in electricity costs due to the reduced NIT_{eff}. For the studied 750,000 p.e. WRRF, a 1 mg/L NH sensor bias (consistently during a full year) would correspond to a substantial annual cost increase, equivalent to employing eight full-time instrument technicians (!). Similarly, the added energy cost for a biased DO sensor would correspond to one additional instrument technician, despite the small change as measured in percentages. The motivation for sensor maintenance and purchasing the best available NH sensor is obvious.

In practice we should only expect biased sensors for part of the time. The probability of having a biased sensor is not easy to estimate as it would require a redundant and accurate reference sensor. A cost benefit analysis of condition-based sensor maintenance, in contrast to the current time-based sensor maintenance, would motivate further studies about the probability of biased sensors in practice.

The results demonstrated that cost and energy reduction enabled by advanced automatic control, can be easily lost using inaccurate on-line sensor measurements. This is rarely considered during benchmarking of new control strategies. Therefore, we suggest a critical review of aeration control strategies, with respect to their sensitivity towards biased sensor measurements.

2.4.2 Sensor bias direction matters

Only a negative bias in DO sensors and positive bias in the NH and SS sensors had a negative impact in NIT_{eff} and N_{rem}. It is problematic with an undesirable negative DO-bias since the common problem with biofilm formation on electrochemical membrane-based DO sensors has been shown to cause such negative drift (Samuelsson et al. 2018). In Samuelsson et al. (2017), there were indications of a negative drift in an ion-selective NH sensor, although the drift direction was not studied in detail. In general, research about sensor drift direction in practice has been limited to a few studies (Ohmura et al. 2019, Samuelsson et al. 2018, Thürlimann et al. 2019). For that reason, we also lack knowledge about whether different sensor technologies result in different drift direction, for example caused by fouling. The findings here emphasise that

knowledge about sensor drift direction is essential. Further studies are needed, especially for NH and SS sensors.

Apart from the bias direction, it was remarked in Section 2.3.3, that the location of a biased DO sensor had an impact on NIT_{eff} and N_{rem}. This knowledge should be considered when sensor maintenance routines are developed to prioritize the sensor maintenance order. During such prioritization, large interactions between two biased sensors should also be considered, which would here apply to NH and DO6.

As mentioned, it is still not fully understood whether drift direction can differ between different sensor technologies, and even between different sensor makes. If so, the expected drift direction should be considered already during the sensor procurement. The expected drift direction would be essential product information.

2.4.3 The interplay between controller setpoints and sensor bias

Ammonium cascade feedback control is reliant on DO setpoint-limits to avoid undesirable DO concentrations. The influence from how these setpoint-limits are assigned increases when sensor bias is considered. The common strategy is to assign tight setpoint-limits, e.g. limit DO between 1 and 2 mg/L. This will avoid unfavourable excess aeration during peak loads at the cost of reduced disturbance rejection rate. Tight limits will also reduce the impact from biased sensors as a negative 1 mg/L DO bias in practice instead would lead to setpoint-limits between 2 and 3 mg/L. Using only a fixed DO-setpoint of e.g. 2 mg/L would minimize the influence from a DO bias to +/- its magnitude. A better approach would be to tighten only the setpoint-limit that will be affected by the expected drift direction. For example, an expected negative drift direction for the DO sensor would lead to too high DO concentrations. A slightly lower maximum DO setpoint of e.g. 1.5 mg/L could counteract unnecessary aeration in the presence of bias, at the cost of reduced disturbance rejection capability.

In this study a proper anti-windup has been applied. Problems related to absence of anti-windup are expected to increase with biased sensors. The reason is that the NH-controller would operate at its setpoint-limits during longer periods due to the sensor bias.

2.4.4 Sensor bias impact on effluent permits

Sensor bias will have an impact on achieving effluent permits. Many WRRFs have permits for the maximum effluent ammonium concentration. The timescale for this maximum differs between countries and WRRFs. In Sweden, yearly or monthly maximum mean values are prevalent, but other countries require appliance for shorter timescales. This will influence which amplitude and time period that can be accepted with sensor bias, while still satisfying

effluent requirement. Thus, the importance of sensor bias is clearly both WRRF and regulatory specific and the methodology applied here could readily be extended to include such aspects.

Seasonal variations, such as cold wastewater temperature during wintertime in Sweden, could impact the nitrification efficiency substantially. When approaching the minimum SRT, the consequence of a biased SS-sensor would be of increasing importance. Similarly, an (undesired) reduced aeration due to bias in either a DO or an NH-sensor could then become more critical than was observed here. A relevant future study would therefore identify controller setpoints that lies in the borderline for critical SRT, minimum temperature and maximum influent flow. We expect that the optimization tools in RSM would be feasible for this purpose and could be an extension of the methodology applied here.

2.4.5 Benefits and risks with response surface methodology

The main benefit with using response surface methodology (RSM) compared to evaluating scenario by scenario is that a good overview of the key influential sensor bias can be obtained with less effort. This can guide new rules-of-thumb that can be used in practice, for example, “avoid negatively biased NH and SS sensors both for cost and nitrogen removal reasons”.

One drawback of the RSM is that the obtained regression coefficients only indicate the average effect, and that there may exist combinations of biased sensors that produce results deviating from what the regression model predicts. This risk increases when a full factorial experiment is reduced by e.g. Box-Behnken design which was applied here.

The regression coefficients cannot be interpreted separately but require a causal interpretation from the simulations. Otherwise, the possibility to transfer the insights to similar systems will be limited.

The RSM methodology resembles a sensitivity analysis. A sensitivity analysis for optimizing controller setpoint values could also have been used to analyse sensor bias impact. Thus, the dual goal of process optimization and critical sensor analyses can be performed simultaneously and would likely increase the motivation for executing similar studies in practice.

2.4.6 Preventive actions to mitigate negative effects from biased sensors

Based on the results from this simulation study some practical advice can be given:

1. The most common drift direction and expected magnitude should be assessed for the current plant conditions and seasonal variations;

and sensor makes. This will make it easier to identify critical problems in practice. If simulations are to be conducted, this will also reduce the need to simulate and interpret non-existing sensor bias combinations.

2. The impact from the DO controller's setpoint-limits should be studied and assigned while considering the expected sensor drift direction.
3. A multi-criteria analysis should be conducted to identify which of the expected harmful effects from biased sensors that are the most important. A trade-off between treatment costs, treatment efficiency, and achieving effluent permits ought to be identified. From such analysis, the most critical sensor(s) with respect to bias can be identified. Consequently, the maintenance of these sensors should be prioritized.
4. It could also be possible to facilitate detection of biased sensors by transferring knowledge about fault symptoms indicated in the simulations to the operator or fault management system. For example, if a positive drift in the NH sensor is expected to produce a higher DO-setpoint than desired, then the operator should monitor the duration of maximum DO-setpoints during peak loads, as they will be affected by such bias.

2.5 Conclusions

There is an obvious need to assess the reliability of on-line sensor data used for automatic control. This aspect is not commonly included in control system benchmarking but is critical to assure that the real system is optimized at realistic conditions. It is concluded that:

- Biased sensors and influent variations considered as separate factors have a large impact on nitrification energy efficiency and less impact on total nitrogen removal. The impact from biased sensors do not however increase as influent diurnal flow and concentration variations increase.
- To implement preventive measures, it is important to know the expected sensor bias direction. Positive bias in NH and SS sensors and negative bias in DO sensors should be avoided to maintain a high total nitrogen removal and energy efficient nitrification.

2.6 Chapter acknowledgements

We gratefully acknowledge discussions and recommendations from Håkan Fridén at IVL Swedish Environmental Research Institute concerning the RSM methodology and how to avoid the one-at-the-time approach. The M.Sc. thesis by Ahlström (2018) was an inspiration for continuing the research on the topic of sensor bias impact on control.

3 Fault signatures and bias progression in dissolved oxygen sensors

Biofilm fouling is known to impact the data quality of sensors, but little is known about the exact effects. In this chapter we study the effects of artificial and real biofilm fouling on dissolved oxygen (DO) sensors in full-scale water resource recovery facilities, and how this can automatically be detected. Biofilm fouling resulted in different drift direction and bias magnitudes for optical (OPT) and electrochemical (MEC) DO sensors. The OPT-sensor was more affected by biofilm fouling compared to the MEC-sensor, especially during summer conditions. A bias of 1 mg/L was detected by analysing the impulse response (IR) of the automatic air cleaning system in the DO sensor. The IR is an effect of a temporal increase in DO concentration during the automatic air cleaning. The IRs received distinct pattern changes that were matched with faults including: biofilm fouling, disturbances in the air supply to the cleaning system, and damaged sensor membrane, which can be used for fault diagnosis. The results highlight the importance of a condition based sensor maintenance schedule in contrast to fixed cleaning intervals. Further, the results stress the importance of understanding and detecting bias due to biofilm fouling, in order to maintain a robust and resource efficient process control.

3.1 Introduction

The dissolved oxygen (DO) concentration is a key measured variable in Water Resource Recovery Facilities (WRRFs). Two measurement technologies are common, membrane electrochemical (MEC) and optical fluorescent (OPT) measurement techniques. The MEC type was originally described in (Clark 1959) and the OPT type was introduced by Demas et al. (1999).

On-line DO measurements have been commercially available since the 1970s and have enabled the development of automatic DO control (Olsson et al. 2005). Typically, the DO concentration is maintained at different DO set-points at different zones using multiple DO sensors to enhance biological nitrogen removal. Moreover, multiple parallel treatment lines result in a large total number of installed DO sensors. As an example, the two WRRFs in this

study, Bromma (Sweden) WRRF (about 300,000 p.e.) and Henriksdal (Sweden) WRRF (about 750,000 p.e.) make use of 30 and 21 DO sensors, respectively.

Regardless of measurement technology, a sensor has to be clean to provide accurate measurements. Inaccurate measurements used in a feed-back control loop may result in an undesired DO concentration and potentially reduced treatment efficiency or unnecessary aeration (with associated increasing costs). Therefore, visually inspecting the sensor, its manual cleaning, and readings verification are needed on a regular basis to guarantee accurate readings. Current sensor verification practice, as in (International Organization for Standardization 2012), rely on assessing sensor readings under predefined conditions. For dissolved oxygen sensors, oxygen free water solutions and water saturated air are commonly used. Such actions are however time-consuming, especially for WRRFs with many and possibly remotely located DO sensors. In sewage measurements, specific systems have been developed to resist biofilm fouling and pro-long maintenance intervals (Li et al. 2017). The need for prolonged sensor maintenance is also highlighted in (Thürlimann et al. 2018), where a qualitative soft-sensor approach was developed to replace an ammonium sensor with high maintenance requirements.

There are multiple sensor fouling sources in WRRFs that originate both from the influent wastewater stream and the treatment process itself. The fouling sources include: solids deposition (biofilm formation, chemical precipitation, sludge, and plastic products), hair and fibres, and grease (WEF 2013). In this study, we consider DO sensors located in the activated sludge process where the main fouling substance is from biofilm formation. In the following we use the term biofilm fouling to distinguish fouling from biofilm formation and other fouling sources.

In WRRFs, most sensor manufacturers provide optional automatic air cleaning that extends the required time interval for manual cleaning. The air cleaning results in a temporal increase in the DO concentration, which we further denote as an air cleaning IR. Andersson and Hallgren (2015) showed that the IRs contain information about the degree of biofilm fouling. A similar approach was suggested by Spanjers and Olsson (1992), where a change in the time constant of the DO sensor was shown to be a good indication of an artificially fouled DO sensor.

Although there are reasons to believe that the IR of an automatic air cleaning system can be used to detect biofilm fouling, we lack knowledge about the robustness and sensitivity of using the IRs to detect different levels of biofilm thickness. Andersson and Hallgren (2015) detected a bias in an OPT-sensor of -0.6 and -0.8 mg/L due to biofilm fouling during two one-month experiments. Additional experiments are required to study the IRs during clean and fouled conditions in order to extend the knowledge about IRs for biofilm fouling detection. In this study, we were therefore interested to investigate:

- The lowest detectable bias due to biofilm fouling using the response time method
- The variations in IR patterns of repeated biofilm fouling procedures and at different process conditions
- Whether both MEC- and OPT type DO sensors are applicable to bias detection with IRs

It is a common assumption that biofilm formation on a DO sensor affects its readings. As an example, Yoo et al. (2008) assumed that sludge clogging of the DO sensor can result in a complete sensor failure as a part of a simulation study. Hsu and Selvaganapathy (2013) used yeast and nutrient mixture in laboratory experiments to compare the effect of biofilm growth (in terms of protein absorption) on Teflon coated DO sensor membranes with silicone rubber alternative. Janzen et al. (2007) found indications of negative drift due to biofilm fouling in a redesigned MEC-sensor positioned in the ocean. Carlsson and Zambrano (2016) studied how bias in DO-sensors could be detected by comparing the air-flow ratios. However, there are few studies, apart from (Andersson and Hallgren 2015), that have investigated the effect of biofilm fouling on DO sensors under real conditions in WWRFs. Our general understanding about the effects of biofilm fouling is further complicated by: WRRF-specific process conditions, seasonal variations, and differences between sensor types and brands. These aspects are fundamental to understand in order to design a cost-effective sensor maintenance schedule. Therefore, we were also interested to study the bias progression due to biofilm fouling under real conditions for MEC- and OPT type DO sensors.

In this study, we conducted long-term experiments under full-scale conditions with artificial and real biofilm fouling with two sensor techniques: MEC- and OPT DO sensors. Further, we studied the applicability of the response time method to detect biofilm fouling and worn out sensor membranes.

3.2 Materials and Methods

Two experiments were conducted, one with artificial biofilm fouling (grease) and one with real biofilm fouling. The purpose of evaluating both artificial and real biofilm fouling was to study two aspects: variation in IRs during different process conditions and long-term time effects on IRs and bias from biofilm fouling. The two aspects require different studies since long-term studies are difficult to repeat under different conditions. Both aspects need to be considered to evaluate the usefulness of IRs for fouling detection. The experimental set-ups differed between the two experiments and are detailed in the following two sections.

3.2.1 Experiment 1 - Artificial biofilm fouling experiments in Henriksdal WRRF

The first experiment was conducted in Henriksdal WRRF with artificial biofilm fouling to investigate:

- The shape change of a fouled sensor's IR at a small bias (<0.2 mg/L), i.e. the detection sensitivity of using IR for biofilm fouling detection.
- The effect of different process conditions on the IRs, including varying DO and suspended solids concentration.
- Variation in the response time during clean and fouled conditions during repeated experiments.
- The difference between MEC and OPT-sensors in terms of a response to artificial biofilm fouling.

The artificial biofilm fouling experiments were carried out in Henriksdal WRRF during summer conditions with a wastewater temperature of about 19 °C. Four locations in the activated sludge process (ASP), both in aerated and unaerated zones, were considered to study the impact of different DO concentrations. Parts of the measurements were conducted in the return sludge channel to study the impact of high suspended solids concentration on the IRs.

3.2.1.1 Artificial biofilm fouling

In the artificial biofilm fouling experiments, the goal was to decide and apply a fouling substance to the DO sensor that fulfilled the following three criteria:

- It should be simple to repeatedly apply and remove without damaging the sensor.
- It should be similar to organic biofilm, or at least result in a small negative bias.
- It should remain fixed to the sensor during repeated air cleaning impulses.

During one experiment, the fouling procedure consisted of the following three steps:

1. Compare test sensor measurements with reference DO sensors;
2. Repeat IR measurements with clean test sensor;
3. Manually foul test sensor with an artificial biofilm fouling substance and repeat IR measurements during fouled conditions.

3.2.1.2 Sensor set-up and data collection

Five DO sensors (Cerlic O2X DUO) were connected to a data acquisition system with hardware and software from National Instruments, with the software

LabVIEW. Data were stored in a PostgreSQL database in the same laptop computer. Each DO sensor could be switched between MEC- and OPT measurement technology by simply changing the top part of the sensor. A photo of the experimental equipment is given in the Supplementary materials S3.1.

Two of the DO sensors, one OPT- and one MEC-sensor, were used to study IRs (test sensors) and two were used as references (both MEC-sensors). The third reference sensor was used as a back-up in case of a failing sensor. All sensors were mounted at rods according to the manufacturer's instruction at a slight angle (5-30°) and at 0.5 m depth. All membranes were replaced with new ones and calibrated in the beginning of the experiments. The length of an air cleaning impulse was set to 15 s at 2 bar for the two test sensors, which was expected to be sufficient to obtain a clear IR, even for high DO concentrations (4 mg/L).

3.2.1.3 Data pre-processing

Data were sampled with 8 Hz and later down-sampled to resemble full-scale conditions. First, the data were low-pass filtered (anti-alias filter) and afterwards down-sampled to 1 Hz. Details about the data pre-processing are given in the Supplementary materials S3.1.

A reference DO concentration was calculated from the two reference sensors by their variance weighted least squares (WLS) estimate, see e.g. (Kay 1993). Then, the bias for a test sensor for a given IR was calculated as the difference between the test sensor and WLS-estimate for the time interval between two IRs.

3.2.1.4 Measurements

The measurements were conducted at two positions in the aerated zone, in the anoxic zone, and in the return sludge channel. At each position, the three-step fouling procedure was repeated multiple times.

3.2.2 Experiment 2 – Real biofilm fouling on a full scale in Bromma WRRF

In the second experiment conducted in Bromma WRRF, real biofilm growth was studied for 7 months to investigate:

- fault progression in terms of bias due to biofilm fouling;
- difference in bias magnitude between MEC- and OPT-type sensors due to biofilm fouling;
- sensitivity to detect bias using the response time of an IR;
- variation in the response times for IRs, both for normal and fouled conditions;
- impact of seasonal variations on biofilm growth and bias.

Bromma WRRF had a conventional ASP operated at 6 parallel lines with seven zones per line. The sensors in the ASP experience severe biofilm growth compared to other facilities in the city. Despite the automatic air cleaning system in the DO sensors, manual cleaning is required and conducted between one time a week or month, depending on the biofilm growth magnitude.

3.2.2.1 Sensor set-up and data collection

One MEC- and one OPT-sensor were used as test sensors. The sensors were positioned in the ASP in zone 5 (OPT-sensor) and zone 6 (MEC-sensor) about 1 m apart of the existing DO sensors that were used as references (MEC-sensors, Cerlic O2x DUO). Both zones were continuously aerated, but their DO set-points differed slightly (4.0 mg/L in zone 5 and 3.5 mg/L in zone 6).

Both test and reference sensors were equipped with automatic air cleaning system, performing a cleaning cycle every second hour. The air pressure was set between 0.8 and 1.1 bar for each sensor. Both test and reference sensors in each zone were cleaned simultaneously so that their initial DO concentration would be equal.

Data from test and reference sensors were stored in the existing process database with 1s sampling time.

3.2.2.2 Measurements

Each experimental period consisted of monitoring the biofilm growth on two test sensors until both received a bias larger than 0.3 mg/L. The bias was calculated as the mean difference in DO concentration between the reference- and test sensor during the period bounded two consecutive IRs, i.e. 2 hour mean values. The reference sensors were manually cleaned and inspected in addition to the automatic air-cleaning which was assumed to be sufficient to remove biofilm growth on the reference sensor. We expected a time between 1 and 4 weeks to obtain a bias above 0.3 mg/L in the test sensor.

3.2.2.3 Software calculations and data availability

The response time of an IR was defined as the time to reach 63 percent of the peak amplitude, see *Figure 3.1(a)* for an illustration. The initial DO concentration was calculated as the average DO concentration 5s before the IR. A short function computing the response time is provided in the Supplementary materials S3.1. All calculations and data pre-processing were performed off-line in MATLAB. The pre-processed data from both artificial and real biofilm growth experiments are available at www.ivl.se under Creative Commons 0 license (CC0).

3.3 Results

The results from experiments in Henriksdal and Bromma WRRF are described in Section 3.3.1 and Section 3.3.2, respectively.

3.3.1 Experiment 1 - Artificial biofilm fouling experiments in Henriksdal WRRF

The results include illustrations of the selected artificial biofilm fouling substance, motivation of discarded experimental data, and graphs showing variations in IRs during non-faulty- and different faulty conditions.

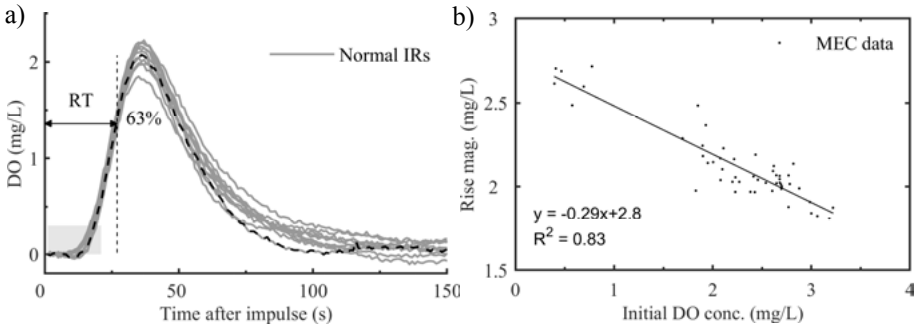


Figure 3.1. **a**) Ten IRs from a non-faulty MEC-sensor (solid grey lines) and the estimation of the response time (t_{63}) for one of the curves (dashed black line). All IRs were normalised by subtracting the initial DO concentration of its IR for illustrative purposes. The response time (RT) is indicated as the time from the start of the cleaning procedure (time after impulse = 0), to reach 63 percent of the peak value. **b**) Correlation between the initial DO concentration and the IR rise magnitude of clean MEC-sensors. The straight line, which is the least squares fit indicates a small negative correlation (-0.29 mg/L per 1 mg/L).

A wide variety of fouling agents were evaluated. A mixture of ball-bearing grease and floating grease from the pre-sedimentation fulfilled the three criteria for artificial biofilm fouling described in the Materials and Methods section. Both MEC- and OPT-sensors obtained a negative bias when fouled with the grease mixture. The same effect was seen for the OPT-sensor with organic biofilm fouling (Andersson and Hallgren 2015). Similarly, our own (unpublished) experience suggests that a completely fouled MEC-sensor display 0 mg/L, thus a negative bias as well. Photos of artificially fouled sensors can be compared with real biofilm fouling in *Figure 3.2*.

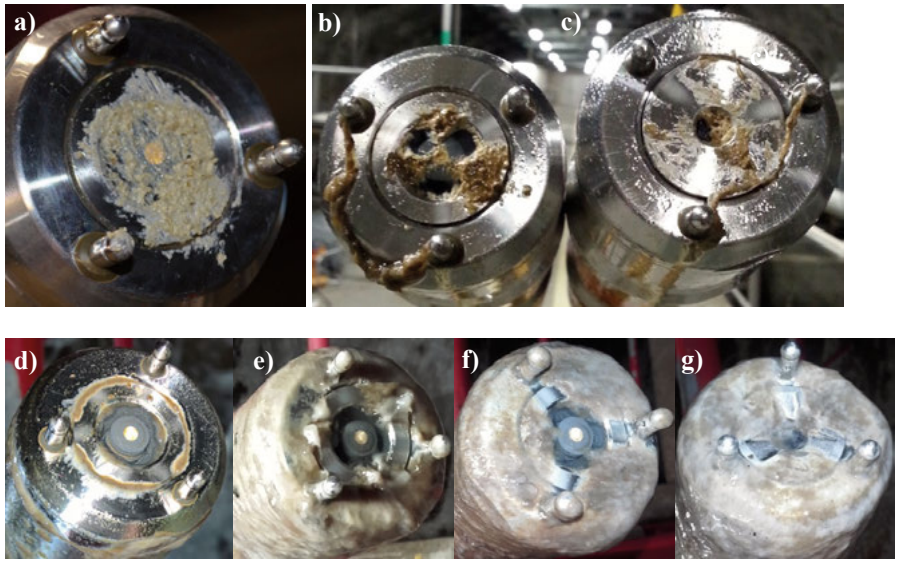


Figure 3.2. **a**) Artificially fouled MEC-sensor with a mixture of ball bearing grease and pre-sedimentation grease before an experimental procedure (2015-07-24). **b**) MEC- and (**c**) OPT-sensor after repeated air cleaning impulses (2015-08-04). Note the effect of the three air cleaning nozzles that resulted in a partly cleaned MEC-sensor in (b). **d-g**) Different stages of biofilm fouling on MEC-sensor (d-f) and OPT-sensor (g). Fouling time: d) 12 days (18/10), e) 27 days (14/11), f) 62 days (19/12), and g) 62 days (19/12). Note that the biofilm starts to grow in the circular gap between the senor and the sensor body (d), and further expands over the entire sensor and rod (e-f).

During the experimental start, the OPT-sensor was damaged and data from the OPT-sensor were not further considered. Instead, both test sensors were of MEC type which allowed the simultaneous comparison of fouled and clean MEC-sensors.

A majority of the fouling procedures for the MEC test sensor resulted in the desired bias magnitude between 0 and 0.2 mg/L with clear IRs. However, for two of the initial fouling procedures, the average bias was larger (-1.7 and -0.35 mg/L). Similarly, a negative bias between -0.27 and -0.51 mg/L was also seen for the clean test sensor during the same time period. The data were studied in detail, which revealed that part of the bias was an effect of large variations in the reference measurements (see the Supplementary materials S3.1). Since it was not possible to explain the reason for the variations, data with an absolute bias larger than 0.2 mg/L were not further evaluated. Also, seven IRs showed deviating shape with double peaks. Those IRs were discarded and removed from the normal dataset. The double peak behaviour was noticed after repeated manual fouling and subsequent cleaning which could have affected the membrane. The potentially damaged membrane was replaced after noticing the double peak behaviour.

The remaining data were pre-processed and the IRs were extracted from the pre-processed data and grouped according to their status (normal or fouled). This resulted in 50 normal IRs and 2628 IRs for fouled conditions.

3.3.1.1 Variations in IRs during normal conditions

Well-defined IRs were obtained for the MEC-sensor. A typical set of IRs at non-faulty conditions were shown in *Figure 3.1*. The rise magnitude variation (the difference between initial and maximum DO concentration during one IR) was about 0.5 mg/L (*Figure 3.1(a)*). Part of this variation was a consequence of a changing initial DO concentration (*Figure 3.1(b)*).

3.3.1.2 Identification of common sensor faults

In addition to the artificial biofilm fouling, accidental faults occurred over the experimental period, common faults that could happen during everyday operations. The accidental faults resulted in distinct changes in the IRs, specific to the different accidental faults. The mean values for the faulty IRs are visualised in *Figure 3.3*.

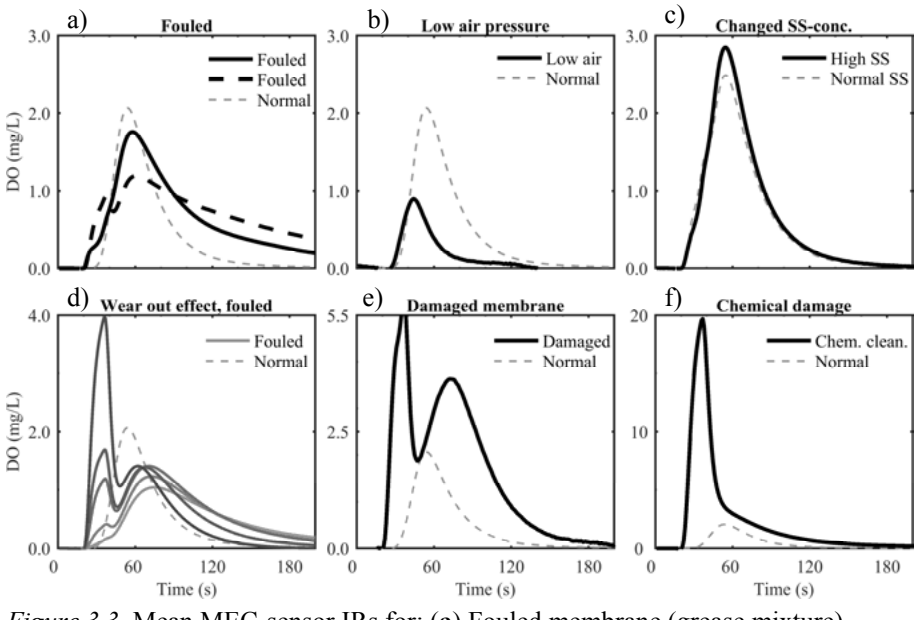


Figure 3.3. Mean MEC-sensor IRs for: (**a**) Fouled membrane (grease mixture). Mean IR of fouled test data (black solid and dashed lines, n=2628). (**b**) Low air pressure during the automatic air-cleaning procedure (n=9). (**c**) The effect of increasing the suspended solids from 2500 mg/L (normal SS, n=9) to 8500 mg/L (high SS, n=10). (**d**) A potential gradual wear out effect by repeated IRs on a fouled MEC-sensor. An increased wear out of the membrane is indicated by darker grey (n=2180). (**e**) A mechanically perforated membrane during manual cleaning (n=2). (**f**) The effect of using harsh cleaning liquid to remove excess grease (n=98). The air cleaning impulse was between time 20 and 35s.

Fouling the MEC-sensor with grease mixture resulted in a dampened IR with an extended time to regain the original DO concentration (*Figure 3.3(a)*, solid and dashed black lines). The small tweak during the impulse rise was not a typical pattern, but was merely an effect of combining IRs with two different shapes: one part of the dataset of the fouled IRs displayed a "double peak behaviour" similar to *Figure 3.3(d)*, whereas one set had a straight increase with dampened IRs similar to *Figure 3.3(b)*.

In the end of the experimental period, a long-term test for 11 days was conducted, where the MEC-sensor was fouled and subject to repeated cleaning events about 200 times per day. This was far more than the recommended amount of cleaning procedures, which resulted in a gradual change of the IR, potentially due to wearing out the membrane (*Figure 3.3(d)*). It is interesting that the shape of the IR changed from a dampened IR to exhibit increasingly pronounced double peak behaviour. A double peak was also seen for the damaged membrane (perforated during manual cleaning) (*Figure 3.3(e)*). We have no clear explanation for the double peaks although they were present more frequently in faulty data, especially where the membrane was mechanically damaged. The first peak coincided with the duration of the air cleaning impulse whereas the second peak was delayed compared to a normal IR.

A decreased air pressure for the automatic air cleaning system resulted in a dampened peak (*Figure 3.3(b)*). Measurements in the return sludge channel with high suspended solids (SS) only indicated a slight increase in the peak height of the same order of magnitude as the normal variation (compare *Figure 3.1* and *Figure 3.3(c)*). This indicates that the SS-level had none or minor impact on the IR. A large increase of the IRs peak value after cleaning the MEC-sensor with harsh cleaning liquid (*Figure 3.3(f)*) indicates that the membrane became more sensitive.

3.3.2 Experiment 2 – Real biofilm fouling experiments in Bromma WRRF

The results first illustrate the impact of biofilm fouling on bias progression and the correlation with IR response time values. Then, variations in the response time values for clean sensors are shown together with the impact on the IRs of damaged sensors. Last, uncertainties in the results due to sensor maintenance during the experiments are detailed.

3.3.2.1 Bias progression due to biofilm fouling

Different stages of biofilm growth fouled sensors can be seen in *Figure 3.2* with resulting bias progression in *Figure 3.4*.

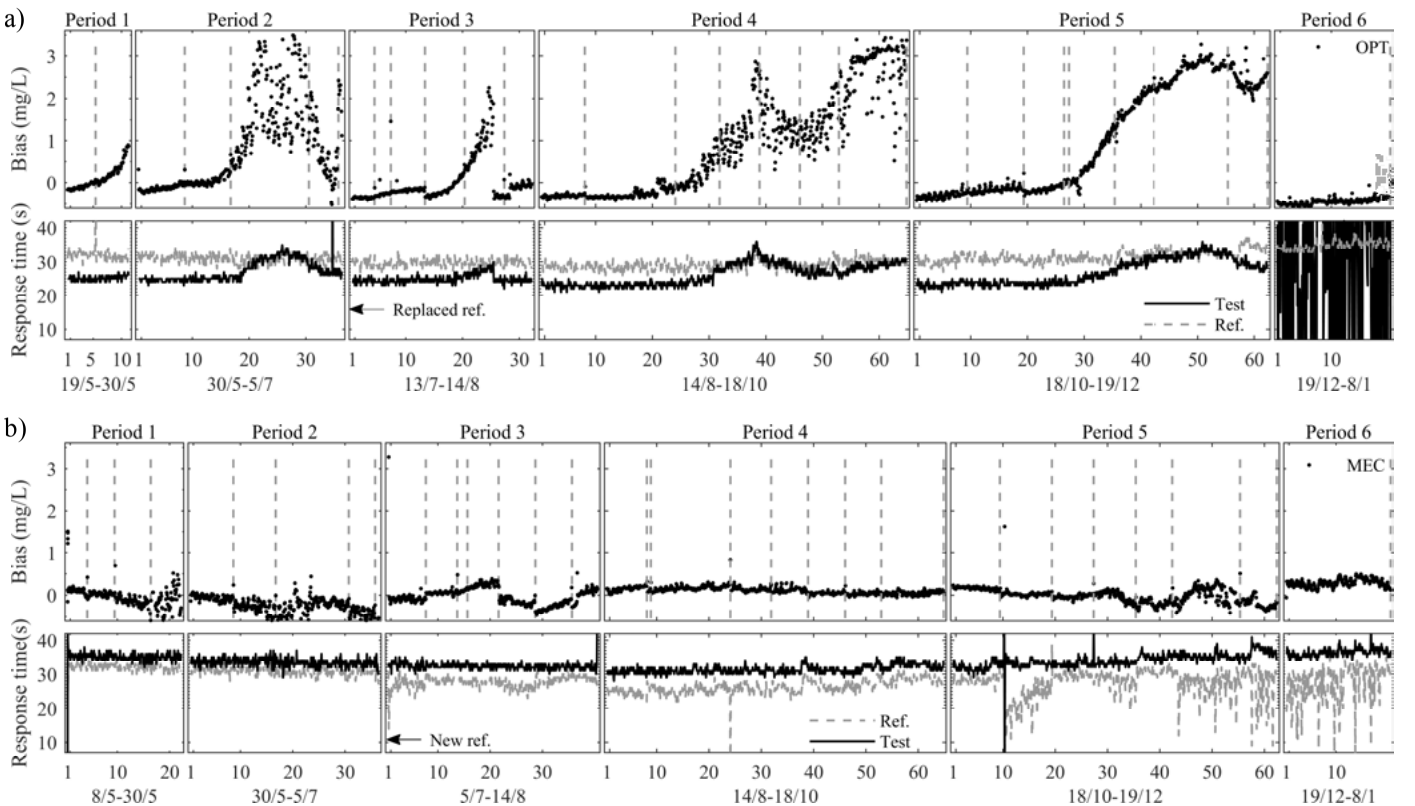


Figure 3.4. Bias and response time values during experimental periods 1-6 in Bromma WRRF for (**a**) Optical sensor (OPT) and (**b**) membrane sensor (MEC). Manual cleaning of reference sensor is indicated with dashed vertical lines.

The bias progression differed between the MEC- and OPT-sensors in several aspects (compare top graphs in *Figure 3.4(a)* and *(b)*). Firstly, the OPT-sensor had an increasing bias in all six periods in contrast to the MEC-sensor which mainly had a decreasing bias (period 1, 2, and 5). Secondly, the bias magnitude was larger for the OPT-sensor compared to the MEC-sensor. Lastly, the OPT-sensor showed a transition in bias progression from linear to exponential increasing after 10-14 days during period 1-3 and after 28 days during period 4-5. The MEC-sensor showed a linear bias trend throughout all periods. Note also that the bias in the OPT-sensor temporary decreased after reaching a large value (above 2-3 mg/L) in period 2-5.

3.3.2.2 Bias detection with the Response time estimation method

The changes in response times due to bias (bottom graphs in *Figure 3.4(a)* and b) were not as evident as expected. For the OPT-sensor, there was a clear change in period 2-5 for bias larger than 1 mg/L and a correlation between bias and the response time values was identified. For the MEC-sensor, there was no obvious change in response time that correlated to the bias magnitude in any of the periods (*Figure 3.4(b)*). Only a slight increase in the variance for the response time values in period 2 (days 15-20) and period 5 (days 35-60) was indicated.

3.3.2.3 Variations in response time values

Surprisingly, the response times also changed for the reference sensors during the experiments. First, a small trend of decreasing response times with about 1s per 30 days can be seen for the OPT-sensor's reference (period 1-3) and for the MEC-sensor's reference (period 1-2). This is similar to the artificial biofilm fouling experiments where the MEC-sensor had a decreasing response time during the wearing out experiment (*Figure 3.3(d)*). A similar decrease can also be noted for the OPT test sensor by comparing the response times during the first days in each test period (period 1-5). Note that there is an opposite trend with increasing response time values for the OPT-sensor's reference and for the MEC test sensor (period 4-6). Second, the new MEC-sensor's reference had larger variations in its response time values, with a large increase in variation during period 5-6.

The OPT-sensor had in general a lower response time than the MEC-sensors (recall that all reference sensors were MEC-sensors). Note also that the MEC sensors had different response times at clean conditions with new membranes.

The response time values for the OPT test sensor in period 6 deviated considerably from previous periods. The reason was an accidental kink in the air supply hose that blocked the air supply to the sensor's air-cleaning system disabling all IRs after period 5. The estimated response time values were therefore randomly obtained. At the same time, its reference sensor was instead given a surplus of air resulting and an increase in IR peak value (see the

Supplementary materials S3.1). Surprisingly, there was no large change in the response time for the reference sensor (*Figure 3.4(a)*, period 6) despite the increased airflow rate and pressure.

3.3.2.4 Damaged sensors

After the final experimental period, both test sensors were deliberately damaged to study mechanically worn out sensors. The MEC-sensor's membrane was perforated with a needle and later with a screw-driver and the OPT-sensor's fluorophore coating was scratched with an iron brush. Photos of the damaged sensors together with corresponding IRs are provided in the Supplementary materials S3.1.

For the MEC-sensor, both the needle and screw-driver perforation resulted in a double peak behaviour as was seen in *Figure 3.3(d)*, although with a smaller initial peak than previously observed. About half of the IRs with needle perforation showed a dip instead of a double peak. There was only a minor change in bias due to the perforation events (before perforation 0.37 mg/L, after needle perforation 0.49 mg/L, and after screw-driver perforation 0.42 mg/L).

For the OPT-sensor, the IRs maintained a one-peak shape characterised by decreasing response time values with increasing amount of scratches. The first scratches did not result in any bias change although the following scratches, which removed > 50% of the fluorophore coating, resulted in a large negative bias (before scratching -0.54 mg/L, after first scratching -0.51Mg/L, and after second scratching -1.54 mg/L).

3.3.2.5 Sensor maintenance

After period 2, a new membrane was installed in the MEC reference sensor. The existing membrane was moved to the OPT-sensor reference sensor. The reason for replacing the membrane already after 2 months was to assure that the observed small bias was an effect of biofilm fouling and drift in the test sensor, and not of a drift in the reference sensor.

In period 3, the MEC test sensor showed a segmented line with both increasing and decreasing bias trends within the segments. The root cause for the segmentation was the manual cleaning of the reference sensor (vertical dashed lines) which coincides with the line segments (*Figure 3.4(b)* period 3). Since the bias changed when the reference sensor was cleaned, this indicates that the reference was also affected by biofilm fouling. This was not seen in the remaining periods where a weekly manual cleaning interval seemed to be sufficient.

For period 3, it is therefore hard to draw conclusions about the bias direction since we are not certain whether it was the reference or test sensor (both of MEC type) that was actually drifting. In addition, the last three manual cleanings were conducted during the vacation period by personal unfamiliar

with the experiments. This resulted in an uncertainty whether only the reference sensor or both the test and reference sensors were manually cleaned. The third last cleaning (day 21 for MEC-sensor and day 12 for OPT-sensor) introduced a change in bias, similar to what would be expected after cleaning the test sensor. In addition, the exact time locations for the last three manual cleanings events were not obtained, but only the actual date. Therefore, these cleanings were marked at 12:00 and should be interpreted with a ± 4 hours' uncertainty.

3.4 Discussion

The outcomes of Experiment 1 (artificial biofilm fouling) and Experiment 2 (biofilm fouling) are discussed and compared with existing results. Further, implications of the findings are discussed with the perspective of using IRs for fault detection in a full-scale application. This includes the aspects of: bias progression linked to sensor maintenance, DO process control, and factors with an impact on the IRs and their potential limitations on a full-scale fault detection and diagnosis application.

3.4.1 Bias progression in oxygen sensors due to biofilm fouling

The purpose of using grease in the artificial biofilm fouling experiments was to resemble organic biofilm growth. The results showed that the MEC-sensor received a negative bias for both artificial and real biofilm fouling. However, the effect on the IRs was larger for small bias with grease compared to the real biofilm. It is difficult to conclude the true explanation for the difference. Although we can intuitively assume that grease resulted in a denser film compared to a water permeable biofilm, which may have contributed to the difference.

For the OPT-sensor, all real experiments resulted in a positive bias which is in contrast to the results by Andersson and Hallgren (2015). Despite the difference in bias direction, the detection sensitivities were in the same range (1.0 mg/L compared to 0.6-0.8 mg/L in the previous study). The OPT-sensor had in addition a faster increase in bias than the MEC-sensor. No results for the MEC-sensor were reported by Andersson and Hallgren (2015), but their unpublished data suggest that the bias was small for the MEC, in contrast to the OPT-sensor, which agrees with this study. A potential explanation for the large impact on OPT-sensor is that the small fluorophore area was more easily covered by biofilm, compared to the larger membrane area in the MEC sensor (compare *Figure 3.2(g)* with *Figure 3.2(f)*). Additional studies are needed to verify whether this is valid in general or only for the specific sensor make in this study. Future studies should compare parallel treatment lines and different WRRFs as this may contribute to variations in a biofilm growth.

As expected, both sensor types had a faster bias increase during summer conditions compared to winter conditions. However, the magnitude of this seasonal effect was larger than expected. The time to reach a bias of 0.3 mg/L was less than one week for the OPT-sensor during summer conditions, and beyond 60 days for the MEC-sensor during wintertime. The large span shows the importance of designing an adaptive sensor cleaning schedule when compared to a fixed interval of one or two weeks, which is common practice today.

3.4.2 Oxygen sensor bias implications for process control

As seen from a process control perspective, knowledge about the bias progression is important. That is, does a fouled sensor result in a positive or negative bias, or even alternate between the two? When a DO sensor is used in a feedback control loop (which is de facto standard), a bias in the sensor will lead to different consequences depending on: the bias direction, the controller structure, and the controlled process.

Consider a MEC sensor with a strictly negative bias that is operating in a feed-back loop with a fixed DO-set point. The true DO concentration will be underestimated resulting in excess air supply with a higher DO concentration than desired. Consider instead the same MEC-sensor but in an ammonium cascade controller where the effluent ammonium adjusts the DO-set point. In such situation, the exact DO concentration will be less important since the bias will be partly compensated by the DO-set point given by the ammonium master controller. The opposite argumentation applies for a strictly positive bias as was indicated for the OPT-sensor. Whether a positive or negative drift direction is bad or worse depend on the process configuration. This raises the question whether the most likely bias direction for a specific sensor should be included in the early process design.

3.4.3 Factors affecting the impulse responses and response time values

One of the questions in this study was to investigate the impact of changed process conditions on the IRs and the corresponding response time values. This was mainly conducted for the MEC-sensor due to the damaged OPT-sensor in Experiment 1. Some of the factors, such as varied SS or initial DO concentration were well studied during the artificial biofilm fouling experiments, but the interpretation of normal variations in response time values remain uncertain. The main problem was that the response time values differed between the clean MEC-sensors. The variation in response time values for the new membrane in the MEC-sensor's reference sensor in period 3 was larger than for the other MEC-sensors. In addition, it was not possible to define the

reason for the long-term small change in response time values with both decreasing and increasing values. As seen in the artificial experiments, decreasing response time values could be a consequence of wearing out the membrane, but the results in Experiment 2 indicate that the decrease could equally well be related to a temperature, seasonal, or unknown effects. Only a few sensor individuals were used in the study, which raises the question whether some of the noted changes are effects of variation among sensor individuals. Note that it is not clear from the results whether variations in the IR and its related response time are due to factors that are independent of sensor bias or e.g. seasonal effects that only affect the dynamics of the IR.

Future studies should therefore both study potential seasonal effects in the response time as well as multiple parallel sensors to distinguish variations in sensor individuals from general effects. In future studies it is also important to avoid ambiguities in bias estimation about whether it is the reference-, test sensor or both that are affected by biofilm fouling. Regular checks of the sensor reading in known zero-valued and saturated oxygen conditions should be used as supplement to the reference measurements.

3.4.4 Fault detection and diagnosis based on impulse responses

The lowest detectable bias with the response time method was larger than the desired 0.3 mg/L. It is therefore interesting to further study if other fault detection algorithms can improve the detection sensitivity. Based on the bias progression, biofilm fouling detection would be most valuable during summer condition when the biofilm growth is fast with a high likelihood of obtaining a bias. Summer is also the time of year when personal resources may be limited due to vacations at the same time as ammonium effluent permits can be stricter than during wintertime (in Sweden). Therefore, automatic biofilm fouling detection in DO sensors has the largest potential to improve process treatment during summertime.

The results also showed that the IRs contained information about different faults including: reduced air supply in the air-cleaning system of the sensors and damaged sensor membrane. These faults gave rise to distinct pattern changes in the IRs, fault signatures, extending the possibility of fault detection to diagnosis. Most importantly, double peaks were evident in both Experiment 1 and 2. By studying unpublished data in the study by Andersson and Hallgren (2015), we noted that both the double peak behaviour (*Figure 3.3(d)*) and the extreme peaks (*Figure 3.3(f)*) were present in that dataset. The results suggest that a double peak indicates a damaged membrane, although the size of the first peak differed for different wearing or perforation causes. We have no physical explanation for the double peaks, although it seems like the first peak is aligned with the 20s long air-cleaning phase. As repeated perforation experiments are costly, the existence and occurrence of double peaks should be evaluated on a full scale, studying a more natural occurrence of double peaks.

An important question to answer is how early before complete sensor failure, a double peak arises.

3.5 Conclusions

The results have improved our knowledge of how biofilm fouling impacts the data quality of DO sensors. The bias progression speed due to biofilm fouling differed between sensor types which need to be considered in fall-back strategies for process control. In addition, the bias progression was faster during summer conditions compared to winter conditions. The results could be used to design effective sensor maintenance routines and to detect and diagnose sensor faults. This is a step towards an increased robust wastewater treatment with decreased environmental impact.

The results showed that IRs and related response time values contained information about the status of both MEC- and OPT-sensors. Bias due to biofilm fouling was detected for bias above 1 mg/L in OPT-sensor but not for values up to 0.8 mg/L in the MEC-sensor. Surprisingly, changes in the pattern of IRs were matched to common sensor faults in the MEC-sensor, suggesting that such fault signatures can be used for fault diagnosis.

The OPT-sensor was affected by biofilm growth to a greater extent than the MEC-sensor, with a positive bias compared to the MEC-sensor with mainly a negative bias. Without manual cleaning or fault detection, it is likely that a DO-controlled process will be operated at a different (unknown) DO concentration than desired, especially when an OPT-sensor is used. Whether this is a general draw-back for the OPT-sensor compared to the MEC-sensor or not should be studied in future studies. Future studies should also consider the following questions:

- Can other fault detection methods be applied to the IRs and improve the bias detection sensitivity?
- Which type of faults can be diagnosed based on IR data, and which fault diagnosis methods are suitable for this task?
- What is the normal variation in response time values and shapes of IRs with respect to seasonal variations, among multiple sensor individuals of the same brand, and in relation to sensor bias?

3.6 Chapter Acknowledgments

We gratefully acknowledge the support of Simon Robertsson (IVL Swedish Environmental Research Institute/Uppsala University) during the artificial biofilm fouling experiments. During full-scale experiments the assistance from staff personnel at Bromma WRRF was invaluable including Anders Påhlsson and Kristin Ahlström. We received valuable comments for setting up the experimental plan regarding the artificial biofilm fouling experiments from Erik Lindblom (Stockholm Water), Henri Spanjers (TU Delft), and Gustaf Olsson (Lund University). Kris Villez (Eawag) provided valuable feedback on the article content and manuscript. Finally, we would like to acknowledge Sofia Andersson and Fredrik Hallgren at IVL Swedish Environmental Research Institute for sharing data and their experiences from previous experiments on DO sensors. Sensor equipment was kindly provided by Cerlic controls AB.

4 Practicability of mass balance-based data reconciliation for process rate monitoring in water resource recovery facilities

Accurate flow and concentration measurements are central for process monitoring in water resource recovery facilities (WRRFs). Data have become abundant, but also difficult to overview and assess, with a frequently questioned accuracy. Data reconciliation based on mass balances has emerged as one solution to these issues. However, the experience with data reconciliation has been limited to a few studies where short-term (weekly) updates needed for process monitoring have not been considered. Additionally, there are no guidelines for tuning the underlying data reconciliation parameters such as the error covariance matrix. These shortcomings make it difficult to appreciate the practical usefulness of data reconciliation in the context of process monitoring. In this study, the merits and pitfalls with data reconciliation applied to process monitoring were evaluated in a seven months long monitoring campaign. Challenges to balance theoretical assumptions with opposing practical limitations were identified, and advices to mitigate pitfalls were provided. A structured approach was proposed for assigning the error covariance matrix and further compared with a common ad-hoc approach. The results indicated that the error covariance matrix may have a major impact on the reconciled values and gross errors detection, which has not yet been made clear in the literature before. Further, the data reconciliation successfully detected measurement errors, inconsistencies in mass balance assumptions, and large process rate changes. Ultimately, the findings imply that the added value of data reconciliation is first, to bring structure to data. Next, to detect gross errors, and last, to improve the accuracy of data.

4.1 Introduction

Data availability has transitioned from being a bottleneck, to currently overwhelm process operators with an abundance of on-line sensor and laboratory (lab) data. The usage of on-line sensor data has diverged from the original

drivers of automatic control (Olsson 2012), to multiple purposes with an impact on all parts of the water resource recovery facility (WRRF) organization (Corominas et al. 2018). For this reason, it has become increasingly important to identify and remove inaccurate data from what is used for further analysis. This is especially important when big sets of data are combined for critical decision support such as process monitoring.

Process monitoring involves early warning of unexpected process upsets, which for example can be caused toxics in the influent (Ren 2004). The follow up of process rates constitutes a similarly important monitoring tasks. For instance, controlling the denitrification rate by addition of external carbon source is associated with costs (Wang et al. 2017) and greenhouse gas emissions (Willis et al. 2017). Nitrification and denitrification process rates (referred to hence forth as de-/nitrification) can be determined by lab experiments and batch tests (Loosdrecht et al. 2016), although such tests are time-consuming and not routinely performed. Instead, the abundance of data is used for monitoring the conversion of nitrogen *concentrations* (ammonium, nitrate, and nitrite) throughout the process. A better overview of current process rates would be obtained if *mass flows* were monitored instead of concentrations. This, however, is rarely possible because part of the flows and concentrations remain unmeasured (despite the abundance of data). It is also common that some measurements are unreliable due to the harsh measurement environment and therefore cannot be used. Thus, if mass flows are to be evaluated, they need to be estimated from existing, possibly inaccurate data.

Mass balance-based data reconciliation has been successful in addressing the occasionally low data quality in WRRFs (Le 2019). Data reconciliation has also shown useful for detecting and correcting gross errors in WRRF applications (Puig et al. 2008). The method has been well-described in e.g. (Romagnoli and Sanchez 1999) and has also been demonstrated useful with real data in other industries (Câmara et al. 2017, Moreno et al. 2019, Özyurt and Pike 2004). In fact, data reconciliation is available as commercial software packages for industries (Câmara et al. 2017) and has not been limited to mass balance-based methods. Gaussian mixture models (Xie et al. 2018), principal component analysis (Fuente et al. 2015), and kernel principal component regression (Marimuthu and Narasimhan 2019) are examples of methods that have also been used for data reconciliation. The most common method is however to minimize estimation errors in a (non)linear system of constraints with weighted least squares (Câmara et al. 2017). This method has been dominating in WRRF applications (Le 2019, Puig et al. 2008, Spindler 2014), and will also be considered in this study.

One limitation of applying (the weighted least squares) data reconciliation is that current guidelines for assigning the error covariance matrix, Σ, in the least squares approach are inapplicable for WRRF data. Two approaches have been suggested where Σ is either: i) estimated from data alone (Direct ap-

proach), or ii) estimated from data in combination with linear constraints (Indirect method) (Romagnoli and Sanchez 1999). However, both approaches fail if data contain gross errors (i.e. biased measurements) and assume a process in steady-state. Benqlilou (2004) pinpoints that the key issue in real industrial environments (for calculating Σ), is to separate process dynamics from measurement errors.

In WRRF-related studies so far, expert knowledge has been the dominating strategy to assign Σ, and no studies seem to clearly describe their reasoning behind this essential step. Le (2019) reviewed four data reconciliation studies and found that in none of them there was a clear motivation for how and why Σ was defined. One possible reason for the lack of transparency is that Σ has so far been assigned ad-hoc, which can be difficult to motivate in a research paper (although this is not necessarily wrong).

This study evaluates the practical usefulness of mass balance-based data reconciliation in the context of process monitoring. A 7 months long measurement campaign was conducted on a full-scale pilot WRRF (65,000 p.e. pilot in Käppala WRRF, Sweden) with a whole range of available on-line sensors and lab measurements (sludge-, water phase and off-gas) to assess a realistic data complexity. We propose a transparent method for assigning the error covariance matrix and compare it with an existing ad-hoc method, and the overall influence of the error covariance matrix on the data reconciliation results and gross error detection capability. Ultimately, this study aims to contribute to better data usage in WRRFs by revealing pitfalls and their solutions for bringing data reconciliation into practice.

4.2 Material and Methods

The methodology (Section 4.2.1) describes how the study was conducted to evaluate the practicability of data reconciliation. The process configuration is described in Section 4.2.2 with targeted process indicators in Section 4.2.3. Then, the data reconciliation procedure and gross error detection tests (Section 4.2.4) and mass balance constraints (Section 4.2.5) are described along with the measurements (Section 4.2.6) and assumptions (Section 4.2.7) that underlies the mass balances. A new method to assign the error matrix is proposed in Section 4.2.8, which is later compared with the ad-hoc approach (Section 4.2.9). Last, the data processing and a data quality index are described in 4.2.10.

4.2.1 Methodology

Five process performance indicators related to de-/nitrification were studied during a 6 months long pilot plant measurement campaign during different

denitrification process modes. The usefulness of data reconciliation for process monitoring was assessed by, first, evaluating how well the reconciled process performance indicators could describe changes in denitrification process modes and known process disturbances. Second, by comparing the agreement between theoretical, reconciled values and one laboratory measurement campaign. Last, the reconciled denitrification rate (as one representative process indicator) was compared with its analogous rate computed from unreconciled data. The impact from the error covariance matrix, on the reconciled denitrification rate and gross error detection distribution, was evaluated by comparing three different methods for assigning the error covariance matrix. The contribution from an increased redundancy and measurements was finally evaluated by evaluating the change in gross error detections, due to successively adding mass balance groups as constraints in the data reconciliation.

4.2.2 Process configuration and operation

The pilot plant (*Figure 4.1*) is a treatment line separated from the full 510,000 p.e. Käppala WRRF. The original purpose with the pilot plant was to evaluate the pre and post-denitrification capacity in a future process solution. In this study, three different process modes were considered (Table 4.1), which simulated different nitrate loads with resulting changes in pre and post-denitrification rates. The three modes were produced by adjusting the nitrate recirculation to the pre-denitrification and the addition of external carbon to the post-denitrification. In *high carbon*, nitrate recirculation was shut down to force all denitrification to the post-denitrification zone, and simulate a high nitrate load and resulting external carbon consumption. The *normal carbon* mode used both pre- and post-denitrification. The *low carbon* mode only used pre-denitrification and no added external carbon. External carbon was only added to the post-denitrification zone in normal and high carbon mode. More details about the pilot plant are provided in Supplementary materials S4.1.

Table 4.1. *Process modes used to produce different denitrification rates during the measurement campaign. The process mode low/normal/high carbon refers to the amount of external carbon added to the post-denitrification (zone 8).*

Process mode	External carbon dosage	Nitrate recirculation
Low carbon	None	Active (3-5 x *Qasp*)
Normal carbon	Fixed amount	Active (3-5 x *Qasp*)
High carbon	Maximum pump capacity	Off

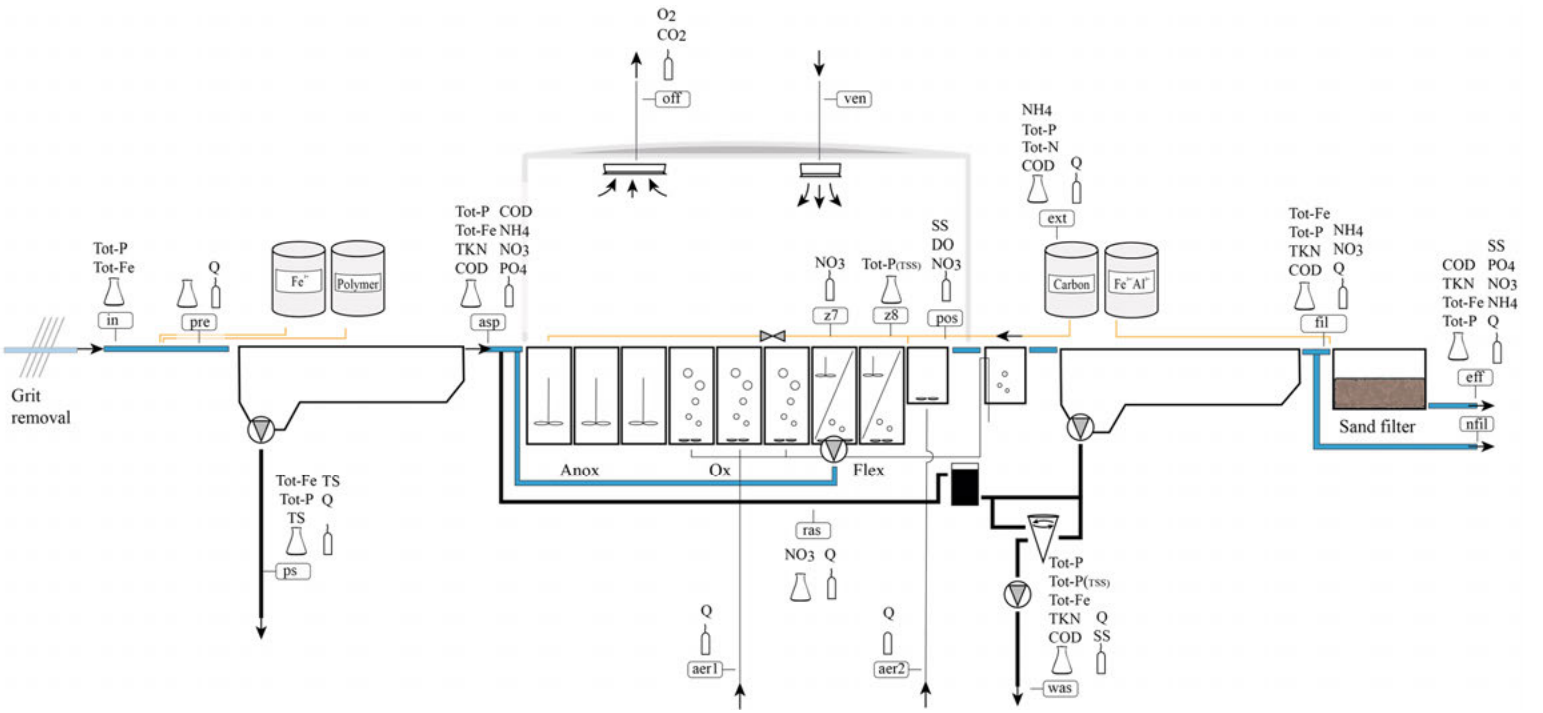


Figure 4.1. Käppala pilot plant including on-line sensors (sensor illustration), lab measurements (conical flask illustration) used during data reconciliation.

