Data-Driven Approaches to Improve Operations of Biological Processes in WRRFs: Analytics, Model and Optimization

by

Cheng Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Environmental Engineering and Scientific Computing)
in The University of Michigan
2022

Doctoral Committee:

Professor Glen T. Daigger, Chair
Dr. Evangelina, Belia
Associate Professor Branko Kerkez
Associate Professor Raj Rao Nadakuditi

Cheng Yang
yangche@umich.edu
ORCID iD: 0000-0002-9352-0675

To my family and friends

ACKNOWLEDGEMENTS

It is a genuine pleasure to express my deep sense of appreciation and gratitude to a lot of people for helping me complete one of the greatest milestones in my life. Without you, I won't be able to reach this point.

First of all, I'd like to thank my advisor Prof. Glen Daigger, who is a mentor for both my life and work during my Master's and PhD journey at University of Michigan. Whenever I struggled for research ideas or got stuck with research progress, Glen always responded to me with a gentle smile and said, 'A good project takes twice of the time it was planned'. His confidence in smiles is truly infectious, calming me down and restoring my excitement about my research. He listened carefully and provided constructive and straight-to-the-point suggestions to guide me walk through these difficulties. He never hesitates to share his resources and sponsor activities that would be beneficial to his students. I can still remember his connecting me to different people to address my research needs and sponsoring me to conferences where I even don't have presentations. In life, as a foreign student, cultural differences are usually a challenge, but I suffered a little, thanks to Glen who is open and willing to answer my questions about appropriate behaviours and etiquette. I'd also like to thank Patty Daigger, the spouse of Glen, one of the kindest women I have ever met. I can't remember how many times Patty has hosted festival dinners for students like us who can't go home with family, and how many times she told us that Glen and she are so proud of us. I am so lucky that having Glen and Patty always be there for me.

I'd like to thank my committee members Dr. Lina Belia, Prof. Branko Kerkez,

and Prof. Raj Rao Nadakutiditi, who assisted me in improving my dissertation through feedback, course instruction, and conversations. Lina is so wonderful and knowledgeable in process engineering from both research and practise perspectives. Her experience and expertise has enriched the contents of this dissertation and has ensured they are tangible to practice. Words can't express how grateful I am to have her on board and work with her. Branko's and Raj's courses grounded the concepts of signal processing and machine learning techniques used in this dissertation. I can't help myself connecting the word 'cool' with Branko. He is a cool guy and he always uses 'cool' to describe interesting projects. I can sense his openness, enthusiasm and excitement about novel research ideas. I admired and learned such an attitude from him. Raj's course has cultivated me the needed mathematical mindset for this dissertation and the machine learning applications he shared contributes a lot in visioning how these tools could be used for WRRFs. I'd like to quote a sentence he told me here, 'Cheng, if you plan to try something hard in your life, then PhD is the best time to do it'. This echoes with me throughout my academic journey.

I have been privileged to work alongside amazing peers in the Daigger Research Group, the Real-Time Water Systems Lab and the Environmental Biotechnology Group at University of Michigan. They are amazing groups! It was a wonderful experience to mutually learn form each other through thoughtful and patient feedback, questions and comments. I'd like to also express my gratitude to the friends who have accompanied me throughout my life in Michigan. The most important period in my life is pleasant and colorful because of you!

Finally, I can never sufficiently thank my wonderful family. To my parents, your personalities and family dedication has shaped my values, which I will carry on throughout my entire life. I am so fearless because I know you will always love, respect and support me and you will always be a harbor for me. As the single child of the family, I know how badly you wish I could be near around but you never hesitate

to encourage me to travel abroad to take a look at the world. To Yuhang Zhang, my classmate, my roommate and my friend, who I have met for 13 years. Although we are not bonded by blood, you are already a brother, a family to me in my heart. When in high school, we were already good friends. Coincidentally we both came to U of M for master programs and became roommates, and finally we both transited to PhD and completed it. I am so grateful to have you by my side as a family for the past six years. May the friendship lasts for eternity.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

LIST OF ABBREVIATIONS

AD Anaerobic Digestion

ANN Artificial Neural Network

ASM Activated Sludge Model

AUC Area Under Receiver Operating Characteristics Curve

BOD_5 5-day Biochemical Oxygen Demand

BSM1 Benchmark Simulation Model No. 1

CCOD Colloidal Chemical Oxygen Demand

COD Chemical Oxygen Demand

CSTR Continuously Stirred Tank Reactor

DE Differential Evolution

DO Dissolved Oxygen

EA Evolutionary Algorithm

ES Evolutionary Strategy

GA Generic Algorithm

GLWA Great Lakes Water Authority

HPO High Purity Oxygen

IWA Internation Water Association

MLE Modified Ludzack-Ettinger

MLVSS Mixed Liquor Volatile Suspended Solids

NSGA-II Non-dominated Sorting Genetic Algorithm II

ORP Oxidation-Reduction Potential

OLS Ordinary Least Squares

PCA Principle Component Analysis

PCOD Paticulate Chemical Oxygen Demand

RAS Return Activated Sludge

RL Reinforcement Learning

RMSE Root Mean Square Error

ROC curve Receiver Operating Characteristic Curve

SCCOD Soluble and Colloidal Chemical Oxygen Demand

SCOD Soluble Chemical Oxygen Demand

SRT Solids Resident Time

TP Total Phosphorus

TSS Total Suspended Solids

WRRF Water Resource Recovery Facility

ABSTRACT

Wastewater treatment plants are being repurposed towards water resource recovery facilities (WRRFs), addressing nutrient recovery and energy neutrality to deal with stricter emission regulation, increasing water scarcity and rapid urbanization. With the growing ubiquity of sensors and meters, utilities can launch their digital transformation towards more sustainable and intelligent WRRFs. However, the unprecedented amount of data, which are various in source, magnitude, type and frequency, created a "data-rich equals information-poor" dilemma for wastewater professionals. It can be challenging to effectively collect, process, analyze and utilize these data and transform them into actionable intelligence. Outstanding knowledge gaps exist in what exact problems can be solved with these data and related data-driven tools and how to utilize them in the context of the wastewater treatment domain. "Data → Information → Knowledge → Intelligence" describes how data flow through a data pipeline and represents the increasing levels of understanding and ability to solve engineering problems in pursuit of building intelligent WRRFs. To embody this data pipeline, this dissertation investigates the critical steps of it and provides a holistic vision about transforming data into intelligence. Specifically, Chapter 2 evaluates how data analysis and modelling can improve data collection procedures. Chapter 3 evaluates how data pre-processing can be enhanced by coupling data-driven tools with engineering judgements. Chapter 4 develops an adaptive grey-box model whose model parameters can be self-updated with data streams, leveraging benefits from both black-box and white-box models. Chapter 5 develops an intelligent agent that

can assist water professionals with process design and operation. They provide vivid examples of where and how tools from the data science field can be used to advance wastewater treatment and what improvements could be achieved if they are coupled with WRRF domain knowledge, from the beginning to the end of the data pipeline. The holistic investigation of the data pipeline is significant at the current stage because the digital transformation journey of WRRFs is just launched and is still at an early phase. This dissertation helps identify the opportunities and challenges in the digital transformation journey, deepens understanding of methodology development, and demonstrates that promising outcomes that can be achieved with data-driven approaches.

CHAPTER I

Introduction

Wastewater treatment plants are facilities designed to remove contaminants from wastewater and convert it into an effluent that has acceptable impacts on environmental and public health. In recent years, they are being repurposed towards water resource recovery facilities (WRRFs) for water reuse, nutrient recovery and energy neutrality, as wastewater is considered as a resource for water, energy, heat and chemicals [2, 3, 4]. While both the academic and industrial water sectors are dedicating to sustainability by developing newer treatment technologies, digital solutions that embrace big data and artificial intelligence (AI) emerge and attract broad attention [5, 6].

An international survey (2019) of the global urban wastewater management community identified the emerging opportunities and threats introduced by digital transformation, ubiquitous sensing and new data sources [7]. A satisfactory response with a total of 309 surveys was received, covering the academic sector (60%), environmental consultants (20%), utilities (15%), manufacturers, government, and students ($<$ 5%). Europe (67%) and North America (21%) dominated the responses, and only a few were from South America, Africa, and Australasia. **The top ten novel top-**

ics [1] **(out of 35) were identified where Ontology ranked third.** Ontology is defined as a systematic representation of the available urban water management knowledge, whose key elements include the construction of digital intelligent agents that successfully deploy the knowledge available to make autonomous decisions or suggestions to an expert user. This rank indeed reveals the demand and tendency of calling for research in digital solutions to advance wastewater treatment, although their developments are still within the early stage.

Historically, the ontology of wastewater treatment was often constrained by difficulties in collecting data, however, the growing ubiquity of sensors and meters exposes utilities to an unprecedented amount of data [5, 7, 8]. For instance, small facilities (˜20,000 Population Equivalents, PE) can generate up to 400 signals, whereas large ones produce more than 30,000 [8, 9]. Meanwhile, the data is acquired from various sources in WRRFs: laboratory analysis, online sensor measurements, operation and maintenance logs and others. Each source creates data that are different in magnitude, types (numerical, categorical, textual) and frequency (from minutes to months). Due to the size and complexity of datasets generated by WRRFs, and the common lack of data science background for water professionals, a vast amount of raw data remain buried in 'data graveyards', as has been recognized that data-rich is all too often equivalent to information-poor [1, 8, 9, 10, 11]. **How to transform these data into actionable knowledge to advance wastewater treatment and what roles can data-driven tools play in the transformation** are fundamental and ongoing research topics.

The wastewater treatment community has been actively exploring approaches to take full advantage of WRRF data and has created a large volume of studies

[1] The top ten topics are (in a descending order): Linking Aquatic Ecology to Emissions; Reinforcement Learning; Ontologies; Cybersecurity; Complexity: Blind Trust; Micropollutant and Pathogen Monitoring; Environmental DNA (Biomonitoring of Natural and Engineered Aquatic Systems Using Environmental DNA); Increasing Risk of Global Transition Which Could Disrupt the Performance of Urban Wastewater Systems; Secondary Health Benefits; Onsite High-Resolution Mass Spectrometry.

focusing mainly on three types of applications: (1) **Monitoring**. The data are used to monitor and assess the current state of a process and help diagnose causes if unintended performance deviations occur. (2) **Modelling**. The data are used for model development, either for training purely data-driven models or for calibrating mechanistic models. The models encode relevant process information and knowledge, whose predictions greatly enhance design and operations of WRRFs. (3) **Control**. Control increases the stability of processes to ensure good performance at all the time and to optimize the usage of resources. Control design relies heavily on the data.

It is recognized that these three applications generally include most of the ongoing research studies in the interdisciplinary field of data and wastewater, however, bringing research into practice requires a more holistic vision, where more components need to be integrated and considered. Therrien *et al.* [1] shared such a vision about how data flow through a data pipeline in WRRFs as showed in Figure 1.1. "Data → Information → Knowledge → Intelligence" represents the increasing levels of understandings and ability to solve engineering problems in pursuit of building 'smart' wastewater systems. Critical steps include collection, pre-processing, storage and access, data mining, modeling, comprehension, and intelligent actions, which are listed with essential professions. Although Figure 1.1 pictures the conceptual pathway from data to intelligence, significant knowledge gaps still exist in methodology development, which hampers a more rapid digital transition of WRRFs [1, 7, 8, 10, 11, 12]:

- It has been generally recognized that a digital era of water and wastewater sector is coming with the rapid development of data and artificial intelligence. However, its scope, pathway and outcomes within the wastewater community are still ambiguous. For instance, sensor manufacturers and machine learning specialists promise substantial benefits of collecting and mining data, yet all too often leave out which exact challenges can be solved with these data (and which ones cannot).

- Data quality is the first and foremost challenge. Although cheaper memory storage and more reliable sensors make data collection easier, expanding collection in a 'brute force' manner is insufficient to guarantee data quality. Therefore, it would be of great importance to investigate collecting data in a smart, cost-effective way or extracting information from existing data by integrating modelling and analysis.

- Once good quality data are available, the next step is transforming them into actionable knowledge to advance wastewater treatment. Numerous tools are available with the recent development of data science and artificial intelligence, which post challenges to method selection and developments. What engineering problems could be solved with these borrowed tools and what improvements and benefits could be obtained by using them?

- Many methods are available from data science, however, realizing their full potential needs to take the WRRF particularities (e.g. physical constraints, process knowledge) into consideration.Yet no well-defined guidelines or rules have been explicitly proposed. For instance, where and how to adopt data-driven approaches with these particularities? What improvements can be achieved with these coupling approaches compared to traditional approaches or purely data-driven approaches?

This dissertation, leveraging knowledge and insights from both data science and wastewater process engineering, addresses these gaps and contributes to the digital transformation towards WRRFs. While a single dissertation is not able to address every aspect, this work covered major critical steps in the data pipeline to provide an integrated investigation. The research work is divided into four chapters (Chapter 2-5), each of which is also tagged in Figure 1.1. The dissertation ends with a conclusion chapter (Chapter 6), where the results, contributions and future directions are

provided:

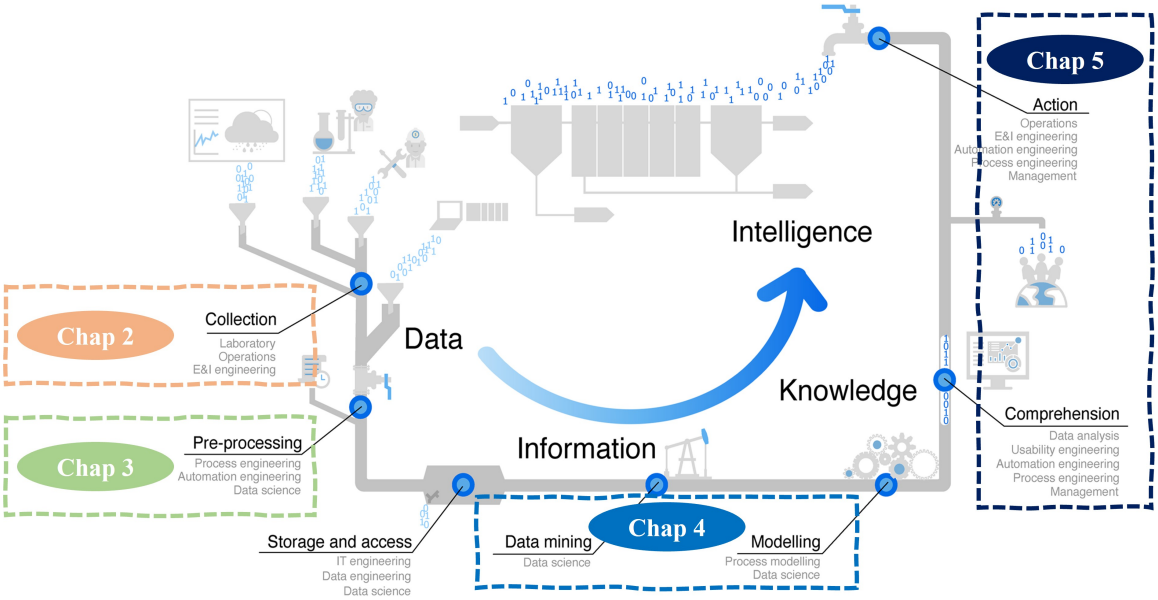

Figure 1.1: A road map of how data flow through a data pipeline and finally are transformed into intelligence. For each step in the pipeline, the most essential professions are listed. The figure was adopted and revised from Therrien *et al.* [1]. Chapters in this dissertation address critical steps along the data pipeline

- **Chapter-2**: This chapter evaluates how data analysis and modeling can improve data collection procedures. It introduces a case study where comprehensive data analysis and modeling were applied to quantify the variability and uncertainty of a fundamental data source (wastewater characteristics) in WRRFs and simulated consequences of insufficient data collection. It provides insights on minimal requirements for data collection and proposes a strategic plan to improve the data collection procedures.

- **Chapter-3**: This chapter evaluates how data pre-processing can be enhanced if WRRF-particularities are incorporated into data-driven tools. It introduces a case study where sensor measurements are flawed and seemingly useless. Appropriate usage of data-driven tools instructed by physical factors (e. g. prior process knowledge, physical constraints, phenomenon observations) extracted

useful information from signals, whereas performance of purely data-driven ones were unsatisfactory. It demonstrates that WRRF knowledge is indispensable to the appropriate and efficient usage of data-driven tools in data pre-processing.

- **Chapter-4**: This chapter develops an adaptive process model whose model parameters can be self-updated with data streams in WRRFs. In this chapter, a simply structured and intuitive grey-box model was reduced from first principal models (white-box models), then was tested with synthetic datasets. Like black-box models, its parameters were estimated from data in real-time, rather than being set upfront as white-box models do. It demonstrates that grey-box modeling is an efficient approach to boost the transformation from data into information and knowledge, addressing the data mining and modeling steps in the pipeline.

- **Chapter-5**: This chapter develops an intelligent agent that can assist water professionals with WRRF process design, operation and optimization. In this chapter, a computer agent based on genetic algorithms was build and connected with SUMO models to complete a process design task. Results showed that the agent was able to complete the task and propose reasonable designs for a new process, whose design criteria are not fully established in practice. This work demonstrates how decisions could be made with the assistance of artificial intelligence, addressing final steps in the data pipeline.

- **Chapter-6**: This chapter concludes this dissertation. Conclusions and contributions regarding each chapter are summarized and future directions are proposed.

Chapter 2: Improving Data Collection Procedures in WRRFs with Data Analysis and Modeling

One important type of data collected in WRRFs is wastewater characteristics. These data are used to instruct design, upgrade and operation of WRRFs. Moreover, they are critical inputs for process modelling. Generally, more data is desired for robust and reliable information because limited data may introduce biases. However, due to the laborious nature and limited budget for such data collection campaigns, in practice, often a limited amount of data is collected. Thus, it is important to have guidance concerning "how much data is enough".

The specific research question of this chapter is "How much data is required for a robust and reliable wastewater characterization?". With a one-year-long wastewater characterization dataset from the Great Lakes Water Authority (GLWA) WRRF as the experimental subject, the results demonstrated that typical evaluation metrics and techniques in data science (e. g. Monte Carlo Simulation) helped better determine the needed sample size. Results show that a minimum of 20 samples randomly distributed throughout the year is needed for a robust wastewater characterization. An adaptive data collection approach for wastewater characteristics was further proposed.

The main contribution of this chapter to the general wastewater community is that it identifies that the current approaches of wastewater characterization are insufficient in sample size for a robust estimation, and data collection should be planned wisely with adaptability in order to be cost-effective.

Chapter 3: Improving Sensor Data Processing in WRRFs by Coupling Data-driven Approaches with Physical Factors

Increased availability and affordability of sensors, especially water quality sensors, is poised to improve process control and modelling in water and wastewater systems. Sensor measurements are often flawed by unavoidable influent complexity and sensor instability, making extraction of useful signals difficult. Although a natural solution

is to put extra effort into sensor maintenance to achieve more reliable measurements, it is believed that useful signals can be extracted from those unqualified signals by appropriate usage of data-driven tools.

The specific research question is "How to couple physical factors (e. g. observations, physical constraints, process knowledge et al.) with data-driven tools to enhance the ability to extract useful sensor signals?". The specific task was to distinguish two influent sources from a highly flawed sensor signal, where standard data-processing methods yielded unsatisfactory results. With physical factors such as periodicity of diurnal pattern and the non-negative nature of influent mass, a customized algorithm based on Fourier series and regularized least squares was developed and proven to be useful. Logistical regression on statistics of pattern similarity, motivated by phenomenon observations, further separated sensor data that was truly bad due to sensor issues or seemingly flawed due to the inference of one influent source. Additionally, a well-calibrated neural network model was used to further support the signal processing and classification results. Discussions of how to couple physical factors into data-driven tools were also presented.

This chapter demonstrates that coupling data-driven tools with WRRFs particularities can help solve a current challenge that most of WRRFs are facing – data-rich does not guarantee information-rich. The methodology presented and discussed in this chapter provides problem-solving ideas to address this challenge.

Chapter 4: Developing an Adaptive Real-time Grey-box Model with Data Streams in WRRFs

Grey-box models, which combine the explanatory power of first-principle models with the ability to detect subtle patterns from data streams, are gaining increasing attention in the wastewater sector. Simple structured but fit-for-purpose grey-box models that capture time-varying dynamics by adaptively estimating parameters are

desired for real-time process optimization and control. The extracted real-time intuitive information will help WRRF staff make punctual adjustments in operations and management, and therefore, avoid downtime and failure of WRRFs.

The specific research question in this chapter is "how to realize an adaptive real-time model for advanced control and process optimization, whose parameters could be updated with data streams?" Following the concepts of grey-box models, this chapter presents the identification of such a grey-box model structure and its further conversion into an extended Kalman filter (EKF). The EKF can estimate the nitrification capacity and ammonia concentrations of a typical Modified Ludzack-Ettinger (MLE) process. The EKF was implemented and evaluated in real-time by interfacing Python with SUMO (Dynamita™), a widely used commercial process simulator. The EKF was able to accurately estimate the ammonia concentrations in multiple tanks when given only the concentration in one of them. In addition, the nitrification capacity of the system could be tracked in real-time, which provides intuitive information for facility managers and operators to monitor and operate the system.

This chapter provides a methodology on how to realize adequate online models with data streams in WRRFs for advanced real-time control and optimization, enriching the toolkit for transforming data into actionable information from the modelling perspective.

Chapter 5: Automating Process Design and Operations by Coupling Mechanistic Models with Genetic Algorithms

In recent years, the development of commercial software and simulators has progressed to assist engineers to optimize design, operation, and control of wastewater treatment processes. However, the methodology of using them is still primary. Commonly, manual trial-and-error approaches with engineering experience or exhaustive searches are used to find candidate solutions with simulators. These approaches are

becoming less favorable because of the increasingly elaborate process models, especially for new and innovative processes whose process knowledge is not fully established.

This study introduced a case study that coupled genetic algorithms (GAs), a subfield of Artificial Intelligence (AI), with a commercial simulator (SUMO) to automatically complete a design task, upgrading a plant from a Modified Ludzack-Ettinger (MLE) process to a new and complicated process - hybrid membraned aerated biofilm reactor process (Hybrid MABR). Results demonstrated that GAs can (1) accurately estimate influent wastewater fractions with common regular measurements (e. g. sludge yield, aeration supply and routine water quality concentrations) from the MLE process, and (2) propose reasonable designs (membrane surface area, tank sizes and sludge retention time) for the hybrid MABR process that reduce aeration, pumping and footprints with significantly improved effluent nitrogen quality.

This study demonstrated that tools from AI promote efficiency in wastewater treatment process design, operation and optimization by searching candidate solutions both smartly and automatically, as compared with the traditional manual trial-and-error approach. To the best of the authors' knowledge, it was one of the earliest studies that couple AI tools with commercial simulators. Because of its flexibility and efficiency, it is a promising approach that could be adopted widely in the industry, contributing to the ongoing developments of smart WRRFs.

CHAPTER II

Improving Data Collection Procedures in WRRFs with Data Analysis and Modeling

Published as:

Cheng Yang, Wendy Barrott, Andrea Busch, Anna Mehrotra, Jane Madden, and Glen T. Daigger. How much data is required for a robust and reliable wastewater characterization? *Water Science and Technology*, 79(12):2298–2309, 07 2019

2.1 Introduction

Process modeling based on the Internation Water Association (IWA) Activated Sludge Models (ASMs) has become the standard technique for the design of Water Resource Recovery Facilities (WRRFs) [14, 15, 16, 17]. These models depend on a detailed characterization of the influent wastewater that goes beyond the general simple lumped parameters, such as 5-day Biochemical Oxygen Demand (BOD_5) and Chemical Oxygen Demand (COD) typically collected for plant operation. Robust and valid characterization is essential for process modeling, as inaccurate wastewater composition inputs can lead to significant modeling errors [16]. The profound effect of wastewater characterization on modeling outputs has been demonstrated many times [18, 15, 19], and includes but is not limited to the followings:

- Sludge production is influenced by the estimated inert particulate COD.
- Oxygen demand is influenced by the estimated total bio-degradable COD.
- Anoxic denitrification rate and anaerobic phosphorus release are influenced by the estimated readily biodegradable COD.
- Effluent COD is influenced by the estimated inert soluble COD.

In practice, wastewater characterization is conducted mainly via two methods: (1) physical-chemical and (2) respirometric. STOWA [20] proposed simple and easy to implement guidelines based on physical-chemical methods. WERF [21] provided a state-of-the-art and frequently used method for measuring key influent wastewater characteristics, kinetics and stoichiometric parameters covering both methods. BIOMATH [22] developed a protocol for activated sludge model calibration, with influent wastewater characterized by the respirometric method. Recent attempts at integrated characterization suggested a combination of both methods [23]. These various methods were compared by Gillot & Choubert [24] and Fall *et al.* [25], where significant gaps were found in results.

Despite lack of agreement on the best characterization method, the choice should fit the purpose for which the model is being developed. Due to its time-consuming and labor-intensive nature, wastewater characterization is often conducted intensively within one or a limited number of short duration campaigns. While these data allow a simulation model to be set up, concerns exist when the model is to be used to simulate future performance. For example, 'Are sufficient data collected to robustly characterize the wastewater on a long-term basis?' and 'Do wastewater characteristics vary on a seasonal or more random basis?' Non-representative wastewater characterizations can lead to significant cost implications when model results are used to make decisions on facility upgrades/expansions and operation.

On-going work at the Great Lakes Water Authority (GLWA) WRRF in Southeast Michigan provided an opportunity to conduct detailed wastewater characterization over an annual cycle. Building on this long-term data set, an assessment of variations in wastewater characteristics and impacts of different strategies for wastewater characterization campaigns was conducted.

This chapter evaluates alternative wastewater characterization campaign designs, mainly focusing on campaign size and timing. Following physical-chemical guidelines provided by WERF [21], detailed wastewater fractionation and characterization was conducted every week for a one-year period. Characterization results were fed into a standard ASM1 model, modified as described below, and different practical campaign strategies were evaluated. Based on these investigations, suggestions about obtaining robust and reliable wastewater characterization estimates by campaign design are proposed. Bioreactor Mixed Liquor Volatile Suspended Solids (MLVSS) concentration, which responds in a straightforward fashion to process operating conditions and the relative fractions of biodegradable and non-biodegradable particulate matter in the influent wastewater, was used as the modeled response variable, compared to actual daily values. GLWA uses the High Purity Oxygen (HPO) activated sludge process operated with an average 2.3-day Solids Resident Time (SRT), making MLVSS concentration responsive to variations in wastewater characteristics.

2.2 Materials and Methods

2.2.1 Description of the plant

The GLWA WRRF is a 3,560,000 m3/day (940 MGD) peak flow (secondary treatment) facility serving 3.1 million residents in Southeast Michigan. The liquid process treatment train consists of influent pumping and preliminary treatment (screening and grit removal), conventional primary treatment with ferric chloride addition for

phosphorus removal, HPO activated sludge, and effluent disinfection. Flows above 3,560,000 m3/day and up to 4,500,000 m3/day receive primary treatment with ferric chloride addition. Secondary treatment requirements apply, along with seasonally varied monthly effluent Total Phosphorus (TP) limits of 0.7 mg-P/L (October to March) and 0.6 mg-P/L (April to September). The plant routinely meets all discharge standards. Solids are thickened, dewatered, and either subject to drying or incineration and landfill.

2.2.2 Wastewater fractions

Flow-proportioned 24-hour composite samples are collected daily for influent wastewater, secondary influent (primary effluent) and secondary effluent by GLWA WRRF staff. There are actually three separate influent streams to the GLWA facility, and each is sampled separately. While a combined primary effluent stream is conveyed to secondary treatment, it passes through two different pumping stations to secondary treatment, and each secondary influent stream is sampled separately. Return Activated Sludge (RAS) is combined and conveyed to the HPO bioreactors, resulting in a 'single' biological population, but two separate sets of secondary clarifiers exist and each set is sampled separately. Detailed wastewater fractionation was conducted weekly on all seven streams on samples collected on random weekdays over the period from October 19, 2017 to October 17, 2018. Wastewater fractionation generally followed the physical-chemical guidelines provided by WERF [21], and consisted of stepwise filtration through the standard glass fiber filter (1.2 μm nominal pore size) and an 0.45 μm membrane filter. Filtrate through the glass fiber filter (1.2 μm) was defined as the sum of Soluble and Colloidal Chemical Oxygen Demand (SCCOD). Filtrate through the 0.45 μm membrane filter was defined as Soluble Chemical Oxygen Demand (SCOD). The difference between these two filtrates was defined as Colloidal Chemical Oxygen Demand (CCOD). Paticulate Chemical Oxygen Demand (PCOD)

was defined as the difference between the total COD and SCCOD. COD and BOD_5 analyses were conducted by GLWA staff according to Standard Methods [26].

Flocculation and filtration [27, 20] is more generally applied to determine the soluble fraction of wastewater. Previous work [28] had indicated that, for this wastewater, there was no significant difference for COD and BOD_5 between 0.45 μm membrane filtrate and the results with flocculation and filtration per the WERF protocol. Note that ferric chloride is added prior to the primary clarifiers for phosphate removal, and this may function, to a certain extent, to achieve the flocculation of colloidal organic matter present in the influent wastewater. An independent wastewater characterization effort was conducted during this period in connection with an on-going master planning effort [29] that reached similar conclusions. In that study, they performed six days of COD characterization at the GLWA WRRF following standard physical-chemical guidelines [26], and these results generally support that use of simple membrane filtration, rather than the more complicated flocculation and filtration procedure, is reasonable to characterize soluble organic constituents for this wastewater. Secondary influent (primary effluent) data were used in this study for modeling purposes. Not including flocculation and filtration of the samples collected from the several locations each week also facilitated the significant duration of the sampling program and became a practical consideration in proceeding with the characterization campaign.

2.2.3 Mapping measured wastewater fractions into model inputs

As showed in Figure 2.1, required IWA ASM inputs include readily biodegradable COD (S_s), slowly biodegradable COD (X_s), soluble inert COD (S_I) and particulate inert COD (X_I) [14], which were calculated as fractions of total COD. As discussed below, colloidal COD was found to be insignificant for this wastewater and, therefore, was incorporated into the particulate COD fraction. The soluble inert COD

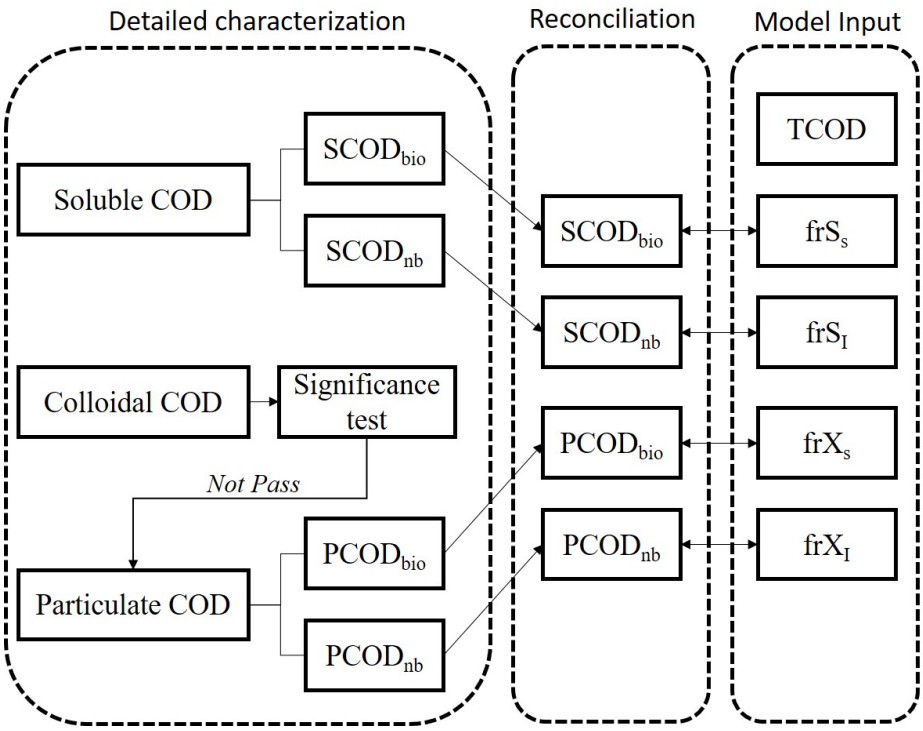

Figure 2.1: The summary of Mapping measured wastewater fractions into model inputs.

(S_I) was determined directly as the measured secondary effluent membrane filtered COD ($SCOD_{nb}$). The readily biodegradable COD (S_s or $SCOD_{bio}$) was calculated as the difference between the total soluble COD ($SCOD$) and $SCOD_{nb}$. The total biodegradable COD ($SCOD_{bio} + PCOD_{bio}$) was determined using the measured BOD_5 following STOWA guidelines [20] and using a *biodegradable* COD/BOD_5 ratio of 1.73 mg COD/mg BOD_5. The slowly biodegradable COD (X_s or $PCOD_{bio}$) was determined as the difference between the total biodegradable COD and $SCOD_{bio}$. The final remaining COD ($PCOD_{nb}$) was then the particulate inert COD (X_I). A manual reconciliation process, including mass balance check, specific ratio check, non-negativeness check etc. [30] was applied to the four wastewater component data, and records with apparent abnormalities were omitted. The reconciled COD concentrations were converted into fractions and then fed into the model.

2.2.4 Biological process modeling

HPO process bioreactor MLVSS concentrations were calculated using a standard IWA ASM1 [14], modified as described below and implemented in MATLAB®, with measured secondary influent total COD and fractions determined as above as input. Secondary influent was used in the model for two reasons. One is that it represents the direct input to the secondary treatment process and, consequently, the impacts of upstream treatment on wastewater constituents need not be included in the model. Secondly, GLWA measures secondary influent total COD daily, so a several-year database was available for extensive evaluation of model performance based on various approaches for analyzing the fractionation results, as described below. Two long-term data sets were used for modeling and model evaluation. Daily data for the period of 19 October 2017 to 17 October 2018, corresponding to the year over which detailed wastewater fractionation occurred, were used as the model training set. Daily data from 18 October 2013 to 17 October 2017 were used for model evaluation and verification.

A simplified model based on a single completely-mixed bioreactor was used to compute the MLVSS, the response model variable which was compared to the measured MLVSS concentration. This simplified model facilitated process modeling and data analysis (around 50 times reduction on run-time). A more complete model of the entire liquid treatment process had previously been developed in SUMO (Dynamita). Comparison of the results from the two models demonstrated that use of the simplified bioreactor configuration did not materially affect MLVSS predictions. Further details of the model used include the following:

- Biochemical processes included growth, decay and hydrolysis. Because biomass prediction was the main objective of this study, only these highly biomass-related reactions were considered.

- Heterotrophic biomass was used to estimate the overall biomass. As is typical for HPO processes used for secondary treatment due to the relatively low SRT (average = 2.3 days) and the reduced bioreactor pH due to the retention of CO_2 in solution, nitrification does not occur in the full-scale system.

- Since it is an HPO process, where oxygen is not limiting, oxygen limiting terms in reaction rate expressions were not included.

- Standard stoichiometric and kinetic parameters and temperature correction factors from the literature [31, 17, 32] were used, as summarized in Table 2.1

- Standard checks on the data, such as the mass balance over the secondary clarifier, were performed for the entire data set and confirmed the integrity of the data for its intended use (data not shown).

2.2.5 Model performance evaluation

Mean and standard deviation values were calculated for model predictions and actual MLVSS data, and the Root Mean Square Error (RMSE) between model predictions and actual MLVSS concentrations was calculated to evaluate model performance. Our evaluation focused particularly on instances where model predictions appeared to differ noticeably from measured values, as they suggested periods of lack of fit for the model. We defined two types of deviations, namely outliers and spikes. Outliers were defined by comparison of individual model predictions to individual actual values where the deviation exceeded $\pm$ 3 standard deviation from the actual MLVSS (corresponding to a probability of occurrence of 0.3% based on the assumption of a normal distribution). Spikes were defined by deviations exceeding $\pm$ 2 standard deviation of actual MLVSS (corresponding to a probability of 4.6% [33]).

Table 2.1: Stoichiometric and Kinetic Parameter Values and Temperature Correction Factors Used in Model

Type	Symbol	Parameter	Unit	Value(20°C)*	Factor θ
Kinetics	μ_H	Maximum specific growth rate of Heterotrophs	d^{-1}	0.6	1.04
	K_s	Substrate half saturation for heterotrophs	$\text{mg-COD} \cdot \text{L}^{-1}$	20	1.03
	f'_D	fraction of active biomass contributing to debris	$\text{mg-COD} \cdot \text{mg-COD}^{-1}$	0.08	-
	b_H	Aerobic decay coefficient for heterotrophs	d^{-1}	0.63	1.03
	k_h	Hydrolysis rate coefficient	d^{-1}	2.2	1.03
	K_x	Hydrolysis half saturation coefficient	$\text{mg-COD} \cdot \text{mg-COD}^{-1}$	0.15	1
Stoichiometries	Y_H	Yield of Heterotrophs on substrate	$\text{mg-COD} \cdot \text{mg-COD}^{-1}$	0.67	1
Partitioning Coefficients	$i_{VSS,B}$	COD/VSS ratio of biomass	$\text{g-COD} \cdot \text{g-VSS}^{-1}$	1.42	1
	$i_{VSS,XI}$	COD/VSS ratio of particulate inert	$\text{g-COD} \cdot \text{g-VSS}^{-1}$	1.5	1
	$i_{VSS,Xs}$	COD/VSS ratio of particulate substrate	$\text{g-COD} \cdot \text{g-VSS}^{-1}$	1.8	1
	$i_{VSS,XD}$	COD/VSS ratio of biomass debris	$\text{g-COD} \cdot \text{g-VSS}^{-1}$	1.3	1

*Temperature Correction: $x(T) = x(20°C) \cdot \theta^{(T-20°C)}$.

