Influent generator: Towards realistic modelling of wastewater flowrate and water quality using machine-learning methods

Thèse

Feiyi Li

Doctorat en génie des eaux
Philosophiae doctor (Ph.D.)

Québec, Canada

Influent generator: Towards realistic modelling of wastewater flowrate and water quality using machine-learning methods

Thèse

Feiyi Li

Sous la direction de :

Peter A. Vanrolleghem, directeur de recherche

Résumé

Depuis que l'assainissement des eaux usées est reconnu comme un des objectifs de développement durable des Nations Unies, le traitement et la gestion des eaux usées sont devenus plus importants que jamais. La modélisation et la digitalisation des stations de récupération des ressources de l'eau (StaRRE) jouent un rôle important depuis des décennies, cependant, le manque de données disponibles sur les affluents entrave le développement de la modélisation de StaRRE.

Cette thèse vise à faire progresser la modélisation des systèmes d'assainissement en général, et en particulier en ce qui concerne la génération dynamique des affluents. Dans cette étude, différents générateurs d'affluent (GA), qui peuvent fournir un profil d'affluent dynamique, ont été proposés, optimisés et discutés. Les GA développés ne se concentrent pas seulement sur le débit, les solides en suspension et la matière organique, mais également sur les substances nutritives telles que l'azote et le phosphore. En outre, cette étude vise à adapter les GA à différentes applications en fonction des différentes exigences de modélisation. Afin d'évaluer les performances des GA d'un point de vue général, une série de critères d'évaluation de la qualité du modèle est décrite.

Premièrement, pour comprendre la dynamique des affluents, une procédure de caractérisation des affluents a été développée et testée pour une étude de cas à l'échelle pilote. Ensuite, pour générer différentes séries temporelles d'affluent, un premier GA a été développé. La méthodologie de modélisation est basée sur l'apprentissage automatique en raison de ses calculs rapides, de sa précision et de sa capacité à traiter les mégadonnées. De plus, diverses versions de ce GA ont été appliquées pour différents cas d'études et ont été optimisées en fonction des disponibilités des données (la fréquence et l'horizon temporel), des objectifs et des exigences de précision.

Les résultats démontrent que : i) le modèle GA proposé peut être utilisé pour générer d'affluents dynamiques réalistes pour différents objectifs, et les séries temporelles résultantes incluent à la fois le débit et la concentration de polluants avec une bonne précision et distribution statistique; ii) les GA sont flexibles, ce qui permet de les améliorer selon différents objectifs d'optimisation; iii) les GA ont été développés en considérant l'équilibre entre les efforts de modélisation, la collecte de données requise et les performances du modèle.

Basé sur les perspectives de modélisation des StaRRE, l'analyse des procédés et la modélisation prévisionnelle, les modèles de GA dynamiques peuvent fournir aux concepteurs et aux modélisateurs un profil d'affluent complet et réaliste, ce qui permet de surmonter les obstacles liés au manque de données d'affluent. Par conséquent, cette étude a démontré l'utilité des GA et a fait avancer la modélisation des StaRRE en focalisant sur l'application de méthodologies d'exploration de données et d'apprentissage automatique. Les GA peuvent

donc être utilisés comme outil puissant pour la modélisation des StaRRE, avec des applications pour l'amélioration de la configuration de traitement, la conception de procédés, ainsi que la gestion et la prise de décision stratégique. Les GA peuvent ainsi contribuer au développement de jumeaux numériques pour les StaRRE, soit des système intelligent et automatisé de décision et de contrôle.

Mots-clés : apprentissage automatique, pilotage par la donnée, exploration des données, eaux numériques, la conception et le contrôle des StaRRE

Abstract

Since wastewater sanitation is acknowledged as one of the sustainable development goals of the United Nations, wastewater treatment and management have been more important then ever. Water Resource Recovery Facility (WRRF) modelling and digitalization have been playing an important role since decades, however, the lack of available influent data still hampers WRRF model development.

This dissertation aims at advancing the field of wastewater systems modelling in general, and in particular with respect to the dynamic influent generation. In this study, different WRRF influent generators (IG), that can provide a dynamic influent flow and pollutant concentration profile, have been proposed, optimized and discussed. The developed IGs are not only focusing on flowrate, suspended solids, and organic matter, but also on nutrients such as nitrogen and phosphorus. The study further aimed at adapting the IGs to different case studies, so that future users feel comfortable to apply different IG versions according to different modelling requirements. In order to evaluate the IG performance from a general perspective, a series of criteria for evaluating the model quality were evaluated.

Firstly, to understand the influent dynamics, a procedure of influent characterization has been developed and experimented at pilot scale. Then, to generate different realizations of the influent time series, the first IG was developed and a data-driven modelling approach chosen, because of its fast calculations, its precision and its capacity of handling big data. Furthermore, different realizations of IGs were applied to different case studies and were optimized for different data availabilities (frequency and time horizon), objectives, and modelling precision requirements.

The overall results indicate that: i) the proposed IG model can be used to generate realistic dynamic influent time series for different case studies, including both flowrate and pollutant concentrations with good precision and statistical distribution; ii) the proposed IG is flexible and can be improved for different optimization objectives; iii) the IG model has been developed by considering the balance between modelling efforts, data collection requirements and model performance.

Based on future perspectives of WRRF process modelling, process analysis, and forecasting, the dynamic IG model can provide designers and modellers with a complete and realistic influent profile and this overcomes the often-occurring barrier of shortage of influent data for modelling. Therefore, this study demonstrated the IGs' usefulness for advanced WRRF modelling focusing on the application of data mining and machine learning methodologies. It is expected to be widely used as a powerful tool for WRRF modelling, improving treatment configurations and process designs, management and strategic decision-making, such as when transforming a conventional WRRF to a digital twin that can be used as an intelligent and automated system.

Key words: machine learning, data-driven models, data mining, digital water, WRRF design and control

Table of contents

List of Figures

List of Tables

List of Abbreviations

Acronyms	
ASM	Activated Sludge Model
ANN	Artificial neural network
BOD5	5-day biochemical oxygen demand
BSM	Benchmark Simulation Model
COD	Chemical oxygen demand
CDH	Clos de Hilde
DWF	Dry weather flow
GA	Genetic algorithm
IG	Influent generator
IUWS	Integrated Urban Wastewater System
KL	Kullback-Leibler (divergency)
LSTM	Long short-term memory
MPE	Mean Percentage Error
N	Nitrogen
NH4	Ammonia
P	Phosphorus
PDF	Probability density function
PE	Person equivalents
RDII	Rainfall-dependent infiltration and inflow
RMSE	Root Mean Square Error
RNN	Recurrent neural network
RTC	Real-time control
TSS	Total suspended solids
VFA	Volatile fatty acids
VSS	Volatile suspended solids
WRRF	Water resource recovery facility
WWF	Wet weather flow
WWTP	Wastewater treatment plant

< *Long ago, when I started,*
the snowflakes fly.
Now that I turn back,
the willows spread their shade. >

Acknowledgements

Accompanied by the most beautiful snow, I arrived in this country on a memorable day in January 2018 to start on this incredible journey pursuing a PhD. Without a doubt, it was and will always remain one of my life's fantastic experiences.

First, I would like to express special thanks to my supervisor, professor Peter A. Vanrolleghem. Peter, thank you for letting me be a part of the wonderful modelEAU group. Your academic mind, enthusiastic attitude, and your immense kindness have influenced me so much while pursuing my PhD. In addition to the planned intensive discussions, you also participated in spontaneous ones and still managed to provide clear and valuable inputs at the same time. Your understanding of the field of Urban Water Engineering encouraged me to make this contribution and finish this thesis.

Furthermore, I would also like to express my gratitude to the reviewing committee members: Dr. François Anctil, Dr. Paul Lessard, and Dr. Thibaud Maruéjouls, whose in-depth comments were not only very much appreciated but also greatly improved the content of this thesis.

Of course, I would also like to thank the entire modelEAU research team for all your support. I was glad to be part of such a friendly and creative team. Thanks, Julia, for collaborating with me on the MOSAIQUE project and especially with your help at the beginning of my PhD. A special thanks to the pilEAUte and Daqua teams including Christophe, Gamze, Jean-David, Jeffrey, Karen, Maryam, Nathalia, Niels, Rania, Romain, Sovanna, and also other colleagues, friends, interns in the team, including Andreia, Asma, Bernard, Cyril, Elena, Kamilia, Marcello, Majid, Maxine, Morgane, Queralt, and Thomas, who have supported me constantly.

I would like to express my gratitude to our research partners, Quebec City and the SUEZ Le Lyre research center in Talence, France, who were indispensable towards the completion of this research. I would also like to thank the Natural Science and Engineering Research Council (NSERC) and SUEZ Canada for the financial support of this PhD research.

Last but certainly not least, I would like to thank my parents, my family and my friends in China, France and Canada, who have always supported me in any difficult time despite of the distance. This includes Stefano, who has always been by my side and supported me in every possible way.

Foreword

This dissertation was written in the framework of the research project MOSAIQUE, the French abbreviation of 'MOdélisation du Système d'Assainissement Intégré basée sur la QUalité des Eaux', which stands for modelling the integrated urban wastewater system based on water quality.

This research project was funded by a Collaborative Research and Development grant of the Natural Sciences and Engineering Research Council (NSERC) and Suez Treatment Solutions Canada. The research was conducted by the modelEAU research group and CentrEau, the Quebec Water Research Centre funded by FRQNT. Additional important technical support was obtained from Quebec City, Bordeaux Metropole, Société de Gestion de l'Assainissement de Bordeaux Métropole (SGAC) and Suez's Le LyRE research center.

This research project included three case studies with both theoretical and practical work phases. The data collection at the pilEAUte WRRF installed at Université Laval (Québec, QC, Canada), was performed by the modelEAU research group under the supervision of P.A. Vanrolleghem. For the Quebec City case study, data were collected from the WRRF EST with the collaboration of Y. Lanthier, D. Dufour and F. Cloutier, employees of Quebec City. For the case study of Bordeaux, data collection work was conducted by J.M. Ledergerber, PhD student at modelEAU, at the Clos de Hilde (CdH) water resource recovery facility (WRRF) in Bordeaux, France, in collaboration with the research center Le LyRE, Suez, Talence, France, under the supervision of T. Maruéjouls.

The dissertation is presented in paper format with eight chapters, four of which are published as journal or conference papers. Instead of presenting the work chronologically, the presented papers were redrafted and ordered in a logical manner, allowing successive chapters to build upon previous ones.

The following peer-reviewed papers are included as chapters in the dissertation:

Chapter 4: Li, F., Vanrolleghem, P.A., 2022. An essential tool for WRRF modelling: A realistic and complete influent generator for flow rate and water quality based on data-driven methods. Water Sci. Technol. wst2022095.

Chapter 5: Li, F., Vanrolleghem, P.A., 2022. An influent generator for WRRF design and operation based on a recurrent neural network with multi-objective optimization using a genetic algorithm. Water Sci. Technol. 85, 1444–1453.

Chapter 6: Li, F., Vanrolleghem, P.A., 2022. Including snowmelt in influent generation for cold climate WRRFs: comparison of data-driven and phenomenological approaches. Environ. Sci.: Water Res. Technol., 2022.

Chapter 7: Li, F., Vanrolleghem, P.A., 2022. Data-driven influent generator for WRRF database gap filling and model-based control evaluation. *Submitted to IWA World Water Congress & Exhibition, Copenhagen, September 2022*

Feiyi Li (FL) is the first author of each paper in this thesis. The author contributions are as follows:

Chapter 4: FL: data acquisition and preparation, model conceptualization, methodology and analysis, writing-review-edition; PAV: data acquisition, model conceptualization, result interpretation, project supervision and administration, funding, edition and review.

Chapter 5: FL: model conceptualization, methodology optimization, writing-review- refining. PAV: methodology, result interpretation, project supervision and administration, funding, edition and review.

Chapter 6: FL: data preparation, model conceptualization, methodology, writing, review & editing. PAV: methodology, result interpretation, project supervision and administration, funding, edition and review.

Chapter 7: FL: data acquisition and data treatment, model conceptualization, methodology and writing, visualization, refining and edition. PAV: methodology, result interpretation, project supervision and administration, funding, edition and review.

Introduction

Problem statement and objectives

Nowadays, water resource protection is becoming increasingly critical, as an important part of the Sustainable Development Goals (Malik et al., 2015). As an important method to protect the public health and receiving waters from pollution, wastewater treatment plants (WWTPs) receive the urban sanitary wastewater, industrial waste and storm water from the sewer system. Recently, the transition from traditional WWTPs to water resource recovery facilities (WRRFs) is increasingly taking place.

WRRF modelling has been developing over decades and different WRRF modelling works are available based on the activated sludge model (ASM) family extended with other biological and physical process frameworks.

More and more models are used for practical and research projects. Adequate models demonstrate their usefulness for WRRF design, upgrade, operation optimization, cost reduction and the transition of WWTPs into WRRFs. In recent years, the digital twin has increasingly gained attention for smart real-time control, cost reduction, energy efficiency increase and climate change mitigation (Pedersen et al., 2021). Although it is known that WRRF model performance depends on a decent influent description (e.g. influent flow and load), many issues of the modelling of the input time series have been exposed.

In conventional design guidelines, most often the influent variables are taken under steady state conditions (Talebizadeh, 2015). Dynamics and stochastics are not often incorporated in the average values, with daily or monthly-based peaking factors accounting for diurnal and seasonal variability in flows and wastewater strength. These peaking factors are extracted from historical data and from empirical equations, or based on engineering judgement, which might cause under- or overestimation of influent load. Usually, under steady state conditions, a wastewater treatment plant will perform correctly, but most of the observed problems occur during important load variations. In fact, the influent, as input of WRRF models, is an important source of uncertainty (Refsgaard et al., 2007).

On the other hand, the modelling for operation optimization (controller design, digital twin modelling) needs high frequency and long-term time series. However, typically only an insufficiently complete dataset is available, that is imprecise and only gives an undetailed description of the dynamics and in this way leads to difficulties for the model training and calibration.

In addition, advanced control systems (e.g. an ammonia-based controller) usually require high frequency and real-time forecasted influent data, this becoming a critical prerequisite for control strategy development and application.

Hence, it is crucial and challenging to develop a reliable influent generator (IG) to predict the water quantity and quality entering a WRRF, enabling a better simulation of the system's hydraulic configuration requirements (volume, flows) and its dynamic behaviour under disturbances.

Although different IG models are already available, as will be shown in the literature review, to achieve a practical and reliable IG model, different challenges remain:

Challenge 1: Description of the influent dynamics that correctly integrates wastewater production and transport processes.