4.2.3 Process performance indicators

Two types of process performance indicators (referred to hence forth as process indicators) were defined to exemplify variables used for process monitoring. The process indicators reflect *Process rates* and *Efficiency* (Table 4.2). The de-/nitrification rates were normalized with respect to volatile suspended solids concentration, hence the specific de-/nitrification rates. Note in Table 4.2 that the theoretical specific nitrification rate r_{nit} and the optimal specific denitrification rate r_{deni} were temperature adjusted with a correction factor θ_T. The purpose was to enable a fair comparison with the rates based on reconciled values, which were obtained for different temperatures. The correction factor θ_T is detailed in Appendix A4.1 and was suggested by Henze et al. (2002).

4.2.3.1 Process rates

The theoretical specific nitrification rate r_{nit} was estimated using traditional process stoichiometry, literature coefficients and mean process values and is detailed in Appendix A4.1.

The largest practically feasible (referred to as 'optimal') specific denitrification rate was estimated from batch lab experiments. These showed 4.4 and 9.2 g N_2/kg,VSS,h for pre- and post-denitrification, when compensated to the same temperature (20°C) by applying a standard temperature correction factor as in (Henze et al. 2002). The optimal specific denitrification rate in practice, r_{deni}, was then defined as the average value of the two rates, weighted by the (mass) external carbon to the post-denitrification ($mCODext_{Z8}$) that would give a complete external carbon usage. The proportion of external carbon was based on the chemical oxygen demand (COD) with the optimal denitrification rate defined as

$$r_{deni} = 9.2 \left(\frac{mCODext_{Z8}}{2.86\, mDENI}\right) + 4.4 \left(1 - \frac{mCODext_{Z8}}{2.86\, mDENI}\right) \quad (4.1)$$

4.2.3.2 Efficiency process indicators

Three efficiency process indicators for de-/nitrification and carbon source utilization were defined. The first two measures simply describe the percentage de-/nitrification compared to complete de-/nitrification (Table 4.2). The third measures carbon efficiency as the *efficiency factor* in (Henze et al. 2002) and indicates the percentage of carbon (in terms of COD) that was used for denitrification and biomass production, in proportion to the total carbon removal (including oxidation of COD into CO_2 ($mCODrem$)). Note that the efficiency factor in Table 4.2 only considers biologically removed carbon in the ASP, and neglects COD variations in the effluent and primary sludge.

Table 4.2. *Process indicators used for process monitoring of de-/nitrification. Denitrified nitrogen ($DENI$), nitrified nitrate (NIT), volatile suspended solids (VSS), sludge production (COD_{was}), temperature corrected optimal rate (θ_T), Kjeldahl nitrogen load to activate sludge process (TKN_{asp})and TKN in waste sludge (TKN_{was}). All computations were based on the mass of each component, indicated with prefix m.*

Process indicators					
	Nitrification		**Denitrification**		**Carbon utilization**
	Specific	Theoretical	Specific	Optimal	
Process rates	$\frac{mNIT}{mVSS, day}$	$\theta_T r_{nit}$	$\frac{mDENI}{mVSS, day}$	$\theta_T r_{deni}$	
Efficiency	$\frac{mNIT}{mTKN_{asp}}$		$\frac{mDENI}{mNIT}$		$\frac{2.86\, mDENI + mCODwas}{2.86\, mDENI + mCODwas\ +\ mCODrem}$

4.2.4 Data reconciliation and gross error detection

The weighted least square data reconciliation problem is defined as

$$\begin{aligned} &\underset{x}{arg\,min}(\bar{y} - x)^T \Sigma^{-1}(\bar{y} - x) \\ &s.t.\ Mx = 0 \end{aligned} \tag{4.2}$$

where $\bar{y}$ is the mean of the measured data, x is the estimated (or reconciled) values, Mx is a set of linear constraints in the form of mass balances, and Σ^{-1} is the inverse of the related error covariance matrix, see (Romagnoli and Sanchez 1999) for further details. The mean value $\bar{y}$ (and the reconciled value) are obtained for a certain time window, which was here one week.

The constraint $Mx = 0$ is an equality, which can be solved with the Lagrange multiplier method. Extensions exist, which also include inequality constraints that can be solved using non-linear programming as in (Lid and Skogestad 2008). The constraints can also be a set of non-linear equations, such as bi-linear constraints. An extensive comparison of different numerical methods for solving (4.2) with different constraint variants is given by Câmara et al. (2017).

The MATLAB implementation by Le et al. (2018) was used to solve (4.2), which uses the Lagrange multiplier method (Romagnoli and Sanchez 1999). The constraint equations Mx contain both linear and bi-linear mass balance equations and are described in Table 4.4. One key feature of the used implementation is that the subset of mass balance constraints that only contain measured variables so-called active constraints are identified (here denoted active constraints). The active constraints facilitate the initial data reconciliation procedure by providing direct estimates of the unmeasured variables. The steps to executed in this study are given as a typical sequence in Table 4.3, although the process was iterative in practice.

Table 4.3. *The steps in this study for planning, executing, and evaluating data and related process indicators with data reconciliation.*

Data reconciliation workflow
Data and problem preparation
1. Define key variables that are required to compute process indicators
2. Identify current process layout and measurements
- Identify mass balance constraints from the process layout
- Identify existing measurements of variables in mass balances
3. Make assumptions that can replace unmeasured variables
4. Assign error covariance to all measurements
5. Run experimental design (Le et al. 2018) and identify redundant and unmeasured variables that needs to be measured
6. Complement existing measurements, with additional sensors and lab measurements based on results in Step 5
7. Process data
- Pre-process time series and replace outliers with Not-a-Number (NaNs)
- Make concentration measurements flow proportional
- Compute mean values for a 7 days long time window
Execute data reconciliation
8. Reconcile data
- Estimate unmeasured flows using linear constraints
- Derive active constraints
- Make initial estimate of unmeasured variables based on active constraints
- Perform bi-linear data reconciliation using the Lagrange multiplier method using the method in (Le et al. 2018)
Evaluate results
9. Test for gross errors
- Perform gross error test and identify potentially biased measurements
- Take action to mitigate potential sensors errors
10. Assess process indicators
- Compute process indicators from reconciled values
- Assess data quality used to compute process indicators
- Interpret process indicators and monitor the process

Three conventional gross error detection tests were considered, all described in (Narasimhan and Jordache 2000) and their test statistics are given for the linear criterion in (4.2) for clarity.

The global test gives an overall indication of the presence of any gross error in the full system of constraints. The global test statistic γ_G is defined as

$$\gamma_G = r^T V^{-1} r \tag{4.3}$$

where r is a vector of residuals between all constraints, i.e. $r = M\bar{y} - Mx$, and $V = M\Sigma M^T$.

The nodal test statistic is defined as

$$\gamma_{N,i} = \frac{|r_i|}{\sqrt{(V_{ii})}} \tag{4.4}$$

for each active constraint i.

The measurement test performs one test for each measurement n using the test statistic

$$\gamma_{M,n} = \frac{|e_n|}{\sqrt{\Psi_{nn}}} \tag{4.5}$$

for the residuals $e = \bar{y} - x$, where Ψ_{nn} is the n-th diagonal element covariance matrix of the vector of residuals e. Note that the measurement test can only be performed for variables that are reconciled and measured.

Both the measurement-, and nodal test were adjusted to 95% simultaneous confidence level using the Bonferroni correction (Shaffer 1995). In short, the Bonferroni correction change the simultaneous confidence level, α_{sim}, compared to the initial individual confidence limit $\alpha_{ind} = 0.05$ as $\alpha_{sim} = \frac{\alpha_{ind}}{n_{ind}}$, where n_{ind} in this study was number of independent measurement tests. Here, n_{ind} was about 30 to 40 depending on how many mass balances and measurements that were considered in the measurement test.

Note also that, for every time window, the global test can produce one error, the nodal test can raise one error per active constraint i, and the measurement test can produce at the most one error per measurement n.

An ad-hoc gross error detection approach was used as complement to the conventional measurement test (4.5). The ad-hoc error test straightforwardly compares the absolute percentage change between measured and reconciled values, with a threshold at 20% to decide gross error existence. In addition, an 0.5 absolute difference between measured and reconciled values was also required to raise an error to avoid false alarms for variables with measurements close to zero.

4.2.5 Mass balance constraints

In total, 28 mass balance equations were defined (Table 4.4) to describe expected mass flows of COD, total phosphorous (Tot-P), nitrogen, iron (Tot-Fe) and total mass flows (water, sludge, and air) for the process layout illustrated in *Figure 4.1*. The mass flows describe the mass during a certain time interval (mass/day). The mass balances were further grouped based on the modelled components (middle column 'Group' in Table 4.4). For example, mass balance equations related to water flows were in *flowMB* and mass flows of Tot-P were in *tpMB*. The group *coreMB* contained mass balances needed to compute the process indicators.

Note that Table 4.4 contains more mass balances than needed to compute the process indicators, i.e. there is purposefully redundant information. In fact, a large effort was conducted to identify all available measurements, to maximize the usage of available data by defining the mass balances related to the process indicators. As an example, both *fe-* and *tpMB* were included to allow an improved reconciliation of the flow measurements, which were not redundant. Also, the off-gas measurements (equation 15-17, Table 4.4) were added to produce a redundant measurement of the denitrification rate.

As an example of how the process indicators were obtained from the mass balances, consider the nitrification rate (*mNIT*), which is included in mass balance 9,10,13,14,17, and 20 in Table 4.4. All other variables contained in these seven mass balances contribute to how accurate the computed *mNIT* will be. The workflow for increasing the accuracy in *mNIT* is to fuse all available measurements via the mass balances. This gives a reconciled estimate of *mNIT* that has used all available measurements and process knowledge, given the measurements uncertainties encoded in the error matrix.

Variables in bold text in Table 4.4 were estimated from the mass balances whereas variables in normal text were measured. Before executing the reconciliation in (4.2), the active constraints were identified as the minimum number of unique mass balances obtained from a Gauss elimination on the full set of mass balances, see (Le et al. 2018) for details. This step includes separating measured from unmeasured variables that needs to be estimated.

Table 4.4. *Mass balances around unit processes in Figure 4.1. All mass balances should sum to zero when theoretical and practical assumptions are fulfilled. Bold components were estimated during data reconciliation, italic components indicate an assumed constant value and normal font indicate measured components. The abbreviations are detailed in the bottom of the table. Note that the table covers two pages.*

	Type	Group	Mass balance
1.	Flow (ASP + post-sed. + sand filter + effluent)	*flowMB*	***Qasp*** *- Qwas - Qfil -* ***Qnfil***
2.	Flow (pre-sed.)		*Qpre -* ***Qasp*** *- Qps*
3.	Flow (ASP)		***Qpos*** *-* ***Qasp*** *- Qras*
4.	Flow (sand filter)		*Qfil - Qeff*
5.	Airflow	*coreMB*	*Qoff - Qven - Qaer1 - Qaer2*
6.	COD (asp + post-sed)		***Qasp*****CODaspW + Qext_z1*CODextG+ Qext_z8*CODextG - (****Qnfil****+Qfil)*CODfilW - Qwas*SSwas*SS2CODwas -* ***mCODrem*****1000 - 2.86*****mDEN*****1000*
7.	COD (sand filter)		***mCODsf*** *- Qfil*CODfilW/1000 + Qeff*CODeffW/1000*
8.	COD (sensor equals lab sample)		*CODasp - CODaspW*
9.	TKN nitrogen (asp)		***Qasp*****(NaspW-NOaspW) + Qras**** SSras*****SS2TKNwas -* ***Qpos*****(NfilW - NOfilW) -* ***Qpos*****SSpos*SS2TKNwas -* ***mNIT*****1000 + Qext_z1*NextG+ Qext_z8*NextG*
10.	TKN nitrogen (asp + post-sed)		***Qasp*****(NaspW-NOaspW) - Qwas*SSwas*SS2TKNwas -* ***mNIT*****1000 - (Qfil+****Qnfil****)*(NfilW - NOfilW) + Qext_z1*(NextG-NOextG)+ Qext_z8*(NextG-NOextG)*
11.	Nitrate nitrogen (sensor equals lab sample)		*NOfilW - NOfil*
12.	Nitrate nitrogen (sensor equals lab sample)		*NOasp – NOaspW*
13.	Nitrate nitrogen (asp)		***Qasp*****NOasp + Qras*NOpos -* ***Qpos*****(NOpos) +* ***mNIT*****1000 -* ***mDEN*****1000*
14.	Nitrate nitrogen (asp + post-sed)		***Qasp*****NOasp -* ***mDEN*****1000 +* ***mNIT*****1000 - (Qfil+****Qnfil****)*NOfil - Qwas*NOwas*
15.	Nitrogen gas (asp)		***1000*mDEN - Qoff*********N2off*** *+* ***(Qaer1+Qaer2+Qven)*925.08***
16.	Nitrogen gas (off)		***N2off*** *- (1 000 000ppm - Aroff - CO2off - O2off)*
17.	Oxygen gas (asp)		*(Qaer1+Qaer2)* O2aer + Qven* O2aer - Qoff*O2off -* ***mCODrem*****1000 - 4.57****mNIT*****1000+2.86****mDEN*****1000*
18.	Denitrification pre/post		***mDEN*****1000 -* ***dNOpre*****(****Qasp****+QNOasp+Qras)-* ***dNOpost*****(****Qasp****+Qras)*

19. Post-denitrification (concentration change)		***dNOpost - NOz7** + NOpos*
20. Nitrate nitrogen (Nitrification volume (part of ASP))		***Qasp***NOasp + Qras*NOpos **-dNOpre*(Qasp**+QNOasp+Qras) +**mNIT***1000 - **NOz7*(Qasp**+Qras)*
21. Net oxygen requirement		***OCnet** - (**mCODrem** + 4.57***mNIT** - 2.86*(**mDEN**))*
22. Suspended solids	*ssMB*	*SSpos***Qpos** - **SSras***Qras - Qwas*SSwas - (Qfil+**Qnfil**)*CODfilW/SS2CODwas*
23. Total phosphorus (pre-sed)	*tpMB*	*Qpre*TPinW **- Qasp***TPaspW - Qps*DSps*DS2TPps*
24. Total phosphorus (pre-sed+asp+post-sed)		*Qpre*TPinW - Qps*DSps*DS2TPps - Qwas*SSwas*SS2TPwas - (Qfil+**Qnfil**)*TPfilW + Qext_z8*TPextG + Qext_z1*TPextG*
25. Total phosphorus (sand filter)		***mTPsf** - Qfil*TPfilW/1000 + Qeff*TPeffW/1000*
26. Iron (pre-sed)	*feMB*	*Qpre*FEinW - **Qasp***FEaspW - Qps*DSps*DS2FEps*
27. Iron (pre-sed+asp+post-sed)		*Qpre*FEinW- Qps*DSps*DS2FEps - Qwas*SSwas*SS2FEwas - (Qfil+**Qnfil**)*FEfilW*
28. Iron (sand filter)		*Qfil*FEfilW - Qeff*FEeffW - **mFEsf***1000 + QFEfil*FEext_fil*

Note: The total phosphorous and iron are abbreviated with TP and FE in the table in contrast to the main text where Tot-P and Tot-Fe are used. Additional abbreviations used are: asp = activated sludge units; post-sed = post-sedimentation; pre-sed = pre-sedimentation; fil = water to sand filter; Q = flow (m3. d-1), m = mass flow of components (kg. d-1); d = concentration change (mg L-1); TKN = Kjelldahl nitrogen; NIT = Nitrified; DEN = denitrified; rem = removed; NO = Nitrate + nitrite nitrogen. SS2TKN = a fixed ratio and conversion factors for converting SS to TKN. The suffix 'W' in NOfilW indicates a weekly lab sample in contrast to NOfil which indicates an on-line sensor measurement.

The mass balances describe mass flows during a given time interval $t = \{1, \dots, T\}$ and form the data reconciliation constraints M in (4.2). For any component with concentration C and flow Q, the empirical expected mass flow $\mathrm{E}[M]$ is given by

$$E[M] = \tfrac{1}{T}\int_{t=1}^{T} Q(t)C(t)dt. \tag{4.6a}$$

Equation (4.6a) can be approximated by its discrete time equivalent expression, with sampling time Ts, when sufficiently fast sampling is applied

$$\begin{aligned}
&\frac{1}{T}\int_{t=1}^{T} Q(t)C(t)dt \;\approx\; \frac{1}{KTs}\sum_{k=1}^{K} Q(k)C(k) = \\
&= \frac{1}{KTs}\left(\sum_{k=1}^{K} Q(k)\sum_{k=1}^{K} C(k) + \sum_{k=1}^{K}(Q(k)-\bar{Q})(C(k)-\bar{C})\right) = \\
&= \frac{1}{KTs}\sum_{k=1}^{K} Q(k)\,\frac{\sum_{k=1}^{K} C(k)Q(k)}{\sum_{k=1}^{K} Q(k)} = \\
&= \bar{Q}\,\frac{\sum_{k=1}^{K} C(k)Q(k)}{\sum_{k=1}^{K} Q(k)}
\end{aligned} \tag{4.6b}$$

where k is an integer denoting discrete time and $\bar{Q}$ is the mean flow during $\{1, \dots, T\}$. Note that the factor in the last term $\frac{\sum_{k=1}^{K} C(k)Q(k)}{\sum_{k=1}^{K} Q(k)}$ is the flow proportional mean concentration, which differs from the mean concentration $\frac{1}{KTs}\sum_{k=1}^{K} C(k)$.

4.2.6 Measurements

Existing measurements, mass balances and the targeted variables used to obtain the process indicators were analysed for redundancy using the experimental design method in (Le et al. 2018). The outcome from the experimental design resulted in installing additional on-line sensors:

- an UV-vis sensor (S:can, Spectrolyzer) for measuring COD and nitrate (NO) at the influent to the ASP.
- oxygen (O_2) sensor (Servomex Multiexact 4100), and carbon dioxide (CO_2) sensor (Vaisala, GMP343) measuring the off-gas composition.
- a suspended solids sensor (Cerlic ITX) measuring the waste activated sludge (WAS).

In total, 36 on-line sensors were used (sensor type and make are given in Supplementary materials S4.2), see *Figure 4.1* for sensor locations.

The existing weekly lab sampling programme was also extended, mainly to analyse the composition of sludge streams and external carbon source. Sludge analyses for WAS and primary sludge (PS) were needed to obtain sufficient data for closing the mass balances. Four grab sludge samples were taken about one month apart targeting dry weather occasions. A set of conversion factors were then computed from the four samples to convert suspended solids to the components COD, Tot-Fe, Tot-P, TKN, and NO. These conversion factors were included in the mass balances as variables (Table 4.4). As an example, the *SS2CODwas* was the factor to convert suspended solids measurements in WAS to COD. A complete list of lab samples and on-line sensor data is provided in Supplementary materials S4.2, and also indicated with abbreviations and symbols (conical flask or sensor symbol) in *Figure 4.1*.

The reason for limiting the grab samples to only four was the high cost and that these kinds of samples are not commonly analysed for all sludge streams in the WRRFs in Sweden. The need for more extensive sludge sampling campaigns to enable data reconciliation in a realistic setting would therefore be questioned by practitioners. Thus, here we used the minimum practically feasible sludge analyses.

4.2.7 Assumptions and constants

The mass balances were simplified with regards to data collection and mass flow modelling by making 14 assumptions (Appendix A4.3).

In short, the assumptions included the following. The WAS composition in terms of COD, NO, and TKN was assumed to be constant in relation to the suspended solids content. Changes in the WAS composition were assumed to be caused by changes in biological composition of the activated sludge. Such changes are slow in contrast to PS composition that is strongly affected by precipitation mode. Both Tot-Fe and Tot-P content in PS and WAS can change fast depending on the current precipitation mode and it was therefore uncertain whether four sludge sample were enough to obtain representative conversion factors for Tot-Fe and Tot-P.

The constant airflows for incoming fresh air and off-gas ventilation in shafts were measured manually with a portable air velocity logger and estimated to a constant value as detailed in Supplementary materials S4.1.

4.2.8 A systematic approach to assigning the error matrix

The inverse of Σ in (4.2) is commonly denoted as the weight matrix $W = \Sigma^{-1}$, as it weighs the residual errors during reconciliation. Due to the lack of WRRF applicable guidelines for defining Σ we propose a transparent and systematic method that is applicable to real conditions and limitations in WRRF data. The

method assumes that all error magnitudes can be defined as percentages of the mean, representing a one standard deviation if a multivariate Gaussian error distribution is assumed. The method is outlined in Table 4.5 and further described in the following section.

Table 4.5. *The suggested stepwise method to define the error covariance matrix.*

A method to assign the error covariance matrix in data reconciliation
Step 1
Define and assign error magnitudes for the error sources in terms of 1-std of the mean (assuming Gaussian error distribution)
a) **Analytical error**, lab sample (5 – 15% depending on analyte)
b) **Sampling error**, lab sample
- Weekly sample (20% - volume proportional sampling)
- Grab sample (30% - manual grab sample)
- Grab sampling campaign (40% - several grab samples taken to estimate a certain property for sludge that is expected to be constant)
c) **Sensor error** (0 – 30%, for sensors that lack a reference measurement)
d) **Operator experience error** (0 – 50%)
Step 2
Identify correlated sampling and sensor errors.
If correlations are unknown, assume zero correlation.
Step 3
Assign the diagonal of Σ_{eff} with the squared effective standard deviation from Step 1, and the covariance from Step 2 as off-diagonal elements. The effective standard deviation, S_{eff}, was defined as, $S_{eff}{}^2 = S_{1,precision}{}^2 + S_{1,accuracy}{}^2 + \dots S_N{}^2$, where the standard deviation for N error sources (both precision and accuracy) were summed as in (Narasimhan and Jordache 2000).

The sampling error magnitude estimates in Step 1 in Table 4.5 represent the total sampling error, see (Petersen and Esbensen 2005) for a detailed description. Here, the tool from (Rossi et al. 2011) was used to estimate the sampling errors from the theory of sampling (TOS) with settings detailed in Supplementary materials S4.1 and S4.2. The main contribution to the total sampling error was the so-called Point materialization error (PME), which describes how representative the sample is for the sampled stream. The PME was indicated in (Rossi et al. 2011) to lie between 20% for good sampling conditions in sewer (e.g. turbulent, well mixing), and up to 100% (laminar flow). In this study we therefore assumed 20% PME for weekly lab samples, 30% for grab samples in water phase, and 40% PME for grab sludge samples. Further, the sampling errors were assumed to be fully correlated for samples when multiple analyses were performed on the same sample (Step 2). Note that the sampling errors are larger than the analytical errors, which is common for heterogeneous systems. To our knowledge, there exist no practical studies that have evaluated whether these large sampling error are realistic in a WRRF, and the assumed values were therefore considered to be the best available estimates.

Most on-line sensors in WRRFs are calibrated and adjusted based on lab reference measurements. Therefore, the equivalent procedure was applied for

assigning error magnitudes to on-line sensor measurements. The exception was the flow sensors, which were not calibrated with a reference measurement. For those sensors, the manufacturers accuracy values were used as error estimates (<1%) with a possible drift range (<3%). We also added a subjective error term, see Step 1 in Table 4.5, for sensors with poor historic measurement records, potentially due to reduced maintenance or an unreliable sensor. Note that no precision sensor error was assumed since mean values in the order of one week, computed from on 1-minute sampling time, would smear out any random sensor noise.

In the final Step 3, all error sources were combined to effective standard deviations (Narasimhan and Jordache 2000) that defines the Σ_{eff}. All assumed error magnitudes in this study are given in Supplementary Materials S4.2.

We should note that the effective standard deviation allows a combination of both bias and random errors. The Σ_{eff} will therefore deviate from the statistical assumption were only random (variance) errors are assumed. However, we argue that this is not a practical concern as long if we adjust our interpretation of the data reconciliation outcome accordingly.

4.2.9 The ad-hoc error approach

As comparison to the suggested method in Table 4.5, we also defined the minimal error matrix Σ_{min} as the diagonal error matrix with the smallest 1-std error of the mean that did not indicate a global gross error on verification data. We define verification data as a period of data without any known measurement errors, which is in contrast to the evaluation data were gross errors might be detected. The Σ_{min} can be said to roughly indicate the combined quality of measurements and mass balances where a small value indicate only minor adjustments during reconciliation. The approach is ad-hoc but common in industry.

4.2.10 Data treatment and software

The complete data collection and treatment process is detailed in Supplementary materials S4.1 and in short, all raw data were exported from the WRRF's historical database with 1-minute sampling time and thereafter pre-processed and analysed off-line in MATLAB version R2020a.

Quality checks were performed on all raw on-line sensor data with the goal to detect and remove datapoints that were most likely biased. Univariate data quality tests similar as in (Olsson et al. 2005) were applied including: missing data, running variance (low variance-identifying frozen values, high variance - detect too large signal variance), and rate of change (outlier detection). Detected anomalies by the quality tests were excluded from the mean value computations by replacing them with Not-a-Number (NaN).

A Data error index was further developed to enable simultaneous assessment of the data reconciliation as bad signal data quality. The data error index combines three signal error indications: 1) the percentage of variables flagged in the measurement gross error detection test, 2) the percentage of variables that contained more anomalies (NaNs indicated by the quality tests) than a threshold parameter $\varepsilon_{0.1}$ (here set to 10%), and 3) the percentage of variables with missing data, i.e. 100% NaNs from the quality tests. When all data were missing, the mean value in the previous time window was used as replacement.

Concentration measurements of on-line sensor measurements were initially made flow proportional before the mean values were computed. This was however problematic because the flow measurements contained too many anomalies in the quality checks. This was especially a problem for flows that were not directly measured but estimated from several other flow measurements. As an example, *Qpos* is obtained as the sum of *Qasp* and *Qras* (mass balance row 4, Table 4.4). All anomalies in *Qasp* and *Qras* were then transferred to *Qpos* in the form of NaNs. This resulted in an accumulation of NaNs in *Qpos*, which in turn made it infeasible to compute a flow proportional mean value on the reduced data signal. The difference between the mean value for a complete concentration measurement, compared to the mean value for a flow-proportional-compensated concentration measurement was occasionally large, which indicates that a bias was introduced in the flow proportional mean value due to missing (or NaN) data. For this reason, the on-line sensor concentration mean values were instead computed directly (without flow proportional compensation). This approach is the conventional approach previously used in e.g. (Le 2019), which has shown sufficiently accurate for improving the data quality and detecting gross errors. Weekly lab sample measurements were already obtained volume proportionally as the time window matched the weekly sampling interval.

4.3 Results

The results first evaluate how well the process indicators from reconciled data reflected the different process modes and observed process variations during the measurement campaign (Section 4.3.1). This represents the last step in Table 4.3 of interpreting the data reconciliation results. Then, the data reconciliation validity with respect to gross errors (Section 4.3.2) is evaluated along with how the data quality impact the reconciled values (Section 4.3.3). Last, gross error detections due to combination of different mass balance groups (Section 4.3.4) and from different choices of the error covariance matrix is studied (Section 4.3.5).

4.3.1 Tracking process modes with process indicators

The process indicators (Table 4.2) computed from reconciled measurements and the *coreMB* are given in *Figure 4.2(a)* and *Figure 4.2(c)*, during the different process modes (*Figure 4.2(b)*, Table 4.1). The periods with high data error index (*Figure 4.2(d)*), mainly due to gross error detections indicates unreliable reconciled values and are considered separately in Section 4.3.2.

The nitrification efficiency was almost complete (>95%, see *Figure 4.2(a)*) apart from an initially low de-/nitrification efficiency in week 1 (*Figure 4.2(a)*). The low de-/nitrification efficiencies during that week were caused by heavy rain and snow-melting, which required a temporally change of the aerated and unaerated flex zones (*Figure 4.1*), to maximize the available nitrification capacity and reduce the pre-denitrification volume accordingly. The aerated zones (nitrification volume) were increased by 20 percent, which was still not sufficient to fully compensate for the reduced nitrification efficiency. The specific nitrification rate was larger than the theoretical one during week 1-15 (*Figure 4.2(c)*). During this period, there was a trend in increased theoretical nitrification rate (thin black line, *Figure 4.2(c)*) due to an increase in temperature (data not shown). A similar trend during the same time period was also seen in the reconciled nitrification rate, but with an off-set (thick black line, *Figure 4.2(c)*).

Considering denitrification, the effect from the high carbon dosage mode (H1-H3) was most evident during H3 where the denitrification efficiency increased (*Figure 4.2(a)*). By contrast, during H1 the denitrification efficiency remained constant. This was similar also for the weeks before H1 with successive changes in the process modes (*Figure 4.2*(a-b)). During H2, the nitrate recirculation pump was only shut down during a short period of time, which makes H2 more similar to the normal carbon mode and therefore unsuited for comparison with H1 and H3.

During both H1 and H3, the specific denitrification rate remained constant and similar to the preceding weeks with low and normal carbon dosage modes (*Figure 4.2(c)*). This was despite the shut-down of the nitrate recirculation (QNO/Qasp, *Figure 4.2(b)*), which was expected to substantially decrease the pre-denitrification and thereby the overall specific denitrification rate. To further assess this, five grab samples of the in- and effluent nitrate to the post-denitrification were analysed during H1. The mean value of the grab samples (circle, *Figure 4.2(c)*) indicated a lower denitrification rate of 0.016 gN/g,VSS, day, compared to the 0.065 gN/g,VSS, day as indicated by the reconciled rate (*Figure 4.2(c)*). Note that the specific rate for the data reconciliation (and the grab samples) was based on the total denitrification volume, including the pre-denitrification. This results in a low overall specific denitrification rate when, in principle, the pre-denitrification volume was inactive. Finally, the total reconciled denitrified nitrogen mass during H1 and H3 was large (900–1000 kg/day) compared to what could have been obtained with

only *Qras* (20–50 kg/day) in combination with the post-denitrification (about 100 kg/day), contributing to nitrate reduction. In the end, these findings indicated a substantial and unexpected pre-denitrification, possibly due to nitrate recirculation despite inactive nitrate recirculation pumps. Once this was realized, the operational staff confirmed similar experiences and that the recirculation pipe could have operated as a syphon since no valves were closed after the pumps. The syphon effect was not possible to verify after the measurement campaign since the nitrate recirculation pumps were moved. A test was however performed one year after the measurement campaign, where the nitrate recirculation was shut down (including the following valves in the pipe). This resulted in a doubled nitrate concentration in the effluent (7 mg/L to 16 mg/L), and likewise in the return sludge. Due to this high nitrate concentration, a 50 percent denitrification was obtained due to the return sludge recirculation. However, during the measurement campaign and H1 and H3, the nitrate effluent concentration was low (3–6 mg/L) and would therefore have needed a higher recirculation flow to achieve the observed denitrification rates. Altogether, these indications supported the hypothesis that the suspected syphon effect was the reason for the discrepancy between specific denitrification rate of grab samples and the reconciled values.

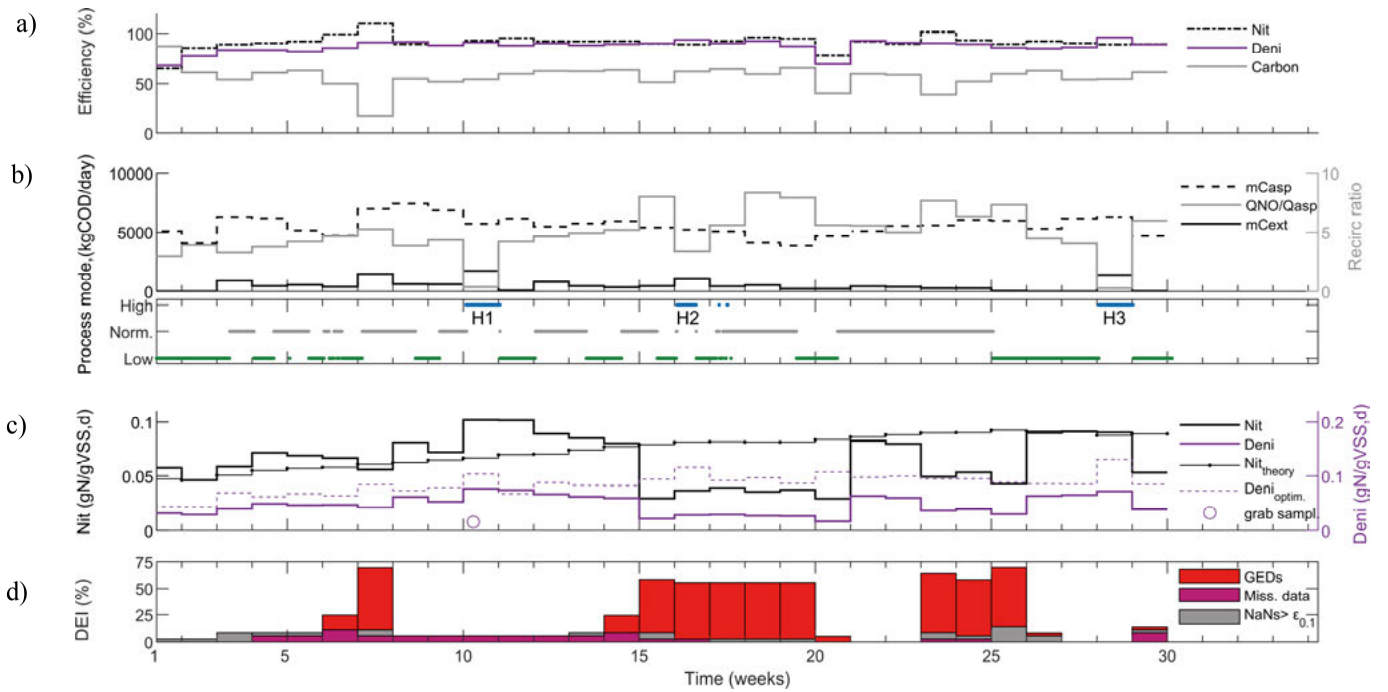


Figure 4.2. Weekly reconciled process indicators (Table 4.2) efficiency measures (**a**) and process rates (**c**) for nitrification (Nit) and denitrification (Deni) during different process modes (**b**). The data error index (DEI) (**d**) show the three components gross error detections (GEDs), missing data (Miss. Data), and proportion of measurements with more than 10 percent Not-a-Number indications (NaNs) caused by anomalies in the original time series signals. The three periods without nitrate recirculation (QNO/Qasp) and with high external carbon dosage (mCext) are highlighted with H1, H2, and H3, where the influent carbon (mCasp) is indicated as comparison. Note that H2 only covers part of the weekly time window. The measurement campaign ranges from the 11th of March (week 1), until the 1st of October (week 30) during 2019.

4.3.2 Gross errors in the flow rate and off-gas measurements

The abundance of gross error detections during week 6-7, 14-20, and 23-25 for the c*oreMB* were evident in *Figure 4.2*, but also for all other mass balance group combinations (*Figure 4.3*(a-i)). This indicates that the flow measurements were the underlying reason (since they were the common variable in all mass balance groups). Indeed, the gross error detections raised in the *ssMB* (that contained only flow and suspended solids measurements) were caused by flow measurements. Their causal problems in practice were identified in the process engineer's logbook as maintenance work and process changes. These problems influenced the flow measurements and are further described in the following sections.

Gross error detections in week 6-7, 20 and 23-25 were caused by a temporary shut-down of the sand filter, which reduced the sand filter flow (*Qeff* and *Qfil*) from 5000 m^3/day to about 100-200 m^3/day. The reasons for closing the sand filter was to replace sand (week 6-7) and to repair a broken hatch (week 20, 23-25). Such changes should not raise any gross error detections since the changed flow is still described by the mass balances and were therefore regarded as a false gross error detection.

By contrast, the gross error detections in week 14-19 were 'true errors' caused by a temporary process change, which in turn, induced a model error. During that time, the flow to the sand filter (*Qfil*) was deliberately mixed with the full WRRF flow to verify the functionality of the effluent phosphate sensor (this was needed for evaluating the pilot plant and should be considered as a known disturbance from the viewpoint of this study). When the flows were mixed, the *Qfil* sensor measured zero at the same time as *Qeff* measured the normal full flow, which clearly deviated from the mass balance (row 4 in Table 4.4). From the viewpoint of the mass balance model, all suspended solids then needed to exit the system via the WAS, which caused unrealistic high reconciled suspended solids concentrations (up to five times larger reconciled values, than the measured values).

The off-gas measurements also induced problems. First, the manual measurement campaign for assessing incoming and effluent ventilation air produced inconsistent results (1,170,000 Nm^3/day incoming air and 800,000 Nm^3/day off-gas airflow rate). This inconsistency was then needed to be compensated for during reconciliation, which resulted in adjustments of the airflow rate rather than in the off-gas component concentrations. Next, the gas composition from the activated sludge process was diluted with a large ventilation airflow, which resulted in a mixed off-gas composition similar to outdoor air. As an example, the expected decrease in O_2 off-gas (compared to ventilation air) was less than 10 000 ppm (1% O_2) during normal nitrification.

Last, the off-gas pump broke down during week 4-14, which resulted in a sustained period with constant O_2 and CO_2 off-gas measurements, i.e. missing data (*Figure 4.2(d)*).

4.3.3 Data quality impact on reconciled values

The impact from the data error index on the reconciled process indicators was, however, unpredictable. During week 14 there were both missing data and gross error detections, but no evident impact on the specific de-/nitrification rates (*Figure 4.2*(c-d)). By contrast, there were clear step decreases in de-/nitrification rates during week 15-21 and week 23-25 (*Figure 4.2(c)*), which, in part, coincided with the true gross error detections during week 14-19, and in part, coincided with the false gross error detections week 20, 23-25. During week 29, a substantial change was seen in the reconciled rates, although only a small number of gross error detections in combination with missing data were present.

4.3.4 Balancing constraint complexity and gross errors

As described in Section 4.2.5, groups of mass balance constraint were added component-wise to increase the redundancy. How an increase of these constraints influences the gross error detections is shown in *Figure 4.3*.

When only part of the constraints and related measurements were used, for example the mass flows of suspended solids (*ssMB*), less gross error detections were obtained compared to when all balances were used (compare *Figure 4.3(a)* and *Figure 4.3(i)*).

The absence of gross error detections in *ssMB*, apart from the time with flow errors, indicates that the suspended solids data were reliable. However, when they were combined with the *coreMB*, gross error detections were raised in both measurement and nodal tests (*Figure 4.3(e)*). During the entire experimental period, the conversion factors *SS2CODwas* and *SS2TKNwas* were reconciled to unrealistic values (even negative). Many of the suspended solids measurements were also reconciled to unrealistic values, but not sufficiently deviating to raise a gross error detection due to their large error variance. These were instead detected with the ad-hoc test (*Figure 4.3(e)*). This indicates that the suspended solids data were not compatible with additional constraints, which made further use of the suspended solids conversion factors.

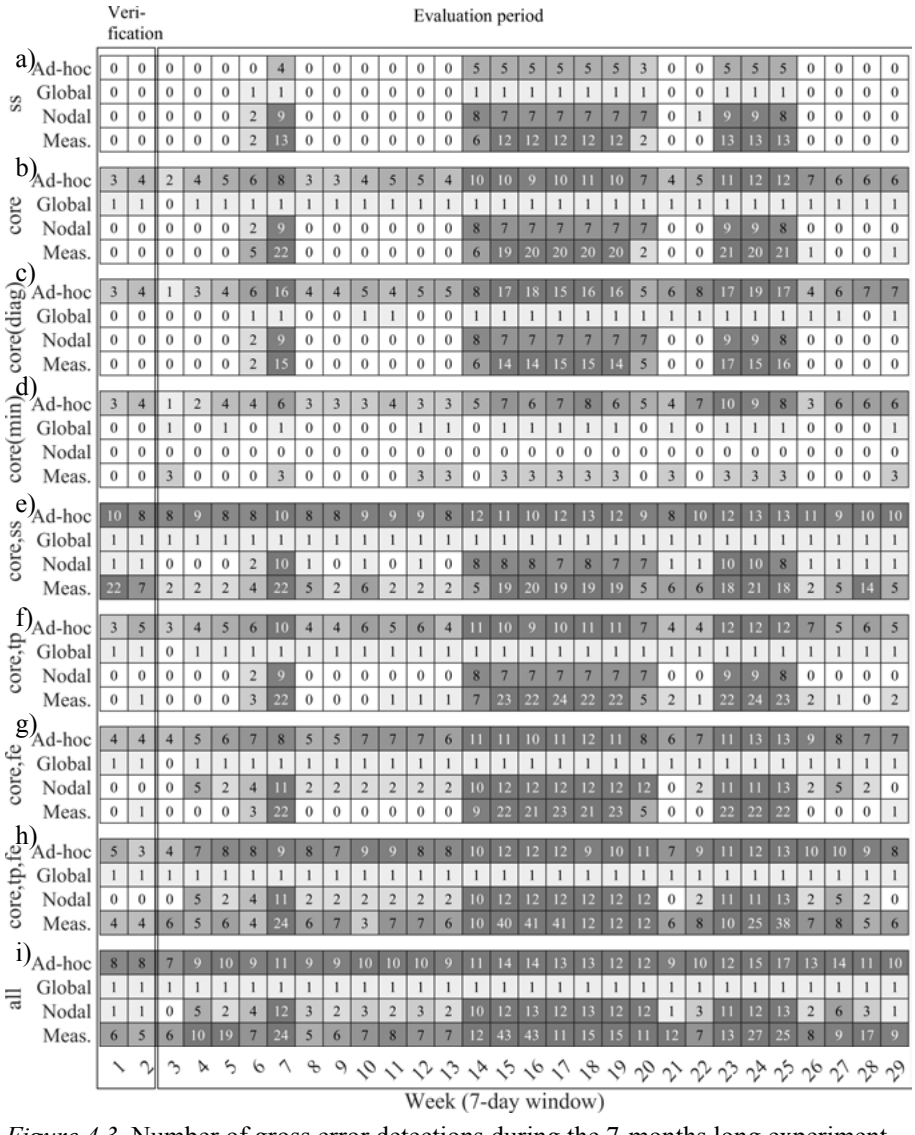


Figure 4.3. Number of gross error detections during the 7-months long experiment for different groups and combinations of mass balances and error matrices. Gross error detections are shown for the Ad-hoc, Global, Nodal, and Measurement (Meas.) test. Results from top (**a**) to bottom (**i**) are shown for increasingly larger sets of mass balance constraints, starting with *ssMB*(**a**), the *coreMB* with error matrix Σ_{eff} (**b**), diagonal error matrix (**c**), and minimum error matrix Σ_{min} (**d**). Then, combinations of the *coreMB* with *ssMB* (**e**), *tpMB* (**f**), *feMB* (**g**), both *tp*- and *feMB* (**h**), and all mass balances in Table 4.4 (**i**) are shown.

Related to the suspended solids and conversion factors were the COD sludge analyses, which were problematic. Initial *CODwas* analyses showed about 1000 mg/L, which was unrealistically low as indicated from the mass balances. The (accredited) lab then re-analysed all samples twice, which resulted in more reasonable mean values of about 10 000 mg/L. This illustrates that the conversion factors relating suspended solids to COD were affected by uncertainties in both its constraint assumptions (A4.3, Table 4.7) and analytical procedures.

The raised ad-hoc gross error detections for the *coreMB* (*Figure 4.3(b)*) indicated large adjustment in the measurements for combined COD and nitrate sensor, the ventilation and off-gas airflow rates, and the *SSwas*. All adjustments were identified as true errors and possible to explain, apart from the *SSwas* that instead supports the suspicion of an inconsistency in the *ssMB*. The COD and nitrate sensor had a known drift due to biofilm fouling (despite weekly to twice-weekly manual cleaning), which was compensated for by the weekly lab samples. The inconsistency in airflow rates was explained in Section 4.3.2.

Any additional combinations of *core-*, *ss-*, *tp-*, and *feMB* (*Figure 4.3(e)-(i)*) resulted in severe measurement and model inconsistencies as indicated by the presence of gross error detections. This further indicates that neither of the combinations in *Figure 4.3(e)-(i)* were feasible for computing reconciled process indicators.

4.3.5 The influence from the error matrix on gross error detection

The ad-hoc approach for assigning the error matrix (core(min)) produced deviating results, both in terms of number of gross error detections and their distribution, compared to the suggested method described in Section 2.8 (compare *Figure 4.3(b)* and *Figure 4.3(d)*). The minimal standard deviation for the mean (std) was identified to 0.36 std during the verification period, which was much larger than the assumed 0.01 std for the flows. This is in-line with why fewer gross error detections were identified during the periods with errors related to the flow measurements.

The effect from introducing correlated sampling errors in the covariance error matrix (Step 2, Table 4.5) had only a minor impact on gross error detections (compare *Figure 4.3(b)* and *Figure 4.3(c)*). By contrast, the reconciled denitrification rate was substantially lowered when only the diagonal of the error covariance matrix was considered (compare solid and dashed black lines *Figure 4.4*). In fact, the change in reconciled denitrification rate due to a diagonal error covariance matrix was, for a part of the time, larger than change from the combination of the *core-*,*fe-,* and *tpMB* (which produced abundant

gross error detections). The denitrified mass was used as an example of a reconciled process indicator in *Figure 4.4* (by contrast to the specific rate normalized by VSS), due to the unfeasible reconciliation of suspended solids measurements, which produced occasional negative values. The basic option to only consider measurements and mass balances (without data reconciliation) is given with red dashed line in *Figure 4.4* as comparison.

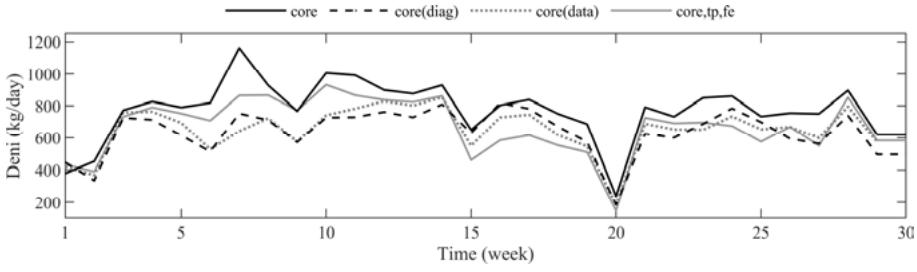


Figure 4.4. Reconciled denitrification (Deni) rate obtained from using *coreMB* with a correlated error covariance matrix (solid black line, 'core'), diagonal error covariance matrix (dashed black line, 'core(diag)'), and without reconciliation (red dotted line, 'core(data)'). The combinations of *coreMB*, *tpMB* and *feMB* are given as examples and for comparison of the impact of using a constraints with abundant gross error detections (grey solid line, 'core,tp,fe').

4.4 Discussion

First, we discuss the challenge with the abundance of gross error detections (Section 4.4.1) and whether they were caused by assumptions related to the suspended solids measurements and mass balances (Section 4.4.2). Second, we consider the value with off-gas measurements in perspective to data reconciliation (Section 4.4.3). Third, we reflect upon how violation of theoretical assumptions impacts the practicability of data reconciliation and the process indicators (Section 4.4.4). Last, we summarize pitfalls and benefits with applying data reconciliation in practice (Section 4.4.5) along with identified shortcomings that need further studies (Section 4.4.6).

4.4.1 Gross error detection

The evident result in Section 4.3.2 was the abundance of gross error detections when several mass balance constraints were combined. Recall from Section 4.2.5 that several mass balances were grouped component-wise (e.g. Tot-P and Tot-Fe) to increase the constraint redundancy and maximize the usage of existing data. The results, however, indicated that it may be a draw-back to make use of large systems of mass balances and related measurements. The

spread of gross error detections in *Figure 4.3* instead indicated a trade-off between maximizing the usage of available process information in terms of measurements and mass balances; and obtaining reconciled variable estimates without gross errors. It is natural that the probability to have one sensor error increases when the number of measurements increase, and it also becomes increasingly difficult to close an increasingly large system of mass balances, without gross errors. But the abundance of gross error detections demonstrates a great benefit with the data reconciliation – its ability to assess the data quality in terms of gross error detections based on process knowledge encoded in the mass balance constraints.

The sensitivity for raising a gross error detection was clearly influenced by the error matrix. The suggested method in Table 4.5 combined with the measurement test was too insensitive and missed a COD sensor drift and unreasonable suspended solids values. These deviations were instead detected by the ad-hoc test. This emphasizes that there is a balance between how pragmatic and theoretically rooted the tuning of the gross error detection test should be. Both the Bonferroni correction and the assumed sampling errors should therefore be reconsidered to enable a more practically relevant gross error detection level.

The large impact from the error covariance matrix on the reconciled values highlights that it is infeasible to use an ad-hoc approach for assigning the error covariance matrix. Such ad-hoc approaches could produce reconciled results, which would doubtfully be better than the original measured values. Note that it is the ratio between the variances in Σ and equation (4.2) for all variables, which weigh how the errors in data reconciliation are reconciled. Thus, having reasonable *ratios* between the errors for the individual variables is of main importance, rather than the *absolute* error magnitudes when considering reconciled values.

How to diagnose the root cause fault that produce gross error detections needs further studies. For the flow errors, the error was smeared out to additional sensors, which complicates the identification of the root cause error. It would be interesting to see how a sensitive analysis could complement the interpretation of gross error detections and the error root cause analysis. A sensitivity analysis would also be valuable to increase our understanding of the data reconciliation with all the nested variables in the mass balance constraints.

4.4.2 Assumptions about suspended solids

The assumption of constant suspended solids conversion factors needs to be further evaluated as the *ssMB* were not possible to combine with the remaining balances without gross error detections and unrealistic reconciled values. It is likely that errors in suspended solids balances (either the assumptions or measurements) produced consequential errors via the conversion factors *SS2TKN*

and *SS2COD*. Also, the representativity of using sludge grab samples should be questioned. The sampling errors and measurement errors, for example, for COD in WAS are likely large but to our knowledge unquantified. This therefore needs to be better assessed in practice to increase the realism of mass balance-based data reconciliation.

Similarly, reconciliation based on information from the sludge streams and Tot-P and Tot-Fe mass flows induced both nodal and measurement test gross error detections. It is likely that the changes in different process modes in the pilot-line opposed the assumptions about a constant Tot-P content in the sludge. Reconciling mass flows of Tot-P to increase data quality is otherwise recommended prior modelling studies (Rieger et al. 2012), which rely on more stable conditions.

The problems related to suspended solids conversion factors emphasize that the sludge streams measurements were insufficient. For the suspended solids, either on-line sensors or automatic samplers could reduce the sampling error. For the nitrogen components, Tot-P, and Tot-Fe in the sludge, additional samples are needed. It is still uncertain whether such action would be enough to assure that measurements in the sludge stream would contribute to the overall data reconciliation accuracy. Ultimately, the results demonstrate the importance of selecting and combining only those mass balance constraints and related measurements that increase the accuracy of the reconciled measurements and the process indicators (instead of the opposite). In the end, the data reconciliation is no exception from the general assumption that poor input data also produce poor output conclusions.

4.4.3 Off-gas measurements for data reconciliation

The off-gas measurements provide an independent estimation of denitrification as complement to liquid phase measurements and have been identified by as key data for mass balance data reconciliation (Puig et al. 2008). Here, the denitrification rate estimated solely from off-gas data produced unrealistic and low values due to the inconsistent flow measurements. Despite these problems, off-gas measurements can be valuable for the purpose of reconciling the off-gas flow rate. The flow rates here were uncertain and can be essential to be reconciled when process indicators related to the off-gas are of interest. Leu et al. (2010) studied oxygen- and carbon dioxide transfer rates to monitor nitrification performance. Such rates could straightforwardly be obtained from the set of mass balances in this study, when the off-gas flow-rate is accurately reconciled. Also, monitoring greenhouse gas mass emission such as nitrous oxide would be facilitated with a reconciled off-gas flow rate.

4.4.4 Process rate monitoring by data reconciling

The results did not show the expected correlation between changes in process mode (*Figure 4.2(b)*) and denitrification process indicators (*Figure 4.2*(a,c)). An increase in denitrification process indicators was expected when, for example, the carbon was added in normal mode compared to the low carbon mode. The reason for this was likely that the addition of external carbon, as compared to the influent carbon, was too small to impact the process indicators. Note that the variations in influent carbon (mCasp) were larger than the variations in external carbon (mCext) between low and high carbon mode (*Figure 4.2(b)*). Only large changes such as the lowered capacity during snowmelting were noticed. Possibly, a toxic shock could be identified. Nevertheless, the identification of the anticipated syphon effect and incomplete shutdown of nitrate recirculation illustrates the value of putting a large set of measurements into context. The data reconciliation allows monitoring of mass flows, which produce a deeper process insight compared to monitoring concentration measurements signal-by-signal.

The belief that data reconciliation can improve the data quality and the process indicators, is reliant on four theoretical requirements that underlies equation (4.2). First, no gross measurement errors are present. It is only random measurement variations that cause inconsistencies in the constraints, which should be compensated for. Second, the random measurement variations can a priori be encoded in the covariance error matrix, commonly assumed to be Gaussian. Third, the constraints are true (i.e. the mass balance constraints are free from modelling errors). Last, the process is in steady-state for the given time window.

This study has demonstrated that these theoretical requirements may be infeasible to fulfil. For that reason, it is relevant to discuss whether the data reconciliation really improves the trueness of data (process information), and what random variations that can be accurately compensated for.

Random variations in time windowed data can be caused by many factors, but in the current study the commonly assumed noise in sensors is not a major error source. The reason is that we here consider mean values for one week, and any high frequent (zero mean valued) sensor noise will be practically eliminated due to the low-pass filtering effect from constructing such mean values. Instead, weekly temporal variations that randomly cause inconsistencies in the mass balance constraints are desired to remove during reconciliation. For example, slow temporal changes in the build-up of Tot-P or Tot-Fe in the sludge is expected, but not modelled, and can be considered as random variations. Here, these variations were likely too large to be used in reconciliation as indicated by gross error detections.

We deliberately violated the first two theoretical assumptions when sensor bias was included in the effective standard deviation, which was further used in the error covariance matrix. Biased sensor measurements are expected in

practice, but if they are considered as gross errors they should, according to theory, not be part of the reconciliation. The equation (4.2) indicates that it is the ratio of errors in the error covariance matrix that decide how an error is smeared out during reconciliation. Thus, we suggest that an initially suspected sensor bias could be accurately compensated for, as long as it is assigned in proportion to the other (variance) errors in the covariance matrix.

In the end, this study verifies the common assumption that sensor and model errors are present in practice. It is within this context the data reconciliation needs to be used, and to be useful, despite that this violates the theoretical requirements. The error covariance matrix construction is in this aspect key for the data reconciliation practicability. First, it can be considered as the tuning parameter for gross error detection sensitivity. Second, it decides how the unavoidable measurement errors are smeared out throughout the mass balance constraints. The trueness of reconciled values is directly linked to how well the error matrix describe the true measurement errors. This is apparently very difficult to know, which was one of the reasons for applying the data reconciliation in the first place. Thus, the data reconciliation feature to detect gross error detections is possibly more useful in practice, than as a tool to improve the accuracy in process monitoring.

4.4.5 Lessons learned

The results in this study demonstrated both benefits and challenges when applying data reconciliation in practice.

- The data reconciliation brings structure to data. It is difficult to overview all data flows and potential measurement errors without data reconciliation, which were here about 40 sensor tags and 30 different lab samples. The application of data reconciliation requires the organization of data in a structured way, where process knowledge in the form of mass balances is used for reconciliation. This enables process insights about mass flows that are not possible from concentration measurements alone, as was demonstrated with the syphon effect. The data quality assessment in terms of gross error detections also reduce the risk to misinterpret biased measurements for process changes, which was clearly demonstrated with the data error index.
- The data reconciliation provides a straightforward way for fusion and validation of equivalent lab and sensor measurements via the mass balances as indicated when the COD lab samples compensated for the COD sensor drift. This enables reusage of existing data.

- The data reconciliation provides a method to highlight doubtful data. This can guide plant operators in prioritizing data quality improvement actions.
- Weak spots in data collection that would be needed to close the mass balances are automatically identified during the data reconciliation construction. Understanding conversions and mass flows of nitrogen, phosphorus, and COD is after all the core business in any WRRF.

At the same time, there were challenges, which are also expected to be important elsewhere:

- The need for additional and reliable lab measurements. Data from the sludge streams were absent and were also difficult to sample and analyse accurately. We expect this unfortunate combination of critical, but absent, data to be common for most WRRFs.
- Sensitivity to temporary failing sensor measurements. Failing sensors are a normal situation in a WRRF, and there is a large likelihood that at least one measurement is missing during normal operations. If the failing sensor is also critical for the mass balances (e.g. a flow sensor), this can limit the usefulness of the data reconciliation completely with abundant gross error detections (worst scenario) or at least require a redefinition of the mass balances.

Ultimately, we see the practical value of data reconciliation as the combination of:

1. Overview of mass flows and component conversions
2. Detection of gross errors and model inconsistencies and data quality assessment
3. Reconciliation and estimation of unmeasured data in a repeatable and structured manner

4.4.6 Future work

This study pinpoints missing knowledge related to i) the theoretical and technical aspects of the data reconciliation method, and ii) knowledge needed in practice for a successful implementation. The following research is suggested to mitigate the identified shortcomings.

- Sensitivity analysis should be studied as a complementary tool for understanding cause and effect of discrepancies in the mass balances during the iterative data reconciliation procedure. Although the data reconciliation brings structure to data once in place, the

road to get there can be overwhelming with a large system of mass balances and measurements. Also, the impact of the many assumptions could easier be ranked in order of importance in the light of a sensitivity analysis.

- The impact from flow proportional concentration mean values was not possible to evaluate in this study due to disturbances in the flow signals. The impact of flow proportional mean values should be better studied in a computer simulation.
- How the covariance error matrix should be assigned needs further studies. This is especially important for the sensitivity during gross error detection.

The following practical aspects needs to be considered before wide-spread usage of the data reconciliation.

- What error magnitudes do water and sludge sampling have in practice, and how can these be reduced by an improved sampling methodology? The sampling error had a major impact here and was not assessed in practice but estimated from literature values.
- In relation to the previous aspect, it is essential to establish a method for defining the number of required samples to obtain representative conversion factors. The conversion factors were critical for how the reconciled errors were distributed in the mass balances.

4.5 Conclusions

Mass balance-based data reconciliation is a promising tool to gain process information from large process datasets, which here enabled computation of process indicators and detection of process faults such as the anticipated syphon effect in nitrate recirculation. The combination of mass balances and reconciliation allows an iterative, but still structured path for an increased understanding of process rates and the related data quality. But data reconciliation is no magic tool. The accuracy of reconciled values will suffer from poor data quality, model inconsistencies, and a poorly chosen error covariance matrix. Nonetheless, it is the capability to reveal these issues that makes data reconciliation valuable in practice.

We conclude that

- The ability to detect gross errors is promising and is possibly more useful than producing reconciled measurements.
- The error matrix construction has a large impact on both reconciled values and gross error detection and needs to be further studied as practical experiences about measurement and sampling errors are missing.
- A high data quality in both on-line sensors and lab measurements is desired to benefit from data reconciliation. Otherwise, the reconciled measurements can produce unreliable reconciled values and process indicators.
- Many measurements and a large system of mass balance constraints can decrease the usefulness of the reconciled values as the possibility to obtain gross errors increases.

Based on the results in this study we recommend the following prior implementing data reconciliation for process monitoring:

- Obtain redundant flow measurements.
- Install sludge samplers and budget for weekly analysis costs if data about historic sludge composition is absent.
- Carefully assess sampling and measurement errors for the current system
- Consider subsets of the complete treatment process for the data reconciliation.

Appendix A4.1 - Theoretical nitrification rate

This section derives the theoretical specific nitrification rate r_{nit} in Table 4.2.

The specific nitrification rate r_N is defined in (Sedlack 1991) as

$$r_N = \frac{\mu_{nit}}{q_N} \tag{4.7}$$

where μ_{nit} is the specific nitrification growth rate and q_N is the nitrification yield constant. The q_N describe the mass of *nitrification bacteria* in the volatile suspended solids, which are formed per mass of nitrate-nitrogen produced during to nitrification. The specific nitrification rate can also be normalized to the *total* volatile suspended solids by correcting for the percentage of nitrifiers δ_{nit} in the volatile suspended solids as

$$r_N \delta_{nit} = \frac{\mu_{nit} \delta_{nit}}{q_N} = r_{nit} \tag{4.8}$$

which result in the specific nitrification rate r_{nit} normalized by the total volatile suspended solids, which was given in Table 4.2.