2.2.6 Practical campaign strategies evaluation

Three averaging strategies, yearly, quarterly and monthly, were applied for conversion of the measured fractionation data to determine model inputs, and then fed into the model to predict the bioreactor MLVSS concentration. This approach was used not only for the period over which detailed wastewater fractionation was conducted (19 October 2017 to 17 October 2018). To further evaluate the general applicability of the fractionation data and averaging strategies, the results from the three different averaging strategies were applied over the preceding four years of data and the resulting bioreactor MLVSS concentrations were calculated. In addition, each single-monthly average fraction value was used to represent whole-year values to evaluate the performance of shorter period characterization campaigns.

2.2.7 Potential indicators of days bad for campaign

Using the yearly-average model for the training data set, individual days were divided into two categories – spikes ($\geq$ 2 STD) and non-spikes. Differences in important plant conventional influent wastewater and operational features for these two data sets were investigated. Unpaired two sample t-tests were conducted over those features to detect statistically significant differences in mean values. Significantly different features can potentially serve as a flag for a bad campaign day.

2.2.8 Campaign size evaluation

Random sampling without replacement was conducted for different sample sizes from the year-long campaign data to determine the effect of sample size on wastewater characteristic estimates. Estimates of COD fractions gained from different sample sizes were averaged and fed into the model for simulation. Fifty iterations were conducted for each sample size. Maximum and mean values for averages of year-long predicted MLVSS, RMSE, number of outliers and number of spikes were calculated

for each sample size.

2.3 Results and Discussion

2.3.1 Determination of model input values based on measured fractionation data

Secondary influent total COD and concentration fraction data for the year over which these data were collected are presented in Figure 2.2. The total COD concentration varied significantly (158 ± 40 mg/L, ranging from 87.5 to 259 mg/L) throughout the year, primarily as a result of dilution during wet weather periods considering the GLWA WRRF is a combined sewage system. Particulate components appeared to be the most varied, covering a range of 27–217 mg/L, while the soluble component fluctuated with a range of 21–104 mg/L. The colloidal component was generally smaller than the particulate and soluble components, and some negative values were recorded. This can arise because the colloidal component is calculated by difference between the measured glass fiber and 0.45 mum filter filtrates. Since any measurement is subject to random errors, a measured value for the 0.45 mum filtrate that is randomly higher than the true value and the measured value for glass fiber filtrate that is randomly lower than the true value can result in the calculation of a negative value. The uncertainties (standard deviations) for total COD and glass fiber filtered COD were 40 and 27 mg COD /L respectively, and the maximum absolute value of the colloidal component was 56 mg/L, mathematically supporting that the colloidal concentration was subject to measurement error. Analysis of the secondary influent wastewater characterization data collected by [29] during this same period suggested that the colloidal fraction is not statistically significant. Thus, it appears likely that the concentration of colloidal COD in the secondary influent (primary effluent) may be small enough that it cannot be accurately measured for this wastew-

ater. Inspection of the data presented in Figure 2.3 also suggests that colloidal COD is a small fraction of the total COD and that it can, perhaps, be neglected as long as it is incorporated into another COD fraction.

Based on the observations above, a one-sample-t-test was conducted with a null hypothesis that the mean value of the colloidal COD is not equal to zero. With 95% confidence, the analysis failed to reject the null hypothesis (p-value = 0.13). In addition, ordinary least square linear regression analysis was conducted for the relationship between colloidal COD and total COD. Results showed that:

(a) both slope and intercept were not significant;

(b) the goodness of fit, R^2 square, was 0.022;

(c) the p-value of the ANOVA test comparing this linear fitting with no fitting was 0.32.

These results all indicate that the colloidal component is sufficiently small that it cannot be measured for this wastewater with this technique. Consequently, this fraction was incorporated into particulate components, as is the typical approach when ASM1 is applied.

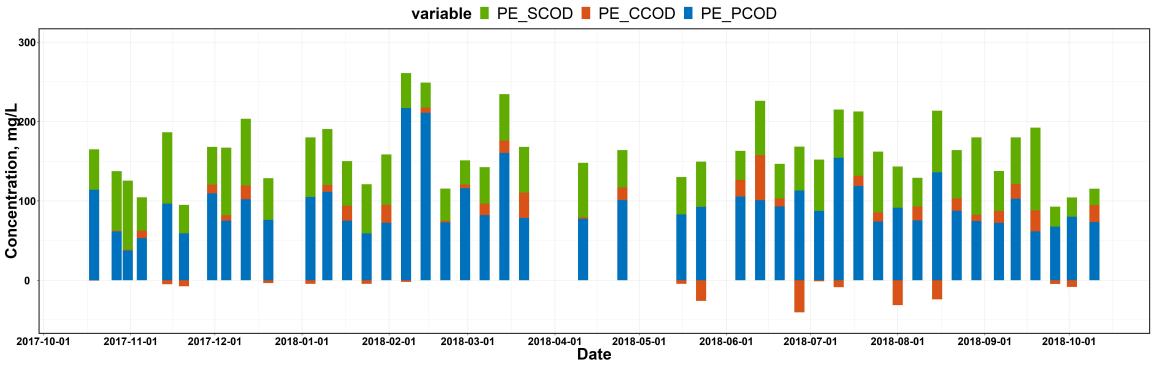

Figure 2.2: Variation of secondary influent (primary effluent) COD concentration fractions based on filtration procedure applied throughout the campaign year.

Figure 2.3 summarizes the reconciled input fractions for each day of the campaign year. There was no obvious pattern throughout the campaign year, and particulate

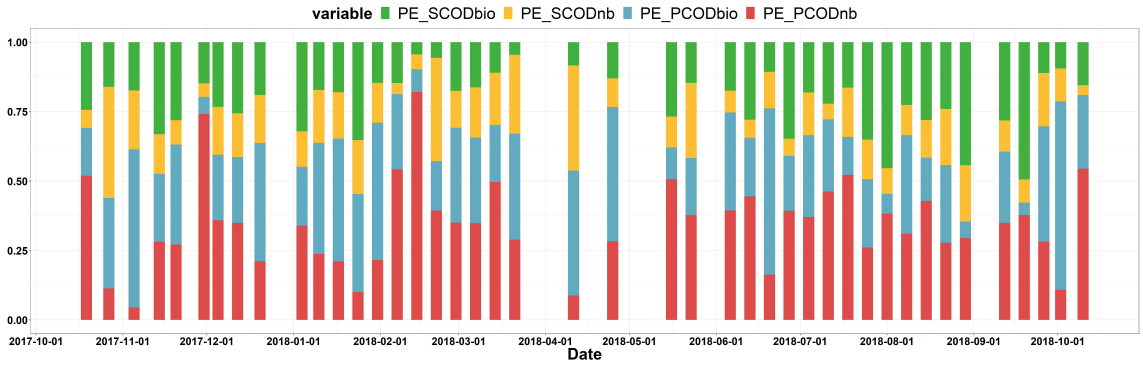

Figure 2.3: COD model input values as a fraction of total COD for the campaign year. Components: biodegradable COD (S_s), slowly biodegradable COD (X_s), soluble inert COD (S_I) and particulate inert COD (X_I).

COD (both biodegradable and non-biodegradable) varied more than soluble COD components. Table 2.2 provides both raw influent wastewater and primary effluent characteristics, as determined in this study, compared to recent literature values. The results for this wastewater are within the range of those obtained with other wastewaters, suggesting that it may be generally representative of domestic wastewater from a large metropolitan area.

2.3.2 Comparison of model results with actual data

The three different methods for averaging the fractionation data were evaluated using the campaign year as the training set, and the preceding four years as the validation set, as described above. Figure 2.4 compares predicted and measured MLVSS concentrations for the three methods for the training set, while Table 2.3 summarizes performance statistics for the training and validation data sets. While variations occur between model-predicted and actual MLVSS values, the model-predicted and measured MLVSS concentrations are generally of the same order of magnitude for all three averaging methods. This is significant as the modeling approach does not include a mechanism to directly calibrate the model results to measured values. Model stoichiometric and kinetic parameters are standard values taken from the literature,

Table 2.2: Comparison of the results from this study for wastewater COD fractions compared to literature values

Source	Total COD	COD Fraction in Percentage				
	mg/L	S_s	S_I	X_s	X_I	X_H
Primary effluent characteristics						
This study[a]	159 ± 41	22 ± 9	15±7	28 ± 13	35 ± 15	-
[25][a]	492	36	5	35	24	-
[34][b]	250	10	8	58	24	-
[35]		29	3	43	11	14
Raw influent characterisitics						
This study[a]	280 ± 85	20 ± 10	9 ± 5	35 ± 15	36 ± 20	-
[29][a]	290	15	9	24	52	-
[23]	540	8-10	1-4	27-40	14-36	23-46
[36]	176-220	19.5-27.8	8.4-12.8	16.1-37.3	13.9-33.4	14.7-18.9
[20][a]	241-827	9-42	3-10	10-48	23-50	
[37]	250-430	7-11	12-20	53-60	8-10	7-15
[35]	400	27	15	40	17	

[a] The fractions were measured and calculated purely with a physical-chemical method.

[b] The fraction was calibrated with estimations from the literature (not directly measured).

as discussed above and summarized in Table 1, and wastewater influent values are based on measured influent values, as described previously. As noted in Table 2.3, actual average MLVSS concentrations compare quite well with model values, irrespective of the averaging method used. Importantly, this suggests that the wastewater characterization method used, along with the use of relatively standard stoichiometric and kinetic coefficients, can lead to a reasonable model to begin with.

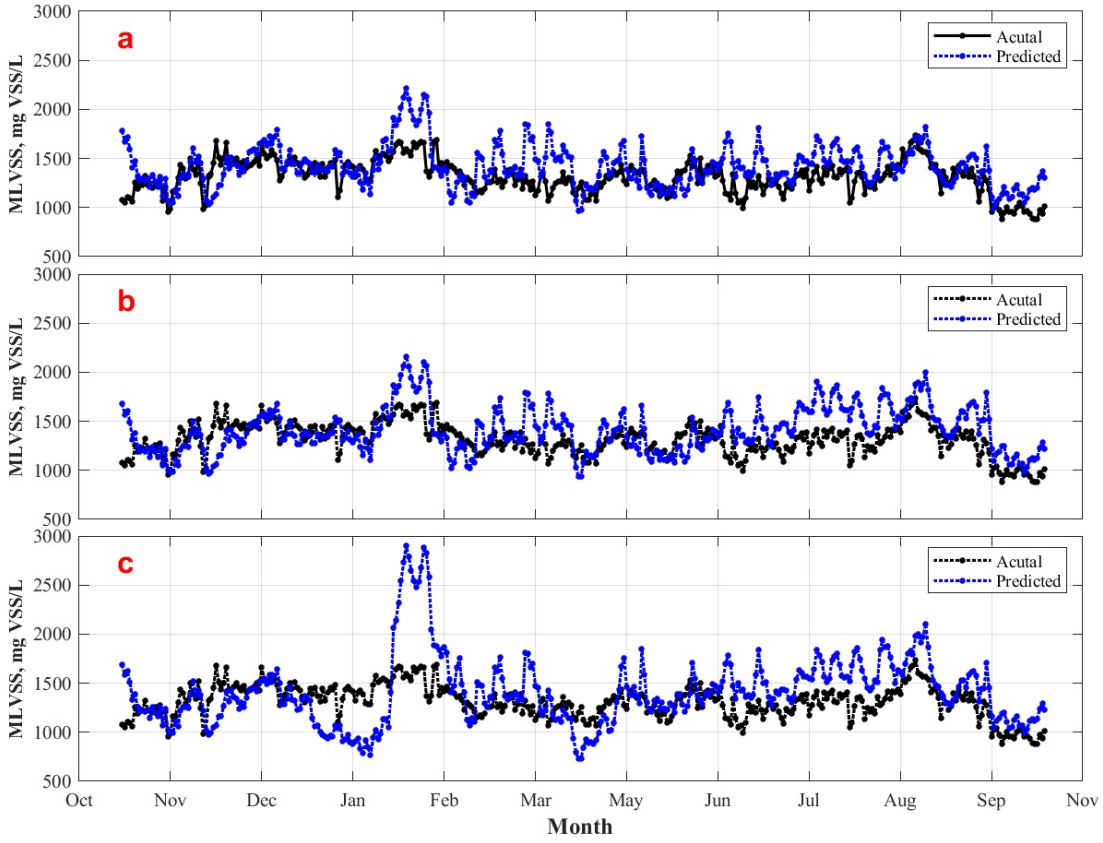

Figure 2.4: Simulation results for the three different fractionation averaging methods for the training data set. (a) Yearly average; (b) quarterly average; (c) monthly average.

Visual inspection of the data presented in Figure 2.4 indicates a noticeable lack of fit from early February to late March. Model predictions consistently exceed actual values, and the deviations exceed the 10% criteria often applied to indicate model lack of fit [16]. Inspection of the individual data during this period indicated that this

Table 2.3: Simulation results for the three different fractionation averaging methods for the training and testing data sets

Set	Average Method	Mean mg VSS/L	STD mg VSS/L	RMSE mg VSS/L	>1 STD	>2 STD	>3 STD
Training	Actual	1311.6	170.9				
(2017/10/18-	Yearly	1419.3	216.5	229.4	38.9%	14.0%	4.4%
2019/10/17)	Quarterly	1411.2	237.5	243.2	45.5%	15.6%	4.9%
Size: 365	Monthly	1413.5	361.7	343.5	55.6%	27.4%	10.7%
Testing	Actual	1165.9	185.9				
(2013/10/18-	Yearly	1119.7	281.0	256.3	44.0%	16.2%	3.0%
2017/10/17)	Quarterly	1112.8	288.1	185.9	47.8%	17.2%	3.8%
Size: 1461	Monthly	1106.9	332.5	312.2	56.0%	24.4%	6.0%

arose because of the nature of the model used. As indicated in Table 2.1, values for the COD/VSS for influent particulate inert material ($i_{VSS,XI}$) and influent particulate substrate ($i_{VSS,Xs}$) of 1.5 and 1.8 gCOD $\cdot$ gVSS^{-1} are used, while the actual measured value for the influent particulate matter throughout the year was 1.9 $\pm$ 0.7, ranging from 0.7–4.1 gCOD $\cdot$ gVSS^{-1}. The ratio for February to April was 2.5 $\pm$ 1.0. In fact, use of higher values in the model for this period resulted in near elimination of this lack of fit. By adjusting $i_{VSS,Xs}$ from 1.8 to 2.3 and $i_{VSS,XI}$ from 1.5 to 2.0, the February spike was eliminated, but the resulting model underestimated the actual MLVSS for March. Overall, the mean predicted MLVSS was improved to 1,361.4 mg/L, along with small improvements of RMSE and standard deviation (221.3 and 204.5 mg/L, less than 8%). From a modeling perspective, the lack of fit during the February to March period did not occur due to variations in wastewater characteristics, but rather because of poor model structure, as the COD to VSS ratio for these individual model components was not formulated as a wastewater characteristic but as a model parameter.

Interestingly, the months of February and March represent a distinct operating period when influent flows tend to be somewhat higher and periods of precipitation occur (this is a combined system, as described above). This unusual operating period may explain why the COD to VSS ratio is higher during this period. The impact of unusual operating conditions is addressed in additional detail below. From a modeling perspective, a priori knowledge concerning this failure of model structure would be required if the model is to be used to predict future performance. From a practical perspective, however, extreme COD/VSS values (around 0.7 or 4.1 for example), can be used to eliminate those days from the data set as they are likely measurement errors.

The results summarized in Table 2.4 address a different question; that is, whether there were better and worse times to conduct fractionation studies. It differs from

the monthly analysis summarized in Table 2.3 and illustrated in Figure 2.1, in that the fractionation results for a single month are used to model the entire year. The results indicate that some time periods are better than others.

The poorest results occur when characterization data from February is used, as might be expected from the results presented immediately above. The difference between the mean predicted and actual MLVSS increases to 46% of the actual value, the RMSE is more than triple the value for yearly average results presented in Table 2.3, the percentage of predictions exceeding one STD increased to 96.2%, and 58.6% exceeded three STDs. On the other hand, the fractionation data from certain months, such as March and October to December, generally performed better in terms of mean values, RMSE, and the percentage exceeding two and three STD (spikes and outliers) as summarized in Table 2.3. Note that the number of fractionation measurements was not the main contributor to improved performance, as larger sample size did not guarantee good performance (August and February) and smaller sample size did not diminish performance (April). It is noted that the period of October to December generally represents a period of lower plant influent flow.

A further analysis of the potential reasons for deviations was conducted by evaluating the differences in operating conditions on days where spikes (difference between modeled and actual MLVSS ≥ 2 STD) occurred, compared to the operating conditions for days when spikes did not occur (Figure 2.5). The hypothesis test results indicate that, within a 95% interval, days with spikes tended to occur on days with lower SRT, MLSS, higher secondary influent BOD_5, COD, TSS and VSS concentration, and higher secondary effluent TSS concentration. In short, efforts should be made to conduct fractionation campaigns during periods of relatively normal influent flow, loading, and operation, and results from periods where these factors are somewhat abnormal should be carefully screened and reviewed.

Table 2.4: Simulation results using fractionation data from an individual month to represent the whole year.

Month	Mean mg VSS/L	Std mg VSS/L	RMSE mg VSS/L	>1 STD	>2 STD	>3 STD	Sample Size
January	947.9	145.7	400.4	86.6%	60.3%	16.7%	5
February	1915.9	292.2	656.2	96.2%	88.8%	58.6%	4
March	1424.2	217.2	230.0	40.0%	12.9%	4.4%	3
April	1068.5	163.2	297.8	72.3%	28.5%	4.7%	2
May	1525.5	232.7	302.2	57.5%	24.1%	8.2%	2
June	1441.8	219.9	242.7	40.8%	15.9%	5.8%	4
July	1512.2	230.7	292.4	54.2%	22.5%	7.4%	4
August	1651.8	251.9	409.7	81.1%	43.0%	20.0%	5
September	1494.5	228.0	279.1	51.0%	20.5%	6.3%	3
October	1337.4	204.0	195.8	34.2%	8.2%	2.5%	4
November	1342.8	205.1	198.8	32.6%	8.2%	2.2%	4
December	1304.2	199.5	194.1	35.6%	8.8%	1.6%	3
Actual	1311.6	170.9					43
Yearly Average	1361.3	326.0	294.4	53.0%	23.0%	8.0%	43

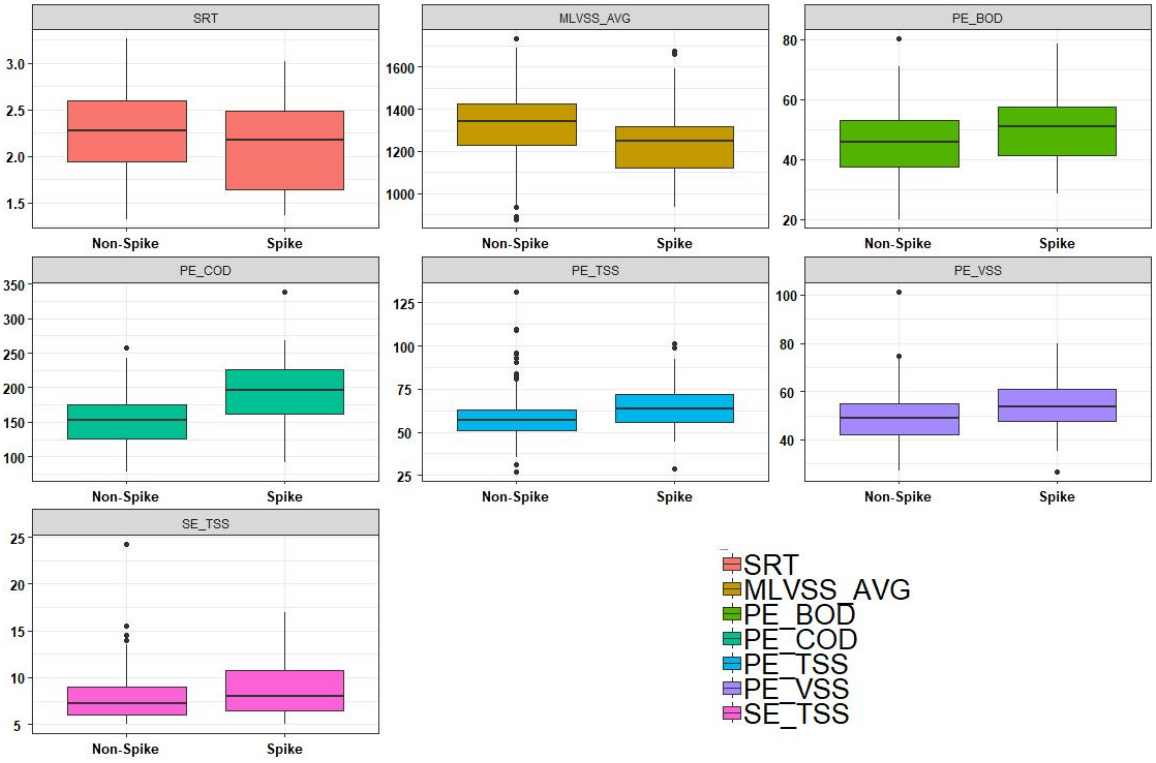

Figure 2.5: Boxplots of Potential Indicators in Spike Days and Non-Spike Days. These indicators were chosen based on 95% confidence interval of unpaired two sample t-test.

2.3.3 Impacts of sample size on wastewater characteristic estimation and model performance

Increased sample size can improve estimated fractionation, but with diminished results, as presented in Figure 2.6. Fifty iterations were implemented for each sample size, with the designated characterization records randomly pooled without replacement. Averaged fractions were fed into the model for simulation, and performance was evaluated. To minimize the error introduced by chance in sampling, both the maximum and average values of the model performance statistics among the 50 iterations were calculated. Average values reflect the overall performance of each sample size, while maximum values indicate the robustness, meaning that the result is not significantly influenced by individual characterizations. As indicated in Figure 2.6, there is a point of diminishing return. As expected, the desired 'elbow point' is controlled by the maximum criteria to achieve robust estimates of wastewater characteristics, making 20 the desired sample size in this instance. A preliminary analysis regressing the inert particulate fraction on the total COD with bootstrap sampling reached a similar conclusion (data not shown).

2.3.4 Implications for wastewater characterization campaigns

These results provide guidance on the number of individual measurements that can result in a robust assessment of wastewater characteristics. The analysis summarized in Figure 2.6 suggests that around 20 measurements represent a reasonable balance between achieving a robust assessment without an excessive number of measurements. The results presented in Table 2.3 also support the conclusion that 'more is better' (yearly average compared to quarterly and monthly) when assessing wastewater characteristics and their impact on model performance. This result conflicts, however, with those presented in Table 2.4, which indicated that even a small number of measurements conducted at 'the right time' (March and October to December

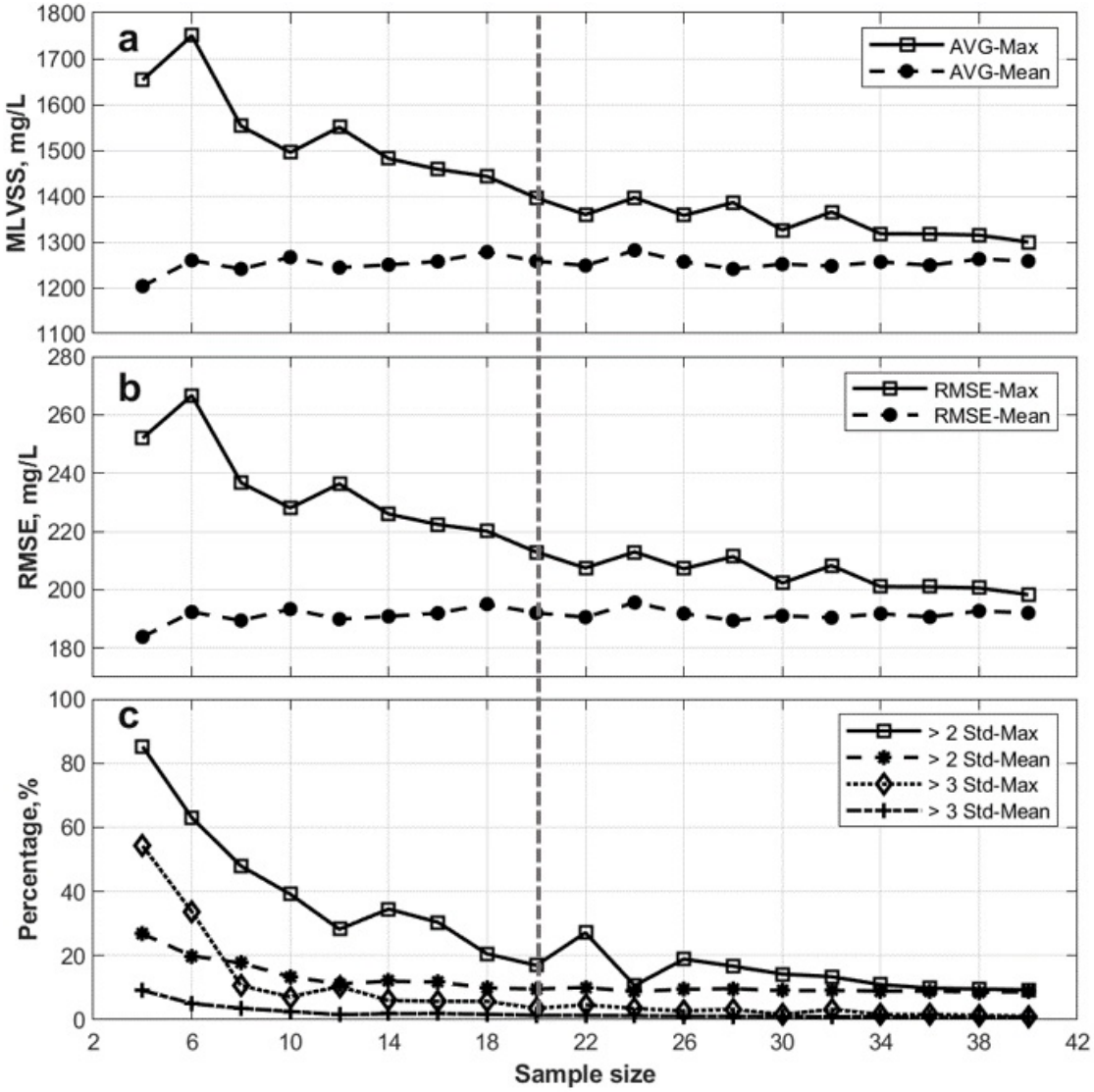

Figure 2.6: Elbow plots to determine sample size. Each sample size was iterated 50 times, and then maximum and mean values for each model assessment parameter were extracted to represent each sample size. The model evaluation parameters used include maximum and average values for (a) mean of predicted MLVSS; (b) RMSE of predicted MLVSS; (c) days with different deviations.

in this case) can result in better characterization of the wastewater relative to model performance. This presents a conundrum for planning wastewater characterization campaigns, as it is not possible to know, a priori, what the 'right time' is. Certainly, periods that are recognized to generally represent unusual conditions can be avoided, but it may not be possible to predict the ideal time. This suggests that an adaptive approach to wastewater characterization may be needed. It may consist of multiple sampling events, each of relatively short duration, with the results carefully evaluated after each event for consistency in model predictions as well as the occurrence of unusual influent or operating conditions. Sampling periods continue until a consistent set of results is achieved. Using this approach, sampling can be terminated when a sufficient number of measurements are obtained during periods of normal operation so that a robust assessment of wastewater characteristics is achieved. Issues related to model structure, as occurred in this instance during February and March of 2018, can also be identified with this approach and addressed appropriately given the objective of the modeling exercise. Use of this approach makes it unnecessary to specify initially the number of measurements required to achieve a robust assessment of wastewater characteristics, as the methodology itself will determine this. A robust budget is needed to account for unforeseen conditions. Given the significant economic impact of poor wastewater characterization in many instances, unnecessarily limiting the wastewater characterization budget may not be a wise use of funds as the economic impact of poor decisions may be orders of magnitude greater than the cost of additional testing.

The system considered and model application used in this work represents perhaps one of the simplest, but one with potentially significant economic impacts. Accurate prediction of the MLVSS concentration translates directly into the required bioreactor and secondary clarifier sizes, which represents a major capital expense for any suspended growth biological treatment system. The colloidal organic matter fraction of

the biological process influent wastewater was found to be negligible in this instance, and the dissolved fraction could be characterized based on membrane filtration rather than flocculation and filtration. Note that GLWA serves a large and diverse metropolitan area, and that a significant portion of the collection system consists of combined sewers, leading to significant variations in influent flows, both seasonal and daily, and significant temperature variations given its location in the Northern USA. In spite of these factors, it was found that one set of wastewater characteristics applied over the entire year. Thus, while the precise numerical results determined for this application may not generally apply, the adaptive approach to wastewater characterization and model calibration described here may be more generally applicable.

2.4 Conclusions

An extended wastewater fractionation study conducted at the GLWA WRRF provided the basis to evaluate alternative wastewater characterization campaign designs. An ideal campaign results in a robust characterization of the wastewater while managing the time and resources required to achieve this result. Wastewater characterization must, of course, be viewed in the context of the objectives of the modeling exercise and the potential impacts of improper model development. The following conclusions can be offered based on this study:

1. The characteristics of this wastewater originating from a large and diverse metropolitan area, as assessed based on predicted versus actual bioreactor MLVSS concentration, did not vary on a seasonal basis. This occurred in spite of significant daily and seasonal influent wastewater flows and seasonal temperature variations due to the fact that the collection system included a substantial combined sewer component.

2. Sampling during periods of normal and stable plant operation results in the

most reliable estimates of wastewater characteristics. Increasing the number of samples can help to partially overcome the adverse impacts on sampling results resulting from occasional periods of unusual plant operation, but the best results will be obtained by avoiding, when possible, sampling during unusual operating periods.

3. For this application, around 20 samples randomly distributed over an annual cycle was found to represents a good trade-off between further increasing the number of samples and the gain in precision in the estimation of wastewater characteristics.

4. An adaptive approach to wastewater characteristics measurement consisting of multiple measurement campaigns, each of limited duration, may provide the best results. Sufficient resources need to be devoted to the campaign to allow for sufficient sampling events to ensure that a reliable and robust assessment of wastewater characteristics is achieved.

5. Attention should be paid to the potential for periods of poor model structure, including numerical values of key parameters, when assessing results. Some redundancy in measured parameters (COD, BOD_5, TSS, VSS) can facilitate identification of such periods.

CHAPTER III

Improving Sensor Data Processing in WRRFs by Coupling Data-driven Approaches with Physical Factors

Published as:

Cheng Yang, Glen T. Daigger, Evangelia Belia, and Branko Kerkez. Extracting useful signals from flawed sensor data: Developing hybrid data-driven approaches with physical factors. *Water Research*, 185:116282, 2020

3.1 Introduction

Continued developments in sensor and computer technology have accelerated application of sensor signals to satisfy tightening effluent quality standards and achieve increasing operation efficiency to lower costs [8, 39, 40]. Nearly all water and wastewater utilities have deployed primary (e.g. flow meters, pH sensors etc.) and advanced (e.g. nutrient sensors/analyser etc.) sensors [6]. Small facilities (20,000 Population Equivalents, PE) can generate up to 400 signals in number, whereas large ones produce more than 30,000 [8, 9]. The proliferation of sensors seems promising but being data rich does not necessarily lead to being information rich [8, 40]. Raw signals must be analysed and processed first, so that measurement faults or abnormal pro-

cess situations can be identified and isolated. Post-processed signals must be subject to quality validation and transformed into actionable information to support decision making and operational control.

The idea of processing signals for useful information has existed for decades. Peddie *et al.* [41] showed that an Oxidation-Reduction Potential (ORP) profile has distinctive features associated with cyclic operation of digesters independent of signal magnitude and range. Numerous similar studies further extended this concept with other sensors such as Dissolved Oxygen (DO), pH and ammonia for control purposes [42, 43, 44, 45]. By linking features in signal profiles with known bio-chemical activities, they succeeded in the application of signals with high temporal variability (dynamics). More recently, studies explicitly explored the possibility of mining information from low-quality sensor signals complicated by both dynamics and sensor instability. Schneider *et al.* [46] developed four soft sensors based on unmaintained sensors to monitor the status of sequencing batch reactors. Thürlimann *et al.* [47] used qualitative trend analysis to increase soft-sensor tolerance to sensor drifts.

This research extends previous research to advanced water quality sensors, such as organic matter concentration and composition, which are of growing interest. For instance, identifying the influent source(s) can help better characterize the key inputs for dynamic modelling with the Activated Sludge Models (Rieger et al., 2012). Accurate fractionation of Chemical Oxygen Demand (COD) for various sources can yield more accurate predictions for decision making and control. However, extracting reliable and useful signals from online water quality sensors is more difficult due to influent complexity and sensor instability. Influent complexity is largely introduced by water and wastewater source variability, and their dynamic nature, which directly or indirectly leads to unstable sensor readings. Sensors themselves are also subject to faults and noise due to design limitations and unavoidable situations such as offline, interference and fouling.

This research extends the incorporation of knowledge into signal processing, rather than just signal use as described above. Although numerous data-driven and/or statistics-based techniques are available for processing signals to improve their quality [48, 49, 11], most are derived by the mathematical features (e.g. variance, frequency etc.) inherent to signals. Their outcomes may not reflect, and may even contradict, real-world physical factors (e.g. mass balances, observations). Simply borrowing data-driven or statistical tools from other fields is not sufficient, given the presence of such systematic constraints [8, 40]. Sensor data quality can be significantly improved by incorporating prior knowledge from operators, process experts, physical constraints and phenomenological observations [8, 40, 10, 11]. Additional knowledge from the physical world is defined herein as physical factors, and an approach incorporating physical factors with available data-driven tools is referred to here as a hybrid signal processing approach.

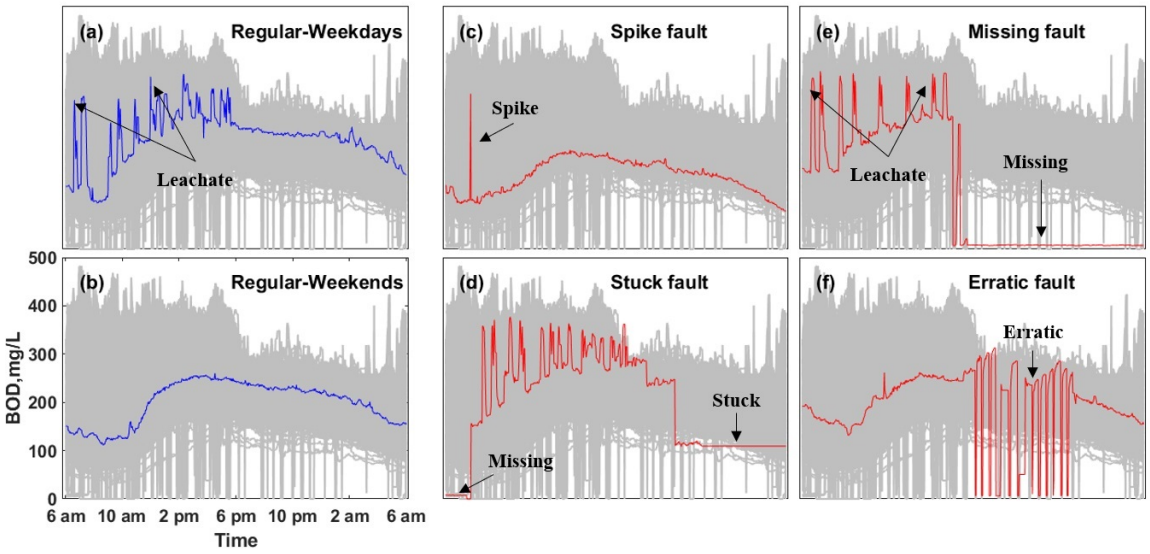

Figure 3.1: Examples of one-day influent wastewater BOD_5 sensor measurements for the Grand Rapids (Michigan, USA) Water Resource Recovery Facility. The shaded areas are all signal profiles stacked together. (a) & (b) are regular signals for weekdays and weekends respectively, referred to as clean/reliable/high-quality signals. During weekdays, landfill leachate is dumped into the WRRF by trucks, causing irregular and highly variant spikes. (c)-(d) are four typical sensor faults regularly observed, referred to as dirty/flawed/low- quality signals.

Figure 3.1 illustrates real-world water quality sensor signals including typical sensor disturbances. For this application a two-year-long raw five-day biochemical oxygen demand (BOD_5) signal was divided into one-day profiles and stacked together, forming the grey area in all six subplots. The signal fundamentally consists of a typical diurnal pattern (illustrated by the regular weekend pattern) but complicated by the discharge of leachate to the plant by trucks during normal working hours on weekdays (illustrated by the regular weekday pattern). Note that the curve formed by the lower edge of the regular weekday signal resembles the weekend diurnal pattern when leachate spikes do not occur. This suggests that actual diurnal patterns should, if not exactly, closely align with the lower edge of the weekday signals. Typical sensor faults (red) also occurred, as illustrated in Figure 3.1 (c)-(f) (definition provided in Section 3.2.1). Due to the similarity of the leachate to some sensor faults (spike and erratic), mathematical features (e.g. variance, frequency, derivatives) are insufficient to clearly separate them.

This dataset provided the basis to investigate the hypothesis that hybrid approaches, as defined above, can lead to improved extraction of reliable and useful signals from flawed sensor data. The hybrid approach was specifically achieved using an assembled system, with each section illustrating at least one aspect of hybridization implementation.

3.2 Materials and Methods

3.2.1 Plant and data

The City of Grand Rapids, Michigan, US, Water Resource Recovery Facility (WRRF) treats an average flow of 38 million gallons a day (144,000 m3/d). The plant inflow not only consists of domestic and industrial wastewater, but also landfill leachate. Landfill leachate is transported on weekdays and discharged to the plant

influent during the normal working hours. Because truck arrival times and intervals between them are random, irregular and variant spikes in the daily influent profile (Figure 3.1 (a) & (b)) are introduced which could be easily confused with sensor faults.