It is known that the water generation process in a combined sewer system is complex and nonlinear. A simple sewer model is typically not adequate for describing the pollutant transport in the sewer processes: for instance, the fate of solids that settle during dry weather and resuspend under rain conditions, a comprehensive description of the dilution and storage, and biodegradation processes that may occur as well. The lack of physical, chemical and biological transformation processes hampers an adequate water quality estimation at the entry of a WRRF, which might therefore not lead to a satisfactory description of the loads, particularly regarding the influent composition (e.g. production of volatile fatty acids affecting Bio-P removal), the correlations between the different influent constituents (COD, TN, TP, TSS) and the recurrence periods of peak phenomena.

Challenge 2: Completed influent variables and temporal resolution.

The existing models' influent time series are insufficiently complete. For example, ionic composition is not included yet (NH_4^+, K^+, conductivity) and the fractionation dynamics are not well studied, even though they are an essential part of biodegradability analysis and coupling with ASMs. On the other hand, the level of temporal resolution (e.g. hourly, 4-hourly, daily input/output) and time horizon (monthly or yearly) depends on the model purpose, such as WRRF design, control system design, data gap filling etc. The optimal level of temporal resolution for the scenario generation needs to be investigated.

Challenge 3: Balance between model complexity and model performance evaluation. In the literature review and practical engineering, different influent submodels are applied in the WRRF simulators, such as WEST by DHI, SIMBA by ifak, BSM provided by IEA-Lund University and WRRF influent generator coupled to CITYDRAIN (Achleitner et al., 2007; Talebizadeh et al., 2016) to name a few. The various influent generators methods have been successful to different degrees. However, in practice, it is difficult to balance model complexity and performance. Targeting engineering applications, the IG model is expected to be designed as user-friendly as possible, in order to offer engineers a simple tool with a credible result.

Moreover, phenomenological or physical models included in IG need considerable catchment details (area, sewershed, soil infiltration properties, etc.), but these parameters are not always available for IG calibration and validation. These difficulties and investments of modelling calibration (effort and time consuming) hinder the usage of IG. Therefore, pursuing a good balance between parsimony and complexity of the IG model becomes critical in improving model performance and reducing the costs of modelling. And most of all, the IG model performance is influenced by many aspects: the model's complexity, its capacity of generalization and transferability, the process interpretability, the data needed, the modelling efforts and precision, the representation of the time series' variability, etc. Therefore, to evaluate the IG's performance there is a need to develop a complete evaluation of a set of the performance criteria.

The objectives of this PhD project are manifold in both scientific and practical perspectives. The research aims at filling some of the research gaps and improve the existing methodology in influent time series generation and analysis. These include the characterization and visualization of influent data, the development of a set of methods for influent variable generation for WRRF design or influent prediction, with different time horizons and data frequency, and with the consideration of the application needs. The IG model aims to be simple, easy-to-use and accurate. Therefore, the model's complexity will be an important aspect to optimize.

The main objectives of this PhD project are represented by the four following sub-objectives.

- **To advance the methodology of efficient characterization, visualization and analysis of the influent collected from combined sewer systems (CSS) at the inlet of a WRRF**

The first objective is the efficient characterization of raw domestic wastewater, i.e. the dynamics of flowrate and pollutants, as well as pollutant composition. The pil*EAU*te plant is situated at the end of a very small sewer catchment. This means there is hardly any dampening effect, thus enabling the characterization of the dynamics of the raw wastewater generation, including COD, TSS, ammonia etc. To integrate with ASM models, the COD characterization will include the detailed analysis of pollutant composition. Finally, the conductivity, alkalinity and a selection of ions (K) will also be analyzed to complete the influent characterization. The measurement stations and dat*EAU*base system allows developing a methodology for influent monitoring, visualization and analysis in real time.

- **To develop a methodology of IG model performance evaluation and assessment**

The IG model performance will be evaluated in terms of result precision and modelling efforts, including prediction precision, statistical distribution of the time series and deviation between the model and observed

data, data needs, model complexity and so on. Using different criteria, influent variables generated by the new data-driven IG model, will be assessed and compared with a standard influent generation under static conditions (by using the average load) and an existing phenomenological model. A methodology will be developed to evaluate the model quality for the dynamic influent time series.

- **To achieve a flexible IG model: Definition, modelling, analysis and validation of IG models considering influent time series prediction, variability, and generalization by applying the proposed IG in different case studies**

The last objective of this research is aimed at developing a new influent generator model of the influent of a WRRF. The proposed influent generator will be able to describe the wastewater generation patterns from different sources (household, industrial, rainfall run-off) and will be able to handle different sewer processes (transport and transformation), while incorporating climate and catchment characteristics. IG models must also adequately capture the influence of the seasons, the time of the year and other sources of variability.

A set of IG-models will be modelled, improved, validated and completed in order to optimize the overall modelling process. The ultimate aim is to get a well-behaved IG-model with limited modelling efforts. The model outputs will include three dynamics: flowrate, overall pollutant (soluble, insoluble, organic, nutrients) concentration, and pollutant fractionation. The output will distinguish different time horizons and temporal resolutions in order to satisfy different user requirements. To validate its performance and versatility, the IG will be applied to catchments of different sizes and under different climate conditions.

Dissertation Outline and Contributions

This thesis has been structured in eight main chapters.

This first chapter is an introductory chapter and is followed by 7 more chapters in which an influent generator is developed for different uses. This introduction provides the problem statement, objectives and the outline of this dissertation.

Chapter 1 gives an overview introduction of wastewater influent characterization, analysis, generation and IG models development and relevant research. First, this chapter provides a critical literature review on the problem statement of influent data and motivation of IG building. Then, it lists different modelling methods for influent generation and highlights the advantages and disadvantages of each of them. Finally, data-driven IG modelling and data mining is reviewed to illustrate the importance of IG's for digitalization of the water field.

Chapter 2 provides the description of the case study and a summary of the methodologies that will be applied in the research, including the description of the data collected during the experimental work as well as the data provided by previous research works and applied in this PhD study.

Chapter 3 develops a methodology for raw urban wastewater influent characterization at source, its analysis and visualization by taking advantage of the case study of the pil*EAU*te facility installed at Université Laval. This WRRF can be considered to be treating the wastewater from a catchment without significant sewer system effect. The case study also allows studying the diurnal variation of the fractionation of COD, which has not been widely studied yet. This chapter also contributes to the originality of the PhD study by integrating conductivity, as well as K, alkalinity and VFA (volatile fatty acids) into an IG, thus providing relevant information for the modelling of physicochemical processes in water resource recovery facilities.

Chapter 4 presents a fast and user-friendly data-driven IG method based on artificial neural networks and polynomial regression. In this chapter, a weather-based IG is developed for the long-term influent generation for the case studies of Québec City (QC, Canada) and Bordeaux (France), which can be used for WRRF and controller design, according to different requirements for time horizon and data frequency. The model result includes wastewater quantity and routine water quality, as well as nutrient concentrations. Based on the model result, it is extended with a stochastic generator in order to better capture the pollutant concentration variability.

Chapter 5 is dedicated to improving the data-driven model of chapter 4. The Long Short-Term Memory (LSTM) architecture is adopted to improve the influent time series precision. A genetic algorithm (GA) is applied to calibrate the LSTM neural network in order to simultaneously optimize multiple IG performance criteria for influent generation.

Chapter 6 shows how the proposed methodology can be adapted and used for the typical climate of Quebec City, which is an example of North American climate with a 6 month period influenced by snow and a long period of snowmelt. The data-driven model is analuyzed and compared with the existing phenomenological Benchmark Simulation Model (BSM) Influent generator. The approach is successfully validated and tested for the case study during the snowmelt period.

Chapter 7 evaluates different influent generation scenarios based on different user demands, which allows confirming the proposed IG's flexibility and practical utility. The IG is applied to two alternative applications: increasing the data frequency of an IG and multi-timestep prediction of wastewater characteristics. This allows confirming the potential benefits of IG for digital water and digital twin development.

The conclusion of this PhD thesis summarizes this research and highlights its results, and provides the perspectives for future research that may tackle some of the unresolved issues.

This PhD research project aims at both an applied and a scientific impact.

From an application point of view, engineering consultants, and their clients, cities and utilities, will get an improved tool for the development of optimized management strategies of their wastewater system. By adding stochastic generation part, the IG-model will help the designer of a WWTP to get a good simulation of the proposed design under realistic temporal variability, especially in conditions for which no dataset is available. An appropriate IG may also make an important contribution to the development and optimisation of process control systems.

Scientifically, the project will improve the understanding of the mechanisms of wastewater production from the source all the way through the urban wastewater system until the entrance of the WWTP. The efforts related to the fractionation of COD will be beneficial to many ASM-based modelling studies. All in all, this research project contributes to the wastewater modelling and water digitalisation and relevant issues in the water engineering field.

Chapter 1. Literature review

This literature review addresses the overall topics of this dissertation. This chapiter will describe the IG background and developments so far. In order to avoid repetition, the literature related to the specific subjects of the thesis will be presented later in each chapter.

WRRF influents have an important impact on WRRF design, modelling and performance evaluation, etc. As input data, the influent generation is required to be dynamic, complete and able to represent the complex generation process. Usually, engineers have to make their decisions on the design of a WWTP under uncertainty, for example living with stoichiometric, biokinetics or influent uncertainty and uncertainty due to the unknown hydraulic behaviour of the plant or the non-predictability of sludge settling. The analysis of uncertainty of the influent is one of the critical elements when simulating a WWTP's performance under dynamic conditions and sensitivity to specific design assumptions (Belia et al., 2021).

The literature review on the IG is studied in five main sections, including municipal wastewater characteristics (flow and quality), the problem and motivation of IG, current IGs with their advantages and shortcomings, followed by background on artificial intelligence in the wastewater field, as well as the use of machine learning and especially influent data mining in water digitalization.

1.1 Characterization of municipal wastewater

This section will introduce different major characteristics of the influent.

- *Flow*

The flow of municipal wastewater is generated by different sources: on the one hand, the sanitary sewage from households and ICI (Industry, Commerce, Institutions), and on the other hand, aquifer infiltration, rain or snowmelt that drains off rooftops, lawns, roads, and other urban surfaces. The typical contributions differ with regard to wastewater source, country, region and climate etc. The long-term dynamics are influenced by weather variations (dry weather, wet weather) and seasonal variations. In resort areas, the seasonal variations are also affected by holiday periods. And melting snow influences the amount of water sent to the sewer system.

The wastewater production has been described in many studies. The quantity of domestic wastewater produced by different appliances has been characterized, and the per capita household wastewater discharge of two countries (England and Malta) was compared to explain the contributions to the different flowrate patterns from the different lifestyles, as shown in Figure 1-1 (Friedler and Butler, 1996).

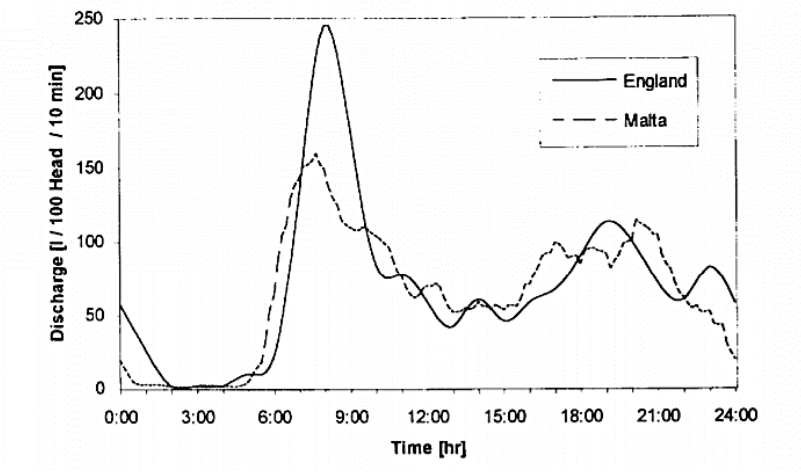


Figure 1-1 Characterization per capita domestic wastewater discharge (Butler et al., 1995)

Besides domestic water, the infiltrations are also important contributors to both combined and separate sewer systems. During rain events, the WWF of combined sewer systems is increased by direct run-off of rainfall, and the WWF of a sanitary sewer system increases due to rainfall-dependent infiltration and inflow (RDII), as shown in Figure 1-2. It also contributes to combined sewer flow (Vallabhaneni et al., 2007). In general, groundwater infiltration also contributes an amount of flow for both types of sewer system.

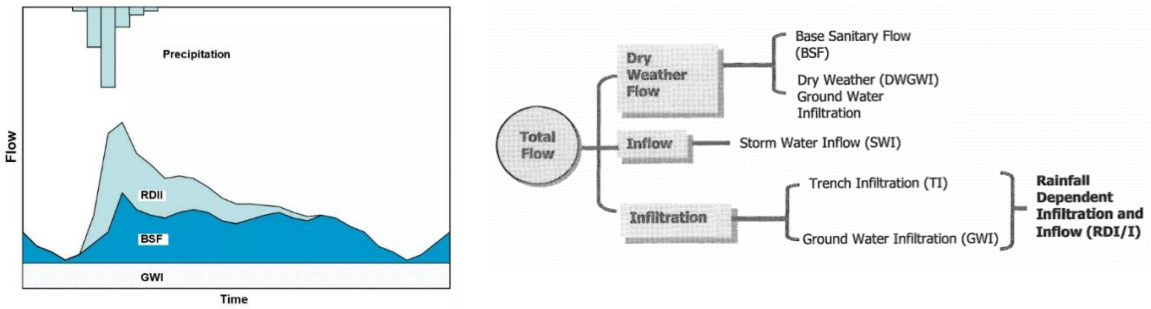


Figure 1-2 Three components of sanitary wastewater flow (Vallabhaneni et al., 2007) *and RDII components* (Dent et al., 2000)

The flow is the one of the most important variables during WRRF design, because it determines process requirements for the facilities such as the best treatment system configuration, the sizing, and the pump capacity etc. (e.g. bypassing of influent flow in wet weather conditions to maintain stable plant operation). Moreover, because the variation of flow also leads to pollutant load variation, this will influence the biological treatment process design, such as the sludge retention time, the food-to-microorganism ratio, not to mention process control system settings, etc.

- *Pollutant concentration and load*

A major source of pollutant into sewer networks derives from domestic sanitary wastewater. The characterization of this pollution is an important step to describe the wastewater. The typical pollutant concentration ranges of untreated domestic wastewater are shown in Table 1-1 and it can be concluded that the composition varies significantly.