The nitrification rate μ_{nit} can be estimated as a Monod-function of the maximum specific growth rate $\mu_{nit,max}$, which is reduced by the nitrification rate limiting substances ammonium $[NH_{eff}]$ and dissolved oxygen $[DO]$, and their corresponding half-saturation constants K_{NH} and K_{DO} as

$$\mu_{nit} = \mu_{nit,max} \frac{[NH_{eff}]}{K_{NH}+[NH_{eff}]} \frac{[DO]}{K_{DO}+[DO]}. \tag{4.9}$$

The nitrification rate is further temperature dependent, which can be described by the van't Hoff correction factor θ_T as

$$\theta_T r_{20} = r_{20}\, e^{(\varkappa(T-20))}, \tag{4.10}$$

with r_{20} being the reaction rate at 20°C, T the current temperature and $\varkappa$ is a temperature coefficient (Henze et al. 2002).

The fraction of nitrification bacteria δ_{nit} was estimated from the reduction in substrate concentrations in the process and yield coefficients from (Henze et al. 2002) using

$$\delta_{nit} = \frac{Y_{nit}\,[NH_{rem}]}{Y_{nit}\,[NH_{rem}] + Y_{het}[COD_{rem}]} \tag{4.11}$$

where, Y_{nit} and Y_{het}, are the yield coefficients, and $[NH_{rem}]$, and $[COD_{rem}]$ are the reduction in ammonium and COD substrate concentrations, for nitrification and heterotrophic bacteria, respectively.

The constants in (4.7)-(4.11) were obtained from (Henze et al. 2002) and substrate concentrations in (4.9) and (4.11) were mean values from the pilot plant, which are all given in Table 4.6.

Table 4.6. *Literature values, computed coefficients and assumed substrate concentrations used to compute the theoretical specific nitrification rate.*

Description	**Variable**	**Value**	**Unit**
Maximum specific nitrification growth rate	$\mu_{nit,max}$	0.7	1/day
Van't Hoff temperature coefficient	$\varkappa$	0.1	1/°C
Yield constants for nitrifier/heterotrophic bacteria	Y_{nit}/Y_{het}	0.15/0.39	g VSS/g substrate
Ammonium concentration effluent/reduced	$[NH_{eff}]/[NH_{rem}]$	0.5/45	mg/L
Dissolved oxygen concentration/COD reduction	$[DO]/[COD_{rem}]$	2/300	mg/L
Nitrifier fraction	δ_{nit}	0.05	-

Appendix A4.2 - Specifying the error covariance matrix

The error covariance matrix Σ for two variables X, Y can be obtained as the sum of two independent error covariance matrices representing the correlated sampling error Σ_{tse}, and the total analytical error Σ_{tae}.

$$\Sigma = \Sigma_{tse} + \Sigma_{tae} \tag{4.9}$$

The Σ_{tae} contain the independent analytical error variances in the diagonal

$$\Sigma_{tae} = \begin{pmatrix} \sigma_{tae,X}^2 & 0 \\ 0 & \sigma_{tae,Y}^2 \end{pmatrix} \tag{4.10}$$

which are obtained from the analytical error in terms of one standard deviation and percent, $\sigma_{1std,tse}$, of the measured value x, $\sigma_{tae,\mathrm{X}}^2 = (x\sigma_{1std,tse})^2$. Similarly, the sampling error can be described by an error matrix

$$\Sigma_{tse} = \begin{pmatrix} \sigma_{tse,X}^2 & \sqrt{(\sigma_{tse,X}\sigma_{tse,Y})^2} \\ \sqrt{(\sigma_{tse,X}\sigma_{tse,Y})^2} & \sigma_{tse,Y}^2 \end{pmatrix} \tag{4.11}$$

with the difference that the off-diagonal elements contain the covariance of the sampling error, which can be shown from the Cauchy-Swartz inequality. This results in the final error matrix

$$\Sigma = \begin{pmatrix} \sigma_{tae,X}^2 + \sigma_{tse,X}^2 & \sqrt{(\sigma_{tse,X}\sigma_{tse,Y})^2} \\ \sqrt{(\sigma_{tse,X}\sigma_{tse,Y})^2} & \sigma_{tae,Y}^2 + \sigma_{tse,Y}^2 \end{pmatrix} \tag{4.12}$$

where Σ is the covariance matrix containing the independent analytical error variance σ_{tae}^2 and the correlated sampling error σ_{tse}^2.

The Cauchy-Swartz inequality for correlated random variables states that

$$|\mathrm{E}[XY]| \leq \sqrt{\mathrm{E}[X^2]\mathrm{E}[X^2]} = \sqrt{\sigma_{\mathrm{X}}{}^2\sigma_{\mathrm{Y}}{}^2} \tag{4.13}$$

With equality iff $X = \alpha Y$ for a constant $\alpha \in \mathbb{R}$. Since X and Y are both Gaussian distributions with variances proportional to the measured mean value

$$X \sim N(x, (x\sigma_{1std,tse})^2, Y \sim N(y, (y\sigma_{1std,tse})^2 \quad (4.14)$$

They can be scaled to equal distributions if α is chosen so that $\alpha X = Y$

$$Y = \alpha X \sim N(\alpha x, (\alpha x \sigma_{1std,tse})^2. \quad (4.15)$$

Appendix A4.3 – Assumptions related to mass balance constraints

The Table 4.7 detail mass balance assumptions in Section 4.2.7.

Table 4.7. *Assumptions made during data collection (DC), component composition (CC) and mass balance modelling (MBs)*

	Assumption	Related to
A1	Nitrate in WAS was zero	DC
A2	Nitrate in RAS was equal to nitrate in the secondary settler	DC
A3	The off-gas air was assumed to be dry without any water vapour as the gas sensor was located after a condensation (dehumidifying) unit.	DC
A4	Incoming fresh air contained 400ppm CO_2, 9300 ppm argon, 78.078% N_2, 20.95% O_2 at a mean temperature of 15 °C.	DC
A5	The suspended solids were conserved as a component (related to mass balance 22)	CC
A6	The ratio of TKN/SS were equal in WAS, RAS and ASP	CC
A7	The ratios between suspended solids; and TKN, COD, Tot-Fe, and Tot-P were constant in WAS.	CC
A8	The ratio between DS; and Tot-Fe and Tot-P, were constant in PS	CC
A9	The composition of external carbon source (Brenntag Brenntaplus) was equal between delivered batches	CC
A10	The biomass and suspended solids in pre-, and post-sedimentation were assumed to be constant and in steady-state, i.e. no build-up of biomass was included in the mass balances.	MBs
A11	No nitrification occurred after the nitrification zone 6, with the resulting assumption that TKN was conserved throughout the secondary settler.	MBs
A12	No denitrification occurred in post-sedimentation, sand filter or in the RAS	MBs
A13	The flow from external carbon source was negligible in comparison to the magnitude of water and sludge flows and excluded from the flow mass balances	MBs
A14	Back-wash flow in sand filter was excluded from flow balance in sand filter	MBs

4.6 Chapter Acknowledgments

Considerate efforts have been performed by the staff at Käppala WRRF in order to collect, install new sensors, and analyse misbehaviour of sensors and equipment. We gratefully acknowledge their efforts, and acknowledge the contributing staff in alphabetical order: Ewelina Basiak-Klingspetz, Sofia Bramstedt, Stefan Erikstam, Frida Jeppson, Olle Lindberg, Tofik Massiri, Michael Médoc, Pontus Nordgren. We also acknowledge the constructive review on the initial manuscript from Adrienne Menitti (Clean Water Services).

5 Gaussian process regression for monitoring and fault detection of wastewater treatment processes

The purpose with this chapter is to evaluate a promising machine learning method, Gaussian process regression (GPR), for WRRF monitoring applications. We evaluated GPR at two WRRF monitoring problems: estimate missing data in a flow rate signal (simulated data) and detect a drift in an ammonium sensor (real data).

The results showed that GPR with the standard estimation method, maximum likelihood estimation (GPR-MLE), suffered from local optima during estimation of kernel parameters, and did not give satisfactory results in a simulated case study. However, GPR with a state-of-the-art estimation method based on sequential Monte Carlo estimation (GPR-SMC) gave good predictions and did not suffer from local optima.

Comparisons with simple standard methods revealed that GPR-SMC performed better than linear interpolation in estimating missing data in a noisy flow rate signal. We conclude that GPR-SMC is both a general and powerful method for monitoring full-scale WRRFs. However, this study also shows that it does not always pay off to use more sophisticated methods. New methods should be critically compared against simpler methods, which might be good enough for some scenarios.

5.1 Introduction

Monitoring methods are important tools for detecting process and sensor faults in water resource recovery facilities (WRRFs). Faults that influence the treatment process have to be detected and corrected to manage increasingly strict effluent demands. At the same time, automatic process control is an essential part of an efficient treatment process and is dependent on reliable sensor measurements.

Large amount of process data has been available at WRRFs for a long time, and the most studied method to exploit the data for monitoring purposes, is principal component analysis (PCA). Although proposed for WRRF applications more than thirty years ago (Aarnio and Minkkinen 1986), PCA is still an

active research area (Haimi 2016). Some of the research has been done to extended PCA to handle non-linear interactions in kernel-PCA (Lee et al. 2004), multiple models in Multiway PCA (Yoo et al. 2007), and in combination with other data processing methods (Qin et al. 2012).

Similarly as at WRRFs, the amount of available data has increased in many other areas, especially where the Internet has been a driving force. Discovering that data are not equivalent with useful information, large efforts have been done to develop computer-based algorithms that extract information from big data. This has resulted in a research area called machine learning, which can be defined as: methods that can learn and make predictions from data (Kohavi and Provost 1998). Machine learning has successfully been applied for monitoring processes other than WRRFs. Some examples include: Gaussian process regression for monitoring the melting index of polyethene in multigrade process (Liu et al. 2015), Deep learning for monitoring product quality in petroleum refinery process (Shang et al. 2014), and Support vector machines and *k*-nearest neighbourhood for classification of faults in a chemical process with incomplete data (Askarian et al. 2016). Although most machine learning methods are not new methods, they might still be new in a WRRF monitoring context and therefore interesting to study.

Gaussian process regression (GPR) is one machine learning method, promising for monitoring applications at WRRFs. The method is flexible (can handle non-linear problems), non-parametric (does not require parameter selection), and probabilistic (predictions include an uncertainty estimate) (Rasmussen and Williams 2005). In fact, these three properties have proved to be useful to solve problems related to monitoring including change detection (Garnett et al. 2010a), process modelling (Ranjan et al. 2016), and fault detection (Boškoski et al. 2015). However, few papers have studied GPR applied to WRRFs, see (Ažman and Kocijan 2007, Južnič-Zonta et al. 2012, Liu et al. 2016), of which only (Liu et al. 2016) studied how GPR could be used for monitoring purposes. Thus, GPR has not been studied to the same extent as PCA despite its promises.

Although GPR is a promising method, it has two potential drawbacks that could limit its usage in full-scale applications. First, a proper kernel (parametric autocovariance function) has to be selected by the user, which is a key step in GPR. This is commonly described as a strength of GPR, since the user can include prior knowledge of the system by selecting a proper kernel (Rasmussen and Williams 2005). However, most papers that apply GPR only use the standard kernel and do not consider other kernels. Also, a separate research field has developed on how to construct new kernels, see e.g. (Duvenaud. 2014). This indicates that kernel selection may not be as straight forward as originally claimed, which could make it hard for a WRRF engineer to adopt GPR as a new method.

The second potential drawback concerns the estimation of a kernel´s parameters. Commonly, the parameters are obtained by maximum likelihood estimation (MLE). However, this estimation method is also known to give bad solutions if several local optima exist (non-convex optimization problem). Although the risk with local optima has been acknowledged in both textbooks and papers, few papers quantify its impact when GPR is applied to real problems. Any full-scale monitoring method has to be robust and deliver correct results every time, and it is therefore crucial to evaluate the impact of local optima of GPR in WRRF applications.

To summarize, we know that GPR is a promising method that could improve monitoring of WRRFs, but we lack knowledge of the impact of some potential drawbacks. Further, we know little about GPR applied to WRRFs and therefore also, how good GPR is compared to existing WRRF monitoring methods.

In this chapter, we give a detailed introduction to GPR and how it can be applied for monitoring applications at WRRFs. Further, we study whether the promises of GPR are valid in a WRRF monitoring context, and whether the potential drawbacks of GPR can be handled.

We introduce the theory of GPR and two variants of GPR: 1) GPR with maximum likelihood estimation (GPR-MLE), which is the standard approach, and 2) a state-of the-art variant of GPR that uses a sequential Monte Carlo approach for estimation (GPR-SMC) (Svensson. et al. 2015). Both GPR variants use the same model structure, a Gaussian process (GP), but they approximate and estimate the GP´s parameter distribution differently.

First, we evaluate the potential drawback with local optima in a simulated case study, where the monitoring problem is to estimate missing data in a flow rate signal. We also study whether a priori of the flow rate signal (periodicity) is easily included in the kernel. For both potential drawbacks, GPR-SMC and GPR-MLE are compared in the simulated case study.

Second, we evaluate the most promising method, GPR-SMC, on real data with the monitoring task to detect a drift in ammonium on-line sensor values. As always, new methods should be critically evaluated and compared against existing methods. Here, we used linear interpolation in the first case study and linear regression in the second to assess the benefit of using GPR compared to existing simple methods.

5.2 Material and Methods

Gaussian process regression (GPR) combines properties from multivariate Gaussian distributions with Bayesian statistics to obtain a regression model that is flexible and give an uncertainty estimate of the predictions. First, we introduce the theory about GPR. Second, we introduce the two GPR variants,

GPR-SMC and GPR-MLE, and their differences in approximating and estimating the kernel parameter distributions. Thirdly, we give a step-by-step instruction for applying GPR-MLE and GPR-SMC in practice. Finally, we describe the two case studies in this chapter.

5.2.1 Gaussian processes

Gaussian processes (GPs) have been popular for two properties. First, a GP is a stochastic process that is completely defined by the multivariable Gaussian probability distribution. Second, the best unbiased prediction is a linear combination of previously observed values.

A GP is a distribution of functions with function values $f(x)$ at inputs x defined as

$$f(x) \sim GP\left(m(x), k(x_i, x_j)\right) = N\left(m(x), k(x_i, x_j)\right), \tag{5.1}$$

where x is size $n \times m$ with n inputs and m variables, and $k(x_i, x_j)$ expresses the covariance between input i and j, where $i, j = \{1, \ldots, N\}$. Thus, the indices i and j are not necessarily time indices, but indicate the different inputs in a specific data set. Note that $k(x_i, x_j)$ is an $n \times n$ matrix regardless of the number of variables. In (5.1),

$$m(x) \triangleq \mathbb{E}\left[f(x)\right] \tag{5.2}$$

is the mean function, and the covariance function is defined as

$$k(x_i, x_j) \triangleq \mathbb{E}\left[f(x_i) - m(x_i)\right]\left[f(x_j) - m(x_j)\right]^T, \tag{5.3}$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator.

The best unbiased prediction $f(x^*)$ at one unobserved x^*, is a linear combination of previous function values $f(x)$. This holds from the general property that two jointly Gaussian vectors, here $f(x)$ and $f(x^*)$

$$\begin{bmatrix} f(x) \\ f(x^*) \end{bmatrix} \sim N\left(\begin{bmatrix} m(x) \\ m(x^*) \end{bmatrix}, \begin{bmatrix} k(x_i, x_j) & k(x_i{}^*, x_j) \\ k(x_i{}^*, x_j)^T & k(x_i{}^*, x_j{}^*) \end{bmatrix}\right) \tag{5.4}$$

are also conditionally Gaussian distributed, with mean

$$m(x^*|x) = k(x_i{}^*, x_j)^T \, k(x_i, x_j)^{-1} \, f(x) \tag{5.5}$$

and covariance matrix

$$k(x^*|x) = k(x_i{}^*, x_j{}^*) - k(x_i{}^*, x_j)^T \, k(x_i, x_j)^{-1} k(x_i{}^*, x_j), \tag{5.6}$$

this can be written as

$$f(x^*|x) \sim N\big(m(x^*|x), k(x^*|x)\big) \tag{5.7}$$

where $m(x^*|x)$ should be read as the mean of x^*, conditioned on x. In other words, given the function values $f(x)$, we obtain the predicted mean $m(x^*|x)$ as a linear combination of the covariance between observed and unobserved inputs $k(x_i^*, x_j)^T k(x_i, x_j)^{-1}$, and of the observed function values $f(x)$. From (5.6) we also get the variance $k(x^*|x)$ of the predicted mean, which is used as an uncertainty measure of the prediction.

One limitation with GPR is that a large $n \times n$ covariance matrix has to be inverted to obtain the predictions in (5.5) and (5.6). The computation time increase exponentially with n^3, which limits the training data to a few thousand data. Approximations for handling large data have been suggested (Quiñonero-Candela and Rasmussen 2005) but were not considered in this chapter.

5.2.2 Covariance functions

In machine learning literature, a covariance function $k_\theta(x_i, x_j)$ is called a *kernel,* where θ refers to its parameters. The most common kernel is the squared exponential kernel

$$k_{Single}(x_i, x_j) = \theta_1 \exp\left[-\frac{(x_i - x_j)^2}{2\theta_2^2}\right] \tag{5.8}$$

with the parameters θ_1 and θ_2. Many other kernels have been suggested, see (Duvenaud. 2014) for a recent work about the properties of different kernels applied to GPR.

5.2.3 Regression with Gaussian processes

In a general regression problem, we observe data $y(x)$ and try to describe the underlying function $f(x)$ that is corrupted by additive noise $\varepsilon(x)$

$$y(x) = f(x) + \varepsilon(x). \tag{5.9}$$

As an example, a linear regression problem would be to estimate θ_1 and θ_2 in $f(x) = \theta_1 + \theta_2 x$. The standard method to solve this problem is ordinary least squares (OLS), which is described in statistical textbooks, see e.g. (Kay 1998).

In GPR, both $f(x)$ and $\varepsilon(x)$ are modelled as two independent GPs, where $\varepsilon(x)$ is described with a constant noise θ_{σ^2}

$$\varepsilon(x) \sim GP\left(0, k_{WN}(x_i, x_j)\right), k_{WN}(x_i, x_j) = \theta_{\sigma^2}\, \mathrm{I}_n, \tag{5.10}$$

where I_n is the identity matrix. Following the law of variances for independent covariance functions, the covariance of the final GP model $\mathcal{M}_\theta(x)$ is the sum of the two covariance functions $k_{Single}(x_i, x_j)$ and $k_{WN}(x_i, x_j)$, therefore we have

$$\mathcal{M}_\theta(x) \sim GP\left(m(x), k_\theta(x_i, x_j)\right), k_\theta(x_i, x_j) = k_{Single}(x_i, x_j) + k_{WN}(x_i, x_j) \tag{5.11}$$

where the regression task is to estimate the parameters θ from a data set X, Y. Here, X and Y include all observations $i = 1, \dots, n$ for x_i and y_i. See (Perez-Cruz et al. 2013) for a comparison between traditional and Bayesian regression with GPR.

5.2.4 Bayesian regression

Bayesian regression relies on Bayes rule which combines prior knowledge $p(\mathcal{M}_\theta)$, together with the likelihood $p(Y|\mathcal{M}_\theta, \mathrm{X})$ of a dataset X, Y, resulting in a posterior probability distribution $p(\mathcal{M}_\theta|X, Y)$ according to

$$p(\mathcal{M}_\theta|X, Y) = \frac{p(Y|\mathcal{M}_\theta, X)p(\mathcal{M}_\theta)}{\int p(Y|\mathcal{M}_\theta, X)p(\mathcal{M}_\theta)d\theta} \tag{5.12}$$

In GPR, the prior knowledge $p(\mathcal{M}_\theta)$ consists of the model $\mathcal{M}_\theta$, see (5.11), with an assumed probability distribution of the parameters $p(\theta)$. For example, a squared exponential kernel (5.8) as prior model, means that the data in X and Y are realizations from (5.8). The posterior distribution describes the probability of the model with parameters, *given the data and prior assumptions*, in contrast to the likelihood where the distribution describes the probability of data, *given the model and parameters*. Note that the posterior distribution (5.12) makes it possible to compare the probability of different models with each other, given the same data set.

Predictions from the posterior model are made by using the fact that $p(\mathcal{M}_\theta|X, Y)$ in (5.12) is a Gaussian process, and likewise a multivariate normal distribution. Therefore, (5.7) can be used to obtain the predictive distribution of the output $y(x^*|X)$ for a desired input x^*. Thus, the predictive distribution is given by

$$p(y^*|X, Y, \mathcal{M}_\theta) \sim N\left(m(x^*|x), k_\theta(x^*|x)\right), \tag{5.13}$$

Gaussian process regression can in theory be non-parametric. To obtain a non-parametric GPR we integrate (5.13) and the prior distribution for all parameters

$$p(y^*|X,Y,\mathcal{M}) = \int p(y^*|X,Y,\mathcal{M}_\theta)p(\mathcal{M}_\theta)d\theta \tag{5.14}$$

where the resulting predictive distribution $p(y^*|X,Y,\mathcal{M})$ does not depend on parameter values. This procedure is known as marginalization and can be seen as a weighting of different models with their respective probabilities. Unfortunately, (5.14) is in general analytically intractable, meaning that it has to be approximated in practical applications. In this chapter, we use two different approximations of (5.14) with two different estimation methods GPR-MLE and GPR-SMC.

5.2.5 GPR-MLE: Maximum likelihood variant

The most common approximation for (5.14) is to replace the integral in (5.14) with the maximum likelihood estimate (MLE). The MLE uses the parameter set $\hat{\theta}_{MLE}$ that maximizes the predictive density $p\left(y^*\middle|x,y,\mathcal{M}_{\hat{\theta}_{MLE}}\right)$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \left(-\frac{1}{2}\ln\left|k_\theta(x_i,x_j)\right| - \frac{1}{2}Y^T k_\theta(x_i,x_j)^{-1}Y - \frac{n}{2}\ln(2\pi)\right) \tag{5.15}$$

That is, the integral of the complete parameter distribution in (5.14) is approximated with a point estimate. This approximation has two drawbacks. First, the solution to (5.15) is a non-convex optimization problem with multiple solutions. Second, the point estimate in (5.15) is a poor estimate for parameter distributions with multiple modes. The impact of the second drawback depends on the optimization problem and cannot be avoided. However, the first drawback could be relieved by using a global optimization method, instead of a local optimization method. Note that GOs are not guaranteed to find the global optimum within finite time, although they commonly provide better solutions than a pure local search. See (Moles et al. 2003) for a comparison of different GOs applied for parameter estimation of biochemical systems.

In this chapter we used a conjugate gradient optimization method implemented in Gaussian Processes for Machine Learning (GPML) toolbox (Rasmussen and Nickisch 2010) to estimate (5.15). The method finds a local minimum based on an initial start guess. We used the method since it is the standard optimization method implemented in the well spread GPML-toolbox for GPR (Rasmussen and Nickisch 2010) and simple to use. A global optimization method would be a feasible alternative for optimization problems with multiple local optima, but would also increase the complexity for the user due to tuning efforts of the GO.

5.2.6 GPR-SMC: Sequential Monte Carlo variant

In this study we also considered a less common approximation of (5.14) and estimation method, namely sequential Monte Carlo for Gaussian processes (GPR-SMC) described in (Svensson. et al. 2015). Here the probability distribution in (5.14) is approximated as

$$\hat{p}(y^*|X,Y,\mathcal{M}) = \sum_{k=1}^{K} w^{(k)} p\left(y^* \middle| X,Y,\mathcal{M}_{\theta^{(k)}}\right) \tag{5.16}$$

where $w^{(k)}$ are K weights that are normalized as $\sum_{k=1}^{K} w^{(k)} = 1$. This approximation differs most notably from (5.15) due to its use of multiple estimates, in contrast to the single point estimate in (5.15). The multiple estimates are supplied by a particle filter, which in short uses a set of estimates (the particles) with different weighting, to approximate the complete probability distribution in (5.14). However, note that both GPR-MLE and GPR-SMC use the same original model structure, the Gaussian process, but mainly differ in how the parameter distribution in (5.14) is approximated and estimated.

One reason for using the more complex estimate (5.16) is that the predictive GPR model becomes non-parametric, or at least less dependent on the GPRs parameter values. This should in theory also result in a better model in terms of mean and variance prediction of the GPR. The drawback of (5.16) is that it is computationally demanding: a) when many particles are used (may be required for a good approximation of the parameter distribution), and b) for high dimensions (many parameter distributions to estimate). It is beyond the scope of this study to introduce the research area of sequential Monte Carlo methods and particle filters and (Schön et al. 2015) and (Gustafsson 2010) respectively provide thorough introductions to the two research areas.

5.2.7 Workflow - GPR in practice

GPR-MLE and GPR-SMC were motivated from comprehensive theory. However, we can apply the methods straightforward once a suitable kernel and parameter limits have been selected. The workflow for applying GPR-MLE and GPR-SMC step-by-step, is given in Table 5.1.

Table 5.1. *Workflow for applying GPR-MLE and GPR-SMC in practice.*

Step	GPR-MLE	GPR-SMC
1)	Select kernel $k_\theta(x_i, x_j)$ and training data X, Y	
2)	Set initial parameter values and optimization settings	Select prior parameter distribution $p(\theta)$ and GPR-SMC settings
3)	Run optimization method, find $\hat{\theta}_{MLE}$	Run GPR-SMC and approximate posterior distribution of parameters
4)	Select x^* where the distribution of y^* will be predicted	
5)	Predict y^* with $\mathcal{M}_{\hat{\theta}_{MLE}}(x^*)$	Predict y^* as a weighted mean of the posterior distribution $\hat{p}(y^*\|x, y, \mathcal{M})$

5.2.8 Case study 1: Missing data in a simulated flow rate signal

The purpose of the first case study was to a) serve as an introductory example of how GPR can be applied to univariate time series, and b) to compare the standard GPR-MLE with GPR-SMC and simple existing methods in a relevant WRRF monitoring problem. The intention was to compare the methods predictive performances, and at the same time study the impact of the aforementioned two potential drawbacks (local optima, and kernel selection). We chose the introductory problem to estimate missing data in a simulated flow rate signal for several reasons.

- A simulated case study enables a simple comparison between true- and estimated values. In addition, simulations allow for repeated well defined experiments.
- Missing data is a common issue in monitoring applications. Data can be missing due to sensor faults, data transfer issues, outliers or stuck values that has to be replaced (Schraa et al. 2006).
- Well established influent flow generators exist that can generate realistic flow rate signals including periodicities (daily-, weekly, and seasonal variations) and stochastic extreme events (rain and snowmelt periods), see e.g. (Gernaey et al. 2011).
- The flow rate signal is a representative process variable for WRRFs since the flow is correlated to many measured nutrient concentrations. Thus, a majority of the monitored process variables at WRRFs exhibit similar characteristics such as periodicities and stochastic extreme events.
- A familiar univariate estimation problem is easy to visualize and interpret as an introductory example.

Two common approaches to replace missing data are either to linearly interpolate over the data gap (referred to as Interpolation), or use the last known

data point before the data gap as estimate (referred to as Last value). Here, we used Interpolation and Last Value to assess the benefit of using GPR-MLE and GPR-SMC.

A simulated flow rate signal, resembling 10 years of influent with 15 min sampling time, was generated with the influent generator described in (Gernaey et al. 2011) with the default parameter settings. Further, we added Gaussian noise with two different standard deviations:

- Low noise level: σ = 135 m^3/h. This was equivalent to 0.5 percent of maximum flow rate, which is the recommended maximum repeatability level for measurements in Parshall flumes (Kulin 1984).
- High noise level: σ = 3000 m^3/h, which was considered to be a noisy but still realistic flow rate signal.

Data were removed at 100 random positions in both the low- and high noise signal. This was repeated four times with gap length of 1, 8, 16, and 24 samples (15min, 2h, 4h, and 6h), which resulted in a total of eight different faulty signals.

The estimation performance was evaluated as normalized root mean squared error (NRMSE) between the true $y(i)$ and estimated value $\hat{y}(i)$ at each missing data position i, normalized with the full signals, Y, standard deviation

$$NRMSE = \frac{\sqrt{\frac{1}{M}\sum_{i=1}^{M}(y(i) - \hat{y}(i))^2}}{std(Y)} \tag{5.17}$$

In (5.17), M is the length of the missing data gap times the total number of gaps. Thus, for a signal with gap length 16, $NRMSE$ was calculated with M equal to 1600.

Two kernels were evaluated: Single kernel (s): the squared exponential kernel in (5.8). The single kernel is indicated with a lower-case *s*, GPR-MLE-s and GPR-SMC-s. Combined kernel (c): a combination of (5.8) and a periodic kernel

$$k_{Combined}(x_i, x_j) = k_{Single}(x_i, x_j) + \theta_3 k_{Single}(x_i, x_j)\, \theta_6^2\, \cos(\frac{2\pi}{\theta_7}(x_i - x_j)) \tag{5.18}$$

and the combined kernel is indicated with a lower-case *c* as, GPR-MLE-c and GPR-SMC-c.

The combined kernel was used to include our prior knowledge about the periodic characteristics of the flow rate signal. In this first case study, time was the input (regressand) expressed as x_i at time instant i, and flow rate was the output (regressor) expressed as $f(x)$.

A uniform distribution was used as prior distribution for parameters in GPR-MLE-s and GPR-MLE-c. The same interval was used to randomly initialize the optimization for GPR-MLE-s and GPR-MLE-c, see step 2 in Table 5.1

Two days before (192 data points) and two days after the data gap were used as training data. Thus, for each signal with 100 data gaps, 100 local GPR-MLE and GPR-SMC models were obtained. Settings for GPR-SMC and prior distributions are given in Supplementary material S5.1.

5.2.9 Case study 2: Detecting abnormal air flow-nitrified ammonium ratios

The intention of the second case study was to study GPR on a real data set and to study whether GPR could be used for model based fault detection. Here, we used GPR to build a simple process model by describing the air requirement as a function of nitrification rate, a relationship that is not necessarily linear. Thus, the promises of GPR allow us to build a well-defined process model based on qualitative process knowledge and process data. Given such a model based on historic data, we could compare new measurement with the model´s predictions and detect deviations in the ammonium measurement.

Drift in ammonium on-line sensor measurements have limited the performance of ammonium based controllers, still, the sensors are common in full-scale control applications (Åmand 2014). In this second case study, we monitored the accuracy of an ammonium on-line sensor positioned in the pre-sedimentation basin ($NH_{4_{in}}$), with GPR-SMC. We only used GPR-SMC since it had better results than GPR-MLE in case study 1.

The GPR-SMC was used to model the ratio between air flow (output/regressor) and nitrified ammonium (input/regressand) according to

$$Q_{air}(t) = f(Q_{flow}(t)\left(NH_{4_{in}}(t) - NH_{4_{out}}(t+5)\right) \tag{5.19}$$

where Q_{air} was the total air applied to the activated sludge process at time t [hours]. We can see from (5.19) that Q_{air} is a function of the wastewater flow Q_{flow}, and the monitored influent and effluent ammonium sensors, $NH_{4_{in}}$ and $NH_{4_{out}}$ respectively. Thus, the regressor variable, the nitrified amount ammonium, was calculated from three variables although the process model was univariate.

We used delayed measurements of the effluent ammonium concentration to compensate for the time-delay between measured influent and effluent ammonium concentration. We used a time-delay of five hours, since this gave the

highest correlation with the influent ammonium concentration. This was similar to the average hydraulic retention time and thus a reasonable approximation of the WRRFs time dynamics.

We used a linear mean function, $m(x)$, and a squared exponential kernel combined with a white noise kernel for the covariance function, i.e.

$$\mathcal{M}_{air/NH_4}(x) \sim GP\left(m(x), k_\theta(x_i, x_j)\right) \tag{5.20}$$

$$k_\theta(x_i, x_j) = k_{Single}(x_i, x_j) + k_{WN}(x_i, x_j) \tag{5.21}$$

where x was amount of nitrified ammonium, and $\mathcal{M}_{air/NH_4}(x)$ was the modelled air flow. The reason for the linear mean function was prior knowledge that an increased amount of nitrified ammonium requires an increase in supplied air. However, since the relationship is not necessarily linear but at least suspected to be smooth, the squared exponential kernel seemed like a reasonable choice.

As benchmark method we used a standard linear regression model with two parameters θ_3 and θ_4 as the regression function $F(x)$

$$F(x) = \theta_3 + \theta_4 x, \tag{5.22}$$

The parameter estimates $\hat{\theta}_3$ and $\hat{\theta}_4$ were obtained by ordinary least squares.

Historic process data from Bromma WRRF in Sweden (approximately 300,000 p.e.) were used in this study. The original dataset was 2 years with one hour sampling time, but we reduced the data set for two reasons. First, we were only interested in periods where we were certain that the influent ammonium sensor was either faulty or correct. Second, we wanted to restrict the study to stable and warm periods to be able to use the simple process model in (5.19).

The first data selection requirement was needed to assess the performance of the fault detection methods. In contrast to simulated data, it is problematic to know the true values in full-scale data. Here, we used weekly lab analyses to decide whether the influent ammonium sensor showed correct or faulty readings. Although lab measurements include uncertainties, this validation procedure is a common approach to validate most online sensors. Since the weekly lab samples were sampled volume proportional, we adjusted $NH_{4_{in}}$ to be flow proportional at the equivalent week and then compared the weekly averages. The 20 weeks with largest- and smallest difference in mean were defined as faulty and correct measurements, respectively. We discarded the

remaining data since we considered the correctness of influent ammonium sensor to be unclear.

An extended multivariate version of (5.19) with suspended solids concentration and temperature could potentially cover a broader process conditions. However, in this introductory study we prefer to use a simple univariate model. We refer the interested reader to (Ažman and Kocijan 2007) for examples of how GPR can be applied in the multivariate case and to (Dolenc et al. 2016) for recent example of GPR applied to fault detection applications.

The second selection requirement was selected to maintain a simple process model. The process model in (5.19) depends on temperature and suspended solids concentration, and is therefore only valid within a specific range. Here, we used stable inflow neglecting heavy rainfalls and snow melting periods (Q_{flow}below 2.2 m^3/h), summer temperatures (temperatures above 15°C), and normal suspend solids concentration (SS) (SS between 3200 and 3900 mg/L).

At the end, the two selection requirements resulted in training and test data of about four months, compared to the original two year large data set. The four months data were categorized as: training data (1200 hours, non-faulty), non-faulty test data (1000 hours), and faulty test data (300 hours), where the training data came from year one and test data from the second year.

Similarly, as it is problematic to use full-scale data to evaluate fault detection evaluation, it may also be problematic to assure that the training data comes without significant errors. In this case study we validated the training data by using the following approaches:

- $NH_{4_{out}}$ was verified by lab measurements similarly as $NH_{4_{in}}$. In contrast to $NH_{4_{in}}$, the ammonium at the effluent was measured by an analyzer which showed good agreement with lab analyses.
- Q_{flow} had no large changes in ratio compared to the summed flow measurements of all parallel flows at the activated sludge basin. Thus, we are confident that no large changes occurred during the two years, although the true value might be biased. A constant bias would however not affect our suggested monitoring approach.
- Q_{air} was hard to validate, and significant errors in training data could result in either to many false alarms or a too restrictive fault detector. However, we expected Q_{air} to be a stable measurement with minor risk to drift since the sensor measures clean air indoor and is unlikely to be affected be e.g. moist or fouling. Also, previous experience suggests that Q_{air} measurements are stable towards large drift within two years' time and a constant bias would not affect our suggested monitoring approach.

We used percentage of detection P_D, and false alarm probability P_{FA} to evaluate the fault detection performance of GPR-SMC and OLS

$$P_D = \frac{\#Alarms}{\#Faulty\ test\ data} \tag{5.23}$$

$$P_{FA} = \frac{\#Alarms}{\#Non-faulty\ test\ data} \tag{5.24}$$

A fault was detected when a new observation was outside a ± 2 standard deviation prediction interval. A Gaussian 2 standard deviations prediction interval should include 95% of new values, given that the process model still applies. We assumed that this was a good trade-off between false alarm rate and detection probability. However, the magnitude of the prediction interval is a user choice that should fit the monitoring purpose. An increase to e.g. 3 standard deviations (which is also a common choice), would give a more restrictive fault detector with lower probability for false alarms with the risk of missing true faults.

5.2.10 Software

We used MATLAB for all simulations and calculations. For GPR-MLE and GPR-SMC the GPML-package (Rasmussen and Nickisch 2010) was used with an add-on toolbox for GPR-SMC (Svensson. et al. 2015). We used the influent generator described in (Gernaey et al. 2011) to simulate the flow rate signals in Case study 1. Computation time on a standard laptop computer was used as indication of computation complexity.

5.3 Results and Discussion

This section presents and discusses the results for the two case studies separately.

5.3.1 Case study 1: Estimating missing data in a flow rate signal

In this case study we evaluated whether GPR-MLE and GPR-SMC could be used to estimate missing data in a simulated flow rate signal. The simulated flow rate signal and the missing data estimation methods were described in the methods section.

The lowest NRMSE was obtained by GPR-SMC, and the largest NRMSE was obtained by Last value in both the low- and high noise signals (*Figure*

5.1). It was interesting to note the benefit of an advanced method (GPR-SMC) compared to a basic method (Last Value) for estimating missing data, although we expected Last value to be the worst method.

Interpolation had the shortest mean computation time of about 1 ms, disregarding Last value that did not involve any computations. GPR-MLE and GPR-SMC were on average 3 and 5 orders of magnitude slower (5s and 100s), respectively. The estimation of parameters (GPR-MLE) and approximating the predictive distribution (GPR-SMC) were the reasons for the high computation time. The computation time did not depend on the estimated gap length, but on the size of training data. Computation time was only used as an indication of the different methods computational complexity, and was not a comparison between the specific software implementations of GPR-MLE and GPR-SMC.

Although GPR-SMC was the most computationally demanding method, it should be fast enough for an off-line batch data estimation application. For estimating data in the low noise signal, one should use Interpolation since it had comparable performance with GPR-SMC, but was both faster and easier to use (*Figure 5.1(a)*).

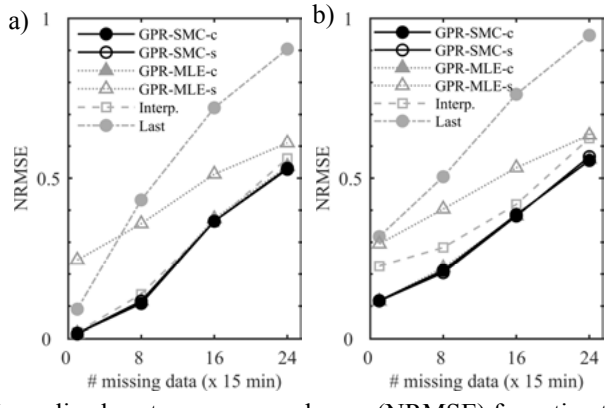


Figure 5.1. Normalized root mean squared error (NRMSE) for estimating 100 missing data gaps during 10 years in a simulated flow rate signal. Four different gap lengths were evaluated with 1, 8, 16, and 24 missing data with 15 minutes sampling time, resulting in data gaps between 15 minutes (1 missing datum) and 6 hours (24 missing data). **a)** Low noise, $\sigma = 135$ m^3/h, and **b)** high noise $\sigma = 3000$ m^3/h. GPR-SMC-s (empty black circle) with single kernel, and combined kernel GPR-SMC-c (filled black circle) gave similar NRMSE and overlap in both a) and b). GPR-MLE-c (filled grey triangle) overlap with GPR-SMC-s and GPR-SMC-c in a) and b).

An increase in the noise level made it more difficult for all methods to estimate missing data (compare *Figure 5.1(a)* and *(b)*). Interestingly, the difference in NRMSE between GPR-SMC and Interpolation increased for the high noise signal, compared to the low noise signal. Thus, GPR-SMC was less sensitive to the noise level compared to Interpolation. GPR-MLE and GPR-SMC estimate both the signal and the measurement noise, i.e. the covariance function

in (5.11). As a result, the mean function estimated will be more accurate (given that the noise estimate is accurate) compared to Interpolation, which is by construction noise sensitive.

Both GPR-MLE and GPR-SMC provided consistently good noise estimates that can be used as a quality check when the true measurement noise is known. With unknown measurement noise, the measurement noise estimate is an additional benefit of GPR-MLE and GPR-SMC. Tracking of the noise estimate could be used to detect changes in sensors measurement noise.

GPR-SMC-s gave good estimates of both the true signal (the signal without measurement noise) and missing data in the high noise signal (*Figure 5.2(c)*). Likewise, the kernels for GPR-SMC-s were close to the estimated auto covariance function (ACF), (*Figure 5.2(d)*). The ACF was estimated from the noisy data. Note that GPR-SMC used several kernels (*Figure 5.2(b)* and *(d)*), one for each particle with a specific weight according (5.16).

In contrast to the good estimates by GPR-SMC-s, GPR-MLE-s gave poor estimates for missing data and the true signal (*Figure 5.2(c)*). The poor estimates were caused by a non-informative kernel with bad parameter values ((*Figure 5.2*(d)), which were obtained from a local minimum during parameter optimization. In this case study, 26% of the optimization rounds of GPR-MLE-s (single kernel) got stuck in a local optimum, and less than 1% of the optimization of GPR-MLE-c (combined kernel). This was the reason why GPR-MLE-s also had a worse NRMSE than GPR-SMC-s (compare GPR-MLE-s and GPR-SMC-s in *Figure 5.1*). Also note that GPR-MLE-c and GPR-SMC-c have similar NRMSE and overlap in *Figure 5.1*. We expected a kernel with many parameters, such as GPR-MLE-c, to be more likely to be stuck in local optima, than GPR-MLE-s. But as the results indicate, for this specific dataset, the combined kernel suffered less than the squared exponential kernel from local optima.

In summary, GPR-SMC performed better than GPR-MLE because it could better account for local optima, for the two different kernels. However, it is not clear from the results whether this was an effect of the particle sampling of (5.16), or due to the non-parametric approximation used in GPR-SMC. In addition, we did not use a Global optimization method (GO) in GPR-MLE, that potentially could have handled the local optima in a better way than the local optimization method in the GPML-toolbox. Thus, a comparison between GPR-SMC and GPR-MLE with a GO would be an interesting future study.

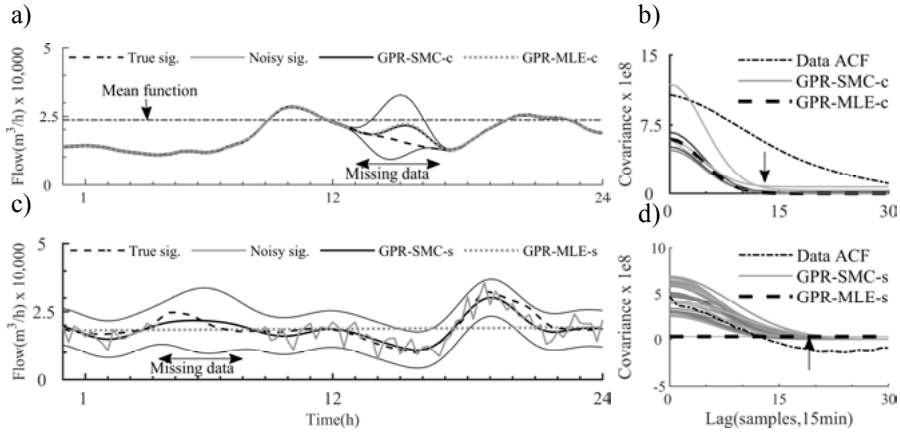


Figure 5.2. **a, b)** Low noise level and **c, d)** high noise level. Time series are in a and c) and corresponding kernels and estimated autocovariance functions in b) and d). Arrows in b) and d) indicate number of time lags when kernels approach zero covariance. Note that GPR-SMC has equally many covariance functions as particles used for the approximation in (5.16). The predictions of GPR-SMC in a) and c) are the sum of the different covariance functions with a weighing provided by the particle filter. **Legends time series a) and c)**: true signal without noise (dashed black), simulated signal with noise (solid grey), estimated signal by GPR-SMC (solid black, thick line for mean prediction and thin lines for 2 standard deviation prediction interval), GPR-MLE (dashed grey), and mean function value of GPR-MLE-c and GPR-SMC-c (dashed dotted black, only in a). **Legends kernels and covariance functions b) and d):** estimated auto covariance function (Data ACF) from simulated signal (dashed dotted black), kernels GPR-SMC-s and GPR-SMC-c (solid greyscale, dark indicate high weight), kernel for GPR-MLE-s and GPR-MLE-c (dashed black).

It was not beneficial to use the combined kernel given in (5.18), compared to the standard squared exponential kernel given in (5.8). GPR-SMC-s and GPR-SMC-c had similar NRMSE (*Figure 5.1*) although the computation time was twice as long for the combined kernel as for the single kernel (126s for GPR-SMC-c and 70s for GPR-SMC-s). The reason for the equal estimation performance was that the periodic part of the combined kernel was given weights close to zero (see θ_3 in (5.18)). In effect, we obtained the same results for both the combined- and the single kernel.

A nice feature of GPR-SMC was its ability to do the trade off in model complexity automatically. In general, a too flexible model would overfit, which did not happen in this experiment. However, one should use the simplest model that serves the purpose, and in this experiment the squared exponential kernel was good choice both for GPR-SMC and GPR-MLE.

Both GPR-MLE and GPR-SMC provide an uncertainty estimate of the mean value predictions, which is given by the covariance in (5.6). The uncertainty estimate is given as prediction interval with ± 2 standard deviations for GPR-SMC-s in *Figure 5.2(a)* and *(c)* by the thin solid black lines. The total covariance is the sum of the estimated measurement noise and the covariance added from the kernel (5.6). We can see that the maximum covariance of a

prediction is obtained if the second term in (5.6) is zero. This corresponds to the lag distance when the kernel goes to zero (see arrows at x-axis in *Figure 5.2(b)* and (d)). Thus, the uncertainty estimate attains its maximum value when no information is used by adjacent points for the predictions. At the same time, the prediction approaches the mean function value which in this study was the mean value of the signal. Both effects can be seen in *Figure 5.2(a)*) were the predictions (solid black line) stretches towards the mean function value (dotted dashed line *Figure 5.2(a)*, indicated with arrow), and the covariance bounds increase at the missing data gap. It is important to remark that the covariance bound is not a measure of agreement between true (unknown) signal value and predicted value, but a prediction interval that we expect to contain the signal.

5.3.2 Case study 2: Detecting anomalous air flow–nitrified ammonium ratios

The test data included a time period with two similar trends were the first one was caused by a drift in ammonium on-line sensor, and the second trend was a true decrease in ammonium concentration (*Figure 5.3*)**.** GPR-SMC successfully detected the drift in the ammonium on-line sensor and had few false alarms during the non-faulty test period. In the full test dataset, GPR-SMC had 74.5% detection percentage and 15.9% false alarm percentage at a ± 2 standard deviation prediction interval. The size of the prediction interval is a user choice, and could be increased in order to e.g. decrease the false alarm percentage.

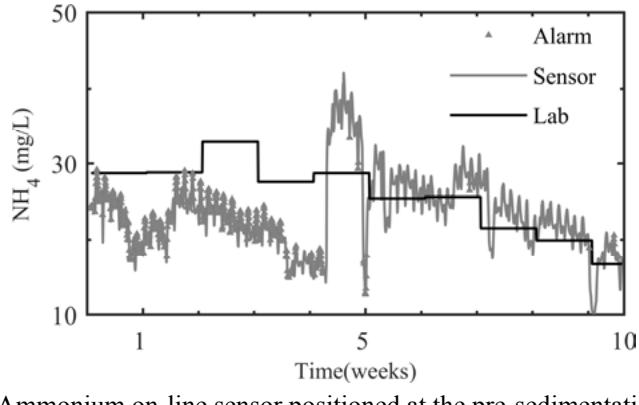


Figure 5.3. Ammonium on-line sensor positioned at the pre-sedimentation basin in Bromma WRRF. The figure shows a part of the test data with faulty period (week 1 to 4.5) and non-faulty period (week 4.5 to 10). Alarms by GPR-SMC are indicated with red triangles, ammonium sensor (solid grey), lab measurements (solid black).

We used the ordinary least squares solution (OLS) to the linear regression model in (5.22) as benchmark method, to evaluate the benefit of using Gaussian process regression. Surprisingly, both the detection percentage and false alarm percentage were equivalent to GPR-SMCs results. The reason for this was that the air flow–nitrified ammonium ratio was almost linear in the range 18–45 mg/L, which contained the majority of test data (*Figure 5.4*). Thus, we did not benefit from GPR-SMCs flexibility in this specific case study.

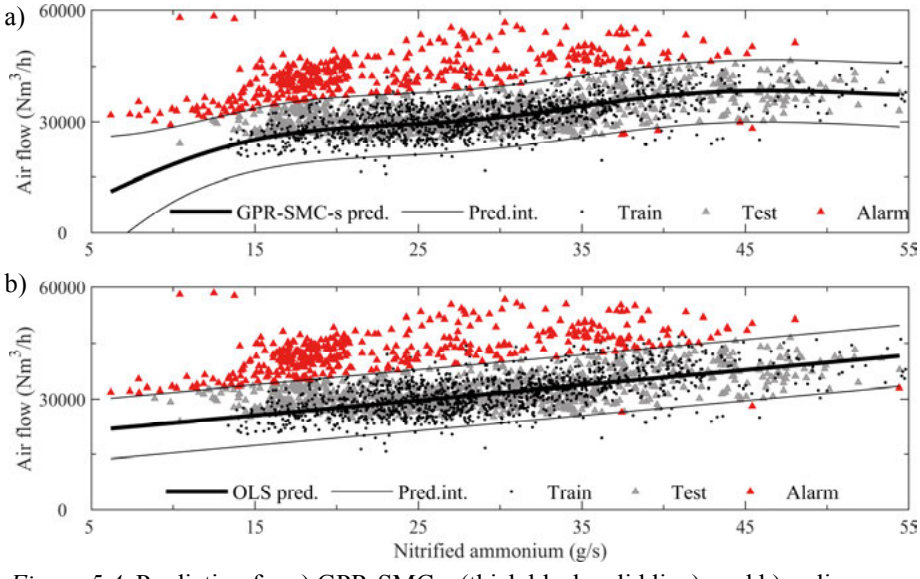


Figure 5.4. Prediction for a) GPR-SMC-s (thick black solid line), and b) ordinary linear regression (OLS) (thick black solid line). Both methods used the same training data (black dots), and test data (grey triangles). Test data outside a ± 2 standard deviation prediction interval (thin black solid line) raised an alarm (red triangles). The training data were approximately linear between 15 and 45 g/s nitrified ammonium.

The magnitudes of the prediction intervals were rather constant for both OLS and GPR-SMC (*Figure 5.4*). GPR is commonly motivated by its data dependent confidence bounds, more exact the prediction intervals (Rasmussen and Williams 2005). In this case study, they did not vary much because of two reasons. First, the variance due to measurement noise was much larger than the variance contribution from the kernel. Thus, the variance from measurement noise would hide potential changes in variance from the kernel. Second, we did not have a change in the predicted variance since we had many data. It was only for predictions far from training data (below 15 g/s and above 55 g/s) that the variance in GPR-SMC increased due to distance to data.

In future studies, we would like to study a regression problem with more profound static non-linearity to see if GPR-SMC would provide better results than linear regression. Further, such study should also be multivariate to assess

GPR-SMCs performance in larger dimensions with multiple parameter distributions to estimate.

5.4 Conclusions

This study gives new insights into the strengths and weaknesses of two variants of the machine learning method Gaussian process regression, GPR-MLE and GPR-SMC. The methods were applied to two WRRF monitoring problems.

The results in the simulated missing data flow rate case showed that GPR-SMC had better performance than the standard GPR-MLE, where GPR-MLE suffered from getting stuck in local minima during kernel parameter estimation. The local optima were mainly a problem for GPR-MLE when the single kernel was used, and indicates that the standard GPR-MLE with a standard kernel was not satisfactory for the specific case to estimate missing data in a flow rate signal. In contrast, GPR-SMC had no problems with local optima and gave at the same time better predictions than linear interpolation. In the second case study with real data, GPR-SMC was as good as linear regression, but this was because the data in the regression problem had an almost linear relationship.

It was not straightforward to include prior knowledge in the kernel, as commonly claimed in e.g. (Rasmussen and Williams 2005). Although an influent flow rate signal is clearly periodic, the combined kernel with a periodic part, did not give better results than the single kernel. Thus, we argue that the common claim that prior knowledge easily can be used by selecting a proper kernel is idealized. We believe that efficient usage of GPR in full-scale applications requires thorough understanding about kernels. In this study we show that a graph of the autocovariance with respect to time lags can be useful to interpret the impact of different kernels. This somewhat demystifies the kernel and simplifies interpretation of the results.

We conclude that GPR-SMC is both a general and powerful method for monitoring full-scale water resource recovery facilities. It did not suffer from local optima in either the simulated or the full-scale case study and gave good results with both kernels.

5.5 Chapter acknowledgements

We thank Erik Lindblom (Stockholm Water) for providing data and valuable discussions in case study 2, and Andreas Svensson (Uppsala University) for providing the software implementation of GPR-SMC and valuable discussions about its usage. Finally, we thank Krist Gernaey (Technical University of Denmark) for providing access to the influent flow rate simulator.

6 Machine Learning Techniques for Monitoring the Sludge Profile in a Secondary Settler Tank

The aim of this chapter is to evaluate and compare the performance of two machine learning methods, Gaussian Process Regression (GPR) and Gaussian Mixture Models (GMM), as two possible methods for monitoring the sludge profile in a secondary settler tank (SST). In GPR the prediction of the response variable is given as a Gaussian probability density function, whereas in the GMM the probability density function is built as a weighted sum of Gaussian distributions. In both approaches, a residual is calculated and a fault detection criterion is implemented via a recursive decision rule. As case study, GMM and GPR were tested using real data from a sensor measuring the suspended solids concentration as a function of the SST level in a water resource recovery facility in Bromma, Sweden. The results suggest that GMM gives a faster response but is also more sensitive than GPR to changes during normal conditions.

6.1 Introduction

Increasing demands on effluent water quality and resource efficient operation are important driving forces for water resource recovery facilities (WRRFs). Process monitoring and detection of abnormal conditions are crucial tasks, since they help in keeping a robust and efficient performance of the plant (Olsson et al. 2014). Furthermore, increasing the number of sensors adds process information but also increases the complexity for plant operators. Hence, the need for fault detection methods is a priority.

The secondary settler tank (SST) is a key part of a WRRF since it provides two functions: clarification and thickening. In the clarification, the aim is the removal of suspended solids (SS) so to get a clarified effluent that meets the effluent SS goal. In the thickening, the aim is to get the concentration of the settled solids to be returned to the bioreactor. The SST uses gravity to separate the sludge (solid) component from the treated water (liquid). Due to the clarification and thickening functions, typically a concentration profile (also called sludge profile) has a low value for the concentration close to the effluent, and

this value increases towards the bottom of the SST. There is a level where the solid concentration abruptly changes, which is called sludge blanket level. This level together with the SS concentration in the bottom or in the return sludge are the typical values to monitor in a SST.

Some examples of methods applied to monitor a SST include image analysis (Grijspeerdt and Verstraete 1997) and model-based approaches (Traoré et al. 2006, Yoo et al. 2002) even though the prediction of the sludge concentration profile is still far from satisfactory (Li and Stenstrom 2014), which makes it problematic to perform a good monitoring of the SST.

Mathematical models have also been used for predicting the sludge profile, see for example 1D model proposed by Diehl et al. (2016) and 3D model proposed by Xanthos et al. (2011) and to estimate the biomass in the activated sludge process (Hedegärd and Wik 2011).

In the last two decades, the research field Machine Learning has gained especial attention since with Machine Learning it is possible to develop methods that can automatically detect patterns in data (i.e. learning), and then to use this information to predict future data (Murphy 2012). There are many different techniques in machine learning including decision trees, data clustering, neural networks, Gaussian process regression, Gaussian mixture models, to mention a few.

From these machine learning techniques, Gaussian process regression (GPR) and Gaussian mixture models (GMM) are two techniques that have started to gain interest in different applications. GPR is a regression method where a prediction of the response variable is given as a Gaussian probability density function. Thus, the predicted value of the response variable comes with a variance estimate, which is interpreted as an uncertainty measure of the prediction (Rasmussen and Williams 2005). It is worth to note that GPR is not a new concept, it was originally known as Kriging, with an origin from geostatistics in the 1950s (Cressie 1990).

GPR has several properties making it useful for monitoring and fault detection, such as: probabilistic prediction including an uncertainty estimate, flexible regression in a non-parametric fashion, and it is relatively simple to implement in common programming languages. GPR has been used for monitoring and fault detection in different applications (Roberts et al. 2013), such as: maritime vessel track analysis (Smith et al. 2012), change point detection (Garnett et al. 2010b) and process monitoring (Serradilla et al. 2011). GPR has also been used in environmental applications, such as monitoring and fault detection in water monitoring signals (Samuelsson et al. 2017), modelling of an anaerobic wastewater treatment system (Ni et al. 2012), modelling nitrification process and biomass growth (Ažman and Kocijan 2007) and control of a sequencing batch reactor (Kocijan and Hvala 2013).

GMM is an alternative machine learning method to GPR for data monitoring. GMM is a parametric probability model for density estimation using a mixture of Gaussian distributions (Bishop 2006). In this way, the GMM can

describe a set of data using the combination of Gaussian distributions. Applications of GMM in data monitoring can also be found in literature, for example in sensor monitoring (Zhu et al. 2014), fault detection and diagnosis (Jiang et al. 2016, Yu 2012). Some other applications include data classification (Bouveyron 2014) image segmentation (Greggio et al. 2012), and many others.

The objective of this study is to compare the performance of GPR and GMM as tools for monitoring and fault detection of sludge profiles in a secondary settler tank. The chapter is organized as follows. First, a general introduction to GPR and GMM is presented. Then, the case study is detailed, and a fault detection criterion based on GPR and GMM is then formulated. Next, results and discussions are shown, and some conclusions are drawn.

6.2 Material and Methods

The GMM is described in Section 6.2.1, whereas the GPR was described in Chapter 5 (Section 5.2.1-5.2.5). The fault detection problem is outlined in Section 6.2.2 and the fault detection criteria for both methods are given in Section 6.2.3.

6.2.1 Gaussian mixture models

Assume that a dataset with the inputs $X = [x_1, \ldots, x_N]$ and outputs $Y = [y_1, \ldots, y_N]$ form a dataset $\boldsymbol{w} = (X, Y)$, using the same nomenclature as in Chapter 5, where the N observations have been obtained independently. One way to model these data is by a mixture of models, where the aim is to represent certain subpopulations from the whole data set by means of a conditional probability density (binomial, exponential, etc.) (Južnič-Zonta et al. 2012). In the case of Gaussian Mixture Models (GMMs), the distribution of the observation $\boldsymbol{w_n}$ is modeled as a sum (or mixture) of several Gaussian distributions (Murphy 2012)

$$p(\boldsymbol{w_n}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\boldsymbol{w_n} | \mu_k, \sigma_k^2), \tag{6.1}$$

where μ_k represents the mean and σ_k^2 represents the covariance matrix of the k-distribution. Hence, (6.1) is a combination of K Gaussian distributions, where each of them has a mixing weight π_k. These mixing weights must satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The resulting function $p(\boldsymbol{w_n})$ is a probability density function from observing the data $\boldsymbol{w_n}$

Once the value K is specified, the GMM parameters π_k, μ_k, σ_k^2 can be inferred by using the iterative Expectation-Maximization (EM) algorithm applied to Gaussian Mixtures (Murphy 2012). EM is a technique used to find

maximum likelihood solutions for probabilistic models containing variables that are not directly observed but can be inferred (Bishop 2006). EM for GMM is summarized in Table 6.1.

Table 6.1. *Expectation maximization for Gaussian mixture models.*

Algorithm 1

1: Collect a dataset $\boldsymbol{w}$ with N observations
2: Initialize $\pi_k^1, \mu_k^1, \sigma_k^1$, set $i = 1$.
3: **while** not convergence of π_k, μ_k, σ_k , **do**
4: Expectation step: Compute

$$\gamma(z_{nk}) = \frac{\pi_k^i \mathcal{N}(w_n|\mu_k^i, \Sigma_k^i)}{\sum_{j=1}^{K} \pi_j^i \mathcal{N}(w_n|\mu_j^i, \Sigma_j^i)}, \tag{6.2}$$

$$for \ n = 1, \ldots, N, \qquad k = 1, \ldots, K. \tag{6.3}$$

5: Maximization step: Compute

$$\mu_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) w_n, \tag{6.4}$$

$$\pi_k^{i+1} = \frac{N_k}{N}, \tag{6.5}$$

$$\Sigma_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \Omega\Omega^T, \tag{6.6}$$

6: $i \leftarrow i + 1$
7: **end while**
where $\Sigma_k^i = (\sigma_k^2)^i$, $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$, $\Omega = w_n - \mu_k^{i+1}$.