The plant is in its digitalization transformation journey, gradually shifting its decision-making and operations from traditional human-driven approaches (e.g. lab measurement, operators' experience) to data-driven digital solutions (e.g. sensor measurement, dynamic modelling and control). Useful signals with sufficient insights in domain knowledge are the fundamentals for its digital transformation to achieve overall energy efficiency, resource recovery and cost reduction. For instance, in dynamic modelling it is anticipated that the characteristics (e.g. solubility and biodegradability) of landfill leachate are sufficiently different from municipal and industrial wastewater therefore, it may be useful to separate these streams to better optimize the downstream treatment processes.

The sensor signals in this study were BOD_5 concentration measurements of plant influent by a LiquID™ Station unit (ZAPS Technologies, LLC.), with 720 measurements per day. The dataset covered from 2016/05/05 to 2018/5/14 (688 days & 495,360 measurements). Daily flow proportioned composite BOD_5 samples were also collected and analysed (688 days).

The most common sensor fault symptoms presented in Figure 3.1 (c)-(d) follow the definitions provided by Jan *et al.* [50], specifically:

- Missing fault – sensor output is missing for a period of time. All missing values in signals are replaced with zeros to ensure signal continuity.
- Erratic fault – variance of the sensor output significantly increases above the usual value.
- Spike fault – spikes are observed in the output of the sensor at fixed intervals.

- Stuck fault – the sensor's output gets stuck at a fixed value (also known as frozen signal).

One author manually labelled all sensor profiles into two categories, and labelling was verified by another author independently before quality classification implementation. It is recognized that manual labelling can lead to biases in subsequent analyses as deviations between the predicted label and the label provided by the human expert is implicitly assumed to be due to the inadequacy of the model and not due to error in the label provided by the human expert. The potential for such biases must be carefully examined. The labels used to train and test the classification algorithms were:

- **Clean signals**: a one-day signal profile that has no or only minor sensor faults of the above types. Disturbance is solely caused by landfill leachate. These represent 72% (505 out of 688) of the available daily profiles.

- **Dirty signals**: a one-day signal profile that has obvious presence of any of the above sensor faults. Disturbances appear to be mainly caused by sensor instability. These represent 28% (187 out of 688) of the available daily profiles.

3.2.2 Data analysis methodology

The data analysis methodology, summarized in Figure 3.2, consisted of three steps: (1) pattern separation, (2) quality classification and (3) data validation. All daily profiles were first subject to pattern separation to extract the regular diurnal pattern. The daily profiles then entered the qualification classification process where similarities of diurnal patterns were calculated pairwise. For instance, a N-day diurnal patterns has N-1 pair-wise similarities. Dirty and clean signals displayed different features in pair-wise similarities, based on which the classification algorithm was trained to automatically classify them. Finally, the identified two classes of data

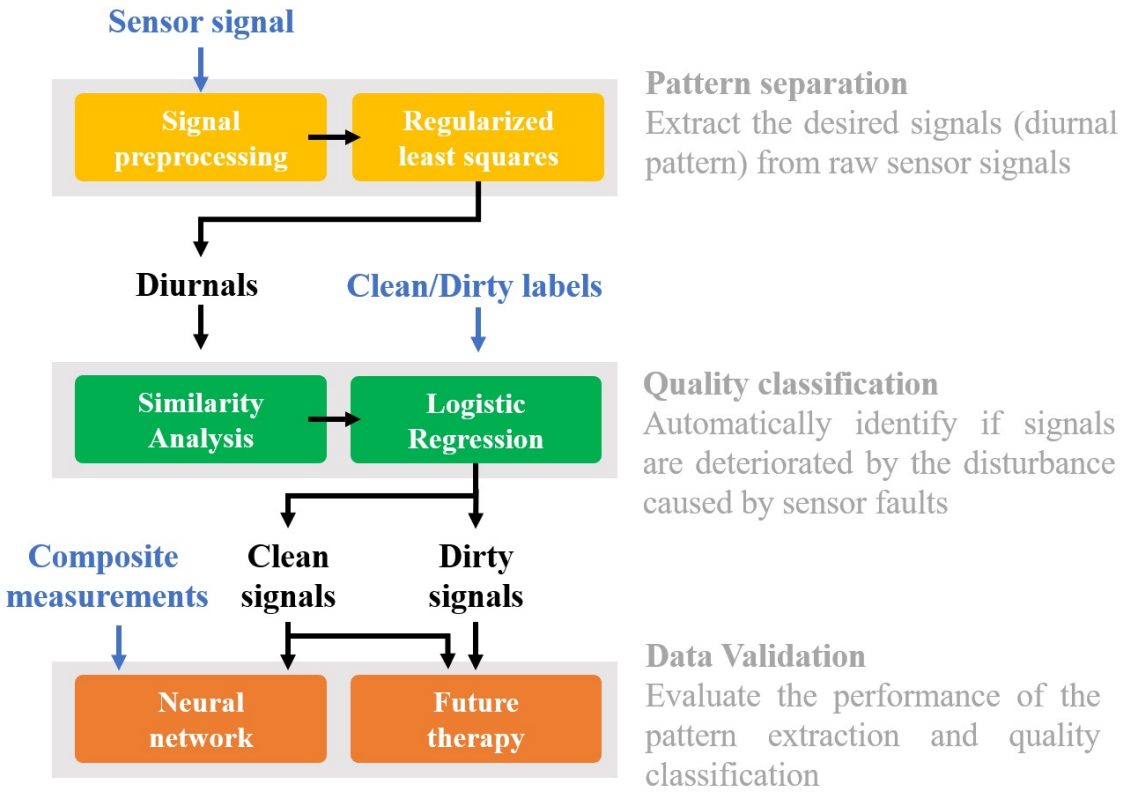

Figure 3.2: Data analysis methodology, with objectives for each section listed. The black arrows represent the process flow, while the blue arrows indicate information provided for each section. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

entered the data validation step to evaluate the success of the pattern separation and quality classification. Details of each step are provided in the following subsections.

3.2.3 Pattern separation

Signal pre-processing consisted of reshaping the signals into daily profiles (signal vectors) and replacing missing values with zeros to ensure signal continuity. The sensor signal vector y, representing 720 measurements of sensor in one day, was viewed as a combination of three parts: (1) the diurnal pattern vector d, (2) the leachate pattern vector l, and (3) the residual vector ε, as showed in Eq. 3.1.

$$y = d + l + \varepsilon \tag{3.1}$$

The pattern separation problem was thus framed as finding the proper diurnal pattern d and leachate pattern l so that the residual ε was minimized. This is a typical least squares problem as formulated in Eq. 3.2. Two physical factors were translated and embedded into the least squares problem: (1) given that the diurnal pattern d has a periodical feature, it was fitted with a 3^{rd} order Fourier series $d = \Gamma x$, where x is the Fourier coefficients vector and Γ is the matrix with sin-cosine elements and (2) the non-negative constraint for the leachate component. Although no obvious pattern exists in leachate, it should be non-negative by nature. To this end, regularization was applied using the shifted Huber-Hinger error function $h(l)$ [51], which penalized the negative elements in l, and forced them to be non- negative.

$$\hat{d}, \hat{y} = \arg\min_{d,l} ||\varepsilon||_2^2 = \arg\min_{d,l} ||y - d - l||_2^2 \tag{3.2}$$

With these specifications, this problem was transformed into a regularized least squares problem, as in Eq. 3.3, where λ is a hyperparameter and is tuned to 0.5 for better performance in this case. Solving Eq. 3.3 required mathematical techniques

which have been provided in the Appendix A.1. Finally, these patterns were identified based on the daily profiles provided.

$$\hat{d}, \hat{s} = \arg\min_{d,l} ||\varepsilon||_2^2 = \arg\min_{x,l} ||y - \Gamma x - l||_2^2 + \lambda \cdot h(l) \tag{3.3}$$

3.2.4 Quality classification

The recurring pattern, typical for water and wastewater systems, was the physical factor used for this purpose. Although daily patterns vary somewhat, the influent daily pattern should be similar in shape if sensor faults are absent. Pearson correlation, as in Eq. 3.4, was used here to quantify similarity. The y_i and y_j are the diurnal pattern vectors for two different days, Day i and Day j, and n is the length of y. The $\bar{(\cdot)}$ is the mean value of that day. The closer the correlation to 1, the higher the similarity. An additional benefit of Eq. 3.4 is that the correlation has been normalized by the standard deviations of the two individual days.

$$\rho(y_i, y_j) = \frac{\sum_{k=1}^{n}(y_{i,k} - \bar{y_i})(y_{j,k} - \bar{y_j})}{\sqrt{\sum_{k=1}^{n}(y_{i,k} - \bar{y_i})^2}\sqrt{\sum_{k=1}^{n}(y_{j,k} - \bar{y_j})^2}} \tag{3.4}$$

Similarity was analysed pairwise between every two days, and the distributions of pairwise similarity were used as predictors for clean/dirty classification. Each day's similarity distribution was plotted into ten-bin histograms, and the frequencies of each bin were fed into a logistic regression model as inputs. An additional variable (stuck index- see Appendix A.2.1) quantifying the severity of the stuck fault was used to improve classification accuracy. The classification model was trained based on 5-fold cross-validation. Results were examined by standard classification evaluation methods, including confusion matrix and Receiver Operating Characteristic Curve (ROC curve) [52]. Specifically, the notations for the confusion matrix were:

- True positive - Dirty signals are predicted as Dirty

- True negative - Clean signals are predicted as Clean
- False positive - Clean signals are predicted as Dirty
- True positive - Dirty signals are predicted as Clean

3.2.5 Data validation

The results of conventional water quality analyses [26] were used to further validate the processing steps and increase confidence in extracted signal quality. The Grand Rapids WRRF routinely collected and analysed flow-proportioned composite influent wastewater samples for BOD_5 in its laboratory. Two approaches were used to convert sensor data into estimated 24-hour composite BOD_5 values: (1) direct calculation using available flow data, defined here as the flow-weighted model and (2) an Artificial Neural Network (ANN) model. Details for the neural network model are provided in the Appendix A.3. Both models estimated the 24-hour composite data, and their performance was used to evaluate data quality for both the clean and dirty data sets. Model performance was evaluated by standard regression evaluation metrics, such as Root Mean Square Error (RMSE). Finally, the problematic portion of the dirty signals was remedied and then fed into the developed neural network model. Algorithms achieving data remediation which the authors refer to as the future therapy showed in Fig. 3.2. In this paper, only preliminary results of data remediation were achieved and displayed.

3.2.6 Implementation

The entire solution developed in this paper, including the full source code and implementation details, are available on an open-source public web repository. While the authors are not at liberty to share all of the raw sensor data, an anonymized example data set is included in the web repository to allow others to evaluate our

approach and implementation. Users should also be able to apply the solutions, with necessary modification, for their own sensor measurements. All analyses were carried out on a Windows OS laptop, and the code was written in MATLAB (2019b edition). The code for regularized least squares used for pattern separation was written by the authors, and the classification and the neural network regression were implemented with Statistics and Machine Learning Toolbox™ and Deep Learning Toolbox™, respectively, in MATLAB. Most of the results are available in the repository under the file '688results'.

Repository: http://github.com/ChengYangUmich/WRsubmission.

3.3 Results

3.3.1 Pattern separation

Several traditional, purely data-driven algorithms, including but not limited to Ordinary Least Squares (OLS), Lowpass filter and Principle Component Analysis (PCA), were used to extract the diurnal pattern. As illustrated by the typical example presented in Figure 3.3, the diurnal patterns extracted using these algorithms were not satisfactory. Specifically, the diurnal pattern extracted by OLS (orange) consistently overestimated, whereas the lowpass filter (green) produced overestimated portions after 2 a.m. and in the daytime when leachate dumping occurred. The diurnal pattern extracted by PCA (purple) was even worse, with large bumps at 7 a.m. and deviations during 10 a.m. to 6 p.m. Theoretically, after processing, the residual signals should mainly be the leachate component, however, it is obvious that the residuals have significant negative values, which contradicts the physical conditions that leachate mass should be positive. In other words, the separated diurnal pattern is a biased estimate. This occurred because these algorithms fitted the data based purely on mathematical features (e.g. frequency, variances), neglecting physical constraints.

The hybrid algorithm, regularized least squares, extracted a more reliable diurnal pattern (blue), which was free from the impacts of the leachate pattern. While Figure 3.3 illustrates the performance of a variety of standard data-driven signal process algorithms, several others were also evaluated with similar poor performance (data not shown).

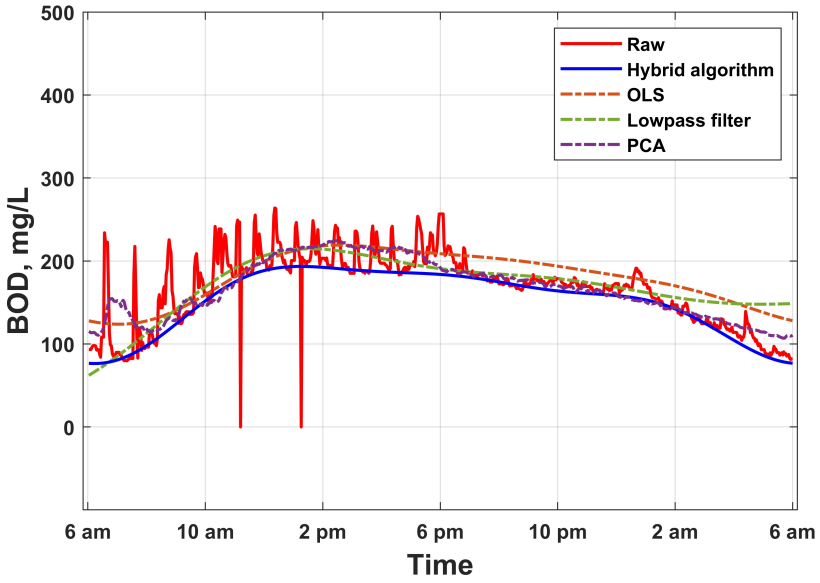

Figure 3.3: Comparison of different pattern separation algorithms.

The performance of the hybrid algorithm is further demonstrated for clean data in Figure 3.4(a). The diurnal pattern (blue) aligns well with the bottom of the original signal (red). As for the separated leachate pattern (black), it occurred mostly in the daytime when leachate dumping happens, and the positive nature of leachate is well-preserved. Although occasional negative values appeared, their magnitude was quite small and acceptable. The separate leachate patterns were further validated by plant personnel by comparing the timing of spikes and trunk entry recordings. Diurnal patterns were also extracted for dirty data, as illustrated in Figure 3.4(b). Distortions were mainly caused by sensor faults, with the algorithm appearing to track the diurnal pattern reasonably well, except in proximity to the onset of the fault. It returned to good correspondence when the fault ended.

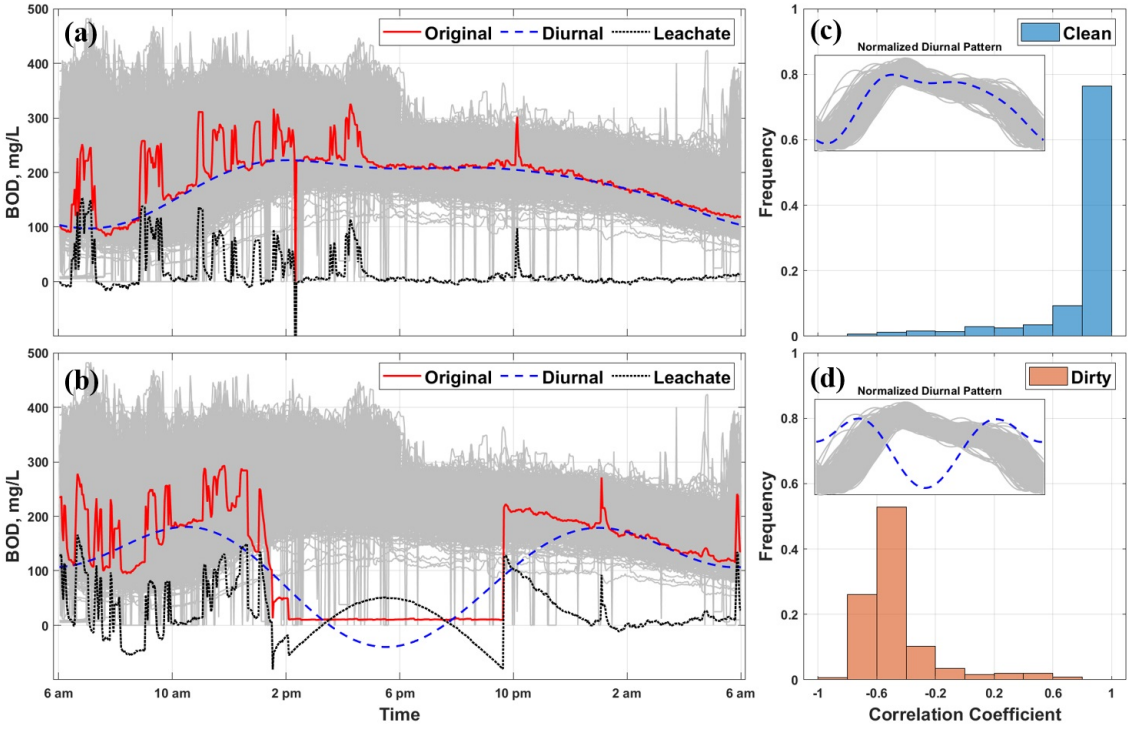

Figure 3.4: Two pattern separation results and their corresponding pairwise similarity distribution for diurnal patterns: (a) & (c) Clean signals; (b) & (d) Dirty signals. (c) and (d) further compare the individual diurnal profiles to all diurnal profiles of clean signals (grey field). All separation results have been provided in the public web repo: http://github.com/ChengYangUmich/WRsubmission.

3.3.2 Quality classification

With removal of the influence of leachate spikes, the remaining diurnal patterns with large dissimilarity to others likely suffer from major sensor faults and should be labelled as dirty days. Each day had 687 pairwise similarities, forming a population of similarities. Traditionally, in statistics, mathematical features of the population such as minimum, maximum and mean are used for classification. For instance, if one day has a population mean much smaller than others, it is classified as an outlier/abnormal. Therefore, the next step was to determine: (1) which mathematical feature should be chosen, and (2) which value of that feature is used for classification.

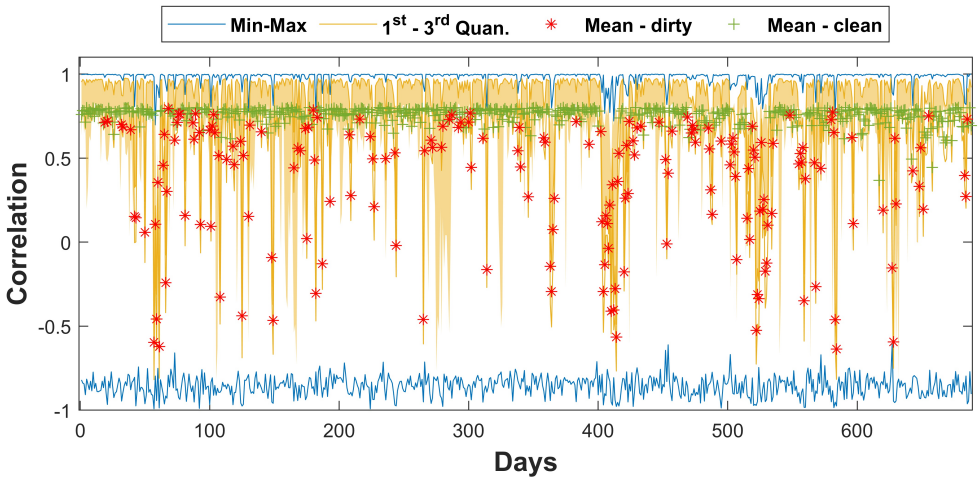

Figure 3.5: Statistics of pairwise correlations for all diurnal patterns, including both clean and dirty data.)

Typical statistics are displayed in Figure 3.5. The blue lines are the minimum and maximum of pairwise similarities, whereas the yellow intervals are the 1^{st} quantile to 3^{rd} quantile. The red stars and green crosses are the mean similarity for dirty and clean signals, respectively. The extrema were not robust choices because, if the same sensor fault was repeated for several days, for instance because sensors are repeatedly offline for maintenance at certain times, those days would have similarity maxima as high as regular days. This occurred frequently, as many days with red stars had no

significant difference in extrema with others (e.g. days in the interval of 200-300 in Figure 3.5).

Means were then investigated. Clean and dirty signals tended to behave differently - the former had a mean at around 0.8, and in contrast, the latter had random values due to different sensor faults. The threshold between clean and dirty signals was trivial, however, as some dirty signals invaded into regions of clean signals (e.g. around Day 80-100) and vice versa (distributed throughout).

Evaluation of quantiles suggested using the whole distribution as the data feature. As indicated in Figure 3.5, days with dirty signals tended to have spiky 1^{st} quantile values, along with deviations in 3^{rd} quantiles. Means were also not lying in the middle of the yellow intervals, indicating highly skewed distributions. Two example distributions are plotted as histograms in Figure 3.4 (c)&(d). The rest have been included in the public web repo (http://github.com/ChengYangUmich/WRsubmission). Distributions for clean signals were exponential-like, with thin tails at the correlation interval of -1 to 0.6, whereas dirty signals had different shapes due to various faults. Not all similarities were concentrated in the last bin because of the existence of minor variations in the daily profiles. Thus, the density of the distribution, which is quantified by the frequencies in the histogram, was used as the classification input, and the decision threshold for the logistic regression was selected based on the ROC curve. With the frequencies of each bin as input, along with the stuck index (explained in discussion), the classification algorithms yielded an accuracy of 95.2%. False positive rate and false negative rate were low, 1.5% (10 out of 688) and 3.3% (23 out of 688) respectively.

3.3.3 Data validation

The daily flow-composite BOD_5 data estimated from the sensor data by both the flow-weighted average model and the neural network model are compared to actual

values in Figure 3.6. Data are compared separately for clean and dirty data. Comparison indicates better performance for the neural network model (red for training and green for testing as defined) than the flow-weighed model as they were closer to the 1:1 line. The dotted lines are the actual value within the range of one standard deviation of measured composite values. In the flow weighed model, a great number of points were overestimated as they exceed the range of one standard deviation of actual values. In contrast, most points fall within the range for ANN. The RMSE, was improved from 52 mg/L to 40 mg/L for the training set and 36 mg/L for the testing set. Composite BOD_5 (actual values) had a range of 182 $\pm$ 43 (54 to 391) mg/L whereas the neural network estimates were 179 $\pm$ 24 (82 to 242) mg/L. In contrast, dirty data showed little correlation with measured values. These results further supported the unreliability of the dirty signals and the validity of the pattern separation and the classification algorithm.

Points deviating from the 1:1 line in Figure 3.6 were found to be signals with major sensor faults. A typical example was provided in Figure 3.7(a). A further test illustrated that the unreliability in dirty signals was due to sensor faults. The unreliable part (missing fault) was remediated based on the information extracted from clean signals and the reliable part of the dirty signals. The remediated signals went through the first two steps in the analytical methodology and were then fed into the neural network. As illustrated in the typical example presented in Figure 3.7(a), the resulting signals appeared to be more reasonable, and the prediction was improved as showed in Figure 3.7(b).

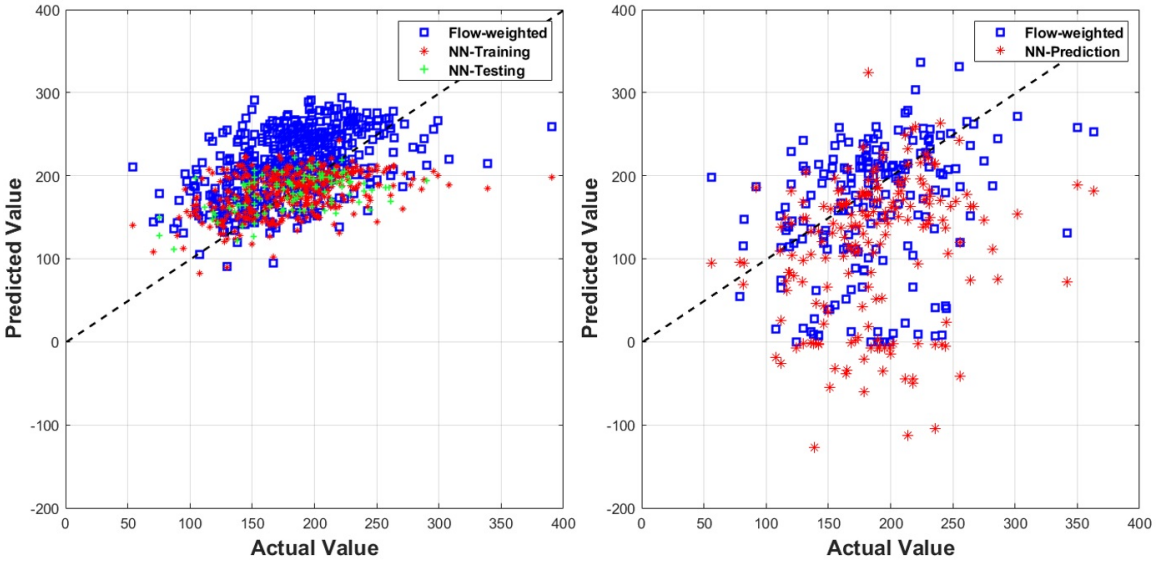

Figure 3.6: Mapping sensor signals into composite measurements. (a) Comparison of artificial neural network and the flow-weighted average. (b) Neural network model predictions with dirty data as input.

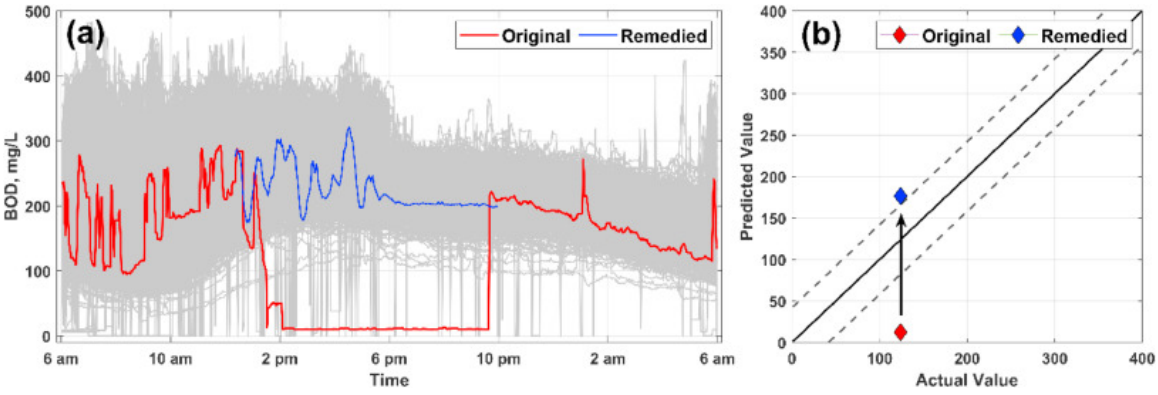

Figure 3.7: An example that problematic part of dirty data could be fixed via therapy algorithm. (a) The reconstruction of dirty data. (b) Prediction improvement achieved by reconstruction, from red (dirty) to blue (remediated).

3.4 Discussion

3.4.1 Hybrid approaches in signal processing

While conventional signal processing techniques are usually sufficient to prepare signals for use, they failed in our case due to the unique characteristics of the signals, as follows:

1. The diurnal and leachate patterns both recur on a daily basis. From the frequency domain view, this indicates their principal frequencies overlap and are hard to separate. Signal filtering is only feasible when distinguishable frequency differences exist between the two patterns.

2. Matrix factorization based on Principal Component Analysis and Independent Component Analysis are two popular pattern recognition and source separation methods [8, 53, 54]. They extract patterns that could explain most mathematical features, including variance and high-order cumulants, respectively. However, the results usually do not assure the inference power for the separated patterns. As showed in Figure 3.3, the bump at the beginning of the diurnal pattern has no physical meaning.

3. Smoothing methods, such as weighted least squares and moving average, are not favourable in this case because it is predictable that, when leachate spikes happen, the separated 'diurnal patterns' will unavoidably be overestimated. This result contradicts the observation in Figure 3.1 that the actual diurnal pattern should align with the lower edge of the signal profiles.

The failure of these conventional techniques led to the exploration of hybrid approaches, inspired by the concept of coupling prior knowledge. Hybrid approaches have been practiced in the water and wastewater engineering field for a long time such as hybrid (grey-box) modelling, soft sensor design and advanced control, but rarely

for signal processing [8, 46, 45, 55, 47, 56, 57]. The hybrid approach was successful, and the following sections will discuss reasons and implications.

Pattern separation – customizing algorithms with physical factors

The success of the pattern separation algorithms relies heavily on hybridizing two physical factors: (1) periodicity of diurnal pattern and (2) non-negative nature of leachate pattern, using a regularized least squares algorithm. The former was abstracted into a Fourier series, a combination of sinusoids with flexible amplitudes and phases but fixed frequencies, while the latter was achieved by applying constraints on the leachate pattern. The implementing algorithm was regularized least squares.

While regularized least squares is widely used in other disciplines to introduce restrictions [58, 59], it has not been used extensively in the water and wastewater field. Instead, other members of the least squares family, such as Partial Least Squares and Generalized Least Squares, are used more frequently [8, 10, 11, 56]. Meanwhile, the theory of Fourier series is well-developed and the non-negative constraints methods are also available [60] and Sutherland-Stacey & Dexter [61] applied non-negative matrix factorization to extract basis spectra from signals from UV/VIS sensors.

Assembling these separate ideas together is innovative and achieved satisfactory pattern separation, which demonstrates the power of coupling physical factors in signal processing. In fact, regularized least squares could be a good beginner algorithm to couple physical factors, thanks to its flexibility. Flexibility arises from the least squares term (the diurnal term) and regularization. The least squares terms can be adjusted with kernel methods or other shapes to embed more assumptions [62, 56, 63]. For instance, a low order polynomial shape could be used rather than a Fourier series to capture gradually changing trends. Various standard regularizations are available for different purposes, for instance, ridge and lasso for screening more important signals and Tikhonov regularization for smoothing [58, 51, 59]. Regularization can also be customized as needed via appropriate error functions, as demonstrated in this

study.

Quality classification – choosing data features that better reflect observations

Numerous similarity measures exist, where two large groups are kernel-based and distance-based [48, 49]. Similarity-based models have been extensively used in literature. For instance, Troutman *et al.* [64] used gaussian processes with a combination of periodic kernel and gaussian kernel to predict the diurnal pattern of missing days. Woo *et. al* [65] used kernel partial least square to overcome collinearity of multivariate features in influent wastewater to better predict effluent water quality. Villez [56] used weighted polynomial regression to smooth signals and obtain the probability distributions of derivatives' signs for later qualitative trend analysis.

It is usually sufficient to directly implement classification based on the values of these similarity measures. However, as shown in Table 3.1, the accuracy based on mean of similarities is lower (reasoned in Section 3.2.3), this is not a good approach in our case. The underlying reasons are: (1) some similar sensor faults repeated for several days, introducing extreme high values that bias the mean and, (2) though close, the similarity of clean diurnal patterns fluctuated within a certain range. Density-based features [48], for instance the number of points within a certain range, better incorporates the above factors and the improvements are shown in Table 3.1.

Different state-of-art classification algorithms were tested but they yielded similar performance. Area Under Receiver Operating Characteristics Curve (AUC) is a standard metric to reveal how well the classifier is able to distinguish between classes [52]. This metric was similar regardless of the algorithm used. The observable performance gain occurred when the stuck index was introduced, which was also instructed by a physical phenomenon. While reviewing misclassified cases, it was found that, if stuck faults coincidentally did not distort the diurnal pattern, they would be falsely identified as clean, as demonstrated in Figure A.5 of Appendix A.4. Introduction of

Table 3.1: Comparison of different algorithms for data qualification.

Algorithms	True positive	False positive	True negative	False negative	Accuracy	Area Under Curve
Directly with similarity values						
Mean	12.5%	0.3%	72.5%	14.7%	85.0%	0.90
Only with similarity distribution as predictors						
Support Vector Machine	20.3%	1.0%	71.8%	6.8%	92.2%	0.93
Decision Tree	21.2%	3.0%	69.8%	6.0%	90.0%	0.90
K nearest neighbour	21.5%	1.9%	70.9%	5.6%	91.7%	0.94
Logistic Regression	21.8%	1.9%	70.5%	5.4%	92.3%	0.95
With similarity distribution and stuck index as predictors						
Logistic Regression	23.8%	1.5%	71.4%	3.3%	95.2%	0.98

the stuck index reduced the cases where dirty signals with stuck faults were diagnosed as clean.

The implication of this section is that appropriate data features are sometimes more important than algorithms. Data features that better incorporate/consider physical factors improve classification performance.

Data validation – choosing appropriate models that users understand

This topic could be framed as a regression problem in which 720 points are weighted and combined into one value. Systematic biases were discovered in the sensor as it tended to overestimate concentrations when leachate spikes occurred (for details see Appendix A.3). Therefore, the flow-weighted model, which places equal weight on every measurement, yielded over-estimations. A natural reaction is to lower weights on those overestimated readings. However, it is hard to manually tune weights and, therefore, machine learning tools, which can automatically select weights by themselves, were used. Neural network models are not critical for this purpose, as other models could equivalently accomplish the auto-tuning function. The authors decided to adopted ANN simply because they understand how ANN works and they are confident that the 'back-propagating' feature of ANN will achieve the goal.

3.4.1.1 Physical factors

The broad definition of physical factors includes but is not limited to prior process knowledge, physical constraints and phenomenological observations, while their contents could be case-specific. Typical physical factors include periodicity, similarity, and mass balance constraints. Table 3.2 summaries the physical factors used in this study and their corresponding actions and improvements. Sources for physical factors include: (1) data mining over raw signals, (2) information from plant operators and professionals, (3) sufficient understanding of systems, (4) input from experts in other disciplines such as statistics and computer engineering, and (5) experimentation.

Table 3.2: Summary of how physical factors were coupled and their improvements.

Physical factors	Actions	Improvements
Pattern Separation		
Observations: • Diurnal pattern aligns with the lower edge of signals • Periods of diurnal and leachate pattern synchronizes **Prior knowledge:** • Periodicity of diurnal pattern • Non-negative nature of leachate	**Customizing algorithms:** • Fourier series • Huber-Hinger regularization • Regularized least squares	**Outcomes:** • Less bias in diurnal pattern • Disturbance caused by leachate separated from sensor faults
Quality Classification		
Observations: • High similarity in shape of clean diurnal patterns and distortions caused by major sensor faults	**Choosing appropriate model:** • Logistic regression • Similarity based models	**Outcomes:** • Increased Classification Accuracy • Fewer false negative cases

• Single statistics in similarities is not sufficient to performance classification • Distributions better represents similarity behaviour • In misclassification cases, stuck faults that incidentally aligned with diurnal pattern were not identified	**Choosing appropriate data features:** • Density based feature - similarity distribution as input • Introduction of a new feature- Stuck-Index	

Data Validation		
Observations:	**Choosing appropriate models:**	**Outcomes:**
• Sensor systematic overestimation in calibration test Experience • Overestimation could be offset by placing lower weights on less reliable measurements • ANN's auto-tuning property- back propagation	• ANN regression	• Improved prediction accuracy • Reduced effort in calibrating sensor system errors

Limited guidance exists on how to couple physical factor with data-driven tools [8]. Based on the experience of this case study, the authors propose three general aspects: (1) customizing algorithms with physical factors; (2) choosing data features that better reflect observations and (3) choosing appropriate models that users understand. The process of coupling is usually not straightforward and requires experimentation.

3.4.2 Improved standard signal processing architecture

Irizar *et al.* [39] proposed a Standard Signal Processing Architecture (SSPA) for online wastewater signals, which has three modules: (1) pre-processing, (2) storage and (3) dedicated post-processing. Raw signals are first sampled and filtered to attenuate the signal noise level in the pre-processing module. The storage module collects data generated in the previous module and analyses it for enriched information, such as time-derivative, average, and variance. Both filtered signals and enriched information are then stored in data repositories for advanced post-processing tools such as monitoring and control.

Based on this work, and as further envisioned by Olsson [40], an improved architecture is proposed in Figure 3.8. Raw signals from probes and analysers are sampled and delivered to the raw data repository for permanent storage (historical data). Sampled signals then go through designed signal processing procedures where target noise-free signals are generated. Both mathematically and physically enriched information is calculated. The former includes standard statistics as listed in SSPA [39], such as average and variance. The latter correspond to physical phenomena, such as the aforementioned similarity, non-negative parameters and stuck index. Quality classification is implemented on the enriched information to differentiate clean and dirty signals as well as to increase confidence of the processed signal. Clean signals go through the validation procedure to be stored in the post-processing repository, while dirty signals could either be discarded or repaired via therapy algorithms for reconstruction.

Once past the validation test, they are also sent to the post-processing repository. Finally, signals for application are directly extracted from the post-processing repository. Throughout the whole process, the Expert Knowledge Support Module is highly involved, including but not limited to determining the sampling frequency, designing relevant algorithms and enriched information extraction.