Table 1-1 Typical composition of untreated domestic wastewater (Metcalf & Eddy et al., 2014)

contaminants	unit	Concentration strength		
		Low	medium	high
Solids, total (TS)	mg/L	390	720	1230
Suspended solids, total (TSS)	mg/L	120	210	400
Biochemical oxygen demand, (BOD_5)	mg/L	110	190	350
Total organic carbon (TOC)	mg/L	80	140	260
Chemical oxygen demand (COD)	mg/L	250	430	800
Nitrogen (total N)	mg/L	20	40	70
Ammonium (NH_4-N)	mg/L	12	25	45
Phosphorus (total P)	mg/L	4	7	12
Bicarbonate (HCO3)	mg/L	50-100		
Potassium (K)	mg/L	7-15		

Different studies have been conducted to measure and characterize wastewater composition. The at-source domestic wastewater quality and quantity have been studied by (Almeida et al., 1999), producing relative daily patterns (pollutographs) for COD_t, PO_4, TSS, NH_3 and NO_3 based on a household appliance use survey. Further, Friedler and Butler (1996) have studied the uncertainty in the quality of domestic wastewater, according to the time, people, appliances and pollutant loads for different pollutants.

Le et al. (2017) studied daily wastewater pollutant dynamics (COD, NH_4-N, etc.) according to the catchment population structure. A comprehensive analysis of COD constituent fractions including typical VFA (volatile fatty acids) composition distributions was also studied by Rössle and Pretorius (2001). Briefly summarized the aforementioned studies found: the WC is a major contributor to domestic wastewater; the daily flow pattern shows two peaks (morning and evening) related to urban activities; the daily patterns of flow and loads follow similar patterns, but the flow and concentration are quite independent.

Besides the major pollutants, the diurnal patterns of micropollutants (methyl-parabens, oxybenzone, etc.) have been studied in domestic wastewater (Alfiya et al., 2018). Eriksson et al. (2009) focused on greywater pollution and loading regarding both macropollutant (COD, TOC, conductivity, nitrate/nitrite and ammonia, ortho-phosphate) and micropollutant concentrations.

It is well known that the pollutant fractions provide varying biochemical substrates for different species of microorganisms for their biological processes. Consequently, fractionation is an important issue to be studied for ASM modelling, as shown in Figure 1-3.

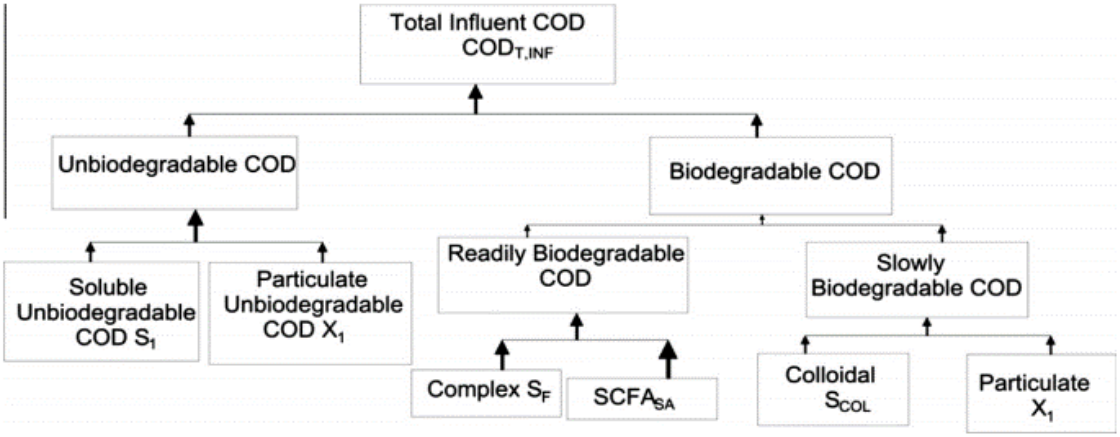
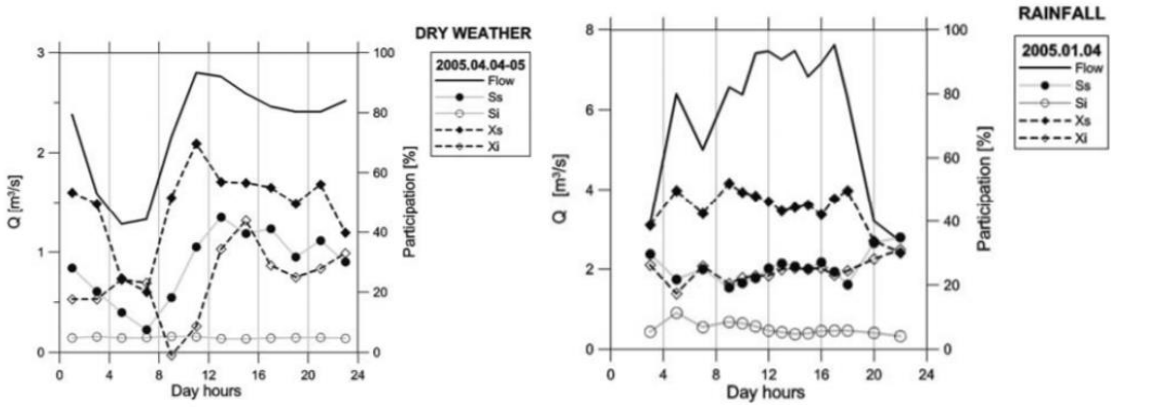


Figure 1-3 Fractions of COD in wastewater (Metcalf & Eddy et al., 2014)

Different fraction characterization studies have been performed in the past: Ginestet et al. (2002) analyzed the biodegradability of raw wastewater as well as different physico-chemical fractions. The COD fraction as well as particularly the typical VFA (volatile fatty acids) composition distribution have been studied as well (Rössle and Pretorius, 2001). Murat Hocaoglu et al. (2010) characterized the biodegradation of domestic wastewater in distinguishing black and grey water fractions and Dixon et al. (2000) have measured and modelled the quality change for untreated grey water focusing on COD, DO and BOD.

Studies also show that combined and separate sewer system have different behaviour in terms of urban runoff pollution load (Brombach et al., 2005; Nasrin et al., 2017). The variability of the fractions of COD and TKN of combined sewage under different conditions has been analyzed for both dry weather and wet weather conditions (Zawilski and Brzezinska, 2009). The range of COD fractions is presented in Figure 1-4, reflecting their variation at different times of the day, as well as in different types of weather. In wet weather, stormwater introduces a considerable inert COD load, while the readily biodegradable COD fraction is strongly diluted. Despite these clear and important phenomena, the dynamics of the fractionation is not included in existing IG models.



COD fractions		S_s		S_i		X_s		X_i	
Type of weather		Range* [%]	Mean ** [%]	Range* [%]	Mean** [%]	Range * [%]	Mean** [%]	Range * [%]	Mean** [%]
Dry weather		6-45	**22.0**	4-9	**5.5**	20-70	**56.0**	4-69	**16.5**
Wet weather	Rainfalls	1-35	**21.3**	3-11	**5.8**	13-79	**49.5**	7-79	**23.3**
	Storms	8-31	**14.0**	5-9	**7.0**	16-70	**47.6**	11-69	**31.2**
	Snowmelt	9-32	**21.1**	7-13	**8.0**	8-93	**26.6**	6-52	**44.3**

* the range for instantaneous samples
** mean values for composite samples weighted by flow

Figure 1-4 Variation and percentage contribution of fractions in total COD of raw dry and wet weather wastewater (Zawilski and Brzezinska, 2009)

On the other hand, the presence of industrial effluents can cause the wastewater to be significantly different from purely domestic effluent. Mhlanga and Brouckaert (2013) focused on the characterization of the carbonaceous fraction of a mixture of industrial and domestic wastewater. By comparing different WWTPs, which receive various industrial and domestic wastewaters, the results confirmed that industrial effluents can remarkably affect the wastewater quality behaviours.

1.2 Problem statement of influent data and motivation of IG

With the development of ASMs (Henze et al., 2000; Vrecko et al., 2006), nowadays, more and more models are used to simulate and evaluate WRRFs for design, operation upgrade, and control strategy evaluation (Regmi et al., 2019). It is known that the influent wastewater composition has a significant impact on WRRF operation and performance. However, because of a lack of adequate input datasets, the full potential of the models remains untapped.

For example, for WRRF design, engineers usually make initial sizing by using design guidelines based on average loads and safety factors (Talebizadeh, 2015). As another example, ammonia forecasting is expected to

improve controller performance, e.g. for ammonia-based aeration control (ABAC) (Newhart et al., 2020), but a high-quality and real-time ammonia prediction is not always available. A critical review of analysing and generating influent data for WWTP modelling has been written by Martin and Vanrolleghem (2014), by classifying different situations of influent data issues from engineering practice.

Even though the influent data is widely needed, and the importance of the data has been demonstrated, issues remain including the following points:

- Lack of data collection

Wastewater quality monitoring is increasingly applied using two approaches: supervisory control & data acquisition (SCADA) systems with modern sensors that aim at high frequency measurements, and 'traditional' grab or composite sampling and subsequent laboratory analysis, which offers short-term and low frequency data, but is easier and more reliable to perform (Schilperoort, 2011). However, the sensor data quality is highly depending on the instrumentation, and it can be influenced by sensor failure, maintenance issues and the complexity of installation due to poor accessibility in the studied system, etc. (Grievson, 2020). Because of the harsh conditions of raw wastewater, it is often difficult to fulfill the frequent maintenance needed for the in-site sensors. The sticky materials of raw wastewater and the heavy deposit of pollutants make their maintenance cost considerable. Given the difficulty and the high cost of data collection in the raw wastewater, the influent data is not always available. A potential solution is to use low cost sensors to generate some useful information under a limited budget (Montserrat et al., 2015).

- Data quality and reliability

Data quality is crucial for modelling, operation and control. The uncertainty of online sensor data (systematic errors or random errors) has been studied in order to quantify the data reliability and evaluate the data quality (Rieger et al., 2005). Simulation results showed that by increasing sensor reliability, system performance can be improved (Hug and Maurer, 2012). However, one of the studies also showed that the sensor performance ranking is depending on the pollutant considered and weather conditions (Lepot et al., 2013). Therefore, improving data accuracy and ensuring the data quality is a key problem for IG modelling.

- Insufficient data collection in terms of frequency

The impact of input data frequency has been studied and the results demonstrate that a WRRF model will lead to different simulation results and accuracy under different input data frequencies (Cierkens et al., 2012). However, in WRRF operation, even though the flowrate data is measured and collected at high frequency, the influent quality data are mostly collected only daily, weekly, or sometims twice per week, depending on the size

of the WRRF. For this reason, an influent generator should be built in a way that a limited data collection can be extended to a higher frequency (Devisscher et al., 2006).

- Comparability between common quality information into ASM family components:

It is important to translate the common wastewater quality information into ASM family components, in order to overcome the barriers between the monitored quality data and the ASM family's modelling approach. Many critical reviews demonstrate the great diversity of techniques and technologies available in both commercial and research laboratories (traditional offline methods as well as online and real time methods) for monitoring the organic or biodegradable fractions of wastewater (Bourgeois et al., 2001; Vanrolleghem and Lee, 2003). Despite the fact that the online short-term BOD (BODst) can be automatically measured in real time, however, the IG models do not in general include this highly dynamic information.

- Database management and data mining

Correct and advanced database management ensures that one can avoid the development of data graveyards and achieve true data mines (Aguado et al., 2021; Plana et al., 2018). The importance of the influent data is crucial for improving current wastewater treatment operation. Moreover, despite the availability of vast amounts of data (data rich situation), they often remain information-poor, and more specific methods are required, which are better adapted to the wastewater field, in order to transfer the data into knowledge for improving treatment (Corominas et al., 2018; Therrien et al., 2020). Some studies have already explored the possibility of mining information from low-cost sensors, unmaintained sensors or flawed sensors to extract reliable and useful water quality data (Schneider et al., 2019; Yang et al., 2020). Soft sensing is becoming a feasible method for collecting data which are unmeasurable or non accurate by the physical sensors (Kadlec et al., 2009; Schneider et al., 2020; Souza et al., 2016). In addition, data-driven models for cost-effective soft sensing based on artificial intelligence are gradually being applied for online monitoring (Dürrenmatt and Gujer, 2012; Shokry et al., 2018). Such expanded data-base provides both challenges and opportunities for IG model development and applications.

In a word, despite the fact that the importance of influent data is widely acknowledged, no efficient and general influent generation model is widely accepted yet. Moreover, since the concept of IG is often not explicitly defined, the motivation to build them is also lacking. The existing IGs are either incomplete or not precise, and are difficult to calibrate. These facts also hamper the IG's developments. Therefore, to solve the problem of influent data mentioned above, a reliable influent generation model is essential to provide input for WRRF modelling (Gernaey et al., 2011; Martin and Vanrolleghem, 2014).

A good and practical IG model should contribute to a WRRF model by being simple to use. Firstly, the model should be able to generate important variables required by engineering standards in the industrial domain. For example, it will determine the hydraulic loading and organic loadings, which are key parameters for a WRRF's process design (Metcalf & Eddy et al., 2014). Moreover, an IG model should also be versatile, which means it should be able to generate WWTP influent disturbance scenarios by providing various time scales and frequencies according to the user's needs. For example, the simulation of operational performance may require a higher time resolution generation than an IG time series needed for a WWTP design. Therefore, IGs can be useful for many diverse applications:

- Faster WRRF design and treatment modelling: Increasing urbanization and population growth generate growing amounts of wastewater, while the environmental protection objectives become more severe, which challenges wastewater utilities. Therefore, a good IG provides a fast and accurate estimate of influent loads for WRRF designers. More importantly, the influent data is one of the main sources of uncertainty in the modelling process, so improving the influent model will help to better assess and evaluate WRRF models (De Keyser et al., 2010). As another example, it has been studied that the disturbance of the influent fractions can affect the production of biogas (CH_4 and CO_2) and the kinetics in WRRF(Solon et al., 2015).

- Treatment performance evaluation and enhancement: An IG is able to provide a credible input profile to anticipate the treatment result under different operating conditions. This information supports the operators in running their treatment processes, with regard to the costs and benefits of advanced control for WWTP performance. Moreover, an IG model can help the transition from a WWTP to a WRRF, such as through the evaluation of its carbon or water footprint (Gómez-Llanos et al., 2020). The identification of the quality and quantity of wastewater at the inlet of a WWTP enables the improvement of energy (heat, carbon) and nutrient (N, P, sludge) recovery and allows to predict the potential products that can be extracted from the wastewater (Solon et al., 2019).