One way to assign a value for the number for Gaussian distributions K is to use the silhouette criterion (Rousseeuw 1987). This criterion calculates a silhouette value S which indicates how similar samples are in one cluster to samples in another cluster. S ranges from -1 (data misclassified) to $+1$ (data well clustered), whereas S close to zero means that the clusters are indistinguishable.

6.2.2 Case study – monitoring a secondary settler in WRRF

The present approach was evaluated using real data from a sensor installed in a SST at Bromma WRRF in Stockholm, Sweden. The sensor goes from the top to the bottom of the settler, passing through the clarification and the thickening zone. In this way the sensor measures the level [m] and the SS concentration [g/L], as shown in *Figure 6.1(a)*. The profile obtained is called *sludge profile*. A typical sludge profile is shown in *Figure 6.1(b)*.

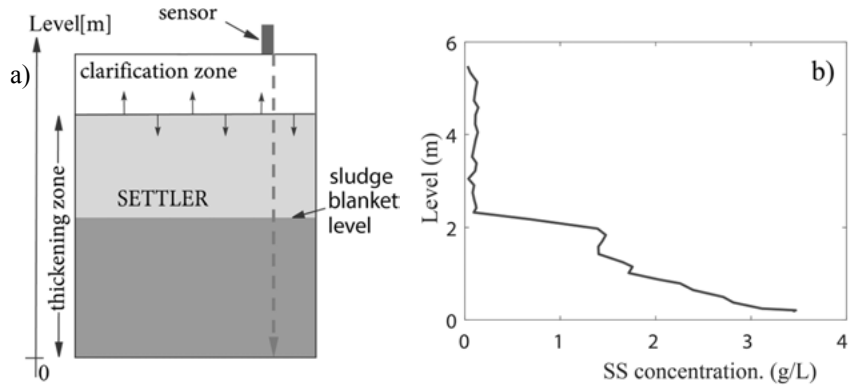


Figure 6.1. (**a**) Experimental setup. (**b**) A typical sludge profile plotted as level versus SS concentration.

The sensor works periodically, which means that a new sludge profile is automatically measured after a certain period of time. The sludge profile can be affected by different factors including: changes in the return and/or excess of sludge flow rates, sludge scape, large variations in the influent flow and composition, and sensor clogging or malfunctioning.

As part of the experiment, two additional measurements were recorded: the level at which the SS concentration was equal to 0.5 g/L (here referred as *fluff level*) and equal to 2.5 g/L (here referred as *sludge level*). We will refer to these levels during the subsequent sections.

6.2.3 Fault detection criterion

6.2.3.1 Decision rule

The implementation of a FDM for a sensor signal y_m involves detecting any significant change in the dynamic of y_m when the sensor is subject to possible clogging or malfunctioning situation. A parameter v related to the dynamic of y_m is assumed to belong to one out of two conditions

$$\begin{aligned} H_0: v &= v_0 \quad (normal), \\ H_1: v &= v_1 \quad (faulty). \end{aligned} \tag{6.7}$$

To decide between H_0 and H_1, two FDMs are proposed. These methods are based on GPR and GMM (see the next subsections), which use the following recursive decision rule (Basseville 1993)

$$\varepsilon_m = \begin{cases} \varepsilon_{m-1} + s_m & if \ \ \varepsilon_{m-1} + s_m > 0 \\ 0 & otherwise \end{cases} \tag{6.8}$$

where $\varepsilon_0 = 0$ is set as the initial value, and $s_m = \ln(p_{v_1}(y_m)/p_{v_0}(y_m))$ is the log-likelihood ratio between the probability density function $p_{v_1}(y_m)$ and $p_{v_0}(y_m)$, m indicates the profile number at which ε_m is calculated.

A common way to detect possible changes in y_m is by analyzing its *residual* r_m, which is related to the distance between data in normal and possible faulty conditions. Therefore, in normal condition r_m is equal or close to zero. Otherwise, it will increase and will belong to a possible faulty condition. Hence, the detection of a change in the residual can be done by detecting a change in the mean value of the sequence r_m, i.e. in this work, the parameter v referred in (6.7) corresponds to the mean value of the residual calculation of y_m.

Assume that r_m follows a Gaussian distribution $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, where $\tilde{\mu}$ and $\tilde{\sigma}$ are the mean and standard deviation of r_m, respectively. The log-likelihood ratio test (Basseville 1993) for a change in $\tilde{\mu}$ is expressed by

$$s_m = \frac{(\tilde{\mu}_1 - \tilde{\mu}_0)}{\tilde{\sigma}^2}\left[r_m - \frac{(\tilde{\mu}_0 + \tilde{\mu}_1)}{2}\right] \tag{6.9}$$

where $\tilde{\mu}_0$ is the mean value in normal condition, and $\tilde{\mu}_1$ is the possible change we want to detect. $\tilde{\mu}_1$ is calculated by collecting data in a moving window. Equation (6.9) can be used in (6.8) to calculate ε_m recursively.

A fault is decided if $\varepsilon_m > h_\varepsilon$, where $h_\varepsilon = \alpha_\varepsilon max(\varepsilon_m)|_{1 \le m \le T_0}$ is a threshold value which is calculated taking the maximum value of ε_m obtained in normal condition for a certain predefined time T_0, and α_ε is a threshold factor.

The next subsections show how to calculate the residual r_m for GPR and GMM, denoted as r_{gp} and r_{gm}, respectively.

6.2.3.2 Residual calculation using GPR

Algorithm 2 in Table 6.2 was used to compute the residual value r_{gp} for GPR.

Table 6.2. *Computation of GPR residuals.*

Algorithm 2

1: Collect M profiles in normal conditions, each of them with N observations, set $m = 1$.
2: Select a covariance function and optimize the hyperparameters by maximizing (5.15)
3: Obtain $p(y_*|(x, y), x_*)$ from (5.13)
4: **while** monitoring a new profile (x_m, y_m) **do**
5: Compute

$$r_{gp}(m) = \frac{1}{N}\sum_{i=1}^{N} |y_{i,m} - \mu_*(x_{i,m})|, \tag{6.10}$$

where $x_m = [x_{1,m}, \ldots, x_{i,m}, \ldots, x_{N,m}]$ and $y_m = [y_{1,m}, \ldots, y_{i,m}, \ldots, y_{N,m}]$

6: $m \leftarrow m + 1$
7: **end while**

Note from (6.10) that the residual is calculated using the distance between the data and the predictive mean μ_*. Then, the more the data in the new profile that are outside the predictive distribution, the larger the residual r_{gp}.

Regarding the covariance function, the selection depends on the case study. A common choice is to use a squared-exponential function. However, if a particular dynamic of the data needs to be captured, a combination of different covariance function should be implemented, such as constant, linear, sinusoidal, etc. See studies by (Lloyd et al. 2014, Wilson 2013) for some examples showing the choice of covariance functions to different datasets. In this case study, the sum of a linear and a squared exponential function was used, that is

$$k_\theta(x_i, x_j) = \theta_1 + \theta_2 x_i + \theta_3 exp\left[-\theta_4 \left(x_i - x_j\right)^2\right], \tag{6.11}$$

where $(\theta_1, \theta_2, \theta_3, \theta_4)$ are hyperparameters.

6.2.3.3 Residual calculation using GMM

Algorithm 3 in Table 6.3 was used for the residual calculation r_{gm} using GMM.

Table 6.3. *Computation of GMM residuals.*

Algorithm 3

1: Collect M profiles in normal conditions, each of them with N observations, set $m = 1$.
2: Set K and compute Algorithm 1 to obtain (μ_k, π_k, σ_k), where $k = 1, \ldots, K$.
3: **while** monitoring a new profile (x_m, y_m) **do**
5: Compute

$$r_{gm}(m) = \left(\sum_{n=1}^{N}\sum_{k=1}^{K} \pi_k \mathcal{N}(w_n|\mu_k, \sigma_k)\right)^{-1} \tag{6.12}$$

where $x_m = [x_{1,m}, \ldots, x_{i,m}, \ldots, x_{N,m}]$ and $y_m = [y_{1,m}, \ldots, y_{i,m}, \ldots, y_{N,m}]$

6: $m \leftarrow m + 1$
7: **end while**

Note from (6.12) that r_{gm} is the inverse of the summation of the probability density function of the entire dataset w. So that the farther the new profile data are from the data in normal condition, the probability density function decrease and the r_{gm} increases.

We decided to choose x representing the level of the sensor and y representing the SS concentration.

6.2.3.4 The training data set

A total of $M = 15$ sludge profiles in normal conditions were used as training dataset for Algorithm 2 and 3. The recursive decision rule (6.8)–(6.9) was implemented with a moving window of 10 profiles, with $T_0 = 4$ d to compute the threshold value h_ε, and $\alpha_\varepsilon = 1.1$ as threshold factor.

6.2.3.5 Software

The software MATLAB was used for all the calculations and simulations. For the GPR implementation, the GPML-toolbox (Rasmussen and Nickisch 2010) was used, whereas the GMM implementation was done with the MATLAB function 'gmdistribution'.

6.3 Results and Discussion

This section shows the comparison between the GPR and GMM performances applied to the case study. First, the training is shown, next the detection perfromance is compared between GMM and GPR. Last, the findings are discussed and put into practical context.

6.3.1 Training the methods

The *Figure 6.2(a)*-(b) shows the sludge profiles in normal conditions used for training. The optimized GPR hyperparameters (6.11) were obtained to: $\hat{\theta}_1 = 3.88$, $\hat{\theta}_2 = -0.95$, $\hat{\theta}_3 = 0.35$, $\hat{\theta}_4 = 1.13$. *Figure 6.2(c)* shows the predictive mean value (red line) along with $\pm 2\sigma_*$ (the predictive distribution of the standard deviation).

For GMM, the highest silhouette value was $S = 0.77$ with $K = 3$, indicating that three was the optimal number of clusters, as shown by the three Gaussian distributions in *Figure 6.2*(d).

Note in *Figure 6.2 (a)*-(b) that a typical sludge profile shows an abrupt change in the SS concentration from values close to zero to values larger than $1 g/L$, and then this concentration keeps increasing as the sensor approaches the bottom of the SST. This change in the concentration was captured by the GPR and GMM. For GPR, the mean predictive decreases from top to bottom of the SST, passing through the data set, with the predictive standard deviation covering almost all the points. In the case of GMM, it classifies data before and after the jump as two separate Gaussian distributions. Note also that data for high concentration and low SST level were classified with another Gaussian distribution.

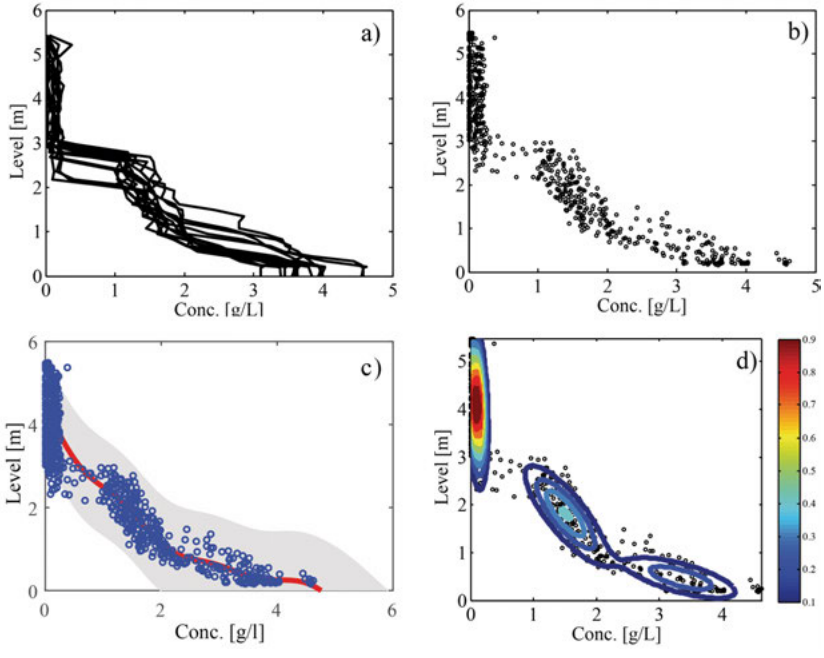


Figure 6.2. (**a**) Sludge profiles used as training data set. (**b**) Sludge profiles in (a) plotted using dots. (**c**) Predictive distribution over the training data set, showing μ_* (red line) and $\pm 2\sigma_*$ (gray zone); (**d**) Contour of the GMM probability density function, with values indicated with colour scale.

6.3.2 Monitoring the secondary settler

Several trials were done to monitor the settler. As illustration, we show one trial which lasted for 23 days, where a new sludge profile was collected every 15 minutes, giving a total of 2208 profiles. The progression of the profiles during time is shown in *Figure 6.3(a)*-(d) after 5, 10, 15 and 20 days of the experiment, respectively.

Figure 6.4 shows the profile of the fluff and sludge level, and the profile of the residuals (r_m) and the decision rule (ε_m) calculated via GPR and GMM approach.

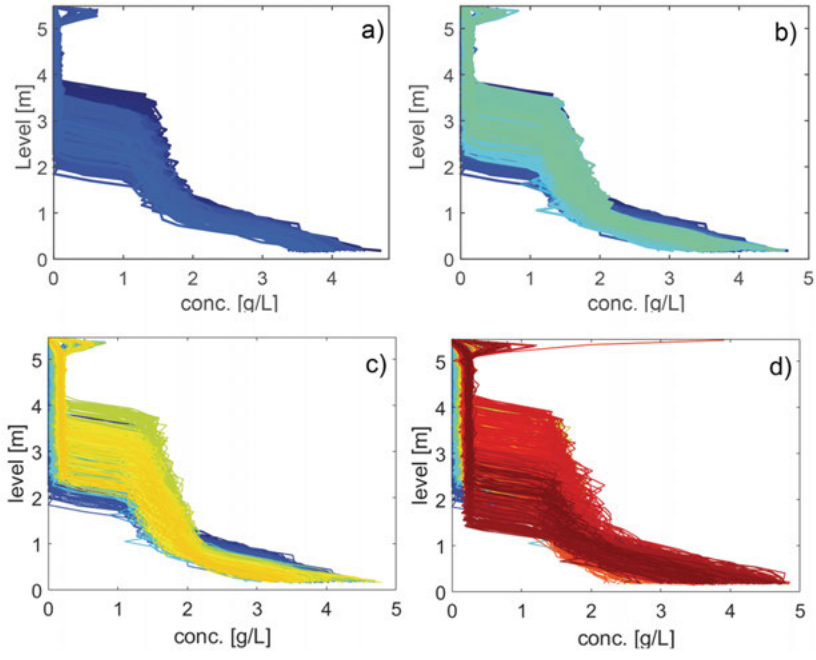


Figure 6.3. Sludge profiles during the SST monitoring. (**a**) after 5 days, (**b**) after 10 days, (**c**) after 15 days, (**d**) after 20 days. First days are coloured in dark blue, last days in dark red.

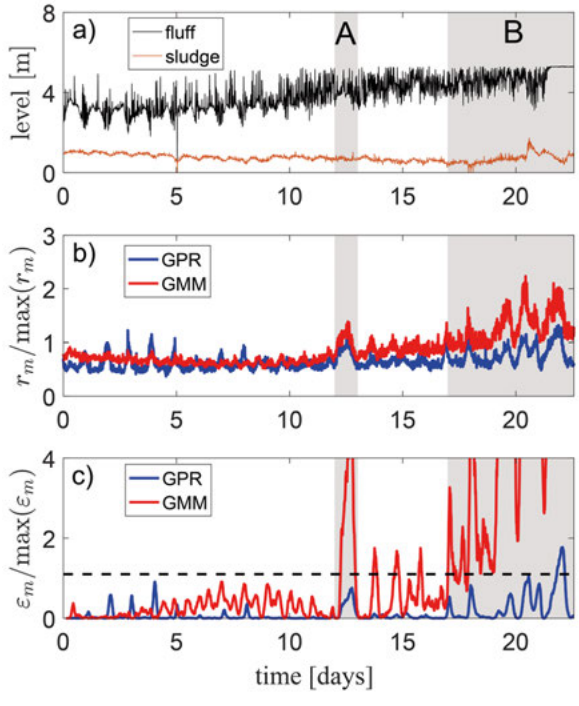


Figure 6.4. (**a**) Fluff level (black line) and sludge level (brown line), (**b**) Normalized residual response for GPR (blue line) and GMM (red line), (**c**) Normalized decision rule response for GPR (blue line) and GMM (red line), threshold for $\alpha_\varepsilon = 1.1$ (black dashed line). Period A and B are shown as grey zones.

There are two periods to highlight in *Figure 6.4*, marked as Period A and B. The sludge profiles of these periods are shown in *Figure 6.5*, which also includes the GRP predictive mean, and the GMM probability density function. Period A corresponds to variations in the residual profiles observed in days $12 - 13$. The profiles of this period show that the concentrations between $1 - 2$ g/L are in a higher level with respect to the level in the training data set, see *Figure 6.5(a)* and *(c)*. This gave a certain variation in the fluff level as seen in *Figure 6.4(a)*. This behaviour was mainly captured by the GMM approach, see *Figure 6.4(c)*.

Period B refers to an event related to sensor clogging occurring after day 17. This event was confirmed by an in-situ ocular inspection of the sensor and the presence of floating sludge at the surface level of the settler, causing sludge scape in the effluent. The profiles of this period are shown in *Figure 6.5*(b) and (d), where concentrations in the range of $0 - 2$ g/L are far from the GMM probability density region and far from the GPR predictive mean. As seen in

Figure 6.4(c), this event was first captured by the GMM which shown a persistent increasing in the decision rule. This event was also captured by the GPR after day 20.

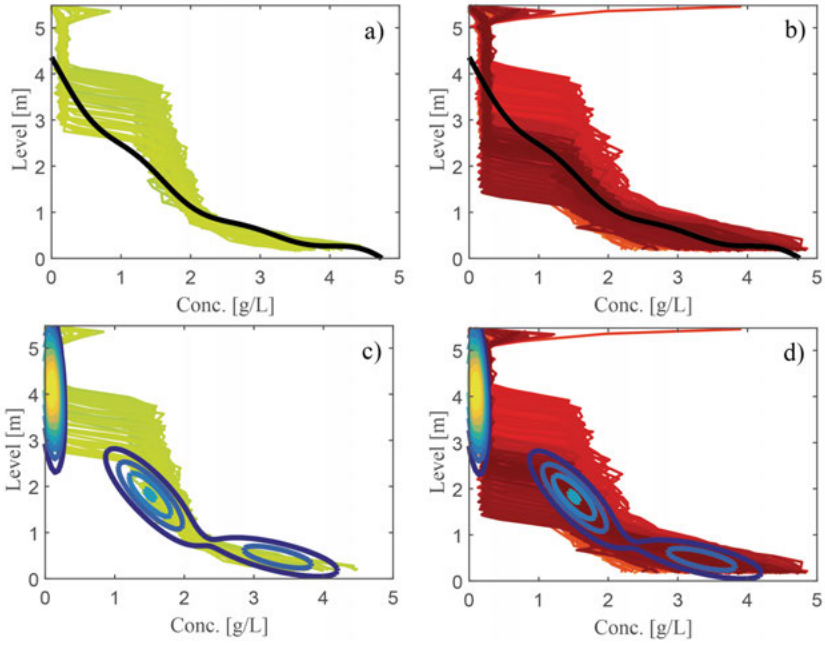


Figure 6.5. Group of sludge profiles for periods indicated in *Figure 6.4(a)*. For GPR: (**a**) Period A, (**b**) Period B. The plots include the predictive mean of the training data set (black line). For GMM: (**c**) Period A, (**d**) Period B. The plots include the contours of the probability density function of the training data set.

The sludge level profile (*Figure 6.4(a)*) did not show any significant change during Period A, however it showed a change late in Period B after day 20. Compared to the fluff level (level at 0.5 g/L), the sludge level (level at 2.5 g/L) did not fluctuate from its initial position, as it can be observed from the sludge profiles in *Figure 6.3*, where most of the level fluctuations in the sludge profiles were in the range of $0 - 2$ g/L.

When comparing the performances of GPR and GMM, the GMM gives a more fluctuate dynamics during the experiment. This can be seen in the behaviour of the residuals in *Figure 6.4*(b).

See that when using $\alpha_\varepsilon = 1.1$, GMM produces higher values than GPR during normal conditions (i.e. false alarms), see the responses between days 13 to 17 (*Figure 6.4(c)*. However, this behavior will depend on the α_ε used for the evaluation of the decision rule. A higher α_ε might reduce the events of false alarms. To see this more in detail, the performance of GMM and GPR was also evaluated for different values of α_ε using the following indicators:

- Delay of fault detection (Δt_{FD}): time spent by ε_m to reach the threshold.
- Time in fault detection (t_{FD}): duration when the detector correctly indicates a fault, i.e. time spent by ε_m above threshold in fault detection condition.
- Number of false alarms (FA).
- Time in false alarm (t_{FA}): duration when the detector incorrectly indicates a fault, i.e. time spent by ε_m above threshold in false alarm condition.

The evaluation was done for Periods A, B and the rest of the experiment time. The results are summarized in Table 6.4.

Table 6.4. *Summary of GMM and GPR performance for different* α_ε.

		$\Delta t_{FD}[d]/t_{FD}[d]$		$FA[\#]/t_{FA}[d]$
α_ε	Method	Period A	Period B	Rest of time
1.1	GMM	0.25/0.65	0.01/5.22	3/0.51
	GPR	-/-	4.81/0.35	0/-
1.25	GMM	0.26/0.62	0.02/4.74	3/0.21
	GPR	-/-	4.91/0.22	0/-
1.5	GMM	0.29/0.58	0.04/4.04	0/-
	GPR	-/-	-/-	0/-
1.75	GMM	0.41/0.40	0.06/3.66	0/-
	GPR	-/-	-/-	0/-
2.0	GMM	0.5/0.25	0.96/3.34	0/-
	GPR	-/-	-/-	0/-

See that the table includes the case shown in *Figure 6.4* when $\alpha_\varepsilon = 1.1$. As expected, when α_ε increases, the time delay to detect a fault also increases, whereas the time the decision rule expends in faulty condition decreases. See also that when α_ε increases, the time in false alarm decreases. When $\alpha_\varepsilon = 1.5$ the GMM performance is superior to GPR, i.e. it gave fault detections in both periods with relatively short time delay and with no false alarms. A threshold factor above 1.5 will also avoid false alarms but will also increase the fault detection time of GMM.

6.3.3 Concluding remarks

An important aspect in the GMM method is the definition of the amount of Gaussian distributions that describe the data set. In a given data set, the parameters involved in the GMM should be determined together with a value that indicates how well clustered is the data set, in our case by using the

silhouette criterion. This criterion should be evaluated for different number of clusters, in order to find the optimum data clustering.

Regarding the GPR method, a key aspect for determining the predictive mean is the selection of the covariance function. In our case study, the covariance function (6.11) was formed by two functions: a linear and a squared-exponential function. A linear function was required to capture the shape of typical sludge profiles. The squared-exponential function can be seen as a smoothing function. For any other process profiles, a new definition of the covariance function should be made.

Note that the data in the sludge profiles include outliers, defined as sharp changes between two successive data points. These outliers mean that the measured concentrations were far from the values shown by profiles in normal condition. If there are few outliers in a profile, a possible task in the fault detection is to perform data correction, i.e. to replace outliers using data from the training data set. In this work, the correction of outliers was not part of the study.

Another possible situation in data monitoring is missing data. This situation did not happen in our case study but it is common in other process monitoring applications. As discussed for the case of dealing with outliers, a missing data can be reconstructed by using data from the training data set. See Chapter 5 where the case of missing data is evaluated for some GPR-based approaches.

It is important to recall that two sensors measuring the same variable in the same reactor will give two non-identical data sets. It means that each sensor will have a unique predictive mean and standard deviation for the case of GPR, as well as a unique probability density function for the case of GMM. It follows that the present methodology has the advantage that it can be applied to sensors in diverse areas.

One of the several applications of the current approach could be to use the decision rule of the FD algorithms as a tool for control actions. Therefore, new control strategies could include this variable as useful information to perform further tasks, for example, changing the recycle flow rate of the WRRF in order to keep the sludge profiles in a predefined level, or to give an early alarm that the SS sensor of the settler tank might need a supervision or a cleaning action.

6.4 Conclusions

This study tested two machine learning techniques, GPR and GMM, for monitoring the sludge profiles (level vs. suspended solid concentration) of a secondary settler tank in a water resource recovery facility. The main idea was to train these two methods by using a set of sludge profiles in normal conditions, and then perform the test by monitoring new sludge profiles.

The results showed that GMM gave a fast fault detection than GPR, but GMM also showed to be more sensitive to false alarms. Nevertheless, it was possible to avoid the false alarm condition with a proper setting of the threshold factor.

Both methods have shown to be potential tools for monitoring sludge profiles. They could be applied for getting useful information about the performance of processes with repetitive profile data and to detect possible abnormal conditions.

7 Automated active fault detection in fouled dissolved oxygen sensors

Biofilm formation causes bias in dissolved oxygen (DO) sensors, which hamper their usage for automatic control and thereby balancing energy and treatment efficiency. In this chapter we analyse if a dataset that was generated with deliberate perturbations, can automatically be interpreted to detect bias caused by biofilm formation. We used a challenging set up with realistic conditions that are required for a full-scale application. This included automated training (adapting to changing normal conditions) and automated tuning (setting an alarm threshold) to assure that the fault detection methods (FDMs) are accessible to the operators. The results showed that automatic usage of FDMs is difficult, especially in terms of automatic tuning of alarm thresholds when small training datasets only represent the normal conditions, i.e. clean sensors. Despite the challenging set-up, two FDMs successfully improved the detection limit to 0.5 mg DO/L bias caused by biofilm formation. We showed that the studied dataset could be interpreted equally well by simpler FDMs, as by advanced machine learning algorithms. This in turn indicates that the information contained in the actively generated data was more vital than its interpretation by advanced algorithms.

7.1 Introduction

In this chapter, we study how fault detection methods (FDMs) automatically can detect bias in dissolved oxygen (DO) sensors caused by biofilm fouling. We further evaluate how a limited, but from an operator perspective, realistic amount of quality-checked training data impacts the training and performance of FDMs. The dataset in (Samuelsson et al. 2018) is used, which was generated in line with the concept Active fault detection (AFD), which is so far an unexplored concept for wastewater applications.

7.1.1 The importance of verifying DO-sensor readings

Monitoring the condition of sensors and their related data quality in water resource recovery facilities (WRRFs) is essential, especially when used for automatic control. The DO-sensor is a key sensor to monitor for two reasons: i)

a bias has a direct effect on aeration energy and/or treatment efficiency, and ii) the usage of DO-sensors is widespread with multiple sensors in a standard WRRF. The workload for DO-sensor maintenance is thus expected to make up a large proportion of the total sensor maintenance time.

7.1.2 Pushing state-of-the-art for monitoring DO-sensors

In Chapter 3 we studied biofilm formation (fouling) and its effect on the two common measurement types of DO-sensors, namely membrane electrochemical (MEC) and optical (OPT) sensors. The results showed that a bias above 1.0 mg/L due to biofilm formation could be detected with a simple rise time estimation method (RT) for the OPT-sensor. We concluded that other FDMs than RT needs to be evaluated to detect lower bias levels and biofilm fouling faster, which is the purpose of this study. In this study we aim to improve the detection limit to 0.5 mg/L bias, which we believe is satisfactory for an early warning system, but still challenging in the light of previous results.

7.1.3 Informative data from Active fault detection

A novel part of the work in Chapter 3 was that the data from the DO-sensors were generated in line with the concept AFD, a concept recently reviewed by Puncochár and Skach (2018). The AFD is centred around designing a *test signal,* which is physically implemented (operationalized) to perturb the studied system in a controlled manner. The purpose of the test signal is to *increase the diagnostic information content in data* without interfering with normal operation. In Chapter 3, the test signal was a short impulse and implemented as pressurized air from the sensor's automatic air-cleaning system. The short impulse temporary increase the sensor's ambient DO concentration and has been denoted an impulse response (IR).

Similar ideas as AFD have been introduced including: system identification (Ljung 1999), on-line modelling (Vanrolleghem 1994), adaptive automatic control (Seborg et al. 1986), and reinforcement learning (Buşoniu et al. 2008). At the heart of these ideas lies experimental signal design to produce information that is optimal according to some criterion, see (Fedorov 2010) for an overview of optimal experimental design (OED). The close relationship between OED, system identification and adaptive automatic control has been reviewed in (Pronzato 2008), although fault detection applications were not considered.

The AFD has been used to improve classical fault detection strategies: i) detection of changed model parameter (Nikoukhah and Campbell 2008), ii) diagnosis with multiple failure models (Campbell et al. 2002), iii) and residual generation (Padilla and Choinière 2015). We expect that OED can be useful for further developing AFD since optimality criteria have been described both

for optimal parameter estimation that relate to i) (A-, D-, and E-optimal (Fedorov 2010)) and model discrimination (T-optimal (Atkinson and Fedorov 1975)) that relate to ii). The merits of using OED would however be limited to optimizing models representing normal conditions, as faulty data are infeasible to obtain in practice, especially during an iterative OED procedure.

We believe that AFD can be especially suitable for WRRF applications. A dedicated test signal with high and controllable excitation should be able to produce data with large signal-to-noise ratio. Here, noise refers to the normal daily-, seasonal-, and influent flow-dependent variations that in general makes it hard to detect sensor anomalies during normal operation. Conducting automated test signals enables exact repeatability and precision, which contrasts with manual checks performed by operators.

7.1.4 Automatic fault detection requires annotated training data

FDMs valuable in practice both need to have good detection performance (high detection rate, low false alarm rate) and to be practically feasible. From an operator's perspective, practically feasible means simple to use under existing conditions with little added workload. Automated training (adapting to training data) and tuning (setting the alarm threshold) are therefore desirable. However, an automated training/tuning procedure in turn requires timely access to accurately annotated training data. Here, we refer to annotated data as data labelled either normal or faulty that represent the system's (sensor) state. The degree of required annotation ranges from complete annotation (both faulty and normal classes annotated) to unannotated data, depending on which FDM that is used. So-called *supervised* classification methods exploit annotated data from both classes in contrast to *unsupervised* methods that exploit data without annotation, see e.g. (Bishop 2006). FDMs that only use data from one class only (commonly from the normal mode) are called *one-class classification* methods (Pimentel et al. 2014). In short, the higher degree of annotation in training data, the easier the fault detection task becomes since more information is available about the difference between a fault and the normal condition.

Annotated data unfortunately are scarce in practice. The reason for lack of faulty annotated data is clear; faults should be rare. But even when faults are present, annotation is problematic:

- It is difficult and time-consuming to identify the starting time of e.g. a sensor fault.
- There is seldom a clear-cut transition between the normal and the faulty mode.
- The definition of what is faulty is subjective and depends on the context.

Also, normal annotated data are challenging. Data considered to be normal during daily operations, often contain unknown flaws, which make them unrepresentative for the normal state and therefore also unsuitable for FDM training.

7.1.5 Challenges and our solutions

The goal with our study is to improve the detection level to 0.5 mg/L bias on DO sensors for the data in (Samuelsson et al. 2018). Further, we conduct our study with realistic constraints that are needed to be solved in a full-scale implementation, which include:

- timely access to annotated training data
- automated tuning of alarm thresholds using a realistic amount of annotated normal data
- automated training to allow the FDMs to adapt to changing normal conditions such as worn sensor and/or seasonal variations

We address these challenges by:

- labelling data from the 24-h period after manual sensor verification as being normal
- using small (24h) normal annotated data for FDM training
- only consider automated training and tuning procedures with minimal data pre-processing and end-user input decisions

We assess how both supervised and unsupervised FDMs can be modified to analyse unconventional data with IRs generated with AFD. Note that only normal training data are considered. Further, we combine the performance measure receiver operating characteristics (ROC) see (Kay 1998), with a data transformation model in (Venkatasubramanian et al. 2003) to get insights in the interplay between data and FDMs.

The majority studies considering AFD have been theoretical simulation studies, with few real-world applications. This study is one of few where real data are considered, and to our knowledge the first one with a WRRF application. This study therefore pushes the state-of-the-art knowledge of the applicability of AFD to real systems.

We finally stress that AFD is a different approach compared to previous research in the field, which has so far been driven by interpreting and combining existing data, see e.g. the review by Corominas et al. (2018).

7.2 Materials and Methods

This section is structured as follows. Section 7.2.1 gives an overview of the datasets in Chapter 3 and Section 7.2.2 describe how they were evaluated here in terms of training and test data. Section 7.2.3 explains introduce a data transformation model that was used to assess the performance of the FDMs. The choice of FDMs is motivated in Section 7.2.4 and a technical description of the FDMs can be found in Appendix A7.1.

7.2.1 Data

An overview of the datasets used in this study is given in this section, but details can be found in Chapter 3.

An IR is here defined as a 60 seconds long data sequence with temporary elevated DO-values from a DO-sensor, which has been subjected to a 20 seconds long impulse of pressurized air (*Figure 7.1(a)*). The 60 second IR time frame starts jointly with the air impulse. The qualitative shape of an IR has been shown to be informative about the extent of biofilm fouling on the sensor (*Figure 7.1*(b)). It is the task for the FDMs to differentiate between normal, and fouled IRs that cause a bias.

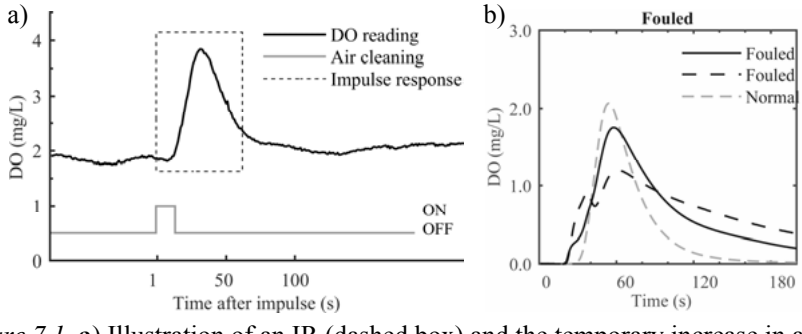


Figure 7.1. **a**) Illustration of an IR (dashed box) and the temporary increase in ambient DO concentration (black) after an air cleaning event (grey). **b**) Typical differences in the IR shapes between normal (black, solid line) and fouled MEC-type DO-sensor (dashed lines), from Chapter 3.

Two datasets produced in Chapter 3 were used in this study, one with artificial (grease) fouling denoted D_{art}, and one with real biofilm formation denoted D_{real}. The D_{art} was produced by recording IRs at different locations in a conventional activated sludge process (ASP) with varying degrees of manually applied grease film, as an imitation of biofilm formation. In total, D_{art} contained 50 normal and 2628 fouled IRs. The D_{real} was produced in a different WRRF than D_{art} where real biofilm formation on sensors is problematic. The D_{real} contained six experimental periods, each one with a length between 10

and 62 days. Each experimental period started with a clean sensor, and the extent of biofilm formation increased until the sensor was manually cleaned, and the next experimental period begun. In both datasets, the bias at different levels of fouling was obtained as the difference between the measured DO concentration in the fouled sensor (before an IR), and the value from a (clean) reference DO-sensor that measures close to the fouled sensor. In D_{art} only DO-sensors of MEC type were fouled, but in D_{real} both MEC and OPT sensors were fouled and evaluated.

7.2.2 Realistic and automated training procedure

As mentioned in Section 7.1.4 it is essential to develop practically feasible FDMs, meaning that realistic constraints need to be tackled. Three such constraints are critical to automate the usage of FDMs:

1. Timely access to annotated normal data
2. Automatic training to adapt to changing normal conditions
3. Automatic tuning of alarm threshold.

Note that we distinguish between training (adapting the FDM to normal conditions) and tuning (setting the alarm threshold for a desired trade-off between sensitivity and robustness).

In the following five sections we motivate why the constraints in our study are realistic and reflect solutions available for a full-scale implementation. The two datasets in this study were used for different purposes with different training and evaluation procedures, which are described in Section 7.2.2.4 and Section 7.2.2.5. Note that all FDMs were given the exact same raw data and conditions for tuning, training, and evaluation. Minimal data pre-processing was applied, where only outliers (less than 10 IRs in total) were initially removed. The removed IR outliers contained partly frozen signals caused by data transfer problems.

7.2.2.1 Timely access to annotated normal training data

We defined the 24 hours after an approved manual sensor verification as normal data and argue that within 24 hours it is reasonable to assume that no fouling or sensor wearing has occurred. Here, a verification consisted of manual cleaning and compensation for potential bias due to drift after the last verification. Sensor verification was conducted in the beginning of each experimental period in D_{real} and for all normal IRs in $D_{art.}$ In full-scale sensor maintenance information system, date and time for manual cleaning and calibration are stored and provide the necessary information for sensor verification. Such data can automatically be transferred to a FDM as a realistic solution for providing the required normal annotated data.

The number of training IRs was limited to 12 to only use realistic (small) training datasets. The reason was that 24 hours of normal annotated training data only contain 12 IRs per sensor when the sensor supplier's recommendation of 2 hour's cleaning interval is applied. Too frequent air-cleaning wears the sensor out and would also reduce the time when the DO-sensor can be used for automatic control (during the 2-3 minutes of air-cleaning, the DO-sensor is not sensing the true DO concentration).

7.2.2.2 Automatic training to adapt to changing normal conditions

It is important to have automatic procedures to compensate for non-stationarity in WRRF process data (seasonal variations) and wear in sensor that may cause changes in its IR. In this study, and for a potential full-scale application, we assumed that the 12 IRs (24h) after a sensor verification were accessible for automatic re-training the FDMs. In practice, sensor verifications (and re-training) are expected to occur in intervals of 2-4 weeks, which is enough to adapt to seasonal variations in data and wearing in sensor.

7.2.2.3 Automatic alarm threshold tuning

We used the off-line approach described by Gustafsson (2000) to set the alarm threshold for all FDMs. The approach requires the user to specify a desired false alarm rate, provide a normal dataset, and an acceptable range for the alarm threshold. Thereafter, an automated procedure is conducted where all alarm thresholds in the acceptable range are evaluated with respect to false alarm rates on the normal dataset. The alarm threshold closest to the desired one is chosen for the on-line application. The implicit assumption for the approach is that the normal training dataset is representative for the full probability distribution of the normal mode. This means that a training dataset with 12 IRs is too small to use for tuning purposes. We however expect that it is a realistic effort for an operator to produce a larger training dataset (about 1 week's data) once every year with the purpose of tuning the alarm threshold. It is an open question whether each DO-sensor would need individual tuning, or if the same dataset and tuning could be used for all sensors of equal types.

7.2.2.4 Dataset Dreal – training, tuning and performance assessment

The D_{real} was evaluated to imitate realistic and automated full-scale usage of the assessed FDMs. The first experimental period, Period 1, was used as off-line normal data to tune the FDMs to a 10% false alarm rate. The 10% threshold was chosen slightly higher than what is desired in a full-scale application where 1% - 5% is more reasonable. The purpose with a 10% alarm limit was to simplify the interpretation of the results, e.g. it is not possible to evaluate whether a FDM that have much lower false alarm rates on test data than 1%.

For the experimental periods 2-5, the first 12 IRs in each period where used for training as described in Section 7.2.2.2, and the remaining IRs for evaluation. False alarm rate and detection rate were used as performance measures

for the evaluation. Note that the 10% desired false alarm rate was the only part specified by the researchers, all remaining training and tuning steps were conducted in an automated way without subjective interference.

7.2.2.5 Dataset Dart – training and performance assessment

The D_{art} was used to assess the performance of the FDMs with equal amount of training data as for D_{real} (12 IRs) but without tuning the alarm threshold. Instead, the full range of possible alarm thresholds was evaluated for both false alarm and detection rates and visualized in a graph. A graph where each point on the curve corresponds to a false alarm probability (x-axis) and detection probability (y-axis) for a specific alarm threshold is known as receiver operating characteristics (ROC), see (Kay 1998). The ROC enables an assessment of detection performances between FDMs with optimal alarm thresholds, where only the interplay of training data and FDMs are assessed. We use the term optimal alarm thresholds since the value for the alarm threshold can differ for different training and test datasets, which is not applicable in practice. The ranges of alarm threshold for all FDMs are provided in the Supplementary Materials S7.1 in Table 2. The ROC curve can be evaluated in terms of a scalar measure by integrating for the area underneath the ROC curve (AUC). Here, we considered the partial area under curve, $pAUC_{\alpha}$, up to a false alarm ratio α. The reason for only considering part of the AUC area was that it is only FDMs with low false alarm rate that are relevant for practical implementations. The actual evaluation was conducted by randomly selecting 12 training IRs from the 50 normal IRs. The remaining 38 normal and 2628 faulty IRs where used to compute results for the ROC curve. The evaluation was repeated four times with different training IRs to get indications about how sensitive the results were based on the selection of training data.

7.2.3 Transforming data to diagnostic information

As described by Venkatasubramanian et al. (2003), any diagnostic decision making process can be described by its underlying data transformation process, see *Figure 7.2*. Traditional FDMs are designed to maximize the diagnostic information content from existing raw (process) data. In AFD, the information content in raw data is increased by a priori knowledge of the system and diagnostic problem which is used to design an appropriate test signal. Further, it is the subsequent combination of data transformation steps that decide the overall diagnostic performance of any FDM. Note that the information flow is in one direction from left to right in *Figure 7.2* Thus, lost information in one early data space cannot be recovered in subsequent steps. This highlights why the AFD has potential since it is the first step (excitation space) in the data transformation process.

In this study, we used the data transformation process as a tool to separate different parts of a FDM and thereby allow increased insights into why some

FDMs fail or not. One relevant example concerns the alarm threshold (class space) and the ROC curve (feature space). If good performance for a FDM is obtained in decision space but poor performance in class space, this indicates that it is the tuning of the alarm threshold that fails. The characteristics and separation of data spaces for all FDMs are described in Table 1 in Appendix A7.1.

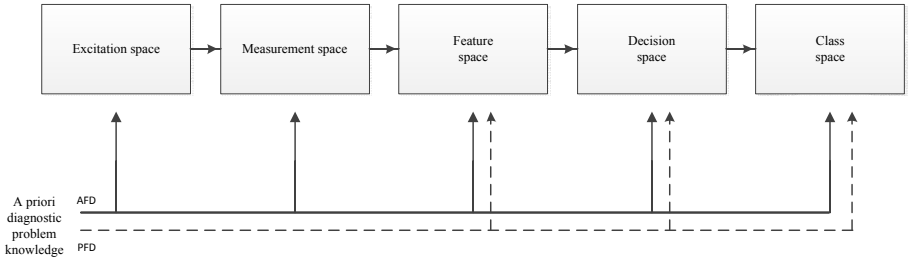


Figure 7.2. The difference in data transformation steps between traditional (passive) fault detection (PFD) and active fault detection (AFD). In AFD (solid lines), prior knowledge is also used to increase diagnostic information in raw process data (measurement space) in contrast to traditional fault detection (dashed lines). The figure has been adapted from (Venkatasubramanian et al. 2003) where the different data spaces are further exemplified.

7.2.4 Fault detection methods

In this section we motivate the choice of FDMs. A detailed description of the evaluated FDMs is given in Appendix A7.1.

State-of-the-art FDMs potentially give better detection performance, at the risk of being more complicated compared to simple FDMs. Complicated FDMs are however more costly with regards to operator training (method usage and tuning) and on-line integration (implementation, debugging, and data storage). Thus, a cost effective FDM is a trade-off between complexity and detection performance. We therefore compared four FDMs of varying degree of complexity, namely: the pragmatic engineering approach (PEA), response time estimation (RT), principal component analysis (PCA), and Gaussian process regression (GPR).

The PEA was a straight-forward usage of the descriptive statistics mean and standard deviation. The RT have been described and evaluated in Chapter 3 and thus serves as baseline performance for the other FDMs. The PCA is the most cited FDM in WRRF applications (Corominas et al. 2018) and one of few FDMs that has actually been implemented on-line in WRRFs, see (Wade et al. 2005). The mathematics of PCA are well known and have been described e.g. in the tutorial review by (Bro and Smilde 2014) and specifically for WRRF applications in (Haimi 2016). Thus, PCA is expected to be familiar to

some WRRF operators and relevant as benchmarking method. The most advanced FDM in this study was GPR which is a state-of-the-art pattern recognition algorithm that has been widely spread in the Machine learning community, see (Rasmussen and Williams 2005) and its citations. The GPR has also been described and introduced to WRRF applications in Chapter 5. We expected GPR to be the best method since it is properties should make it ideal to model changes in the shape of an IR. In addition, GPR produce uncertainty intervals in its default from that are useful for automatically tuning alarm thresholds.

We expected that the ambient DO concentration before an IR, denoted initial DO concentration, should impact the height of the IR. Therefore, we added the feature initial DO concentration to all FDMs and thereby creating several variants of the four FDMs, see Appendix A7.1 for a detailed summary.

We were especially interested in evaluating different variants of GPR since it was the most advanced method that was also the least studied FDM for WRRF applications. For the GPR, the initial DO concentration in combination with the IR was used to construct a two-dimensional non-linear regression model for the shapes of IRs. We expected that this should be a flexible and successful model that could differ any changes in the IRs due to fouling from normal IRs. We compared two, two-dimensional GPR variants (GPR_{2se} and GPR_{2ard}, see Appendix A7.1 for details) with two, one-dimensional variants (GPR_{1seBF} and GPR_{1seLR}). The one-dimensional variants had a different detection mechanism in terms of decision variables where the likelihood ratio between training IRs and test IRs was used as decision variable. This is not the standard way to use GPR for FD and we were curious whether such measure could be better than the two-dimensional variants. We provide a detailed description of the GPR variants in Appendix A7.1.

7.3 Results and Discussion

The results for all FDMs on D_{real} and D_{art} are jointly shown and discussed. We focus on three key topics:

1. detection performances of the FDMs in terms of detecting the desired 0.5 mg/L bias (Section 7.3.1, 7.3.3, 7.3.4, and 7.3.7) and avoiding false alarms (Section 7.3.2)
2. complications that are related to automatic tuning and training in Section 7.3.2.1-7.3.2.2, 7.3.5, and 7.3.6.
3. transferability to a full-scale applications and related aspects (Section 7.3.8).

The results on D_{art} (*Figure 7.3*) give an overview of the general performance ordering of all FDMs. The detection performances on D_{real} are provided for

the two sensor types OPT and MEC in *Figure 7.4* and *Figure 7.5*, respectively. Detailed results such as tuning and detection performances for all methods and datasets are given in Table 7.1.

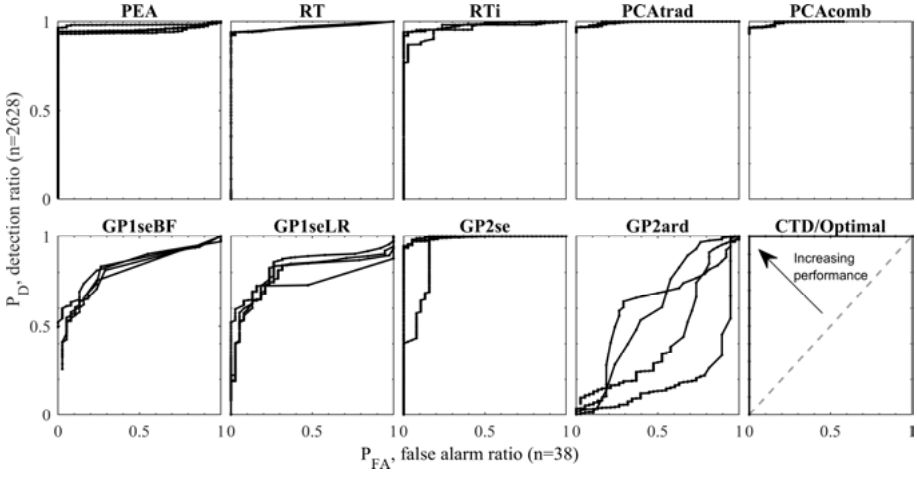


Figure 7.3. Receiver operating characteristics (ROC) for the FDMs applied to D_{art} with four different training datasets (iterations). The ROC consists of false alarm ratio (P_{FA}) on x-axis and detection ratio (P_D) on the y-axis for different threshold values. The coin toss detector performance (CTD, grey dashed line, bottom right), gives the result of randomly raising an alarm; and the optimal detector (black line, bottom right) are provided for comparison.

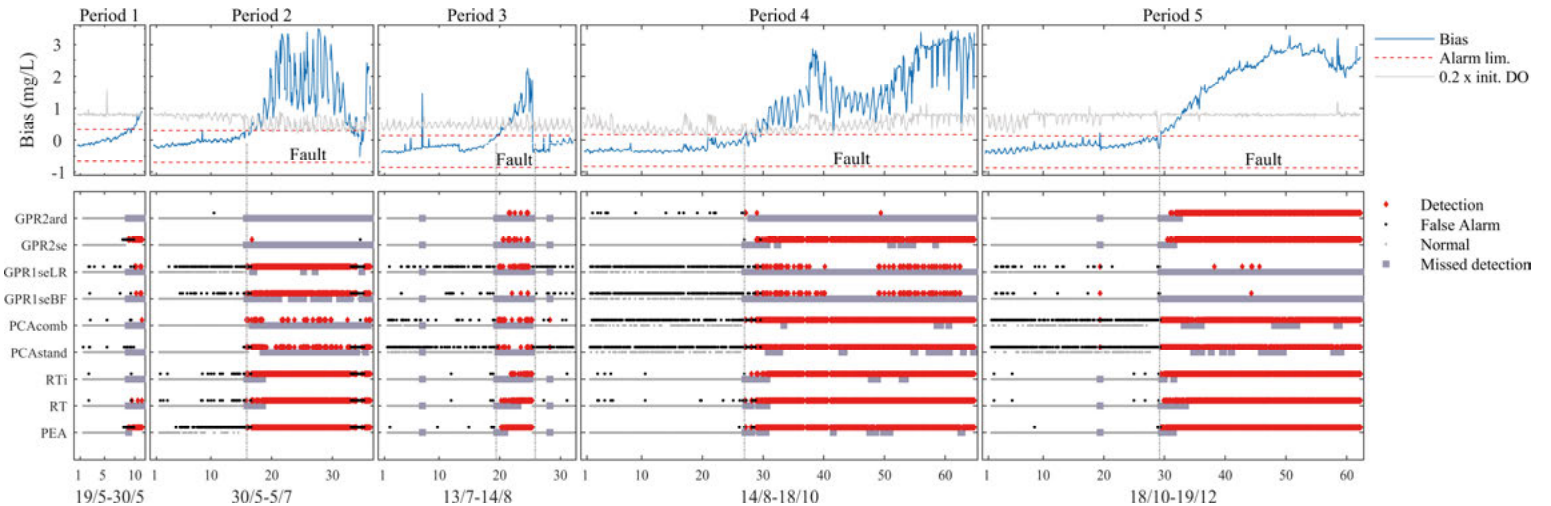


Figure 7.4. Detection results for the nine FDMs on real biofilm data (D_{real}) for fluorescent sensor (OPT). The first period was used for tuning alarm thresholds (Period 1) and the remaining four test periods (Period 2-5) for evaluating detection performance. The four different outcomes at each IR are indicated as either: detection of a fault (red diamond), false alarm (black dot), no fault and no alarm (grey dot), and missed detection – fault but no alarm (grey-blue square). Bias magnitude (blue) and 0.5 mg/L alarm limit (red dashed) are shown together with the DO concentration in the ambient wastewater (grey). Note that the ambient DO concentration has been rescaled by multiplication with 0.2 to increase the graphs readability. The faulty periods are marked with vertical grey dashed lines.

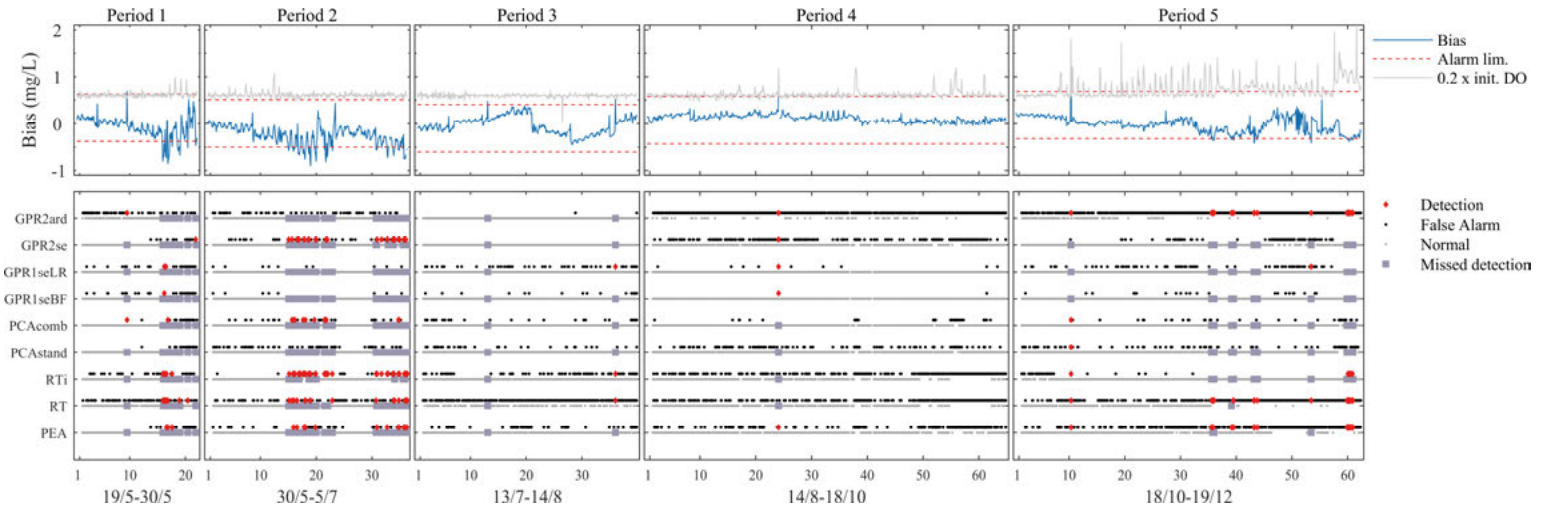


Figure 7.5. Detection results for the nine FDMs on real biofilm data (D_{real}) for electrochemical membrane sensor (MEC). The first period was used for tuning alarm thresholds (Period 1) and the remaining four test periods (Period 2-5) for evaluating detection performance. The four different outcomes at each IR are indicated as either: detection of a fault (red diamond), false alarm (black dot), no fault and no alarm (grey dot), and missed detection – fault but no alarm (grey-blue square). Bias magnitude (blue) and 0.5 mg/L alarm limit (red dashed) are shown together with the DO concentration in the ambient wastewater (grey). Note that the ambient DO concentration has been rescaled by multiplication with 0.2 to increase the graphs readability. No fault periods are marked to avoid blurring of the exact location of faults in Period 2 and Period 5.

Table 7.1. *Detection performance statistics for fluorescent DO sensor (OPT, grey rows) and membrane electrochemical sensor (MEC uncoloured rows) sensor on D_{real} (row 1-2) and on D_{art} (row 3-5). Number of non-faulty (n_{nf}) and faulty (n_f) IRs in D_{real} are given within parenthesis in the first column. The false alarm (P_{FA}) and detection ratios (P_D) are shown in columns two to ten alongside with the alarm threshold values obtain during tuning in Period 1 (P1). The threshold values have the same acronyms as in Table 1 Appendix A7.1. Note that the table is split on two pages.*

	Period #/ # of iteration		**PEA**		**RT**		**RT_i**		**PCA_{stand}**	
	Period	**(n_{nf}/n_f)**	**P_{FA}/P_D**	**β**	**P_{FA}/P_D**	**γ**	**P_{FA}/P_D**	**γ_i**	**P_{FA}/P_D**	**α**
Dreal	P1	(93/25)	0.10/0.96		0.021/0.12		0.021/0.0		0.11/0.0	
	P2	(201/218)	0.65/1.0		0.22/0.93		0.21/0.98		0.14/0.48	
OPT	P3	(298/70)	0.020/0.83	6.1	0.020/0.59	2.2	0.017/0.26	1.2	0.40/0.23	1e-8
	P4	(330/423)	0.036/0.96		0.082/0.98		0.036/0.94		0.67/0.96	
	P5	(340/390)	0.038/0.98		0.032/0.93		0.032/0.98		0.64/0.93	
	P1	(224/24)	0.094/0.13		0.29/0.38		0.12/0.21		0.098/0.08	
	P2	(359/60)	0.095/0.22		0.22/0.28		0.036/0.78		0.18/0.0	
MEC	P3	(458/2)	0.10/-	4.1	0.62/-	1.7	0.16/-	0.3	0.07/-	1e-13
	P4	(747/1)	0.24/-		0.43/-		0.41/-		0.16/-	
	P5	(707/23)	0.45/0.83		0.58/0.96		0.062/0.26		0.16/0.0	
	Iteration	**(n_{nf}/n_f)**								
Dart **Low P_{FA}**	1	(38/519)	0.11/0.98	14	0.11/0.94	1.5	0.11/0.88	5.0	0.11/0.98	1e-13
	2		0.11/0.93	9.5	0.13/0.94	2.2	0.11/0.94	1.5	0.11/0.97	1e-6
	3		0.11/0.95	9.6	0.11/0.94	1.8	0.11/0.95	1.6	0.11/0.97	1e-3
	4		0.11/0.93	11	0.13/0.94	1.0	0.11/0.94	1.0	0.079/0.97	1e-3
Medium P_{FA}	1	(38/519)	0.21/0.98	14	0.16/0.94	1.4	0.21/0.96	3.0	0.18/1.0	1e-4
	2		0.21/0.93	9.1	0.18/0.94	1.3	0.21/0.95	1.2	0.21/0.99	0.06
	3		0.21/0.95	9.2	0.16/0.94	1.0	0.21/0.95	1.1	0.18/0.98	0.04
	4		0.21/0.94	11	0.24/0.94	0.7	0.21/0.95	0.8	0.21/0.99	0.05
			% of optimal detection performance (100% is optimal)							
$pAUC_{0.1}$	1	(38/519)	89	-	81	-	72	-	59	-
	2		82		81		82		**97**	
	3		83		82		82		**97**	
	4		81		82		82		**97**	

	Period #/ # of iteration		PCA_{comb}		GPR_{1seBF}		GPR_{1seLR}		GPR_{2se}		GPR_{2ard}	
	Period	**(n_{nf}/n_f)**	α	P_{FA}/P_D	α	P_{FA}/P_D	$B_{f_{crit}}$	P_{FA}/P_D	LR_{crit}	P_{FA}/P_D	δ	P_{FA}/P_D	δ
Dreal													
OPT	P1	(93/25)	1e-8	0.053/0.04	1e-2	0.032/0.12	34	0.11/0.20	12	0.13/1.0	11	0.0/0.0	12
	P2	(201/218)		0.015/0.23		0.21/0.81		0.48/0.97		5.0e-3/4.6e-3		5.0e-3/0.0	
	P3	(298/70)		0.11/0.37		0.078/0.057		0.22/0.31		0.0034/0.23		0.0034/0.10	
	P4	(330/423)		0.61/0.99		0.55/0.17		0.54/0.15		0.030/0.91		0.052/0.012	
	P5	(340/390)		0.58/0.92		0.047/2.6e-3		0.091/0.013		0.018/095		0.018/0.93	
MEC	P1	(224/24)	1e-13	0.10/0.083	0.35	0.13/0.042	43	0.13/0.083	30	0.067/0.042	6.2	0.27/0.042	2.5
	P2	(359/60)		0.087/0.18		0.022/0.0		0.028/0.0		0.13/0.47		0.17/0.0	
	P3	(458/2)		0.059/-		0.033/-		0.10/-		0.0/-		0.0043/-	
	P4	(747/1)		0.056/-		0.0013/-		0.012/-		0.29/-		0.90/-	
	P5	(707/23)		0.062/0.0		0.031/0.0		0.099/0.043		0.14/0.0		0.83/1.0	
Dart	**Iteration**	**(n_{nf}/n_f)**											
Low P_{FA}	1	(38/519)	1e-13	0.080/0.97	1e-7	0.079/0.59	207	0.11/0.58	400	0.11/0.64	16	0.11/0.021	2.2
	2		1e-6	0.11/0.97	1e-6	0.11/0.56	138	0.11/0.60	300	0.11/0.97	4.5	0.11/0.034	31
	3		1e-3	0.11/0.97	6e-3	0.11/0.61	138	0.079/0.60	300	0.11/0.98	3.6	0.11/0.13	28
	4		1e-3	0.13/0.97	1e-4	0.079/0.63	161	0.11/0.65	90	0.079/0.99	3.1	0.11/0.070	5.5
Medium P_{FA}	1	(38/519)	1e-4	0.21/1.0	0.03	0.18/0.68	69	0.21/0.72	30	0.21/1.0	4.8	0.18/0.14	2.0
	2		0.06	0.21/1.0	0.2	0.24/0.75	23	0.21/0.74	30	0.21/0.99	2.5	0.21/0.043	22
	3		0.04	0.21/0.99	0.2	0.18/0.66	69	0.21/0.73	30	0.21/0.99	2.5	0.21/0.17	15
	4		0.05	0.21/0.99	0.01	0.24/0.81	46	0.16/0.72	-27	0.21/1.0	2.0	0.21/0.41	2.8
			% of optimal detection performance (100% is optimal)										
$pAUC_{0.1}$	1	(38/519)	-	96	-	39	-	39	-	40	-	*0.91*	-
	2			**97**		40		39		83		2.8	
	3			**97**		43		42		84		9.0	
	4			**97**		59		52		84		*4.3*	

7.3.1 The best performing fault detection methods

A bias larger than 0.5 mg/L was successfully detected in D_{real} for the OPT sensor with a low number of false alarms with the three simplest FDMs: PEA, RT, and RTi (*Figure 7.4*). Only GPR2se gave good results (disregarding Period 2) on the same data among the more advanced GPR and PCA methods. Similarly, the PEA, RT, RTi, and GPR2se also performed well on D_{art} (*Figure 7.3*). The MEC sensor on D_{real} (*Figure 7.5*) only obtained a bias above 0.5 mg/L for minor parts of Period 2 and 5, and it was therefore mainly false alarm rates that could be evaluated.