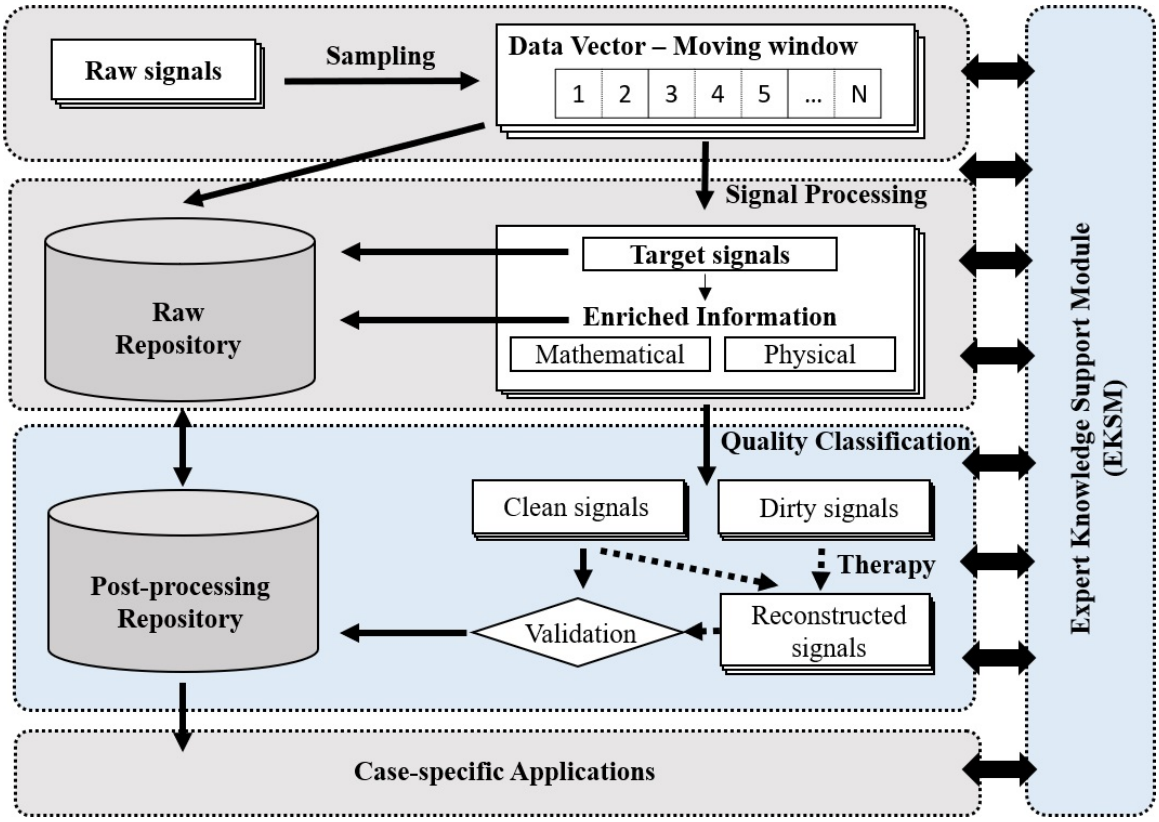

Figure 3.8: General schema of the improved standard signal processing architecture. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Comparing with the SSPA [39], this architecture introduces two new modules, as highlighted in blue. The left added module is to better address the requirements proposed by Olsson [40] – an ideal monitoring system should be able to detect and isolate measurement faults or process abnormal situations (quality classification) and simulate the consequences of operational adjustments (validation). It also provides an alternative solution for dirty signals – instead of burying them in the graveyard,

their residual value is maximized. The added module on the right emphasizes the importance of expert knowledge, involving it in signal processing to maximize the quality of extracted signals. For example, sampling frequency is critical for control purposes [40], and a realistic sampling frequency should be chosen based on sufficient understanding of wastewater dynamics.

3.5 Conclusions

A hybrid approach coupling physical factors with state-of-the-art data-driven tools demonstrated improved ability to extract useful signals from a flawed water quality signal.

Regularized least squares, customized according to physical factors, proved successful in separating diurnal patterns and leachate discharge profiles from sensor signals. In the following quality classification, standard classification metrics (e.g. mean and extrema of similarities) proved insufficient to identify the profiles with quality deteriorated by sensor faults (e.g. missing, stuck, erratic and spike). Classification performance improved using more appropriate data features - the distribution of pairwise similarities and the stuck index. Separated signals were further validated by mapping daily sensor profiles into daily values of 24-hour composite samples, which were measured with conventional water quality analyses. The results indicated that mapping by a neural network model was superior to a direct flow-weighted average model to overcome systematic errors caused by sensor non-linearity. The comprehensive case study demonstrated the ability of a hybrid approach to extract useful signals that comply with validation requirements from flawed sensor data. The success of the hybrid approach provided insights, which led to formulation of an Improved Standard Signal Processing Architecture (ISSPA).

CHAPTER IV

Developing An Adaptive Real-time Grey-box Model with Data Streams in WRRFs

Published as:

Cheng Yang, Peter Seiler, Evangelia Belia, and Glen T. Daigger. An adaptive real-time grey-box model for advanced control and operations in WRRFs. *Water Science and Technology*, 84(9):2353–2365, 09 2021

4.1 Introduction

Wastewater treatment plants are being repurposed to water resource recovery facilities (WRRFs), addressing nutrient recovery and energy neutrality to deal with stricter emission regulation, increasing water scarcity and rapid urbanization [1]. The practice of WRRF design, operation, and control needs adequate models that encode knowledge and information relevant to its designed objectives [12, 1]. In recent decades, the wastewater research community has established generally acknowledged mechanistic and phenomenological models (white-box models) such as the Activated Sludge Model (ASM) family [30, 14], Biofilm Models[67] , and Anaerobic Digestion Model No. 1 (ADM1) [68]. Meanwhile, purely data-driven models (black-box models), mainly focusing on prediction power rather than interpretability, are becoming

more widely developed and used, thanks to advances in big data analytics and artificial intelligence. A comprehensive overview by Haimi *et al.* [10] surveyed and reviewed applications of black-box models in biological processes. While black-box models have often been criticized for their lack of transparency, a mixture of fundamental white-box structure and empirical black-box components has emerged and gained increasing attention, namely the grey-box models. These models integrate components of first-principle models with the data-driven schemes of black-box models for improved ability to estimate unmeasurable variables and kinetics, capture unmodelled dynamics, and predict system performance [10, 11, 1]. Typical integrations include: (1) using black-box models to reconstruct needed but not available information for white-box models, for instance, influent data [69], kinetic parameters [69, 70]; (2) using only partial white-box model structure, relying on data to complete the model (similar to black-box models) [71, 72].

Models are developed for various purposes, including but not limited to design of new plants, upgrade and optimization of existing ones, prediction of future behaviour, education, and process control. The intended use determines the modelling approach and its associated model complexity. For instance, white-box models embedded with as many biochemical reactions as possible are promising teaching tools to educate operators and young water professionals, but not appropriate for real-time control design (e.g., model predictive control) because of their computational intensiveness and unmeasurable parameters. Similarly, black-box models yield good prediction power but are case specific and depend heavily on data availability and quality, in addition to significant effort required for model selection and training. A simply-structured but fit-for-purpose dynamic grey-box model, with data-driven techniques to complete the model, is a candidate solution for real-time prediction and control. One common downstream application is advanced control design based on the developed models. For instance, Stare *et al.* [73] developed a reduced non-linear nitrification model by

modifying expressions in ASM1 for an attached growth pilot plant, with model parameters estimated from real measurements. Different control strategies were then compared based on the identified model. Another common application is soft sensors (also known as state estimators and observers), a virtual asset acting like sensors for prediction and control. Stentoft *et al.*[72] rewrote ASM1 into a simpler stochastic grey-box model and developed an online soft sensor to predict ammonium and nitrate removal in a small recirculating WRRF facility. Nair *et al.* [71] developed a soft sensor based on a grey-box model to estimate volatile fatty acids, phosphate, ammonia and nitrate concentrations based on inputs from inexpensive sensors such as pH and dissolved oxygen. Rich real-time information extracted by soft sensors can improve the efficiency of control and operation.

One potential drawback of a simple grey-box model structure is the need to adaptively update the estimated parameters. Grey-box model identification (the procedure of estimating parameter values from data) often occurs offline, with a limited series of input and output data collected in advance. Estimated parameter values often vary substantially over time in WRRFs, due to slow changes that shift system equilibriums over weeks or months (e.g., biomass, temperature, and wastewater compositions). Adaptive approaches can be used to overcome this issue. One is the Moving Horizon Estimator [74, 75], where the states and parameters are re-estimated periodically after collecting sufficient new measurements. Another is the extended Kalman filter (EKF) [76], where parameter and state estimates are updated efficiently and recursively with new measurements in a continuous mode. Busch *et al.* [77] compared and demonstrated the effectiveness of both approaches in estimating unmeasurable states in the Benchmark Simulation Model No. 1 (BSM1) [78].

In this chapter, the EKF method was selected. Once the grey-box model is equipped with an adaptive scheme, the simplicity of its model structure becomes an advantage for its wide compatibility for other processes, especially for control de-

sign of newly emerging biological wastewater treatment processes whose mechanisms are not fully understood.

It has become a usual practice to develop a white-box model of treatment facilities, often based for instance on the IWA ASMs and ADM1, either as part of the initial design or major upgrade of a facility. If available, the white-box model could be used to simulate a wide range of facility loading and operating conditions, from start-up to design influent flows and constituent loadings, plant operating conditions, and seasonal factors such as wastewater temperature variations. In this paper, such a white-box model was used as a digital twin to develop and evaluate grey-box models. The aim of this paper was to develop an adaptive real-time dynamic model (also known as soft sensors and observers) for advanced control and operations in WRRFs, and evaluate it comprehensively with the various scenarios simulated with its corresponding white-box model. This paper presents the development of the model based on grey-box modelling and EKF. SUMO (Dynamita), an extensively used commercial simulator, was used to simulate a typical and well understood bioprocess, acting like a virtual WRRF to generate data. A grey-box model structure was identified and validated under different scenarios. This model structure was then converted into an EKF to overcome the adaptivity issue. Finally, performance of the EKF was evaluated by comparing the outputs of the EKF-based model to SUMO simulation results.

4.2 Materials and Methods

The realization pathway of this paper is shown in Figure 4.1. SUMO was used as a virtual plant for data generation. Plant performance simulated with SUMO, at design influent flow and loads and different operational scenarios, was used as a reference to evaluate the performance of the grey-box model. The study was divided into two phases: (I) Grey-box modelling, and (II) Implementation of EKF. In Phase

I, input and output data under different scenarios were collected from SUMO simulations, and a grey-box model structure was identified and validated in MATLAB offline. In Phase II, the grey-box model structure was converted into an EKF, the adaptive dynamic model, in Python. Critical steps included discretizing the grey model structure in Phase I and setting parameters to estimate as new states. Influent flow, loads and operations data streams were then generated in Python with noise and fed into SUMO, and performance data streams were retrieved from SUMO, also with added noise. Noise addition was intended to further simulate real sensor signals. The same noisy data streams were used as input to the EKF, and outputs from the EKF were compared with SUMO simulation results. Intuitive information, which requires less professional knowledge to understand, interpret and take actions, was transformed and updated from data produced by the EKF-based model for operations and control.

Phase II can be viewed as an upgrade of Phase I in the following aspects:

(1) Phase II was a real-time implementation while Phase I was offline;

(2) Phase II used EKF to adaptively estimate the parameters in the grey-box model by setting them as new states;

(3) Phase II further reduced the number of required sensor signals for estimation.

4.2.1 Virtual plant

4.2.1.1 Model configuration

The virtual plant process simulated was a typical Modified Ludzack-Ettinger (MLE) activated sludge process, which is widely used for biological wastewater treatment [31]. It consisted of a bioreactor with an anoxic zone, an aerobic zone with three sequential stages, and a mixed liquor recirculation from the end of the bioreactor to

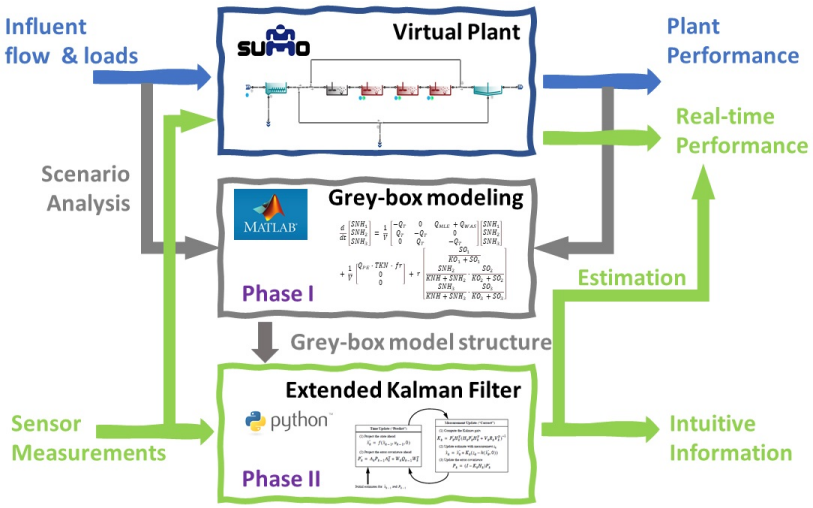

Figure 4.1: The realization pathway of the adaptive real-time grey-box model in this paper. The blue arrows represent data in and out of the virtual plant. The grey arrows represent inputs and outputs of the virtual plant under different scenarios that were used in Phase I for identifying a grey-box model structure. After validation, the grey-box model structure was used to develop an EKF. Simulated sensor data streams were fed into both the virtual plant and the EKF in real-time and outputs from both were compared. Additionally, the EKF generated intuitive information for plant operation and management.

the anoxic zone. The layout as represented in SUMO is depicted in Figure 4.2. The anoxic zone and aerobic zones were represented as Continuously Stirred Tank Reactors (CSTRs) in series, each with a volume of 2,000 m^3. The primary and secondary clarifiers were modelled as ideal separators. The primary clarifier had a suspended solid removal efficiency of 60%, and the secondary clarifier had a fixed effluent solids concentration of 10 mg total suspended solids per liter. The recirculated activated sludge flow rate (QRAS) and the internal recirculation flow rate (QMLE) were set to 23,400 m3/d (roughly 100% of QPE) and 36,000 m3/d (roughly 150% of QPE), respectively. Diurnal patterns are considered in the flow rate, which was observed in primary effluent (QPE). The average value of QPE was around 23,400 m^3/d. Influent total Kjeldahl nitrogen (TKN) varied on a dynamic basis, with an average around 56.6 mg-N/L. Dissolved oxygen levels ($SO_{2,i}$) varied from 0.5 - 3 mg/L. Detailed statistics are provided in Table B.1 in the supplementary material. Biological kinetics were left at their default values (full plant model-SUMO1, Version 20-nb201104), except for the half saturation of ammonia for nitrifiers, which was set to 0.5 mg-N/L, and the half saturation of oxygen for nitrifiers in the three aerobic tanks, which were 0.6, 0.4, 0.2 mg-O_2/L, respectively. Adjustments to half saturations of oxygen for nitrifiers are often found useful in practice and reflect decreasing competition of nitrifiers for dissolved oxygen with heterotrophs because of decreasing heterotrophic activity through the bioreactor (B. Johnson, G. Daigger, personal communication, March 4, 2019). All other settings not mentioned, including the default COD-based wastewater characteristics, were left at the default values in SUMO. A Excel file (openloop.xlsx) including all needed information to reproduce the virtual plant has been provided in the (github.com/ChengYangUmich/SupplementaryMaterialForEKFpaper).

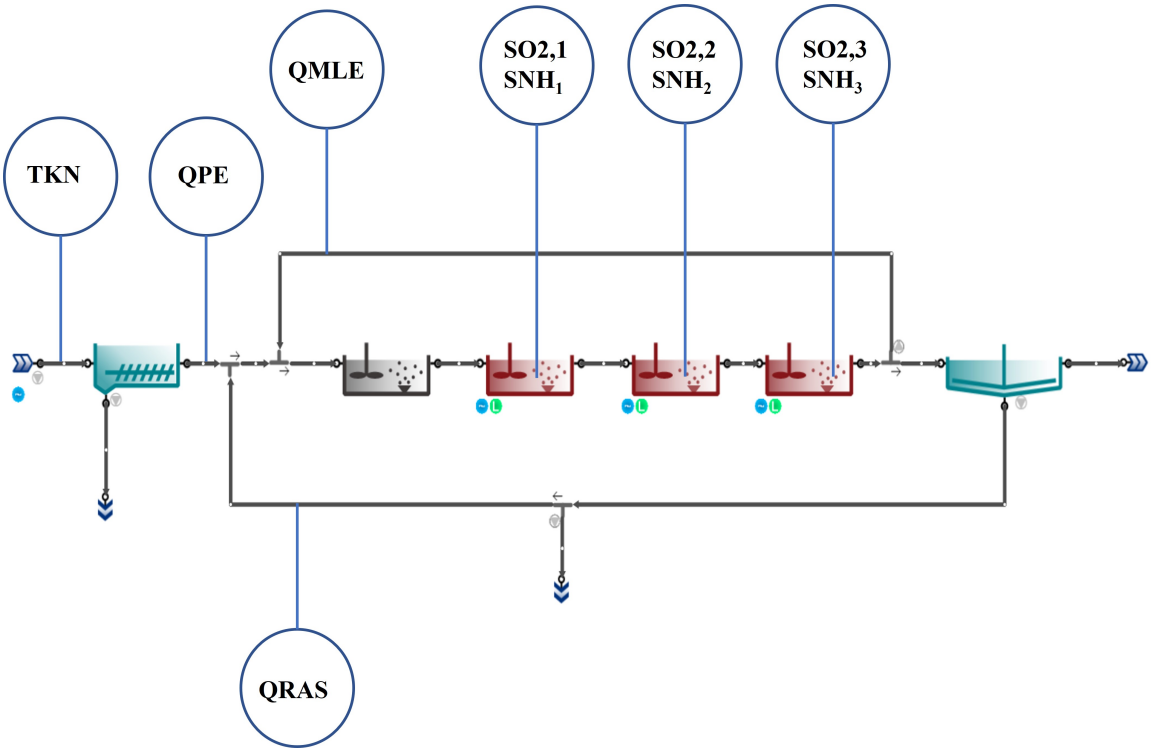

Figure 4.2: The process layout and locations of assumed sensors and meters.

4.2.1.2 Data management

Sensors and meters assumed to be present and used in the analysis are denoted by the circles in Figure 4.2. In total, three flow rates (Q) were measured: (1) the primary effluent, QPE; (2) the internal recirculation from the last aerobic tank to the head of the anoxic tank, QMLE; and (3) the recirculated activated sludge, QRAS. The soluble ammonia into the system was needed for the grey-box modelling. However, in this study, the TKN was used as input because the chosen SUMO influent unit did not include soluble ammonia as input. The soluble ammonia signal was indirectly simulated as 70% of the TKN. In practice, ammonia sensors are used as TKN sensors are not available. Dissolved oxygen ($SO_{2,i}$) and soluble ammonia concentrations ($SNH_{,i}$) in the aerobic tanks were extracted from SUMO for model fitting and evaluation. The sampling interval for all measurements was 10 minutes, and the simulation step size in SUMO was around 1 minute.

4.2.2 Grey-box model structure, identification and validation

4.2.2.1 Grey-box model

The development of a grey-box model typically begins with the general mathematical model for biochemical reactions in one reactor as described by the mass balance for substrates and biological reaction rate, as in Equation 4.1:

$$\frac{dC}{dt} = \frac{1}{V}(Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out}) + \rho \tag{4.1}$$

where:

- $\frac{dC}{dt}$ is the net concentration change rate of the substrate.
- V is the reactor volume.
- Q_{in} and Q_{out} are flows in and out of the reactor.
- C_{in} and C_{out} are substrates concentrations in and out of the reactor.
- ρ is the biochemical reaction rate.

For a white-box model ρ would be a complex function of many other model state variables (such as the nitrifier biomass concentration) and model parameters (such as kinetic and stoichiometric factors). The authors considered a wide number of simplifications of the function as the basis for the grey-box model presented here. Ultimately, the authors selected a highly simplified form consisting of a maximum reaction rate (r) multiplied by relevant Monod functions modifying the maximum reaction rate for DO ($SO_{2,i}$) and soluble ammonia (SNH_i) concentration in each aerobic reactor. The complete model for the three aerobic reactors is depicted in matrix form in Equation 4.2 where one can observe the biochemical reactor rate described above:

$$\frac{d}{dt}\begin{bmatrix} SNH_1 \\ SNH_2 \\ SNH_3 \end{bmatrix} = \frac{1}{V}\begin{bmatrix} -Q_T & 0 & Q_{MLE}+Q_{WAS} \\ Q_T & Q_T & 0 \\ 0 & Q_T & -Q_T \end{bmatrix}\begin{bmatrix} SNH_1 \\ SNH_2 \\ SNH_3 \end{bmatrix} + \frac{1}{V}\begin{bmatrix} Q_{PE}\cdot TKN\cdot fr \\ 0 \\ 0 \end{bmatrix} + r\begin{bmatrix} \frac{SO_{2,1}}{KO_1+SO_{2,1}} \\ \frac{SNH_2}{KNH+SNH_2}\cdot\frac{SO_{2,2}}{KO_2+SO_{2,2}} \\ \frac{SNH_3}{KNH+SNH_3}\cdot\frac{SO_{2,3}}{KO_3+SO_{2,3}} \end{bmatrix} \quad (4.2)$$

where:

- SNH_i, $SO_{2,i}$ are the soluble ammonia and dissolved oxygen concentrations in the i^{th} aerobic tank, respectively, which are measured in these simulations by sensors.
- Q_{PE}, Q_{MLE}, Q_{RAS} are the measured volumetric flow rates as depicted in Figure 4.1, and Q_T is the sum of Q_{PE}, Q_{MLE} and Q_{RAS}.
- TKN is the measured total Kjeldahl nitrogen.
- KO_i, KNH and r are parameters to be estimated, which represent half saturation concentrations for oxygen and ammonia and the maximum ammonia change rate, respectively.
- V and fr are fixed parameters that stand for the reactor volume and fraction of ammonia in TKN.

Essentially, the first term in Equation 4.2 represents internal ammonia transportation between tanks, and the second term is the external ammonia loading into the system. The last term is the ammonia changes due to biological reactions.

Several assumptions were made for this model structure:

- The change of ammonia due to biochemical reactions, including hydrolysis, growth and decay of both heterotrophs and nitrifiers, is combined in the last term, in the form of the maximum rate times the Monod Saturation terms. Non-dominant kinetics are reflected in the estimated KO_i, KNH and r

- The influence of factors such as wastewater composition, active biomass and temperature is implicitly embedded in the estimated maximum nitrification rate (r). Section 3.1 of this paper discusses how this assumption caused model mismatches due to variations in r. The adaptive nature of the EKF proved to be effective in accounting for these variations.

- The ammonia concentration in the first aeration tank is often much greater than the ammonia half saturation coefficient. Therefore, the half saturation expression for ammonia is discarded in the first aeration tank to reduce non-linearity in the grey-box model.

Note that, since the biological reaction term is comprised of a generalized rate, the maximum reaction rate (r), modified by relevant half saturation terms, in its general form it can represent a variety of specific biological transformations. Consequently, Equation 4.2 may be viewed as a general biological reaction model with the constituents and half saturation functions adjusted to the particular biological reaction being considered.

4.2.2.2 Scenario analysis

Several scenarios were used to evaluate the ability of the model structure presented in Equation 4.2 to estimate ammonia concentrations in the virtual plant model and to identify when model parameters needed to be adjusted. A summary of the scenarios evaluated is provided in Table 4.1, and detailed statistics are provided in Table B.1 in the Appendix B. Scenarios 1 and 2 provided a baseline evaluation. Scenario

3 evaluated the sensitivity of model parameter estimations to measurement noise. Scenarios 4 and 5 investigated how estimated parameters changed when factors that are expected to affect r varied, like sludge retention time (SRT) and temperature.

Table 4.1: Input statistics summary of different scenarios

Scenario Index	Flow m^3/d	TCOD mg/L	TKN $mg-N/L$	$SO_{2,i}$ mg/L	Scenario Feature
1	Constant	Constant	Constant	Varied	Constant loadings
2	Varied, Periodic	Varied	Varied	Same as 1	Varying loading
3	Same as 2 + noise				Measurement noise
4	Same as 2				Varying SRT
5	Same as 2				Varying Temperature

Scenario with * indicates parameters were re-estimated with the testing set,when the conditions (temperature and SRT) have changed.

For each scenario, a 7-day period was simulated in SUMO and data was collected and divided equally into two sets: training and testing. The training set was used for parameter estimation and the testing set was used for model validation. The estimated parameters were accepted only when (1) no substantial deviations (within 2 standard deviations for most of the time) in the grey-box modelling prediction were observed, and (2) The Normalized Root Mean Square Error (NRMSE) and R^2 of the testing set were equal or close to that of the training set.

4.2.2.3 Performance evaluation metrics and implementation

Performance of the grey-box model structure was evaluated by comparison of the virtual plant performance and the grey-box modelled values for ammonia concentrations in the three aerated tanks. The metrics used were the goodness of fit (R^2) and the NRMSE, as defined in Equations 4.3 and 4.3, respectively. Greater R^2 and smaller NRMSE generally indicate better performance:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.3}$$

$$NRMSE = \frac{RMSE}{Max(y) - Min(y)} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{Max(y) - Min(y)} \tag{4.4}$$

where:

- y is the true value.
- $\bar{y}$ is the mean of the true value.
- $\hat{y}$ is the estimated value.
- RSS and TSS are residual sum of squares and total sum of squares, respectively.

Estimation of the parameters in the grey-box model was accomplished offline in MATLAB (R2020a) with the System Identification Toolbox Estimate Nonlinear Grey-Box Models (https://www.mathworks.com/help/ident/ref/idnlgrey.html). The search method for estimation was 'lsqnonlin' (Optimization toolbox).

4.2.3 Implementation of the extended Kalman filter

4.2.3.1 Discrete non-linear grey-box model

A general discrete non-linear grey-box model with noise can be written in state-space form as shown in Equation 4.5, with a subset of variables that could be measured as in Equation 4.6:

$$x_{k+1} = f(x_k, u_k) + v_k \tag{4.5}$$

$$z_k = h(x_k, u_k) + w_k \tag{4.6}$$

where:

- $(\cdot)$ denotes variables at the time step k.
- x_k, z_k, u_k denote the vectors of state, observed output, and input at time step k, respectively.
- $f(x_k, u_k)$ is the process model function, which is discretized from Equation 4.2 and $h(x_k, u_k)$is the measurement model function.
- $v_k \sim N(0, Q)$ is the gaussian process noise with zero mean and covariance Q, and the same for $V_k \sim N(0, R)$, the measurement noise.

4.2.3.2 Extended Kalman Filter (EKF)

The Extended Kalman Filter is a recursive method for state-estimation of non-linear process models and measurement models (Eq. 4.5 and Eq. 4.6). It is an optimal state estimator in the sense that it minimizes the error variance. More details about its mathematical origin and practical implementation can be found in [76]. The general steps for the EKF update equations are listed in the sequence of its implementation, as follows:

$$\hat{x}_k^- = f(\hat{x}_{k-1}^-, u_{k-1}) \tag{4.7}$$

$$P_k^- = F_k P_{k-1} F_k^T + Q \tag{4.8}$$

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R)^{-1} \tag{4.9}$$

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - h(\hat{x}_k^-, u_k)) \tag{4.10}$$

$$P_k = (I - K_k H_k) P_k^{-1} \tag{4.11}$$

where:

- $\hat{x}_k^-$ and $\hat{x}_k$ denotes a *prior* and *posterior* rate.

- P_k^- and P_k are the covariance matrices of a prior and a posterior estimation error.

- K_k is the Kalman filter gain and I is the identity matrix.

- F_k and H_k are the Jacobian matrix of partial derivatives of f and h with respective to x, which are evaluated with estimates x and inputs u at time step k, that is, $F = \frac{\partial}{\partial x} f(x,u)|_{x=\hat{x}_k, u=\hat{u}_k}$ and $H = \frac{\partial}{\partial x} h(x,u)|_{x=\hat{x}_k, u=\hat{u}_k}$

- Q and R are covariance matrices of the process and measurement noises respectively. They can be seen as the 'tuning knobs' that determine how much one can trust in the measurements and the process.

4.2.3.3 Online implementation

Determination of variation in the maximum nitrification rate, r, was accomplished by including it as a new state in x, as $x = \begin{bmatrix} SNH_1 & SNH_2 & SNH_3 & r \end{bmatrix}^T$, and Equation 4.1, Equations 4.5 - 4.11 were adjusted correspondingly. In addition, unlike Phase I in which all three ammonia measurements are used ('measured'), the EKF can estimate all states by only measuring SNH_3, therefore, $h(x,u) = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} x$ and $H = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$. In Phase II, noise was added to all measurements before being fed into the EKF.

Communication between SUMO and Python was achieved by the SUMO-Python interface module developed by Dynamita. The Jacobian matrix calculation was accomplished with the Python toolbox SymPy (https://www.sympy.org/en/index.html). The EKF was modified from the toolbox FilterPy (https://github.com/rlabbe/filterpy). The Python scripts used in this chapter are shared in Appendix B.2.

4.3 Results and Discussion

4.3.1 Scenario analysis results

The estimated parameters and model performance metrics are presented in Tables 4.2 and 4.3, respectively. Scenario 1 represents constant ammonia loadings and operations, providing reference information about the nitrification capacity of the system. The estimated parameters were close to the values set in SUMO. Small deviations were expected given that unmodelled dynamics might act as corrections for the kinetic parameters. In Table 4.3, consistently large R^2 and small $NRMSE$ in both the training and testing sets indicated little overfitting. It is important to note that, in general, the NRMSE of ammonia in the last tank is expected to be larger than the previous two. Because the ammonia concentration in the last tank has a relatively smaller range, the same absolute error results in larger NRMSE after normalization.

Table 4.2: Estimated parameters under different scenarios

Scenario Index	**r**	**KNH**	**KO_1**	**KO_2**	**KO_3**
TRUE	-	0.5	0.6	0.4	0.2
1	-325	0.49	0.56	0.32	0.18
2	-334	0.48	0.58	0.30	0.17
3	-321	0.35	0.50	0.30	0.10
4	-322	0.46	0.58	0.30	0.10
4*	-244	0.45	0.52	0.34	0.20
5	-323	0.46	0.57	0.30	0.10
5*	-303	0.50	0.63	0.37	0.27

Scenario with * indicates parameters were re-estimated with the testing set,when the conditions (temperature and SRT) have changed.

Table 4.3: Grey-box model performance under different scenarios

Scenario Index	SNH_1				SNH_2				SNH_3			
	Training Set		Testing Set		Training Set		Testing Set		Training Set		Testing Set	
	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE	R^2	NRMSE
1	0.92	0.10	0.93	0.10	0.94	0.08	0.94	0.09	0.91	0.13	0.90	0.23
2	0.97	0.06	0.97	0.06	0.97	0.07	0.97	0.05	0.95	0.07	0.96	0.33
3	0.97	0.06	0.97	0.06	0.96	0.07	0.97	0.06	0.95	0.13	0.95	0.53
4	0.97	0.06	0.88	0.24	0.97	0.06	0.86	0.32	0.95	0.08	0.84	1.08
4*	0.95	0.23	0.93	0.08	0.91	0.48	0.92	0.09	0.86	1.37	0.93	0.29
5	0.97	0.06	0.94	0.11	0.97	0.06	0.93	0.16	0.96	0.08	0.90	0.82
5*	0.97	0.10	0.95	0.07	0.96	0.17	0.94	0.08	0.94	0.35	0.92	0.35

Scenario with * indicates parameters were re-estimated with the testing set,when the conditions (temperature and SRT) have changed.

Scenario 2 investigated model performance when ammonia loadings varied. The loading patterns are shown in Figure 4.3(a). The influent flow (QPE) was designed to have a daily pattern with small shifts, and the TKN concentration fluctuated. Other inputs were kept the same as Scenario 1. Grey box model performance is shown in Figure 4.3(b) for both the training data and the testing data. By visual inspection, the model captured system dynamics, as further supported by the metrics in Tables 4.2 and 4.3 (large R^2 and small $NRMSE$). Scenario 3 added noise to all measurements to investigate the parameter estimation sensitivity to measurement noise. Although differences were found in estimated parameters (Table 4.2), model performance was acceptably good by visual inspection of Figure B.1 and similar R^2 and $NRMSE$ as Scenario 2 in Table 4.3.

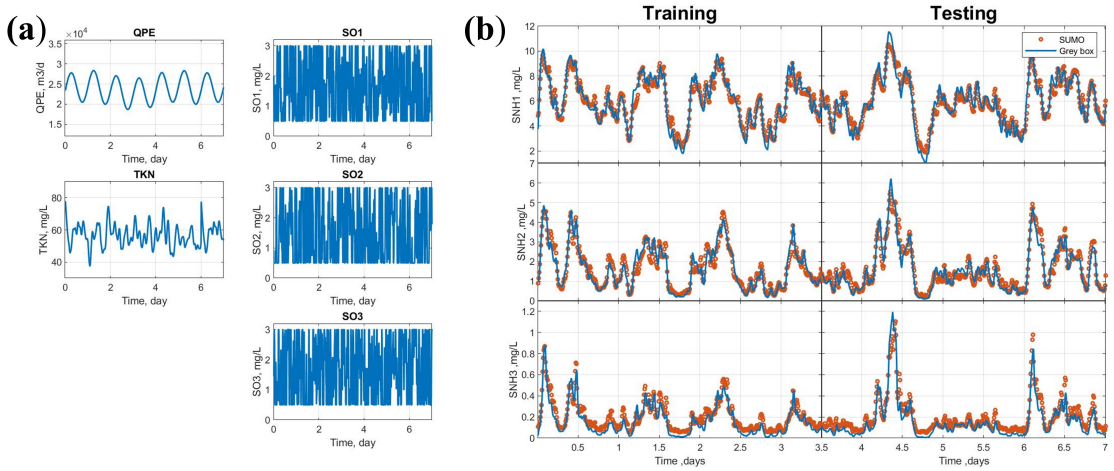

Figure 4.3: An example of scenarios when the grey box model performs well (Scenario 2): (a) model inputs. QMLE and QRAS are flat lines therefore are not shown. (b) Predictions.

Scenarios 4 and 5 investigated the effect of parameters known to change system performance, specifically temperature and SRT, on grey-box model performance. These scenarios investigate situations where the wastewater temperature (Scenario 4) or SRT (Scenario 5) in the actual treatment plant (simulated here using the virtual plant in SUMO) changes. The grey-box model is calibrated to system performance for the previous operating condition (higher temperature or higher SRT) during the

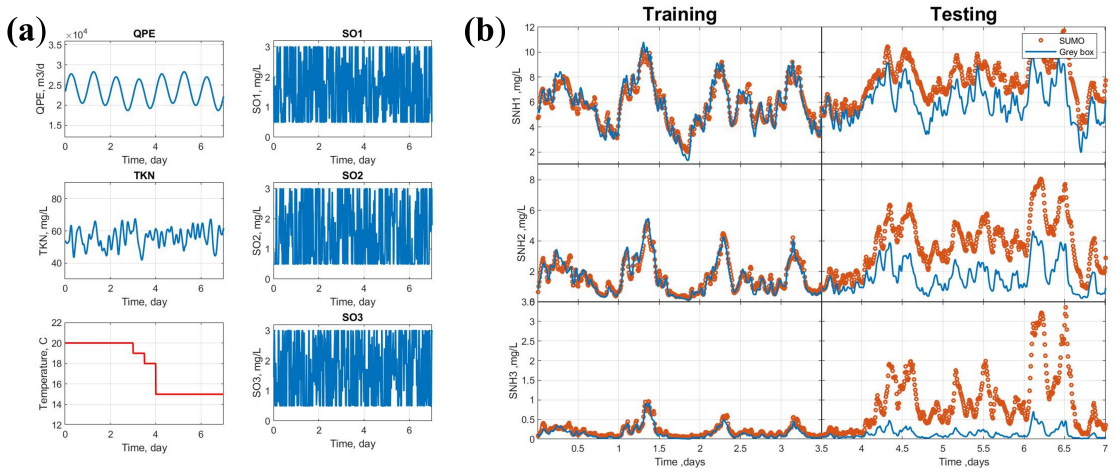

Figure 4.4: An example of scenarios when the grey-box model fails (Scenario 5): (a) model inputs. QMLE and QRAS are flat lines therefore are not shown. (b) Predictions of three ammonia concentrations. Training set (early 3.5 days) is for parameter estimation.

training period and then the performance of the previously calibrated grey-box model is evaluated compared to the performance of the virtual plant when the temperature or SRT decreases. The results indicated that model performance deteriorated when the system temperature and SRT changed. This is illustrated for Scenario 4 in Figure 4.4. When the temperature decreased, the parameters estimated from the training set (20 °C) no longer remained valid for the testing set (15 °C), as large deviations were observed between the SUMO simulation and the grey-box model prediction in the testing set. The R^2 dropped and $NRMSE$ increased dramatically for the testing set, as indicated by the results presented in Table 4.3. Grey-box model performance returned to previous levels when it was calibrated to the virtual plant performance for the altered operating conditions (lower temperature or SRT), as illustrated by Scenario 4* (results for the test data set are shown in Figure B.3). However, in the training set, the grey-box model performance deteriorated for the 3.5 days prior to the temperature change. The major difference in estimated parameters for the grey-box model prior and after the change of temperature was the maximum ammonia reaction rate, r, whose absolute value dropped from 311 to 244 $mg - N/L/day$, im-

plying that the system had smaller nitrification capacity. This is a logical outcome as it is commonly acknowledged that the correlation between biomass activity and temperature follows $r_T = r_{20} \times \theta^{(T-20)}$,where θ is between 1.03 to 1.1 in the literature [31, 16, 14]. In this study, θ equalled 1.05, obeying the temperature dependence. Similar results were observed when the SRT decreased (Figure B.2 and B.4). These results suggested that the parameters re-estimation of the grey-box model to plant data can detect changes in system dynamics (input-output mapping), as expressed in r. With different values of r, the same inputs (DOs) result in different outputs (ammonia concentrations). In this case, the change in r is expected and interpretable with known process knowledge. What was needed was recursive estimation of the time-varying model parameter, which motivated development of the EKF.

4.3.2 Performance of the extended Kalman filter

The EKF was implemented in real time in the SUMO–Python interface for different loading and operating conditions. An example, demonstrating the ability of the EKF to track the maximal nitrification rate, is presented in Figure 4.5. The temperature decreased from 20 to 15 °C on day 5, without changes in other conditions. Other examples are provided in Figure B.5 and Figure B.6. Noise was added for every other measurement. Another difference from Phase I was that only one ammonia measurement (last tank, the green line in Figure B.5(b) bottom left panel) was fed into the EKF.