- Database management and gap filling: Data observation and collection become more and more important in wastewater management. An IG enables quality evaluation of a dataset collected from online measurements. It can do this in different ways, e.g. identifying errors from clogged sensors, detecting extreme measurement values, etc. An IG can also complete missing data (gap filling) due to the removal of invalid data (Patry et al., 2020) and interpolate a low frequency series into a denser series. By data analysis and data mining, an IG model may enable to better understand the internal correlations between different pollutants.

- Control strategy optimization and real time control (RTC): The WWRF is a dynamic system with large disturbances and uncertainty, therefore, the identification of influent uncertainty and dynamics will enhance control strategy optimization, providing the important benefit of energy saving (Qiao and Zhang, 2018). As another example, RTC is essentially active under dynamic loading conditions (Borsányi et al., 2008). RTC stabilizes and optimizes the treatment processes in WWTPs. The IG model can be used to provide reliable predictions for inflow control, especially in case of combined sewage treatment in order to maximize the efficiency of currently available capacities (Seggelke et al., 2013).

Compared to conventional techniques (such as traditional datasets collected experimentally at the plant), creating an IG model does not need excessive measurement campaigns, which are time-consuming, labour-intensive and expensive. An IG can provide a reliable estimate of the influent description with relatively few measurements. Moreover, an IG model can provide a more complete description of the dynamic disturbances, at different spatial and temporal resolutions.

1.3 State of the art of IG development

Different types and versions of IG models already exist in literature.

Based on the availability of process knowledge and incorporation of data, the current IG models can be categorized in two major categories: data-driven models and phenomenological (or mechanistic) models. The data-driven model is also known as black box model, while the mechanistic model is a grey model or white model depending on the availability of process knowledge and incorporation of data. There are advantages and disadvantages in both types of models (Price and Vojinovic, 2011). Compared with a data-driven model, a phenomenological model contains more details of the influent generation process, which leads to an explicit result and a good estimation beyond the calibration range. However, it may usually need more modelling efforts for the calibration.

1.3.1 Data-driven models

Data-driven models depend entirely on the provided dataset, and varying degrees of model complexity have been studied, such as harmonic function models, statistical models, empirical models, and machine learning approaches and so on, according to different complexities of the models.

- Harmonic function:

The Fourier-based harmonic function models are generally used to describe the wastewater patterns under dry weather conditions, thus providing a basic foundation for a lot of different IGs.

Langergraber et al., (2008) developed a simple, reliable generation of diurnal variations for dry weather influent data, whose mathematical formulation is based on a 2nd-order Fourier series, with the following equations:

$$Q_{\mathrm{inf}(t)} = Q_{inf} = Const$$

$$Q_u(t) = Q_u + a_1 \sin(\omega) + a_2 \cos(\omega) + a_3 \sin(2\omega) + a_4 \cos(2\omega) \qquad 1\text{-}1$$

$$Q_d(t) = Q_d + b_1 \sin(\omega) + b_2 \cos(\omega) + b_3 \sin(2\omega) + b_4 \cos(2\omega)$$

where ω=2π/T and T= 1 day, a and b are constant parameters and Q_{inf} is the infiltration rate, Q_u is the mean urine flowrate, and Q_d is the domestic wastewater flowrate without urine.

Based on the same principle, Mannina et al., (2011) generated an input variable Y with harmonics multiplied by daily average values:

$$Y(t) = \mu \cdot (1 - (\beta_1 \cdot \sin(\omega_1 \cdot t + \Phi_1) + \beta_2 \cdot \sin(\omega_2 \cdot t + \Phi_2) + \beta_3 \cdot \sin(\omega_3 \cdot t + \Phi_3))) \qquad 1\text{-}2$$

where β, ω, ϕ are the series' parameters, t is time and μ is the average value.

These models are very easy to use and powerful to obtain hourly values given average daily data. However, the harmonics function is only based on diurnal variation, and does not include rain events. Therefore, this method can only be used for the description of the dry weather flow (DWF) situation, not for describing wet weather flow (WWF) condition.

- Empirical model:

The data-driven model can also be based on an empirical formula, with only very empirical knowledge support. Unlike the harmonic function, empirical models often introduce some relevant variables to make the model more meaningful.

One of the more traditional approaches to estimate water quantity is to provide the average dry weather flowrate and measure the pollutant concentration by assuming a daily flow and pollutant production per capita. One of the traditional civil engineering methods for getting the design flow capacity of a WWTP uses the per capita dry-weather assumption, and a peak factor (PF) to calculate the peak flow (Metcalf & Eddy et al., 2014). The peak dry weather flow is derived from the following formula:

$$PF = \frac{\text{peak flowrate (hourly, daily)}}{\text{average long} - \text{term flowrate}} \qquad 1\text{-}3$$

The statistical analysis of flowrates, constituent concentrations and mass loading are based on measurement series, and probabilities (Metcalf & Eddy et al., 2014). However, relying on the peaking factors indicated in the guidelines, may result in oversizing some treatment units, which should be verified in the engineering design (Corominas et al., 2011).

Langeveld et al., (2017) also used an empirical sewer water quality model based on measured hydraulic dynamics. The model distinguishes small, medium and large storm events with thresholds, and providesthe pollutant concentrations for different patterns (onset of WWF, dilution and recovery) during storm events calculated from the DWF concentrations at different times of the day.

- Statistical and probability theory:

The typical black box methodology of building an IG is based on statistical theory. The statistical method has been widely used for modelling the correlation between overflow volume with pollutant loads and rainfall data, according to different return periods (Veldkamp and Wiggers, 1997). For the WRRF influent, Ahnert et al. (2016) developed a statistical method for the generation of a continuous time series of influent quality, based on the Weibull-distribution. The concentration probability depended on the flowrate, as shown in Figure 1-5. Weibull-distributed random data were fitted to the available routine data, such that the resulting distribution of influent quality data showed identical statistical characteristics. Missing concentrations could be generated from the distribution function. This method was based on incomplete routine data, estimated the shape and scale parameters of the Weibull distribution, and the resulting IG was applied to fill the parts with missing due to a lack of measurements. A similar methodology had already been used for establishing the correlation between pollutant and flow, and the inherent uncertainty of influent characteristics was used as input to WWTP models to analyse the probability of exceeding the effluent limits (Rousseau et al., 2001).The Bayesian approach has

also been applied to estimate the CSO volumes during storm, as it impacts the influent generation (Hutagalung, 1967).

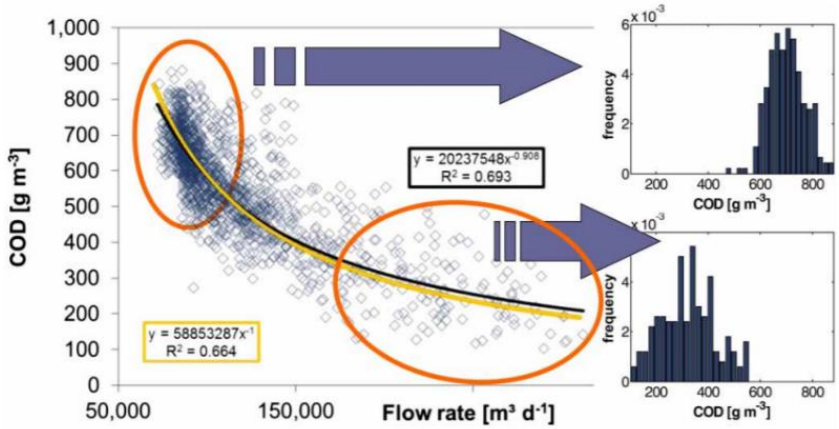


Figure 1-5 Influent data (blue dots) with two different regression lines and histograms of data (Ahnert et al., 2016)

Recently, other black box methods such as artificial neural networks, or Gaussian processes are also being studied to predict wastewater behaviour. For example, Troutman et al., (2017) developed an automated toolchain based on a data-driven methodology to predict the dynamics in combined sewer system flow. The model can be adjusted by 'learning' the error between model output and input data. This process is depicted as the toolchain shown in Figure 1-6. The dry weather flow is filtered by a Butterworth low bandpass filter and modelled with a Gaussian process (black box method). Wet weather flow is derived from rainfall intensity data, and the black box model is identified by maximizing the likelihood.

Although the pure statistical model is useful for influent generation, it heavily depends on the quality of the data set, and it is difficult to explain the generation mechanism.

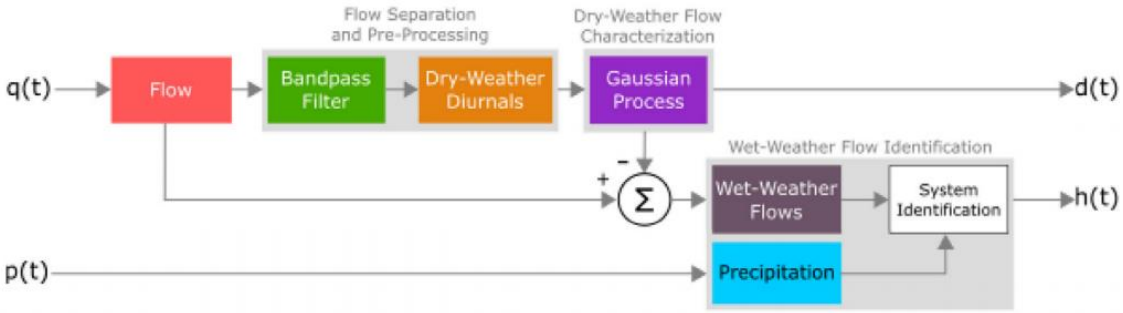


Figure 1-6 Automated toolchain for flow rate prediction in combined sewer systems (Troutman et al., 2017)

1.3.2 Phenomenological and mechanistic models

Mechanistic models can be divided into different types, according to the complexity of the sub-model involved, going from simple grey models, phenomenological models to complex mechanistic models with detailed catchment and sewer system description (Benedetti et al., 2013).

Different processes occurring at catchment level influence the wastewater arriving at the inlet of a WWTP: rainfall events, industrial activities, seasonal effects, etc. Compared to the data-driven method, the phenomenological model partially integrates important generation processes to build the model. This type of model is widely used for influent generation.

- Simple grey box model

The simple grey box model is one of the simplest phenomenological models, enabling the identification of a model for a dynamical system based on simple physical assumptions with statistical tools. Phenomenological models include multiple phenomena of a sewer system, from source to influent. Different sub-models are integrated, and more detailed dynamics are captured.

The simplest grey box model is based on introducing a rainfall event and predicting the run-off flowrate. For example, Carstensen et al., (1998) created a flowrate model consisting of two components to predict the hydraulic load: the diurnal variation approximated by a 2nd-order Fourier function, and a rainfall-runoff transfer function used to represent the wet weather flow, based on autoregressive moving average (ARMA) transfer function modelling(Novotny and Zheng, 1989):

$$Q_{runoff,t} = \frac{\omega(B) \cdot B^{10}}{1 - \varphi B} \cdot I_{rain,t} \qquad 1\text{-}4$$

where the backshift operator of the time series is expressed by B ($BX_t = X_{t-1}$).

A grey box model can also be applied to describe the first flush during a rain event (Bechmann et al., 1999). The model describes the diurnal profile, which depicts the pollution that enters the sewer system, the gradual deposit of solids in dry weather and the flush of sediments in wet weather:

$$\hat{y}(t) = a_0 + \sum_{k=1}^{n}\left(a_k sin\left(\frac{2\Pi kt}{24h}\right) + b_k cos\left(\frac{2\Pi kt}{24h}\right)\right)\underbrace{-a(\hat{x} - \bar{x}) - b(\hat{Q} - \bar{Q})}_{sewer\ process\ part} \qquad 1\text{-}5$$

The first part i.e., the harmonic term is the model for the diurnal variation and the second part refers to the deposit and flush processes. $\hat{y}(t)$ is the pollution flux predicted by the model at time t, $\hat{x}$ denotes the deposition of pollutant in the sewer at time t, $\bar{x}$ and $\bar{Q}$ are the mean values; a and b are unknown parameters, which are assumed to be negative. Hence, for the sewer process part, a flow larger than the average during the period decreases the accumulation rate of the pollutants and a lower one increases it. Another grey box example combines the harmonic function to represent the diurnal variation under DWF and the drift term and diffusion term to represent the sewer behaviour for WWF (Breinholt et al., 2011). Similarly based on the harmonic function, the online ammonium concentration forecasting can be modelled based on a daily profile and online sensors (Vezzaro et al., 2020). Another example of IG is based on hydrological modelling, and was successfully used for a case study, by including infiltration and conceptual CSO modelling, to reproduce the flow in DW and the peak in WW (Coutu et al., 2012).

- Phenomenological model

The mechanistic models often include different sub-models, where the sewer system modelling is an important part when trying to describe the transformations occurring between the wastewater source and the inlet of the WWTP. The sewer system is usually modelled by a configuration of numerous variable storage volumes connected with pipes of different dimensions (van Luijtelaar and Rebergen, 1997). The pipe network of a collection system usually resembles the branches of a tree (2-D dendritic structure).

Saagi (2017) compared different sewer hydraulic models and transport process models. The hydraulic behaviour in the sewer network can be described with either detailed hydrodynamic models (the Saint-Venant equation) or with simplified conceptual modelling tools (linear reservoir models, multi-linear reservoir models, nonlinear reservoir models, etc.). The hydrodynamic models pose challenges in terms of model complexity, which can be overcome by the use of the conceptual reservoir models, such as:

$$\frac{dh_2}{dt} = \frac{1}{A_2} * (Q_{in} - Q_{out}) \qquad 1\text{-}6$$

$$Q_{out} = C * {h_2}^n$$

where h_2 presents the water level in the tank and Q_{in} is estimated by the model user, n and C are model parameters, and A_2 is the water surface of the reservoir. These parameters are often estimated on the basis of high frequency data or a detailed model (Meirlaen et al., 2001).

Actually, this simplified sewer system model has already been successfully used in many phenomenological IG, as shown in Figure 1-7 (Gernaey et al., 2005).

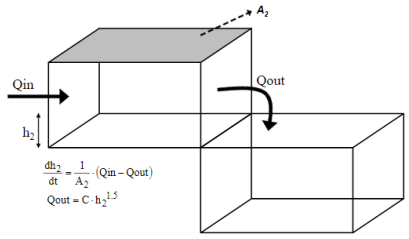


Figure 1-7 Principle of the variable volume tank model for sewer system model (Gernaey et al., 2005)

Different sources, i.e., sanitary sewage, runoff, in-sewer stock etc., will contribute to the quality change in the combined sewer system, especially during wet weather (Michelbach, 1995). Physical, chemical or biological transformations may take place during the transport process within the sewer (Nielsen et al., 1992), which includes different mechanisms: advection, diffusion, accumulation or washout, settling and resuspension (Michelbach, 1995), sediment erosion and also biological transformation, depending on the characteristics of the sewer (Rammal, 2016). As an example, the TSS dynamic behaviour during rain events can be described by including the sedimentation factor (Rossi et al., 2005). The transport and degradation processes have been modelled and calibrated for pharmaceuticals (Coutu et al., 2016) and other micropollutants (Vezzaro et al., 2014, 2010).