We believe that it is interesting that the advanced methods were not superior to the basic methods. We expected both PCA and GPR to improve the interpretation of the IRs compared to the basic RT that only use one point from the IR, whereas PCA and GPR analyse the full IR. Our initial assumption was that when only small training datasets are available, it should become even more important with advanced FDMs that can squeeze out any available information. The results here do not support our initial assumption and indicate that the added workload of using advanced FDMs was not motivated for this application.

7.3.2 False alarm distribution on MEC sensor data

A general challenge with FDMs is to reduce false alarms and understand their source. There are many potential explanations for the abundance of false alarms for the MEC sensor data in *Figure 7.5* where the most obvious are:

- influence from varying initial DO concentration during test periods (grey line, *Figure 7.5*)
- non-representative tuning data resulting in an inappropriate alarm thresholds obtained during Period 1 (the tuning period).

We focused on understanding the source of false alarms for the best performing FDMs as reported in Section 7.3.1, i.e. PEA, RT, RTi, and GPR2se.

7.3.2.1 Impact of initial DO concentration

Both RTi and GPR2se had fewer false alarms in Period 5 compared to Period 4, despite a dramatic variation in the initial DO concentration (between 3 and 6 mg/L Period 5, see *Figure 7.5*). This is the opposite of would be expected if the initial DO concentration during test periods was the reason for the false alarms. During day 38 in Period 4, all four FDMs gave false alarms when the DO concentration was abruptly increased. These results indicate that sudden changes in the DO concentration can produce false alarms, although it was not the main reason for false alarms for RTi and GPR2se. Our explanation is that

both RTi and GPR2se made use of the initial DO concentration as a feature and could therefore avoid false alarms due to varying DO concentrations. This explanation is strengthened by the fact that RT and PEA that did not use initial DO concentration as a feature (Table 1 in Appendix A7.1) and had more false alarms in Period 5.

7.3.2.2 Interaction between tuning, training data and alarm threshold values

To study whether the training data were representative for normal conditions, we compared the variation in the 12 training IRs for all experimental periods, Period 2-5. The training data differed in initial DO concentration (see Supplementary materials S7.1, Figure 1-5), and interestingly, Period 4 with most false alarms had the smallest spread in initial DO concentration in its training dataset.

To test the influence of spread in initial DO concentration in training data, on false alarm rates, we replaced three IRs in the original training data, with IRs with either a smaller or larger spread in initial DO concentration. The spread was increased in Period 4 and reduced in Period 5 and we then repeated the evaluation (see Supplementary Materials S7.1, Figure 6-7). If insufficient spread in training data was the underlying cause for the false alarm rates, an increase in the spread for training data would reduce the false alarm rate and imply that larger variations in the IRs in test data are accepted as normal. The opposite argumentation applies for Period 5 where we reduced the spread, compared to the original training data. The revaluated results showed that changing the spread in initial DO concentration in training data had a clear effect on the false alarm rates (see Figure 8 in Supplementary Materials S7.1). The GPR2se and RTi dramatically decreased their false alarm rates in Period 4 (GPR2se from 29% to 8% and RTi from 41% to 1%). In period 5 the false alarm rates increased from 14% to 18% (GPR2se) and from 6% to 52% (RTi) when the spread instead was decreased. Neither PEA nor RT were affected by the spread in in initial DO concentrations in training data. This was reasonable since neither of them used the initial DO concentration explicitly in feature space and they should therefore not be affected by a changed spread in the initial DO concentration.

We conclude that sufficient spread in initial DO concentration in training data was necessary to reduce false alarms for both GPR2se and RTi. We hypothesize that "sufficient" should be put in relation to the ratio in variations between training and tuning data.

7.3.3 The one-dimensional GPR variants

The one-dimensional GPR variants GPR1seBF and GPR1seLR had better performance than GPR2ard, but worse than all other FDMs on D_{art} (*Figure 7.3*).

The results showed that the marginal likelihood was not a good feature to separate normal from fouled IRs. The separation between normal and faulty distributions in feature space sets an upper limit for the performance in the next data spaces (decision and class space). It was therefore not surprising that none of the one-dimensional GPR variants gave good results on D_{real} (*Figure 7.4*-*Figure 7.5*) where also class space is included. In addition, they had non-robust performance on D_{art} and on the test periods similarly as the GPR2ard variant. In the end, the one-dimensional GPR variants had worse performance than the simple RT and PEA and were therefore not considered further.

7.3.4 Two dimensional GPR variants

We previously noted that GPR2se had good detection performance but missed the faults in Period 2 (reconsider the OPT sensor data in *Figure 7.4*). Similarly as on the OPT sensor data, GPR2se had one iteration with deviating (poor) results on D_{art}. Also, GPR2ard had varying performance, although with poor results in general. We repeated the evaluation for GPR2se and GPR2ard 10 times on the OPT sensor data, which showed that their performance varied (Supplementary Materials S7.1, Figure 9). Since it was only the FDM training (albeit the same training data) that differed between the 10 iterations, we conclude that the varying performance for both GPR2se and GPR2ard was an effect of the stochastic properties in the method. This was surprising since the particle filter variant of GPR showed robust results in Chapter 5. Note that the only difference between GPR2ard and GPR2se was that the ARD kernel allowed separate parameter values for both dimensions in contrast to the SE kernel which used equal parameter values for both dimensions, see Appendix A7.1 for the technical description of GPR. The additional flexibility in the ARD kernel was clearly a drawback and decreased the GPR2's performance from good (GPR2se), to worse than what would have been obtained by chance (compare GPR2ard with CTD in *Figure 7.3*).

7.3.5 The tunability problem

In this section we define the *tunability problem* as when the desired false alarm rate is not reached for a FDM on a given dataset, regardless of alarm threshold value that is used. The automatic off-line tuning procedure described in Section 7.2.2 relied upon obtaining a user specified false alarm rate, to tune the FDMs. The results however showed that this was not applicable for all FDMs.

An example of the tunability problem was for RT and RTi which were tuned to 2% false alarm rate in contrast to the desired 10% (compare false alarm rates in P1 first row in Table 7.1). The reason for RT and RTi, not to be tuned at 0.1 was because it was not possible to separate 10% of the normal IRs in the tuning dataset in Period 1 while iterating within the given range of alarm threshold values (i.e. the rise time limit). The tunability problem was present

in other FDMs as well (GPR1seBF and GP2ard) in the OPT sensor data and also in the MEC sensor data (RT, GPR2se, GPR2ard) if considering a ±5 percent unit deviation from desired false alarm rate as acceptable. For the RT, the reason was a too low sampling time of 1 second which in practice round off all rise time values to full seconds, that in turn made it difficult to separate small differences in IRs.

To conclude, it was the insufficient separation of (normal) training data in feature space that was the cause for the undesired tuning and not poorly chosen threshold values in class space. We expect the tunability problem to increase if lower false alarm rates such as 1% are required. Also, if 1% alarm rate is desired with one detected IR, the minimal number tuning dataset would be 100 IRs (8 days). An alternative would be to assume a probability distribution, and estimate its parameters based on the normal data. In such scenario, the 1% level could be identified in the estimated distribution. From the results it seems like the tuning dataset also need to contain IRs spread out in the assumed normal distribution (we don't necessary refer to a Gaussian distribution). We conclude that the suggested approach for tuning is inappropriate for this application, and likely also similar applications with small tuning datasets.

7.3.6 Inconsistent detection performance

Disregarding the tunability problem and that not all FDMs obtained 0.1 false alarm rate, it is still desirable that they at least obtain similar false alarm rates during the test periods (Period 2-5), as during the tuning period (Period 1). Such behaviour indicates predictable false alarm rates based on the tuning and provides confidence that data transformation from measurement space to class space is consistent with different data. However, none of the nine FDMs produced consistent results with regards to both false alarms and detection rates at all test periods (Table 7.1). A subjective assessment should also be made by studying the location of false alarms and detection rates with respect to their proximity of fault- and normal periods. We would get more confident in FDMs producing early detections close to the fault period compared to randomly spread out false alarms. Based on such subjective assessment, we can see that only PEA, RT and RTi in the OPT sensor data (*Figure 7.4*) gave reasonable results and none of the FDMs gave convincing results for all test periods in the MEC sensor data (*Figure 7.5*).

The inconsistent performance would be a major drawback in a full-scale application since it is critical that the operators gain trust in the FDM at hand to use it. Potentially the consistency could be improved with more representative training data as discussed in Section 7.3.2.2. The fact that all studied FDMs gave inconsistent performance supports the indication that the problem is related to training data, rather than the FDMs.

7.3.7 The PCA variants

The two PCA variants were the best performing FDMs on D_{art} and had close to optimal $pAUC_{0.1}$ performance (Table 7.1 and *Figure 7.3*), but gave poor results (too many false alarms) on D_{real} (*Figure 7.4-Figure 7.5*). The reason for the discrepancy was unsatisfactory tuning of alarm threshold values. Here, the unsatisfactory tuning refers to too tight threshold value with respect to the non-faulty training data resulting in many false alarms in D_{real}.

The discrepancy can be understood by first reconsidering the near optimal performance in $D_{art.}$ The key to this result was that the threshold value was varied to obtain all possible detection performances (combinations of false alarm- and detection ratios), and not fixed as in D_{real} (compare the threshold values in Table 7.1 for PCA in D_{art} and D_{real}). The ROC evaluation thus provides a measure of separation between normal and faulty distributions (for a given combination av data transformation steps). An additional indication that PCA produced good mode distribution separation was that neither of the PCA variants suffered from the tunability problem and achieved the desired false alarm rates (within +/- 5 percent units) in all datasets. This was however obtained with the cost of also changing the threshold values between the iterations in D_{art} (Table 7.1). Especially the first iteration had deviating threshold values in several orders of magnitude, compared to the other iterations. One IR with a different shape was present in the training data in the first iteration (see Supplementary Materials S7.1, Figure 10) and was assumed to be the reason for the deviating threshold values. In addition, the varying threshold values were non-intuitive with values in the range of 1e-3 to 1e-13 and far from the commonly used 0.95 confidence level.

To conclude, the results in this study indicate that PCA successfully separated the faulty and normal data distributions, although it was not possible in practice to use a fixed alarm threshold for different training and test datasets. The PCA needs to be further studied on how to tune and set alarm threshold values in relation to training data.

7.3.8 Topics for DO sensor monitoring on a full scale

Which FDM had sufficiently convincing results, and are there problems that need to be solved to put the FDMs in production on a full scale?

It was clear that the automatic tuning procedure needs to be improved and that the false alarm rates need to be suppressed, which are key problems to handle in a full-scale application.

We suggest three solutions to reduce the problems with false alarms: i) require a certain number of consecutive alarms before an alarm is raised. This would remove random alarms, at the cost of a delayed detection ii) two or more FDMs are combined in an ensemble monitoring system to increase the robustness towards false alarms iii) the test signal should be improved to take

into account that training and tuning data are produced so that a wide range of initial DO concentrations are include in the training data. If possible, despite the manufacturer's recommendation, the size of training datasets should be increased.

From the results, it is clear that achieving automatic tuning of FDMs with small training and tuning datasets is difficult. As mentioned in Section 7.2.2.2 an increased focus on producing training datasets with larger spread in initial DO concentration could improve the training procedure, and probably also the tuning procedure. We therefore believe that it is more important to improve the test signal and the information content in the IRs, rather than developing better FDMs. Unfortunately, a more advanced design procedure for obtaining annotated training and tuning data is unrealistic for a full-scale application since more time and efforts are required. More research on strategies for automatic tuning with small (normal annotated) datasets is needed.

Future studies should also compare how a human expert can interpret the IR's in comparison to the best FDM. Such comparison would serve as a sanity check for how good detection performance that at least can be obtained from the data. That is, if a human expert outperforms the best FDM, there is room for improving the FDM. Recent research has demonstrated that algorithms can produce better detection performance than human experts, see e.g. (Haenssle et al. 2018), and it would be interesting to study whether similar results could be obtained for the given data in this study as well.

The costs and benefits of using the FDMs need to be put into relation to the time for the cost of simply increasing the manual sensor maintenance interval. We expect the cost for a false alarm (in terms of too early sensor cleaning) to be small compared to a using a DO-sensor with 0.5 mg/L bias for aeration control. The cost in terms of lost confidence from the operators due to false alarms is however expected to be high.

Even if the human operator can make a good visual control about the biofilm formation on a sensor, we believe that automated FDMs still are valuable tools. As an example, consider a FDM that provide reliable information about the sensor condition (and likewise its data quality). Such information can be of great use to prioritize which sensor to check and maintain when time is scarce. Further, automated generation of data quality indices for many sensors can also provide quantifiable information about e.g. which lines or zones that are more suspect to fouling, which in turn can be used to improve a fixed sensor maintenance schedule. Automatic and objective storage of the condition for many sensors is time consuming and task that is better suited for automated FDMs, rather than human operators.

7.4 Conclusions

Active fault detection can provide timely information about biofilm formation on DO sensors and it is a promising concept to improve sensor maintenance. Two fault detection methods, RTi and GPR2se, reached the goal to improve the detection limit to 0.5 mg/L bias caused by biofilm fouling. The detection performance was more related to how good the training data represented the normal variations, rather than how advanced the fault detection method was. Further, the study illustrated that it is difficult to conduct automatic fault detection, especially to tune the alarm threshold when only small normal training datasets are available. More research is needed to solve how training data can be made representative for the normal conditions and included in the active fault detection framework.

Appendix A7.1 – Fault detection method descriptions

A7.1.1 The pragmatic engineering approach (PEA)

The PEA used two basic calculation routines: the mean and standard deviation, and should be intuitive to understand and straight-forward to implement. The PEA first normalizes the training IRs by subtracting their respective initial DO concentration. Then, the mean, IR_{mean}, and standard deviation values, IR_{SD}, at each time point are calculated to define the normal behavior of an IR. The final step before running the algorithm on test data, IR_{test}, is to decide the number of standard deviation values by tuning the threshold value β that defines the detection limit. The relation between these steps and the data spaces is given in Table 7.1.

A7.1.2 Rise time estimation (RT)

The RT computes the time for an IR to reach 63 percent of its peak value, see details in (Andersson and Hallgren 2015, Samuelsson et al. 2018). In this study, we include an additional variant (RT_i) that also made use of the initial DO concentration, DO_i, of training and test IRs. To compensate for a potential effect from different initial DO concentrations, the RT_i used the least squares linear regression of the correlation between initial DO concentration and time constant values for the 12 training IRs. The decision limits were then based on the prediction intervals from the least squares estimate according to

$$RT_i(DO_i)_{lim.} = \overline{RT_i}(DO_i) \pm \gamma_i \sqrt{SS_{RT_i}(1 + \frac{1}{12} + \frac{(DO_i - \overline{DO_i})^2}{SS_{DO_i}})}, \qquad (7.1)$$

where the detection limits depend on the initial DO concentration of a test IR. The $\overline{RT_i}(DO_i)$ is the mean predicted rise time based on the initial DO concentration of a test IR, $DO_i$and the γ_i is a threshold value to be tuned. The term

after γ_i is the standard error of prediction, which contain the mean squared error for the regression residuals SS_{RT_i}, and the mean squared sum of the training initial DO concentrations SS_{DO_i}.

A7.1.3 Principal component analysis (PCA)

In short, PCA extracts key linear correlations from the sample covariance matrix of correlated variables. Deviations from the normal covariance structure are used as indication for a fault. In this study, PCA was applied to the IRs where each time step from an IR (1 to 60 seconds) was treated as an individual variable. The data were standardized to zero mean and standard deviation one. Two principal components were used as they contained at least 95% of the variation during a 10-fold cross validation based on training data from Period 1 in D_{real}.

Two PCA variants with different Decision spaces were considered. The standard variant (PCA_{stand}) used the common Hotelling's T^2 index and squared prediction error (SPE) as decision variables whereas the second variant (PCA_{comb}) combined T^2 and SPE to one decision variable denoted φ as in (Yue and Qin 2001). The PCA_{stand} was implemented with threshold value with χ^2-distribution for the critical T^2 limit and the robust SPE limit from (Varmuza and Filzmoser 2009). The threshold value for φ was based on F-distribution for the T^2 limit and χ^2-distribution for the SPE, which gave less restrictive limits compared to the PCA_{stand}. The confidence level, α, was used as tunable alarm threshold value for both PCA variants, see Table 7.1.

A7.1.4 Gaussian Process Regression (GPR)

The GPR has been detailed in e.g. (Rasmussen and Williams 2005) and represents the state-of-the-art FDM in this study and is therefore the least studied algorithm applied to WRRF applications and AFD data. It is hence of interest to understand how variants of GPR can be applied to AFD data and impact its performance.

In short, GPR can be described as a non-linear regression method that performs a smooth interpolation of training data. The GPR also produces a data dependent prediction interval were both the mean prediction and the prediction interval are based on the assumption that data can be described as a Gaussian process (i.e. a multivariate normal distribution).

The GPR was modified to four variants that differed in both feature- and decision space (Table 7.1). The first two variants (GPR_{1seBF} and GPR_{1seLR}) only made use of the IRs as feature (one dimension), whereas the remaining two (GPR_{2se} and GPR_{2ard}) also included the initial DO concentration as a second feature variable (two-dimensional, indicated with a lowercase 2). The difference between the latter two variants was two different kernels: the standard squared exponential (SE) and the Automatic Relevance Detection kernel (ARD). The difference is that the ARD have separate kernel parameter for every dimension in contrast to the SE-kernel that uses the same length- and

variance parameters for all dimensions. Note that since both the SE and the ARD have same kernel construction, the GPR_{1se} would be identical to GPR_{1ard}. The kernel is an important design choice in GPR and can be described as a parametrized covariance matrix, see (Bishop 2006) for an introduction to kernels.

A critical step in training GPR is to infer the kernel parameters from data. The most common approach relies on Maximum Likelihood estimation (MLE) (Rasmussen and Williams 2005). Various Bayesian approaches also exist that make probabilistic approximations of the kernel parameter distributions, e.g. by using particle filters. In this study we used both approaches where the MLE approach from (Rasmussen and Nickisch 2010) was used in GPR1seLR, and the particle filter implementation GPR-SMC in (Svensson. et al. 2015) was used in GPR2se, GPR2ard, and GPR1seBF. To allow a fair comparison, the MLE approach was repeated equally many times as the number of particles in the Bayesian approach. Also, the GPR-SMC was modified to add a small probability value of 1.0e-14 to zero probabilities in the in uniform distribution during kernel parameter sampling to avoid numerical problems. The settings for the GPR variants are provided in the Supplementary Materials S7.1, in Table 7.1.

The difference between GPR_{1seLR} and GPR_{1seBF} was in the Decision space, which also differed from the two-dimensional variants. For the two-dimensional variants, the standard GPR approach was applied which make use of the prediction intervals, i.e. the mean prediction, GP_{mean}, and the standard deviation, GP_{SD} of the predictive normal distribution as decision limits (Table 7.1). The width of the prediction interval (alarm threshold value) was tuned by deciding the number of standard deviations, δ, in the prediction interval. For the one-dimensional variants, it was not straight-forward how to infer the kernel parameters, especially since all training data had equivalent time values (t=1...60). Here, we assumed that training data (12 IRs) were independent realizations from one underlying Gaussian process

$$p(Y_1, \dots, Y_N|\mathcal{M}) = p(Y_1|\mathcal{M}) \dots p(Y_N|\mathcal{M}) \tag{7.2}$$

where $p(Y_1, \dots, Y_N|\mathcal{M})$ is the joint probability of datasets 1 to N given a GPR model $\mathcal{M}$.

The decision variable for GPR_{1seBF} was defined as the Bayes factor

$$BF = \frac{(p(Y_1|\mathcal{M}^*) \dots p(Y_N|\mathcal{M}^*))^{1/N}}{p(Y_{test}|\mathcal{M}^*)} \tag{7.3}$$

which is the likelihood ratio of training data with the most likely model

$$\mathcal{M}^* = max\, p(Y_n|\mathcal{M}_n) \tag{7.4}$$

and a test IR, Y_{test}. Note that a BF larger than 1 indicates that training (normal) IRs have a better fit to the model, compared to test data and therefore indicate a fault. The marginal likelihoods of training data were obtained by approximating the integration of kernel parameters θ

$$p(Y_n|\mathcal{M}) = \int p(Y_n|\theta, \mathcal{M})p(\theta|\mathcal{M})d\theta \tag{7.5}$$

by a weighted sum of the contribution from the particle filter

$$\hat{p}(Y_n|\mathcal{M}) = \sum_{p=1}^{P} w^{(p)} p\left(Y_n\middle|Y, \mathcal{M}_{\theta^{(p)}}\right) \tag{7.6}$$

The GPR_{1seLR} used a similar decision criterion but where the average likelihood value of training data was instead based on MLE kernel parameter values. The decision variable was then defined as the likelihood ratio, LR, between the average likelihood of training data and the average likelihood of test data given all MLE models from training data

$$LR = \frac{(p(Y_1|\mathcal{M}_1) \dots p(Y_N|\mathcal{M}_N))^{1/N}}{(p(Y_{test}|\mathcal{M}_1) \dots p(Y_{test}|\mathcal{M}_N))^{1/N}} \tag{7.7}$$

For numerical reasons, logarithm values of LR were evaluated rather than LR explicitly.

Table 7.2 The five data spaces (excitation-, measurement-, feature-, decision-, and class space) and the related characteristics for the four FDMs and the nine variants thereof. Acronyms in column three (alarm threshold) refer to alarm threshold variables for the FDMs that are defined in Section A7.1.1- A7.1.4. Acronyms in column six (feature scape) indicate the features used by each FDM where t_{63} is the rise time, IR(t=1,...,60) refers to the full IR data sequence, PC 1-2 are the first two principal components, SE-kernel and ARD-kernel are the kernels described in Section A7.1.4, and init. DO stands for the ambient DO concentration just before an IR. The column seven (decision space) shows the decision variables for all FDMs that are described in their context in Section A7.1.1- A7.1.4. Column eight (class space) show when an alarm is raised, i.e. when the decision variable exceeds an upper or lower critical limit. The critical limits and related variables are described in Section A7.1.1- A7.1.4 .

			Data space				
Method	**Variant**	**Alrmthres**	**Excitation**	**Measurement**	**Feature**	**Decision**	**Class (fault indication)**
PEA	PEA	β	Piecewise constant test signal (air-cleaning impulse)	IR-data, DO-measurement, ASP-data	norm. IR-data	$\varepsilon_{PAE} = \mathrm{IR}_{\mathrm{test}} - \mathrm{IR}_{mean}$	$\beta\,\mathrm{IR}_{SD} < \varepsilon_{PAE} < -\beta\,\mathrm{IR}_{SD}$
RT	RT	γ			t_{63}	$\varepsilon_{RT} = \overline{t_{63,train}} - t_{63,test}$	$\gamma\,SD(t_{63,train}) < \varepsilon_{RT} < -\gamma\,SD(t_{63,}$
	$\mathrm{RT_i}$	γ_i			t_{63}, init. DO value	$t_{63,test}$	$RT_i(DO_i)_{lim.} < t_{63,test} < RT_i(DO_i$
PCA	$\mathrm{PCA_{stand}}$	α,			IR(t=1,...,60), PC 1-2	T^2, SPE	$T^2 > T_{crit}(\alpha)$ or $SPE > SPE_{crit}(\alpha)$
	$\mathrm{PCA_{comb}}$				IR(t=1,...,60), PC 1-2	$\varphi = f(T^2, SPE)$	$\varphi > \varphi_{crit}(\alpha)$
GPR	$\mathrm{GPR_{1seBF}}$	$B_{f\,crit}$			IR(t=1,...,60), SE-kernel	BF	$BF > BF_{crit}$
	$\mathrm{GPR_{1seLR}}$	LR_{crit}				LR	$LR > LR_{crit}$
	$\mathrm{GPR_{2se}}$	δ			IR(t=1,...,60), init. DO value, SE-kernel	$\mathrm{IR}_{\mathrm{test}}$	$GP_{mean} - \delta\,GP_{SD} < \mathrm{IR}_{\mathrm{test}} < GP_{mean} +$
	$\mathrm{GPR_{2ard}}$				IR(t=1,...,60), init. DO value, ARD-kernel		

7.5 Chapter acknowledgments

We gratefully thank Andreas Svensson (Uppsala University) for providing the code for GPR-SMC and for valuable discussion about its usage.

8 Monitoring diffuser fouling with grey-box modelling

Fouling of fine-pore diffusers can cause substantial aeration energy wastage. It remains challenging to monitor the condition of the diffuser and thereby decide on the optimal time for labour-intensive and costly cleaning actions. In this chapter we show that data from basic sensors (airflow rate, pressure, dissolved oxygen concentration and airflow valve position) can be used to assess the diffuser's condition. We propose a grey-box estimation method that use a sequence of valve positions, which improve the information content in the generated data, and, enables estimates of dynamic wet pressure (DWP), oxygen transfer, respiration rate and the joint alpha fouling factor (αF). The method was evaluated in a water resource recovery facility over 18 months and successfully detected a reduced cleaning effect from a reverse flex cleaning action as a change in the estimated DWP. This demonstrates its usefulness for diffusor condition monitoring. Further research is needed to assess the accuracy of the method for monitoring α and F. Humid air on airflow rate sensors were found to have a negative impact on the estimation procedure, which has not been appreciated before. Additional sensor related disturbances were identified that demonstrates that the method also can facilitate detection of sensor data quality issues.

8.1 Introduction

Despite technical improvements in fine-pore diffusers and advanced control systems, the aeration system remains the single most energy-demanding operational unit in water resource recovery facilities (WRRFs). Regular cleaning of diffusers is a key action for preventing deteriorating energy efficiency (Garrido-Baserba et al. 2017, Rosso et al. 2008). Garrido-Baserba et al. (2017) reported a 24% increase in energy consumption for the common ethylene-propylene diene monomer (EPDM) disc diffuser due to fouling and aging during the first 14 months of diffuser operation in a low rate loaded WRRF. Fouling progression was even quicker for high rate loadings.

The decrease in energy efficiency has been attributed to fouling and aging of the diffusers (Krampe 2011, Rosso and Stenstrom 2006), which both lead to decreased oxygen transfer efficiency (OTE), whereas the underlying causal

effects may differ (Garrido-Baserba et al. 2017, Jiang et al. 2017, Rosso and Stenstrom 2006). Aging and chemical fouling (scaling) both cause the diffuser membranes to become less flexible and increase diffuser pressure resistance, i.e. an increase in dynamic wet pressure (DWP) (Rosso et al. 2008, Wang et al. 2020). Both aging and fouling change the membrane properties, for example, by changing bubble size and distribution, which results in a decreased OTE (Garrido-Baserba et al. 2016, Garrido-Baserba et al. 2018, Odize et al. 2017).

8.1.1 Timely diffuser maintenance is essential but difficult to implement

Timely countermeasures are essential to minimize energy wastage due to deteriorated diffuser efficiency. Fouling can be suppressed through different cleaning actions, such as top-hose cleaning, manual mechanical cleaning, gas sparging, chemical/acid cleaning and reverse flex (RF) cleaning (Jiang et al. 2020, Odize et al. 2017, Rosso 2018). Cleaning actions are costly, resource demanding and can require a temporary reduction in treatment capacity. For example, top-hose cleaning requires a temporary shutdown of the current activated sludge process (ASP) treatment line.

It is difficult to accurately predict when cleaning is needed, since the fouling progression depends on factors that vary between sites (Garrido-Baserba et al. 2017, Odize et al. 2017) . Garrido-Baserba et al. (2017) reported a state-of-the-art prediction model for linking diffuser fouling to reduced OTE using quantification of deoxyribonucleic acid (DNA) on the diffusor surface. Such DNA analyses are, however, not viable in practice. The use of DNA analysis was motivated by the inconsistent prediction results produced by historically used predictors such as mixed liquor content (MLSS) and solids retention time (SRT). By contrast, monitoring changes in DWP in situ is straightforward (EPA 1989a, Rosso et al. 2008), but not sufficient on its own, since changes in DWP are only partly correlated with changes in OTE. Thus, direct monitoring of the OTE is also needed.

8.1.2 Assessing diffuser fouling from off-gas measurements

The OTE can be determined directly using off-gas measurements, as originally described by Redmon et al. (1983). Although the off-gas method provides accurate data, it has three drawbacks for diffuser fouling monitoring:

- The method is reliant on dedicated measurement equipment (i.e. a floating hood, vacuum pump and sensors), which are costly and require maintenance.

- A representative hood coverage is representative for only a part of the monitored diffusers. The hood needs to be physically moved to change the monitoring position, which is labour-intense.
- Changes in OTE obtained from off-gas measurements cannot separate effects from changes in α-value due to changed diffuser properties. The change in OTE for a used diffuser relative to the new diffuser is a factor F (Metcalf and Eddy 2004, Rosso 2018). Note that OTE is proportional to the product αF, and both α and F describe how effective oxygen in air is converted to dissolved oxygen (DO).

The α-value relates oxygen transfer rate in clean water and process water (Rosso 2018). In general, the α-value is difficult to monitor, as it is influenced by several factors (Amaral et al. 2019), it is dynamically changing (Jiang et al. 2017) and is influenced partly in unknown ways (Rosso 2018). For this reason, the only established method to separate changed diffuser properties (F) from changes in α is to conduct clean water OTE measurements on the actual fouled diffuser. This approach is obviously not feasible during real-time operations.

8.1.3 Estimating oxygen transfer and respiration rate from in situ oxygen dynamics

An alternative and well-studied approach for assessing OTE is to estimate its proxy, the oxygen mass transfer function ($k_L a$). This can be done recursively using, for example, a Kalman filter as in (Holmberg and Olsson 1985, Holmberg et al. 1989, Lindberg 1997). The main challenge in using this approach is to excite the system (change the airflow rate and DO concentration) sufficiently to make it possible to separate the estimation of respiration rate (r) from $k_L a$. Several approaches have been used to excite the system, including square waves (Holmberg and Olsson 1985), oscillating DO setpoints (Holmberg et al. 1989) and on/off aeration (Irizar et al. 2009). These approaches have been tested on a full scale, but only for a limited time (several days). There is a lack of long-term studies demonstrating the practicability and usefulness of in situ estimation of r and $k_L a$.

8.1.4 The suggested staircase excitation method

Here, we extend existing estimation methods by refining the airflow excitation into a series of steps that we call a ‘staircase’ to obtain information about the diffuser’s condition. This enables a more controlled estimation of $k_L a$, r and DWP, which are all related to the diffuser’s condition. The suggested approach is in line with the concept of active fault detection, which has recently

been shown to be useful for detecting biofilm formation on DO sensors (Samuelsson et al. 2019).

8.1.5 Practicability study

Reducing energy wastage due to fouled fine-pore diffusers remains a challenge because the existing method to decide a time point for cleaning requires costly and complicated off-gas measurements. Here, we propose a method for in situ condition monitoring of fine-pore aeration diffusers. The goal is to automatically obtain information about the condition of the diffuser during real-time operations using only common instrumentation. The proposed method extracts data during an airflow rate sequence, which is automatically executed via the control system. The potential of the method is demonstrated with an 18-month-long experiment, during which fouling and changes in αF and DWP were monitored in four aerated zones in a full-scale WRRF.

8.2 Material and Methods

The development, implementation and evaluation of the staircase method to assess the condition of diffusers is described here.

8.2.1 Methodology

The usefulness of the staircase method for diffuser condition monitoring was assessed by comparing the progression of estimates of DWP and αF at four positions in an ASP. Diffusers in different aeration zones were assumed to have different fouling rates depending on their location, diffuser age and aeration mode (continuous or intermittent). At each position, the effect of twice-weekly RF cleaning was assessed as an indication of the ability of the staircase method to monitor changes in diffuser conditions. The experiment was conducted for 18 months to distinguish seasonal effects from long-term diffuser fouling.

8.2.2 Experimental system description

The experiment was performed in the 350,000 p.e. Bromma WRRF, located in Stockholm, Sweden. The WRRF operated at a medium loading rate (SRT 6 days, MLSS 3,600 mg/L) in six parallel ASP lines, each divided into seven zones.

Four positions (Line 1 and Line 2 at zone 3 and 4, denoted L1z3, L1z4, L2z3 and L2z4) were selected for the experiment.

- The Line 1 diffusers were 37 months old at the start of the experiment.
- The Line 2 diffusers were 6 months old at the start of the experiment.
- Zone 3 was the first aerated zone, and only aerated during winter and was expected to exhibit more fouling than zone 4. Zone 4 was aerated as required, depending on the load situation.

The diffusers were made from EPDM (Jaeger, Jetflex HD340) and supplied air at a water depth of 4.07 m, 0.58 m above the basin bottom. The basin had a sloped bottom (1 m slope on each side of the bottom) with a 3.2 m flat area between the sloped sides. The diffusers were mounted on the flat area, with an active diffuser area density of 28% (zone 3) and 30% (zone 4).

The air supply was controlled via a cascade controller with a fixed DO setpoint for each zone, with underlying slave controllers controlling the airflow rate. The sensors for the controller included electrochemical DO sensors (Cerlic O2X DUO), thermal gas flow rate sensors (Endress+Hauser, AT70), and a magnetic induction sensor for assessing the current valve position. The control actuators included a butterfly airflow valve (GEFA, K19) driven by a positioner (Siemens, Sipart PS2). The air was distributed with constant pressure (setpoint 0.51 bar) through a joint manifold for the entire ASP.

8.2.3 Design of the staircase

Airflow maintenance (airMaint), is a maintenance procedure similar to RF cleaning (Odize et al. 2017), during which the maximum nominal diffuser airflow rate is exceeded by 20% for 10 minutes. The airMaint was originally applied to diffusers in all zones two times per week, but was modified to a sequence resembling a staircase, with 10 different valve positions (each corresponding to a time slot) (*Figure 8.1*). Each of the 10 slots was used to facilitate the estimation of parameters used to monitor diffuser fouling.

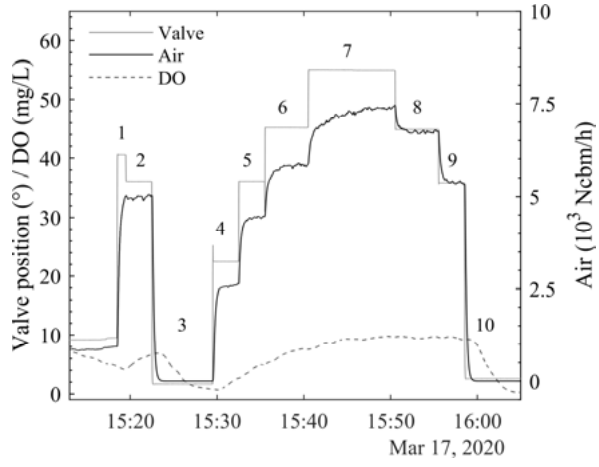


Figure 8.1. A typical response of airflow rates and DO concentrations during the 10 time slots at pre-defined valve positions as defined in the staircase method.

Slot 7 was equivalent to the original airMaint procedure, with the difference that the *valve position* was fixed, in contrast to the original setting where a setpoint for the *airflow rate* was fixed.

Slots 2, 4, 5, 6, 8 and slot 9 were used to simplify $k_L a$ estimation by fixing the airflow rate. Then, the $k_L a$ estimation was made more robust by considering the different airflow rates. Note that slots 2, 4, 5 and slot 6 are located before airMaint (slot 7), in contrast to slots 8 and 9, which were used to assess $k_L a$ after airMaint.

A hysteresis was observed for airflow rate values at a given valve position, depending on the direction of the valve's movement. Thus, the valve position in slot 1 was given a higher value than in slot 2 to ensure that the valve was always performing a 'closing motion' while entering slot 2. Similarly, the fixed staircase pattern enables comparison of airflow rates before and after maintenance for a given valve position obtained in closing motion (slot 2 and 9). To monitor potential changes in the valve hysteresis, the lowest valve position that produced zero airflow rate (in closing motion) was manually identified and used in slot 3. Similarly, the valve position that gave an airflow rate just above zero (in closing motion) was used in slot 10. Any changes in the airflow rates in either slot 3 or 10 were used to indicate an increasing/decreasing valve hysteresis.

The main purpose of slots 3 and 10 was to estimate the respiration rate, as the airflow rate was (essentially) zero in these slots.

The valve positions, at each step, were selected to cover the normal operating range and allow a sufficiently large perturbation and increase in DO. The sequence was repeated at a fixed time and on weekdays (Tuesdays and Thursdays) to minimize the impact of daily variations in the influent and α.

8.2.4 Estimating $k_L a$ and respiration rate from dissolved oxygen dynamics

The theoretical basis for describing DO dynamics in both continuous and discrete time has been described by, for example, Lindberg (1997), and is repeated here for clarity.

The DO dynamics in a completely stirred tank with volume V is given by

$$\frac{dDO(t)}{dt} = \frac{Q(t)}{V}\,(DO_{in}(t) - DO(t)) + k_L a(q_{air}(t))(DO^* - DO(t)) - r(t) \qquad (8.1)$$

where DO_{in} is the DO entering the tank, Q is the flow rate of the influent and effluent, $k_L a$ is the oxygen transfer function at airflow rate q_{air}, DO^* is the DO saturation concentration, and r is the respiration rate at time t. Equation (8.1) needs to be discretized since discrete-time measurements are to be used for estimating $k_L a$ and r. The basic Euler forward discretization was used because of its simplicity. It also produced the same results compared with the more elaborate zero-order-hold (ZOH) method during an initial evaluation. By applying Euler forward discretization, (8.1) is converted to

$$\frac{DO(k+1) - DO(k)}{Ts} = \\ = \frac{Q(k)}{V}\,(\,DO_{in}(k) - DO(k)) + k_L a(q_{air}(k))(DO^* - DO(k)) - r(k), \qquad (8.2)$$

where the time-continuous derivative $\frac{d\,DO(t)}{dt}$ has been estimated as $\frac{DO(k+1)-DO(k)}{Ts}$, where Ts is the sampling time and k is an integer denoting discrete-time samples.

In general, the respiration rate $r(k)$ needs to be estimated simultaneously with $k_L a$, as both quantities vary in time due to load and sludge activity. The $r(k)$ was estimated in the beginning and in end of each step on the staircase (slots 3 and 10), assuming that $r(k)$ from the (variance) weighted least-squares fusion of those two estimates could be approximated as a constant $\hat{r}$ during the 30 minutes between the two estimates.

The respiration rate was estimated by using the prediction error method (PEM) (Söderström and Stoica 1989) using least-squares linear regression, assuming the middle term (containing $k_L a$) on the right side in (8.2) to be zero when the airflow rate is zero (slot 3 and 10). To further simplify the estimation of the respiration rate $\hat{r}$, only DO measurements above 2.5 mg/L were considered. This avoids modelling the influence of DO on $\hat{r}$, which has an effect mainly up to 2–3 mg/L (Loosdrecht et al. 2016). DO measurements were limited to the upper theoretical DO saturation level of 13 mg/L (at 5 °C water temperature).

The $k_L a$ was estimated as a constant, denoted $\widehat{k_L a}$, for each fixed airflow rate. One $k_L a$ estimate was obtained from each of slots 2 and slots 4–9, resulting in a total of seven $k_L a$ estimates. The $k_L a$ estimation was conducted similarly to $\hat{r}$, by minimizing the prediction error. Note that $\hat{r}$ was estimated first, and later used during the subsequent $k_L a$ estimation.

8.2.5 Obtaining αF from $\widehat{k_L a}$

As mentioned in Section 8.1.2, it is difficult to separate α from the fouling factor F. Therefore, the product αF was used as an indicator for diffuser fouling, where F ranges from zero (fully fouled, with no air passing through the diffuser) to one (unfouled new diffuser). The αF can be obtained from $\widehat{k_L a}$, since α is defined as the ratio between $k_L a$ at process conditions and at clean water conditions. That is, $\alpha F \triangleq \frac{k_L a_{process}}{k_L a_{clean}}$, where F is equal to one and is omitted in the standard definition of α. The $k_L a_{clean}$ for the same conditions as $\widehat{k_L a}$ was obtained by rearranging the standard definition of standard OTE (SOTE) during clean water tests and is derived in Appendix A8.1.

8.2.6 Estimating diffuser dynamic wet pressure (DWP)

The DWP can be obtained by measuring the differential pressure over the diffuser. Here, we used only existing measurements, and no pressure measurements were available after the airflow valve (as is common in WRRFs). For this reason, the DWP was instead estimated from the airflow rate through the valve based on information about the valve's position and flow characteristics.

The pressure drop over most commercial valves is experimentally assessed and indirectly available in the form of a k_V table provided by the valve manufacturer. The k_V table consists of k_V values for different valve positions. The standard expression relating k_V to airflow rate Q and differential pressure $\Delta p = p_1 - p_2$ over a valve during subcritical flow (Nesbitt 2007) is given as

$$k_V = \frac{Q}{514}\sqrt{\frac{\rho T}{\Delta p\, p_2}} \tag{8.3}$$

The product of air density ρ and absolute air temperature T was assumed to be constant at an estimated yearly average ($T = 288°K, \rho = 1.2$ kg/m^3). Note that an increase in temperature result in a decrease in air density, which in effect makes their product (essentially) constant in the range 0 - 40 °C.

The pressure p_1 before the valve is the manifold pressure of the air distribution system (both p_1 and p_2 are expressed as absolute pressure in (8.3)). The k_V table for the airflow valve was obtained from the manufacturer and p_2, the

absolute pressure after the valve, was obtained by rearranging (8.3) and solving the resulting second-order equation. The DWP due to fouling, DWP_{foul} was finally obtained as

$$DWP_{foul} = p_2 - p_{atm} - p_{water} - DWP_{diff} \tag{8.4}$$

where the water pressure p_{water}, and the diffuser pressure resistance at the current airflow rate DWP_{diff} (obtained from manufacturer) were subtracted from the gauge pressure, $p_2 - p_{atm}$. The ambient barometric pressure p_{atm} was assumed to be constant at 1 bar.

Since the k_V values are given for a certain valve position, the DWP estimates are sensitive to valve hysteresis. Therefore, the direction of valve motion was considered while computing two parameters, $DWP_{cleaneff}$ and DWP_{avg}, that were used to monitor changes in DWP.

$DWP_{cleaneff}$ indicates the reduction in DWP from RF cleaning. This was defined as the change in DWP before and after airMaint (computed from slot 1 and 7, both in closing motion) as

$$DWP_{cleaneff} = DWP_{foul,1} - DWP_{foul,7} \tag{8.5}$$

The DWP_{avg} was defined as the mean value of the DWP estimates in opening motion (slot 3–4) before airMaint

$$DWP_{avg} = \frac{1}{2}\sum_{s=3}^{4} DWP_{foul,(s)} \tag{8.6}$$

The DWP_{avg} was a good estimate for long-term monitoring of changes in DWP, as it is based on the mean DWP just before an airMaint procedure, when the diffuser is most fouled.

8.2.7 Software, implementation and data processing

The staircase sequence was implemented in the WRRF's control system (Siemens, PCS7). Data were exported from the historical database with 1-second sampling time for all signals except water flow rate and temperature, which were exported with 1-minute sampling time. Data evaluation was performed off-line in MATLAB version R2020a.

Data from all slots were truncated by removing the first 60 seconds to remove effects from lowpass filters in the control system. Additionally, 60 seconds were removed from slots 3 and 10 to allow the decrease in DO concentration to stabilize. The different slots in each step were automatically identified by first smoothing the valve position data, and then detecting the change points using the MATLAB function 'ischange'. The smoothing was performed by applying the MATLAB function 'smoothdata' twice, using moving median ('movmedian') and linear regression ('lowess').

8.3 Results and Discussion

The results first present the estimated DWP and αF, the key parameters linked to diffuser condition. Next, the underlying estimated parameters used to obtain αF ($\hat{r}$ and $\widehat{k_L a}$) are analysed in detail. Finally, the method is assessed from a practical perspective, and its shortcomings are discussed along with suggested improvements and extensions for future applications.

8.3.1 Monitoring diffuser dynamic wet pressure

We first analyse the trends and patterns in the estimated DWP_{avg} (*Figure 8.2*). Then, we analyse the factors that potentially caused a negative offset in DWP_{avg}.

8.3.1.1 Insufficient RF cleaning in Line 1, zone 3

There was a marked increase (0.15 bar) in DWP_{avg} during 2019 in L1z3 (black dots, *Figure 8.2*). As a comparison, there was a 0.03 bar increase in DWP in polyurethane diffusers during their first year of operation (Odize et al. 2017). At the same time as the marked increase, the effect of the cleaning action was reduced as indicated by the decrease in $DWP_{cleaneff}$ (blue dashed line, *Figure 8.2*). The reason for the reduced cleaning effect was that after June 2019, the targeted maximum airflow rate was not reached during airMaint. The maximum airflow rate was particularly low (<1,000 Nm^3/h, compared with the desired 6,600 Nm^3/h) between June and October 2019 due to a 'flapping valve error', which is explained in Section 8.3.4. From January 2020 onwards, the valve was fully open during airMaint, without producing the desired airflow rate. This indicates that the diffusers in L1z3 were fouled, most likely due to the absence of RF cleaning. A higher air pressure in the distribution system would have been needed to retain the cleaning effect. The increase in DWP_{avg} was unlikely to be caused by aging effects, as L1z4 had diffusers of the same age but remained able to reach the maximum airflow rate.

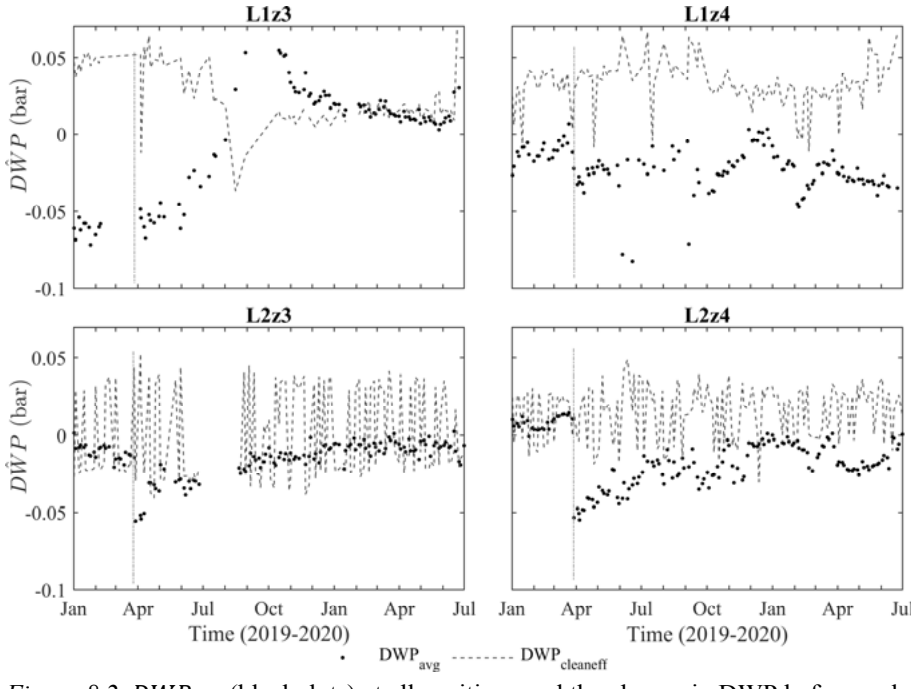


Figure 8.2. DWP_{avg} (black dots) at all positions and the change in DWP before and after airMaint, DWP_{clean} (blue dashed line). Vertical grey dashed lines indicate manual cleaning of the airflow rate sensors. All data are pre-processed and cleaned from known disturbances as detailed in Section 8.3.4. Line 1 have older diffusers than Line 2.

A negative correlation was found between the airflow rate during airMaint and DWP_{avg}, demonstrating the importance of a sufficiently high airflow rate during RF cleaning. This was evident for L1z3 (black dots, *Figure 8.3*) with the mentioned decrease in maximum airflow rate during airMaint. There is also a negative correlation for the other positions in *Figure 8.3*, although less evident. Note the two isolated groups of data for L2z3 (purple diamonds, *Figure 8.3*), which were caused by overestimated airflow rates due to humid air (described in Section 8.3.5) in combination with the difficulty to manually set the valve position to result in a desired maximum airflow rate. During one airMaint procedure, the airflow rate increased (slot 7, *Figure 8.1*), which made it difficult to predict the maximum air flow rate in the end of slot 7. Ideally, all observations in *Figure 8.3* should have been aligned at the desired maximum airflow rate during airMaint on the X-axis.

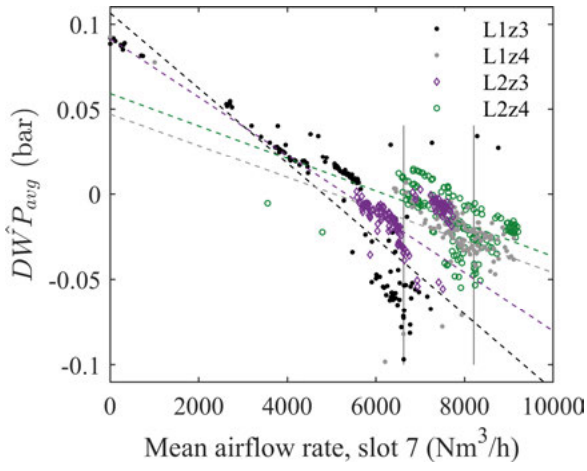


Figure 8.3. Negative correlation between mean airflow rate during airMaint (slot 7) and DWP_{avg}. Data with airflow rate below 6,900 Nm^3/h were used to interpolate L2z3. The two grey vertical lines indicate the desired airMaint of 6,600 Nm^3/h (L1z3,L2z3) and 8,200 Nm^3/h (L1z4, L2z4).

There was also a periodic variation in $DWP_{cleaneff}$ in Line 2 (noisy blue dashed line in *Figure 8.2*). This was caused by unevenly distributed (in time) cleaning actions. Since the RF was executed every Tuesday and Thursday, its effect and, similarly, the DWP_{clean} were larger on Tuesdays, when there had been a longer interval since the last cleaning. For the short interval, the $DWP_{cleaneff}$ was close to zero or showed even negative values. By contrast to Line 2, there was no large periodic variation in DWP_{clean} in Line 1. The $DWP_{cleaneff}$ was instead constantly large at 0.05 bar, without the periodic decrease. The reason for this difference is not fully understood, but we hypothesize that the RF cleaning had a larger effect on older diffusers (Line 1) with more rigid EPDM, compared with newer and more flexible diffusers (Line 2).

8.3.1.2 Impact from manual cleaning of the airflow rate sensors

There was a sudden decrease in DWP_{avg} at the end of March 2019 (vertical grey dashed line, *Figure 8.2*) in all positions. This was caused by manually cleaning the airflow rate sensor from dirt, resulting in an increase in airflow rate (1,300 Nm^3/h to 1,900 Nm^3/h at 20% valve opening) for L2z4, which had the dirtiest sensor. The other positions showed a smaller increase or decrease of 100–200 Nm^3/h. Operational staff indicated that the airflow rate sensors were not routinely cleaned, and had not been cleaned for at least 10 years. The impact from biased airflow rates on DWP_{avg} is evident in (8.3), which demonstrates that accurate airflow rate measurements are essential for the suggested estimation approach.

8.3.1.3 Newer diffusers showed larger increase in dynamic wet pressure

Considering only data from the cleaned airflow rate sensors, there was a larger trend for increasing DWP_{avg} for Line 2 compared to L1z4 (disregarding L1z3 due to the insufficient cleaning action) (*Figure 8.2*). The DWP_{avg} was also on average larger in Line 1 than in Line 2. This agrees with our expectations that newer diffusers in Line 2 would show a clearer aging effect and increase in DWP, compared to Line 1.

The results however differed from our expectations regarding the absolute values of DWP_{avg}. All positions showed negative DWP_{avg} on the order of -0.05 bar (*Figure 8.2*). This indicates a general bias, either in the model assumptions (8.3-8.4), or in a joint measurement (air temperature and air distribution pressure). The offset could also have been caused by bias in the individual measurements at the four positions (airflow rate and valve position), although this is less likely because it assumes a similar bias magnitude at all positions. Ultimately, we cannot identify the root cause for the offset in DWP_{avg} since one faulty measurement alone could not have caused such large bias. Nevertheless, the DWP estimates provided useful information about the diffuser's progressive condition and lack of sufficient RF cleaning. Future studies and practical applications should address the accuracy of the absolute DWP value when new (unfouled) diffusers are installed.

8.3.2 Monitoring changes in αF

The trends for αF partly differed from those for DWP_{avg} (compare *Figure 8.2* with *Figure 8.4*). Notably, there was no change in αF for L1z3 due to decreased RF cleaning as was seen for DWP_{avg}. This observation is in line with (Odize et al. 2017), in which an increased DWP was not linked to a decrease in aeration efficiency. The sudden change in airflow rate due to manual cleaning was not as clear in αF at L2z4, as for DWP in *Figure 8.2*.

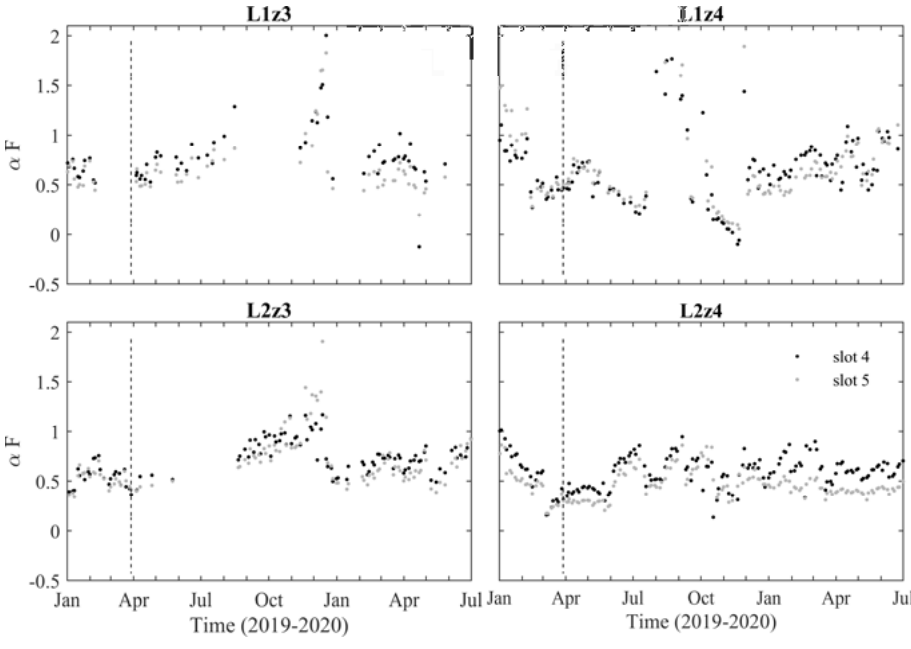


Figure 8.4. Estimated αF values without disturbances for all positions. Slot 5 had a higher airflow rate than slot 4. Dashed lines indicate cleaning the airflow rate sensor.

Considering only data after cleaning of the airflow rate sensors, there is a small trend towards *increasing* αF at all positions apart from L1z3. This contrasts with our expectations, as we assumed there would be an increase in fouling and thus a *decrease* in aeration efficiency over the experimental period. Such decrease was observed in (Garrido-Baserba et al. 2017) and the contradicting results here could either be that the suggested method was incapable of assessing changes in aeration efficiency, or that there the decrease in aeration efficiency was negligible. Furthermore, we assumed that the newer diffusers (Line 2) would show a larger (better) αF than Line 1, which was not observed. The αF was also above 1.0 (its theoretical maximum) for short periods. Since the current WRRF has a medium loading rate leading to extensive sensor fouling (Samuelsson et al. 2018), we suggest that the unexpected results indicate model or measurement errors, rather than negligible diffusor fouling. Note that any error in the estimated underlying parameters ($\hat{r}$, $\widehat{k_L a}, k_L a_{clean}$) would be reflected in the αF.

It is likely that also large variations in the α factor — due to influent composition variations —mask changes in F, which are expected to lie in the interval; αF = 0.11–0.79 (Metcalf and Eddy 2004). It is known that variations in α are large, and occasionally periodic. We still expected a slow trend in $\widehat{\alpha F}$ due to diffusor fouling, that would have been clear despite periodic variations as indicated in L2z4 (*Figure 8.4*). Therefore, further studies are needed to assess whether changes in the fouling factor F can be distinguished from influent

variations with the proposed method. Such assessment would require extensive measurements including off-gas measurements, respirometry measurements, laboratory analyses and clean water tests of fouled diffusers.

8.3.3 Estimating respiration rate

$\hat{r}$ is a critical parameter since any error in the estimate will be transferred to $\widehat{k_L a}$, and in turn to αF. At all positions, there was a tendency for a higher respiration rate at the end of the staircase (slot 10) compared with the initial estimate (slot 3), see Table 8.1. We hypothesize that this increase in respiration rate was due to the intense aeration during airMaint. This hypothesis is supported by the observation that the difference was larger for zone 3 than for zone 4. Zone 3 was unaerated for long periods and thus would probably be more affected by sudden aeration.

Table 8.1. *Mean (±1 std) estimated respiration rates (g O_2/m3,h) during slot 3 ($\hat{r}_3$) and slot 10 ($\hat{r}_{10}$) during the experimental period (January 2019—May 2020).*

	$\hat{r}_3$	$\hat{r}_{10}$
L1z3	98 (± 66)	124 (± 37)
L1z4	67 (± 46)	94 (± 44)
L2z3	86 (± 25)	104 (± 32)
L2z4	57 (± 22)	64 (± 20)

$\hat{r}$ is also informative about the load distribution. Table 8.1 shows that, on average, the estimated respiration rate was higher in Line 1 than in Line 2, indicating an uneven load distribution. Further, the respiration rate was higher for zone 3 than for zone 4, which was expected due to the initial reduction in biodegradable substances.

To perform a rough check on the reliability of the estimated respiration rates, the estimated *specific respiration rates* (obtained by normalizing to the assumed volatile suspended solids (VSS) content (75% of TSS)) were compared with literature values (*Figure 8.5*). The estimated values were within or above the suggested normal range at all positions except L2z4 (*Figure 8.5*). It is unclear why the respiration rate was so low at L2z4, especially as the respiration rate with the same sludge was within the normal range in the preceding zone, L2z3.