In Figure 4.5(b), reasonable estimations of the ammonia concentrations in the previous two tanks were observed as the estimation curves converged to simulated 'true' values in SUMO. The noisy fluctuation is due to measurement noise propagating through the EKF. Fine-tuning or extra filtering may further reduce such noise. Moreover, the maximal nitrification rate, r, followed the trend of the temperature drop as the curves started dropping at around day 5, indicating a good tracking of

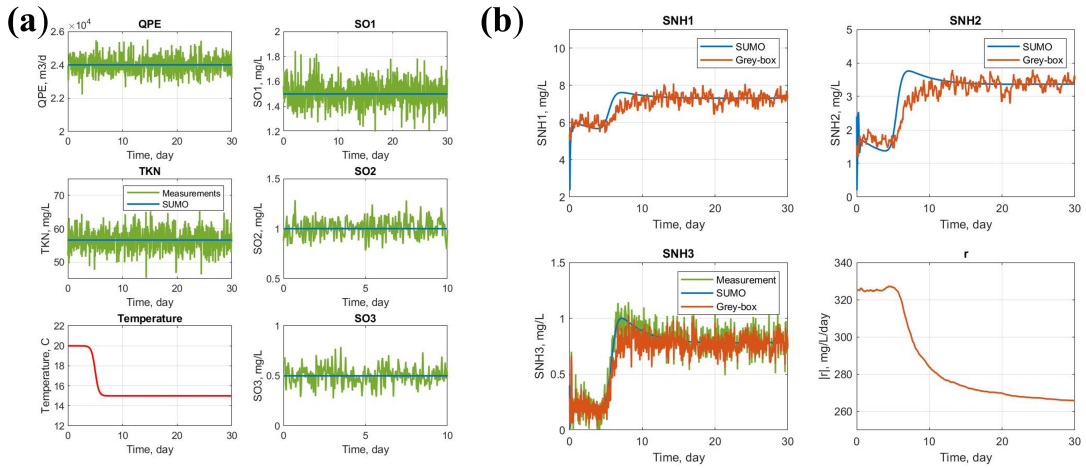

Figure 4.5: The performance of EKF when temperature drops. (a) Model inputs. (b) Estimation of three ammonia concentrations and maximum nitrification rate, r.

the r value.

Performance of the EKF was satisfactory in the sense that:

(1) It was able to provide reasonable estimates of the ammonia concentrations and parameters in all three tanks with fewer signals. This can help reduce cost in practice because the number of sensors employed directly relates to installation and maintenance cost.

(2) It recursively re-estimated r in real time and, therefore, remedied the deficit shown in Section 4.3.1 (the estimated parameters were no longer valid when system equilibriums changed).

(3) It enabled downstream applications, for instance advanced control design like model predictive control and full-state feedback.

4.3.3 Considerations for the implementation of extended Kalman filters

4.3.3.1 Valid grey-box model structure

The structure of the grey-box model needs to be validated before implementation of the EKF, as was done in this study. A suggested approach would be to start with

the full white-box model and remove less relevant components step by step based on appropriate assumptions [71, 79]. Another approach is to start from the major biochemical reactions and add extra components if models exhibit lack of fit with plant data [73]. Grey-box model structure should be fit-for-purpose. In this study, the aim was to model nitrification alone, therefore, minimal but sufficient model components were used. For other processes such as simultaneous nitrification/denitrification and biological phosphate removal, more reaction equations and state variables may be needed.

4.3.3.2 Tuning the extended Kalman filter

The convergence rate and trade-offs between process and sensor noise depend on the two tuning parameters, Q and R, which are the error covariance of the state and measurements. In practical applications, these are not known precisely and, therefore, trial-and-error tuning must be expected. In this paper, noise in flows, TKN, and DO propagated through the process model as process noise. Therefore, Q was written as $J_k M J_k^T$ where M is the covariance matrix of inputs (i.e. flows, TKN, and $SO_{2,i}$) and J is the Jacobian matrix of partial derivatives of f with respect to u evaluated at time step k, $J = \frac{\partial}{\partial u} fx, u|_{x=\hat{x}_k, u=\hat{u}_k}$. Therefore, the 'tuning knob' in this paper was M and R, which were chosen from measurement covariance without further fine tuning.

4.3.3.3 Observability test

Observability is a system property which guarantees that, given inputs (u in Equation 4.5) and measurements (z in Equation 4.6), it is possible to estimate the state values. When a system is observable, the EKF can be developed and incorporated into control design as an observer. However, since not all systems are observable, observability tests should be performed in advance. For non-linear systems, observability tests are usually performed on the linear approximation of the system via

conventional methods, such as the Popov–Belevich–Hautus (PHB) rank test [80]. In this chapter , investigations on the observability revealed one ammonia signal was sufficient for full-state estimation. It is beyond the scope of this paper to discuss observability, but Busch *et al.* [77] provide an integrated approach for observability testing for large-scale wastewater treatment plants.

4.3.4 Significance of intuitive information in real-time grey-box model

The EKF is not new in the wastewater sector, but most uses focus on estimating unmeasurable states instead of yielding intuitive information such as the maximum nitrification capacity as in this study [81, 73, 77, 71]. A monitoring system yielding intuitive information is valuable in that:

1. It supports decision making. A well-educated process engineer may quickly translate data into information based on knowledge and experience. Plant operation staff, especially those in their early career, might not have the same level of knowledge. Intuitive information is easier to understand and therefore prompt decision could be made. For instance, redundant capacity in r allows shorter SRT. It should be noted that the proper actions taken in response to a change in r still require basic process knowledge.

2. It assists system monitoring. Even if not used for automatic control, the intuitive information itself is still valuable in the sense of monitoring the system. In this case, the trend of r could be viewed as a soft sensor monitoring the nitrification capacity of the system. A dramatic change in r without explainable causes could be an alert for anomalies, warning operators to diagnose issues in operations and instrumentation.

3. It improves the control design by adaptive tuning of controller gains. The sensitivity of ammonia to dissolved oxygen (DO) relies on the r values. In

other words, the same change in DO results in different changes in ammonia depending on system maximum nitrification capacity as reflected here in the numerical value of r. It was demonstrated in Section 3.2 that r is a time-varying parameter due to changes in temperature and operational settings such as SRT, among other potential factors. A controller tuned for one r value may be inappropriate as the value of r varies. With the adaptively estimated grey-box model, controller gains could be retuned once significant variation in r is observed. Alternatively, plant operating conditions can be adjusted to return to an r value that allows for better control.

The development of the EKF relies on an adequate grey-box model structure. In this study, as shown in Equation 4.2, the grey-box model is relatively simple – mass balance plus reaction rates governed by Monod kinetics. Although the grey-box model lacks many traditional model components (e.g., biomass concentration, COD) when compared with state-of-art ASM models, the adaptive scheme can incorporate impacts from those neglected variables into the r values, ensuring the accuracy of the model. One benefit of this simplicity is the increased applicability to other processes, especially those that are not fully understood yet.

4.4 Conclusions

This chapter presents the development of a grey-box model able to adaptively estimate the nitrification capacity of an MLE process in real time. Although simple, the grey-box model structure was designed with target information embedded. The model was completed by estimating its parameters from data and was validated under different scenarios. Results of scenario analysis revealed the need to update parameters adaptively to address the changing system dynamics.

An EKF was therefore developed with the grey-box model structure. With Python

interfacing SUMO, a widely used process simulator, EKF performance was evaluated in real time. Results showed that the EKF was able to observe and track the nitrification capacity accurately with fewer sensor signals. Such an adaptive real-time model is valuable in that: (1) it provides intuitive information for decision making on operations; and (2) it enables advanced control design (e. g. model predictive control and state feedback control) and adaptive tuning for controller gains.

CHAPTER V

Automating Process Design and Operations by Coupling Mechanistic Models with Genetic Algorithms

Manuscript submitted as:

Cheng Yang, Evangelia Belia, and Glen T. Daigger. Automating Process Design by Coupling Genetic Algorithms with Commercial Simulators: A Case Study for Hybrid MABR processes. *Water Science and Technology*, 2022

5.1 Introduction

Water resource recovery facilities (WRRFs) design, operation and control are complicated tasks that involve trade-offs between a number of conflicting economic and treatment objectives. Mathematical modelling and simulation play a critical role in evaluating candidate solutions, and engineering judgements are needed to determine a preferred one. Over the decades, process modelling based on the International Water Association (IWA) Activated Sludge Models (ASMs) has become the state-of-art modelling technique [14, 16]. Adapting the first principles in ASMs, commercial simulators and software (e. g. SUMO, SIMBA#, Biowin, GPS-X, BSM1 etc.) implement these models and provide enhanced functionality to assist engineers to

optimize design, operation and control. Typical examples include equipment models (e.g., clarifier model, blower models) and new process units (e.g., membrane biofilm bioreactors, membrane aerated biofilm reactors [MABRs]). Significantly, these updates pose new challenges in WRRF optimization, because of the steadily increasing degrees-of-freedom in models, the expectation of having to simultaneously satisfy a variety of objectives and the need to tackle these challenges along with recognized uncertainty [83, 84]

Generally, these simulators are used in a trial-and-error fashion. A stepwise manual calibration procedure is often adopted, which involves the following steps: (1) the user first selects the wanted processes and layout; (2) the user selects a parameter subset that depends on comprehensive guidelines [16, 67, 83], which aim at minimal calibration efforts; (3) the user runs the model with reference values and compares the results to actual data; (4) the user changes one parameter value at a time and re-runs the simulation and compares the result again; and (5) the user iterates step 4 until sufficient goodness of fit is obtained. This procedure is subjective because the change in parameter values depends on the user's experience, and termination criteria differ based on the user's judgment on different objectives. Meanwhile, the obvious tedious nature of the procedure also limits the total number of simulations performed manually, leading to less efficient exploration in the parameter space.

An alternative and powerful paradigm is to formalize the problem into an optimization question and rely on computers to find the optimal solutions. The user provides formal mathematical specifications for goals and constraints, and a computer agent iteratively completes the trial-and-error steps. Among all the optimization tools, a subfield of artificial intelligence (AI), namely Genetic Algorithms (GAs), attract significant research attention. GAs, which were first introduced by Holland in 1975 [85], search for good solutions by emulating Darwin's natural selection concept on a population of potential solutions. Over the course of evolving genera-

tions, better and better solutions are generated and converged to the optimal. They are proficient in dealing with problems that are non-linear, discontinuous, discrete, non-differentiable, full of constraints and with multiple objectives, as is the case for model-based WRRF problems where classic optimization algorithms (e.g., gradient-based methods, linear and non-linear programming, etc.) often yield poor results [86].

5.2 Current Status of GA Applications in the WRRFs

Early studies directly applying GAs with activated sludge models date back to the 2000s. GAs were coupled with neural network models to optimize fuzzy logic controllers [87, 88]. Coupling GAs with mechanistic models for design and model calibration also started around the same time. Doby *et al.* [89] applied GAs to select the most cost-effective biological nutrient removal (BNR) design, including the presence of primary clarification, the choice from four state-of-arts BNR process configurations and optimal operations. Kim *et al.* [90] applied GAs to calibrate growth kinetics in the ASM1 model and obtained good estimation based on steady-state simulations. Urban and Szetela [91, 92] investigated what output variables are needed in objective functions for GAs to ensure good identification of ASM1 kinetics. These were early proofs of concept demonstrating the potentials of GAs in WRRF applications, even though the process models used were relatively simple and lacked wide applicability. In these studies, the GAs could be viewed as single-objective GA (SOGA) where several objectives (mainly effluent limits and energy cost) were combined into one with different weights.

Demands for more integrated assessments (technical, economic and environmental etc.) led to multi-criteria decision analysis and required new tools to assess trading-offs between various objectives [93]. Consequently, the applications of multiple objectives GAs (MOGAs) quickly developed since the 2010s, such as optimizing pro-

cess designs [93, 94, 95, 96], optimizing process operations and model parameters [94, 96, 97, 98] and optimizing control design [94, 97, 99, 100, 101]. MOGA generates non-dominated solutions where none of their objectives could be improved without degrading others. The collection of non-dominated solutions forms a Pareto front, which allows engineers to solely focus on and pick a most preferred solution by carefully balancing other trade-offs that are not included in the optimization problem.

5.3 Motivations to Couple GAs with Commercial Simulators

In recent years, the prevalence and maturity of WRRF simulators have encouraged more and more engineers to utilize them to support decision making in design, operation and control. At the same time, new treatment processes that offer significantly improved overall environmental performance are being developed rapidly. Full-scale implementation of them requires understandings of numerous interacting components and assessments of apparent trade-offs and potential synergies, which are eventually incorporated into established guidelines. Such understanding has developed for existing processes through their applications over their long history of use. The absence of such understanding and guidelines for the newer generation of process technologies represents an important barrier to their subsequent applications. Process simulation offers an important supplement to experimental investigations and full-scale implementations, which can accelerate the knowledge acquisition needed to achieve effective development and implementation of emerging and evolving processes. Collaborating with users, companies that offer process simulators release innovative process modules for comparative analysis relative to existing technologies. In contrast to the fast development in process simulators, approaches used in their calibration, design and optimization remain relatively static and mismatch the demand created by the increasing degrees-of-freedom and complicated interactions in newer processes. Consequently, artificial intelligence, namely GAs, can be utilized to address the demand

automatically and efficiently in this study.

Although GAs have demonstrated their capabilities in different tasks in the above-mentioned studies, the models used in most of them were either self-written simplified models or based on benchmark models such as BSM1 [78] whose configurations are fixed. Few studies investigated the combination of GAs with commercial simulators. To the best of the authors' knowledge, there is only one study in the literature - Ludwig *et al.* [102] coupled GPS-X with the MATLAB GA toolbox to find the optimal relaxation and filtration times for a membrane plug flow reactor and the solution found by the GA was validated at a technical reference plant.

This chapter aims to further evaluate the power of coupling GAs with commercial simulators to accelerate the learning pace for emerging and complicated biological processes by automating model calibration, process design, and operation optimization using GAs. The chosen process is an innovative and recently commercialized process – Hybrid Membrane Aerated Biofilm Reactor (Hybrid MABR) process [103, 104], whose trade-offs in design and operations are not fully explored yet. The chosen simulator is SUMO (Dynamita) with its corresponding GA toolbox (Geatpy) in Python. No literature has been reported on this conjugation yet. A virtual process upgrade design task was assigned to GAs to find the optimal design and operations for the hybrid MABR process automatically. The task covered key steps of process design, including influent fractionation, process sizing and operation determination.

5.4 Materials and Methods

5.4.1 Hybrid MABR processes

MABRs represent a recently commercialized biofilm biological wastewater treatment process that effectively achieve nitrification and denitrification of wastewater while transferring necessary oxygen much more efficiently than using other oxygen

transfer systems [103, 105]. Air or pure oxygen is supplied to the inside of a gas permeable membrane and diffuses through the membrane and into the attached biofilm, thereby allowing the aerobic growth of bacteria in the biofilm. Substrates in the bulk liquid diffuse into the biofilm from the opposite direction, consequently creating a counter-diffusional process with distinct concentration profiles, resulting in various growth niches for different bacteria.

More recently, MABR units are being incorporated into the anoxic zone of conventional suspended growth BNR processes, allowing the nitrate produced by nitrification in the MABR biofilm to serve as an electron acceptor for heterotrophs in the suspended growth, reducing the need for mixed liquor recirculation for total nitrogen removal [103, 106, 107]. The combined MABR/suspended growth processes are referred to here as hybrid MABR processes, which have advantages such as smaller footprint, less energy cost, higher oxygen transfer efficiency, and better effluent quality[105, 108]. The design criteria for hybrid MABR processes in full scale are not yet fully established however, given the following complicate trade-offs:

- Hybrid MABR processes allow the size of downstream aerobic zones to be reduced as a result of interactions with the upstream nitrifying biofilms [103, 109]. How to allocate nitrification contributions to these two process components to achieve overall cost-benefit performance is still under investigation.

- Higher soluble organics loadings reduce the activity of nitrifying membrane-aerated biofilms [110], which could be relieved by achieving good bio-flocculation by the suspended growth portion. How to minimize sludge retention time (SRT) and size reactors while still achieving needed bio-flocculation and hydrolysis remains unclear.

- The relative fraction of anoxic and aerobic volumes in the suspended growth component determines the final composition of effluent total nitrogen, and such

trade-offs are often observed in simultaneous nitrification and denitrification (SND) processes and are further affected by aeration control. How to choose a ratio and aeration mode that maximizes the use of generated nitrate (electron acceptor) from nitrification with the available carbon source (electron donor) in wastewater is important in design and operation because it reduces external carbon addition and excess aeration.

It is important to note that the abovementioned design considerations are internally connected with each other such that it is not reasonable to conduct a sequential one-by-one investigation and analysis. A simultaneous investigation taking them into considerations is preferred.

5.4.2 A Virtual Process Design Task

A process design task was assigned to GAs to evaluate their ability for assisting in the preliminary process design with SUMO acting as a virtual plant. The task aimed at upgrading a current Modified Ludzack-Ettinger (MLE) process into a hybrid MABR system for better nitrogen removal with smaller footprint and shorter SRT. Following real-world procedures, two major steps, namely influent characterization and process sizing, were implemented in this study to investigate where and how GAs could automate these steps and find an integrated solution for preliminary design.

The plant configurations and operational settings prior to and after the upgrade are displayed in Figure 5.1, where the upper panel is the MLE process, and the lower panel is the equivalent hybrid MABR process to be designed. Dimensions and operations are listed in Figure 5.1 with the to-be-designed parameters marked as red.

The MLE total tank volume is 100 m^3, with a bioreactor configured hydraulically as two anoxic tanks and three aerobic tanks in series. The maximum SRT is 7.7 days to limit the MLSS concentrations to less than 3,600 mg-TSS/L.

In the hybrid MABR system, MABR cassettes (Suez Zeelung 2.0 modules), were

installed in the first anoxic tank. The total media surface area, which determines the maximal ammonia removal in MABR was to be determined by the employed GA. The volumes of the downstream anoxic and aerobic tanks are two additional decisions to make. The SRT of the hybrid MABR system was also assigned to the GA algorithm to optimize. The general design objective was to achieve effluent ammonia and nitrate less than 0.5 mg-N/L and 5 mg-N/L respectively.

Default biological model and kinetics (SUMO1) were used for both steps. Flow rates were fixed throughout the simulations and are shown in Figure 5.1. Influent concentrations and fractions mentioned in Table 5.1, Section 5.4.5.2 and Section 5.4.5.3 were adjusted, otherwise, their values were left at default. Step I was implemented with steady-state simulations. Step II was implemented with dynamic simulations with varying influent total Chemical Oxygen Demand (COD) and Total Kjeldahl Nitrogen (TKN) (Section 5.4.5.3 - Diurnal patterns. Detailed settings for the two processes are provided in the standard SUMO output excel files in the Appendix C).

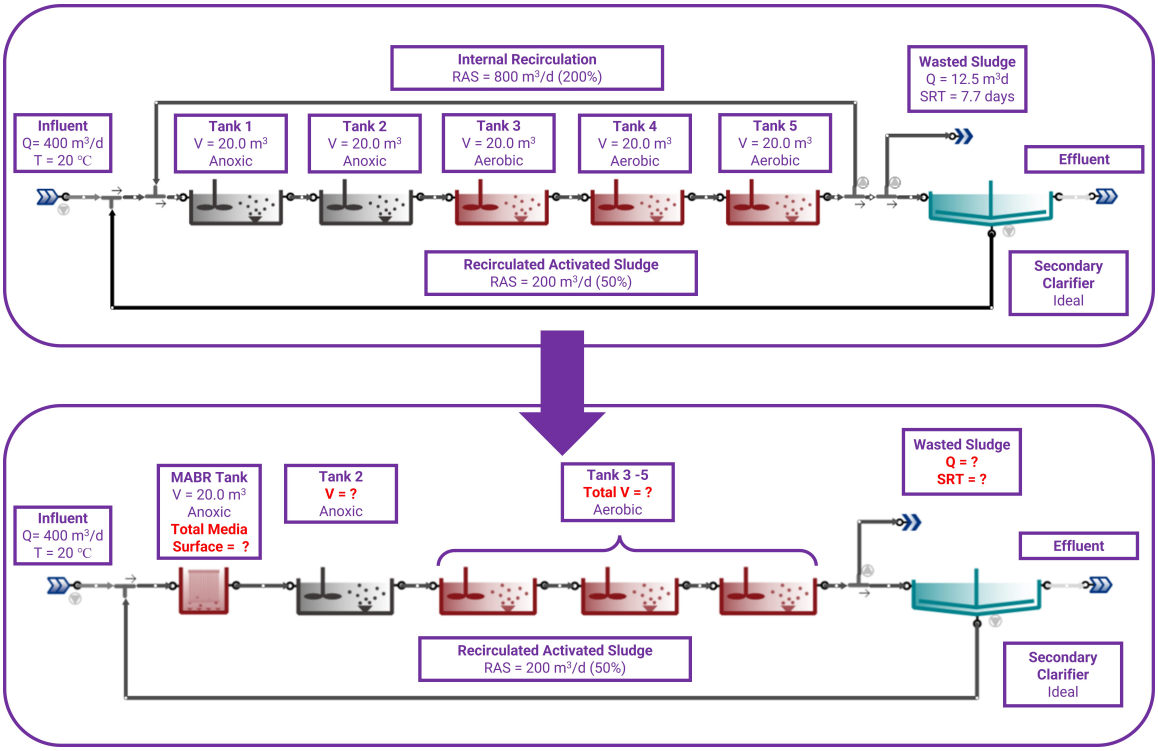

Figure 5.1: The layouts of WRRFs before and after plant upgrade

5.4.3 Step I: Calibration of Influent Fractionation

The first and foremost step to design a biological process is characterizing the biodegradability and fractions (dissolved, colloidal, and particulate) of the influent wastewater organic matters. The profound effect of wastewater characterization on modelling and design has been demonstrated many times in the literature [19, 18, 15, 13]. Fractions substantially influence the influent composition, effluent water quality, oxygen supply & demand and sludge production. Conventionally, wastewater characterization is conducted intensively in a short period of time with limited samples. However, as the authors identified in a previous study [13], these fractions are usually subject to large uncertainties and require much larger sample sizes for a representative estimation. Fortunately, because these fractions are directly correlated with regular measurements in WRRFs, it is possible to reverse-engineer them based on regular measurements. In practice, experienced process engineers often adopt this approach when wastewater characterization data is not available. The objective of this step is to recreate this reverse engineering procedure with GAs to yield a sufficiently good estimation of the wastewater fractions.

The Five influent fractions (listed as the Decision Variables [DVs] in Section 5.4.5.2) used in this study are reasonable values different from the SUMO default. These values are unknowns to GAs, and the GAs need to estimate them by processing regular measurements generated with these fractions in SUMO. The eleven regular measurements were listed in Section 5.4.5.2 - Performance Variables [PVs]. In addition, due to uncertainties in real-world measurements, error tolerances (listed in Section 5.4.5.2 – Error Tolerance [Ws]) were set for PVs. As long as the differences of predictions of PVs were within the error tolerances of the true values, the estimated DVs are considered sufficiently good. Other influent inputs that are not mentioned were left at at SUMO default.

5.4.4 Step II: Hybrid MABR Sizing and SRT Optimization

Once the wastewater characteristics are quantified, the next step is to determine the dimensions and operations for the hybrid MABR process with the same influent fractions as in Step I. In this process nitrification mainly occurs on the biofilm growing on the MABR media whereas the suspended biomass mainly accomplishes denitrification. Nitrifiers are present in the suspended growth because excess nitrifiers grown in the MABR biofilm slough off and accumulate in the suspended growth. Consequently, nitrification can occur in the suspended growth in a polishing aerobic zone to attenuate peak patterns and remove residual ammonia after the MABR. Several important considerations, stated earlier in Section 5.4.1, are re-stated here:

1. Determination of the total MABR biofilm surface area and the last aerobic zone volume to allocate the relative contributions of the biofilm and the suspended growth biomass to nitrification.

2. Selection of the minimum SRT that is required to achieve good bio-coagulation and meet the need of hydrolysis organics for denitrification.

3. Selection of the anoxic/aerobic volume ratio to achieve overall TIN removal.

Manual exploration for optimal solutions on these could be time-consuming and challenging, especially when these considerations are internally connected. Therefore, the objective of this step was to demonstrate that GAs can smartly explore the solution space, taking all considerations together and find feasible solutions for this complicated process, sparing process engineers from an exhaustive search on solution space and finally speeding up the process design.

5.4.5 Problem Formulation

5.4.5.1 General Mathematical Form

In essence, the aforementioned steps can be generalized into an optimization problem – minimizing the objective functions by adjusting the decision variables subject to constraints. Its mathematical form is showed in Eq. 5.1.

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & Obj_1(y_1, f(X)), \cdots, Obj_i(y_i, f(X)) \\ \text{subject to :} \quad & lb \leq X \leq ub \\ & c(X) \leq 0 \end{aligned} \tag{5.1}$$

where:

- X is the vector form of Decision Variables (DVs), which represents the inputs to optimize.
- y_i's are predictions from model simulations, defined herein as Performance Variables (PVs).
- Obj_i's are the objective functions based on DVs and PVs.i is the dimension of the objectives. $i = 1$ indicates a single-objective optimization problem, while when $i \geq 1$, it becomes a multiple-objective optimization.
- lb and ub are the lower and upper bounds for X respectively, and $c(X)$ are constraints that X should meet, where $c(X)$ can take a linear or non-linear form. They are usually derived from physical constraints and process knowledge.

5.4.5.2 Step I: Calibration of Influent Fractionation

Problem Statement: Estimate the five influent fractions that minimize the weighted sum of errors between model predictions (with guessed inputs) and observations (with true inputs) on the eleven regular measurements

Model layout: MLE

Simulation type: Steady state simulation

Objective function: (Single)

$$obj(X) = \sum_{i=1}^{5} \frac{|y_i - f_i(X)|}{W_i} \tag{5.2}$$

Where $f_i(X)$ are the outputs from SUMO simulations and details are listed below:

Decision variables, X (n =5):

- x_1 fraction of VSS in TSS (frVSS_TSS)
- x_2 fraction of filtered COD in total COD (frSCCOD_TCOD)
- x_3 fraction of filtered and flocculated COD in total COD (frSCOD_TCOD)
- x_4 fraction of soluble unbiodegradable organics in filtered COD (frSU_SCCOD)
- x_5 fraction of particulate unbiodegradable organics in total COD (frXU_TCOD)

Performance, y's (n = 11):

- y_1 Influent BOD_5 (Inf_BOD5) (Error tolerance W_1 = 10 mg/L)
- y_2 Influent TSS (Inf_TSS) (Error tolerance W_2 = 5 mg/L)
- y_3 Tank 3 MLSS (Tank3_MLSS) (Error tolerance W_3 = 100 mg/L)
- y_4 Tank3 ammonia (Tank3_SNHx) (Error tolerance W_4 = 0.5 mg-N/L)
- y_5 Tank 3 air supply (Tank3_Qair) (Error tolerance W_5 = 24 Nm^3/d)
- y_6 Tank 4 air supply (Tank4_Qair) (Error tolerance W_6 = 24 Nm^3/d)
- y_7 Tank 5 air supply (Tank5_Qair) (Error tolerance W_7 = 24 Nm^3/d)
- y_8 Effluent BOD_5 (Eff_BOD5) (Error tolerance W_8 = 0.1 mg/L)
- y_9 Effluent Total COD (Eff_TCOD) (Error tolerance W_9 = 20 mg/L)
- y_{10} Effluent ammonia (Eff_SNHx) (Error tolerance W_{10} = 0.1 mg-N/L)
- y_{11} Effluent nitrate (Eff_SNOx) (Error tolerance W_{11} = 1.0 mg-N/L)

Constraints, $C(X)$'s (n = 6):

- (n=5) All fractions are within [0,1], $0 \leq X \leq 1$
- The soluble fraction is smaller than soluble and colloidal fraction, $x_3 - x_2 \leq 0$

Evolutionary Algorithm: Strengthen Elitist Generic Algorithm [111]

5.4.5.3 Step II: Hybrid MABR Sizing and SRT optimization

Problem Statement: Minimize the daily average effluent ammonia and nitrate concentrations given the chosen total media surface area, volume distribution of aerobic and anoxic zones and total SRT for the hybrid MABR system during dynamic simulations. Specifically, the effluent average ammonia concentration should be smaller than 0.5 mg-N/L and the average nitrate concentration should be smaller than 5 mg-N/L. Sinusoidal patterns of influent nitrogen and total COD were used to replicate the diurnal pattern in the simulation. The flow rate was kept constant.

Model layout: Hybrid MABR

Simulation type: Dynamic simulation

Diurnal Pattern:

- **Influent TCOD:** $240 + 40 \times \sin(2\pi t)$
- **Influent TKN:** $34 + 5 \times \sin(2\pi t)$

Objective function:

Two kinds of GAs with different objective dimensions, namely single objective (SOGA) and multiple objectives (MOGA), were tested. Similar to Step I, the first framework includes a single but carefully designed objective function as showed in Eq. 5.3. A penalty function was applied to the effluent ammonia concentration. Basically, when the effluent ammonia exceeds the selected limit (0.5 mg-N/L), a large penalty value proportional to the exceeding amount is added to the objective function. On the contrary, the penalty equals zero when no violation is detected. Eq. 5.3 guides

the GA to be more sensitive to limit violations and corrects the magnitude differences between effluent ammonia and nitrate. In MOGA, 24-h sums of effluent ammonia and nitrate concentrations are straightforwardly used as the final objective function values (Eq. 5.4). For one day simulation, data are collected hourly ($\Delta t = 1h$).

$$obj(X) = \frac{1}{24}\sum_{i=1}^{24}[10 \times \max(y_{1,i\times\Delta t} - 0.5, 0) + y_{2,i\times\delta t}] \tag{5.3}$$

$$\begin{aligned} obj_1(X) &= \frac{1}{24}\sum_{i=1}^{24} y_{1,i\times\Delta t} \\ obj_2(X) &= \frac{1}{24}\sum_{i=1}^{24} y_{2,i\times\Delta t} \end{aligned} \tag{5.4}$$

Decision variables, X (n =4):

- x_1 Total anoxic volume, m^3.
- x_2 Total aerobic volume, m^3.
- x_3 MABR packing density, m^2/m^3.
- x_4 Total SRT, days.

Performance Variables, y_i's (n =2):

- y_1 Effluent ammonia, $mg - N/L$.
- y_2 Effluent nitrate, $mg - N/L$.

Constraints, $C(X)$'s (n = 4):

- $20 \leq$ Total anoxic volume $(x_1) \leq 80$
- $10 \leq$ Total aerobic volume $(x_2) \leq 100$
- $100 \leq$ MABR packing density $(x_3) \leq 500$
- $2 \leq$ Total SRT $(x_4) \leq 7.5$

Evolutionary Algorithm:

- **Single Objective:** Strengthen Elitist Generic Algorithm [111]
- **Multiple Objectives:** Non-dominated Sorting Genetic Algorithm II, NS-GAII. [112]

5.4.5.4 Implementation

All simulations were implemented in Python by coupling the GAs and SUMO numerical core file (sumoproject.dll) via the SUMO-Python API. Necessary modifications in Python-API and Geatpy templets were made to ensure the GA and SUMO core are interacting with each other in real time.

Hardware and software information for this study: SUMO21 (21.0.2); Geatpy2.6 (2020, www.geatpy.com); Processor Info: Intel(R) Core (TM) i7-7700HQ CPU @ 2.80GHz, 2081 Mhz, 4 Core(s).

5.5 Results and Discussion

5.5.1 Step I: Calibration of Influent Fractionation

A standard GA with elitism preservation was applied in this step with a population size of 24 evolving for 40 generations. In analogy to natural evolution, if the fittest solution was not generated during the forty-generation evolution, the found solution would be near-optimal. Therefore, three batches of trials were first conducted to verify the optimality, and a fourth trial with a narrower solution space with shrinking bounds was conducted based on the prior results. The convergence curves for the four trials are displayed in Figure 5.2. In all figures, the objective function values of the best individual and the whole population converged to the same value. This indicates the populations are 'matured', homogeneous and converged to a local optimum. However, that the final converged values of the first three trials are different and far from the theoretically ideal value (zero), reveals that the GA stopped too early at a

suboptimum. This is called 'Premature Convergence', which is commonly observed in GA-related studies. As suggested by Andre *et al.* [113], an adaptive reduction in the definition intervals of the decision variables helps avoid such a situation, as was done in Trial 4.

Table 5.1 shows the decision variables solved by GAs and Table 5.2 includes their corresponding performance variables. In the initial three trials, the estimated fractions deviated from the true value significantly as can be seen in Table 1, however, their predicted PVs are mostly within the error tolerances (Table 5.2). It is commonly acknowledged that measurements of wastewater characteristics are usually subject to uncertainty [83]. Thus, specified ranges are often used to evaluate the quality of estimation. In general, PV values that are within the specified uncertainty ranges are considered sufficiently good. From this perspective, this GA approach provides overall acceptable estimation for fractions. Trial 4 further demonstrated the ability of GAs to find solutions closer to the optimum. Though the final objective function didn't converge to zero, the estimated fractions and predictions were almost identical to their true values. Consequently, there is diminishing return in further narrowing the DVs' bounds or continuing the evolution by increasing the number of generations.

Table 5.1: The GA-solved decision variables in Step I and their corresponding true values and bounds.

Decision Variables							
Name	Real (%)	Bound	Trial 1	Trial 2	Trial 3	Bound	Trial 4
frVSS_TSS	69	[0, 100]	50	75	75	[50 , 90]	69
frSCCOD_TCOD	35	[0, 100]	57	30	25	[20 , 50]	35
frSCOD_TCOD	33	[0, 100]	38	30	25	[20 , 50]	35
frSU_SCCOD	18	[0, 100]	13	20	23	[0 , 20]	18
frXU_TCOD	14	[0, 100]	18	14	16	[0 , 20]	15

Table 5.2: The results of performance variables predicted by GA-solved fractions and their corresponding true values in Step I.

Performance Variables							
Name	Unit	Real	Error Tolerance	Trial 1	Trial 2	Trial 3	Trial 4
Inf_BOD5	mg/L	107.4	10.0	106.3	107.3	101.2	110.0
Inf_TSS	mg/L	133.5	5.0	131.9	131.3	141.2	134.2
Tank3_MLSS	mg/L	3555.0	100.0	4567.1	3263.1	3437.5	3561.6
Tank3_SNHx	mg-N/L	3.6	0.5	3.6	3.6	3.7	3.6
Tank3_QAir	Nm^3/d	1671.6	24.0	1710.1	1637.0	1606.5	1686.2
Tank4_QAir	Nm^3/d	1030.9	24.0	1040.5	1013.2	1022.4	1016.1
Tank5_QAir	Nm^3/d	477.9	24.0	457.8	473.3	488.6	463.1
Eff_BOD5	mg/L	1.8	0.1	1.4	1.9	1.9	1.7
Eff_TCOD	mg/L	19.6	2.0	21.5	19.2	18.7	19.5
Eff_SNHx	mg-N/L	0.1	0.1	0.1	0.2	0.1	0.1
Eff_SNOx	mg-N/L	7.4	1.0	7.4	8.0	7.4	7.6
Time Spent, minutes				44.4	50.0	42.7	29.5
Objective Function Value				18.92	13.52	15.45	7.3

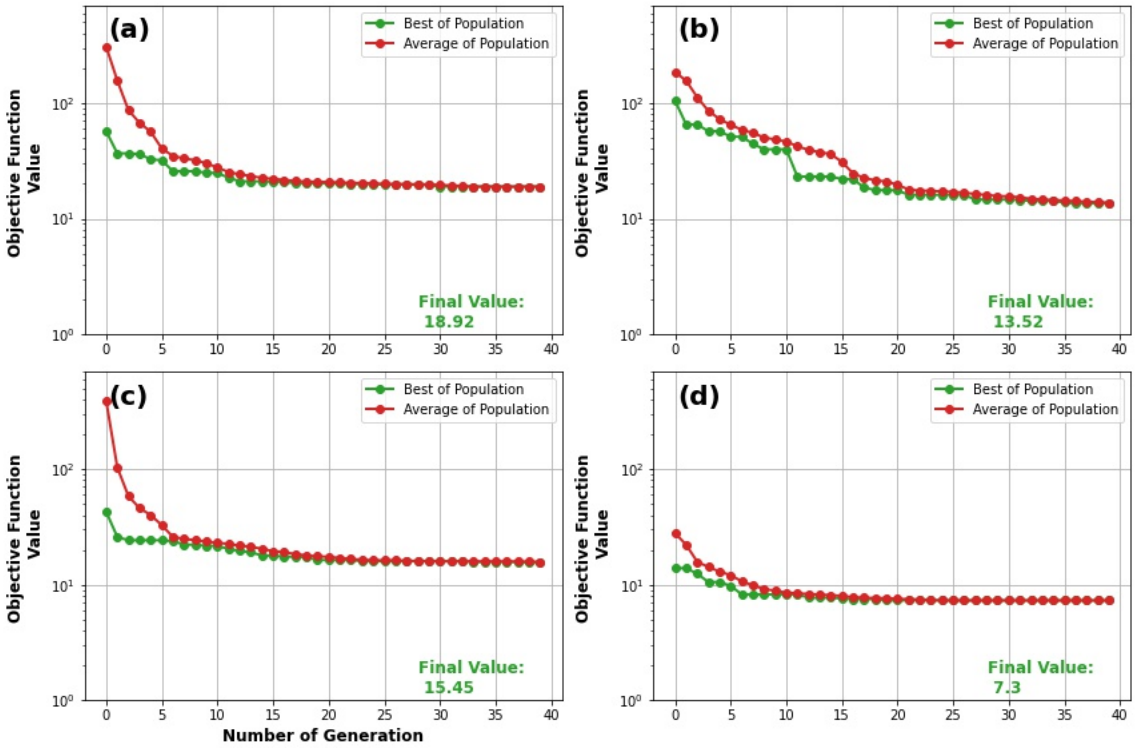

Figure 5.2: The convergence curves of four trials in Step I. (a) Trial 1; (b) Trial 2; (c) Trial 3; (d) Trial 4.