Different mechanistic descriptions can be integrated in the IG. However, these complex nonlinear processes require more modelling efforts, which should also be considered during the IG modelling.

Gernaey et al. (2005) developed a phenomenological model with limited complexity, by focusing on flowrate scenarios and neglecting the complex deterministic model of the urban drainage system. Flores-Alsina et al. (2012) extended this model. Figure 1-8 depicts the individual model blocks of this model. To simplify the model blocks, the production models were based on per capita equivalents, defined by the user, and the sub-model blocks were described using mass balances. As a result, the influent generator can feature diurnal phenomena, the weekend effect, seasonal phenomena and holiday periods as well as rain events.

Flores-Alsina et al. (2014) applied this phenomenological dynamic influent pollutant disturbance scenario generator (DIPSDG) (Gernaey et al., 2011) to two case studies with full-scale data. These applications demonstrate the usefulness of this phenomenological model.

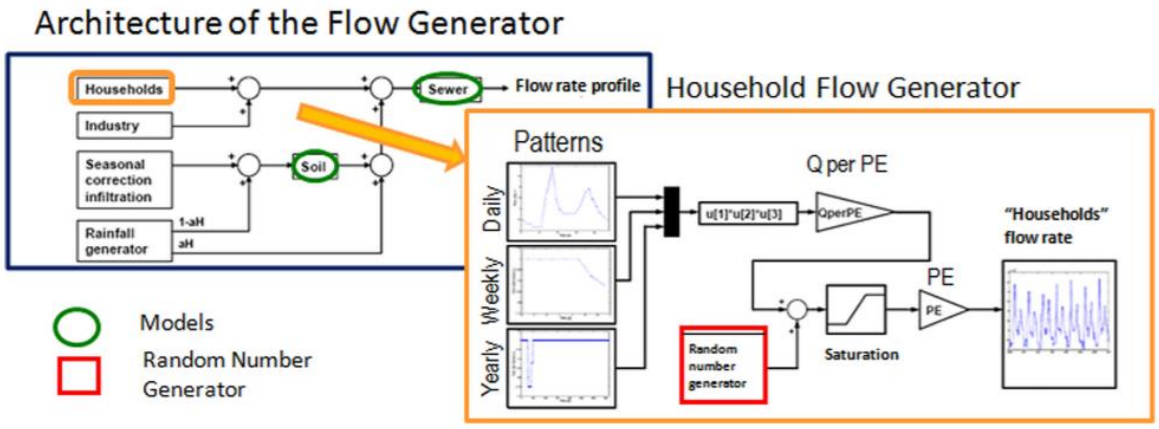


Figure 1-8 Architecture of the flow generator in the influent disturbance model of (Gernaey et al., 2005)*, further developed by* (Flores-Alsina et al., 2012a)

Based on similar phenomenological model principles, Béraud et al., (2007) constructed a simplified model for the case study of a full scale WWTP based on daily average measurements in order to optimize online data information.

Talebizadeh et al., (2015) proposed a hybrid of statistical and conceptual modelling techniques for synthetic generation of influent time series, schematically shown in Figure 1-9. In this model, the statistical part includes DWF and WWF generation. A multivariate auto-regression model is applied for DWF generation taking into account the daily periodic variation, auto-correlation, and cross-correlation. A weather generator is trained on rainfall time series to calculate the sequence of dry and wet weather days by a Markov Chain. The phenomenological modelling part is the model created to calculate the wet weather influents, using the CITYDRAIN catchment model as the conceptual model to estimate effective rainfall during WWF.

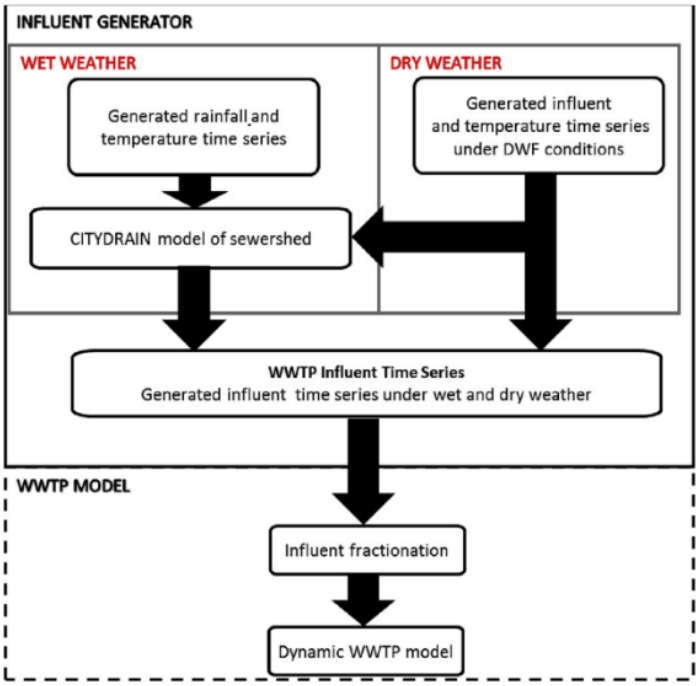


Figure 1-9 Schematic of the influent generator of Talebizadeh (2015)

- Integrated models and IG

A mechanistic IG model can represent the output of a sewer system model in the context of a mechanistic integrated urban wastewater model (Mitchell et al., 2007). In fact, the mechanistic sewer model itself would be the most detailed and complex model for influent generation. Different sewer system modelling tools such as Mike Urban, SWMM, Infoworks, and so on, were compared and many applications have been reviewed and classified (Bach et al., 2014; Elliott and Trowsdale, 2007).

There is a vast amount of sewer modelling research based on commercial sewer models, to simulate the hydraulics and wastewater quality in a complete sewer system. For instance, SWMM (Storm water management model) has been applied to simulate the hydraulics and turbidity under wet weather conditions (García et al., 2017; Rossman, 2015). These sewer system modelling works can inspire IG model developments. For instance, the dynamics of micropollutants in the sewer can be modelled with the GEONIS Sewage software (GIS based) by calculating the hydraulics together with advection-dispersion, which represents the sewer transport process (Ort, 2006). By including biochemical reactions in the sewer and simulate water quality, an sewer conceptual model was calibrated and optimized based on SWMM simulations, allowing for long-term and large scale modelling (Guo et al., 2019). Detailed hydrodynamic modelling for sewer management regarding the issues of CSO, and especially regarding the pollutant concentration dynamics, which can be achieved by using the

SAMBA software and MOUSE (DHI Water & Environment, Denmark) with Advection-Dispersion extended modules (Schlütter and Mark, 2003).

An integrated urban wastewater model library can be found in the software WEST (DHI Water & Environment, Denmark), in which the urban catchment and sewer system are conceptually modelled (Ledergerber et al., 2019; Vanhooren et al., 2003a). Another phenomenological influent generator example focused both on generic pollutants such as COD, but also on priority pollutants such as hydrocarbons, by using a typical emission pattern, and combining it with the integrated catchment model in the WEST modelling and simulation software (De Keyser et al., 2010). Besides, the fate of organic micropollutants has been also studied by using the conceptual sewer models based on WEST (Delli Compagni et al., 2019; Vezzaro et al., 2014).

- Stochastic aspects of IG

The influent generator model can also be stochastic. Compared to deterministic models which give exactly the same output every simulation run, the stochastic model improves realism by adding a stochastic element (noise term), which can describe the random behaviour of, for instance, flow (Carstensen et al., 1998). The noise model can reduce the correlation between different pollutants and produce some additional randomness during influent generation.

Different studies have used white noise to create a stochastic model. For example, Rodriguez et al. (2013) used white noise, i.e. a normal distribution with zero mean and unit standard deviation in a 3rd order Fourier series model, to generate dry weather flows and loads. Similarly, Breinholt et al. (2011) used a normal distribution with zero mean and unit variance to distinguish observation errors from input and model structural errors. Bailey et al. (2018) have also used a stochastic model to identify a diurnal household discharge model by using SIMDEUM® (https://www.watershare.eu/tool/water-use-info/), which is a software tool for water demand modelling. The changes in the sewer system to flow, nutrients and temperature, are simulated by the model based on a combination of the SIMDEUM® and InfoWorks®WS/ICM packages, under different water conservation scenarios (Bailey et al., 2020). Another stochastic model was used to represent the dynamic fluctuations of specific pollutants (Ort et al., 2005). Taking advantage of a stochastic model, probabilistic modelling can be performed, and the uncertainty on its parameters can be understood for sewer management (Thorndahl, 2009; Bartosz et al., 2018).

- Summary

It is evident that detailed sewer modelling can generate a complete wastewater profile, but a parsimonious approach is also expected to be useful within the scope of IG applications, as it allows minimizing the difficulty

of modelling, data collection and calibration efforts. In general, a phenomenological model is more detailed than a simple grey box model, but more concise than a complex mechanistic model, that may include such phenomena as the seasonal effect, rainfall events, etc.

1.4 Data mining and machine learning models

Nowadays, wastewater databases are becoming larger, in terms of the 5V's: volume, velocity, variety, veracity and value (Demchenko et al., 2013). Data mining, also known as knowledge discovery in data, is the process of exploring and harvesting the information, extracting and discovering patterns, and understanding the data and their interpretation (Fayyad et al., 1996). Machine learning (ML), as an important branch of artificial intelligence (AI), focuses on the development of algorithms and the use of data to imitate the way that humans learn. As a modelling tool, an overview of the strengths, weaknesses, opportunities, and threats (SWOT) of ML methods has been studied and critically discussed (Dobbelaere et al., 2021).

Both data mining and machine learning are not an emerging technology in the wastewater field. Even though the utilization of big data is to some extent limited by social, ethical or public authority considerations, vast amounts of data collection and monitoring accelerate the developments in data mining and AI applications in the wastewater field (Manny et al., 2021). In the meanwhile, the value of metadata is drawing attention in data mining, such as for annotating sensor signals, indicating the data quality and so on (Aguado et al., 2021; Plana et al., 2018; Therrien et al., 2020). In addition, since machine learning and deep learning have started to be applied in the water domain, many approaches are increasingly developed for data monitoring, unstaffed WWTP operation, reinforcement learning for traditional and increasingly advanced process control, anomaly detection, and short-term prediction of certain process variables (Eerikäinen et al., 2020; Gopakumar et al., 2018; Muharemi et al., 2019; Mullapudi et al., 2020; Russo et al., 2021; Schneider et al., 2020; Shokry et al., 2018). Moreover, AI is playing an increasingly important role in strategic foresight to achieve the water-related Sustainable Development Goals (SDG) (Mehmood et al., 2020).

In the concrete, many data-driven modelling techniques are applied to support wastewater treatment processes, detection of faulty values in real time (Russo et al., 2020), and provide valuable additional information, such as by soft sensor building (Dürrenmatt, 2011). AI-based decision making has also been applied to wastewater treatment process (Hadjimichael et al., 2016; Han et al., 2020). Different critical reviews of the techniques used in the wastewater field were already published (Corominas et al., 2018; Newhart et al., 2019) i.e. principal component analysis (PCA), clustering, artificial neural networks, or support vector machines (SVM), self-organizing maps (SOM), decision trees, etc. On the other hand, the data-driven model developments can

enhance mechanistic models through hybrid modelling, i.e. merging data-driven and mechanistic models, resulting in a more powerful and flexible approach in the wastewater field (Lee et al., 2005, 2002; Quiza et al., 2012).

Recently, machine learning is increasingly being studied for WWTP influent generation, e.g. for the prediction of flow rates over a short time horizon, based on an artificial neural network (ANN) (El-Din and Smith, 2002), or to create a WWTP influent flow model that also considers spatial features, by using a multi-layer perceptron neural network (Wei and Kusiak, 2015). The ANN can not only be used for flow but also for pollutant concentration predictions (Aminabad et al., 2013).

Other predictive model approaches have been successfully developed and applied: for instance, the predictive model for influent flowrate based on the autoregressive integrated moving average (ARIMA) (Zhang et al., 2019), the estimation of water quality characteristics by using artificial neural networks and canonical correlation analysis (CCA) with PCA-based data preparation (Khalil et al., 2019, 2011), an autoregression model (AR) using wavelet analysis and power spectral density (PSD) analysis for ammonia concentration prediction (Ma et al., 2014), a multiple linear regression or a nonlinear autoregressive exogenous model (NARX) for flow forecasting (Banihabib et al., 2019), and the prediction of non-deposition sediment (suspended solids) transport in the sewer based on the random forest approach (Montes et al., 2021). The data-driven model can also be hybrid, such as by combining autoregressive moving average models (ARMA) and a nonlinear outlier robust extreme learning machine (ORELM) to predict COD, BOD and TSS concentrations for a full-scale treatment plant (Lotfi et al., 2019). ARMA models have also been combined with vector auto-regression (VAR) for COD load forecasting (Man et al., 2019). More detailed information on the methodologies will be presented in the introduction of the chapters discussing the specific contributions of this PhD thesis.

Chapter 2. Case study and methodology overview

This chapter gives a very general description for case study and methodology overview of the thesis. Various data sets are required and tested during the IG creation. Datasets include those collected by dedicated measurement campaigns, as well as those available from previous studies of the model*EAU* research team.

2.1 Case study introduction

Firstly, the influent characterization is studied from a case study of the pil*EAU*te research facility at Université Laval, which considers a system with a small catchment as the source, and with a short sewer system. The pump station feeds the pil*EAU*te system, with a pump capacity of 10m3/h. The combined sewer system between the Lacerte residence and the pil*EAU*te WWTP is sufficiently short, so that it can be considered that no transformation occurs during transport, see details description in next chapiter. This plant consists of a pumping station, a storage tank, a primary settler and two parallel biological treatment lines followed by secondary clarifiers. Each biological line has five reactors: the first two anoxic and the last three aerobic.