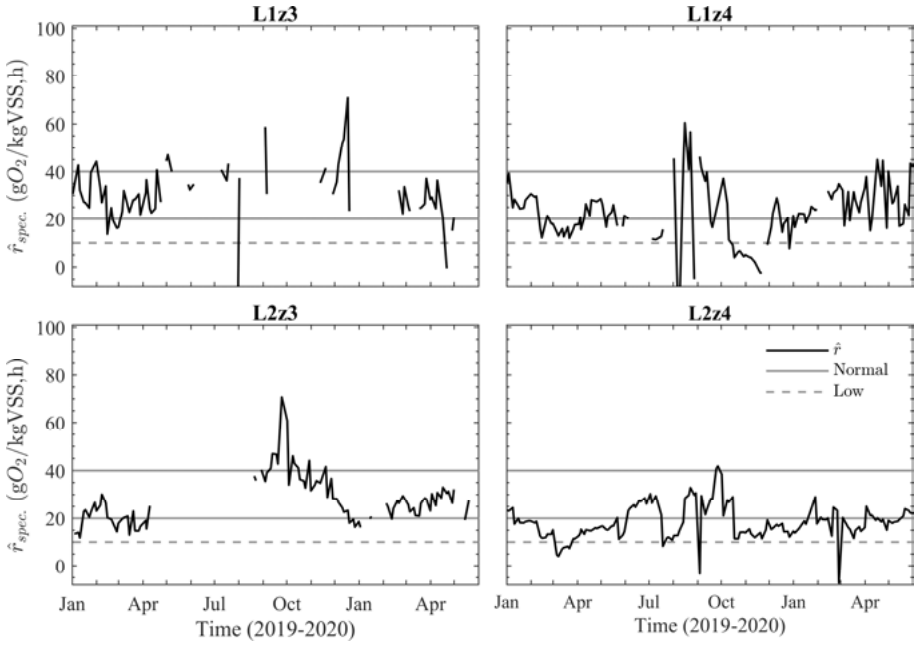


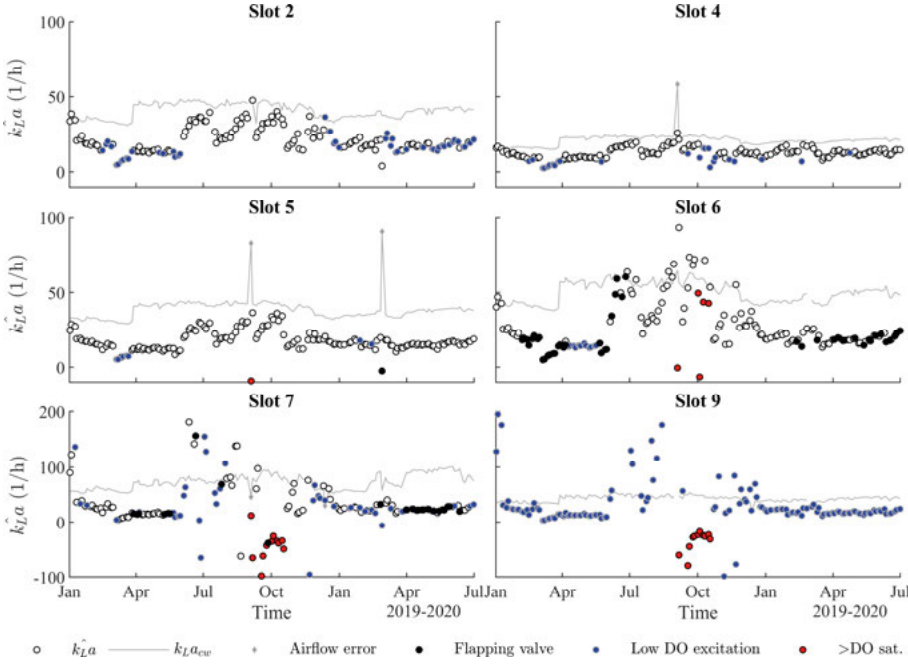
Figure 8.5. Specific respiration rates with bounds for the interval of normal (grey solid line) and low sludge activity (red dashed line) as indicated in (Henze et al. 2002). The VSS content was assumed to be 0.75 of the measured TSS.

8.3.4 Estimating $k_L a$

Line 2 showed the most stable $\widehat{k_L a}$ and close to the expected values, especially in slot 4 and 5 (*Figure 8.6(a)*) when compared with other positions (see Supplementary Materials S8.1 for a complete set of figures). There was a temporarily increase and periodic variation in $\widehat{k_L a}$ for the period July to October (empty circles, *Figure 8.6(a)*), which was not predicted from the theoretical clean water value $k_L a_{clean}$ (grey line, *Figure 8.6(a)*). This sudden increase was seen at all positions (Supplementary Materials S8.1), indicating that it reflects a change in the water composition (either influent or activated sludge), rather than in the diffusers.

A step increase in the theoretical $k_L a_{clean}$ is seen in the beginning of April in all slots (*Figure 8.6(a)*). This was caused by manually cleaning the airflow rate sensor, which also resulted in an increase in airflow rate, as mentioned in Section 8.3.1. Note that in general, the predicted variations in $k_L a_{clean}$ are small compared to $\widehat{k_L a}$. Also, the impact from $k_L a_{clean}$ (and cleaning airflow rate sensors) have a much smaller impact on $\widehat{\alpha F}$, than the variations in $\widehat{k_L a}$ (compare the dip in $\widehat{\alpha F}$ in early March, with the manual cleaning in end of March for L1z4 and L2z4, *Figure 8.4*). This supports the indication that influent variations causing variations in α may be too large, and mask the smaller

variations in F. This difficulty that can severely limit the usage of $\widehat{k_La}$ to represent OTE of diffusors has not been demonstrated in previous studies (Holmberg and Olsson 1985, Irizar et al. 2009, Lindberg 1997) because they only evaluated their approach for a short time (days).



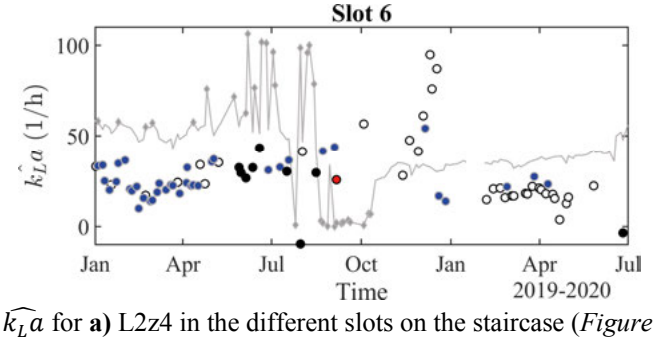


Figure 8.6. $\widehat{k_La}$ for **a)** L2z4 in the different slots on the staircase (*Figure 8.1*). Note that slot 7 and 9 have a different Y-axis scale compared to other slots. **b)** L1z3 with slot 6.

In addition to the variations, potentially caused by the influent, all positions and slots were also affected by disturbances (filled markers, *Figure 8.6(a)*). The disturbances, however, could be explained and are further analysed in the following sections, including:

- a flapping valve (black dots)
- erroneous airflow rate (grey diamonds)
- insufficient DO excitation (blue dots)
- biased DO sensor measurements (red dots)

8.3.5 Flapping valve error

A large variation, or flapping, in both airflow rate and valve position was observed for several measurements, as shown in *Figure 8.7* and by the black dots in *Figure 8.6*. The root cause for this disturbance was that the airflow rate sensor was biased by humid air, which made the measured airflow rate exceeding its maximum value. This in turn triggered the control system to rapidly close the valve. Once closed, the valve was opened again, which caused an excessive airflow rate that was repeated until the airflow sensor was either dry or the staircase entered a slot with a lower airflow rate. The impact from humid air was concluded with a separate experiment where the flapping valve error disappeared when water condensate was removed from the aeration system.

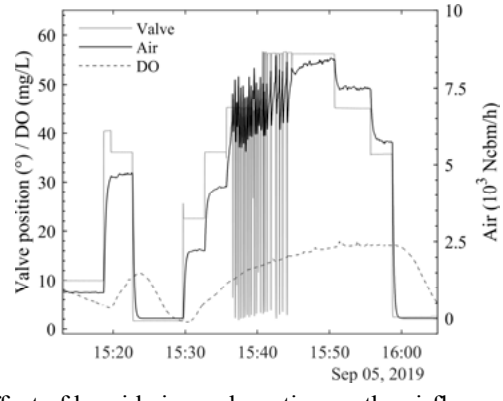


Figure 8.7. The effect of humid air condensating on the airflow sensor. The effect of the flapping valve causing fast variation in airflow rate is evident.

The phenomenon of humid air condensation in the air distribution system is a known problem, which is managed by manually emptying water condensate (ranging up to a few hundred litres per zone and week during early autumn in the current plant). Water and other substances impinging on a thermal airflow mass sensor are also known to affect the measured output. However, to our knowledge, it is not widely appreciated that the accuracy of an airflow rate sensor can be affected by the outdoor weather (the current WRRF is situated underground). Unexpectedly, the problem with condensing water vapour was seen throughout the year, and not only during autumn (slot 6 *Figure 8.6*).

The flapping valve issue was mainly seen in slot 2, probably when the air distribution system contained the largest condensate water volume, and in slot 6 when the airflow rate was high and close to the maximum limit (see Supplementary Materials S8.1 for plots equivalent to *Figure 8.6*A for all positions). Further, there were more flapping valve indications in zone 3 compared with zone 4 (Figure 1 in Supplementary Materials S8.1). We suggest that the configuration of the air distribution system and the intermittent aeration of zone 3 were the main factors contributing to this difference. The magnitude of the bias in the airflow rate sensor due to humid air and whether it has practical implications during normal operations is a subject for future studies.

Surprisingly, the flapping valve was unsuccessful in suppressing DWP despite the resulting frequent changes in airflow rate (compare the black dots around July in *Figure 8.6(b)* with DWP_{avg} in L1z3 in *Figure 8.2*. This indicates that the maximum airflow rate was more important than the variation for an effective RF cleaning. This suggestion is reinforced by the negative correlation previously shown in *Figure 8.3*.

8.3.6 Erroneous airflow rate

During the period from June to October, the $k_L a_{clean}$ in L1z3 exhibited a series of peaks followed by a dip (grey diamonds, *Figure 8.6*(b)). Similar events were seen for several slots in zones 3 and 4, Line 1 (Supplementary Materials S8.1). These were caused by a seemingly erroneous airflow rates that had a direct impact on $k_L a_{clean}$. We identified two different reasons for the deviating airflow rates.

First, the peaks in *Figure 8.6(b)* were caused by a confusion in the slot identification algorithm that was induced by the flapping valve. Due to the flapping valve, the variations in valve position were large, which made it difficult to identify when a new slot began. As mentioned in Section 8.2.7, data for the different slots were extracted based on a change detection algorithm that identified the different slots from valve position data. As a result, the airflow rates from wrong slots were used, which showed as 'erroneous' airflows.

Second, the period with low or zero $k_L a_{clean}$ in L1z3 was caused by a low airflow rate. In the staircases, it was clear that the airflow rate was zero at all valve positions, indicating either a frozen or broken airflow rate sensor or a closed valve with faulty position indication. The DO measurement oscillated between zero (or very low concentration) and normal values. This somewhat mysterious behaviour was explained by a play in the airflow rate sensor, which was solved in November by replacing the sensor, resulting in normal values for $k_L a_{clean}$.

8.3.7 Insufficient DO excitation

Slots 7–9 showed insufficient DO excitation (blue dots, *Figure 8.6(a)*). This was found to be caused by the intensive aeration during airMaint. The airMaint resulted in a DO concentration close to, or even above the maximum saturated value, which limited the available increase in DO concentration in subsequent slots. This made it difficult to estimate $k_L a$ accurately as the model (8.2) may provide a poor estimate when $\frac{DO(k+1)-DO(k)}{Ts}$ is small. The other positions showed similar results, with insufficient DO excitation as in L2z4 (Supplementary Materials S8.1).

A related problem when the DO is close to its saturated value, is that the oxygen is transferred to the off-gas instead of to the water. This is not captured in (8.1) and introduces a bias in $\widehat{k_L a}$. From *Figure 8.6(a)*, it is not clear whether the high DO concentration caused a sufficiently large bias to be problematic, as there is no clear difference between normal (empty circles) and potentially biased estimates (blue circles), for example, in slot 2–5.

8.3.8 Biased DO measurements

Occasionally, the measured DO was even above its theoretical saturated concentration. This result in a negative difference $DO^* - DO(t)$ in (8.1), which in turn produce negative $\widehat{k_L a}$ (slot 7 and 9 in *Figure 8.6(a)*).

DO measurements above the theoretically saturated value (more than 1 mg/L above) were flagged as biased measurements (red dots, *Figure 8.6(a)*). The biased measurements were mainly found in L1z4 (39% in slot 7, see Supplementary materials S8.1), which indicates that the DO sensor in L1z4 produced inaccurate measurements. For comparison, the L2z4 indicated 8% biased measurements for the same slot (Figure 7*(a)*, slot 7). The effect of the biased DO sensor on the estimated $\widehat{k_L a}$ values was unclear because we lack information about the bias type and magnitude.

Similarly, as for the erroneous airflow rates, we note that the staircase method can facilitate detection of data quality issues, here in terms of a biased DO sensor. The biased DO sensors were not noticed in the WRRFs normal routines, which shows the potential for the staircase method to also indicate data quality issues.

8.3.9 Improving the staircase

The many disturbances with a negative impact on $\widehat{k_L a}$ are expected to be mitigated if the following improvements of the staircase are made.

- The airMaint should be positioned at the beginning of the sequence to remove as much condensate water as possible and reduce the risk of a flapping valve.

- During airMaint, the valve should be controlled to an airflow rate setpoint instead of a predefined valve position. This would also avoid the flapping valve error during airMaint.
- A sufficient DO excitation could be obtained if the DO was allowed to settle to about 4 mg/L between each aeration slot. For example, the second slot with zero airflow rate (slot 10) should be placed after the airMaint to lower the DO concentration sufficiently.

8.3.10 Recommendations for replicating the implementation

In a future implementation, the results would be less uncertain if the observed challenges in this study were considered. We recommend:

- Measuring variations in temperature and humidity both before and after the blowers (close to the valve) to estimate the impact of condensate water. An even better solution would be to automatically remove the condensate water early in the air distribution system.
- Verifying that the airflow rate sensors are clean and provide accurate measurements, as this is not commonly part of sensor maintenance procedure.
- After calibrating and validating the DO sensors, comparing the measured maximum DO saturation concentration in buckets with clean and process water. This will be an additional validation of the measured accuracy at the (abnormally) high DO concentrations used for estimating $k_L a$.

The method should be further validated by comparing estimated values with respirometry and off-gas measurements to assess its effectiveness for estimating r, α and F. The DWP should be estimated when new diffusers are installed to identify the DWP off-set for clean diffusers. The accuracy in estimated DWP should be assess with additional pressure measurements, before and after the diffuser.

8.3.11 Extending the method to additional applications

The staircase method was inspired by the active fault detection concept (Puncochár and Skach 2018). In line with that concept, the staircase could be used to obtain additional information about the process and sensors.

Estimating respiration rate has many potential applications as exemplified in (Olsson 2012), and the suggested method would be applicable for in situ respirometry at several (all) zones and lines without the need for added instrumentation.

As indicated previously, the maximum DO concentration is reached (or can be assumed to be reached) during airMaint. This value could be used to assess

the deviation in DO measurements to quantify bias direction and magnitude in the DO sensor.

Finally, the large change in DO during airMaint can be considered to be a controlled disturbance that will impact the DO in the subsequent zones. By monitoring how quickly the subsequent DO controllers suppress the disturbance, a measure of their disturbance-rejection ability could be obtained using for example the method by Petersson et al. (2002).

8.4 Conclusions

Diffuser condition and the reverse flex cleaning (RF) effect can be monitored using only existing instrumentation in combination with the proposed staircase method. An increase in dynamic wet pressure (DWP) due to insufficient RF cleaning was detected as changes in DWP_{avg}, which is suggested as an early warning indicator for diffuser fouling. By contrast, the estimated αF showed no trends of increasing diffusor fouling, but mainly variations that were suspected from unavoidable variations in $\widehat{k_L a}$ and the influent composition. This crucial challenge for using $\widehat{k_L a}$ for diffusor condition monitoring has not been emphasized before and needs to be considered in future studies.

- This study exemplifies that new insights can be enabled by a simple process disturbance that in turn change the information content reflected in data.
- The staircase method is promising as it is simple and available with minor costs.
- The staircase method, and possibly grey-box modelling in general, can also be valuable for data quality assessment. Disturbances causing bias in airflow rates and DO sensor measurements were indicated, which would have been undetected without analysing the data produced by the staircase method.

Ultimately, these findings motivate further research on how modelling in combination with repetitive process disturbances can provide useful information without needing additional instrumentation.

Appendix A8.1 Derivation of SOTE and $k_L a$ during clean water conditions

The SOTE is given by

$$SOTE_{stp,20} = k_L a_{clean}\,(20, q_{test}) \frac{DO_{\infty,20}{}^{*}\, V_{test}}{W_{O_2,test}} \tag{8.7}$$

In (8.7), $W_{O_2,test}$ is the oxygen mass flow, V_{test} is the tank volume and $DO_{\infty,20}{}^{*}$ is the average saturated DO obtained during SOTE clean water tests. Note that $DO_{\infty,20}{}^{*}$ is usually higher than the surface DO saturation in clean water, $DO_{20}{}^{*}$, due to the liquid column pressure of water above the diffuser. $DO_{20}{}^{*}$, but not $DO_{\infty,20}{}^{*}$, was available from the supplier's clean water test data; therefore, $DO_{\infty,20}{}^{*}$ was estimated by assuming that the effective saturation depth, d_e, was 50% of the basin depth (Metcalf and Eddy 2004); d_e is defined as

> "the depth of water under which the total pressure (hydrostatic + atmospheric) would produce a saturation concentration equal to $DO_{\infty,20}{}^{*}$…" (EPA 1989b).

The subsequent conversion is straightforward: $DO_{20}{}^{*} = \frac{DO_{\infty,20}{}^{*}}{\delta}$, where $\delta = \frac{(p_{atm}+d_e)}{p_{atm}}$ and p_{atm} is the atmospheric pressure. By further compensating for temperature T with the common correction factor θ=1.024 (Metcalf and Eddy 2004, Rosso 2018), the $k_L a_{clean}$ is obtained as

$$k_L a_{clean}\,(T, q_{air}) = \frac{\theta^{(T-20)} SOTE_{stp,T}(q_{air}) W_{O_2,process}}{V_{process}\, \delta\, DO_{20}{}^{*}} \tag{8.8}$$

In (8.8), the $SOTE_{stp}(q_{air})$ was obtained by interpolating SOTE test measurements at five airflow rates, which were provided by the manufacturer.

8.5 Chapter acknowledgments

We gratefully acknowledge the efforts made by staff at Bromma WRRF in implementing the staircase method in the control system and for data acquisition: Christer Persson and Anders Påhlsson; and on-site experience and metadata provision by Hanna Gottås, Ove Sigfridsson and Lena Jonsson. We also thank Gustaf Olsson (Lund University), Jesús Zambrano (The Mathworks) and Lena Jonsson for providing valuable comments on the manuscript, Magnus Arnell (Lund University) for discussions related to the modelling of SOTE, and Rico Schuhmann (Jaeger) for providing data and discussions about the manufacturer clean water SOTE tests.

9 Summarizing discussion

In this chapter I discuss general challenges and implications from the results in Chapter 2-8. The major part of this work concerns fault detection as such, and how this relates to data quality is discussed in Section 9.1. Next, the role of annotated data for fault detection is discussed (Section 9.2-9.5) from the perspectives of training, tuning (Section 9.3) and performance assessment (Section 9.5). The discussion also includes how data can get a higher information content with deliberate perturbations (Section 9.4). Last, WRRF data characteristic challenges that needs to be tackled by successful methods (Section 9.6) and general challenges for bringing FDMs into practice (Section 9.7) are discussed.

9.1 Improved data quality and fault detection

The long-term goal of this research project was to improve the data quality in WRRFs. Through the studies here, it has become progressively clearer that this needs to be considered in a wide context. Data quality should not be limited to the bias and variance of a sensor signal, which was the implicit original assumption. Instead, it also needs to consider how well described the data are, e.g. if data are annotated for normal or faulty time periods.

The lack of both annotated data and automated quality assessment methods limited the development of FDMs. The hypothesis for this thesis was that the sensor data can be improved if faults can be quickly identified with FDMs, and subsequently corrected (Chapter 1). This hypothesis, however, contains a Catch-22 (Heller 1961) when the existing methods for assessment need to be improved, in order to develop the actual FDMs. For this reason, the basic tasks related to sensor maintenance (cleaning and calibration) were identified as key activities for improving the data quality, and in turn, the applicability of data driven FDMs (Chapter 2-8).

The studies in this thesis further emphasize the need for automated data quality assessment. Again, this should be considered in a broad context including accuracy, precision and annotations. This time-consuming task was conducted manually in this thesis and will need to be facilitated for a widespread adoption.

There was a surprising knowledge gap about data quality in practice concerning the basic insights about common sensor faults. Sensor drift direction

in DO sensors due to fouling was studied in Chapter 3, where the effect from fouling differed between the two studied measurement techniques. Despite that such a study is straightforward to conduct, to the best of my knowledge, neither in research, nor in practice are such data commonly recorded. The information obtained is highly valuable for developing FDMs, but also for basic understanding about how this impact the treatment process. Chapter 2 showed that the sensor drift direction has a direct impact on how the treatment process is affected from the bias. However, the expected error occurrence (e.g. due to drift) is unknown, which was a draw-back for tuning and applying mass balance-based data reconciliation in Chapter 4. Access to such information was identified as a missing piece of information for FDMs to come to fruition in practice.

9.2 Challenges with annotating data

The success or failure with the experimental studies in this thesis was reliant on access to accurate data annotations although this was not explicitly emphasized in all studies. Further, the availability of data from the normal and the faulty mode guides how FDMs are selected, trained, tuned, and assessed. This was illustrated in the different studies in Chapter 3-8 and is further discussed in the following three sections.

An annotated dataset was produced in Chapter 3 representing clean and fouled DO-sensors. Many unsuccessful (unpublished), efforts were made to artificially imitate biofilm formation on DO-sensors, which indicates the challenge to represent faulty modes in sensors in a realistic way. In Chapter 3, long-term experiments in a real process finally produced such faulty data. Note, however, that the value in final dataset was reliant on the accurate annotations and bias quantification. Without time for manual sensor cleaning, sensor replacement, reliable reference sensor measurements and points in time for known disturbances, the dataset would not have been useful for the subsequent study in Chapter 7.

Also, in the other experimental studies it was the access to both normal and faulty data that allowed us to interpret the methods performances. The access to data containing known sensor faults such as the drifting ammonium sensor in Chapter 5 was essential to demonstrate the usefulness of the Gaussian process regression as FDM.

The most evident example of the need for data annotations was the abundance of a priori unknown disturbances in the studies. For example, worn-out sensor membranes (Chapter 3), erroneous airflow rate readings (Chapter 8), a broken off-gas pump (Chapter 4), and the syphon effect (Chapter 4) had a great impact on data. If these disturbances had remained unexplained (after the experiments), the results would have been uncertain and difficult to interpret.

There are technical challenges to store and retrieve desired annotated data. Proper storage and precise experimental notes (here regarded as data annotations) have been a key actions for successful research as long as scientific experiments have been conducted. This is still an essential step in research. Databases, digital cameras and digital logbooks are tools that have simplified the procedures for documenting experiments. At the same time, the complexity in terms of available data sources has increased, which poses new challenges for storing and retrieving annotated data. Related to data annotations in the experiments here, the two main time-consuming tasks were:

1. Analysing the cause for anomalies in data produced by process changes, sensor maintenance actions, or sensor disturbances
2. Merging different data sources into a joint dataset

First, basic process changes conducted by operators, such as shutting down a pump or changing the aeration, were only partly recorded and hence difficult to analyse a posteriori. Such changes influence the data, partly in unknown ways. In the large WRRFs considered here, many different people were involved in the operations and it was not always possible to obtain an explanation for a historic process change. This was especially a challenge when history refers to weeks or even months back in time. Unfortunately, it was months after the executed experiment (or even a year) as details and data anomalies became clear and needed and an explanation. In Chapter 8, it was noticed that the airflow rate sensor had been replaced when an explanation for the deviating $k_L a$ estimates was searched for. In Chapter 4 the suggested syphon phenomena with nitrate recirculation was only realized while finishing the manuscript during the final interpretation. Current data management systems have not been implemented to support this kind of information, which is in practice instead exchanged orally on a daily basis, and at the best, recorded in a logbook or meeting notes.

Regarding sensor maintenance, it was time-consuming to manually merge data about sensor maintenance that were commonly located in spreadsheets separately from the time series database. Data describing the adjustments during calibration of the DO-sensors in Chapter 3 were stored locally in the sensor signal processing unit, and hence needed to be manually transferred to a logbook at each calibration occasion. Such manual actions are prone to errors and automated procedures would be desirable.

From a research perspective it is clearly a problem with ambiguous annotations. It becomes especially difficult to distinguish effects in data from influent variations, process changes, and sensor faults to draw correct conclusions. The need for formalizing such historic process operator knowledge into a knowledge-based system - essentially putting words on the operator's experience and actions - has been realized since long (Olsson et al. 1989) and is still an active area of research (Ye et al. 2020). Today, systems are lacking

that automatically can (or at least facilitate manual efforts to) annotate data, by integrating and combining existing data sources. Such tailored IT-systems would improve the interpretability of existing data and be a step towards developing knowledge-based systems. I expect that such systems would be valuable for the practitioner as well as the researcher.

9.3 Training and tuning FDMs with annotated data

As mentioned in Chapter 1, the access to annotated training data is in general so central that machine learning algorithms originally were grouped into the two groups of unsupervised (no annotations) and supervised methods (with annotations). In general, the training procedure consists of several sequential data pre-treatment steps that generate features (characteristics in data), which have a large impact on the final classification performance. This was verified in Chapter 7 with the data transformation model in (Venkatasubramanian et al. 2003). In Chapter 7, the training (feature space) was separated from tuning (decision space), which illustrated that even excellent features that produce good normal and faulty mode separation as evaluated by receiver operating characteristics, can fail when automated tuning is required. The problem was that the training dataset (12 impulse responses) was too small to represent the normal variations, which resulted in poor tuning.

The grey-box model in Chapter 8 was based on process theory, and not only on patterns in data, which reduced the need for training. Still, the compensation of the constant off-set in dynamic wet pressure estimates can be said to be similar to training. If the model predictions would have been extended to a FDM, this bias would have been needed to compensate for, for example when tuning the alarm limit. The benefit with the grey-box model over completely data driven methods is that such bias can be physically interpreted and possibly corrected as its root cause is easier to identify. The interpretability of mechanistically rooted models is generally suggested as one main benefit.

The data reconciliation in Chapter 4 was tuned in terms of specifying the error covariance matrix from theoretical assumptions and practical experience related to sampling error. This was done before accessing any training data and is a different approach from the data driven FDMs. The theoretical error assumptions had a large impact on both which gross errors that were detected and on which reconciled values that were obtained. This made the tuning critical and demonstrated that more research is needed about how the data reconciliation should best be tuned.

The lack of a dataset with strictly unbiased sensor measurements in Chapter 4 can be considered as a lack of normal training data. This was a common challenge in all full-scale studies. As indicated in Chapter 7, small training data subsets obtained at the normal conditions, can limit the FDM training.

This is a general challenge to use small datasets for representing reality as was demonstrated in *Figure 1.5*.

With regards to tuning, I suggest that it is the tails of the underlying normal (fault free) distribution that are most relevant to obtain data from. These are obviously difficult to obtain since they are likely close to or overlaps with the faulty mode distribution. Thus, they will have a large impact on false alarm and true detection rates.

The challenge to estimate the normal mode distribution increases when the process is time-varying (non-stationary), which then shifts the distribution location. Re-training and re-tuning is then needed, which further increases the needed amount of training data. It can be possible to compensate for a non-stationary behaviour if this can be modelled. A basic strategy for dealing with non-stationary nature of the WRRF is to correct for temperature changes in the influent (which is one of the main causes for the time-varying behaviour). This was used in both Chapter 4 and Chapter 8. The results in Chapter 4 indicated that the general temperature correction (Henze et al. 2002) needs to be adapted for current WRRF conditions. In Chapter 5, only dry weather was considered, which is one option to deal with non-stationarity. This has the obvious draw-back that only part of the plant's condition can be monitored unless multiple models are used with different tuning of a FDM. In theory this is feasible, but would require extensive model training and validation, in combination with detection of when the different models are applicable. Such approach is similar to gain-scheduling in PID controllers and has been suggested for FDMs based on principal component analysis and partial least squares models (Haimi 2016, Rosén 2001).

Although there are challenges to obtain sufficient and representative normal training data, it can sometimes be straightforward, for example, for sensor fault detection. In Chapter 7, normal data were defined as the 24-hour period after manual sensor cleaning, inspection and validation. It further simplifies that many identical sensors are used in either parallel treatment lines or at different zones. In Chapter 3, up to 30 identical DO sensors were used in the current WRRF. If data from such 24-hour normal periods would be combined, we expect that the ambiguities possibly related to seasonal trends or sensor wear would be possible to quantify. Even better would be to share such normal data openly between different WRRFs to further gain experience in the spread of the normal sensor behaviour.

All methods considered in this thesis were only trained on data from the normal mode. Such (supervised) one-class classification methods require an explicit definition of what is normal. The task to make this, often subjective definition explicit, has a positive side effect that it forces the researcher or process operator to critically assess what deviation, e.g. bias magnitude, that can be allowed. Any ambiguity and subjective discrepancies will then be brought out into the open for discussion. An additional benefit with only training and tuning on the normal mode is that any deviation, regardless of type of

fault, can be detected. If instead a binary supervised classification method is applied, the training and tuning will target a certain fault, which can make the method insensitive for other faults. A challenge with one-class classification is the tuning and to set an alarm threshold. Probabilistic FDMs were used in Chapter 5-7, which provide a natural tuning, while still only trained on normal data. Setting the alarm limit, for example, for observations outside a two or three standard deviations limit under a Gaussian assumption is in line with traditional Statistical process control methods (Kourti and MacGregor 1995, Shewhart 1930) and was used in Chapter 5-7.

The major draw-back of only having access to training data from the normal mode is that it becomes difficult to know which features that are the most sensitive for the faulty mode. All data transformation steps suffer from this. It is clearly easier to detect a certain anomaly if its characteristics are known. The masking influent and process variations makes this even more important. The reason for not considering faulty training data, and supervised methods in the training and tuning, is that data from the faulty mode are even more challenging to obtain than from the normal mode (Pimentel et al. 2014). Different sensor faults and disturbances have, however, been frequent in the studies in Chapter 3, and chapters 4-8. The challenge with these faulty data is that their start time is unknown and not annotated. Therefore, it can still be problematic to obtain enough faulty training data of sensor faults. On the other hand, supervised (two-class) classification methods could become feasible and facilitate sensor fault detection, if sensor faults could be better understood by collection of datasets containing such sensor faults.

9.4 Deliberate perturbations to increase information in data

Chapter 3, and chapters 7-8 studied how deliberate system perturbations could produce data that are more informative about when the sensor or process changes from normal to faulty state. In Chapter 3 and 7, these more informative data were obtained from the sensor membrane dynamics and related change in its response time. In Chapter 8, the information content was instead improved by fixing the system at various constant levels (valve positions). Then, the desired model parameters were estimated from the DO which was excited using steps in the airflow rate. These studies confirm that deliberate excitations in the excitation space impact the measurement space and facilitate the subsequent data transformation steps. This has been well studied for model identification (Ljung 1999) and also fault detection and diagnosis (Puncochár and Skach 2018).

One key challenge in fault detection in WRRFs is solved when process variations can be separated from sensor errors. Methods such as active fault

detection that increase the signal to process noise relation is clearly one way to approach this challenge. In addition, the impulse response (Chapter 3,7) and staircase (Chapter 8) generates patterns with new patterns (features) that differs from the conventional time-series. How these patterns can be analysed with machine learning such as Gaussian process regression was explored in Chapter 7. The results favoured the simple rise time method over the more advanced Gaussian process regression, but the intense research activities in artificial intelligence is expected to improve the knowledge in pattern recognition (Ye et al. 2020), which is suited for interpreting data that are produced by repeated and identical disturbances.

The findings in Chapter 3 further showed that different patterns matched different fault types. It was possible to visualize sensor wear in terms of progressive changes in data patterns, and not only from a visual sensor inspection. This approach contrasts with how time series have historically been used for example to train and use data driven methods such as principal component analysis and partial least squares for process monitoring in WRRFs (Haimi 2016), but is instead in-line with a conventional classification workflow. Thus, deliberate disturbances for definite (short) time periods are promising for supervised classification methods, especially if fault patterns can be produced.

9.5 Performance assessment with annotated data

Access to annotated data is essential for FDM performance evaluation, and even more important than for the training and tuning procedures. Performance assessment becomes difficult if it is not possible to interpret whether an alarm is true or false.

Performance metrics derived from the confusion matrix (Chapter 1, Figure 1.6) are central, including receiver operator characteristics and precision-recall. It should be noted, however, that the relevance of these measures is reliant on how well the data represent the underlying probability distributions, and especially the tails of the distributions. In fact, I suggest that from a performance assessment perspective, the best data are those that lie in the borderline between normal and the faulty mode. These will be the most difficult for the method to separate and will effectively reveal its performance. This can be realized from *Figure 1.5* where the overlapping region between the two distributions have a major influence on the alarm threshold, and thereby also the performance. A good estimate of this part of the probability distributions will reveal the performance at low false alarm rates, which is essential.

A basic benchmark FDM can be used to indicate whether the data used for performance assessment pose a too simple problem for evaluating the methods at hand. A simple method that produce similar performance as advanced ones indicates that: either the detection problem is simple, or the data representing

the problem are easy to classify. When the true detection problem can sufficiently well be solved with a basic method, there is no need for more advanced methods. A basic benchmark method that is simple to understand and interpret can therefore serve as a sanity check that indicates what is gained from more advanced methods. For example, the following examples are given in this thesis: process indicators derived from primary data, in comparison to data reconciliation (Chapter 4); least squares linear regression compared to Gaussian process regression (Chapter 5); the pragmatic engineering approach compared to machine learning methods (Chapter 7). These performed well in comparison to the more advanced methods, which supports the benefit of always including such benchmarks during performance assessment.

There is commonly an imbalance in available observations for the two modes (normal and faulty). This is simply because the normal mode is more common and therefore produce more data (Pimentel et al. 2014). This is undesired for performance measures that implicitly assume that the data are estimated from the underlying probability distribution (Bishop 2006). This requirement was opposed in Chapter 7, where only part of the data was considered for computing the partial area under curve. The reasoning was that FDMs that produce false alarm rates above a certain limit, here 10 percent was chosen, will not be considered in practice. In effect, only the 10 percent in the tail of the normal data were used for performance assessment. This reasoning favours performance assessment on data that are representative for the border line between the two modes, rather than obtaining a balanced dataset in terms of number observations.

The available data and related degree of annotations decide which performance measures that are applicable. Time-based measures such as mean time to detection were not considered in the studies here because it requires knowledge of the start time for the fault. Such detailed information is feasible in simulations, but less feasible in practice. The simulation study in Chapter 2 showed that small changes in nitrogen removal, down to a few percent, easily can be identified and analysed in a simulation. This would not have been possible to identify in practice. Practical experiments are however needed to, for example, assess drift direction as in Chapter 3. In the end, a thorough performance assessment of FDMs should therefore include both simulation studies and practical experiments, to assess both theoretically and practically relevant aspects.

The reliability and generalization of FDM performance assessment is tightly linked to scientific rigour of the data producing experiments. The internal validity relates to clear and accurate annotations of faulty and normal data, and further that sensor validation procedures in practice are reliable. From a scientific methodology perspective, only the study in Chapter 3 can be regarded as a controlled experiment. The main problem for conducting experiments in WRRFs is to control all background variables, and limit interventions to only deliberate ones. In fact, in all studies there were undeliberate

disturbances or faults that were analysed a posteriori as part of the performance assessment. These disturbances were partly useful and partly time-consuming to assess. This occasional lack of internal validity suggest that the studies here should instead be regarded as natural observational studies. To compensate for this flaw, long-term studies are needed that replicate potential seasonal variations, which then can be distinguished from other variations.

As mentioned, it was a great challenge to produce annotated data (both normal and faulty) for performance assessment. This will be even more difficult in practice when time is scarce and the main goal is process operations rather than detailed performance evaluation of FDMs. For this reason, we expect that only the raised alarms will be feasible for performance assessment in practice. That is, the operators will not prioritize investigating whether normal indications (no alarms) are missed detections or not. Instead, I expect that they will base their decision about the method's usefulness on its ability to detect true faults, and amount of false alarms. Otherwise, every time instant would be needed to be assessed for its (normal or faulty) condition. Note that when only data from alarms are considered, the statistics based on the confusion matrix (Figure 1.6) become infeasible. For example, the false alarm rate cannot be obtained since this would require information about how many of the observations that did not cause an alarm and belong to the normal mode. Also, assessment of the true detection rate requires abundance of faults to be evaluated, which is undesired and therefore scarce in practice.

Instead, the tuning method proposed by Gustafsson (2000) and used in Chapter 7 could be adapted as a performance measure that is based on alarm detections, which are practically attainable. In the original setting, tuning is done by successively changing the alarm threshold until the desired false alarm rate is obtained. In an on-line setting, the occurrence of false alarms is used as proxy for the methods sensitivity to detect faults. The threshold is set at the maximum acceptable false alarms level. Then, performance assessment is conducted by assessing the ratio between the number of true detections and the number false alarms. Note the difference between the *number of* false alarms and the false alarm *rate*. Such performance measure, it can be called relative detection frequency, could be used without additional efforts on-line for an iterative tuning and performance assessment.

Dedicated validation experiments can still be used on-line, which comes with the same efforts for annotations as during the experiments here. These efforts could be motived for sensor errors that are simple to imitate as in Chapter 3, and don't have negative impact on the treatment process.

9.6 Characteristic challenges for WRRF fault detection

It is observed from the results that the WRRF, as a data producing source, exhibits some characteristics that indicate both challenges and opportunities

for fault detection. In this section I highlight these characteristics, which should be considered while developing FDMs. The described characteristics can also serve as a roadmap to identify useful methods, based on how well they can deal with the following challenges and opportunities.

Challenges

- Noise in data.
  Data from WRRFs can be perceived as noisy due to the variation mentioned in Section 1.3.2. This can reduce the precision in model predictions (Chapter 5). Variations from unpredictable influent variations and sensor disturbances needs to be considered as a general challenge for obtaining training data, tuning the method and for performance assessment.
- Missing data.
  As in most industrial processes, WRRF data are not complete and about one or a few percent of the sensor values do not reach the database. This can be caused by transferring issues or problems in the sensor data collection procedure. In addition, sensor disturbances, such as in the flow measurements in Chapter 4, can produce anomalies that are beyond doubt biased and therefore unusable. Such measurements can in practice be regarded as missing data, which makes the challenge with missing data more prevalent.
- Small annotated training datasets.
  The studies here emphasized the challenge to obtain training data, both from the normal and faulty mode. The non-stationarity of the WRRF process makes it even more complicated to obtain training data that are representative for the full range of seasonal and operational conditions (like dry and storm weather conditions). The limitation to small training datasets, in turn, makes tuning of FDMs to challenging task (Chapter 7).
- Lacking knowledge about sensor faults.
  The impact from sensor faults on data, in the viewpoint of drift, bias, miscalibration, impact from fouling, sensor wear and their frequency, is not generally understood (Chapter 3). This is a limitation for tailoring FDMs to detect faults.

Opportunities

- Slow WRRF process dynamics.
  Slow process dynamics (on the order of minutes to days) makes FDMs less time critical and allows time-consuming computations

in real-time. Such time-consuming computations could, for example, be model simulations or computationally demanding machine learning methods.

- Periodic patterns.
  The influent variations are not only a challenge, but also pose opportunities. The different periodic influent variations are reflected in the WRRF processes, which in theory provide information about, for example correlated data patterns. This information could be used to distinguish sensor disturbances from process variations. So far, and to the best of my knowledge, the literate lack examples that make explicit and efficient usage of these patterns.
- Robust treatment process.
  The common activated sludge process is robust as it can manage large influent variations. For this reason, the process can be subjected to short manual disturbances that produces tailored diagnostic information.
- Redundant processes.
  Many WRRF processes operate in parallel or serial along with the related instrumentation and measurements. This allows multiple comparisons to detect deviating sensor and/or process behaviour.
- Detailed process knowledge.
  The activated sludge process has been extensively studied for long, and have also been encoded in dynamic process models (Henze et al. 1987). This knowledge has not yet come to fruition in terms of monitoring (Newhart et al. 2019).
- Open data.
  Wastewater treatment is not a competitive business area, at least in Sweden, which allows sharing data. This can be beneficial for developing, for example, probabilistic one-class classification methods that rely on normal datasets.

Ultimately, a robust fault detection procedure could include more than one method. The detection rate can be improved, and the false alarm rate lowered by combining several methods that are independent in relation to which data they use. This can be conducted when the conditional probabilities for a fault, given different methods, are combined using a naïve Bayes model. This is inline with how several different (and independent) medical diagnostic tests are used to make a diagnose more accurate (Bishop 2006). For that reason, one may not choose between purely data-driven machine learning methods or mechanistic grey-box modelling, but instead combine them and interpret their joint output. In addition, data can be obtained from independent data producing processes when deliberate disturbances are used as in Chapter 6 and 10.

9.7 Practicability of fault detection methods

Wastewater research is applied by definition, and its purpose is to improve how we treat our wastewater. Any research output should therefore target usefulness in the water utilities, as well as increase the knowledge in the domain. From that perspective it is surprising that less than one percent (2 out of about 500) of the research articles between 1980–2017 targeting fault detection for WRRFs were evaluated with a full-scale implementation (Samuelsson 2017).

The gap between state-of-the-art methods published in literature and their practical adoption has been acknowledged since long (Olsson et al. 1989). Even today, the lack of on-line implementations in literature is reflected by the absence of FDM usage by practitioners. For example, data reconciliation methods are commercially available and have been applied in many industries (Câmara et al. 2017) and have been demonstrated to be useful also for WRRFs (Le 2019, Puig et al. 2008, Spindler 2014), but there has been no widespread adoption and only a few full-scale implementations such as (Lumley 2002). Other FDMs that have been prevalent in the WRRF literature include Principal component analysis (Rosén 2001, Rosén et al. 2003) and predictive soft-sensors (Haimi et al. 2015) as indicated by Corominas et al. (2018). These methods have neither reached practitioners and daily operations, which raises the question

> "What is limiting a widespread adoption of fault detection methods in WRRF applications?"

A number of challenges that limit adoption of fault detection have been suggested including the following

- Lack of education and knowledge about statistical methods.
  This was suggested by Rosén et al. (2003) and has also been emphasized as barrier for adopting analogous methods such as advanced instrumentation, control and automation (ICA)-methods by Olsson (2006).
- Lack of economic incentives.
  This has been recognized as a limiter for adopting online control (Yuan et al. 2019). Coping with effluent permits is priority, and reduced operational costs are not the main driving force, which potentially could be reduced with improved sensor accuracy enabled by FDMs.
- Mismatch between the current and desired fault detection performance.
  The methods in Chapter 7 could successfully detect a 0.5 mg/L bias. However, an earlier detection would have been need to use the indications for condition based maintenance and correction without a negative impact on the process. Similarly, the data reconciliation and gross error detection in Chapter 4 missed sensors with large bias (> 50%). This needs to be improved, for FDMs to be valuable in practice.

- Lacking incentives to bridge the detection performance gap.
  The research in Chapter 3 and 7 can be said to have increased the technology readiness level (TRL) from TRL 4 to TRL 6 (see Li et al. (2019) for the TRL concept). Most published FDMs are similarly at TRL 6 (prototype demonstration in relevant environment). From a research perspective, low TRL means high novelty and is therefore a higher driving force than research replication. At the same time, the interest from commercial actors have been low to push the TRL, given the few commercially available FDM software's.
- The social context.
  Adoption of evidence-based care in practice was studied in (Grol and Wensing 2004) where aspects of the social context were suggested to impact the adoption rate of new knowledge. These aspects included dedication and incentives to adopt new methods and the leadership. Analogous studies are yet to come in a WRRF context, but I expect similar challenges between the not-for-profit utility and care unit.
- Software and data integration.
  Rigid technical data structures make it difficult to annotate data, integrate different data sources, and implement advanced computations in real-time. Cyber security and robust performance certainly motivate robust IT-systems, although they may reduce the adoption rate of data-driven decision support.

Answering these wide questions is outside the scope for this thesis. The answers are likely both technical and non-technical. In addition, many of the hypothesized barriers were identified by researchers trained in the natural sciences, which may therefore be biased. The perspectives from social sciences, including economists, would be valuable to reveal insights into what main barriers that limit an introduction of fault detection methods to practice. This question is important on its own, because clearly defining the barriers could guide future development of FDMs towards a larger impact on current practices. Ultimately, I stress that we should not be discouraged about this vast array of challenges. Their solutions pose excellent opportunities to improve the way how FDMs can improve wastewater treatment.

10 Conclusions and future work

I finalize this thesis with a chapter containing some general conclusions and suggested topics for future research. For a complete description of the conclusions from this thesis it is suggested to revisit the conclusions in the preceding chapters.

10.1 Conclusions

The main goal with this thesis was to improve data quality by applying FDMs and thereby enable correction of detected faults. Many of the tested methods were successful and revealed insights into sensor and process faults. The results, however, also indicated that an initial poor data quality induced by, for example, insufficient sensor maintenance, may complicate fault detection. Thus, it is recommended to combine condition-based sensor maintenance and fault detection to iteratively improve the data quality.

This thesis demonstrated practical opportunities and challenges in applying data-driven fault detection in WRRFs. Further, the results have increased the understanding about how sensor faults impact the treatment process, and how faults in a wide range of applications can be detected by analysing data. Specifically, three 6-18 months long experiments on pilot and full-scale plants revealed insights into known and unknown sensor and process disturbances and condition-based maintenance of diffusers. The observed sensor faults included damaged and fouled dissolved oxygen sensor membranes, and humid and fouled airflow rate sensors. Additionally, indications of drift direction due to fouling were identified in a nitrate and chemical oxygen demand sensor, two types of dissolved oxygen sensors, and an ammonium sensor. The importance of understanding sensor drift direction to maintain a resource efficient operation was demonstrated by process simulations, which has not been emphasized previously in literature or practice. A few of the indicated process disturbances were influent load variations, uneven load distribution in parallel treatment lines, changed respiration rate and aeration efficiency, a syphon effect in the nitrate recirculation, insufficient sludge sedimentation, and reduced denitrification and nitrification rates. These important and wide insights were obtained by analysing the output from several detection methods, many of which had not previously been applied in a WRRF context. The made insights underline the usefulness of analysing data through fault detection methods, as

these more informative data can be used for improving operational decision support.

Deliberate sensor and system disturbances (excitations) were shown useful for detecting fouling and faults in a dissolved oxygen sensor, and to facilitate grey-box model estimation that was used for diffuser condition monitoring. These disturbances increased the signal-to-noise level, where the noise was commonly large in the studied system due to influent variations that were reflected in data. Thus, such deliberate disturbances in sensors circumvented the difficult problem to distinguish sensor disturbances from process disturbances and is therefore a promising approach for WRRF applications. Similar methods, denoted active fault detection, have been applied in other applications, but seldom on a real system or in a WRRF context.

Training and tuning the alarm threshold in a fault detection method was confirmed to be an essential step to obtain good detection performance. Probabilistic one-class classification methods were shown to facilitate this, although simple methods showed similarly good performance in many of the studied problems. This result demonstrates the value of benchmarking new methods with simple methods. At the same time, all methods were limited by small normal training datasets, and in particular, the data-driven machine learning methods. A mitigating solution to this problem was proposed for sensor fault detection, by using information about the sensor maintenance.

The majority of the studied methods were evaluated on real processes and are therefore applicable for usage in practice, although more research is desired to improve their detection performances. Many of the results replicate the usefulness of data-driven methods in the context of WRRF fault detection. This is important since few examples exist with long-term studies on full-scale processes.

10.2 Future work

The research in this thesis has increased the knowledge and answered (a few) questions related to fault detection in WRRFs. An abundance of new and unanswered research questions has emerged from the studies, which demonstrates the need to further explore how data can be analysed to detect faults in WRRFs. A priority should be given to bridge identified knowledge gaps that are (fairly?) straightforward to obtain in practice (Section 10.2.1). Finally, I suggest some topics for what I think would be relevant future studies (Section 10.2.2).

10.2.1 Identified knowledge gaps

This thesis has demonstrated knowledge gaps that are needed to be closed in order to improve data quality and enable fault detection. Basic knowledge

about sensor behaviour - drift and accuracy - in a realistic environment needs to be completed for the sensor types used in WRRFs. This also includes the prevalence of sensor drift and bias at different sensor maintenance routines.

To assess sensor accuracy, the variance and bias error composition of laboratory samples needs to be better understood.

Knowledge in the form of annotated datasets are needed for different sensor fault modes, which should not be limited to sensor drift and bias.

Methods that facilitate and automate the collection and identification of annotated data are important IT-tools that are needed.

10.2.2 Suggested topics for future studies

Knowing the unknowns in Section 10.2.1 would enable the relevant studies of: i) How sensor maintenance and calibration can be optimized to produce a data quality that balance resources (human efforts and natural resources) and effluent water quality, and ii) How probabilistic one-class classification methods can be applied with improved annotated data. In terms of sensor fault detection, the inclusion of rare sensor faults for improved tuning (resembling supervised classification by contrast to one-class classification) would be an attractive opportunity to study.

Simple deliberate perturbations for improved diagnostic information content is an area of research worthwhile continuing, and specifically, how such can simplify normal mode identification. This would be especially useful for sensor fault detection, which does not disturb the treatment process.

How existing dynamic process models can be used for fault detection, possibly in combination with pattern recognition, is an interesting topic that should be further explored. The mechanistic grey-box modelling was promising for describing changes in diffuser condition, but mechanistic models for fault detection in WRRFs have been less studied than purely data-driven (black-box) methods.

The influent daily and seasonal patterns have in general been a challenge. How these stochastic, but still partly predictable, patterns better can be modelled with probabilistic methods would be a useful component of many FDMs.

The studies in Chapter 2-8 also leaves some unanswered questions that could direct future studies. These suggestions are not repeated here, and it is recommended to revisit the previous chapters for these suggested studies.

Finally, a qualitative research study considering the social context would be valuable as complement the technical studies in revealing the root cause barriers for bringing out the values of fault detection into practice.

11 Supplementary materials

This chapter contains supplementary data and illustrations related to the Chapters 3,4,5,7, and Chapter 8.

S3.1 – Fault signatures and bias progression in dissolved oxygen sensors

This section provides supplementary information about experimental details in Chapter 3.

MATERIAL and METHODS

Experimental set-up

Figure 1 is a photo of the experimental set-up in Henriksdal WRRF.

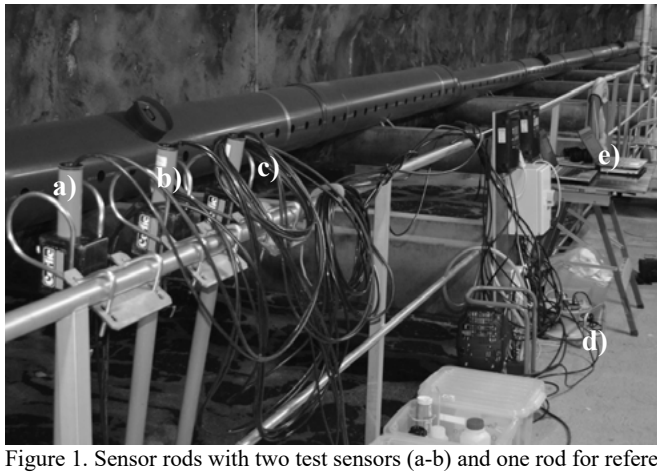


Figure 1. Sensor rods with two test sensors (a-b) and one rod for references (c). Air compressor (d), sensor signal processing units (e), and data collection hardware (f).

Data pre-processing

During the artificial fouling experiments in Henriksdal WRRF the data were down-sampled from 8 Hz to 1 Hz in two steps. Firstly, a low pass filter (anti-alias filter) was applied and secondly the data were decimated (i.e. using only every 8^{th} sample). Figure 2 illustrates the down-sampling for one of the impulses.

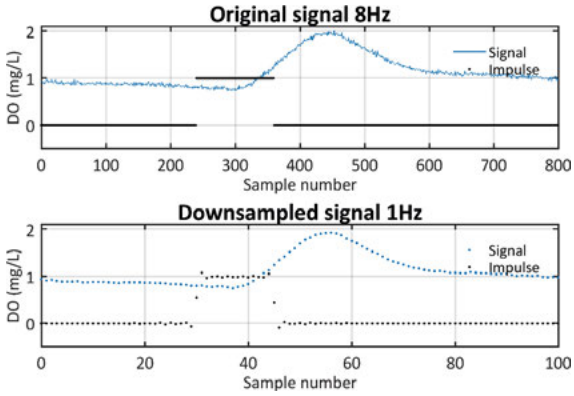


Figure 2. The down-sampling was conducted with the MATLAB function "decimate" which first low-pass filters the signal at 80% of the Nyquist frequency (anti-alias filter) and then decimates the filtered signal by keeping only every 8th sample.

Response time method

Below is a simple function that calculates the response time, considering that the input data have the specific structure of 20s before an impulse and 200s in total length.

```
function [tc] = TimeConstant(yvec)
% Calculates the time constant, 63%, for one impulse.
% yvec should be 1Hz and 20s measurement before impulse, total
length 200s

init=mean(yvec(15:19));
peak=max(yvec);
tcval=0.63*(peak-init)+init;

thigh=find(yvec>=tcval,1); % first second above thresh.
tlow=thigh-1;

% Round off and find closest second to tc
if tlow==0
    tc=-1; % indicate no found time impulse
elseif (yvec(thigh)-tcval)/(tcval-yvec(tlow))>=1
tc=tlow-20;
elseif (yvec(thigh)-tcval)/(tcval-yvec(tlow))<1
tc=thigh-20;
else
    % some kind of error
    error('error in function TimeConstant')
end

end
```

RESULTS

Variations in references measurements during initial measurements

Part of the first fouling procedure is shown in Figure 3 where large variations in the reference sensors (compared to the clean and fouled test sensors) are evident. It is also clear that the bias (the difference between test sensor and reference sensors) for both the clean and fouled test sensors is larger than the desired 0.2 mg/L.

A detailed comparison of the variations for the two references show that they coincide, thereby indicating that they are an effect from variations in the measured media, and are not due to sensor measurement noise (Figure 4).

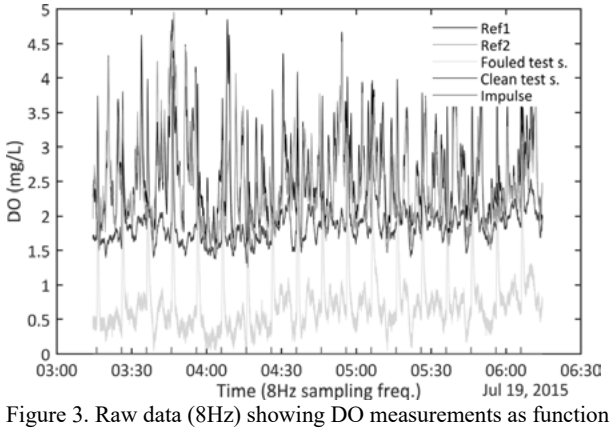


Figure 3. Raw data (8Hz) showing DO measurements as function of time for part of the initial fouling experiment.

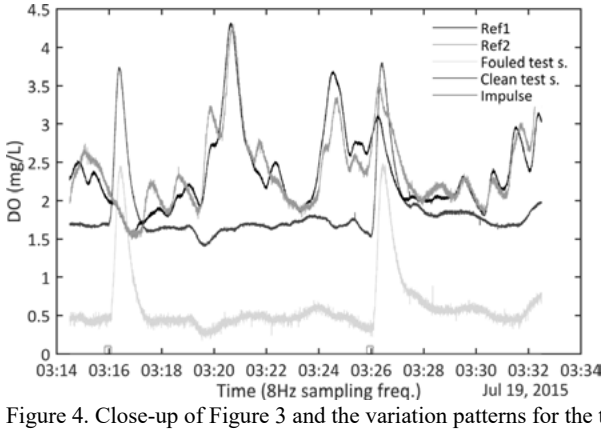


Figure 4. Close-up of Figure 3 and the variation patterns for the two reference sensors. DO measurements as function of time.

During the second experiment, which was conducted at the same location, the variations in the reference measurements decreased at the same time as the bias decreased for the clean test

sensor, see Figure 5. Recall that the bias was obtained as the difference between a test sensor and the WLS-estimate from the two reference measurements.

The correlation between the decreasing variation in the references and the bias is further detailed in Figure 6 where the variance for one reference sensor within a moving average window (15 minute wide window) is shown together with the bias.

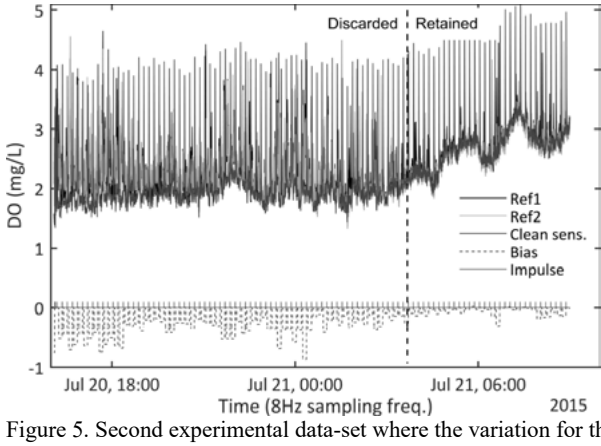


Figure 5. Second experimental data-set where the variation for the reference sensors decrease. Dashed line show borderline between discarded data and retained data used for further evaluation. DO measurements as function of time.

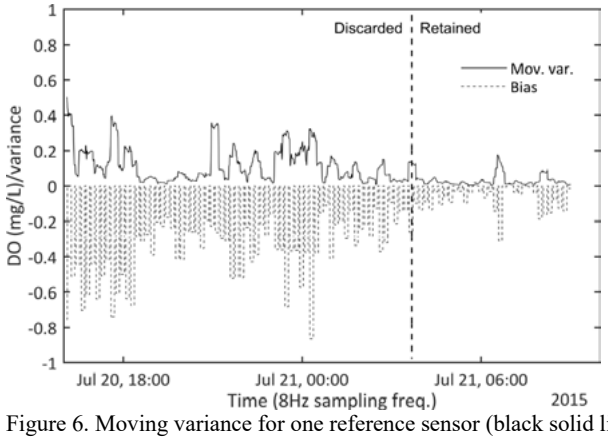


Figure 6. Moving variance for one reference sensor (black solid line) and bias (blue dashed line) as function of time for data in Figure 5. Vertical black dashed line show borderline between discarded and data used for further evaluation.

The last part of the data set exhibited characteristics equivalent with the subsequent data sets that did not show a bias for the clean test sensor, see Figure 7.

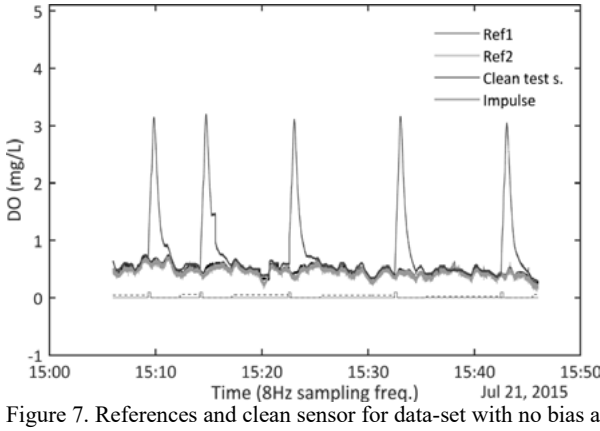


Figure 7. References and clean sensor for data-set with no bias and no variation in reference measurements. DO measurements as function of time.