The results in Step 1 clearly demonstrate the possibility of estimating fractions from regularly measured data with the help of GAs, which is of great practical importance. For most plants, due to budget and resource limitations, it is not realistic to regularly monitor the time varying wastewater characteristics that are critical to regular operations throughout the year. The approach described in this step provides an alternative method to extract this information in a less expensive way. The values of the PVs could be replaced by either long-term average values or a set of representative values for steady-state periods. The automatic nature of this approach, along with the fact that most WRRFs have accumulated abundant data, enables vast repetitions and yields practical benefits. On the one hand, more data produces more sets of estimations, which can help assess the precision and uncertainty of the retrieved fractions. On the other hand, retrieved fractions from different periods throughout the years allows quantification of seasonal variations, which is often costly if done by experimental analysis.

5.5.2 Step II: Hybrid MABR Sizing and SRT Optimization

SOGA was first tested by combining the requirements on effluent ammonia and nitrate into a single objective function with weights. Forty batches (population size = 24, maximal generation = 20) were simulated with the same GA used in Step 1 under the dynamic simulation mode. Twenty-nine out of 40 candidate solutions satisfied the effluent peak ammonia concentration of lower than 1 mg N/L, and 5 out of 40 satisfied the effluent peak ammonia concentration of lower than 0.5 mg N/L. All effluent nitrate concentrations were lower than 5 mg N/L. Not all solutions were feasible due to 'premature convergence' as also observed in Step I. The 29 simulations results (Table C.1) are shown in Figure 5.3. It is important to note that there are multiple candidate solutions satisfying the desired objectives, which is frequently seen in process designs. Incorporating more considerations into objective function design,

such as energy consumption, will drive the algorithms to lean towards more qualified candidate solutions [114]. The distributions of the DVs are showed in Figure 5.4, which illustrates the uncertainty of estimation as in similar studies [32, 90]. The mean anoxic volume, total SRT and packing density are concentrated around 50 m^3, 5 days and 375 m^2/m^3, revealing a region where further design exploration is needed.

The MOGA, NSGA-II [112], was tested in parallel with a population size of 100 with 20 generations. Unlike SOGA, the objective functions target the daily average effluent ammonia and nitrate concentrations separately without lumping them together. In MOGA, objective functions cannot conceptually be compared directly with each other. For instance, the cost of removing 1 mg N/L of ammonia and nitrate, as well as their tolerance of violations may vary from region to region, and this information was not incorporated into the objective function design. Therefore, it is inappropriate to conclude 0.5 mg/L ammonia and 3 mg/L nitrate is superior to 0.7 mg/L ammonia and 2.8 mg/L nitrate, though they have the same effluent Total Inorganic Nitrogen (TIN) concentration. Correspondingly, a Pareto front was generated and displayed in Figure 5.5 to reveal trade-offs between objectives. Points on the Pareto front are a set of solutions possessing a special property - none of their objective functions can be improved without degrading others. In other words, effluent nitrate and ammonia can no longer decrease together at the same time. This quantifies the maximal nitrogen removal capacity for the hybrid MABR system. The Pareto front in Figure 5.5 clearly shows a trade-off between the ammonia and nitrate concentrations. The solutions satisfying effluent limits are marked green and, similar to SOGA, there are multiple candidate solutions. Figure 5.6 displays the decision variables on the Pareto front, where several correlations for feasible solutions (green dots) are observed, consistently matching results of SOGA and process knowledge:

1. The negative correlation between aerobic volume and packing density in Figure 5.4(b). Increased biofilm surface area promotes nitrification in the MABR and

therefore, less residual ammonia passing through and less aerobic volume is required for nitrification by the bulk liquid.

2. The positive correlation between anoxic volume and packing density in Figure 5.4(d). As reasoned above, more biofilm surface area means more nitrate generated, which requires more anoxic volume for denitrification and nitrate removal.

3. A weak negative correlation between aerobic and anoxic volume in Figure 5.4(f). The COD is mostly degraded in the suspended growth part, with either nitrate or oxygen as electron acceptors. Nitrate in the anoxic zones and oxygen in the aerobic zones are competing for the electrons from organic substrates.

4. No apparent relationship between total SRT with other variables because the current range of SRT is sufficient for hydrolysis and bio-coagulation for the suspended biomass to stay health and efficient.

Both SOGA and MOGA demonstrated success using GAs to derive a primary design for the hybrid MABR process, a relatively new process whose design criteria are not well defined at the current stage of development.

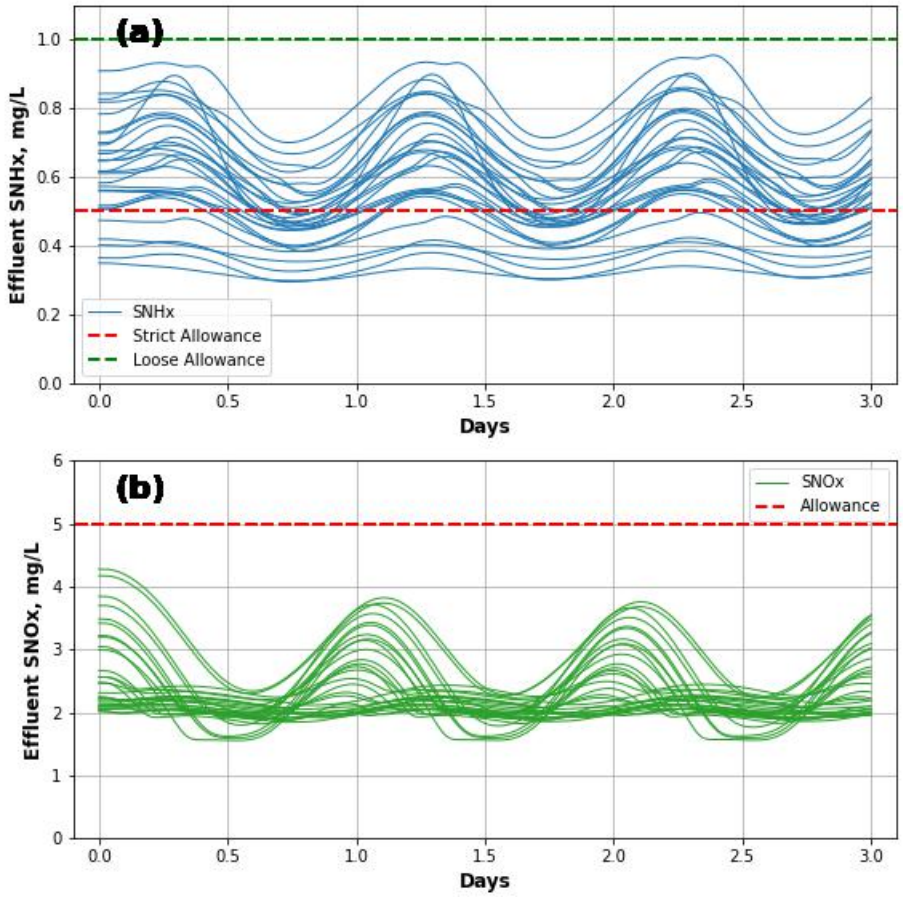

Figure 5.3: The effluent nitrogen profiles with the 29 GA-optimized candidate solutions in Step II. (a) Ammonia; (b) Nitrate and Nitrite.

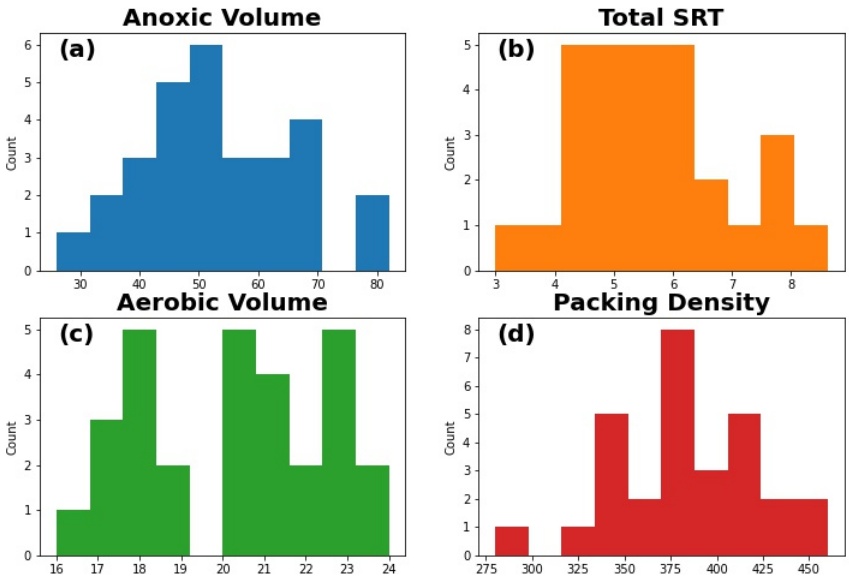

Figure 5.4: The distributions of decision variables in Step II. (a)total anoxic volume, m^3; (b) total SRT, days;(c) total aerobic volume, m^3; (4) Packing Density, m^2/m^3

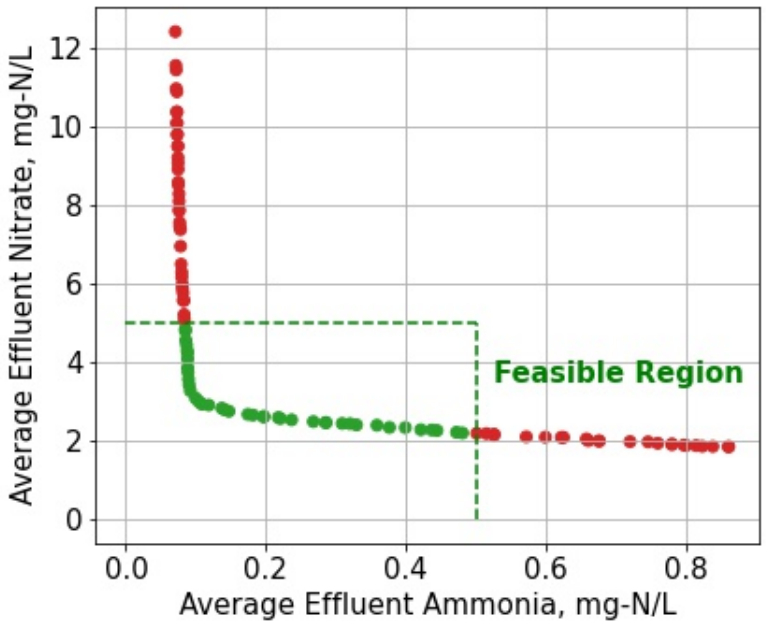

Figure 5.5: The Pareto front solved by multi-objective GAs in Step II.Green dots are predictions from the feasible solutions while red ones are not.

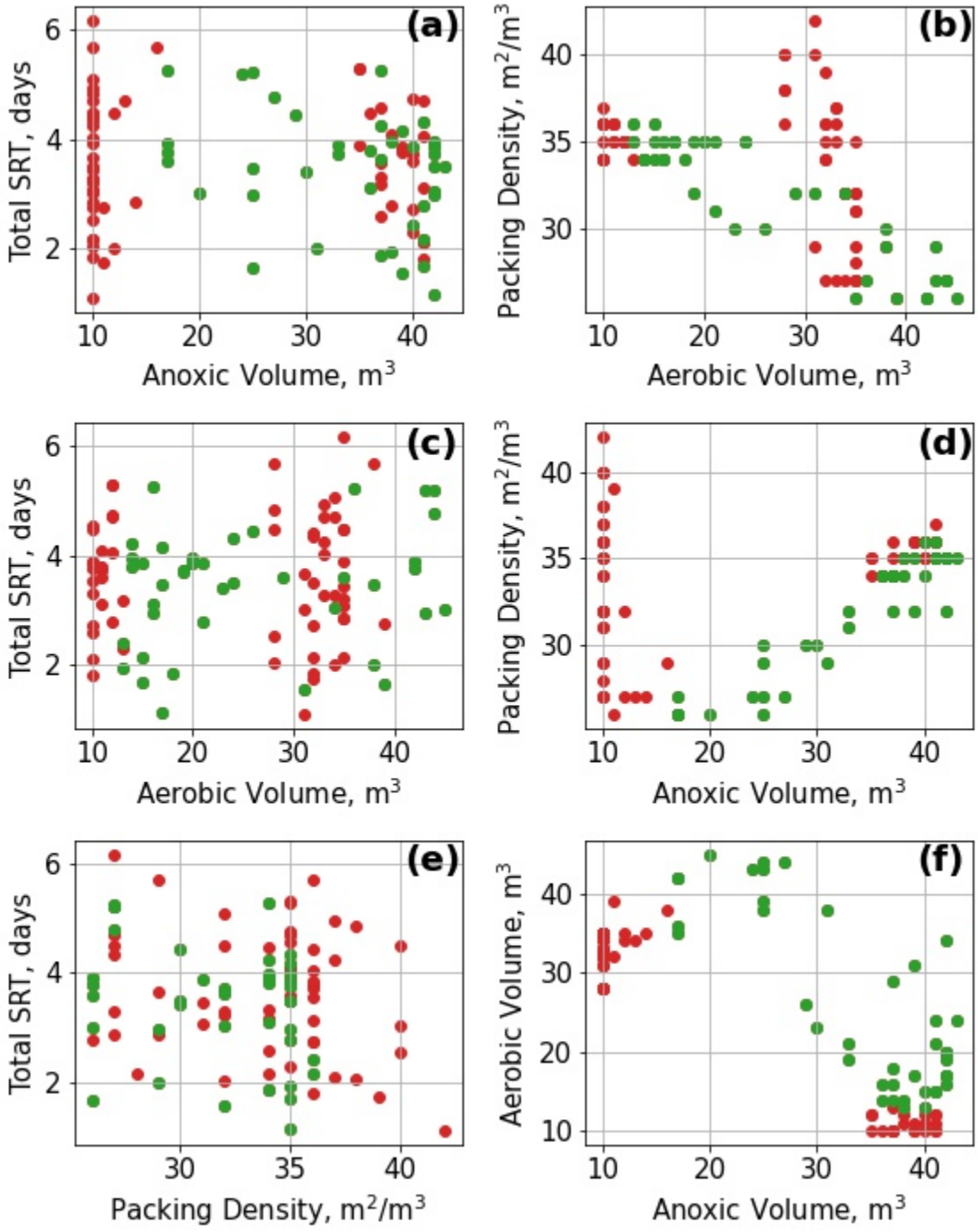

Figure 5.6: The pairwise correlations of Decision Variables solved by multi-objective GAs in Step II. Green dots are predictions from the feasible solutions while red ones are not.

5.5.3 Comparison between the MLE and hybrid MABR processes

Overall, the hybrid MABR achieved better performance - smaller tankage, shorter SRT, less aeration and better nitrate removal, which was expected based on other studies [103]. A rough model-based comparison is provided in Table 5.3, with adjusted aerobic and anoxic volumes for the MLE process to match the effluent ammonia concentrations. The total tank volume was decrease by 17% and the internal recirculation pump was no longer needed. The effluent nitrate concentration improved from 7.4 to 2.5 mg/L while keeping the ammonia concentration at the same level. The mixed liquor concentration is modestly lower and meets the constraint introduced by the secondary clarifier solids loadings. The excess sludge production is lower (dropped from 47 kg/d to 39 kg/d). Aeration decreased by 57% compared to the MLE process, which represents a significant reduction, especially when much of the oxygen needed for the hybrid MABR process would be supplied by the highly efficient MABR unit. These gains come at the expense of installation and maintenance costs for the MABR units. Further monetary analysis is needed to finally determine the better design. Note that these monetary considerations could be further incorporated into the GA by designing proper objective functions.

5.5.4 Significance of Coupling GAs with Commercial Simulators

The mechanistic models embedded in the simulators can be incorporated into a plant digital twin, the virtual replicas of the infrastructure assets, allowing the analysis of data and monitoring of systems to avoid problems before they even occur, prevent downtime, or plan for the future [1, 5]. However, applications of digital twins are as yet at early stages [6]. One of the limiting factors is the traditional manual trial-and-error approach in using these simulators. Coupling GAs with commercial simulators will greatly advance applications of digital twins given its three major features: highly automated, efficient in search and flexible for processes.

Table 5.3: Comparisons between designs in MLE and Hybrid MABR systems when the effluent ammonia concentration is controlled at the same level.

	Unit	**MLE**	**Hybrid MABR**
Sizing			
Anoxic Volume	m^3	55	43
Aerobic Volume	m^3	45	20
MABR tank volume	m^3	0	20
MABR media surface	m^3	N\A	7000
Pumping			
Influent	m^3/d	400	400
RAS	m^3/d	200	200
Internal Recirculation	m^3/d	800	N\A
SRT	day	7.7	5.7
Aeration			
MABR	Nm^3/d	N\A	672
Suspended growth	m^3/d	3017	633
Wasted Sludge	m^3/d	47	39
Performance			
Effluent SNHx	mg-N/L	0.47	0.47
Effluent Nitrate	mg-N/L	7.4	2.3

Highly automated. A well-calibrated model is the first and foremost component for effective simulator applications, based on which operations and control of processes are optimized. Model calibration is often tedious and time-consuming, requiring modelers to interact with simulators frequently to evaluate the results and determine the parameters to change. The combination of GAs and simulators in this study replaced the modelers with a 'computer agent' to interact with the simulator so as to eliminate the manual interactions. Use of GAs in this application allows the modelers to be released from the laborious work of model calibration and to mainly focus on how to establish tuning criteria (e. g. objective functions, DVs, PVs) for the computer agent. Similar procedures are applicable for the optimization of operations and control of the processes.

Efficient in search. Sensitivity analysis followed by brute force search is the most common approach used in simulator applications [115]. There are several limitations for this traditional approach. Frist, the biological processes are often non-linear and sensitivity analysis depends highly on the current values, in other words, they identify local sensitivities [116]. A typical example is the selection of DO setpoints for operations. When the DO concentration is greater than 3 mg/L, an increase in the DO setpoints will have smaller influence on pollutant removal, compared to the case where the DO concentration is less than 1 mg/L. Decisions made from potentially local sensitivity can be biased and lead to ineffective actions. Second, as the degrees-of-freedom in processes increases, especially novel ones, the efficiency of Brute force search decreases dramatically. Assuming there are m decision variables with n grids. The total number of simulations needed is n^m, which grows exponentially either with the increase of decision variables or the grid sizes. Last but not least, optimization is often difficult when conflicting objectives exist, or decision variables are internally related. The users' insights, knowledge and experience determine efficiency of the solution search. For existing treatment technology, such knowledge accumulates from

long-term trial and error. However, for newer processes when knowledge is absent, the efficiency will be low. In contrast, the GAs greatly enhance the efficiency in searching for solutions. The algorithms themselves are designed for multiple objectives optimization for non-linear systems with a large solution space. They can search for the near-optimum solutions more efficiently based on the wisdom from nature selection. Meanwhile, the total number of simulations needed are manageable by determining the population size and number of generations by the users.

Flexible for processes. The flexibility is mainly reflected in the highly modularized feature of the commercial simulators. In earlier proof-of-concept studies, mathematical models were either based on benchmark processes (e. g. BSM1, IAWQ-ASM1) or written by authors. They were case-specific and required long-term effort for developing toolboxes for modeling. Using commercial simulators passes the tool development responsibility to simulator companies and allows users to focus on solving engineering problems instead of building the tools. A wide range of treatment options are available in commercial simulators because these companies are actively updating process units and adding new features. Another flexible aspect of this coupled approach is that mechanistic models could be easily replaced with data-driven/AI models if no adequate mechanistic models are available, as has been demonstrated by Bagheri *et al.* [117] and Mirbagheri *et al.* [118]. In addition, the GAs component is modularized, and the coupling simply requires establishing the data communication between simulators and GA modules. Other optimization toolboxes could also be used to replace the GAs, which provides another layer of flexibility.

5.5.5 Considerations in Coupling GAs with Commercial Simulators

5.5.5.1 Practical Solutions to 'Premature Convergence'

Premature convergence is a problem that was encountered in this study and similar studies [32, 70], and its occurrence is predictable in practical applications. Andre *et*

al. [113] proposed two methods to handle the premature convergence: (1) adaptively reduce the definition interval of each variable, and (2) add a scaling factor to the probability of crossover (an operation in GAs to generate new solutions) to ensure diversity within the population. From a practical perspective, the first method was easier to implement and produced good results in this study. Application of this method is straightforward while the latter needs modifications in the GA module, which requires expertise. This may make reducing the definition interval of each variable the preferred approach. Another two easy alternatives are (1) to increase the size of the population to ensure the diversity of solutions so that initial guesses are more likely to cover the region of feasible solutions; and (2) to repeat the coupled processes for sufficient times until a converged and reasonable results are obtained.

5.5.5.2 SOGAs v. s. MOGAs

Both SOGA and MOGA can be coupled with simulator models and the better choice depends on the project needs. Both could be used when multiple objectives are involved. A general rule suggested by the authors is, if the problem involves trade-offs that could not be clearly quantified, then MOGA is preferred, because it allows engineers to balance trade-offs posteriorly. For SOGA, combining multiple objectives into one function needs sufficient understandings of the process and priority of goals to choose weights carefully so that the population evolves in the wanted direction. For MOGA, it is relatively more straightforward because the Pareto front reveals a set of best solution in the collective sense. However, one challenge of MOGAs is that if the objective dimension is greater than three, it will be less intuitive to visualize and explain the Pareto front.

5.5.5.3 Process Knowledge is Extremely Important

It is appealing that tools from artificial intelligence can be used to accomplish tasks such as model calibration and searching for optimal operations by themselves, but it does **not** mean that the input from process engineers is no longer required. On the contrary, success can never be achieved without guidance from process engineers throughout the whole procedure. First and foremost, formulating the problem into a mathematical form that GAs can understand requires expertise in order to define decision and performance variables and their corresponding boundaries. Design of the objective functions is also a critical step where process engineers should be highly involved. In the literature, several studies constructed the subset of decision variables purely based on mathematical results [119, 91, 92]. The outcome, though feasible, consumed significant computation resources. On the contrary, guidelines [83, 16, 67] that incorporate common process knowledge help reduce unnecessary trials and errors and further enhance the applicability of these AI tools. Last but not the least, evaluations from process engineers are needed to validate the GA-developed solutions with insights from practice. It is important to remember solutions provided by GAs are a starting point but not an endpoint.

5.6 Conclusions

Commercial wastewater treatment software and simulators are becoming the new norm for process design, optimization, operation, and control. The traditional approach, manual interaction with simulators by trial-and-error, is becoming less efficient in applications, given the increasingly elaborate simulator models and embedded process knowledge. Coupling genetic algorithms with commercial simulators significantly advances how simulators could be used in an automatic and efficient manner. In this study, this coupling is found to be effective in estimating influent fractions

and proposing feasible designs for an innovative and complicated process – the hybrid MABR process. The highly automated feature in the combination reduces the need for manual interactions and the search procedure for candidate solutions using GAs are more efficient. This combination has a broad application scope because a variety of treatment options are available in commercial simulators. This study contributes to the ongoing developments of digital twin applications in WRRFs and is a promising demonstration how artificial intelligence tools can help accelerate the learning pace of treatment processes whose process criteria are not well-defined.

CHAPTER VI

Conclusions, Contributions and Future Research Directions

6.1 Conclusions and Contributions

With a data-driven emphasis, this dissertation contributes to the broader knowledge of transforming data into intelligence to advance wastewater treatment. By providing four case studies that address the critical steps in the data pipeline (Figure 1.1), this work provides a holistic roadmap about how and where data-driven tools could be used. Moreover, it illustrates how these tools could better realize their full potential when coupled with domain-specific knowledge. Specifically:

Chapter 2 focuses on the data collection step in the data pipeline, which is the first place where tools from data science could be used to evaluate and improve conventional approaches. It highlights the importance of definite data goals, valid evaluation metrics and adaptive planning in the design of this step. As was illustrated in this chapter, the distinct temporal variation of wastewater characteristics was first revealed. Then, based on model simulations, it tested the current wastewater characterization strategies and identified that they were insufficient in sample size for robust estimation. Finally, it addressed implications on how to conduct data collection- sufficient in size, wisely planned and fit-for-purpose.

Chapter 3 focuses on the data pre-processing step in the data pipeline, which is the second step in the data pipeline. It provides a potential solution to the current 'data graveyard' dilemma. This chapter introduces a case study where useful signals can be extracted from flawed sensor data by coupling data-driven tools with WRRF particularities (e. g. observations, physical constraints, process knowledge et al.). The coupled approaches outcompeted purely data-driven approaches, being able to separate influent sources, classify good or bad signals and further validate these results with an additional source of data. Three methodology guidelines about how to couple data-driven tools with WRRF particularities were proposed and laid the foundations for this dissertation: (1) customizing algorithms with the WRRF particularities; (2) choosing appropriate data features that better reflect the WRRF particularities, and (3) choosing appropriate data-driven tools that users understand.

Chapter 4 focuses on the data mining and modelling step in the data pipeline, which is the most popular step in the data pipeline where data-driven tools are applied. This chapter contributes to a heated research area in the interdisciplinary field of data science and wastewater treatment - grey-box modelling. It introduces a methodology development about how the grey-box model could be realized, which enriches the toolbox of wastewater process modelling. The Extended Kalman Filter was used to construct the grey-box model, whose model structure was reduced from white-box models (first-principle models) and parameters were updated with data streams as black-box models (data-driven models) do. The grey-box model has inherited advantages of its parental models - in addition to the equivalent prediction accuracy, not only system dynamics (as are reflected in key model parameters) could be tracked in real-time, but also intuitive information is generated for facility managers and operators.

Chapter 5 focuses on the comprehension and action steps in the data pipeline, which are located at the end of the data pipeline. It provides a case study answering

what could be fulfilled with artificial intelligence tools and how they can advance wastewater treatment. In this chapter, genetic algorithms, a subfield of Artificial Intelligence, were coupled with a commercial simulator to automatically complete a design task. The task was to upgrade a conventional process to a new one whose design criteria are not fully established. The algorithms estimated needed information from the old process and proposed reasonable designs for the new process that reduce footprint, aeration and pumping with improved effluent quality. It reduces laborious manual efforts and is widely applicable. The methodology developments in this chapter contribute to constructing digital intelligent agents that is able to automate model-calibration, process design and operation optimization, and therefore assist water professionals.

In conclusion, the four case studies in this dissertation provide vivid examples of where and how data data-driven tools can be used to advance wastewater treatment, from the beginning to the end of the data pipeline. The holistic investigation of the data pipeline is significant at the current stage because the digital transformation journey of WRRFs is just launched and is still at an early phase [5, 6]. This dissertation helps identify the opportunities and challenges in the digital transformation journey, deepens understandings about methodology development to deal with them, and demonstrates promising outcomes that can be achieved. Throughout the dissertation, the importance of coupling WRRF particularities with the borrowed tools is addressed and insights to ensure their successful adoption are shared.

6.2 Future Directions

Despite the progress made in this dissertation, further work remains to be done to solidify the digital transformation of WRRFs. Given the nascent nature of the roadmap shown in Figure 1.1, research and practice should be expanded to the following engaging topics:

Extending the developed methodology to other components in WRRFs. In this dissertation, the scope and implementation of data-derived solutions was concentrated on the biological processes of WRRFs, whereas a holistic digital transformation journey requires many other components [6], including but not limited to other physical-chemical processes (e. g. clarifiers, sludge handling, disinfection) and actuating equipment (e. g. blowers, pumps). Fortunately, because the framework of the data pipe is highly general, it is reasonable to assume that such a framework is equally applicable to other WRRF components by embedding proper new particularities. Therefore, a natural next step is to validate such a framework for the other components in WRRFs and test the methods developed in this study. Once achieved, researchers and practitioners can unite all particularities and components together and implement them as a whole throughout individual WRRF projects.

Establish criteria and guidelines for adopting borrowed tools from data science and AI. As pointed out by the leaderships of both academia and practice, no standards or guidelines are currently available for methodology selection and developments, which is one of the major barriers to a more rapid digital transition [8, 11, 12, 120]. From a retro-perspective, the four case studies in this dissertation all share the same concept – coupling domain knowledge with borrowed tools is important and beneficial. Although this could be one contribution to the guideline development, more work remains to explicitly establish concrete standards and guidelines.

First of all, promising tools can be explored as much as possible to build the approach pool. Diversity is needed before the extraction of common criteria, and it requires the collective effort of the digital wastewater community. The approaches proposed in this dissertation are only a subset of methods, solely to enrich the toolbox for data-based solutions. The are many other potential approaches to explore and exploit. For example, Reinforcement Learning (RL) could contribute to the action step in the data pipeline. RL is a generic approach designed to automatically devise

a good decision policy or control strategy by interacting with systems with trials and errors. Based on reward feedback from each action, the control strategy is learned and improved. Recent developments in combining artificial neural networks and RL (so-called Deep RL) showed very good performance for a variety of different systems [121, 122, 123], whereas only a few studies have applied it to the wastewater treatment processes [124, 125].

Secondly, based on the available methodology pool, comparative studies can further extract common patterns for standards and guidelines. In all case studies of this dissertation, comparisons were only made to traditional approaches or currently adopted approaches without comparison to similar coupled approaches. It is important to compare similar approaches with fair metrics (e. g. performance improvement, computational cost) to assess under what scenarios one is better than the other.

Last but not least, different people need to be engaged throughout for comprehensive solution and guideline development, from researchers to practitioners and from wastewater professionals to data scientists. It has been frequently pointed out that the lack of in-depth interactions between different groups has become a major barrier to the faster development of digital solutions [8, 11, 12, 120]. From the experience of this dissertation, a functional committee includes the following roles: (a) Mediators, who should have sufficient exposure to both data science and wastewater field and are responsible for coordinating communications and efforts of others; (b) Wastewater professionals, who are the main executors and are from both researcher and practitioner perspectives. Researchers are often better at formulating research questions, developing solutions and conducting analysis, whereas practitioners are crucial to specify WRRFs particularities and evaluate solutions from practical perspectives (e. g. real-world demands, constraints, and expectations). These two roles may overlap depending on the expertise and experience of individuals. (c) Data scientists, who are important to provide insights into available tools and assist in implementing pro-

posed methods and criteria. This collaboration mode could be further tested on other projects so that a balance could be suggested as cooperation guidelines.

APPENDICES

APPENDIX A

Supplementary Information for Chapter 3

A.1 Regularized Least Squares

A.1.1 Diurnal pattern approximation

The diurnal term was approximated as the 3^{rd}-order Fourier series, as showed in Eq. A.1,

$$y_j = x_1 + \sum_{k=1}^{3} x_{2k} \cos(k\omega j) + \sum_{k=1}^{3} x_{2k+1} \sin(k\omega j) \tag{A.1}$$

Where y_j is the diurnal component at moment j (the j^{th} element in the diurnal vector), x_{2k} and x_{2k+1} are the Fourier coefficients, and ω is angular velocity . Above formula can be transformed into a matrix form as Eq. A.2:

In this study, the length of diurnal vector $n = 720$ and $\omega = \frac{2\pi}{T} = \frac{2\pi}{720}$, because the sensor takes 720 samples per day.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & \cos(1\cdot\omega\cdot 1) & \sin(1\cdot\omega\cdot 1) & \cos(2\cdot\omega\cdot 1) & \sin(2\cdot\omega\cdot 1) & \cos(3\cdot\omega\cdot 1) & \sin(3\cdot\omega\cdot 1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & \cos(1\cdot\omega\cdot j) & \sin(1\cdot\omega\cdot j) & \cos(2\cdot\omega\cdot j) & \sin(2\cdot\omega\cdot j) & \cos(3\cdot\omega\cdot j) & \sin(3\cdot\omega\cdot j) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & \cos(1\cdot\omega\cdot n) & \sin(1\cdot\omega\cdot n) & \cos(2\cdot\omega\cdot n) & \sin(2\cdot\omega\cdot n) & \cos(3\cdot\omega\cdot n) & \sin(3\cdot\omega\cdot n) \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} \tag{A.2}$$

$$\Leftrightarrow \mathbf{d} = \mathbf{\Gamma x}$$

A.1.2 Shifted Huber-Hinge loss function

In machine learning, the huber loss and hinge loss are two different functions added to the regular loss function for regression and classification problems [126, 127, 128]. In order to meet the need of penalizing only the negative difference between actual and fitted values, both loss functions are combined into one into a huber-hinge loss function [51]. Then the huber-hinge function was shifted to ensure $h(L_i = 0) = 0$. The mathematical expression of the huber-hinge function and its derivative is showed in Eq. A.3 and A.4, and their function shapes were showed in Figure A.1. In this study, the δ was set to 10.

$$h(x) = \begin{cases} -x - \frac{\delta}{2} & , \quad x \leq -\delta \\ \frac{1}{2\delta}x^2 & , \quad -\delta < x \leq 0 \\ 0 & , \quad x > 0 \end{cases} \tag{A.3}$$

$$\frac{d}{dx}h(x) = \begin{cases} -1 & , \quad x \leq -\delta \\ \frac{1}{\delta}x & , \quad -\delta < x \leq 0 \\ 0 & , \quad x > 0 \end{cases} \tag{A.4}$$

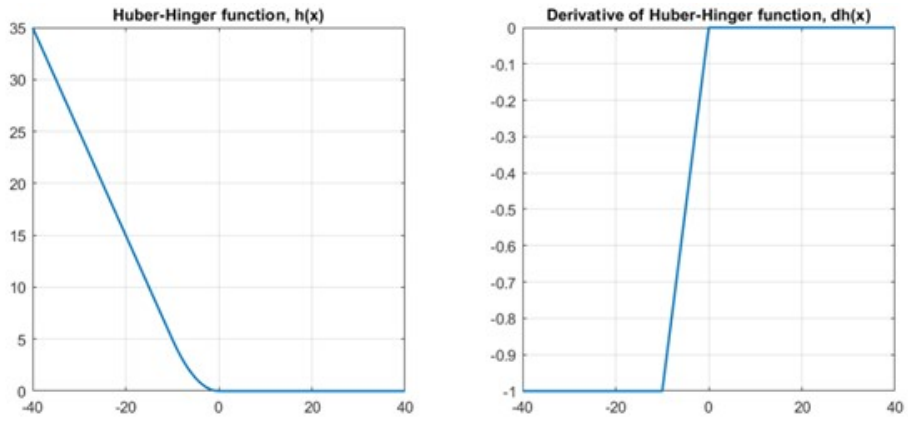

Figure A.1: The huber-hinge function and its derivative function

A.1.3 Formatting the loss function into a solvable form

As mentioned in the pattern separation section in the main article, the question could be framed as Eq.A.5(Eq. 3.3 in main text), where Γx is the diurnal pattern vector and l is the leachate pattern vector, λ is a hyperparameter (=0.5) and $h(x)$ is the huber-hinge loss function.

$$\hat{d}, \hat{s} = \arg\min_{d,l} ||\varepsilon||_2^2 = \arg\min_{x,l} ||y - \Gamma x - l||_2^2 + \lambda \cdot h(l) \tag{A.5}$$

However, this form is hard to solve by gradient descent methods since the analytical form of gradient is not easy to be written out. A little mathematical transformation is needed so that the gradient could be explicitly written out.We can rewrite Eq.3.1 as in Eq.A.6, where $\mathbb{1}$ is an identity matrix, $\tilde{A}$ is the horizontal concatenation of Γ and $\mathbb{1}$, and $\tilde{x}$ is the vertical concatenation of x and l . Therefore, Eq. A.5 can be rewritten as Eq. A.7, where C is a matrix formed by replacing the first seven diagonal elements of identity matrix with zero. The first seven elements in $\tilde{x}$ are the Fourier coefficients and the rest elements of $\tilde{x}$ are the leachate, since $l = C\tilde{x}$. The analytical form of gradient of Eq. A.8 is easy to written and showed in Eq. A.9. Finally, this regularized least squares problem was solved by Nesterov's accelerated gradient descent [129].

$$\begin{aligned} y &= \Gamma x + l + \varepsilon = \Gamma x + \mathbb{1} l + \varepsilon \\ &= \begin{bmatrix} \Gamma & \mathbb{1} \end{bmatrix} \begin{bmatrix} x \\ l \end{bmatrix} + \varepsilon = \tilde{\mathbf{A}}\tilde{\mathbf{x}} + \varepsilon \end{aligned} \tag{A.6}$$

$$\hat{x} = \arg\min_{\tilde{x}} ||y - \tilde{\mathbf{A}}\tilde{\mathbf{x}}||_2^2 + \lambda \cdot h(C\tilde{x}) \tag{A.7}$$

$$Loss = ||y - \tilde{\mathbf{A}}\tilde{\mathbf{x}}||_2^2 + \lambda \cdot h(C\tilde{x}) \tag{A.8}$$

$$\nabla Loss = \tilde{\mathbf{A}}^T \cdot (\tilde{\mathbf{A}}\tilde{\mathbf{x}} - y) + \lambda \cdot C^T \cdot \frac{d}{dx} h(C\tilde{x}) \tag{A.9}$$

A.2 Quality classification

A.2.1 Stuck Index

Stuck Index is introduced to improve the classification accuracy. It measures the fraction of how many measurements are not changed from its previous measurement. The mathematical formula is given below, where n is the length of signal vector y and i is the index of measurements.

$$SI = \frac{\sum_{t=1}^{n}(y(i) - y(i+1))}{n-1} \tag{A.10}$$

A.3 Data validation

A.3.1 Reason to choose Artificial Neural Network

The problem in this section can be framed as mapping a vector of length 720 (daily profile) into a scalar (the composite sample measured by wet-chemistry method in plant laboratory). The first instinct was to average all the sensor readings in the way that the composite samples are collected, which yielded the flow-proportion averaged model. As showed in Figure A.2 of the manuscript, it tends to overestimate the composite values.