Figure 2-1 Localization of pilEAUte treatment plant

Then, the data-driven model has been built and improved based on two databases, including Quebec City (Canada) and Bordeaux (France). Quebec City East WWTP has capacity of around 270 000 PE (person equivalents), consisting of screens, grit chamber, chemically enhanced primary treatment followed by biofilters,

secondary clarification and disinfection. The WWTP receives the wastewater from a combined sewer system, as shown in Figure 2-2. Different hydraulic structures intervene in the sewer system and act as influent disturbances (retention tanks, pumps, valves, etc.). The extended winter period leads to a considerable snowmelt and groundwater level change in spring, which is a special phenomenon to be considered in the IG model. The influent data available at the WWTP include flowrate, weather, TSS, COD, BOD_5, phosphate ($o\text{-}PO_4$), and nitrogen (NH_4), the latter three pollutants being measured weekly or monthly.

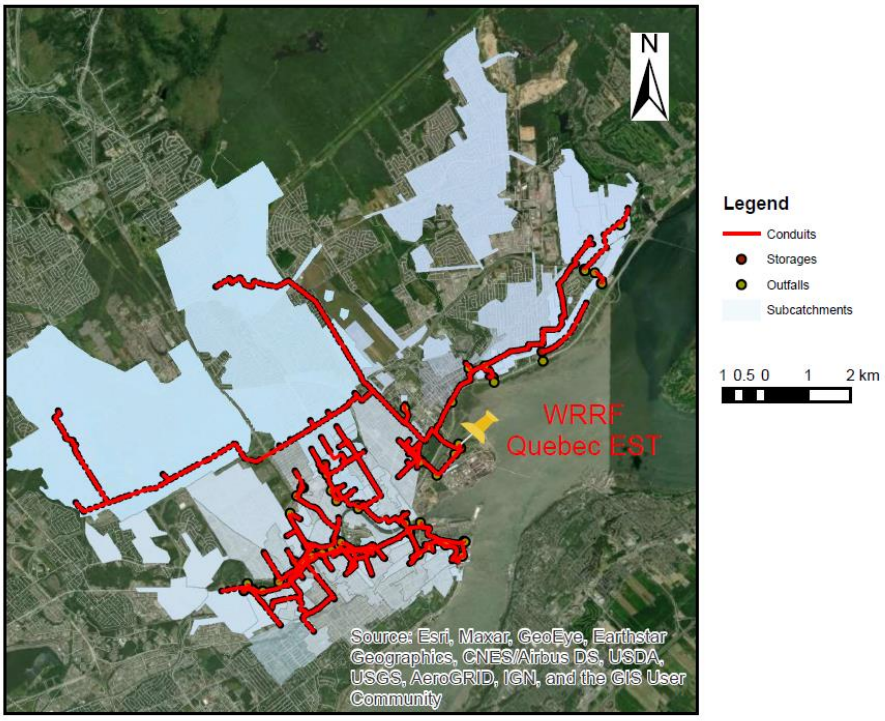


Figure 2-2 The catchment of WRRF EST in Quebec City, the red line represents the combined sewer systems, the pin represents the WRRF location.

The second case study 'Clos de Hilde' (CdH) is located in the city of Bordeaux, in the south-west of France. The CdH catchment is a typical urban catchment, receiving consisting of housing, industrial from both combined (hatched) and separate sewer system (doted), see Figure 2-3. The flowrate has an annual pattern flowrate, which decrease in summer and autumn but increase during winter and spring. The catchment has a typical oceanic climate with evenly distributed rain throughout the year, but storms in summer. The CdH WWTP has a treatment capacity of about 400 000 PE and a flowrate of 100 000 m^3/d.

This case study contributes a higher frequency influent generation. The flowrate data are available at high frequency and the quality data collection focuses on TSS, turbidity and total and soluble COD, which are collected by Ledergerber et al., (2020).

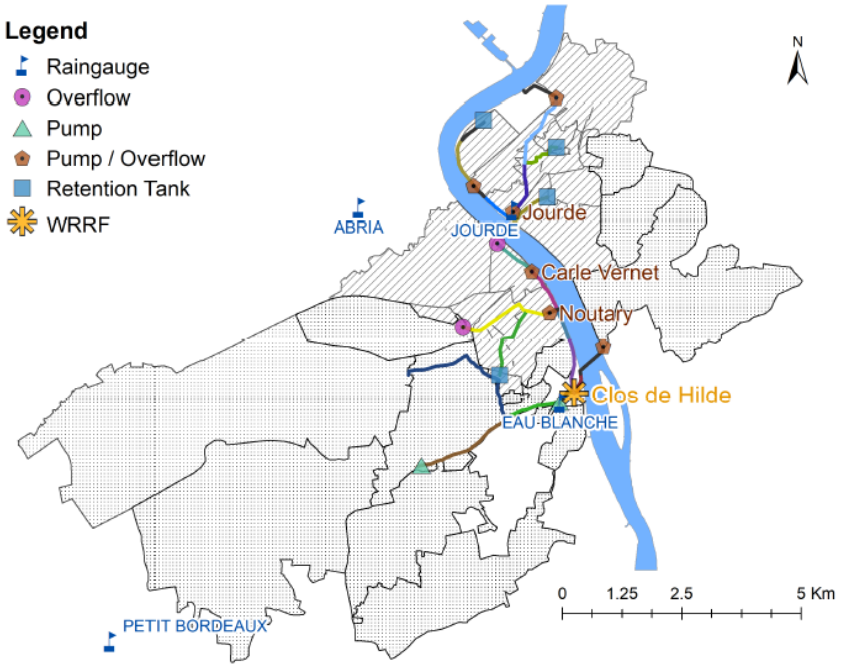


Figure 2-3 Sewer system and WRRF of the CdH catchment (Julia Margrit Ledergerber et al., 2020)

In general, the IG-model calibration and validation will be studied for these three different case studies, in order to get a good performance in stability and extendibility. After developing the IG model, the integrated evaluation and assessment will be discussed according to different models, including the complexity of the processes and the performance of model etc., in order to have an overview conclusion and an outlook for the future studies.

2.2 Model development

After collection the data, these datasets will be analyzed and modelled with software tools including MATLAB, PYTHON environment, e.g., TensorFlow with Keras libraries and WEST with IUWS library.

- MATLAB and SIMULINK

The MATLAB software version 2020b (www.mathworks.com) is used for data treatment including data pre-treatment, data filtering (outlier detection and smoothing) and fault detection, developing of full connected multi-layer perceptron and multivariate regression as well as the development a tool for the performance analysis between different timeseries. The scripts are written by building and compositing a series of functions. The multilayer perceptron is used from the deep learning Toolbox.

The second part of using MATLAB is an application and modification of an open source of influent generator model for benchmark simulation model (BSM). This BSM simulation package in MATLAB/Simulink toolbox possesses an open source code and an interaction graphical interface and code. The original model was developed by Gernaey et al., (2011).

- Python

The machine learning and deep learning model development and simulation are carried out in Python programming language, with mainly the machine learning platform TensorFlow (www.tensorflow.org) and Keras library (Chollet, 2015), which is an open source plate form for machine learning and it is flexible and user friendly for engineering model development. TensorFlow and Keras ensure an efficient tool for modelling architecture design, training and optimization. These libraries provide a wide range of data mining tools which allowed to develop, compile and optimize efficiently machine learning models.

Modelling details will be explained in each corresponding chapters.

- WEST (Version 2016)

WEST (Wastewater treatment plant Engine for Simulation and Training, by DHI) with IUWS Library (www.mikepoweredbydhi.com/products/west) provides a platform for phenomenological catchment modelling, which is used for pil*EAU*te catchment modelling, in order to validate the with experiments result and analyze the catchment properties.

Chapter 3. Characterization, modelling and calibration a conceptual model for urban wastewater influent generation in a pilot scale catchment with a combined sewer system

3.1 Abstract

The WRRF is designed for receiving and treating the wastewater generated by municipal and industrial activities as well as the wet weather flow (WWF) for a combined sewer system. The dry weather flow (DWF) consists the base sanitary flow and groundwater infiltration, and the WWF refers to the rainfall dependent infiltration and inflow. The inflow property change will cause the WRRF system and impact the biochemical process, therefore, the analysis of this inflow behaviour is certainly important for WRRF operation and modelling studies. In this chapter, the influent dynamic (flowrate and pollutant) and COD characterization are monitoring, visualized and analysis for a pilot scale catchment with a combined sewer system in dry weather and wet weather, then the raw wastewater influent dynamic is modelled by a conceptual model based on WEST software consisting of the catchment, conceptual sewer system and a storage, which allows to generate the wastewater flow and concentration dynamics under impact of storm water.

3.2 Résumé

Le StaRRE est conçu pour recevoir et traiter les eaux usées générées par les activités municipales et industrielles ainsi que le débit par temps de pluie (WWF) pour un système d'égout unitaire. Le débit par temps sec (DWF) comprend le débit sanitaire de base et l'infiltration des eaux souterraines, et le WWF fait référence à l'infiltration et à l'afflux dépendant des précipitations. Le changement de propriété de l'afflux entraînera le système WRRF et aura un impact sur le processus biochimique, par conséquent, l'analyse de ce comportement d'afflux est certainement importante pour le fonctionnement du WRRF et les études de modélisation. Dans ce chapitre, la dynamique de l'influent (débit et polluant) et la caractérisation de la DCO sont suivies, visualisées et analysées pour un captage à l'échelle pilote avec un système d'égout unitaire en temps sec et temps humide, puis la dynamique de l'influent des eaux usées brutes est modélisée par un modèle conceptuel basé sur le logiciel WEST composé du captage, du réseau d'égouts conceptuel et d'un stockage, qui permet de générer la dynamique d'écoulement et de concentration des eaux usées sous l'impact des eaux pluviales.

3.3 Background and objective

A wastewater treatment plant (WWTP) is designed to receive and treat the wastewater generated by municipal and industrial activities. For a combined sewer system, it also has to handle wet weather flow (WWF). The dry weather flow (DWF) consists of the base sanitary flow and groundwater infiltration, and the WWF includes the rainfall-dependent infiltration and inflow (RDII) (Dent et al., 2000). The major source of pollutants in DWF originates from domestic sanitary wastewater, including black water, i.e. toilet waste, containing the majority of nutrients, and grey water, originating from bathtubs, washing machines etc. (Hocaoglu et al., 2010). The DWF characteristics are also influenced by the presence of industrial effluents within the catchment, which may lead to a significant difference from purely domestic effluent (Brouckaert and Mhlanga, 2013; Rössle and Pretorius, 2001). The WWF depends on different factors, such as the level of urbanization and the layout of the sewer system, the land use and the wet weather event itself (duration, intensity etc.). The WWF is composed of direct or indirect infiltration, first flush or resuspension of sewer deposits, which makes that the characterization of WWF is different from DWF. The increase of the flowrate and the variability of the influent quality in different weather conditions will bring a different result for the wastewater treatment result, so that the characterization and the modelling of wastewater influent are important for process operation and control, for bioprocess modelling and for digital twin development.

In order to get an accurate model of the influent for a WWTP in different weather conditions, the identification of the inflow and the characterization of influent pollution are carried out. Different studies have been conducted to measure and characterize the wastewater composition. At source domestic wastewater has been studied to obtain a relative daily pattern of pollutographs based on a household appliance use survey (Almeida et al., 1999). Daily wastewater pollutant dynamics are also influenced by population size and catchment structure (Le et al., 2017). Besides the major pollutants, diurnal patterns have also been demonstrated to exist for micropollutant concentrations (Alfiya et al., 2018; Eriksson et al., 2009). It is well-known that the pollutant fractions provide varying biochemical substrates for different species of microorganisms for their biological processes. The variability of the COD and TKN fractions have been studied, demonstrating the variations depending on the sources of the wastewater (Choi et al., 2017; Rössle and Pretorius, 2001). The average fractions vary also depending on the region (Pasztor et al., 2009). Moreover, the fractionation is influenced by weather conditions. In wet weather, the stormwater introduces a considerable inert COD load, while the readily biodegradable COD fraction is strongly diluted (Almeida et al., 1999). Despite these experimental results, as far as known to the author, no research has been conducted to model the dynamics of fractionation so far. On the other hand, the other important factor to consider for influent characterization is the modelling of the sewer system to describe the transformations occurring from the wastewater source to the inlet of the WWTP. The sewer transport process includes different mechanisms: advection, dispersion (Gujer, 2008), settling and resuspension (Michelbach,

1995) and biodegradation. The hydraulic behaviour in the sewer network can be described by either detailed hydrodynamic models such as the Saint-Venant equations or by simplified conceptual modelling tools such as those using multi-linear reservoir models, etc. The multi-reservoirs model has been successfully applied and validated in many IUWS model studies (Ledergerber et al., 2019; Solvi et al., 2006). In the current BSM influent generator model, a simplified sewer system model is used to represent the mass balance (Gernaey et al., 2005).

In this study, the influent dynamics (flowrate and pollutants) and the COD characterization are monitored, visualized and analysed for a pilot scale sewer catchment with a combined sewer system under dry and wet weather conditions. Stormwater inflow leads to an increase in flow and a modification of the pollutant concentrations. The influent data is treated and visualized by Dash, a script developed in Python. The raw wastewater influent dynamics are modelled by a conceptual model implemented in the WEST modelling and simulation software using its integrated urban wastewater systems library (IUWS) (http://www.mikepoweredbydhi.com, Vanhooren et al., 2003). The model includes submodels for the catchment, the conceptual sewer system and the storage tank installed at the inlet of the pilot facility as a continuous stirred tank reactor (CSTR). The calibration is performed on the basis of experimental results. By applying different patterns for soluble and particular pollutants and a random walk stochastic submodel, the overall model can better describe reality.

3.3 Materials and methods

3.3.1 Catchment and wastewater composition

The pilEAUte activated sludge WWTP is located at Université Laval, receiving the sewage wastewater from Pavillon Lacerte, a student residence, and two kindergartens to which stormwater from the parking zone, the green space, and the building roofs is added. The number of population equivalents it serves is around 350-400 persons, and the sewage wastewater is a combined sewer system, with 60% impermeability of the catchment area.