We were not able to explain the initial variation in reference measurements as no deviating process conditions in hydraulic parameters such as influent flow or air flow were evident, see Figure 8.

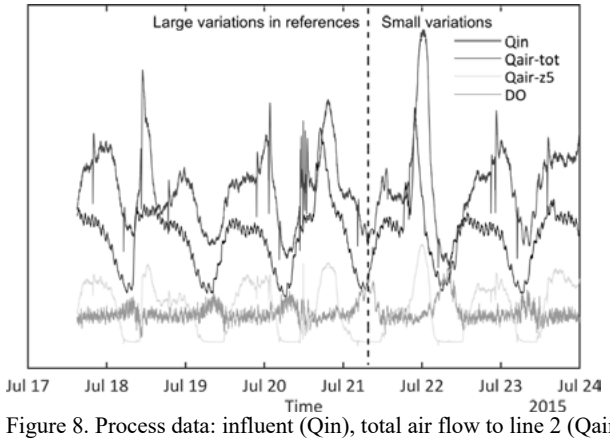


Figure 8. Process data: influent (Qin), total air flow to line 2 (Qair-tot), air flow to zone 5 line 2 (Qair-z5), and dissolved oxygen measurement (DO) from existing sensors in Henriksdal WWRF during the time period during after the large variations in reference measurements. The purpose of the figure is to illustrate the qualitative patterns of process parameters relevant for the DO concentration.

Based on the initial experimental data and the uncertainties about the reason for the variations in the reference measurements that likely caused the bias, we decided to only use the data where the references showed similar characteristics as the subsequent data sets (that did not contain variations in the reference measurements). The borderline between discarded and used data is indicated with vertical dashed line in Figure 5.

Kink on air-supply hose

There was an accidental kink on the air supply hose to the OPT-sensor during cleaning in the end of Period 5. The MEC reference sensor obtained an increase in air since the two sensors

have a common air supply. However, there was no change in response time value despite the altered IR, see Figure 9.

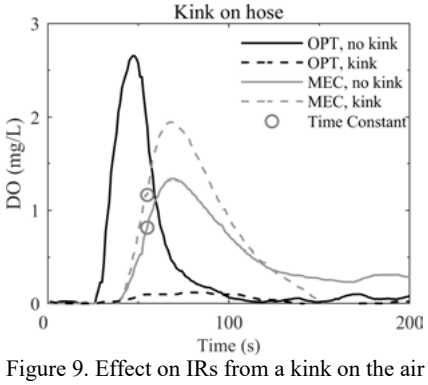


Figure 9. Effect on IRs from a kink on the air supply hose to the OPT sensor. Note that the response time (red circles) did not change despite the change IR peak value.

Damaged sensors

Figure 10 and Figure 11 illustrate the deliberate sensor damage for the MEC and OPT sensor, respectively.

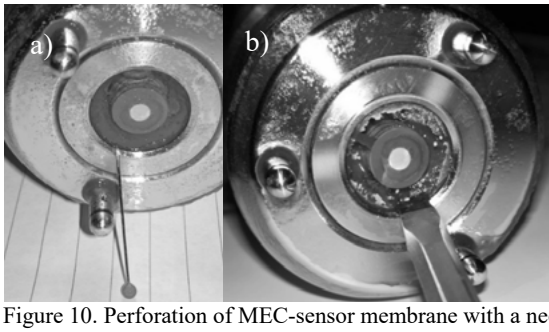


Figure 10. Perforation of MEC-sensor membrane with a needle (a) and a screw-driver (b). Note that the liquid space inside the membrane was emptied when perforated with a screw-driver.

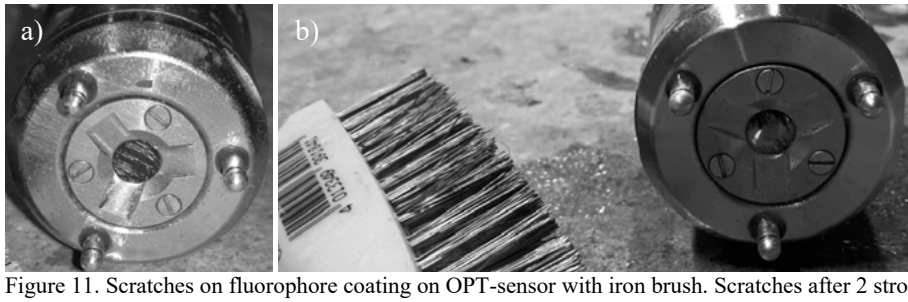


Figure 11. Scratches on fluorophore coating on OPT-sensor with iron brush. Scratches after 2 strokes in (a) and 5 additional strokes in (b).

Experimental data

Pre-processed experimental data are provided in MATLAB.mat format to allow benchmarking of fault detection algorithms, exploiting the IRs. Three datasets are provided at www.ivl.se:
ArtificialFoulingImpulseData.mat (impulse response data)
BiofilmFoulingImpulseData.mat (impulse response data)
BiofilmFoulingTimeSeriesData.mat (time series data "raw data" 1 Hz)

The variable structure for the datasets is described below. All data are free to use under the Creative Commons License CC BY 4.0 (for all purposes).

ArtificialFoulingImpulseData

MEC: Membrane electrochemical DO sensor
Sensor condition
.normal: Clean
.fouled: Fouled with grease mixture
.damaged: Mechanically damaged membrane
.wornout: Repeated air cleaning impulses
.lowair: one loose air hose to cleaning system
.chemicalDamage: Harsh chemical cleaning substance
.highSS: Measuring in return sludge channel
.lowSS: Measuring in unaerated zone, normal suspended solids concentration (about 3000 mg/L)
OPT: Fluorescent DO sensor
Sensor condition
.normal: Clean
.fouled: Fouled with grease mixture

BiofilmFoulingImpulseData

.zone5: Zone where the optical DO sensor was studied, MEC-type reference sensor.
.zone6: Zone where the membrane electrochemical sensor was studied, MEC-type reference sensor.
.MECtest: Membrane electrochemical DO sensor – test sensor (increasingly fouled)
.OPTtest: Fluorescent DO sensor – test sensor (increasingly fouled)
.MECref: Membrane electrochemical sensor – reference sensor (no fouling)
.time: Datetime time format
.indicate: Vector indicating active air cleaning impulse (1= active, 0= no cleaning)
.testPeriod: Period 1-6 as indicated in article

BiofilmFoulingTimeSeriesData

.zone5: Zone where the optical DO sensor was studied, MEC-type reference sensor.
.zone6: Zone where the membrane electrochemical sensor was studied, MEC-type reference sensor.
.time: Datetime time format
.data: Matrix containing reference sensor, test sensor, impulse indication, and Not a Number (NaN) indications.
.metaData: Information about data

S4.1 – Data description of measurement campaign: Practicability of mass balance-based data reconciliation for process rate monitoring in water resource recovery facilities

Pilot plant characteristics

The data in Chapter 4 were obtained from a pilot plant in Käppala Water resource recovery facility (WRRF) during 2019-03-11–2019-10-01. The pilot plant was a part (about 65,000 p.e.) of the full 510,000 p.e. plant and designed for a future process configuration. The pilot plant had the following characteristic data:

Reactor volumes	
Pre-sedimentation	3096 m^3
Activated sludge reactor (total)	18900 m^3
Anoxic zone 1 (pre-denitrification)	1460 m^3
Anoxic zone 2-3 (pre-denitrification)	5780 m^3
Oxic zones 4-6 (nitrification)	6600 m^3
Deox/Oxic flex zones 7-8 (de-/nitrification)	4400 m^3
Deox zone 9 (post-denitrification)	660 m^3
Post-sedimentation	8784 m^3
Reactor depth	
Activated sludge tanks	10 m
Recirculation rates	
Return activated sludge	~ 0.5 Qin
Nitrate recirculation	0.91 m^3/s

Location

Longitude:	59.356289
Latitude:	18.229373
M.a.s.l.	18 (office level)

The process was conventional with possibility to precipitation in both pre-sedimentation and at sand filters with either iron or aluminium. External carbon source (Brenntag, Brentaplus) was available for pre-, and post denitrification although it was only used in the post-denitrification during this experiment.

Data collection

On-line sensors

All on-line sensor data were transferred from the sensor units, to the main process historic database (ABB PGIM) with one second sampling time (or faster). One-minute mean values were

exported from the historic database to files (Microsoft Excel) using a web-based service (Gemit solutions, ACurve). Sensor maintenance was conducted according to existing routines by the plant personnel. All sensors were manually inspected (visual condition and verified location) by the first author prior the experiments (March 2019).

Laboratory samples

All weekly lab samples were collected with a volume proportional sampling device and analysed at the accredited lab at Käppala. The analysis results were stored in the local Lab information system (LIMS) and automatically transferred to the historic database (ABB PGIM). The laboratory data were exported from the historic database to files using a web-based service (Gemit, ACurve).
Four to five grab samples were taken to assess COD:TSS and TKN:TSS ratios in the WAS, external carbon source composition, and sludge (WAS and RAS) composition. The sampling was conducted by the laboratory staff at Käppala and sludge samples (WAS, RAS) were analysed by an accredited external laboratory (Eurofins AB), and carbon source was analysed by the accredited lab at Käppala. The results from the analysis protocols were manually transferred to a spreadsheet.

Air flow measurements

The ventilation air-flow (incoming fresh air and off-gas) were measured using a portable flow velocity sensor (Testo) for a measurement campaign between 2018-03-26 – 2018-04-06. The measurements were carried out by an air measurement expert with assistance from the first author. Three measurements points were considered: Incoming fresh air (1.8 sqm squared channel), off-gas main shaft (1.8 sqm squared channel), and off-gas small circular pipe (0.08 sqm), see Figure 1.

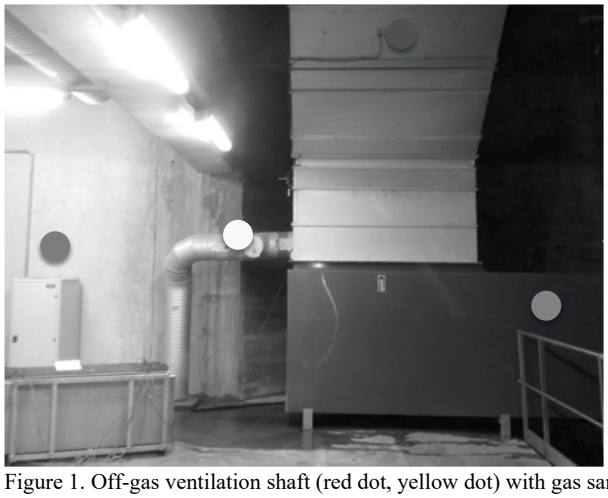


Figure 1. Off-gas ventilation shaft (red dot, yellow dot) with gas sampling pump and sensor cabinet (blue dot) and gas sampling point (purple dot).

For each measurement point a mean air flow was obtained by measuring the air velocity in equally distributed points (drilled holes in the channels) at a location with close to laminar air

flow. For the squared channels, 30 holes were measured and 16 holes for the smaller pipe. The mean values where compared with long-term measurements (> 3days) to obtain an estimate of the daily average for the three measurement positions. The fan speeds in rotations per minute (rpm, tag name: T100_I802/801) for fresh air and off-gas were constant during the experimental period in 2019, and was within 5% of the values for the air velocity measurements 2018.
There was however a noticeable over pressure in the process room for the ASP, which was sealed by automatically closing doors in both ends of the room, apart from the ventilation system.

Data processing

On-line sensors

All on-line sensor data were pre-processed to remove frozen values, outliers, and data segments with unreasonably large variance.
Outliers were identified by the modified Z-score as described in (Boris Iglewicz 1993).
Frozen values were identified as measurements with variance below a certain threshold value in a (about 15-minutes long) moving window. Large variance segments were detected similarly as frozen values, but discarding large variance. All detected anomalies were changed to 'Not a number' and not further included in calculations for the data reconciliation (DR).
The thresholds and parameters for the pre-processing methods were subjectively chosen with the goal to only remove obvious unrealistic data. That is, the pre-processing was conducted in a cautious way. The pre-processing methods were implemented in MATLAB and the applied parameter values were saved together with the raw and pre-processed data.

Specification of sampling error

The procedure and settings used to define uncertainties and errors for both laboratory and on-line sensors are detailed in this section. Sampling errors for laboratory samples were estimated using the tool in (Rossi et al. 2011) which is in turn based on the theory of sampling, see e.g. (Petersen and Esbensen 2005). Time series for flow and concentration measurements from the pilot plant were evaluated to estimate the sampling errors.

The total sampling error rely on four independent error components:

- Point selection error (PSE) depend on the concentration variation in time, and how representative the sample is for the given time series. In this study, weekly flow proportional samples were obtained to have < 1% PSE, grab samples for sensor calibration had about 5% PSE.

- Fundamental sampling error (FSE) depend on the heterogeneity of the studied component in the sampled system. For water soluble components it was assumed to be negligible. For sludge samples is was likely larger, but we lack data for quantifying it. In this study we applied 0% for all samples, despite that it was likely to be larger for sludge samples.

- Point Materialization error (PME) is the sum of several error components that describe how representative the sample is compared to the sampled system. As an indication the value has been assumed to lie between 20% and 100% for correct turbulent sampling conditions or laminar flow conditions in a sewer ((Rossi et al. 2011). Here, we assumed 20% PME for water phase samples sampled with a sampling device (weekly samples), 30% PME for grab water phase samples and 40% for grab sludge samples.

- Sample weighting error (SWE) is an effect of whether the mean values for the sample are obtained with accurate relation to flow variations. Here, we the SWE was zero since the error due to flow measurements is separately specified for each flow measurement and considered when concentration measurements were made flow proportional.

We note that in the end, the total sampling error was based on the PME magnitude, which was chosen based on the indications in (Rossi et al. 2011).

REFERENCES

Boris Iglewicz, D.H. (1993) The ASQC Basic References in Quality Control: Statistical Techniques. Volume 16: How to Detect and Handle Outliers,, ASQC, United States of America.
Petersen, L. and Esbensen, K.H. (2005) Representative process sampling for reliable data analysis - A tutorial. Journal of Chemometrics 19(11-12), 625-647.
Rossi, L., Rumley, L., Ort, C., Minkkinen, P., Barry, D.A. and Chevre, N. (2011) Samplinghelper a web-based tool to assess the reliability of sampling strategies in sewers and receiving waters. Water Sci Technol 63(12), 2975-2982.

S4.2 – Estimated error magnitudes for sensors, laboratory measurements, and conversion factors

This section detail error magnitudes for conversion factors and separately estimated values (Table 1), sensors (Table 2), and laboratory samples (Table 3), which were used to define the error covariance matrix in Chapter 4. The settings used in Sampling Helper (Rossi et al. 2011) are given in Table 4.

Table 1 Error magnitudes for conversion factors and constant (separately estimated) measurements.

Position	Variable name	Unit	1 std error	Analytical error	Sampling error
ven					
	Qven	m^3/s	0.2	-	-
aer					
off					
	Qoff	m^3/s	0.2	-	-
eff					
	SSeff	mg/L	0.15	-	-
	CODeff	mg/L	0.25	-	-
was					
	NOwas	mg/L	0.13	0.13	
	SS2CODwas	-	0.01	0	0.01
	SS2TKNwas	-	0.01	0	0.01
	SS2FEwas	-	0.01	0	0.01
	SS2TPwas	-	0.01	0	0.01
ps					
	DS2TPps	-	0.01	0	0.01
	DS2FEps	-	0.01	0	0.01

Table 2. Sensor location, make and error magnitude measured in terms of one standard deviation of the mean.

Position	Variable_name	Unit	1 std error	Analytical error	Sampling error	Sensor error	Op.experince error	Sensor make
in								
	QFEin	L/h	0.010	0	0	0.01		EH magflow
	QinTot	m^3/s	0.100	0	0	0.01	0.1	Calculated value
pre								
	Qpre	L/s	0.010	0	0	0.01		Nivus, Echo + 2 velocity sensors
	QpreCalc	L/s	0.100	0	0	0.01	0.1	Calculated from total influent flow
asp								
	QNOasp	m^3/s	0.100	0	0	0.01	0.1	Calculated value from pump curves
	PO4asp	mg/L	0.206	0.05	0.20	0		Applikon, analyzer
	CODasp	mg/L	0.206	0.05	0.20	0		SCAN, Spectrolyzer
	NOasp	mg/L	0.224	0.1	0.20	0		SCAN, Spectrolyzer
	NHasp	mg/L	0.245	0.1	0.20	0	0.1	HACH ion selective sensor
z7								
	NOz7	mg/L	0.320	0.1	0.30	0		HACH
pos								
	NOpos	mg/L	0.320	0.1	0.30	0		HACH nitratax
	DOpos	mg/L	0.125	0.125	0	0		WTW
	SSpos	mg/L	0.308	0.05	0.30	0		Cerlic old SS-sensor ssm myp
fil								
	Qfil	m^3/h	0.010	0	0	0.01		Endress Hausser, promag
	NOfil	mg/L	0.304	0.01	0.30	0		HACH ion selective
	NHfil	mg/L	0.313	0.075	0.30	0		WTW
	QFEfil	L/h	0.010	0	0	0.01		Endress Hausser
	Tempfil	Celsius	0.000					
eff								
	Qeff	kbm/h	0.010	0	0	0.01		Endress Hausser, promag F
	NOeff	mg/L	0.320	0.1	0.30	0		Awa 3s
	NHeff	mg/L	0.313	0.075	0.30	0		Applikon, analyzer
	PO4eff	mg/L	0.308	0.05	0.30	0		Bran Lubbe
	pHeff	pH	0.000			0		Stratos ms
	SSeff	%	0.308	0.05	0.30	0		Sigrist, photometer

Position	Variable_name	Unit	1 std error	Analytical error	Sampling error	Sensor error	Op.experince error	Sensor make
ps								
	Qps	L/s	0.010	0	0	0.01		Endress Hausser, promass 1
	DSps	DS	0.308	0.05	0.30	0		Endress Hausser, promass 1
ras								
	Qras	m^3/s	0.010	0	0	0.01		PSM, internationel
was								
	Qwas	L/s	0.010	0	0	0.01		Endress Hausser, promag P
	SSwas	mg/L	0.308	0.05	0.30	0		Cerlic ITX
ext								
	Qext_z8	L/h	0.010	0	0.00	0.01		Endress Hausser, cubemass C
	Qext_z1	L/h	0.010	0	0	0.01		Endress Hausser, cubemass C
	FEext_fil	L/h	0.000					
aer								
	Qaer2	m^3/h	0.010	0	0	0.01		ABB sensyflow
	Qaer1	m^3/h	0.010	0	0	0.01		ABB sensyflow
off								
	CO2off	ppm	0.036	0.03	0	0.02		Vaisala, GMP343
	O2off	%	0.011	0.005	0	0.01		Servomex, multiexact 4100

Table 3. Laboratory samples, their sampling location, and error magnitudes measured in terms of one standard deviation of the mean.

Position	Variable name	Unit	1 std error	Analytical error	Sampling error
in					
	FEinW	mg/L	0.24	0.125	0.20
	TPinW	mg/L	0.21	0.05	0.20
	SAMPLEin				
asp					
	CODaspG	mg/L	0.31	0.05	0.30
	CODaspW	mg/L	0.21	0.05	0.20
	NOaspW	mg/L	0.22	0.1	0.20
	NOaspG	mg/L	0.32	0.1	0.30
	NaspW	mg/L	0.21	0.075	0.20
	FEaspW	mg/L	0.24	0.125	0.20
	TPaspW	mg/L	0.21	0.05	0.20
	PO4aspG	mg/L	0.31	0.05	0.30
z8					
	TPFILTz8G	mg/L	0.31	0.05	0.30
	TPz8G	mg/L	0.31	0.05	0.30
	SSz8G	mg/L	0.31	0.05	0.30
fil					
	CODfilW	mg/L	0.21	0.05	0.20
	FEfilW	mg/L	0.24	0.125	0.20
	TPfilW	mg/L	0.21	0.05	0.20
	NfilW	mg/L	0.21	0.075	0.20
	NOfilW	mg/L	0.22	0.1	0.20
eff					
	CODeffW	mg/L	0.21	0.05	0.20
	FEeffW	mg/L	0.24	0.125	0.20
	TPeffW	mg/L	0.21	0.05	0.20
	NeffW	mg/L	0.21	0.075	0.20
	NOeffW	mg/L	0.22	0.1	0.20
ras					
	NOrasG	mg/L	0.61	0.1	0.60
was					
	TPFILTwasG	mg/L	0.43	0.135	0.40
ext					
	FEext_pre		0.00		
	NextG	mg/L	0.42	0.11	0.40
	TPextG	mg/L	0.43	0.135	0.40
	NHextG	mg/L	0.42	0.11	0.40
	NOextG	mg/L	0.42	0.13	0.40
	CODextG	mg/L	0.41	0.05	0.40

Table 4 Settings used in Sampling Helper to estimate sampling errors.

	Weekly lab samples	**Grab samples**	**Grab sampling campaign**
Number of samples	100	1	1
Sampling strategy	Systematic	Random	Random
Point selection error	**0.01**	**0.05**	**0.05**
Min/Max/Avg concentration	1/700/200	200/500/300	40e3/90e3/65e3
Min/max/avg particle size	1e-8/1e-1/1e-2	0.01/0.2/0.1	0.1/0.1/0.1
Sample size	50mL	500 mL	500mL
Fundamental sampling error	**0.00**	**0.00**	**0**
Point Materialization error	**0.20**	**0.30**	**0.4**
Flow measurement	**0.00**	**0.00**	**0**
Total sampling error	**20.02%**	**30.41%**	**40.31%**

S5.1 – Gaussian process regression for monitoring and fault detection of wastewater treatment processes

This section provides supplementary information about the kernel and GPR-SM settings in Chapter 5.

Kernel parameter intervals

Kernel structure and parameter intervals with equivalent equation numbering as in original paper.

$$k_{Single}(x_i, x_j) = \theta_1 \exp\left[-\frac{(x_i - x_j)^2}{2\theta_2^2}\right] \quad (8)$$

$$k_{Combined}(x_i, x_j) = k_{Single}(x_i, x_j) + \theta_3\theta_4 \exp\left[-\frac{(x_i - x_j)^2}{2\theta_5^2}\right] \theta_6^2 \cos(\frac{2\pi}{\theta_7}(x_i - x_j)) \quad (18)$$

$$y(x) = f(x) + \mathcal{E}(x) \quad (9)$$

$$\mathcal{E}(x) \sim GP\left(0, k_{WN}(x_i, x_j)\right), \quad k_{WN}(x_i, x_j) = \theta_{\sigma^2}\, \mathrm{I}_n \quad (10)$$

	Parameter interval							
Kernel	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_{σ^2}
Single kernel, uniform prior distribution interval	1e2 – 1.5e4	0 – 10	-	-	-	-	-	10 – 5e3
Combined kernel, uniform prior distribution interval	0 – 1e4	0 – 1e2	0 – 5e4	0 – 1e2	0 – 1e4	0 – 10	0 – 1e4	1 – 1e5

Settings for GPR-SMC

The following parameter settings were used for GPR-SMC:

N = 40, number of particles
K = 10, number of MCMC steps
P = 20, number of batches

S7.1 – Automated active fault detection in fouled dissolved oxygen sensors

This section contain supplementary materials, which detail and complement the results (and statements) in Chapter 7.

MATERIAL and METHODS

The settings for the GPR is given in Table 1 along with threshold values for all methods in Table 2.

Table 1. Settings for the four GPR variants.

	GPR1seLL	GPR1se	GPR2se	GPR2ard
Particles/repetitions (N)	40			
MCMC-steps (K)	N.A.	10		
Batches (P)		20		
Parameter interval (uniformly distributed)				
Length scale	[0.0 100]			[0.0 100],[0.0 100]
Variance	[0.0 100]			
Noise	[0.1 100]			
Constant (mean function	-	-	[-10 10]	
Slope (mean function)	-	-	[-10 10], [-10 10]	

Table 2. Range of alarm threshold values to produce receiver operator characteristics (ROC) curves.

Method	Parameter grid
PEA	[(0:0.01:30)]
RT RTi	[(0.1:0.1:30)]
PCAstand.	[0.5;0.4;0.35; 0.3;0.25; 0.2;0.1; 0.11; 1e-1;1e-2;1e-3;1e-4;1e-5;1e-6;1e-7;1e-8;1e-9;1e-10;1e-11;1e-12;1e-13;1e-14;1e-15]
PCAcomb.	[0.5;0.4;0.35; 0.3;0.25; 0.2;0.1; 0.11; 1e-1;1e-2;1e-3;1e-4;1e-5;1e-6;1e-7;1e-8;1e-9;1e-10;1e-11;1e-12;1e-13;1e-14;1e-15]
GPR1seBF	[0: 1e-20; (1e-15:1e-1:1e40); 1e45;1e50;1e55;1e60]
GPR1seLR	[-1e2; (-30:1:30); (30:10:100)]
GPR2se GPR2ard	[0:0.1:30]

RESULTS

Evolvement of IR shapes for MEC sensor data

The Figure 1 to Figure 5 show the variation in the 12 training IRs for Period 1 to Period 5. The figures show the results for claim in Chapter 7

"Their spread in initial DO concentration differed (see Supplementary materials Figure 1-5), and interestingly, Period 4 with the most false alarms had the smallest spread. To study whether training data were representative for normal conditions, we compared the variation in the 12 training IRs for all periods. They differed in initial DO concentration (see Supplementary materials Figure 1-5), and interestingly, Period 4 with most false alarms had the smallest spread."

Period 1

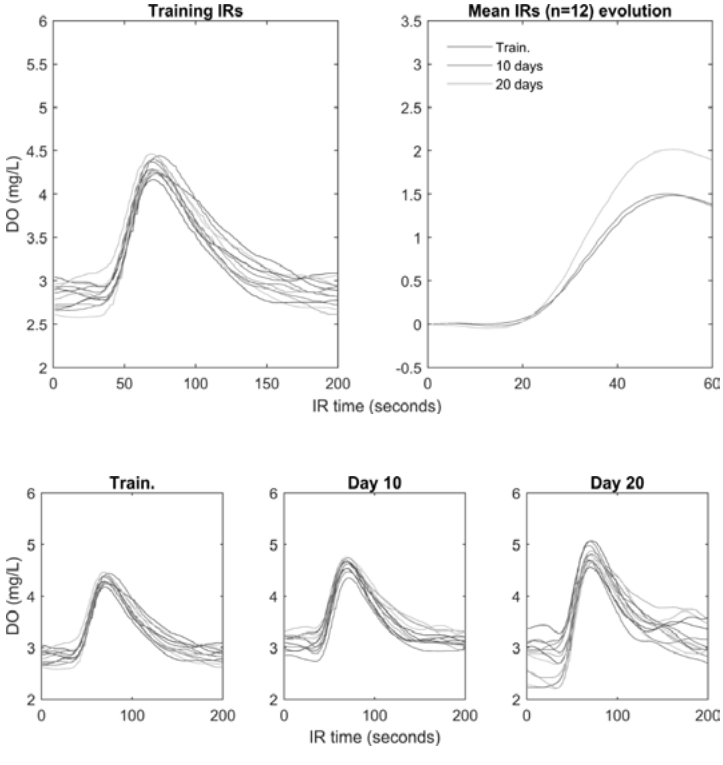


Figure 1. Period 1.

Period 2

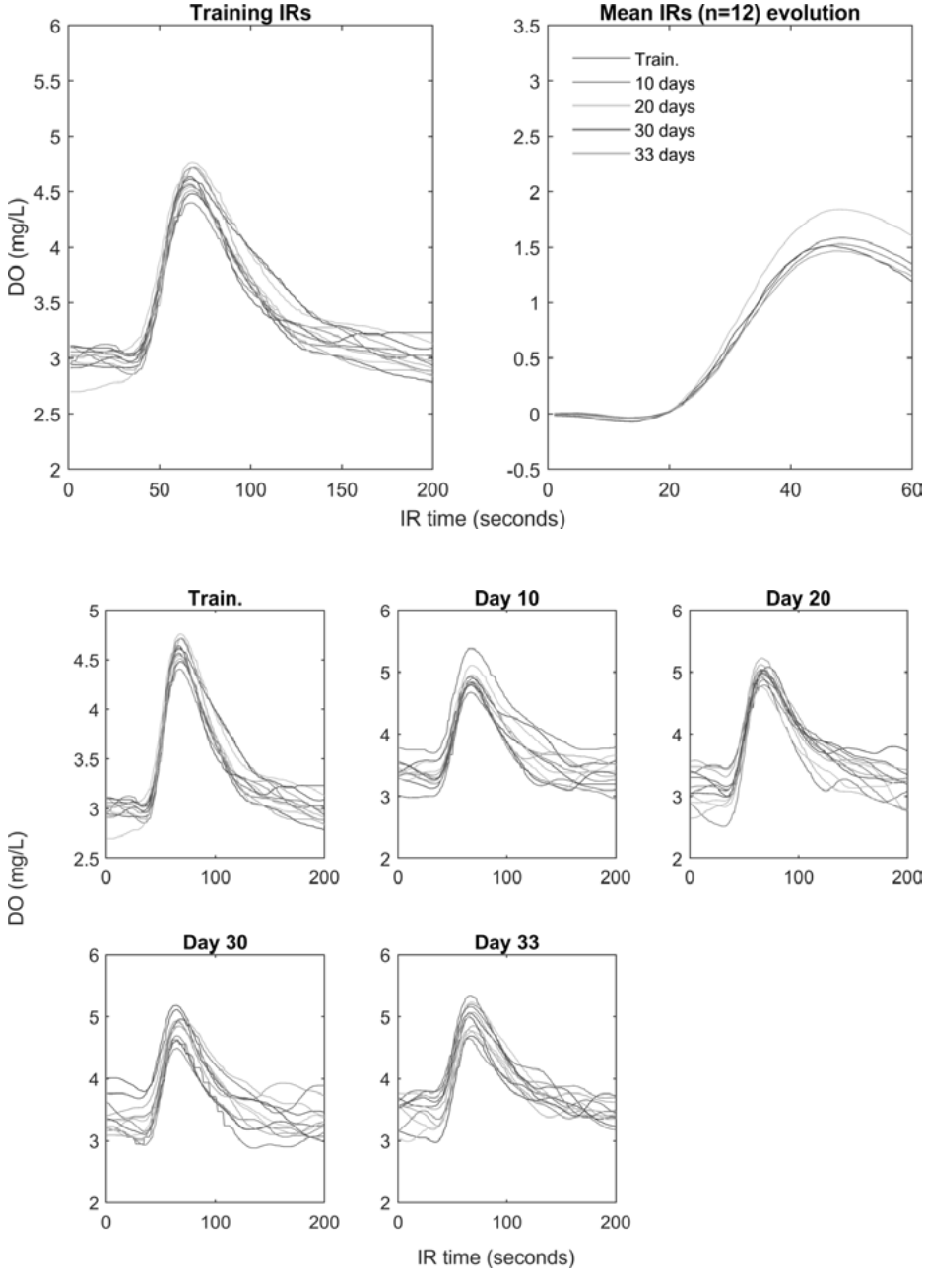


Figure 2. Period 2.

Period 3

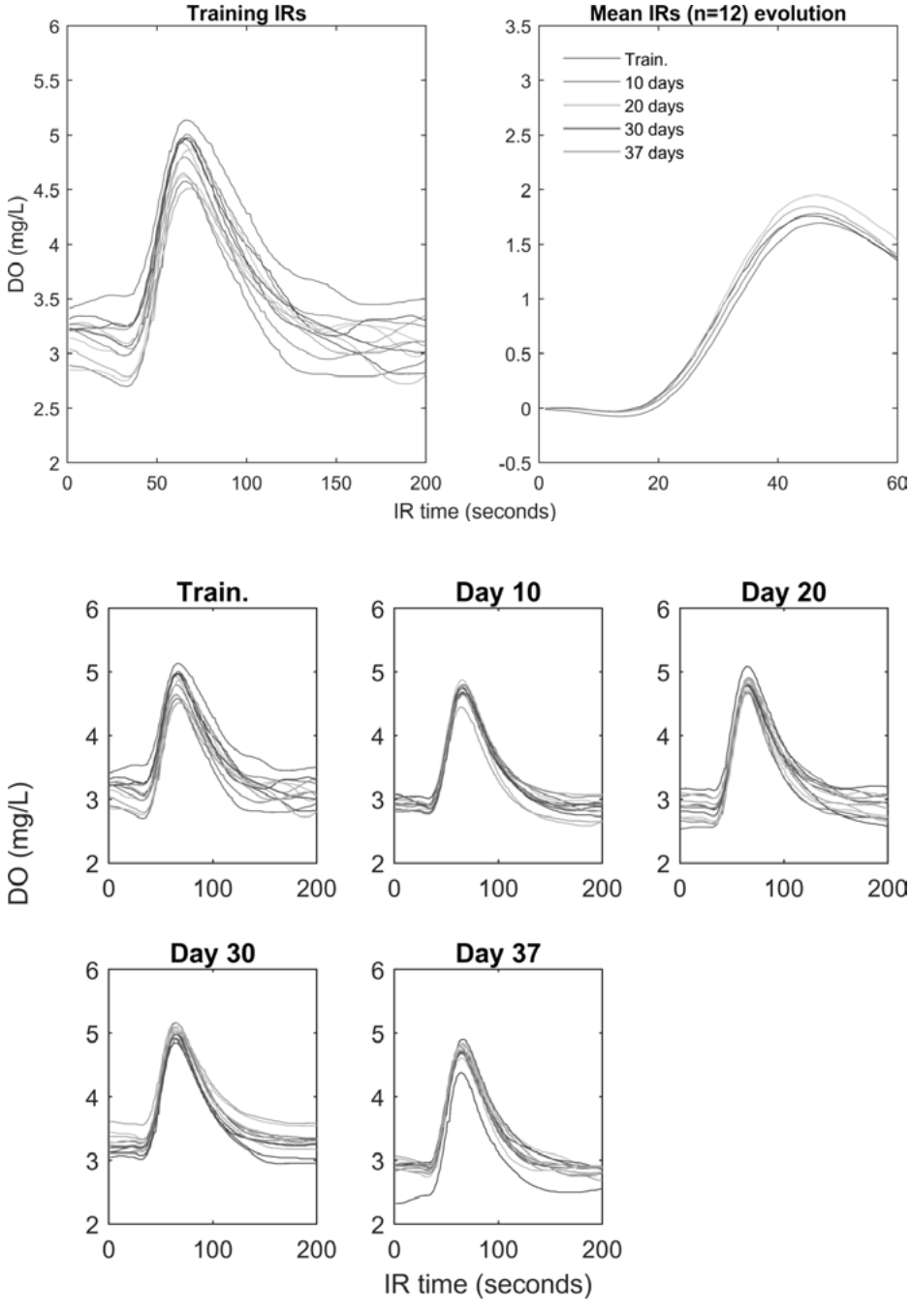


Figure 3. Period 3.

Period 4

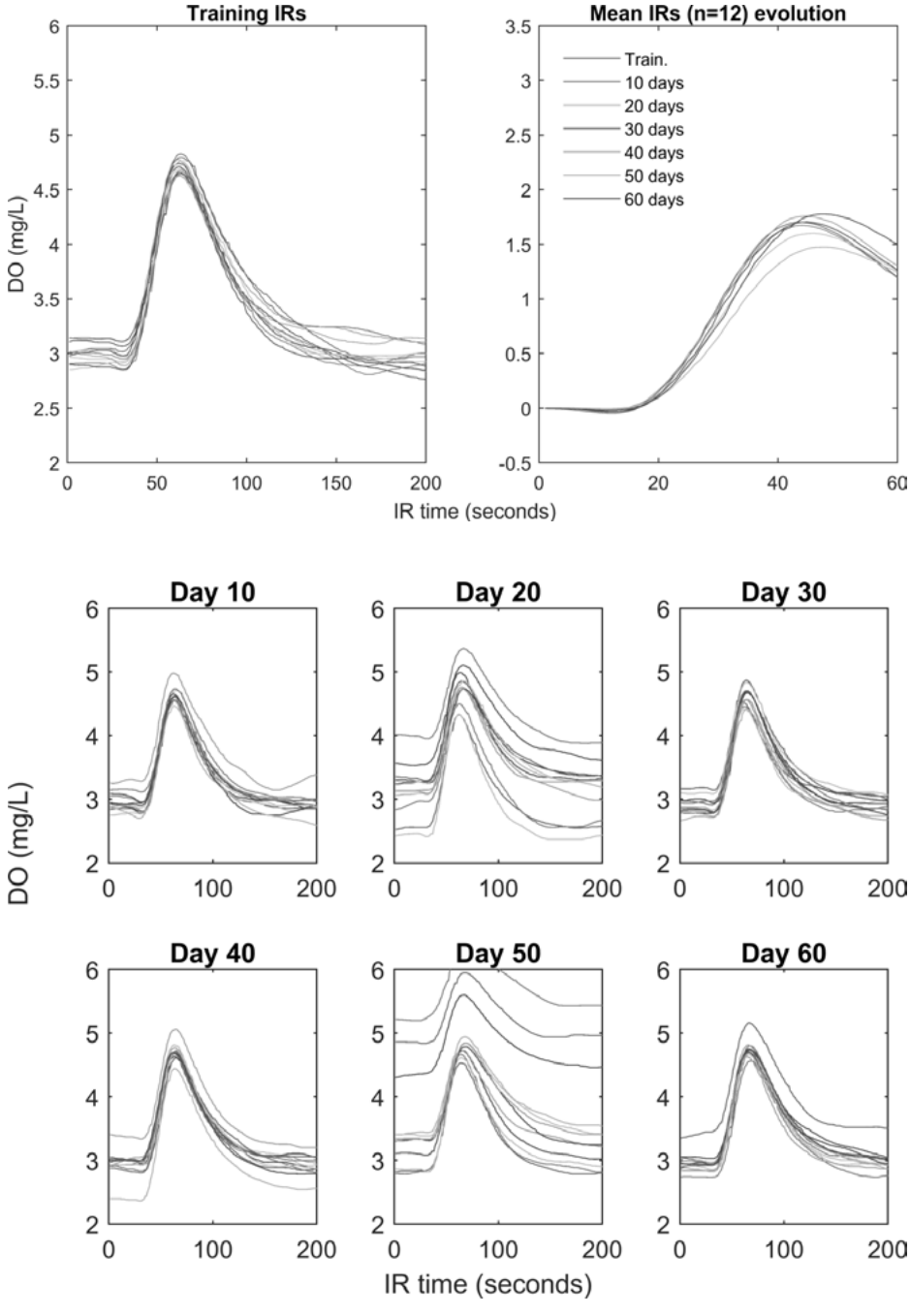


Figure 4. Period 4.

Period 5

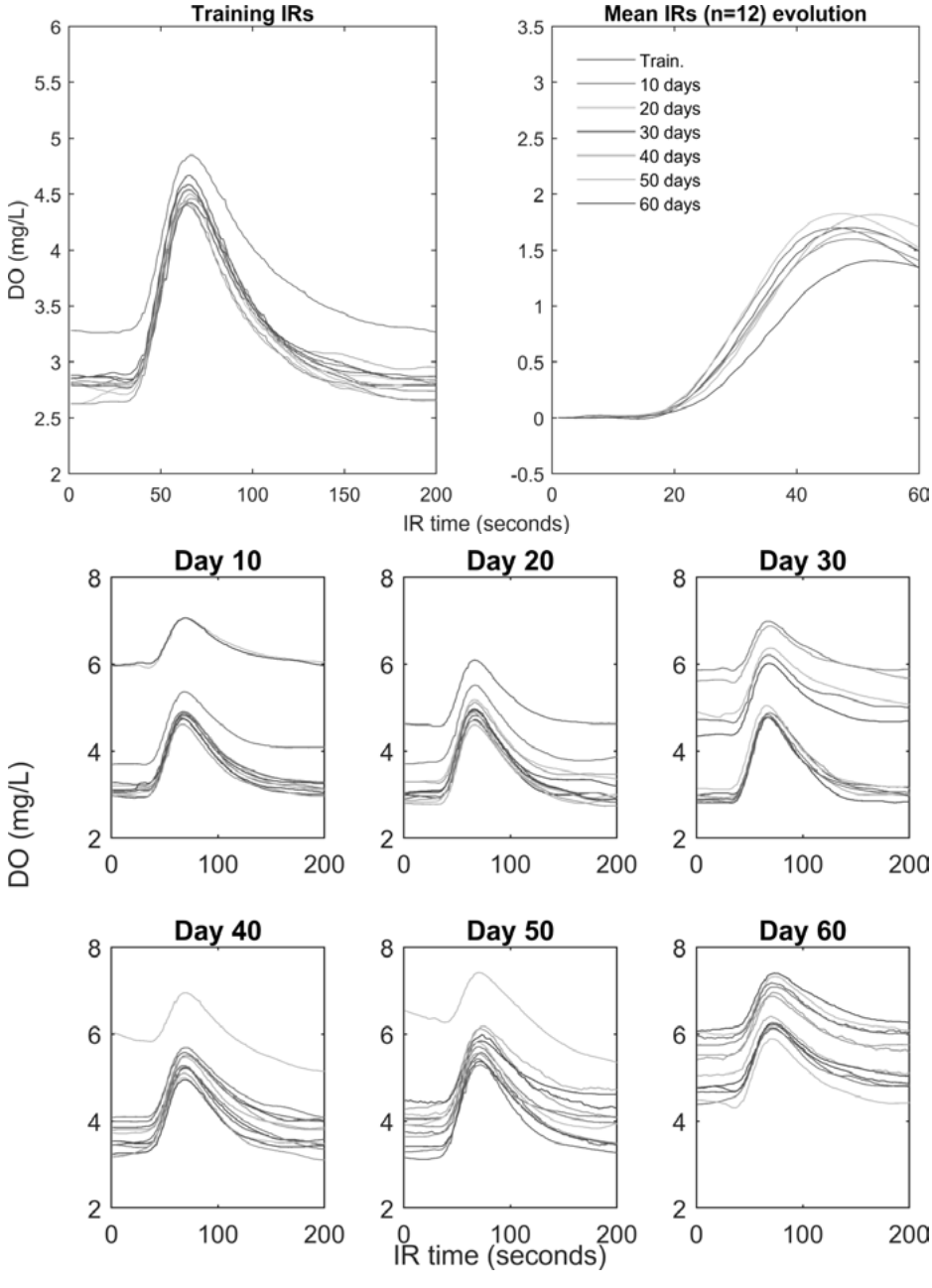


Figure 5. Period 5. Note that the scale of the Y axis in bottom plot is larger than in previous plots.

Intervention on training data

Three IRs in training data in Period 4 and 5 were replaced with IRs from the following two days to increase variation (Figure 6) and decrease variation (Figure 7) in training data. The results show what was stressed in the original article:

"To test the influence of spread in initial DO concentration in training data on false alarm rates, we replaced three IRs in the original data, with IRs from the following day in order to change the spread in training data. The spread was increased in Period 4 and reduced in Period 5 and we then repeated the evaluation (see Supplementary Materials Figure 6-7)."

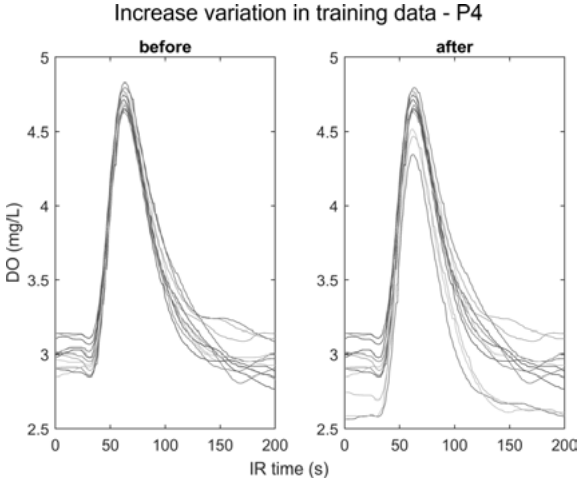


Figure 6. Increased variation in training data Period 4.

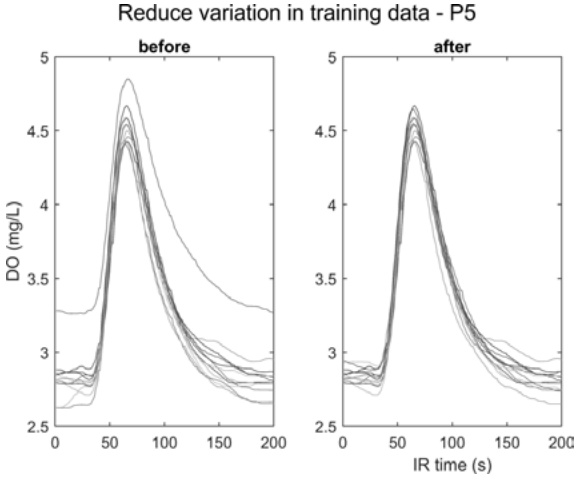


Figure 7. Decreased variation in training data Period 4.

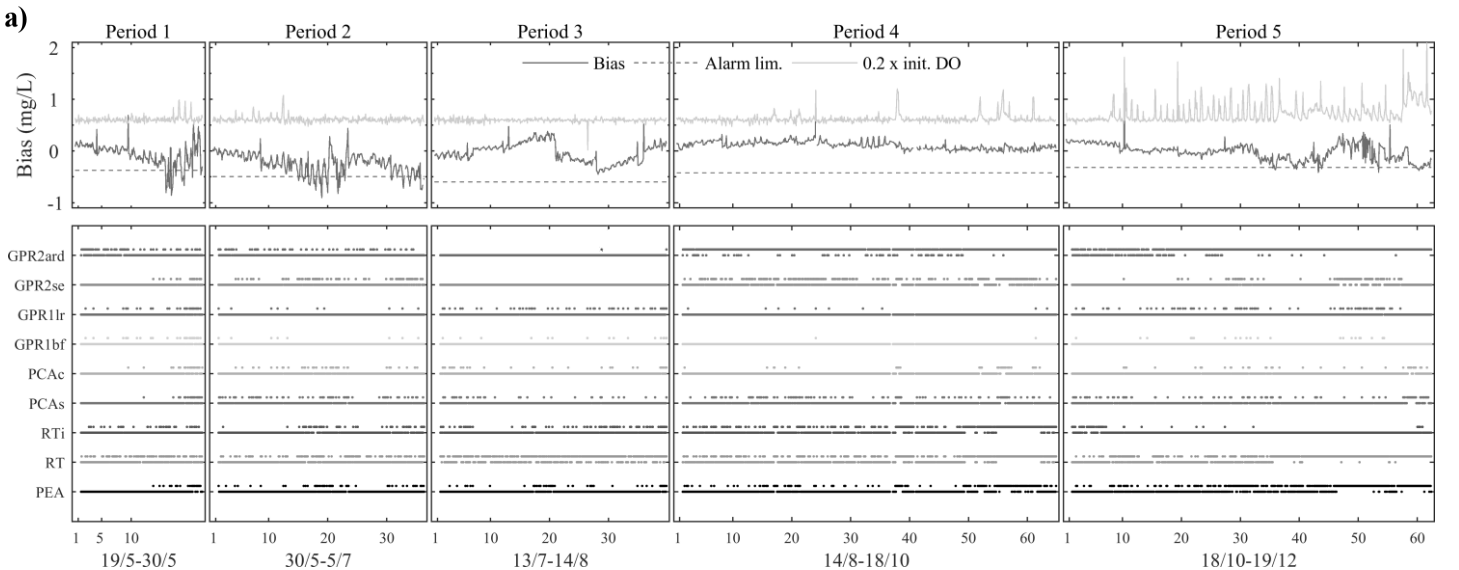
Re-evaluation of results with changed variation in training data

The results in Figure 8 show the change in false alarm rates when the training data were changed according to Figure 6 and Figure 7. Note that alarm (upper dots for each method in Figure 8) and no alarm (upper dots for each method in Figure 8) are shown with the same colours in contrast to the Chapter 7 where false alarms and missed alarms are shown with different markers and colours.

The change in training data gave the expected results for RTi and GP2se with decreased false alarm rate in Period 4 and increased false alarm rate in Period 5. The change in variation in training data had little or negative impact on the remaining methods.

The Figure 8 contain the results for the claimed results in the original article:

“The revaluated results showed that changing the spread in training data had the expected effect (see Figure 8 in Supplementary Materials). The GPR2se and RTi decreased false alarm rates in Period 4 from 29% and 41%, to 8% and 1% respectively. In period 5 the false alarm rates increased from 14% and 6% to 18% and 52% respectively, because of decreased spread in training data. Neither PEA nor RT were affected by the change in training data. This was reasonable since neither of them used the initial DO concentration explicitly in Feature space and they should therefore not benefit from an increased spread.


a)
Period 1
Period 2
Period 3
Period 4
Period 5
Bias
Alarm lim.
0.2 x init. DO
Bias (mg/L)
2
1
0
-1
GPR2ard
GPR2se
GPR1lr
GPR1bf
PCAc
PCAs
RTi
RT
PEA
1 5 10
1 10 20 30
1 10 20 30
1 10 20 30 40 50 60
1 10 20 30 40 50 60
19/5-30/5
30/5-5/7
13/7-14/8
14/8-18/10
18/10-19/12

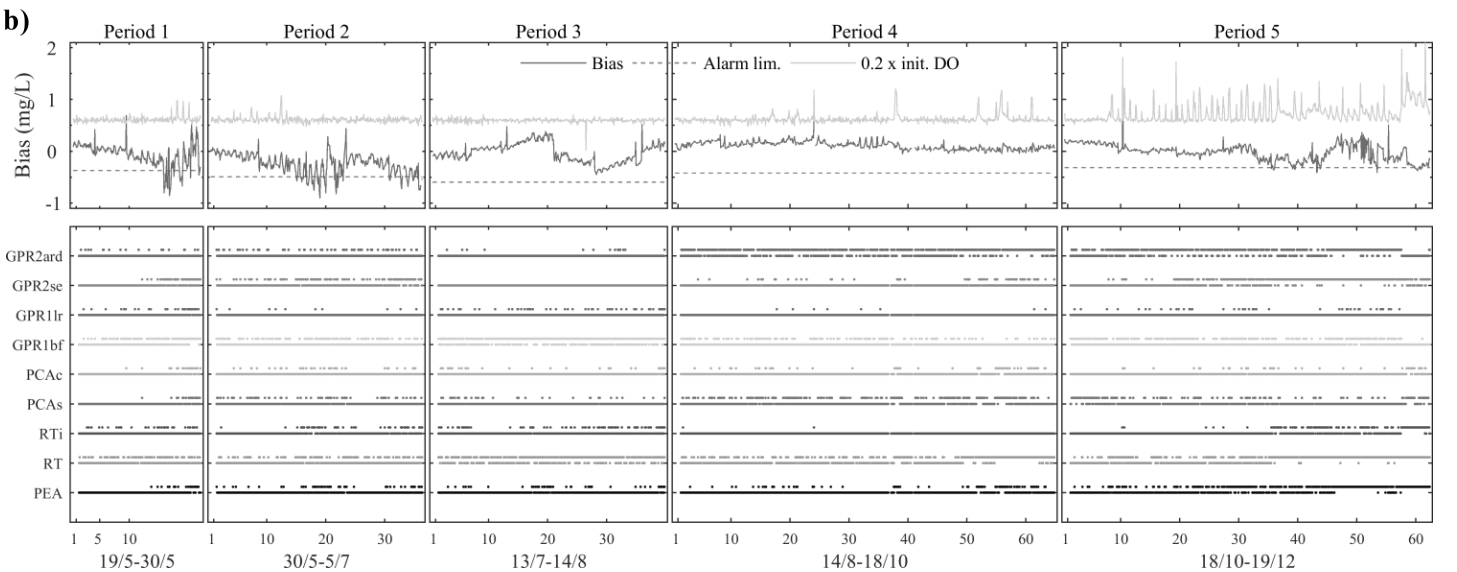


Figure 8. **a**) Original training data with many false alarms in Period 4. **b**) results with increased variation in training data for Period 4 and reduced variation for training data in Period 5. For each method (different colours), no alarm raised (lower line) or alarm raised (upper line) is indicated with dots.

Varying performance for GPR2se

The results in Figure 9 show the variation in performance for GPR2se and GPR2ard while maintaining training- and test data. Thus, the results show that variation in detection performances was only due to the stochastic properties within the GPR method.

Claim in Chapter 7:
”We repeated the evaluation for GPR2se and GPR2ard 10 times on the OPT sensor data, which showed that their performance varied (Supplementary Materials Figure 9). Since it was only the tuning that differed between the 10 iterations, we conclude that the varying performance for both GPR2se and GPR2ard was an effect of the stochastic properties in the method. This was surprising since the particle filter variant of GPR has been shown to produce robust results, see (Samuelsson et al. 2017).”

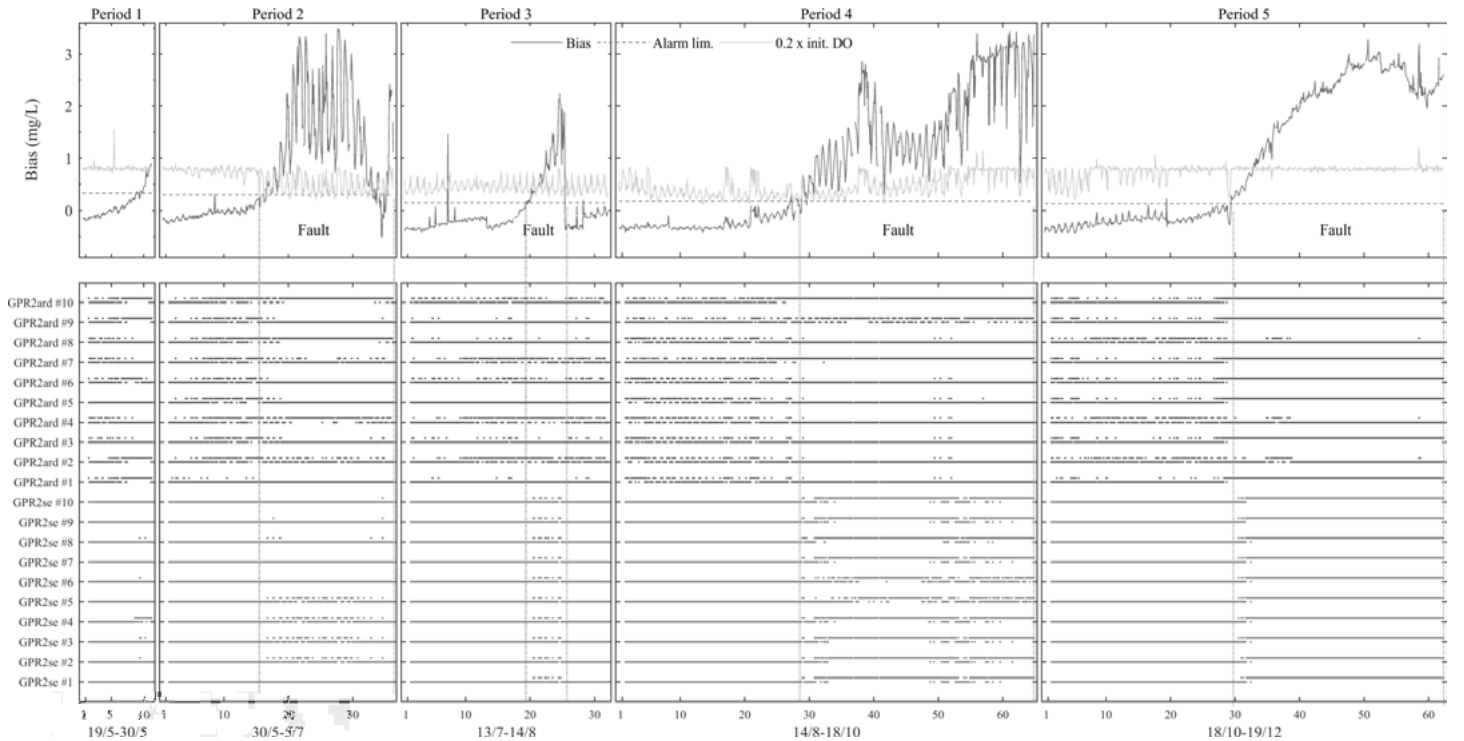


Figure 9. Ten repeated training and evaluations of GPR2se and GPR2ard with same training and testing conditions. For each iteration, no alarm raised (lower line) or alarm raised (upper line) is indicated. Note the difference in amount of false alarms in Period 2 for GPR2se (blue dots).

Training IRs D_{art}

Many of the FD-methods had deviating threshold values for the first iteration in D_{art}. The Figure 10 show the difference in training IRs for the four iterations, where one of the IRs in iteration one deviated (red line).

The results here are the basis for the claim in the original article

“One IR with a different shape was present in the training data in the first iteration (Supplementary Materials Figure 10) and was assumed to be the reason for the deviating threshold values

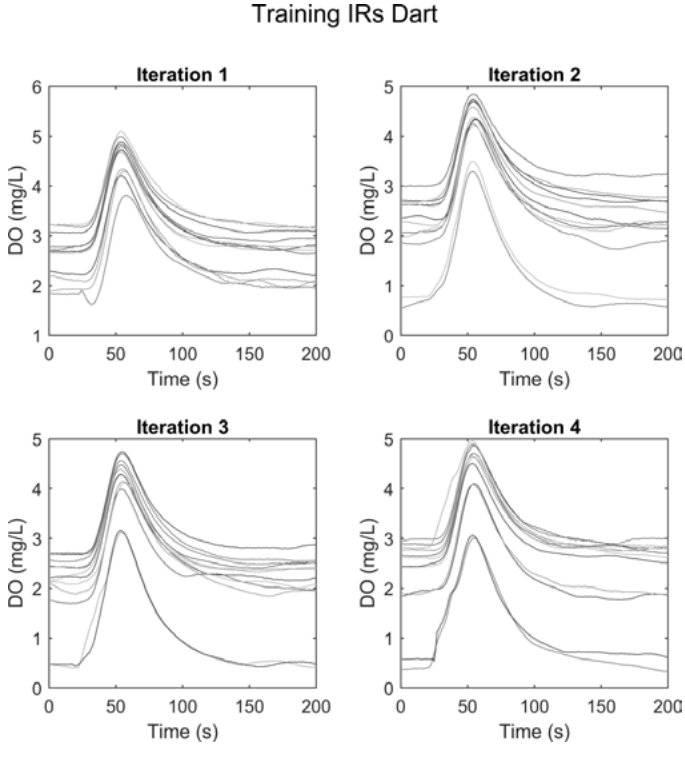


Figure 10. Training IRs in the Dart dataset.

S8.1 – Monitoring diffuser fouling with grey-box modelling

These supplementary materials complement Chapter 8 with additional figures of the estimated $k_L a$ for Line 1 and Line 2.