Results of sensor calibration evaluation experiment are showed in Figure A.2. The x- axis is the lab measurements, and the y-axis is the sensor measurements. The orange points are measurements when leachate spikes happen, and the blue ones are when spikes are absent. From this figure, we concluded that the sensor was not fully linear. Besides, sensors are unable to 'intelligently' switch between two calibration curves. But we hypothesized a solution that by placing lower weights over those

untrustworthy measurements (for instance, measurements with leachate spikes), we can still get a reasonable estimate of the composite data, since the composite data is determined by the combination of 720 instantaneous sensor measurements.

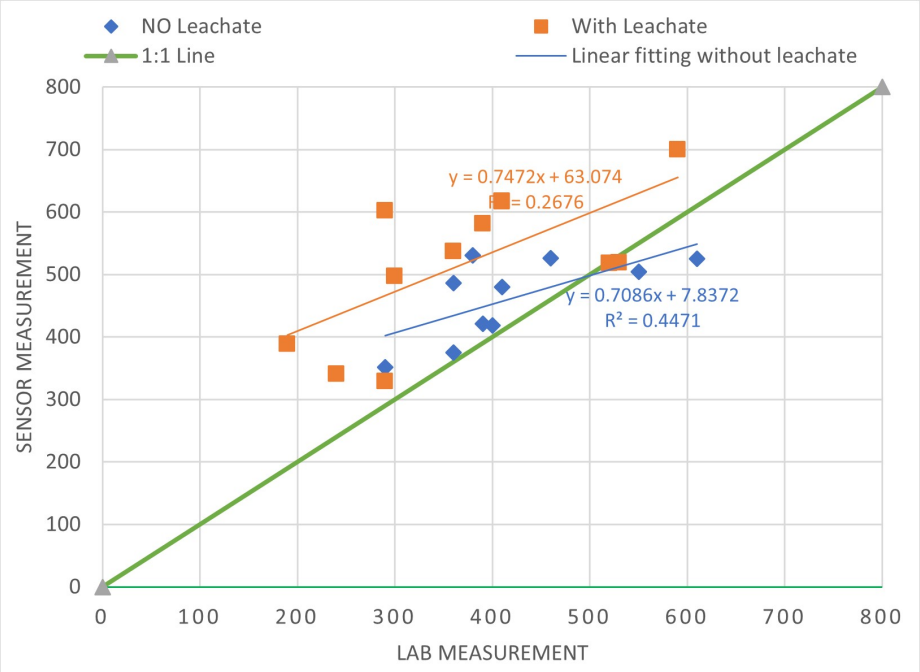

Figure A.2: Results of sensor calibration evaluation experiment

However, it is quite hard for human beings to determine the weights, since the occurrence of spikes and amplitude of them are so random. Therefore, we decided to use ANN as a tool to help adjust the weights 'intelligently' due to its back-propagation property. Another reason is we tried it first and it worked quite well, therefore, we didn't spend time exploring alternatives, since this is just a further step to boost trust in the separated diurnal pattern.

A.3.2 Data preparation

The neural network was trained and tested based on the clean signals. The whole set of clean separated diurnal curves was randomly divided into training set (80%) and testing test (20 %), where the former was used to train the neural network model and the latter to test model performance.

A.3.3 Neural Network Architecture

The Neural Network model was designed with MATLAB (2019b) Deep Network Designer. It is a simple neural network. It consists of one input layer with 720 nodes, one hidden layer with 24 nodes and one output layer with 1 node. The activation function for every node was linear. The layout of the architecture was showed in Figure A.3. It is important to point out that the 'ImageInput' layer does normalization for the input data. Together with 'fc_1' layer, they formed the input layer.

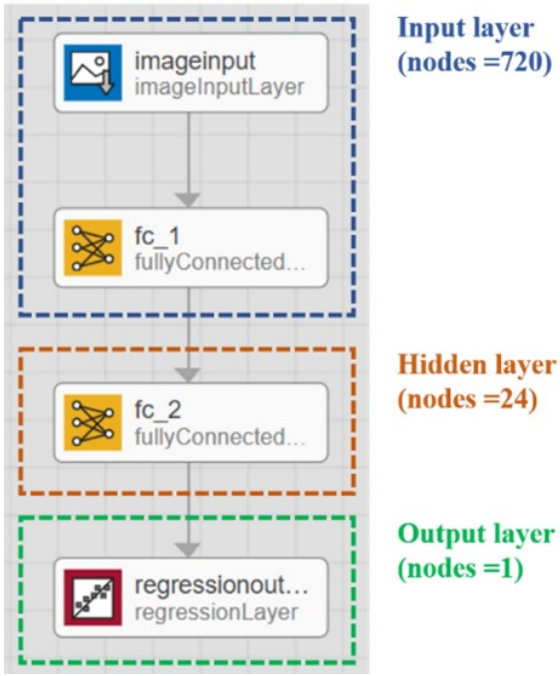

Figure A.3: The architecture of the neural network model

A.3.4 Neural Network Training process performance

As we can see in both plots, the loss and Root Mean Square Error (RMSE) both converged, indicating the neural network was learning the mapping from signals to composite measurement. Meanwhile, the testing (black) set displayed similar trends and finally converged to similar error and loss, indicating the neural network model was not overfitting, which is commonly found in ANN models.

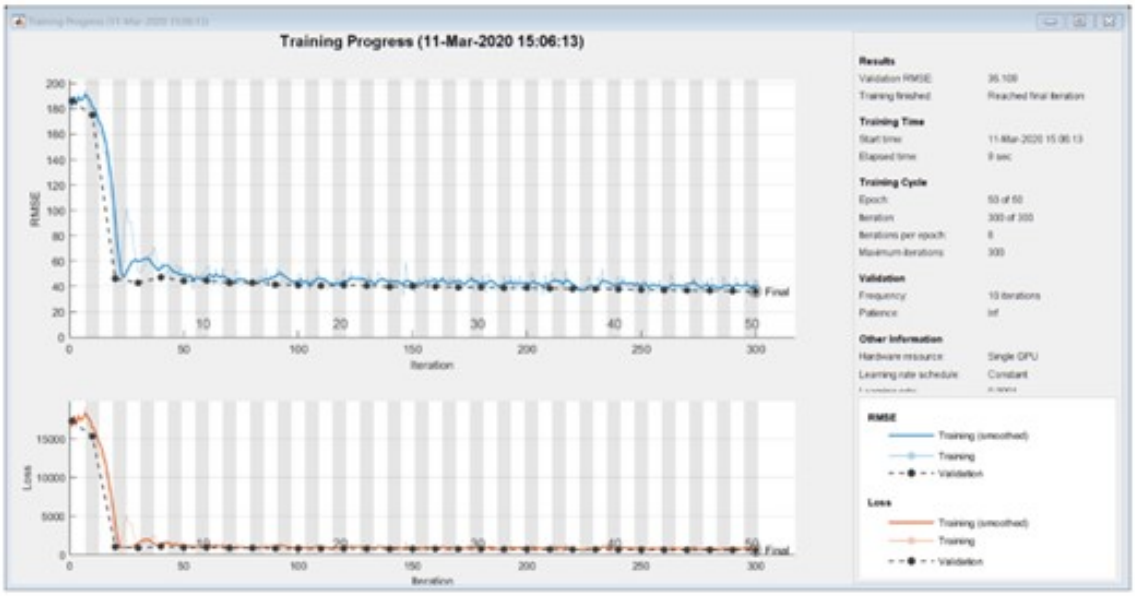

Figure A.4: Neural Network Training process performance

A.4 A mis-classified example with stuck faults

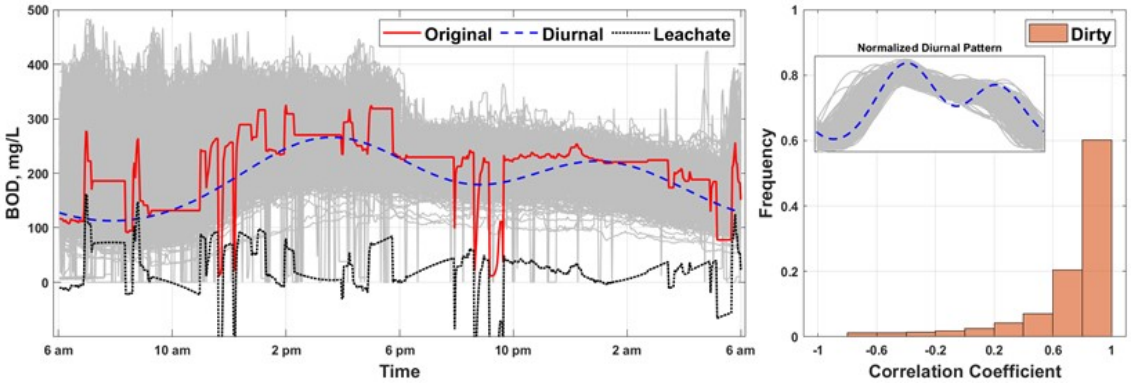

Figure A.5: A mis-classified example where severe stuck faults were not identified

APPENDIX B

Supplementary Information for Chapter 4

B.1 Figures

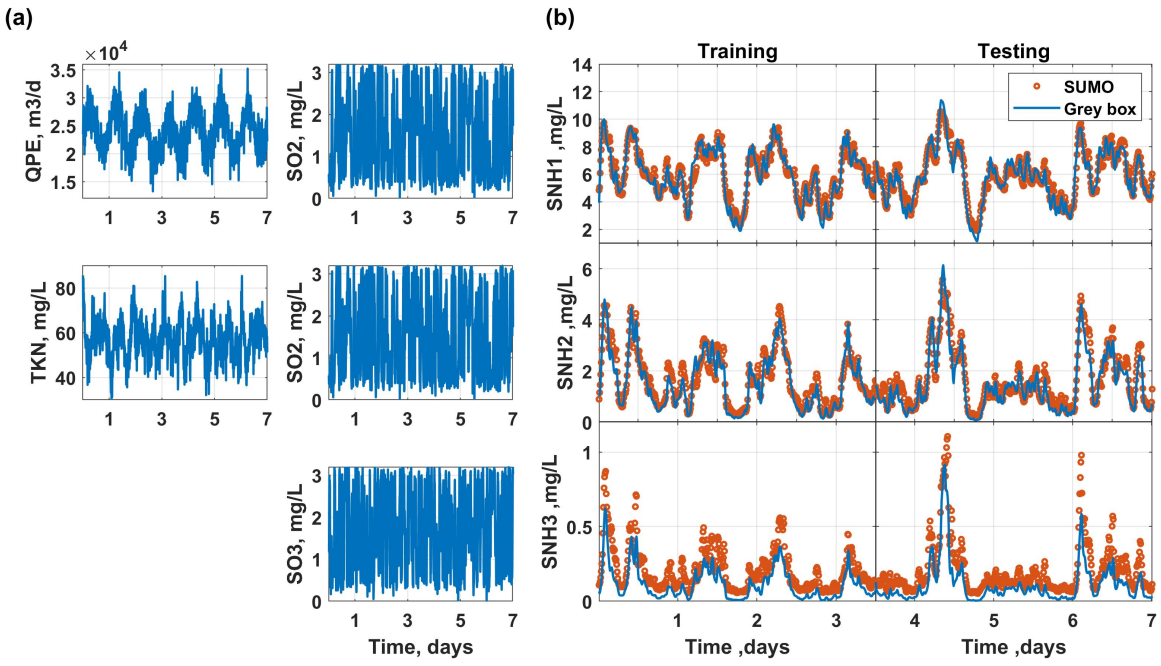

Figure B.1: Grey box model performance - Scenario 3 – with 10% measurement noise. (a) model inputs. QMLE and QRAS are flat lines therefore are not shown. (b) Predictions of three ammonia concentrations. Training set (early 3.5 days) is for parameter estimation, while testing set (late 3.5 days) is for validation of the estimated parameters.

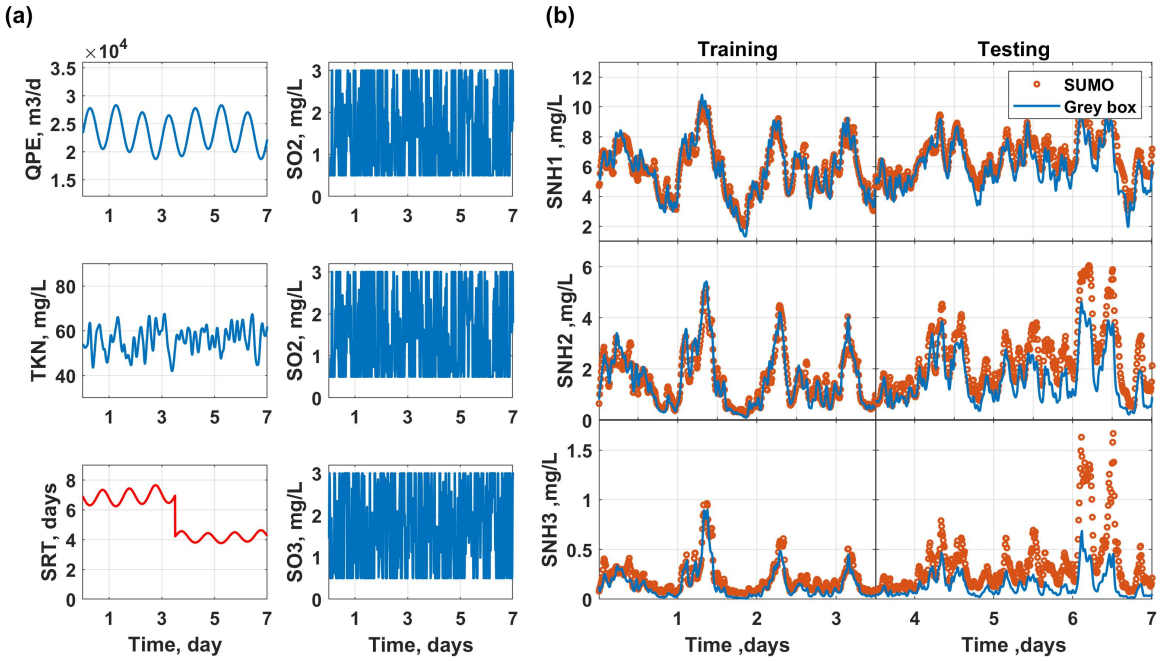

Figure B.2: Grey box model performance - Scenario 5 – SRT drop. (a) model inputs. QMLE and QRAS are flat lines therefore are not shown. (b) Predictions of three ammonia concentrations. Training set (early 3.5 days) is for parameter estimation, while testing set (late 3.5 days) is for validation of the estimated parameters.

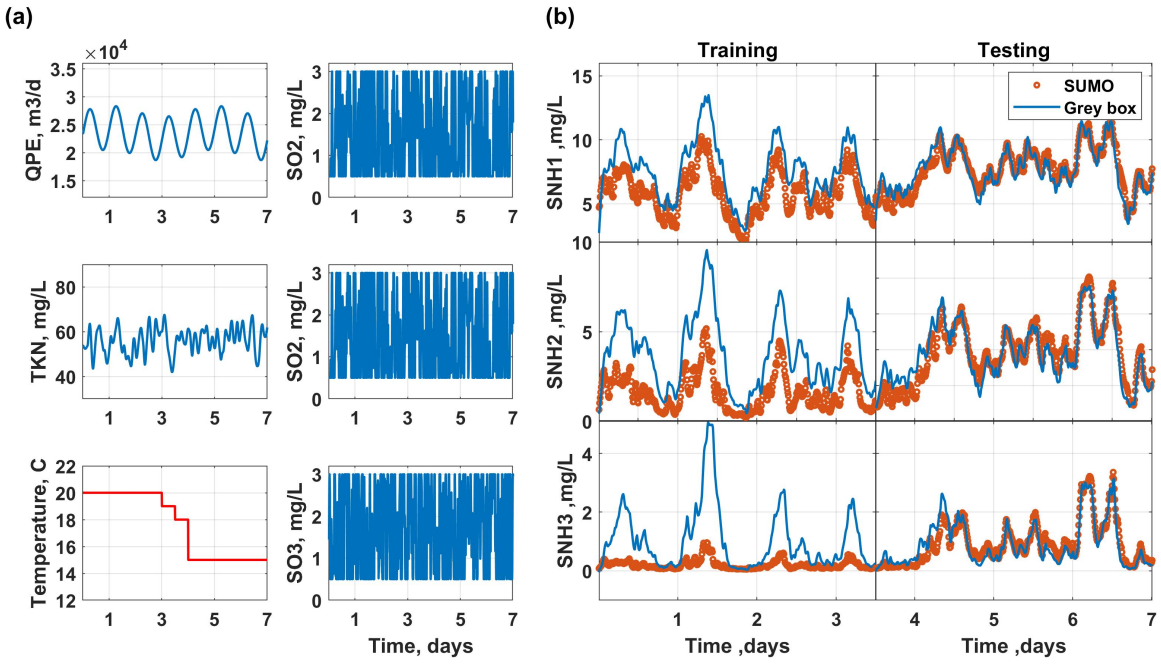

Figure B.3: Grey box model performance - Scenario 4* – temperature drop. (a) model inputs. QMLE and QRAS are flat lines therefore are not shown. (b) Predictions of three ammonia concentrations. Training set (late 3.5 days) is for parameter estimation, while testing set (early 3.5 days) is for validation of the estimated parameters.

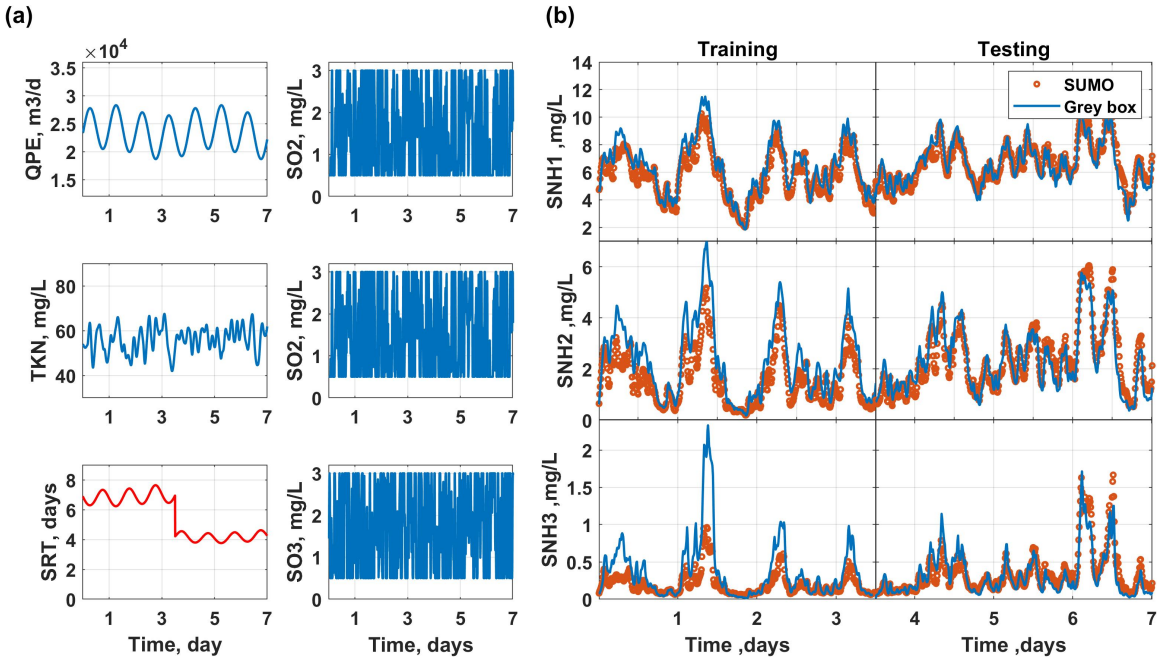

Figure B.4: Grey box model performance - Scenario 5* – SRT drop. (a) model inputs. QMLE and QRAS are flat lines therefore are not shown. (b) Predictions of three ammonia concentrations. Training set (late 3.5 days) is for parameter estimation, while testing set (early 3.5 days) is for validation of the estimated parameters.

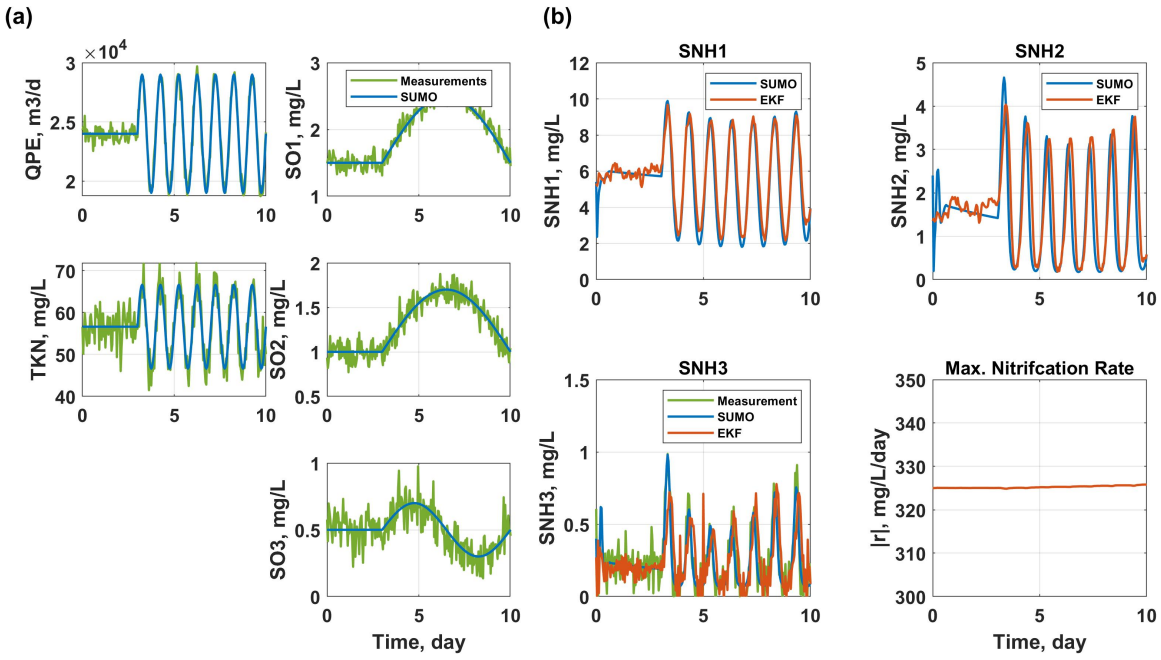

Figure B.5: The performance of EKF when there is no SRT or temperature changes. A trial test. (a) model inputs. QMLE and QRAS are flat lines therefore are not shown. (b) Predictions of three ammonia concentrations.

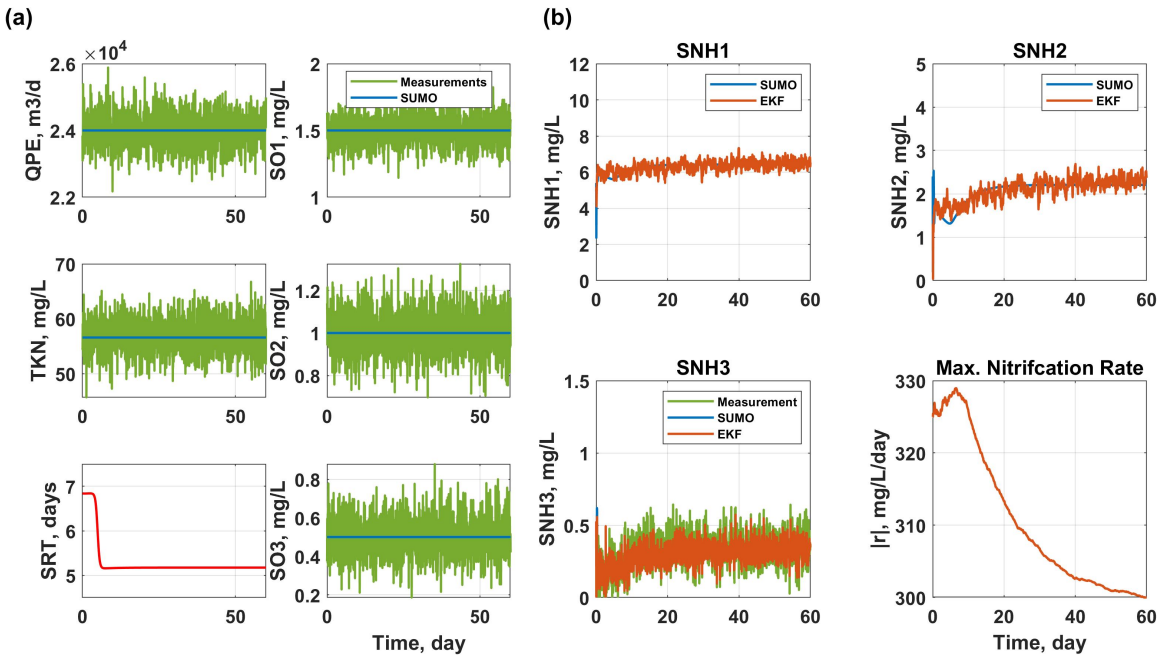

Figure B.6: The performance of EKF when SRT drops. (a) model inputs. (b) Estimation of three ammonia concentrations and maximum nitrification rate, r.

B.2 Code

Codes for the Implementation of this study

Please refer to github.com/ChengYangUmich/SupplementaryMaterialForEKFpaper

Excel file to reproduce the SUMO virtual plant

Please refer to github.com/ChengYangUmich/SupplementaryMaterialForEKFpaper

B.3 Table

Table B.1: Input statistics summary of different scenarios

Scenario	Index	1	2	3	4	5	6
QPE	Min	23473	18667	13252			
	Max	23473	28348	35178			
	Mean	23473	23756	23862			
	Std	0	2810	3657			
QMLE	Min	23400	23400	15800			
	Max	23400	23400	31441			
	Mean	23400	23400	23427			
	Std	0	0	2280			
QRAS	Min	36000	36000	25820			*
	Max	36000	36000	47487			
	Mean	36000	36000	36095			
	Std	0	0	3601			
TKN	Min	56.50	37.51	29.90			
	Max	56.50	77.63	85.57			
	Mean	56.50	56.59	56.80	*	*	
	Std	0.00	6.74	8.95			
SO1	Min	0.50	0.50	0.00			1.19
	Max	3.00	3.00	3.37			1.83
	Mean	1.63	1.63	1.63			1.50
	Std	0.95	0.95	0.98			0.10
SO2	Min	0.50	0.50	0.00			0.64
	Max	3.00	3.00	3.40			1.36
	Mean	1.53	1.53	1.53			1.00
	Std	0.97	0.97	0.98			0.10
SO3	Min	0.50	0.50	0.00			0.22
	Max	3.00	3.00	3.55			0.87
	Mean	1.65	1.65	1.64			0.50
	Std	0.96	0.96	0.98			0.10
Temp	Min	20.0	20.0			15.0	
	Max	20.0	20.0	*	*	20.0	*
	Mean	20.0	20.0			17.6	
	Std	0.0	0.0			2.3	
SRT	Min	6.9	6.2		6.2		
	Max	6.9	7.6	*	7.6	*	*
	Mean	6.9	6.9		6.9		
	Std	0	0.4		0.4		

Note: * means conditions are the same as Scenario 2

APPENDIX C

Supplementary Information for Chapter 5

C.1 SUMO files

For asking SUMO model files (.sumo) and corresponding codes in chapter please contact yangche@umich.edu.

The standard SUMO output excels, which includes model setup, layout, changes in input parameters, kinetics etc., are listed below and available upon email request.

- MLE: SUMO_excel_MLE.xlsx
- Hybrid MABR: SUMO_excel_MABR.xlsx

C.2 Tables

Table C.1: The 29 GA-solved process design for the hybrid MABR system in Step II and their corresponding effluent quality.

ANX V m^3	**AER V m^3**	**Packing Density m^2/m^3**	**Total SRT day**	**AER SRT day**	**ANX SRT day**	**Eff SNHx mg-N /L**	**Eff SNOx mg-N /L**	**WAS m^3/d**
26	**21**	**280**	**3.0**	3.0	2.0	0.64	2.02	20
36	**18**	**320**	**4.4**	3.5	3.2	0.70	2.00	15
		340	**5.0**	4.0	3.7	0.57	2.10	13
38	**16**	**340**	**5.0**	3.5	3.8	0.75	1.97	13
41	**17**	**350**	**5.3**	3.8	4.1	0.67	2.04	13
42	**20**	**360**	**4.6**	3.6	3.3	0.60	2.12	16
43	**17**	**370**	**5.4**	3.8	4.2	0.66	1.98	13
	20	**350**	**5.7**	4.4	4.2	0.47	2.24	13
	21	**340**	**4.4**	3.6	3.2	0.63	2.00	17
45	**17**	**360**	**5.6**	3.8	4.3	0.68	2.04	13
48	**18**	**370**	**5.9**	4.0	4.5	0.61	2.09	13
49	**19**	**370**	**6.0**	4.2	4.6	0.55	2.16	13
50	**20**	**380**	**6.2**	4.4	4.7	0.48	2.19	13
51	**20**	**380**	**6.2**	4.4	4.7	0.48	2.19	13
52	**18**	**390**	**4.4**	2.9	3.5	0.83	2.19	18
	22	**370**	**5.0**	3.7	3.7	0.57	2.19	17
53	**22**	**370**	**3.9**	2.9	2.9	0.74	2.39	22
55	**23**	**370**	**4.9**	3.7	3.7	0.56	2.22	18
59	**18**	**410**	**6.6**	4.0	5.3	0.61	2.11	13
	23	**390**	**4.6**	3.3	3.5	0.61	2.70	20
60	**24**	**390**	**7.2**	5.3	5.4	0.34	2.38	13
61	**21**	**410**	**6.1**	4.0	4.7	0.51	2.46	15
62	**19**	**410**	**5.0**	3.0	4.0	0.77	2.44	18
66	**23**	**410**	**5.8**	3.9	4.5	0.51	2.67	17
67	**23**	**410**	**7.6**	5.1	5.9	0.37	2.37	13
68	**20**	**430**	**6.9**	4.1	5.5	0.52	2.45	14
69	**21**	**430**	**7.6**	4.6	6.0	0.43	2.43	13
81	**24**	**460**	**8.6**	5.3	6.8	0.32	3.00	13
82	**23**	**460**	**8.0**	4.7	6.4	0.38	3.05	14

APPENDIX D

MICDE Requirements - A Short Literature Review on Evolutionary Algorithms

D.1 Introduction and Overview

Many environmental decision-making problems focus on finding a preferred option among different alternatives, where models play a crucial role in supporting these tasks. In the field of wastewater treatment, mechanistic models are extensively used to evaluate the effectiveness of design, operation, and control alternatives for the biological processes in Water Research Recovery Facilities (WRRFs). However, identifying suitable options is generally difficult and time-consuming. On the one hand, the mappings between model inputs and outputs are less intuitive because of the complexity and scale of model equations, therefore, trial-and-error is often the main approach adopted by the environmental modelers. On the other hand, due to the large space of available options, only a subset of representative options can be chosen and evaluated manually, until an acceptable one (may not be the optimal one) is found, the exploration of the options space may be less efficient. Consequently, significant benefits can be achieved by linking models with a modern family of opti-

mization algorithms, namely, the Evolutionary Algorithm (EA), and automating the most suitable alternatives finding.

The EAs use evolutionary principles found in nature, "evolving" to find better solutions to complex environmental problems. There are several attractive features in applying EAs to finding the suitable options:

1. The optimization process is very intuitive in analogy to Darwin's Theory of Evolution. The EAs initiate a population of different solutions and then learn from the outcomes of these trials. The fittest ones survive and pass their information to the children population with minor adjustments (mutations and crossovers) until the final population become homogeneous. This process is highly similar to the process when the modelers manually select and fine-tune the best-so-far solutions.

2. EAs can find (near-) globally optimal solutions. EAs are population-based, which indicates they have an entire search party exploring the entire solution space for the globally optimal solutions, rather than a single agent which tends to be trapped into local optima. In addition, the members of the search party often exchange and share information, enabling promising regions of the search space to be identified more effectively and subsequently enabling the search to be concentrated in these regions.

3. EAs are easily linked with (existing) simulation models. The coupling is straightforward:

   (a) the EAs determine the decision variable values and pass them to the simulation model;

   (b) the simulation model evaluates the corresponding objective functions and constraints and pass them back to the EAs for decision variable update.

4. EAs can straightforwardly handle constraints. Two common available methods are used:

   (a) using penalty functions to transform the problem from a constrained optimization problem into an unconstrained one;

   (b) using feasibility rules to select out the violated solution in the selection procedure.

5. EAs can deal with multiple objective simulations simultaneously. Weighing different objectives and combining them into a single new one is one widely used approach. However, as commonly seen in biological processes, the same single objective value might correspond with multiple solutions, limiting gained insights in the problem and solutions, as trade-offs between objectives are not able to be explored. EAs provide a benefit to approximate Pareto fronts in a single algorithm run, overcoming the limitation brought by the weighting method.

D.2 Evolutionary Algorithms

EAs are a population-based search techniques to find the optimal or near-optimal solutions to complex optimization problems. Individuals in the population represent search points, which are randomly initialized within the solution domain. The progress in the search is achieved by evaluating the fitness of all individuals in the population, selecting the individuals with a better fitness value, and combining them to create new individuals with likely-improved fitness, such that the offspring generations evolve towards the optimal solution. After sufficient generations of evolution, the population converges to the best individual that represents the optimum (or near-optimum) solution.

In analogy to biological evolution, basic concepts include selection, recombination/crossover, mutation and reproduction ([130],Figure D.1):

1. **Selection**. The better candidate solutions are selected based on their fitness, which is quantified by their objective function values.

2. **Recombination/Crossover**. The recombination (also known as crossover) represents the process by which new candidate solutions (children) are generated from two or more candidate solutions (parents) following certain heuristics.

3. **Mutation**. Mutation is applied to only one candidate solution, and it results in one new child.

4. **Reproduction** After executing recombination and mutation, a new generation of candidate solutions are generated and subject to selection.

The basic structure of EA is generally the same, with the the following key steps[114]:

1. Create the first population using random initialization

2. Evaluate the fitness of each individual in the population

3. Repeat the evolution steps until stopping criterion satisfied:

   (a) Select the individuals for reproduction

   (b) Perform genetic operations to generate the offspring

   (c) Evaluate the individual fitness of the offspring

   (d) Replace the least fit individuals with new best fit individuals

4. Report the best solution of the fittest individual.

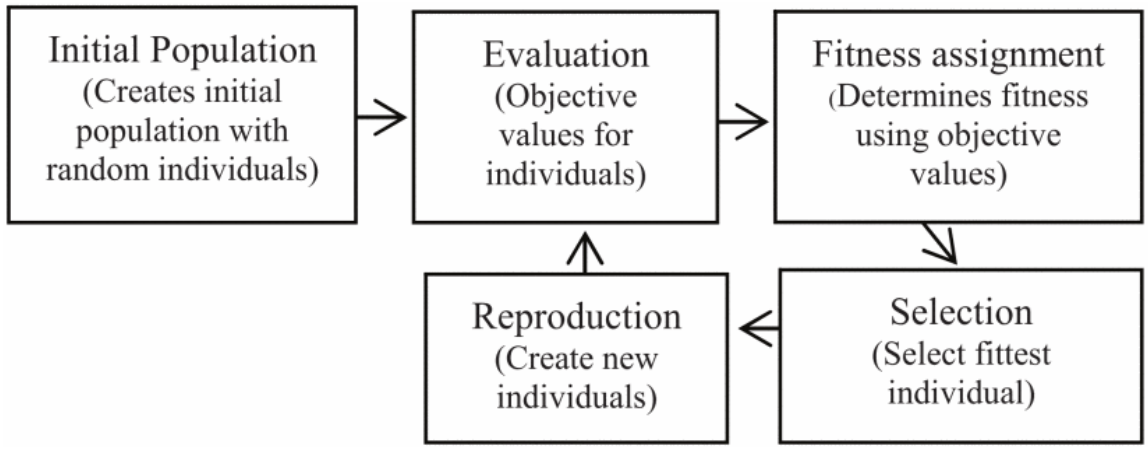

Figure D.1: The general process of evolutionary algorithms. Figure from Vikhar [130]

Janga Reddy[114] summarized the taxonomy of the optimization methods and displayed where the EAs are located as showed in Figure D.2.Three major types of EA are applied in this chapter, namely Generic Algorithm (GA), Evolutionary Strategy (ES) and Differential Evolution (DE).

D.2.1 Genetic Algorithm

The most popular type of EA is Genetic Algorithm (GA), as it exhibits the closest mapping of the natural evolution process onto to computer. It was first proposed by [85], and since then it is widely used for optimization problems. GA uses recombination and mutation operators to seek the solutions of a problem, which are coded in the form of strings of numbers (traditionally binary) i.e. a Bit-String representing the genes.

D.2.2 Evolution Strategy

Evolution Strategy (ES) was first proposed by Rechenberg[131] as an optimization method for complex, multimodal and non-differentiable functions. Key features include that the search space is real-valued and mutation is performed by adding a normally distributed random vector. The mutation strength (i.e. the standard

deviation of the normal distribution) is often governed by self-adaptation. The (environmental) selection in evolution strategies is deterministic and only based on the fitness rankings, not on the actual fitness values.

D.2.3 Differential Evolution

Differential Evolution (DE) was first proposed by Storn[132] for multidimensional real-valued functions but does not use the gradient of the problem being optimized, which means DE does not require the optimization problem to be differentiable, as is required by classic optimization methods such as gradient descent and quasi-newton methods. DE optimizes a problem by maintaining a population of candidate solutions and creating new candidate solutions by combining existing ones according to its simple formulae, and then keeping whichever candidate solution has the best score or fitness on the optimization problem.

D.2.4 The differences of these three major types

- The GA usually uses binary representations for solutions while the ES and DE use real numbers, in other words, GA is usually discrete while the rest are continuous.