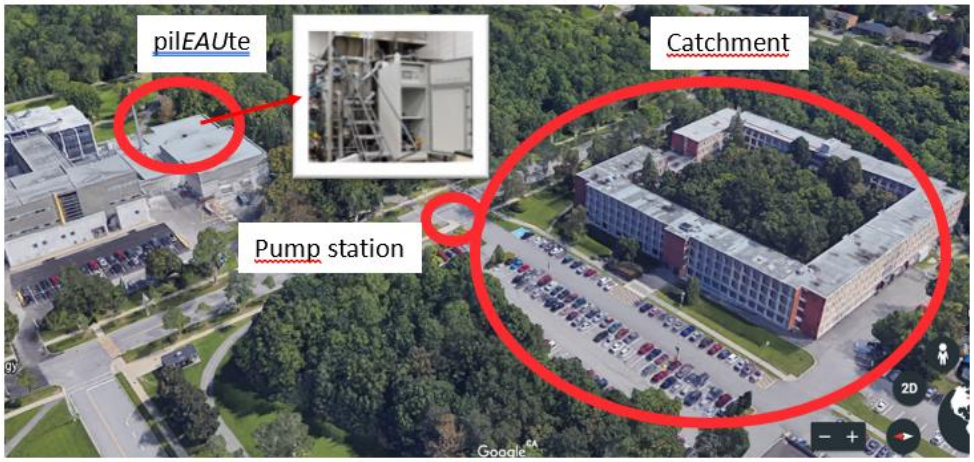


Figure 3-1 Localization of pilEAUte treatment plant on the Université Laval campus in Québec City, QC, Canada

Figure 3-2 shows the plan of the pump station and the sewer system. The pump station feeds the pil*EAU*te system, with a pump capacity of 10m³/h. The combined sewer system between the *Lacerte* residence and the pil*EAU*te WWTP is relatively short, limiting biodegradation during transport, which allows studying the influent dynamics almost at source. The wastewater is conveyed to a storage tank by a pressure pipe (brown line), which is the entry of the pil*EAU*te facility.

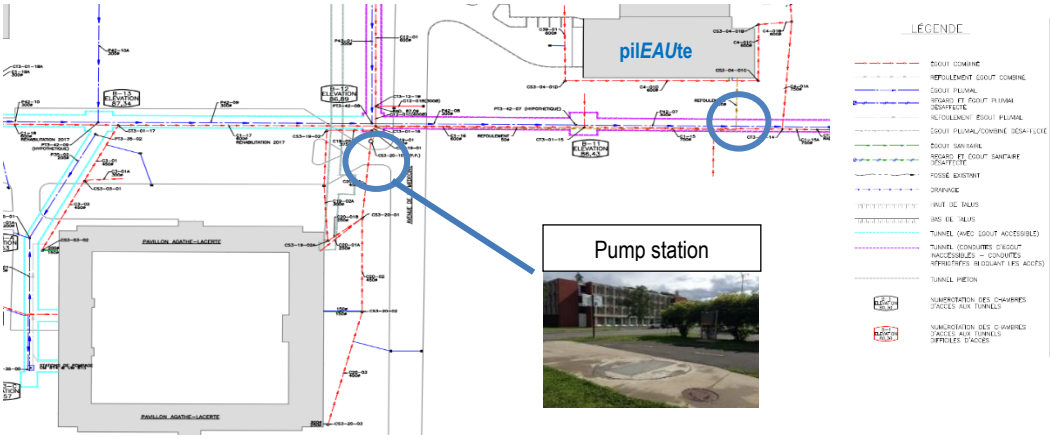
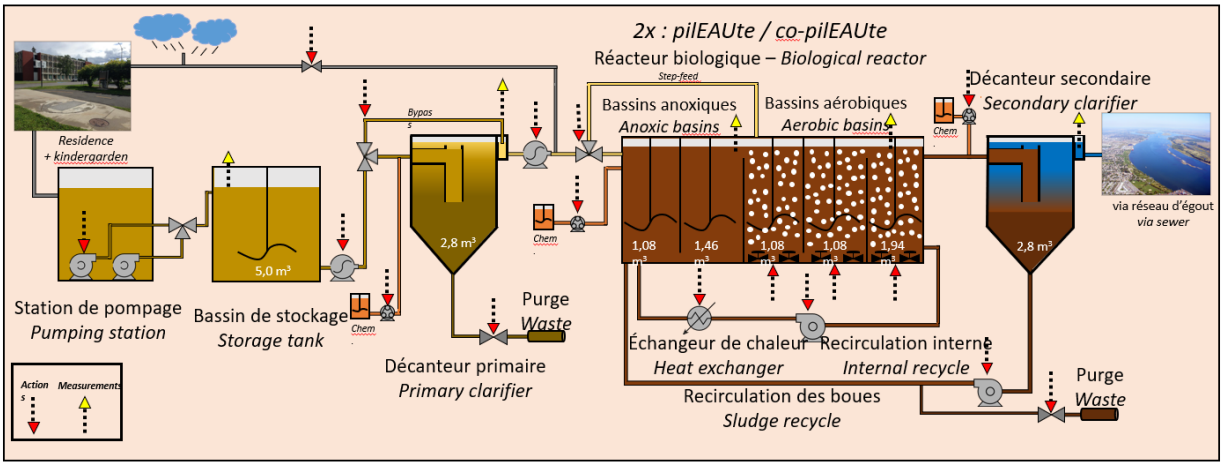


Figure 3-2 Plan of pump station and sewer system for pilEAUte alimentation

shows the overview schematic view of the pilEAUte treatment plant and its complete treatment process. The plant is monitored by a supervisory control and data acquisition system (SCADA). The sampling point is the place where the data are collected by the different sensors and it is the place where the measurement campaign was conducted, which enables the characterization of the raw wastewater and to the identification of the pollutant composition.

Figure 3-3 shows an overview of the pil*EAU*te treatment plant and its complete treatment process. The plant is monitored by a supervisory control and data acquisition system (SCADA). The sampling point is the place where data are collected by the different sensors and it is the place where the measurement campaign was conducted to characterize the raw wastewater and to identify the pollutant composition.



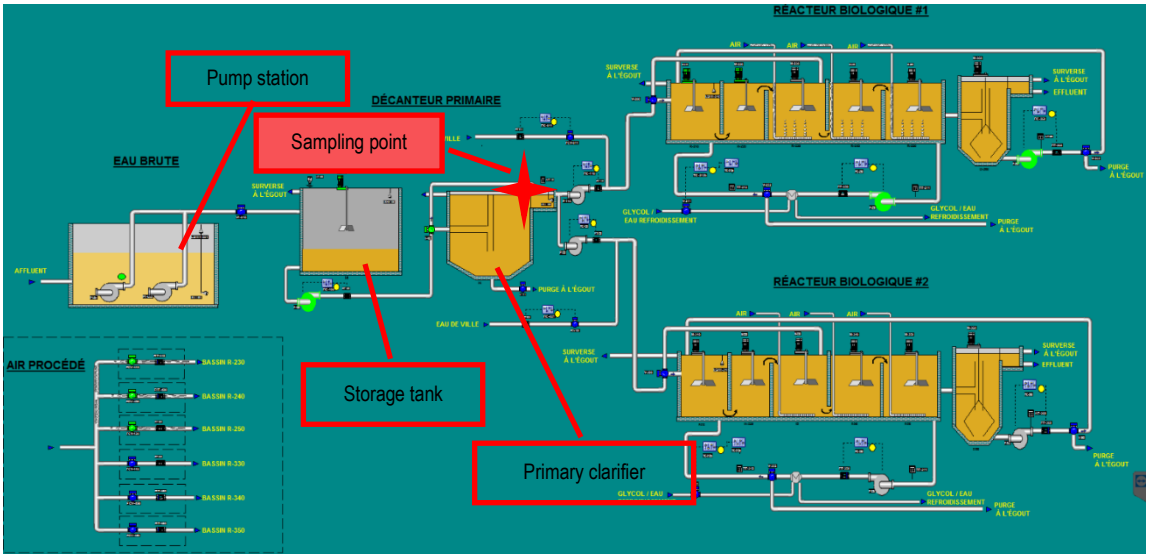


Figure 3-3 Schematic of the treatment process of the pilEAUte treatment plant

3.3.2 Catchment and wastewater composition

3.3.2.1 Flow estimation

The pilEAUte facility unfortunately is not equipped with a flowmeter at the inlet of the storage tank. The only flowmeter available is the flowmeter measuring the pumped flow out of the storage tank to the treatment plant. Still, the inlet flowrate can be calculated by monitoring the rate of filling the storage tank and the knowledge of the outflow from the storage tank. The quicker the height increases in storage tank, the bigger the flowrate is.

$$\frac{dh_{storage}}{dt} = \frac{1}{A_{storage\ surface}} * (Q_{catchment}) - Q_{pump} \qquad 3\text{-}1$$

During the measurement campaign, the storage tank outlet was closed which further simplifies the above mass balance. The inlet flow can then simply be computed from the increase of the water volume in the storage tank monitored by continuously measuring the water level. Given the finite capacity of the storage tank, whenever the water level reaches the maximum storage level, which is around every 4 hours, an emptying operation is performed by opening the valve at the bottom of the storage tank to resume the inflow monitoring procedure.

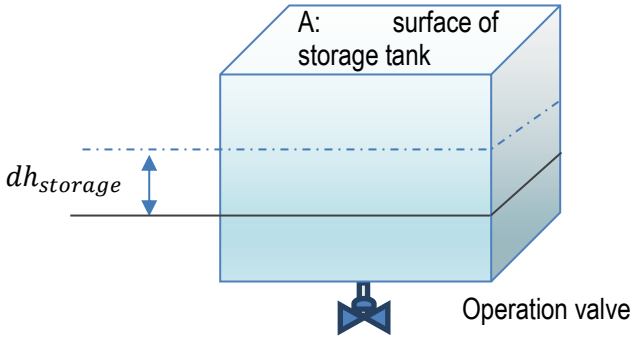


Figure 3-4 Flow measurement by storage tank

3.3.2.2 Online data collection for wastewater composition

For the pilEAUte case study the online data collection system which the system is equipped with, provides continuous data thanks to two measurement systems for monitoring and storing data: a monEAU station (Rieger and Vanrolleghem, 2008) and the pil*EAU*te's SCADA (Supervisory Control and Data Acquisition) system.

Six different sensors are installed at the sampling point at the outlet of the primary clarifier. Various parameters are observed: the spectro::lyser and ammo::lyser (s::can, Vienna, Austria) are connected to the monEAU station, while the Solitax and conductivity meter (Hach, Loveland, Colorado, USA) are monitored by the SCADA system. Finally, a RODTOX on-line respirometer, Kelma, Niel, Belgium (Vanrolleghem et al., 1994) is connected to the

monEAU station as well. The different pollutants listed in Figure 3-5 can be monitored by these online measurements.

spectro::lyser	ammo::lyser	Solitax	Conductivity	RODTOX
•CODsoluble •CODtotal •TSS •Nitrate	•Ammonia •Potassium •Temperature •pH	•TSS	•Conductivity	•BOD •Fraction

Figure 3-5 Online sensors avalaible at the pilEAUte plant and the pollutants they can measure

The monEAU automated measuring station (RSM30, Primodal, Hamilton, Ontario), as shown in Figure 3-6, is a powerful tool to collect and store data with high quality. The monEAU station data are stored in the monEAU station's database in files with extension '.par' and '.tsdb'. The monEAU station also includes sensor calibration and maintenance data to guarantee the reliability of the measurements.

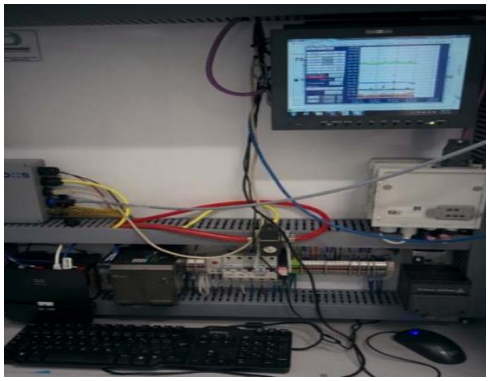


Figure 3-6 monEAU station for data collection

The SCADA system, provided by Veolia as part of the pilEAUte plant, is a system to visualize, manage and control the various unit processes of the pilEAUte plant in an intuitive way. Sensor information is stored in a database under the SQL language (Server Query Language), every minute for most sensors. The frequency is modifiable according to user needs.

The RODTOX is an online and autonomous respirometer (Vanrolleghem et al., 1994), as shown in Figure 3-7. The respiration rate of the microorganisms can be calculated and interpreted by the calculation of the oxygen

consumption and the oxygen mass transfer coefficient (KLa). A mathematical model of the RODTOX operation has been developed in previous research, allowing to estimate some biokinetic coefficients of the activated sludge and the wastewater fractionation (readily biodegradable and slowly biodegradable substrate)(Therrien, 2017).

Figure 3-7 Schematic of RODTOX system

3.3.2.3 Wastewater composition measurement campaigns

The COD fractionation, and the TP and TN (total phosphorus and total nitrogen) concentrations are important to describe the influent. However, no online measurement is available for the nutrients and the different COD fractions. Hence, to complete the characterization of the wastewater pollution and to deal with sensor measurement errors, which may occur when the sensors are fouled, or the calibration is not adequate, additional measurement campaigns were planned.

An autosampler was used to grab samples each 15 min, and the samples will be combined into a composite sample for two hours. As a result, 12 samples are collected each day.

Analyses of the samples was made for all pollutants mentioned in section 3.3.2.2, as well as for the two nutrients (total phosphorus and nitrogen), the COD fractionation, and alkalinity. Alkalinity was measured by a titrator (794 Basic Titrino, Metrohm, Switzerland). The concentration of all other pollutants was realized with Hach (Loveland, CO, USA) kit reagents and measured spectrophotometrically using DR5000 Hach Spectrophotometer (Loveland, Colorado, US).

In order to analyze the difference in dynamics during DWF and WWF conditions, the measurement campaigns were duplicated for both dry and wet weather situations.

3.3.3 COD fraction and biodegradability

According the Activated Sludge Model No2 (Henze et al., 2000), the COD fractionation of the influent can be performed according the following steps:

- Characterize the influent total and soluble COD concentration.
- Determine the readily biodegradable COD by RODTOX respirometry.
- Identify the slowly biodegradable COD (BOD_{28}) concentration by an OxiTop® Control respirometer.
- Identify the fraction of non-biodegradable particulate COD.

The information on the wastewater fractions is obtained by applying the equations in Table 3-1.

Table 3-1 Fractionation for COD (Roeleveld and van Loosdrecht, 2002)

COD_{total}	New notation*	$= S_s + S_I + X_s + X_I$
S_I	S_U	$= sCOD_{eff}$
S_s	$S_F + S_A$	$= sCOD_{inf} - S_I$
X_s	SC_B	$= \frac{BOD_{u,inf}}{1 - f_{BOD}} - S_s$
X_I	X_u	$= COD_{t,inf} - (S_s + X_s + S_I)$

where:

S_A: Volatile fatty acids (acetate)

S_F: Readily (fermentable) biodegradable substrate

S_I: Inert (soluble) non-biodegradable organics

X_I: Inert (particulate) non-biodegradable organics

X_S: Slowly biodegradable substrate

*A new notational framework has been developed to unify the notation, based on subscript levels that provide greater specification (Corominas et al., 2010).