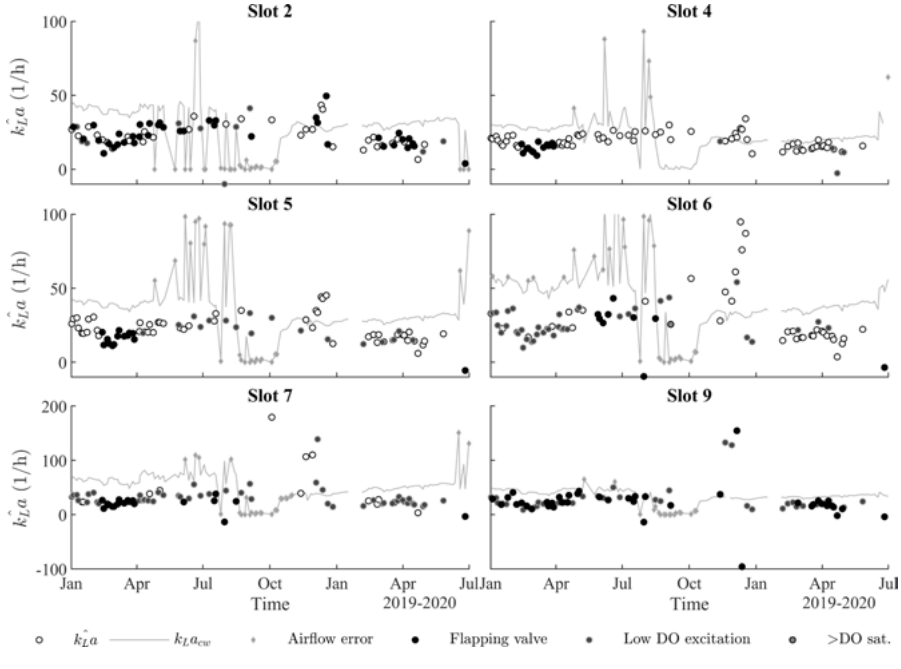
RESULTS and DISCUSSION

Estimating $\boldsymbol{k_L a}$

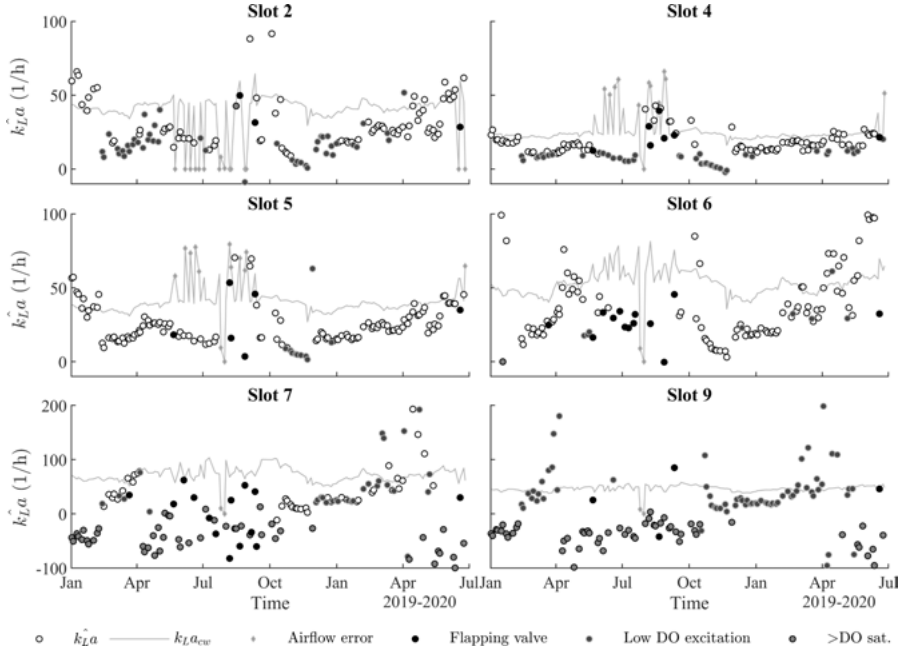
A complete set of figures for all positions with $k_L a$ estimates are shown in Figure 1(a)-(d). In Chapter 8, only Figure 1(d) is shown as a representative example.

Note the large occurrence of potentially biased DO sensor measurements in Line 1 zone 4 (Figure 1(b)), compared to the other positions.

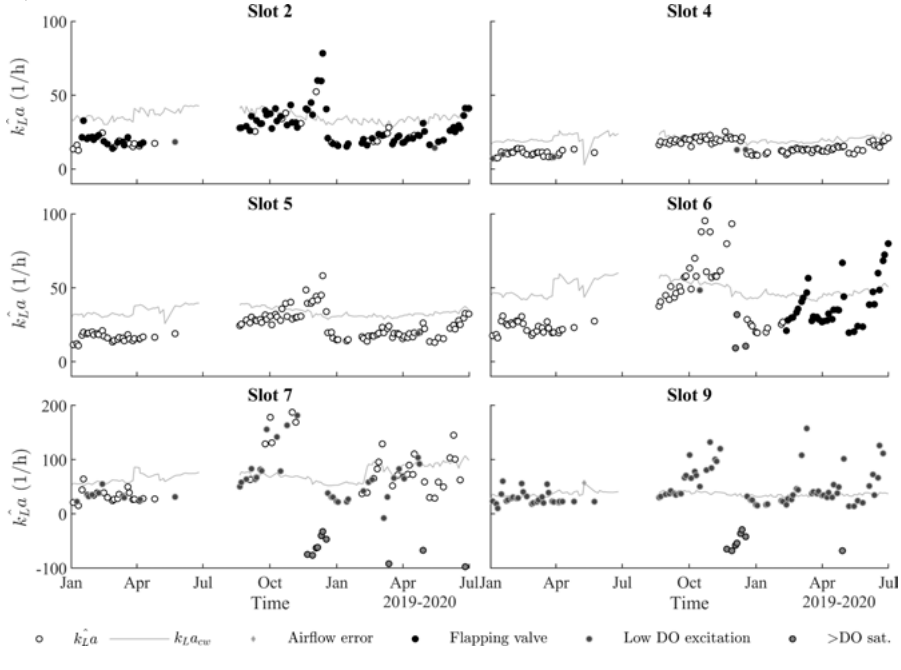
a) L1z3



b) L1z4


Slot 2
Slot 4
Slot 5
Slot 6
Slot 7
Slot 9
$k_L\hat{a}$ (1/h)
Time
2019-2020
Jan
Apr
Jul
Oct
$k_L\hat{a}$
k_La_{cw}
Airflow error
Flapping valve
Low DO excitation
>DO sat.

c) L2z3


Slot 2
Slot 4
Slot 5
Slot 6
Slot 7
Slot 9
$k_L\hat{a}$ (1/h)
Time
2019-2020
Jan
Apr
Jul
Oct
$k_L\hat{a}$
k_La_{cw}
Airflow error
Flapping valve
Low DO excitation
>DO sat.

d) L2z4

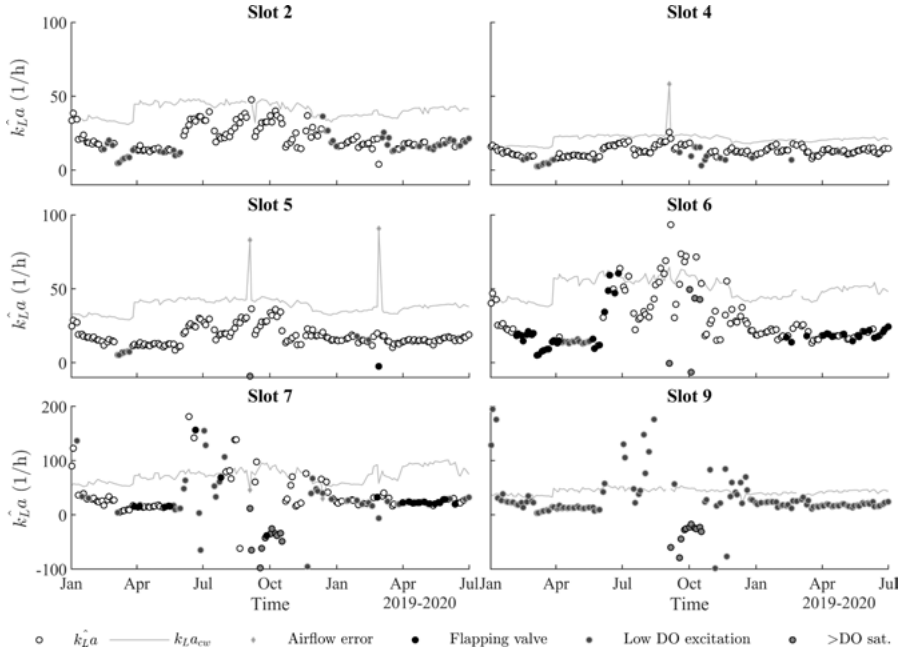


Figure 1. k_La estimates for line 1 and 2, zone 3 and 4 for the different slots in the staircase.

Bibliography

Aarnio, P. and Minkkinen, P. (1986) Application of partial least-squares modelling in the optimization of a waste-water treatment plant. Anal Chim Acta 191(C), 457-460.

Ahlström, M. (2018) Online-instrumentering på avloppsreningsverk: status idag och effekter av givarfel på reningsprocessen (Online sensors in wastewater treatment plants: status today and the effects of sensor faults on the treatment process (eng)), M.Sc. Thesis., Uppsala University, Sweden.

Amaral, A., Gillot, S., Garrido-Baserba, M., Filali, A., Karpinska, A.M., Plósz, B.G., de Groot, C., Bellandi, G., Nopens, I., Takács, I., Lizarralde, I., Jimenez, J.A., Fiat, J., Rieger, L., Arnell, M., Andersen, M., Jeppsson, U., Rehman, U., Fayolle, Y., Amerlinck, Y. and Rosso, D. (2019) Modelling gas–liquid mass transfer in wastewater treatment: When current knowledge needs to encounter engineering practice and vice versa. Water Science and Technology 80(4), 607-619.

Andersson, S. and Hallgren, F. (2015) Sensor fault detection methods applied on dissolved oxygen sensors at a full scale WWTP, Gold Coast, Australia.

Askarian, M., Escudero, G., Graells, M., Zarghami, R., Jalali-Farahani, F. and Mostoufi, N. (2016) Fault diagnosis of chemical processes with incomplete observations: A comparative study. Computers & Chemical Engineering 84, 104-116.

Atkinson, A.C. and Fedorov, V.V. (1975) The design of experiments for discriminating between two rival models. Biometrika 62(1), 57-70.

Ažman, K. and Kocijan, J. (2007) Application of Gaussian processes for black-box modelling of biosystems. ISA Transactions 46(4), 443-457.

Barnabé-Lortie, V., Bellinger, C. and Japkowicz, N. (2015) Active Learning for One-Class Classification, pp. 390-395.

Basseville, M. (1993) Detection of abrupt changes theory and application, Prentice-Hall, Englewood Cliffs, New Jersey.

Basseville, M. and Nikiforov, I.V. (1991) A unified framework for statistical change detection, pp. 2586-2591 vol.2583.

Beaupré, M. (2010) Characterization of on-line sensors for water quality monitoring and process control, Université Laval, Québec.

Benqlilou, C. (2004) Data Reconciliation as a Framework for Chemical Processes Optimization and Control. Ph.D Thesis, Universitat Politècnica de Catalunya.

Bezerra, M.A., Santelli, R.E., Oliveira, E.P., Villar, L.S. and Escaleira, L.A. (2008) Response surface methodology (RSM) as a tool for optimization in analytical chemistry. Talanta 76(5), 965-977.

Bishop, C.M. (2006) Pattern recognition and machine learning, Springer, United States of America.

Boškoski, P., Gašperin, M., Petelin, D. and Juričić, Đ. (2015) Bearing fault prognostics using Rényi entropy based features and Gaussian process models. Mechanical Systems and Signal Processing 52–53(0), 327-337.

Bouveyron, C. (2014) Adaptive Mixture Discriminant Analysis for Supervised Learning with Unobserved Classes. Journal of Classification 31(1), 49-84.
Box, G.E.P. and Behnken, D.W. (1960) Some New Three Level Designs for the Study of Quantitative Variables. Technometrics 2(4), 455-475.
Bro, R. and Smilde, A.K. (2014) Principal component analysis. Analytical Methods 6(9), 2812-2831.
Câmara, M.M., Soares, R.M., Feital, T., Anzai, T.K., Diehl, F.C., Thompson, P.H. and Pinto, J.C. (2017) Numerical aspects of data reconciliation in industrial applications. Processes 5(4).
Campbell, S.L., Horton, K.G. and Nikoukhah, R. (2002) Auxiliary signal design for rapid multi-model identification using optimization. Automatica 38(8), 1313-1325.
Carlsson, B. and Zambrano, J. (2016) Fault detection and isolation of sensors in aeration control systems. Water Science and Technology 73(3), 648-653.
Casella, G. and Berger, R.L. (2001) Statistical Inference, Duxbury Resource Center, United States of America.
Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y. and Ren, H. (2020) Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Research 171, 115454.
Clark, L.C. (1959) Electrochemical device for chemical analysis. U.S. patent No. 2,913,386A.
Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U. and Poch, M. (2017) Turning passive data into knowledge - a review of wastewater treatment monitoring techniques, Quebéc.
Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U. and Poch, M. (2018) Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental Modelling and Software 106, 89-103.
Corominas, L., Villez, K., Aguado, D., Rieger, L., Rosén, C. and Vanrolleghem, P.A. (2011) Performance evaluation of fault detection methods for wastewater treatment processes. Biotechnology and Bioengineering 108(2), 333-344.
Cressie, N. (1990) The origins of kriging. Mathematical Geology 22(3), 239-252.
Demas, J.N., DeGraff, B.A. and Coleman, P.B. (1999) Oxygen sensors based on luminescence quenching. Analytical Chemistry 71(23), 793A-800A.
Diehl, S., Zambrano, J. and Carlsson, B. (2016) Steady-state analysis of activated sludge processes with a settler model including sludge compression. Water Research 88, 104-116.
Dolenc, B., Stepancic, M., Juricic, D., Kocijan, J., Marra, D. and Pianese, C. (2016) Accounting for modelling errors in model-based diagnosis by using Gaussian process models, pp. 525-530.
Duvenaud., D.K. (2014) Automatic Model Construction with Gaussian Processes, University of Cambridge.
Eguchi, S. and Copas, J. (2006) Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma. Journal of Multivariate Analysis 97(9), 2034-2040.
EPA, U. (1989a) Fine Pore Aeration Systems Design Manual. EPA/625/1-89/023,, U. S. EPA, Office of Research and Development, Cincinnati, Ohio.
EPA, U.S. (1989b) Design Manual - Fine Pore Aeration Systems U.S. Environmental Protection Agency, Cincinnati, Ohio.

Farkas, J., Polesel, F., Kjos, M., Carvalho, P.A., Ciesielski, T., Flores-Alsina, X., Hansen, S.F. and Booth, A.M. (2020) Monitoring and modelling of influent patterns, phase distribution and removal of 20 elements in two primary wastewater treatment plants in Norway. Science of the Total Environment 725.
Fedorov, V. (2010) Optimal experimental design. Wiley Interdisciplinary Reviews: Computational Statistics 2(5), 581-589.
Fuente, M.J., Gutierrez, G., Gomez, E., Sarabia, D. and de Prada, C. (2015) Gross error management in data reconciliation. IFAC-PapersOnLine 48(8), 623-628.
Garnett, R., Osborne, M.A., Reece, S., Rogers, A. and Roberts, S.J. (2010a) Sequential bayesian prediction in the presence of changepoints and faults. Computer Journal 53(9), 1430-1446.
Garnett, R., Osborne, M.A., Reece, S., Rogers, A. and Roberts, S.J. (2010b) Sequential Bayesian Prediction in the Presence of Changepoints and Faults. The Computer Journal 53(9), 1430-1446.
Garrido-Baserba, M., Asvapathanagul, P., McCarthy, G.W., Gocke, T.E., Olson, B.H., Park, H.-D., Al-Omari, A., Murthy, S., Bott, C.B., Wett, B., Smeraldi, J.D., Shaw, A.R. and Rosso, D. (2016) Linking biofilm growth to fouling and aeration performance of fine-pore diffuser in activated sludge. Water Research 90, 317-328.
Garrido-Baserba, M., Asvapathanagul, P., Park, H.D., Kim, T.S., Baquero-Rodriguez, G.A., Olson, B.H. and Rosso, D. (2018) Impact of fouling on the decline of aeration efficiency under different operational conditions at WRRFs. Science of the Total Environment 639, 248-257.
Garrido-Baserba, M., Sobhani, R., Asvapathanagul, P., McCarthy, G.W., Olson, B.H., Odize, V., Al-Omari, A., Murthy, S., Nifong, A., Godwin, J., Bott, C.B., Stenstrom, M.K., Shaw, A.R. and Rosso, D. (2017) Modelling the link amongst fine-pore diffuser fouling, oxygen transfer efficiency, and aeration energy intensity. Water Research 111, 127-139.
Gernaey, K.V., Flores-Alsina, X., Rosen, C., Benedetti, L. and Jeppsson, U. (2011) Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. Environmental Modelling and Software 26(11), 1255-1267.
Gertler, J.J. (1988) Survey of Model-Based Failure Detection and Isolation in Complex Plants. IEEE Control Systems Magazine 8(6), 3-11.
Golub, G.H., Heath, M. and Wahba, G. (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2), 215-223.
Greggio, N., Bernardino, A., Laschi, C., Dario, P. and Santos-Victor, J. (2012) Fast estimation of Gaussian mixture models for image segmentation. Machine Vision and Applications 23(4), 773-789.
Grijspeerdt, K. and Verstraete, W. (1997) Image analysis to estimate the settleability and concentration of activated sludge. Water Research 31(5), 1126-1134.
Grol, R. and Wensing, M. (2004) What drives change? Barriers to and incentives for achieving evidence-based practice. Medical Journal of Australia 180(6 SUPPL.), S57-S60.
Gustafsson, F. (2000) Adaptive filtering and change detection, John Wiley & Sons Ltd, West Sussex, England.
Gustafsson, F. (2010) Particle filter theory and practice with positioning applications. IEEE Aerospace and Electronic Systems Magazine 25(7 PART 2), 53-81.

Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A. and Uhlmann, L. (2018) Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology 29(8), 1836-1842.
Haimi, H. (2016) Data-derived soft sensors in biological wastewater treatment, Aalto University, Finland.
Haimi, H., Corona, F., Mulas, M., Sundell, L., Heinonen, M. and Vahala, R. (2015) Shall we use hardware sensor measurements or soft-sensor estimates? Case study in a full-scale WWTP. Environmental Modelling and Software 72, 215-229.
Hauduc, H., Neumann, M.B., Muschalla, D., Gamerith, V., Gillot, S. and Vanrolleghem, P.A. (2015) Efficiency criteria for environmental model quality assessment: A review and its application to wastewater treatment. Environmental Modelling & Software 68, 196-204.
Hedegärd, M. and Wik, T. (2011) An online method for estimation of degradable substrate and biomass in an aerated activated sludge process. Water Research 45(19), 6308-6320.
Heller, J. (1961) Catch-22, The Modern library, New York
Henze, M., Grady Jr, C.P.L., Gujer, W., Marais, G.V.R. and Matsuo, T. (1987) A general model for single-sludge wastewater treatment systems. Water Research 21(5), 505-515.
Henze, M., Harremoes, P., Jansen, J.l.C. and Arvin, E. (2002) Wastewater treatment, biological and chemical processes, Springer-Verlag Berlin, Heidelberg, New York.
Himmelblau, D.M. (1978) Fault detection and diagnosis in chemical and petrochemical processes, Elsevier Scientific Pub. Co. ; distributors for the US and Canada, Elsevier North-Holland, Amsterdam ; New York : New York.
Holmberg, U. and Olsson, G. (1985) Simultaneous On-Line Estimation of Oxygen Transfer Rate and Respiration Rate. IFAC Proceedings Volumes 18(17), 205-209.
Holmberg, U., Olsson, G. and Andersson, B. (1989) Simultaneous DO control and respiration estimation. Water Science and Technology 21(10-11 -11 pt 4), 1185-1195.
Hsu, L.H.H. and Selvaganapathy, P.R. (2013) Development of a low cost Hemin based dissolved oxygen sensor with anti-biofouling coating for water monitoring.
Ingildsen, P., Jeppsson, U. and Olsson, G. (2002) Dissolved oxygen controller based on on-line measurements of ammonium combining feed-forward and feedback. Water Science and Technology 45(4-5), 453-460.
International Organization for Standardization (2012) Water quality - Determination of dissolved oxygen - Electrochemical probe method (ISO 5814:2012).
Irizar, I., Alferes, J., Larrea, L. and Ayesa, E. (2008) Standard signal processing using enriched sensor information for WWTP monitoring and control, pp. 1053-1060.
Irizar, I., Zambrano, J.A., Montoya, D., De Gracia, M. and García, R. (2009) Online monitoring of OUR, KLa and OTE indicators: Practical implementation in full-scale industrial WWTPs, pp. 459-466.
Isermann, R. and Ballé, P. (1997) Supervision, fault-detection and fault-diagnosis methods — An introduction. Control Engineering Practice 5(5), 639-652.
Janzen, C., Murphy, D. and Larson, N. (2007) Getting more mileage out of dissolved oxygen sensors in long-term moored applications.
Jeppsson, U., Rosén, C., Alex, J., Copp, J., Gernaey, K.V., Pons, M.-N. and Vanrolleghem, P.A. (2006) Towards a benchmark simulation model for plant-wide control strategy performance evaluation of WWTPs. Water Science and Technology 53(1), 287-295.

Jiang, L.-M., Chen, L., Zhou, Z., Sun, D., Li, Y., Zhang, M., Liu, Y., Du, S., Chen, G. and Yao, J. (2020) Fouling characterization and aeration performance recovery of fine-pore diffusers operated for 10 years in a full-scale wastewater treatment plant. Bioresource Technology 307, 123197.
Jiang, L.-M., Garrido-Baserba, M., Nolasco, D., Al-Omari, A., DeClippeleir, H., Murthy, S. and Rosso, D. (2017) Modelling oxygen transfer using dynamic alpha factors. Water Research 124, 139-148.
Jiang, Q., Huang, B. and Yan, X. (2016) GMM and optimal principal components-based Bayesian method for multimode fault diagnosis. Computers and Chemical Engineering 84, 338-349.
Južnič-Zonta, Ž., Kocijan, J., Flotats, X. and Vrečko, D. (2012) Multi-criteria analyses of wastewater treatment bio-processes under an uncertainty and a multiplicity of steady states. Water Research 46(18), 6121-6131.
Karpinska, A.M. and Bridgeman, J. (2016) CFD-aided modelling of activated sludge systems - A critical review. Water Research 88, 861-879.
Kay, S.M. (1993) Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall PTR, Upper Saddle, NJ, USA
Kay, S.M. (1998) Fundamentals of Statistical Signal Processing: Detection theory, Prentice-Hall PTR, Upper Saddle River, NJ, USA.
Kocijan, J. and Hvala, N. (2013) Sequencing batch-reactor control using Gaussian-process models. Bioresource Technology 137(0), 340-348.
Kohavi, R. and Provost, F. (1998) Glossary of Terms. Machine Learning 30(2), 271-274.
Kourti, T. and MacGregor, J.F. (1995) Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemometrics and Intelligent Laboratory Systems 28(1), 3-21.
Krampe, J. (2011) Full scale evaluation of diffuser ageing with clean water oxygen transfer tests. Water Science and Technology 64(3), 700-707.
Kulin, G. (1984) Recommended Practice for the Use of Parshall Flumes and Palmer-Bowlus Flumes in Wastewater Treatment Plants, National Bureau of Standards (NEL), U.S. Environmental protection agency, Cincinnati, OHIO, USA.
Käppalaförbundet (2018) Miljörapport (eng. Annual emission declaration report).
Larsen, T.A. (2015) CO2-neutral wastewater treatment plants or robust, climate-friendly wastewater management? A systems perspective. Water Research 87, 513-521.
Le, H.Q. (2019) Mass-Balance-based Experimental Design and Data Reconciliation for Wastewater Treatment Processes. Ph.D., Gent University, Gent.
Le, Q.H., Verheijen, P.J.T., van Loosdrecht, M.C.M. and Volcke, E.I.P. (2018) Experimental design for evaluating WWTP data by linear mass balances. Water Research 142, 415-425.
Lee, J.M., Yoo, C.K., Choi, S.W., Vanrolleghem, P.A. and Lee, I.B. (2004) Nonlinear process monitoring using kernel principal component analysis. Chemical Engineering Science 59(1), 223-234.
Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N. and Nandi, A.K. (2020) Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing 138, 106587.
Leu, S.Y., Libra, J.A. and Stenstrom, M.K. (2010) Monitoring off-gas O2/CO2 to predict nitrification performance in activated sludge processes. Water Res 44(11), 3434-3444.
Li, B. and Stenstrom, M.K. (2014) Research advances and challenges in one-dimensional modeling of secondary settling Tanks – A critical review. Water Research 65(0), 40-63.

Li, B., Udugama, I.A., Mansouri, S.S., Yu, W., Baroutian, S., Gernaey, K.V. and Young, B.R. (2019) An exploration of barriers for commercializing phosphorus recovery technologies. Journal of Cleaner Production 229, 1342-1354.
Li, T., Winnel, M., Lin, H., Panther, J., Liu, C., O'Halloran, R., Wang, K., An, T., Wong, P.K., Zhang, S. and Zhao, H. (2017) A reliable sewage quality abnormal event monitoring system. Water Research 121, 248-257.
Lid, T. and Skogestad, S. (2008) Data reconciliation and optimal operation of a catalytic naphtha reformer. Journal of Process Control 18(3-4), 320-331.
Lindberg, C.-F. (1997) Control and estimation strategies applied to the activated sludge process, Uppsala University.
Lindblom, E., Jeppsson, U. and Sin, G. (2019) Identification of behavioural model input data sets for WWTP uncertainty analysis. Water Science and Technology 81(8), 1558-1568.
Liu, Y., Chen, T. and Chen, J. (2015) Auto-Switch Gaussian Process Regression-Based Probabilistic Soft Sensors for Industrial Multigrade Processes with Transitions. Industrial & Engineering Chemistry Research 54(18), 5037-5047.
Liu, Y., Xiao, H., Pan, Y., Huang, D. and Wang, Q. (2016) Development of multiple-step soft-sensors using a Gaussian process model with application for fault prognosis. Chemometrics and Intelligent Laboratory Systems 157, 85-95.
Ljung, L. (1999) System identification: theory for the user, Prentice Hall PTR.
Lloyd, J.R., Duvenaud, D., Grosse, R., Tenenbaum, J.B. and Ghahramani, Z. (2014) Automatic construction and natural-language description of nonparametric regression models, pp. 1242-1250.
Loosdrecht, M.C.M.v., Nielsen, P.H., Lopez-Vazquez, C.M. and Brdjanovic, D. (2016) Experimental Methods in Wastewater Treatment.
Lu, H., Chandran, K. and Stensel, D. (2014) Microbial ecology of denitrification in biological wastewater treatment. Water Research 64, 237-254.
Lumley, D. (2002) On-line instrument confirmation: how can we check that our instruments are working? Water Sci Technol 45(4-5), 469-476.
Ma, J. and Jiang, J. (2011) Applications of fault detection and diagnosis methods in nuclear power plants: A review. Progress in Nuclear Energy 53(3), 255-266.
Marimuthu, K. and Narasimhan, S. (2019) Nonlinear Model Identification and Data Reconciliation Using Kernel Principal Component Regression. Industrial & Engineering Chemistry Research 58(26), 11224-11233.
Martin, C. and Vanrolleghem, P.A. (2014) Analysing, completing, and generating influent data for WWTP modelling: A critical review. Environmental Modelling and Software 60, 188-201.
Md Nor, N., Che Hassan, C.R. and Hussain, M.A. (2020) A review of data-driven fault detection and diagnosis methods: Applications in chemical process systems. Reviews in Chemical Engineering 36(4), 513-553.
Metcalf and Eddy (2004) Wastewater engineering, treatment and reuse, McGraw-Hill, New York.
Moles, C.G., Mendes, P. and Banga, J.R. (2003) Parameter estimation in biochemical pathways: A comparison of global optimization methods. Genome Research 13(11), 2467-2474.
Moreno, M., Liu, J., Su, Q., Leach, C., Giridhar, A., Yazdanpanah, N., O'Connor, T., Nagy, Z.K. and Reklaitis, G.V. (2019) Steady-State Data Reconciliation Framework for a Direct Continuous Tableting Line. Journal of Pharmaceutical Innovation 14(3), 221-238.
Murphy, K. (2012) Machine learning: a probabilistic perspective, MIT Press, Cambridge, London.

Myers, R.H., Montgomery, D.C., Geoffrey Vining, G., Borror, C.M. and Kowalski, S.M. (2004) Response Surface Methodology: A Retrospective and Literature Survey. Journal of Quality Technology 36(1), 53-78.
Narasimhan, S. and Jordache, C. (2000) Data Reconciliation & Gross Error Detection: An Intelligent Use of Process Data, Gulf Publishing Company.
Naseer, S., Saleem, Y., Khalid, S., Bashir, M.K., Han, J., Iqbal, M.M. and Han, K. (2018) Enhanced network anomaly detection based on deep neural networks. IEEE Access 6, 48231-48246.
Naturvårdsverket (2018) Statistiskt meddelande, MI22 SM 2001, Utsläpp till vatten och slamproduktion.
Nesbitt, B. (2007) Handbook of Valves and Actuators.
Newhart, K.B., Holloway, R.W., Hering, A.S. and Cath, T.Y. (2019) Data-driven performance analyses of wastewater treatment plants: A review. Water Research 157, 498-513.
Ni, W., Wang, K., Chen, T., Ng, W.J. and Tan, S.K. (2012) GPR model with signal preprocessing and bias update for dynamic processes modeling. Control Engineering Practice 20(12), 1281-1292.
Nierychlo, M., Andersen, K.S., Xu, Y., Green, N., Jiang, C., Albertsen, M., Dueholm, M.S. and Nielsen, P.H. (2020) MiDAS 3: An ecosystem-specific reference database, taxonomy and knowledge platform for activated sludge and anaerobic digesters reveals species-level microbiome composition of activated sludge. Water Research 182, 115955.
Nikoukhah, R. and Campbell, S.L. (2008) On the detection of small parameter variations in linear uncertain systems. European Journal of Control 14(2), 158-171.
Odize, V.O., Novak, J., De Clippeleir, H., Al-Omari, A., Smeraldi, J.D., Murthy, S. and Rosso, D. (2017) Reverse flexing as a physical/mechanical treatment to mitigate fouling of fine bubble diffusers. Water Science and Technology 76(7), 1595-1602.
Ohmura, K., Thürlimann, C.M., Kipf, M., Carbajal, J.P. and Villez, K. (2019) Characterizing long-term wear and tear of ion-selective pH sensors. Water Science and Technology 80(3), 541-550.
Olsson, G. (2006) Instrumentation, control and automation in the water industry – state-of-the-art and new challenges. Water Science and Technology 53(4-5), 1-16.
Olsson, G. (2012) ICA and me - A subjective review. Water Research 46(6), 1585-1624.
Olsson, G., Andersson, B., Göran Hellstrom, B., Holmström, H., Reinius, L.G. and Vopatek, P. (1989) Measurements, Data Analysis and Control Methods in Wastewater Treatment Plants–State of the Art and Future Trends. Water Science and Technology 21(10-11), 1333-1345.
Olsson, G., Carlsson, B., Comas, J., Copp, J., Gernaey, K.V., Ingildsen, P., Jeppsson, U., Kim, C., Rieger, L., Rodríguez-Roda, I., Steyer, J.-P., Takács, I., Vanrolleghem, P.A., Vargas, A., Yuan, Z. and Åmand, L. (2014) Instrumentation, control and automation in wastewater – from London 1973 to Narbonne 2013. Water Science & Technology 69(7), 1373-1385.
Olsson, G., Nielsen, M., Yuan, Z., Lynggaard-Jensen, A. and Steyer, J.P. (2005) Instrumentation, control and automation in wastewater systems, IWA Publishing.
Padilla, M. and Choinière, D. (2015) A combined passive-active sensor fault detection and isolation approach for air handling units. Energy and Buildings 99, 214-219.
Perez-Cruz, F., Van Vaerenbergh, S., Murillo-Fuentes, J.J., Lazaro-Gredilla, M. and Santamaria, I. (2013) Gaussian processes for nonlinear signal processing: An overview of recent advances. IEEE Signal Processing Magazine 30(4), 40-50.

Petersen, L. and Esbensen, K.H. (2005) Representative process sampling for reliable data analysis - A tutorial. Journal of Chemometrics 19(11-12), 625-647.
Petersson, M., Årzén, K.E., Sandberg, H. and De Maré, L. (2002) Implementation of a tool for control structure assessment, pp. 435-440.
Pimentel, M.A.F., Clifton, D.A., Clifton, L. and Tarassenko, L. (2014) A review of novelty detection. Signal Processing 99, 215-249.
Piñeiro, G., Perelman, S., Guerschman, J.P. and Paruelo, J.M. (2008) How to evaluate models: Observed vs. predicted or predicted vs. observed? Ecological Modelling 216(3-4), 316-322.
Puig, S., van Loosdrecht, M.C.M., Colprim, J. and Meijer, S.C.F. (2008) Data evaluation of full-scale wastewater treatment plants by mass balance. Water Research 42(18), 4645-4655.
Puncochár, I. and Skach, J. (2018) A Survey of Active Fault Diagnosis Methods, International Federation of Automatic Control, Warsaw, Poland.
Qin, S.J. (2012) Survey on data-driven industrial process monitoring and diagnosis. Annual Reviews in Control 36(2), 220-234.
Qin, X., Gao, F. and Chen, G. (2012) Wastewater quality monitoring system using sensor fusion and machine learning techniques. Water Research 46(4), 1133-1144.
Quiñonero-Candela, J. and Rasmussen, C.E. (2005) A unifying view of sparse approximate Gaussian process regression. Journal of machine learning research 6, 1939-1959.
Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simón, P., Allende, A. and Sánchez, G. (2020) SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. Water Research 181, 115942.
Ranjan, R., Huang, B. and Fatehi, A. (2016) Robust Gaussian process modeling using em algorithm. Journal of Process Control 42, 125-136.
Rasmussen, C.E. and Nickisch, H. (2010) Gaussian Processes for Machine Learning (GPML) Toolbox. Journal of machine learning research 11, 3011-3015.
Rasmussen, C.E. and Williams, C.K.I. (2005) Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, Cambridge, Massachusetts
London, England.
Redmon, D., Boyle, W.C. and Ewing, L. (1983) Oxygen Transfer Efficiency Measurements in Mixed Liquor Using Off-Gas Techniques. Journal (Water Pollution Control Federation) 55(11), 1338-1347.
Ren, S. (2004) Assessing wastewater toxicity to activated sludge: Recent research and developments. Environment International 30(8), 1151-1164.
Rieger, L., Jones, R.M., Dold, P.L. and Bott, C.B. (2014) Ammonia-based feedforward and feedback aeration control in activated sludge processes. Water Environment Research 86(1), 63-73.
Rieger, L., Sylvie Gillot, Günter Langergraber, Takayuki Ohtsuki, Andrew Shaw, Imre Takács and Stefan Winkler (2012) Guidelines for Using Activated Sludge Models, IWA Publishing, London, UK.
Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. and Aigrain, S. (2013) Gaussian processes for time-series modelling.
Romagnoli, J.A. and Sanchez, M.C. (1999) Data Processing and Reconciliation for Chemical Process Operations, Elsevier Science.
Rosén, C. (1998) Monitoring Wastewater Treatment Systems, Lunds University, Lund Sweden.
Rosén, C. (2001) A Chemometric Approach to Process Monitoring and Control - With Applications to Wastewater Treatment Operation. PhD thesis, Lund university.

Rosén, C., Rieger, L., Jeppsson, U. and Vanrolleghem, P.A. (2008) Adding realism to simulated sensors and actuators. Water Science and Technology 57(3), 337-344.
Rosén, C., Röttorp, J. and Jeppsson, U. (2003) Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation. Water Science and Technology 47(2), 171-179.
Rossi, L., Rumley, L., Ort, C., Minkkinen, P., Barry, D.A. and Chevre, N. (2011) Samplinghelper a web-based tool to assess the reliability of sampling strategies in sewers and receiving waters. Water Sci Technol 63(12), 2975-2982.
Rosso, D. (ed) (2018) Aeration fundamentals, performance and monitoring. In: Aeration, Mixing, and Energy: Bubbles and Sparks, IWA Publishing, London, UK.
Rosso, D., Libra, J.A., Wiehe, W. and Stenstrom, M.K. (2008) Membrane properties change in fine-pore aeration diffusers: Full-scale variations of transfer efficiency and headloss. Water Research 42(10-11), 2640-2648.
Rosso, D. and Stenstrom, M.K. (2006) Economic implications of fine-pore diffuser-aging. Water Environment Research 78(8), 810-815.
Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20(C), 53-65.
Russo, S., Lürig, M., Hao, W., Matthews, B. and Villez, K. (2020) Active learning for anomaly detection in environmental data. Environmental Modelling and Software 134.
Saagi, R., Flores-Alsina, X., Kroll, S., Gernaey, K.V. and Jeppsson, U. (2017) A model library for simulation and benchmarking of integrated urban wastewater systems. Environmental Modelling & Software 93, 282-295.
Sabokrou, M., Khalooei, M., Fathy, M. and Adeli, E. (2018) Adversarially Learned One-Class Classifier for Novelty Detection, pp. 3379-3388.
Samstag, R.W., Ducoste, J.J., Griborio, A., Nopens, I., Batstone, D.J., Wicks, J.D., Saunders, S., Wicklein, E.A., Kenny, G. and Laurent, J. (2016) CFD for wastewater treatment: An overview. Water Science and Technology 74(3), 549-563.
Samuelsson, O. (2014) A feasibility study - Fault detection in on-line sensors at Swedish wastewater treatment plants. Internal report.
Samuelsson, O. (2017) Fault Detection in Water Resource Recovery Facilities. Licentiate Thesis, Uppsala University, Uppsala.
Samuelsson, O., Björk, A., Zambrano, J. and Carlsson, B. (2017) Gaussian process regression for monitoring and fault detection of wastewater treatment processes. Water Science and Technology 75(12), 2952-2963.
Samuelsson, O., Björk, A., Zambrano, J. and Carlsson, B. (2018) Fault signatures and bias progression in dissolved oxygen sensors. Water Science and Technology 78(5-6), 1034-1044.
Samuelsson, O., Zambrano, J., Björk, A. and Carlsson, B. (2019) Automated active fault detection in fouled dissolved oxygen sensors. Water Research 166.
Santín, I., Pedret, C., Vilanova, R. and Meneses, M. (2016) Advanced decision control system for effluent violations removal in wastewater treatment plants. Control Engineering Practice 49, 60-75.
Saunders, A.M., Albertsen, M., Vollertsen, J. and Nielsen, P.H. (2016) The activated sludge ecosystem contains a core community of abundant organisms. ISME Journal 10(1), 11-20.
Schraa, O., Tole, B. and Copp, J.B. (2006) Fault detection for control of wastewater treatment plants. Water Science and Technology 53(4-5), 375-382.

Schön, T.B., Lindsten, F., Dahlin, J., Wågberg, J., Naesseth, C.A., Svensson, A. and Dai, L. (2015) Sequential Monte Carlo methods for system identification, Beijing, China.
Sedlack, R. (ed) (1991) Phosphorus and Nitrogen Removal from Municipal Wastewater: Principles and Practice (2nd ed.), CRC Press.
Serradilla, J., Shi, J.Q. and Morris, A.J. (2011) Fault detection based on Gaussian process latent variable models. Chemometrics and Intelligent Laboratory Systems 109(1), 9-21.
Shaffer, J.P. (1995) Multiple Hypothesis Testing. Annual Review of Psychology 46(1), 561-584.
Shang, C., Yang, F., Huang, D. and Lyu, W. (2014) Data-driven soft sensor development based on deep learning technique. Journal of Process Control 24(3), 223-233.
Shewhart, W.A. (1930) Economic Quality Control of Manufactured Product. Bell System Technical Journal 9(2), 364-389.
Showkatbakhsh, M., Shoukry, Y., Diggavi, S.N. and Tabuada, P. (2020) Securing state reconstruction under sensor and actuator attacks: Theory and design. Automatica 116.
Smith, M., Reece, S., Roberts, S. and Rezek, I. (2012) Online Maritime Abnormality Detection Using Gaussian Processes and Extreme Value Theory, pp. 645-654, IEEE.
Sohlberg, B. and Jacobsen, E. (2008) Grey box modelling branches and experiences.
Spanjers, H. and Olsson, G. (1992) Modelling of the dissolved oxygen probe response in the improvement of the performance of a continous respirationmeter. Wat. Res. 26(7), 945-954.
Spindler, A. (2014) Structural redundancy of data from wastewater treatment systems. Determination of individual balance equations. Water Res 57, 193-201.
Stockholm vatten och avfall (2019) Miljörapport (eng. Annual emission declaration).
Svensson., A., Dahlin., J. and Schön., T.B. (2015) Marginalizing Gaussian process hyperparameters using sequential Monte Carlo, Cancun, Mexico.
Söderström, T. and Stoica, P. (1989) System identification, Prentice-Hall, Inc.
Thürlimann, C.M., Dürrenmatt, D.J. and Villez, K. (2018) Soft-sensing with qualitative trend analysis for wastewater treatment plant control. Control Engineering Practice 70, 121-133.
Thürlimann, C.M., Udert, K.M., Morgenroth, E. and Villez, K. (2019) Stabilizing control of a urine nitrification process in the presence of sensor drift. Water Research 165, 114958.
Tian, L. (2010) Confidence interval estimation of partial area under curve based on combined biomarkers. Computational Statistics & Data Analysis 54(2), 466-472.
Traoré, A., Grieu, S., Thiery, F., Polit, M. and Colprim, J. (2006) Control of sludge height in a secondary settler using fuzzy algorithms. Computers & Chemical Engineering 30(8), 1235-1242.
Wade, M.J., Sánchez, A. and Katebi, M.R. (2005) On real-time control and process monitoring of wastewater treatment plants: Real-time process monitoring. Transactions of the Institute of Measurement & Control 27(3), 173-193.
Wang, M., Mo, H., Liu, G.H., Qi, L., Yu, Y., Fan, H., Xu, X., Luo, T., Shao, Y. and Wang, H. (2020) Impact of scaling on aeration performance of fine-pore membrane diffusers based on a pilot-scale study. Scientific Reports 10(1).
Wang, Q., Chen, Q. and Chen, J. (2017) Optimizing external carbon source addition in domestics wastewater treatment based on online sensoring data and a numerical model. Water Science and Technology 75(11), 2716-2725.

Varmuza, K. and Filzmoser, P. (2009) Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Taylor and Francis, Boca Raton US.
WEF (2013) Automation of Water Resource Recovery Facilities: WEF Manual of Practice No. 21. 4th ed., Water Environment Federation, Alexandria, USA.
Venkatasubramanian, V., Rengaswamy, R., Yin, K. and Kavuri, S.N. (2003) A review of process fault detection and diagnosis. Computers & Chemical Engineering 27(3), 293-311.
Wicklein, E., Batstone, D.J., Ducoste, J., Laurent, J., Griborio, A., Wicks, J., Saunders, S., Samstag, R., Potier, O. and Nopens, I. (2016) Good modelling practice in applying computational fluid dynamics for WWTP modelling. Water Science and Technology 73(5), 969-982.
Wilén, B.M., Jin, B. and Lant, P. (2003) Relationship between flocculation of activated sludge and composition of extracellular polymeric substances, pp. 95-103.
Willis, J.L., Al-Omari, A., Bastian, R., Brower, B., DeBarbadillo, C., Murthy, S., Peot, C. and Yuan, Z. (2017) A greenhouse gas source of surprising significance: Anthropogenic CO2 emissions from use of methanol in sewage treatment. Water Science and Technology 75(9), 1997-2012.
Willsky, A.S. (1976) A survey of design methods for failure detection in dynamic systems. Automatica 12(6), 601-611.
Wilson, A.G. (2013) Gaussian process kernels for pattern discovery and extrapolation, pp. 2104-2112.
Xanthos, S., Gong, M., Ramalingam, K., Fillos, J., Deur, A., Beckmann, K. and McCorquodale, J.A. (2011) Performance Assessment of Secondary Settling Tanks Using CFD Modeling. Water Resources Management 25(4), 1169-1182.
Xie, S., Yang, C., Yuan, X., Wang, X. and Xie, Y. (2018) Layered online data reconciliation strategy with multiple modes for industrial processes. Control Engineering Practice 77, 63-72.
Ye, Z., Yang, J., Zhong, N., Tu, X., Jia, J. and Wang, J. (2020) Tackling environmental challenges in pollution controls using artificial intelligence: A review. Science of the Total Environment 699.
Yoo, C., Choi, S. and Lee, I.-B. (2002) Adaptive modeling and classification of the secondary settling tank. Korean Journal of Chemical Engineering 19(3), 377-382.
Yoo, C.K., Villez, K., Lee, I.B., Rosén, C. and Vanrolleghem, P.A. (2007) Multi-model statistical process monitoring and diagnosis of a sequencing batch reactor. Biotechnology and Bioengineering 96(4), 687-701.
Yoo, C.K., Villez, K., Van Hulle, S.W.H. and Vanrolleghem, P.A. (2008) Enhanced process monitoring for wastewater treatment systems. Environmetrics 19(6), 602-617.
Yu, J. (2012) A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. Chemical Engineering Science 68(1), 506-519.
Yuan, Z., Olsson, G., Cardell-Oliver, R., van Schagen, K., Marchi, A., Deletic, A., Urich, C., Rauch, W., Liu, Y. and Jiang, G. (2019) Sweating the assets – The role of instrumentation, control and automation in urban water systems. Water Research 155, 381-402.
Yue, H.H. and Qin, S.J. (2001) Reconstruction-based fault identification using a combined index. Industrial and Engineering Chemistry Research 40(20), 4403-4414.
Zhu, H., Chen, S. and Han, C. (2014) Fusion of Gaussian mixture models for possible mismatches of sensor model. Information Fusion 20(1), 203-212.
Åmand, L. (2014) Ammonium Feedback Control in Wastewater Treatment Plants, Uppsala University, Sweden.

Åmand, L. and Carlsson, B. (2013) The optimal dissolved oxygen profile in a nitrifying activated sludge process – comparisons with ammonium feedback control. Water Science and Technology 68(3), 641-649.
Åmand, L., Olsson, G. and Carlsson, B. (2013) Aeration control - A review. Water Science and Technology 67(11), 2374-2398.
Åström, K.-J. and Hägglund, T. (1995) PID Controllers: Theory, Design, and Tuning, Instrument Society of America, North Carolina, Research Triangle Park.
Özyurt, D.B. and Pike, R.W. (2004) Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. Computers and Chemical Engineering 28(3), 381-402.

Acta Universitatis Upsaliensis
Uppsala Dissertations from the Faculty of Science
Editor: The Dean of the Faculty of Science

1–11: 1970–1975

12. *Lars Thofelt*: Studies on leaf temperature recorded by direct measurement and by thermography. 1975.
13. *Monica Henricsson*: Nutritional studies on Chara globularis Thuill., Chara zeylanica Willd., and Chara haitensis Turpin. 1976.
14. *Göran Kloow*: Studies on Regenerated Cellulose by the Fluorescence Depolarization Technique. 1976.
15. *Carl-Magnus Backman*: A High Pressure Study of the Photolytic Decomposition of Azoethane and Propionyl Peroxide. 1976.
16. *Lennart Källströmer*: The significance of biotin and certain monosaccharides for the growth of Aspergillus niger on rhamnose medium at elevated temperature. 1977.
17. *Staffan Renlund*: Identification of Oxytocin and Vasopressin in the Bovine Adenohypophysis. 1978.
18. *Bengt Finnström*: Effects of pH, Ionic Strength and Light Intensity on the Flash Photolysis of L-tryptophan. 1978.
19. *Thomas C. Amu*: Diffusion in Dilute Solutions: An Experimental Study with Special Reference to the Effect of Size and Shape of Solute and Solvent Molecules. 1978.
20. *Lars Tegnér*: A Flash Photolysis Study of the Thermal Cis-Trans Isomerization of Some Aromatic Schiff Bases in Solution. 1979.
21. *Stig Tormod*: A High-Speed Stopped Flow Laser Light Scattering Apparatus and its Application in a Study of Conformational Changes in Bovine Serum Albumin. 1985.
22. *Björn Varnestig*: Coulomb Excitation of Rotational Nuclei. 1987.
23. *Frans Lettenström*: A study of nuclear effects in deep inelastic muon scattering. 1988.
24. *Göran Ericsson*: Production of Heavy Hypernuclei in Antiproton Annihilation. Study of their decay in the fission channel. 1988.
25. *Fang Peng*: The Geopotential: Modelling Techniques and Physical Implications with Case Studies in the South and East China Sea and Fennoscandia. 1989.
26. *Md. Anowar Hossain*: Seismic Refraction Studies in the Baltic Shield along the Fennolora Profile. 1989.
27. *Lars Erik Svensson*: Coulomb Excitation of Vibrational Nuclei. 1989.
28. *Bengt Carlsson*: Digital differentiating filters and model based fault detection. 1989.
29. *Alexander Edgar Kavka*: Coulomb Excitation. Analytical Methods and Experimental Results on even Selenium Nuclei. 1989.
30. *Christopher Juhlin*: Seismic Attenuation, Shear Wave Anisotropy and Some Aspects of Fracturing in the Crystalline Rock of the Siljan Ring Area, Central Sweden. 1990.

31. *Torbjörn Wigren*: Recursive Identification Based on the Nonlinear Wiener Model. 1990.
32. *Kjell Janson*: Experimental investigations of the proton and deuteron structure functions. 1991.
33. *Suzanne W. Harris*: Positive Muons in Crystalline and Amorphous Solids. 1991.
34. *Jan Blomgren*: Experimental Studies of Giant Resonances in Medium-Weight Spherical Nuclei. 1991.
35. *Jonas Lindgren*: Waveform Inversion of Seismic Reflection Data through Local Optimisation Methods. 1992.
36. *Liqi Fang*: Dynamic Light Scattering from Polymer Gels and Semidilute Solutions. 1992.
37. *Raymond Munier*: Segmentation, Fragmentation and Jostling of the Baltic Shield with Time. 1993.

Prior to January 1994, the series was called *Uppsala Dissertations from the Faculty of Science*.

Acta Universitatis Upsaliensis

Uppsala Dissertations from the Faculty of Science and Technology
Editor: The Dean of the Faculty of Science

1–14: 1994–1997. 15–21: 1998–1999. 22–35: 2000–2001. 36–51: 2002–2003.

52. *Erik Larsson*: Identification of Stochastic Continuous-time Systems. Algorithms, Irregular Sampling and Cramér-Rao Bounds. 2004.
53. *Per Åhgren*: On System Identification and Acoustic Echo Cancellation. 2004.
54. *Felix Wehrmann*: On Modelling Nonlinear Variation in Discrete Appearances of Objects. 2004.
55. *Peter S. Hammerstein*: Stochastic Resonance and Noise-Assisted Signal Transfer. On Coupling-Effects of Stochastic Resonators and Spectral Optimization of Fluctuations in Random Network Switches. 2004.
56. *Esteban Damián Avendaño Soto*: Electrochromism in Nickel-based Oxides. Coloration Mechanisms and Optimization of Sputter-deposited Thin Films. 2004.
57. *Jenny Öhman Persson*: The Obvious & The Essential. Interpreting Software Development & Organizational Change. 2004.
58. *Chariklia Rouki*: Experimental Studies of the Synthesis and the Survival Probability of Transactinides. 2004.
59. *Emad Abd-Elrady*: Nonlinear Approaches to Periodic Signal Modeling. 2005.
60. *Marcus Nilsson*: Regular Model Checking. 2005.
61. *Pritha Mahata*: Model Checking Parameterized Timed Systems. 2005.
62. *Anders Berglund*: Learning computer systems in a distributed project course: The what, why, how and where. 2005.
63. *Barbara Piechocinska*: Physics from Wholeness. Dynamical Totality as a Conceptual Foundation for Physical Theories. 2005.
64. *Pär Samuelsson*: Control of Nitrogen Removal in Activated Sludge Processes. 2005.

65. *Mats Ekman*: Modeling and Control of Bilinear Systems. Application to the Activated Sludge Process. 2005.
66. *Milena Ivanova*: Scalable Scientific Stream Query Processing. 2005.
67. *Zoran Radovic´*: Software Techniques for Distributed Shared Memory. 2005.
68. *Richard Abrahamsson*: Estimation Problems in Array Signal Processing, System Identification, and Radar Imagery. 2006.
69. *Fredrik Robelius*: Giant Oil Fields – The Highway to Oil. Giant Oil Fields and their Importance for Future Oil Production. 2007.
70. *Anna Davour*: Search for low mass WIMPs with the AMANDA neutrino telescope. 2007.
71. *Magnus Ågren*: Set Constraints for Local Search. 2007.
72. *Ahmed Rezine*: Parameterized Systems: Generalizing and Simplifying Automatic Verification. 2008.
73. *Linda Brus*: Nonlinear Identification and Control with Solar Energy Applications. 2008.
74. *Peter Nauclér*: Estimation and Control of Resonant Systems with Stochastic Disturbances. 2008.
75. *Johan Petrini*: Querying RDF Schema Views of Relational Databases. 2008.
76. *Noomene Ben Henda*: Infinite-state Stochastic and Parameterized Systems. 2008.
77. *Samson Keleta*: Double Pion Production in dd→αππ Reaction. 2008.
78. *Mei Hong*: Analysis of Some Methods for Identifying Dynamic Errors-invariables Systems. 2008.
79. *Robin Strand*: Distance Functions and Image Processing on Point-Lattices With Focus on the 3D Face-and Body-centered Cubic Grids. 2008.
80. *Ruslan Fomkin*: Optimization and Execution of Complex Scientific Queries. 2009.
81. *John Airey*: Science, Language and Literacy. Case Studies of Learning in Swedish University Physics. 2009.
82. *Arvid Pohl*: Search for Subrelativistic Particles with the AMANDA Neutrino Telescope. 2009.
83. *Anna Danielsson*: Doing Physics – Doing Gender. An Exploration of Physics Students' Identity Constitution in the Context of Laboratory Work. 2009.
84. *Karin Schönning*: Meson Production in pd Collisions. 2009.
85. *Henrik Petrén*: η Meson Production in Proton-Proton Collisions at Excess Energies of 40 and 72 MeV. 2009.
86. *Jan Henry Nyström*: Analysing Fault Tolerance for ERLANG Applications. 2009.
87. *John Håkansson*: Design and Verification of Component Based Real-Time Systems. 2009.
88. *Sophie Grape*: Studies of PWO Crystals and Simulations of the $\bar{p}p \to \bar{\Lambda}\Lambda, \bar{\Lambda}\Sigma^0$ Reactions for the PANDA Experiment. 2009.
90. *Agnes Rensfelt*. Viscoelastic Materials. Identification and Experiment Design. 2010.
91. *Erik Gudmundson*. Signal Processing for Spectroscopic Applications. 2010.
92. *Björn Halvarsson*. Interaction Analysis in Multivariable Control Systems. Applications to Bioreactors for Nitrogen Removal. 2010.
93. *Jesper Bengtson*. Formalising process calculi. 2010.
94. *Magnus Johansson*. Psi-calculi: a Framework for Mobile Process Calculi. Cook your own correct process calculus – just add data and logic. 2010.
95. *Karin Rathsman*. Modeling of Electron Cooling. Theory, Data and Applications. 2010.

96. *Liselott Dominicus van den Bussche*. Getting the Picture of University Physics. 2010.
97. *Olle Engdegård*. A Search for Dark Matter in the Sun with AMANDA and IceCube. 2011.
98. *Matthias Hudl.* Magnetic materials with tunable thermal, electrical, and dynamic properties. An experimental study of magnetocaloric, multiferroic, and spin-glass materials. 2012.
99. *Marcio Costa.* First-principles Studies of Local Structure Effects in Magnetic Materials. 2012.
100. *Patrik Adlarson*. Studies of the Decay $\eta \rightarrow \pi^+\pi^-\pi^0$ with WASA-at-COSY. 2012.
101. *Erik Thomé.* Multi-Strange and Charmed Antihyperon-Hyperon Physics for PANDA. 2012.
102. *Anette Löfström.* Implementing a Vision. Studying Leaders' Strategic Use of an Intranet while Exploring Ethnography within HCI. 2014.
103. *Martin Stigge.* Real-Time Workload Models: Expressiveness vs. Analysis Efficiency. 2014.
104. *Linda Åmand.* Ammonium Feedback Control in Wastewater Treatment Plants. 2014.
105. *Mikael Laaksoharju.* Designing for Autonomy. 2014.
106. *Soma Tayamon.* Nonlinear System Identification and Control Applied to Selective Catalytic Reduction Systems. 2014.
107. *Adrian Bahne.* Multichannel Audio Signal Processing. Room Correction and Sound Perception. 2014.
108. *Mojtaba Soltanalian.* Signal Design for Active Sensing and Communications. 2014.
109. *Håkan Selg.* Researching the Use of the Internet — A Beginner's Guide. 2014.
110. *Andrzej Pyszniak*. Development and Applications of Tracking of Pellet Streams. 2014.
111. *Olov Rosén*. Parallel Stochastic Estimation on Multicore Platforms. 2015.
112. *Yajun Wei*. Ferromagnetic Resonance as a Probe of Magnetization Dynamics. A Study of FeCo Thin Films and Trilayers. 2015.
113. *Marcus Björk*. Contributions to Signal Processing for MRI. 2015.
114. *Alexander Madsen*. Hunting the Charged Higgs Boson with Lepton Signatures in the ATLAS Experiment. 2015.
115. *Daniel Jansson*. Identification Techniques for Mathematical Modeling of the Human Smooth Pursuit System. 2015.
116. *Henric Taavola*. Dark Matter in the Galactic Halo. A Search Using Neutrino Induced Cascades in the DeepCore Extension of IceCube. 2015.
117. *Rickard Ström*. Exploring the Universe Using Neutrinos. A Search for Point Sources in the Southern Hemisphere Using the IceCube Neutrino Observatory. 2015.
118. *Li Caldeira Balkeståhl.* Measurement of the Dalitz Plot Distribution for $\eta \rightarrow \pi^+\pi^-\pi^0$ with KLOE. 2015.
119. *Johannes Nygren.* Input-Output Stability Analysis of Networked Control Systems. 2016.
120. *Joseph Scott.* Other Things Besides Number. Abstraction, Constraint Propagation, and String Variable Types. 2016.
121. *Andrej Andrejev*. Semantic Web Queries over Scientific Data. 2016.

122. *Johan Blom*. Model-Based Protocol Testing in an ERLANG Environment. 2016.
123. *Liang Dai*. Identification using Convexification and Recursion. 2016.
124. *Adriaan Larmuseau*. Protecting Functional Programs From Low-Level Attackers. 2016.
125. *Lena Heijkenskjöld*. Hadronic Decays of the ω Meson. 2016.
126. *Delphine Misao Lebrun*. Photonic crystals and photocatalysis. Study of titania inverse opals. 2016.
127. *Per Mattsson*. Modeling and identification of nonlinear and impulsive systems. 2016.
128. *Lars Melander*. Integrating Visual Data Flow Programming with Data Stream Management. 2016.
129. *Kristofer Severinsson*. Samarbete = Samverkan? En fallstudie av AIMday vid Uppsala universitet. 2016.
130. *Nina Fowler*. Walking the Plank of the Entrepreneurial University. The little spinout that could? 2017.
131. *Kaj Jansson*. Measurements of Neutron-induced Nuclear Reactions for More Precise Standard Cross Sections and Correlated Fission Properties. 2017.
132. *Petter Bertilsson Forsberg*. Collaboration in practice. A multiple case study on collaboration between small enterprises and university researchers. 2018.
133. *Andreas Löscher*. Targeted Property-Based Testing with Applications in Sensor Networks. 2018.
134. *Simon Widmark*. Causal MMSE Filters for Personal Audio. A Polynomial Matrix Approach. 2018.
135. *Damian Pszczel*. Search for a new light boson in meson decays. 2018.
136. *Joachim Pettersson*. From Strange to Charm. Meson production in electron-positron collisions. 2018.
137. *Elisabeth Unger*. The Extremes of Neutrino Astronomy. From Fermi Bubbles with IceCube to Ice Studies with ARIANNA. 2019.
138. *Monica Norberg*. Engagerat ledarskap för att skapa förutsättningar för allas delaktighet. Utgångspunkter i kvalitetsarbetet. 2019.
139. *Peter Backeman*. Quantifiers and Theories. A Lazy Aproach. 2019.
140. *Walter Ikegami Andersson*. Exploring the Merits and Challenges of Hyperon Physics. with PANDA at FAIR. 2020.
141. *Petar Bokan*. Pair production of Higgs bosons in the final state with bottom quarks and τ leptons in the ATLAS experiment. Search results using LHC Run 2 data and prospect studies at the HL-LHC. 2020.
142. *Carl Kronlid*. Engineered temporary networks. Effects of control and temporality on inter-organizational interaction. 2020.
143. *Alexander Burgman*. Bright Needles in a Haystack. A Search for Magnetic Monopoles Using the IceCube Neutrino Observatory. 2020.
144. *Eleni Myrto Asimakopoulou*. Search for charged Higgs bosons with tau-lepton signatures at the ATLAS experiment of the Large Hadron Collider and development of novel semiconductor particle detectors. 2021.
145. *Oscar Samuelsson*. Sensor Fault Detection and Process Monitoring in Water Resource Recovery Facilities. 2021.