- The mutation operator of GA is bit-flip ($0 \leftrightarrow 1$) and the crossover operator is partial genes exchange between parent genes. The ES uses linear combination of two parent genes for recombination and uses normal distribution to perform mutation. The DE calculates a mutant from three randomly selected individuals (V_1, V_2, V_3) and create a temporal variant by $V^{'} = V_1 + F \times (V_2 - V_3)$ and performs partial genes exchange between V_1 and $V^{'}$ as recombination and mutation.

- The GA and the ES are viewed as random-based given the creation of variants is random, whereas the DE is direction-based given a direction vector $(V_2 - V_3)$

is calculated.

D.3 Applications of Evolutionary Algorithms in WRRFs

Although EA are widely used in studies regarding water distribution and wastewater collection networks, their applications in WRRFs are emerging and limited. This chapter summarized related studies in the recent five years, mainly focusing on the above-mentioned three types.

D.3.1 Modeling

Lariche *et al.* [133] developed a machine learning model to predict the removal of methylene blue, a synthetic pollutant commonly seen in colored industrial wastewater, by nanoparticle adsorption. GA was used to tune the model parameters and yielded a good fitting with $R^2 = 0.999$.

Bonakdari *et al.* [134] developed a polynomial regression model to predict the energy consumption of electrocoagulation to treat industrial wastewater and applied GA to tune the model parameters to fit multiple outputs.

Lin *et al.* [60] developed a neural network model to predicted the effluent Total Suspended Solids (TSS), BOD_5 and COD of the Benchmark Simulation Model No. 1 (BSM1). Three variants of DE were used to tune the model parameters, and all of them provided promising fittings.

Rivera-Salvador *et al.* [135] incorporated a new biological reaction in the conventional Anaerobic Digestion (AD) models and applied DE algorithms to find the proper kinetic parameters of the reactions.

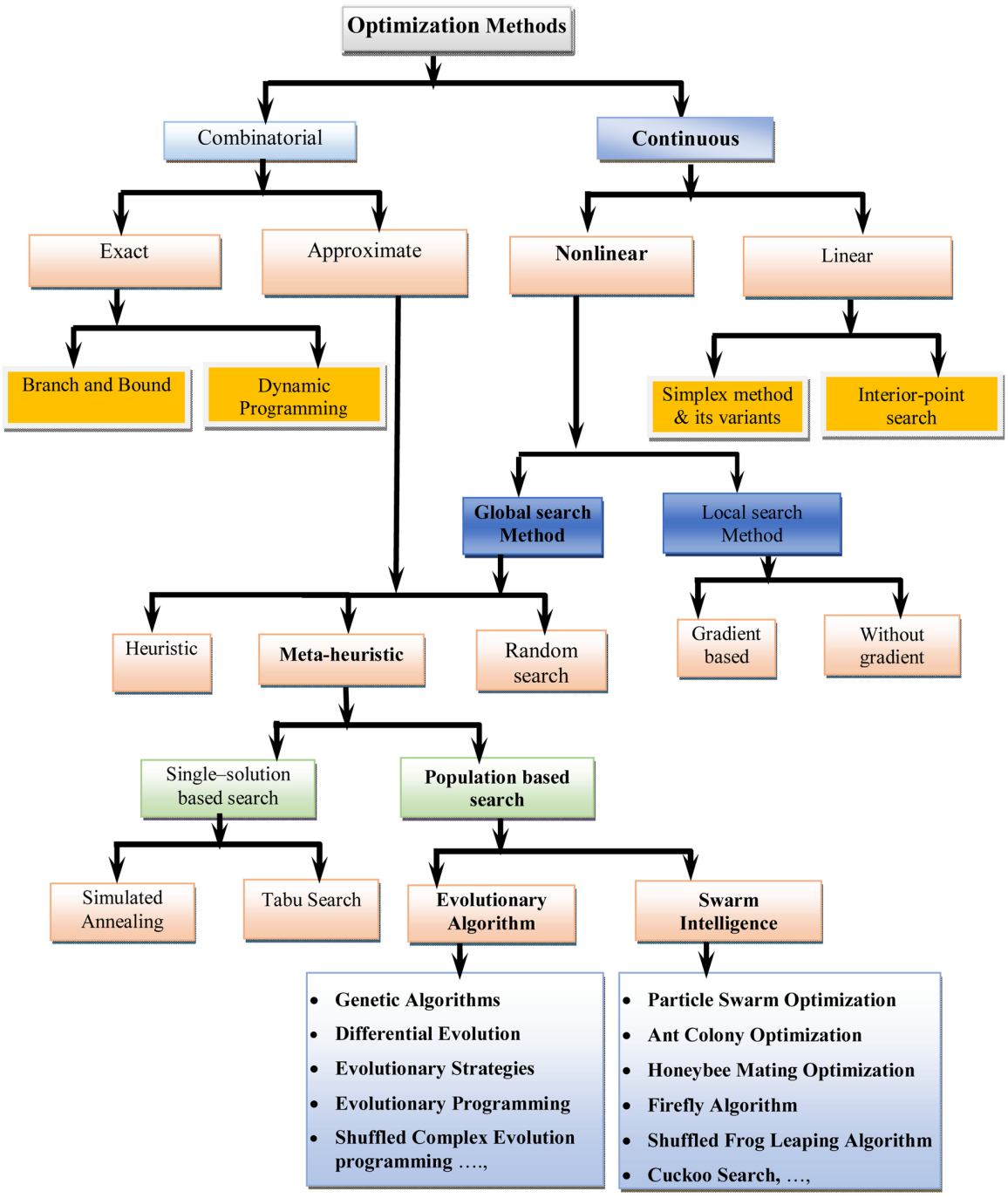

Figure D.2: Taxonomy of optimization methods showing where the Evolution Algorithms locate. Figure from Janga Reddy[114].

D.3.2 Operations and Control

Tejaswini *et al.* [101] applied a multi-objective GA, namely, Non-dominated Sorting Genetic Algorithm II (NSGA-II) [112], to tune the PI controller parameters -gains and time constants- in the BSM1. The optimized PI controller parameters provided improvements in both effluent quality and operational cost. It was observed that when optimizing one objective, others may be compromised, which is common in WRRFs. Simultaneous multi-objective optimization by NSGA-II is more beneficial to decision-makers to evaluate the trade-offs between alternative solutions.

Qiao *et al.* [100] constructed an online neural network model to approximate the mapping from DO and nitrate set-points to plant performance in BSM1. Based on the approximated model, NSGA-II was used to dynamically search for the optimal set-points which were input into PI controllers in BSM1, whose result demonstrated reduced energy consumption. Zhou *et al.* [136] further developed an adaptive multi-objective evolutionary algorithm to replace the NSGA-II in Qiao *et al.* [100] to find the optimal set-points and displayed faster convergence and coupled it with a smart controller, which further saved energy.

Mohd Zain *et al.* [137] proposed an improved DE to optimize the fed-batch substrate feeding rate for several fermentation processes, including methane production from sludge and aerated lagoons treating winery wastewater. The DE proposed is able to optimize the feed rate for higher production yields.

Abimbola *et al.* [138] reviewed and summarized studies about optimization and control strategies using EAs in anaerobic digestion (AD) technologies. Coupling GA with ANN emerged to be one of the most efficient methods in data-based modeling and optimization [139, 140, 141, 142]. Besides, several studies applied DE for process optimization for higher biogas production, better effluent quality and lower operational cost. [143, 135].

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Jean-David Therrien, Niels Nicolaï, and Peter A Vanrolleghem. A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. *Water Science and Technology*, 82(12):2613–2634, 2020.

[2] Stanley B Grant, Jean-Daniel Saphores, David L Feldman, Andrew J Hamilton, Tim D Fletcher, Perran LM Cook, Michael Stewardson, Brett F Sanders, Lisa A Levin, Richard F Ambrose, et al. Taking the "waste" out of "wastewater" for human water security and ecosystem sustainability. *science*, 337(6095):681–686, 2012.

[3] Mark CM van Loosdrecht and Damir Brdjanovic. Anticipating the next century of wastewater treatment. *Science*, 344(6191):1452–1453, 2014.

[4] Water Environment Federation. *Moving Toward Resource Recovery Facilities*. Water Environment Federation, 2014.

[5] Manel Garrido-Baserba, Ll Corominas, Ulises Cortes, Diego Rosso, and Manel Poch. The fourth-revolution in the water sector encounters the digital revolution. *Environmental science & technology*, 54(8):4698–4705, 2020.

[6] Will Sarni, Cassidy White, Randolf Webb, K Cross, and R Glotzbach. Digital water: Industry leaders chart the transformation journey. *International Water Association and Xylem Inc*, 2019.

[7] Frank Blumensaat, João P Leitão, Christoph Ort, Jorg Rieckermann, Andreas Scheidegger, Peter A Vanrolleghem, and Kris Villez. How urban storm-and wastewater management prepares for emerging opportunities and threats: digital transformation, ubiquitous sensing, new data sources, and beyond-a horizon scan. *Environmental science & technology*, 53(15):8488–8498, 2019.

[8] Ll Corominas, M Garrido-Baserba, Kris Villez, Gustaf Olsson, Ulises Cortés, and Manel Poch. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental modelling & software*, 106:89–103, 2018.

[9] Gustaf Olsson, Bengt Carlsson, Joaquim Comas, John Copp, KV Gernaey, P Ingildsen, Ulf Jeppsson, C Kim, L Rieger, Ignasi Rodriguez-Roda, et al. Instrumentation, control and automation in wastewater–from london 1973 to narbonne 2013. *Water Science and Technology*, 69(7):1373–1385, 2014.

[10] Henri Haimi, Michela Mulas, Francesco Corona, and Riku Vahala. Data-derived soft-sensors for biological wastewater treatment plants: An overview. *Environmental Modelling & Software*, 47:88–107, 2013.

[11] Kathryn B Newhart, Ryan W Holloway, Amanda S Hering, and Tzahi Y Cath. Data-driven performance analyses of wastewater treatment plants: A review. *Water research*, 157:498–513, 2019.

[12] Pusker Regmi, Heather Stewart, Youri Amerlinck, Magnus Arnell, Pau Juan García, Bruce Johnson, Thomas Maere, Ivan Miletić, Mark Miller, Leiv Rieger, et al. The future of wrrf modelling–outlook and challenges. *Water Science and Technology*, 79(1):3–14, 2019.

[13] Cheng Yang, Wendy Barrott, Andrea Busch, Anna Mehrotra, Jane Madden, and Glen T. Daigger. How much data is required for a robust and reliable wastewater characterization? *Water Science and Technology*, 79(12):2298–2309, 07 2019.

[14] Mogens Henze, Willi Gujer, Takashi Mino, and Mark CM van Loosdrecht. *Activated sludge models ASM1, ASM2, ASM2d and ASM3*. IWA publishing, 2000.

[15] HM Phillips, KE Sahlstedt, K Frank, J Bratby, W Brennan, S Rogowski, D Pier, W Anderson, M Mulas, JB Copp, et al. Wastewater treatment modelling in practice: a collaborative discussion of the state of the art. *Water Science and Technology*, 59(4):695–704, 2009.

[16] Leiv Rieger, Sylvie Gillot, Günter Langergraber, Takayuki Ohtsuki, Andrew Shaw, Imre Takacs, and Stefan Winkler. *Guidelines for using activated sludge models*. IWA publishing, 2012.

[17] H Hauduc, L Rieger, T Ohtsuki, A Shaw, I Takács, S Winkler, A Héduit, PA Vanrolleghem, and S Gillot. Activated sludge modelling: development and potential use of a practical applications database. *Water Science and Technology*, 63(10):2164–2182, 2011.

[18] Britta Petersen, Peter A Vanrolleghem, Krist Gernaey, and Mogens Henze. Evaluation of an asm1 model calibration procedure on a municipal–industrial wastewater treatment plant. *Journal of Hydroinformatics*, 4(1):15–38, 2002.

[19] Jean-Marc Choubert, Leiv Rieger, Andrew Shaw, John Copp, Mathieu Spérandio, Kim Sørensen, Sabine Rönner-Holm, Eberhard Morgenroth, Henryk Melcer, and Sylvie Gillot. Rethinking wastewater characterisation methods for activated sludge systems–a position paper. *Water Science and Technology*, 67(11):2363–2373, 2013.

[20] PJ Roeleveld and MCM Van Loosdrecht. Experience with guidelines for wastewater characterisation in the netherlands. *Water Science and Technology*, 45(6):77–87, 2002.

[21] Henryk Melcer. *Methods for wastewater characterization in activated sludge modelling*. IWA publishing, 2004.

[22] Peter A Vanrolleghem, Güclü Insel, Britta Petersen, Gürkan Sin, Dirk De Pauw, Ingmar Nopens, Huub Dovermann, Stefan Weijers, and Krist Gernaey. A comprehensive model calibration procedure for activated sludge models. *Proceedings of the Water Environment Federation*, 2003(9):210–237, 2003.

[23] P Lu, X Zhang, and D Zhang. An integrated system for wastewater cod characterization and a case study. *Water Science and Technology*, 62(4):866–874, 2010.

[24] S Gillot and J-M Choubert. Biodegradable organic matter in domestic wastewaters: comparison of selected fractionation techniques. *Water Science and Technology*, 62(3):630–639, 2010.

[25] C Fall, NA Flores, MA Espinoza, G Vazquez, J Loaiza-Návia, MCM Van Loosdrecht, and CM Hooijmans. Divergence between respirometry and physicochemical methods in the fractionation of the chemical oxygen demand in municipal wastewater. *Water Environment Research*, 83(2):162–172, 2011.

[26] APHA. *Standard methods for the examination of water and wastewater*. American Public Health Association, 23rd edition, 2017.

[27] Daniel Mamais, David Jenkins, and Paul Prrr. A rapid physical-chemical method for the determination of readily biodegradable soluble cod in municipal wastewater. *Water research*, 27(1):195–197, 1993.

[28] Jin Yan, Cheng Yang, Zheyi Tian, and Glen T Daigger. Characterizing the performance and operational characterisitics of the bioreactors at the detroit, mi, water resource recovery facility: May, 2017- march 2018 results. Technical report, University of Michigan,Ann Arbor, MI, USA, 2018.

[29] Anna Mehrotra. 2018 BioWin modeling report for GLWA master wastewater plan. Technical report, CDM Smith, Boston, MA, USA, 2018.

[30] Leiv Rieger, Imre Takács, Kris Villez, Hansruedi Siegrist, Paul Lessard, Peter A Vanrolleghem, and Yves Comeau. Data reconciliation for wastewater treatment plant simulation studies—planning for high-quality data and typical sources of errors. *Water environment research*, 82(5):426–433, 2010.

[31] CP Leslie Grady Jr, Glen T Daigger, Nancy G Love, and Carlos DM Filipe. *Biological wastewater treatment*. CRC press, 2011.

[32] Jamal Alikhani, Imre Takacs, Ahmed Al-Omari, Sudhir Murthy, and Arash Massoudieh. Evaluation of the information content of long-term wastewater characteristics data in relation to activated sludge model parameters. *Water Science and Technology*, 75(6):1370–1389, 2017.

[33] John Taylor. *Introduction to error analysis, the study of uncertainties in physical measurements*. University Science Books, 2nd edition, 1997.

[34] H Siegrist, P Krebs, R Bühler, I Purtschert, C Rock, and R Rufer. Denitrification in secondary clarifiers. *Water Science and Technology*, 31(2):205–214, 1995.

[35] Mogens Henze. Characterization of wastewater for modelling of activated sludge processes. *Water Science and Technology*, 25(6):1–15, 1992.

[36] Zhen Zhou, Zhichao Wu, Zhiwei Wang, Shujuan Tang, and Guowei Gu. Cod fractionation and parameter estimation for combined sewers by respirometric tests. *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology*, 83(12):1596–1601, 2008.

[37] J Kappeler and W Gujer. Estimation of kinetic parameters of heterotrophic biomass under aerobic conditions and characterization of wastewater for activated sludge modelling. *Water Science and Technology*, 25(6):125–139, 1992.

[38] Cheng Yang, Glen T. Daigger, Evangelia Belia, and Branko Kerkez. Extracting useful signals from flawed sensor data: Developing hybrid data-driven approaches with physical factors. *Water Research*, 185:116282, 2020.

[39] I Irizar, J Alferes, L Larrea, and E Ayesa. Standard signal processing using enriched sensor information for wwtp monitoring and control. *Water Science and Technology*, 57(7):1053–1060, 2008.

[40] Gustaf Olsson. Ica and me–a subjective review. *Water research*, 46(6):1585–1624, 2012.

[41] Craig C Peddie, Donald S Mavinic, and Christopher J Jenkins. Use of orp for monitoring and control of aerobic sludge digestion. *Journal of environmental engineering*, 116(3):461–471, 1990.

[42] S Plisson-Saune, B Capdeville, M Mauret, A Deguin, and P Baptiste. Real-time control of nitrogen removal using three orp bending-points: signification, control strategy and results. *Water Science and Technology*, 33(1):275–280, 1996.

[43] Sebastià Puig, Lluís Corominas, M Teresa Vives, M Dolors Balaguer, Jesús Colprim, and Joan Colomer. Development and implementation of a real-time control system for nitrogen removal using our and orp as end points. *Industrial & engineering chemistry research*, 44(9):3367–3373, 2005.

[44] María Victoria Ruano, J Ribes, A Seco, and J Ferrer. An advanced control strategy for biological nutrient removal in continuous systems based on ph and orp sensors. *Chemical Engineering Journal*, 183:212–221, 2012.

[45] M Spérandio and I Queinnec. Online estimation of wastewater nitrifiable nitrogen, nitrification and denitrification rates, using orp and do dynamics. *Water Science and Technology*, 49(1):31–38, 2004.

[46] Mariane Yvonne Schneider, Juan Pablo Carbajal, Viviane Furrer, Bettina Sterkele, Max Maurer, and Kris Villez. Beyond signal quality: the value of unmaintained ph, dissolved oxygen, and oxidation-reduction potential sensors for remote performance monitoring of on-site sequencing batch reactors. *Water research*, 161:639–651, 2019.

[47] Christian M Thürlimann, David J Dürrenmatt, and Kris Villez. Soft-sensing with qualitative trend analysis for wastewater treatment plant control. *Control Engineering Practice*, 70:121–133, 2018.

[48] Charu C Aggarwal. An introduction to outlier analysis. In *Outlier analysis*, pages 1–34. Springer, 2017.

[49] Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han. Outlier detection for temporal data. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 5(1):1–129, 2014.

[50] Sana Ullah Jan, Young-Doo Lee, Jungpil Shin, and Insoo Koo. Sensor fault classification based on support vector machine and statistical time-domain features. *IEEE Access*, 5:8682–8690, 2017.

[51] Patrick JF Groenen, Georgi Nalbantov, and Jan C Bioch. Svm-maj: a majorization approach to linear support vector machines with different hinge errors. *Advances in data analysis and classification*, 2(1):17–43, 2008.

[52] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[53] Puripus Soonthornnonda and Erik R Christensen. Source apportionment of pollutants and flows of combined sewer wastewater. *Water research*, 42(8-9):1989–1998, 2008.

[54] Zhijian Yuan and Erkki Oja. A fastica algorithm for non-negative independent component analysis. In *International Conference on Independent Component Analysis and Signal Separation*, pages 1–8. Springer, 2004.

[55] Karl Svardal, Stefan Lindtner, and Stefan Winkler. Optimum aerobic volume control based on continuous in-line oxygen uptake monitoring. *Water Science and Technology*, 47(11):305–312, 2003.

[56] Kris Villez. Qualitative path estimation: A fast and reliable algorithm for qualitative trend analysis. *AIChE Journal*, 61(5):1535–1546, 2015.

[57] Moritz Von Stosch, Rui Oliveira, Joana Peres, and Sebastião Feyo de Azevedo. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering*, 60:86–101, 2014.

[58] Richard C Aster, Brian Borchers, and Clifford H Thurber. *Parameter estimation and inverse problems*. Elsevier, 2018.

[59] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[60] Mei-jin Lin, Cai-xia Zhang, and Cai-hong Su. Prediction of effluent from wwtps using differential evolutionary extreme learning machines. In *2016 35th Chinese Control Conference (CCC)*, pages 2034–2038, 2016.

[61] L Sutherland-Stacey and R Dexter. On the use of non-negative matrix factorisation to characterise wastewater from dairy processing plants. *Water Science and Technology*, 64(5):1096–1101, 2011.

[62] Jens Hainmueller and Chad Hazlett. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168, 2014.

[63] Kris Villez, Venkat Venkatasubramanian, and Raghunathan Rengaswamy. Generalized shape constrained spline fitting for qualitative analysis of trends. *Computers & chemical engineering*, 58:116–134, 2013.

[64] Sara C Troutman, Nathaniel Schambach, Nancy G Love, and Branko Kerkez. An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. *Water research*, 126:88–100, 2017.

[65] Seung Han Woo, Che Ok Jeon, Yeoung-Sang Yun, Hyeoksun Choi, Chang-Soo Lee, and Dae Sung Lee. On-line estimation of key process variables based on kernel partial least squares in an industrial cokes wastewater treatment plant. *Journal of Hazardous Materials*, 161(1):538–544, 2009.

[66] Cheng Yang, Peter Seiler, Evangelia Belia, and Glen T. Daigger. An adaptive real-time grey-box model for advanced control and operations in WRRFs. *Water Science and Technology*, 84(9):2353–2365, 09 2021.

[67] Bruce E Rittmann, Joshua P Boltz, Doris Brockmann, Glen T Daigger, Eberhard Morgenroth, Kim Helleshøj Sørensen, Imre Takács, Mark Van Loosdrecht, and Peter A Vanrolleghem. A framework for good biofilm reactor modeling practice (gbrmp). *Water Science and Technology*, 77(5):1149–1164, 2018.

[68] Damien J Batstone, J Keller, Irini Angelidaki, SV Kalyuzhnyi, SG Pavlostathis, A Rozzi, WTM Sanders, HA Siegrist, and VA Vavilin. The iwa anaerobic digestion model no 1 (adm1). *Water Science and technology*, 45(10):65–73, 2002.

[69] G Zahedi, A Elkamel, A Lohi, A Jahanmiri, and MR Rahimpor. Hybrid artificial neural network—first principle model formulation for the unsteady state simulation and analysis of a packed bed reactor for co2 hydrogenation to methanol. *Chemical Engineering Journal*, 115(1-2):113–120, 2005.

[70] B Shiva Kumar and Ch Venkateswarlu. Estimating biofilm reaction kinetics using hybrid mechanistic-neural network rate function model. *Bioresource Technology*, 103(1):300–308, 2012.

[71] Abhilash M Nair, Abaynesh Fanta, Finn Aakre Haugen, and Harsha Ratnaweera. Implementing an extended kalman filter for estimating nutrient composition in a sequential batch mbbr pilot plant. *Water Science and Technology*, 80(2):317–328, 2019.

[72] Peter Alexander Stentoft, Thomas Munk-Nielsen, Luca Vezzaro, Henrik Madsen, Peter Steen Mikkelsen, and Jan Kloppenborg Møller. Towards model predictive control: online predictions of ammonium and nitrate removal by using a stochastic asm. *Water Science and Technology*, 79(1):51–62, 2019.

[73] Aljaž Stare, Nadja Hvala, and Darko Vrečko. Modeling, identification, and validation of models for predictive ammonia control in a wastewater treatment plant—a case study. *ISA transactions*, 45(2):159–174, 2006.

[74] Douglas G Robertson, Jay H Lee, and James B Rawlings. A moving horizon-based approach for least-squares estimation. *AIChE Journal*, 42(8):2209–2224, 1996.

[75] Christopher V Rao, James B Rawlings, and David Q Mayne. Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations. *IEEE transactions on automatic control*, 48(2):246–258, 2003.

[76] Greg Welch, Gary Bishop, et al. *An introduction to the Kalman filter*. Chapel Hill, NC, USA, 1995.

[77] Jan Busch, David Elixmann, Peter Kühl, Carine Gerkens, Johannes P Schlöder, Hans G Bock, and Wolfgang Marquardt. State estimation for large-scale wastewater treatment plants. *Water research*, 47(13):4774–4787, 2013.

[78] J Alex, L Benedetti, J Copp, KV Gernaey, Ulf Jeppsson, I Nopens, MN Pons, L Rieger, C Rosen, JP Steyer, et al. Benchmark simulation model no. 1 (bsm1). *Report by the IWA Taskgroup on benchmarking of control strategies for WWTPs*, pages 19–20, 2008.

[79] Abhilash M Nair, Blanca M Gonzalez-Silva, Finn Aakre Haugen, Harsha Ratnaweera, and Stein W Østerhus. Real-time monitoring of enhanced biological phosphorus removal in a multistage ebpr-mbbr using a soft-sensor for phosphates. *Journal of Water Process Engineering*, 37:101494, 2020.

[80] Thomas Kailath, Ali H Sayed, and Babak Hassibi. *Linear estimation*. Prentice Hall, 2000.

[81] E Ayesa, J Florez, JL García-Heras, and L Larrea. State and coefficients estimation for the activated sludge process using a modified kalman filter algorithm. *Water Science and Technology*, 24(6):235–247, 1991.

[82] Cheng Yang, Evangelia Belia, and Glen T. Daigger. Automating Process Design by Coupling Genetic Algorithms with Commercial Simulators: A Case Study for Hybrid MABR processes. *Water Science and Technology*, 2022.

[83] Evangelia Belia, Marc B. Neumann, Lorenzo Benedetti, Bruce Johnson, Sudhir Murthy, Stefan Weijers, and Peter P.A. Vanrolleghem. *Uncertainty in Wastewater Treatment Design and Operation*. IWA Publishing, nov 2021.

[84] A. Rivas, I. Irizar, and E. Ayesa. Model-based optimisation of Wastewater Treatment Plants design. *Environmental Modelling & Software*, 23(4):435–450, apr 2008.

[85] John H. Holland. Genetic algorithms and the optimal allocation of trials. *SIAM J. Comput.*, 2(2):88–105, 1973.

[86] Daniel H Loughlin, Troy A Doby, Joel J Ducoste, and Francis L de los Reyes, III. System-wide optimization of wastewater treatment plants using genetic algorithms. In *Bridging the Gap: Meeting the World's Water and Environmental Resources Challenges*, pages 1–10, 2001.

[87] NI-BIN CHANG, WC Chen, and Wen K Shieh. Optimal control of wastewater treatment plants via integrated neural network and genetic algorithms. *Civil Engineering Systems*, 18(1):1–17, 2001.

[88] Hiroki Yoshikawa, Taizo Hanai, Shuta Tomida, Hiroyuki Honda, and Takeshi Kobayashi. Determination of operating conditions in activated sludge process using fuzzy neural network and genetic algorithm. *Journal of chemical engineering of Japan*, 34(8):1033–1039, 2001.

[89] TA Doby, DH Loughlin, FL De Los Reyes III, and JJ Ducoste. Optimization of activated sludge designs using genetic algorithms. *Water science and technology*, 45(6):187–198, 2002.

[90] S Kim, H Lee, J Kim, C Kim, J Ko, H Woo, and S Kim. Genetic algorithms for the application of activated sludge model no. 1. *Water Science and Technology*, 45(4-5):405–411, 2002.

[91] RAFAŁ Urban and Ryszard Szetela. Calibration of the activated sludge model with genetic algorithms. part i. calibration results. *Environment Protection Engineering*, 33(1):31, 2007.

[92] RAFAŁ Urban and Ryszard Szetela. Calibration of the activated sludge model with genetic algorithms. part ii. analysis of results. *Environment Protection Engineering*, 33(1):51, 2007.

[93] X Flores, A Bonmati, M Poch, IR Roda, L Jimenez, and R Banares-Alcantara. Multicriteria evaluation tools to support the conceptual design of activated sludge systems. *Water Science and Technology*, 56(6):85–94, 2007.

[94] Wenliang Chen, Chonghua Yao, and Xiwu Lu. Optimal design activated sludge process by means of multi-objective optimization: case study in benchmark simulation model 1 (bsm1). *Water science and technology*, 69(10):2052–2058, 2014.

[95] Jussi Hakanen, Kaisa Miettinen, and Kristian Sahlstedt. Wastewater treatment: New insight provided by interactive multiobjective optimization. *Decision Support Systems*, 51(2):328–337, 2011.

[96] Jussi Hakanen, Kristian Sahlstedt, and Kaisa Miettinen. Wastewater treatment plant design and operation under multiple conflicting objective functions. *Environmental modelling & software*, 46:240–249, 2013.

[97] Benoît Beraud, Cyrille Lemoine, and Jean-Philippe Steyer. Multiobjective genetic algorithms for the optimisation of wastewater treatment processes. In *Computational Intelligence Techniques for Bioprocess Modelling, Supervision and Control*, pages 163–195. Springer, 2009.

[98] Jawed Iqbal and Chandan Guria. Optimization of an operating domestic wastewater treatment plant using elitist non-dominated sorting genetic algorithm. *Chemical Engineering Research and Design*, 87(11):1481–1496, 2009.

[99] J Guerrero, A Guisasola, J Comas, I Rodríguez-Roda, and JA Baeza. Multicriteria selection of optimum wwtp control setpoints based on microbiology-related failures, effluent quality and operating costs. *Chemical Engineering Journal*, 188:23–29, 2012.

[100] Junfei Qiao and Wei Zhang. Dynamic multi-objective optimization control for wastewater treatment process. *Neural Computing and Applications*, 29(11):1261–1271, 2018.

[101] ESS Tejaswini, Soniya Panjwani, Uday Bhaskar Babu Gara, and Seshagiri Rao Ambati. Multi-objective optimization based controller design for improved wastewater treatment plant operation. *Environmental Technology & Innovation*, 23:101591, 2021.

[102] T Ludwig, P Kern, M Bongards, and C Wolf. Simulation and optimization of an experimental membrane wastewater treatment plant using computational intelligence methods. *Water Science and Technology*, 63(10):2255–2260, 2011.

[103] Avery L Carlson, Huanqi He, Cheng Yang, and Glen T Daigger. Comparison of hybrid membrane aerated biofilm reactor (mabr)/suspended growth and conventional biological nutrient removal processes. *Water Science and Technology*, 83(6):1418–1428, 2021.

[104] GT Daigger, AL Carlson, BR Johnson, and X Chen. Coupled anoxic suspended growth and membrane aerated biofilm reactor process options. In *Proceedings of the 92nd Annual Water Environment Federation Technical Exposition & Conference; Chicago, Illinois*, pages 21–25. Water Environment Federation Alexandria, VA, 2019.

[105] Huanqi He, Brett M Wagner, Avery L Carlson, Cheng Yang, and Glen T Daigger. Recent progress using membrane aerated biofilm reactors for wastewater treatment. *Water Science and Technology*, 84(9):2131–2157, 2021.

[106] Leon S Downing and Robert Nerenberg. Total nitrogen removal in a hybrid, membrane-aerated activated sludge process. *Water Research*, 42(14):3697–3708, 2008.

[107] Dwight Houweling, Jeff Peeters, Pierre Cote, Zebo Long, and Nick Adams. Proving membrane aerated biofilm reactor (mabr) performance and reliability: Results from four pilots and a full-scale plant. *Proceedings of the Water Environment Federation*, 2017(16):272–284, 2017.

[108] Duowei Lu, Hao Bai, Fangong Kong, Steven N Liss, and Baoqiang Liao. Recent advances in membrane aerated biofilm reactors. *Critical Reviews in Environmental Science and Technology*, 51(7):649–703, 2021.

[109] Dwight Houweling and Glen T Daigger. *Intensifying Activated Sludge Using Media-Supported Biofilms*. CRC Press, 2019.

[110] Leon S Downing and Robert Nerenberg. Effect of bulk liquid bod concentration on activity and microbial community structure of a nitrifying, membrane-aerated biofilm. *Applied Microbiology and Biotechnology*, 81(1):153–162, 2008.

[111] Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[112] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[113] Jerome Andre, Patrick Siarry, and Thomas Dognon. An improvement of the standard genetic algorithm fighting premature convergence in continuous optimization. *Advances in engineering software*, 32(1):49–60, 2001.

[114] M. Janga Reddy and D. Nagesh Kumar. Evolutionary algorithms, swarm intelligence methods, and their applications in water resources engineering: a state-of-the-art review. *H2Open Journal*, 3(1):135–188, 06 2020.

[115] JR Kim, JH Ko, JJ Lee, SH Kim, TJ Park, CW Kim, and HJ Woo. Parameter sensitivity analysis for activated sludge models no. 1 and 3 combined with one-dimensional settling model. *Water science and technology*, 53(1):129–138, 2006.

[116] Andrea Saltelli, Ksenia Aleksankina, William Becker, Pamela Fennell, Federico Ferretti, Niels Holst, Sushan Li, and Qiongli Wu. Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental modelling & software*, 114:29–39, 2019.

[117] Majid Bagheri, Sayed Ahmad Mirbagheri, Zahra Bagheri, and Ali Morad Kamarkhani. Modeling and optimization of activated sludge bulking for a real wastewater treatment plant using hybrid artificial neural networks-genetic algorithm approach. *Process Safety and Environmental Protection*, 95:12–25, 2015.

[118] Seyed Ahmad Mirbagheri, Majid Bagheri, Zahra Bagheri, and Ali Morad Kamarkhani. Evaluation and prediction of membrane fouling in a submerged membrane bioreactor with simultaneous upward and downward aeration using artificial neural network-genetic algorithm. *Process Safety and Environmental Protection*, 96:111–124, 2015.

[119] Wenliang Chen, Xiwu Lu, Chonghua Yao, Guangcan Zhu, and Zhuo Xu. An efficient approach based on bi-sensitivity analysis and genetic algorithm for calibration of activated sludge models. *Chemical engineering journal*, 259:845–853, 2015.

[120] Antonia Hadjimichael, Joaquim Comas, and Lluís Corominas. Do machine learning methods used in data mining enhance the potential of decision support systems? a review for the urban water sector. *AI Communications*, 29(6):747–756, 2016.

[121] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[122] Rui Nian, Jinfeng Liu, and Biao Huang. A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, 139:106886, 2020.

[123] Joohyun Shin, Thomas A Badgwell, Kuang-Hung Liu, and Jay H Lee. Reinforcement learning–overview of recent progress and implications for process control. *Computers & Chemical Engineering*, 127:282–294, 2019.

[124] Douglas Alves Goulart and Renato Dutra Pereira. Autonomous ph control by reinforcement learning for electroplating industry wastewater. *Computers & Chemical Engineering*, 140:106909, 2020.

[125] Jorge Filipe, Ricardo J Bessa, Marisa Reis, Rita Alves, and Pedro Póvoa. Data-driven predictive energy optimization in a wastewater pumping station. *Applied Energy*, 252:113423, 2019.

[126] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[127] Jason DM Rennie and Nathan Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, volume 1. Citeseer, 2005.

[128] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

[129] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[130] Pradnya A. Vikhar. Evolutionary algorithms: A critical review and its future prospects. In *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, pages 261–265, 2016.

[131] Ingo Rechenberg. Evolutionsstrategien. In *Simulationsmethoden in der Medizin und Biologie*, pages 83–114. Springer, 1978.

[132] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[133] Milad Janghorban Lariche, Shahrzad Soltani, Hossein Davoudi Nezhad, Hamed Moradi, Sheyda Soltani, Fatemeh Farsayad, and Hossein Moradi Kazerouni. Developing supervised models for estimating methylene blue removal by silver nanoparticles. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 42(10):1247–1254, 2020.

[134] Hossein Bonakdari, Isa Ebtehaj, and Azam Akhbari. Multi-objective evolutionary polynomial regression-based prediction of energy consumption probing. *Water Science and Technology*, 75(12):2791–2799, 2017.

[135] Víctor Rivera-Salvador, Irineo L López-Cruz, Teodoro Espinosa-Solares, Juan S Aranda-Barradas, David H Huber, Deepak Sharma, and J Ulises Toledo. Application of anaerobic digestion model no. 1 to describe the syntrophic acetate oxidation of poultry litter in thermophilic anaerobic digestion. *Bioresource technology*, 167:495–502, 2014.

[136] Hongbiao Zhou and Junfei Qiao. Multiobjective optimal control for wastewater treatment process using adaptive moea/d. *Applied Intelligence*, 49(3):1098–1126, 2019.

[137] Mohamad Zihin bin Mohd Zain, Jeevan Kanesan, Graham Kendall, and Joon Huang Chuah. Optimization of fed-batch fermentation processes using the backtracking search algorithm. *Expert Systems with Applications*, 91:286–297, 2018.

[138] Abimbola M. Enitan, Josiah Adeyemo, Feroz M. Swalaha, Sheena Kumari, and Faizal Bux. Optimization of biogas generation using anaerobic digestion models and computational intelligence approaches. *Reviews in Chemical Engineering*, 33(3):309–335, 2017.

[139] E Martinez, A Marcos, A Al-Kassir, MA Jaramillo, and AA Mohamad. Mathematical model of a laboratory-scale plant for slaughterhouse effluents biodigestion for biogas production. *Applied energy*, 95:210–219, 2012.

[140] Mingzhi Huang, Wei Han, Jinquan Wan, Yongwen Ma, and Xiaohong Chen. Multi-objective optimisation for design and operation of anaerobic digestion using ga-ann and nsga-ii. *Journal of Chemical Technology & Biotechnology*, 91(1):226–233, 2016.

[141] EB Gueguim Kana, JK Oloke, A Lateef, and MO Adesiyan. Modeling and optimization of biogas production on saw dust and other co-substrates using artificial neural network and genetic algorithm. *Renewable energy*, 46:276–281, 2012.

[142] H Abu Qdais, K Bani Hani, and N Shatnawi. Modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm. *Resources, Conservation and Recycling*, 54(6):359–363, 2010.

[143] Abimbola Motunrayo Enitan, Josiah Adeyemo, Feroz Mahomed Swalaha, and Faizal Bux. Anaerobic digestion model to enhance treatment of brewery wastewater for biogas production using uasb reactor. *Environmental Modeling & Assessment*, 20(6):673–685, 2015.