The S_s is the readily biodegradable fraction, including S_A and S_F, which indicates easily available sources of organic carbon in wastewater. The S_s is divided in to S_A and S_F in the BSM model, S_I is the non-biodegradable soluble fraction, which consists of organic compounds that do not take part in the biological treatment processes,

X_s is the slowly biodegradable fraction, which has a major influence on the dynamic behaviour of the activated sludge process, including oxygen demand, and also forms one of the main design parameters of biological nitrogen and phosphorus removal systems and X_I is the non-biodegradable particulate fraction, which will be collected in the mixed liquor suspended solids and will accumulate in the system. f_{BOD} is a correction factor, varying between $0.1 \sim 0.2$ (Roeleveld and van Loosdrecht, 2002).

3.4 Modelling and calibration of the catchment model

The modelling of the pilEAUte plant's catchment is carried out based in the WEST modelling software using the 2016 IUWS library. The conceptual model consists of a catchment generator, the sewer system modelled as a tanks cascade, and the pil*EAU*te's storage tank as a continuous stirred tank reactor (CSTR).

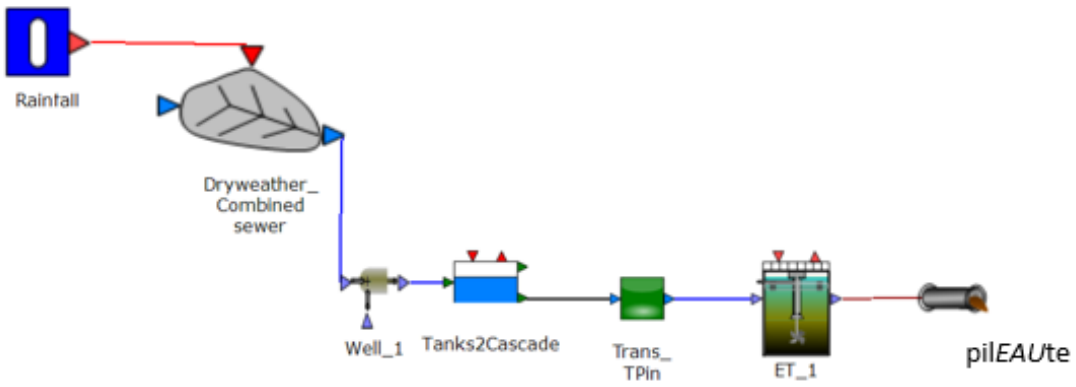


Figure 3-8Model layout

The dry weather combined sewer is the block describing the catchment properties, including the size of the catchment, the dry weather flowrate and the diurnal household discharge patterns (including the number of population equivalents), the infiltration water, etc. The Tanks2cascade block represents the small length sewer system, with a set of linear or non-linear reservoirs (which are in fact stirred tanks in series with variable volume). The Trans_TPin block is included to transform the pollutant concentrations modelled in the sewer system into the wastewater fractions according to the ASM model, as shown in Table 3-2.

The continuous stirred tank reactor (CSTR) mass balance model for water quantity in the sewer pipes is set up as equations 3-2 and 3-3.

$$\frac{\partial h}{\partial t} = \frac{1}{A} * (Q_{in} - Q_{out}) \qquad 3\text{-}2$$

$$Q_{out} = \frac{1}{k} * {h_2}^{1/p} \qquad 3\text{-}3$$

where h presents the water level in the reservoir and Q_{in} is the flowrate generated by an upstream model block, and Q_{out} is the outflow of the reservoir. When the value of the constant p is one, the model corresponds to a linear reservoir, otherwise it would be a non-linear reservoir. The model parameter k is the storage constant of the reservoir, also known as the residence time, that has to be estimated.

Equation 3-4 describes the water quality model:

$$\frac{\partial V.C_{out}}{\partial t} = Q_{in}C_{in} - Q_{out}C_{out} + r \qquad 3\text{-}4$$

where V is the volume of the reservoir and Q_{in} and Q_{out} represent the flowrate at the inlet and outlet of each reservoir, respectively. The C_{in} in and C_{out} are the concentrations of inlet and outlet. Finally, r represents the physico-biochemical processes affecting the concentration, including the effect of sedimentation and suspension, etc.

Table 3-2 Organic matter transformation table in WEST

Organic matter transformation table				
Soluble organic matter			Particulate organic matter	
Inert COD	Readily biodegradable matter (S_S)		Inert COD	Slowly biodegradable
	Fermentable	Fermentation products		
S_I	S_F	S_A	X_I	X_S

3.5 Results and discussion

3.5.1 Online influent data visualization tool

Data visualization, as a tool for understanding and communication, is important for automated WRRF operation and control as well as to support digital twins. An online influent data visualization tool was developed to observe, analyze and monitor the pilEAUte's data in real time.

The data are collected by a data collection station and stored in the datEAUbase server (Plana et al., 2018) , and a web-based visualization interface was created in Dash. The users can define the interface layout by the SQL and Python requests in real time. The interactive visualization interface displays the chosen influent variables in different types of plots and statistics.

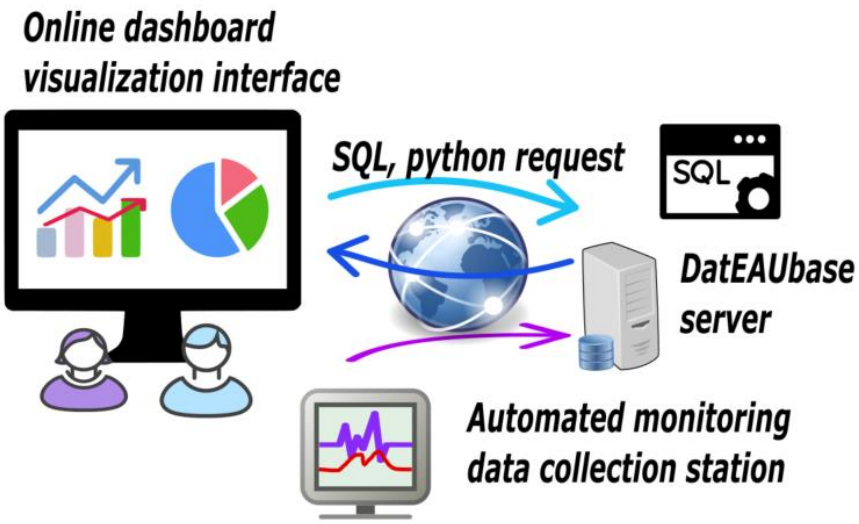


Figure 3-9 Data flow for the pilEAUte's online dashboard visualization interface

The interface is web-based and can be reached from a web browser. The influent loads are represented by violin plots in order to assess their variability and the influent concentrations are represented over time to study historical trends. At the same time, relevant ratios are calculated to quickly diagnose the influent characteristics.

Figure 3-11 shows the violin plot for the influent for a 2 week period, together with the statistical information (mean, median, percentile etc.).

Figure 3-12 shows the concentration evolution over time and the current ratios in a dashboard. The latter can alert in case abnormal influent changes occur and may affect system operation.

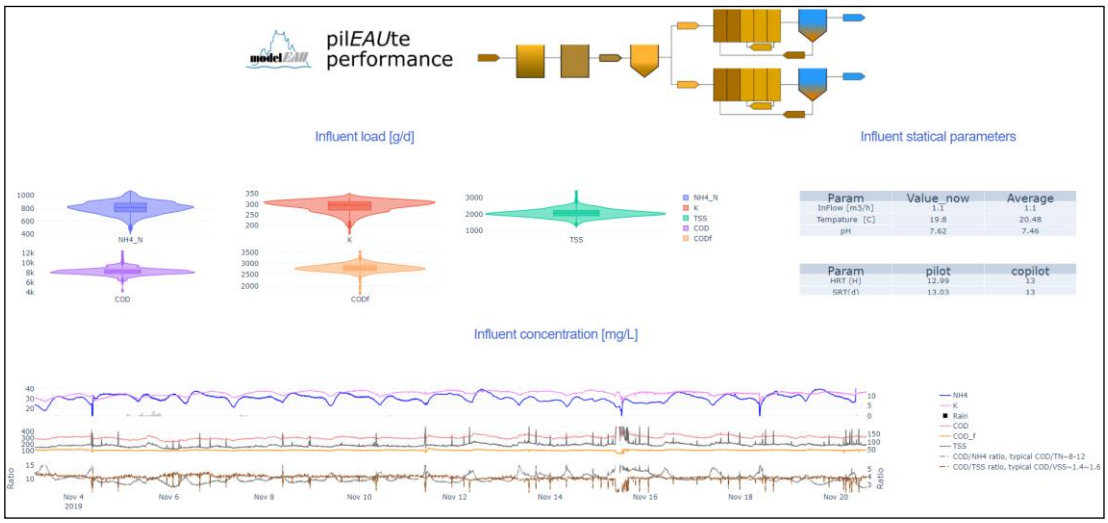


Figure 3-10 Overview of the pilEAUte's influent dashboard for a two-week period

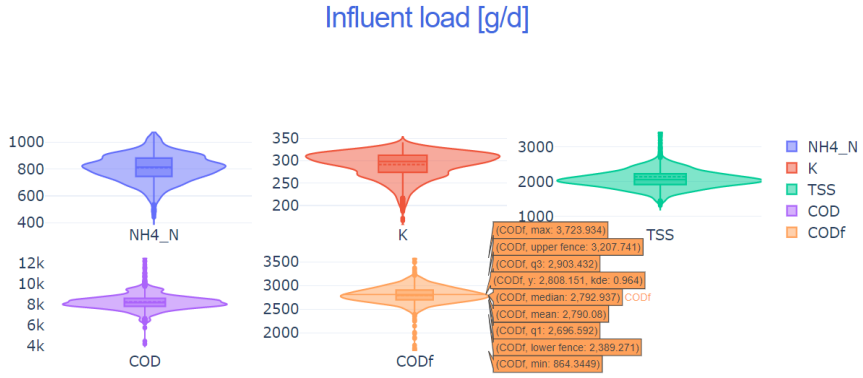


Figure 3-11 Details on influent load variability in the pilEAUte for a two-week period

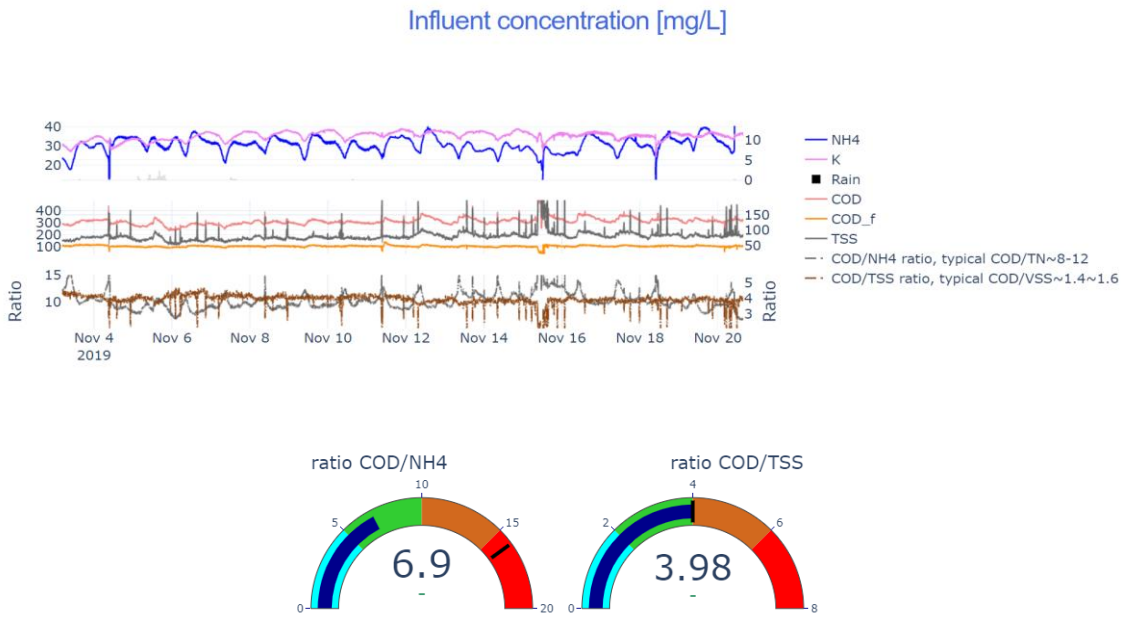


Figure 3-12 Influent concentrations at the pilEAUte for a two-week period and the current ratios for particular wastewater characteristics

3.5.2 Flow measurement and validation

In order to characterize the diurnal and weekly pattern for flow and understand the influence of rain events, 4 measurement campaigns were carried out as shown in Table 3-3.

Table 3-3 Summary of measurement campaign for flowrate

Measurement campaign	Time	Weather	Period of week	Application objectives
1	September, 2018	DW	Weekday	Training phase
2	October, 2018	DW	Weekend	Training phase
3	October, 2018	WW	Weekday	Training phase
4	October, 2018	WW	Weekday	Validation phase

The measurement results are shown in Figure 3-13. The dry weather flows for weekdays (blue line) and weekend days (red line) demonstrate a difference in the weekly pattern that is reflecting the population's lifestyle. At the weekend, the morning peak appears later than on weekdays. The black line represents the average diurnal flow, calculated from three DWF days.

Besides, a wet weather event was surveyed to understand the impact of rain on the typical daily pattern. The two peaks (morning and evening) corresponding to the urban activities under dry weather are impacted by the rain event that increases the inflow. During the night (2h-6h), because the urban activities are very low, the flowrate (about 0.5 m^3/h) represents the infiltration of groundwater, even in this very short sewer. Alternatively, other water leaks in the system may be causing this background flow.

The average flow pattern is used as a diurnal pattern for modelling purposes. Figure 3-14 demonstrates the validation of the flowrate simulations (scenario 4) with the WEST model. The catchment serves about 400 PE with 210L/d average flowrate per person. In this figure, days 1 and 2 represent the DWF and the third day represents the rain event. The quality of the flowrate prediction is evaluated by the normalized root mean squared error (NRMSE), shown in equation 3-5.

$$NRMSE = \frac{RMSE}{\bar{y}} = \frac{\sqrt{\sum_{i=1}^{n} \frac{(y_{observed} - y_{simulated})^2}{n}}}{\overline{y_{observed}}} \qquad 3\text{-}5$$

where $y_{observed}$ is the measured data, $y_{simulated}$ is the simulation result and n is the number of data. The NRMSE in the validation set is 3.21%.

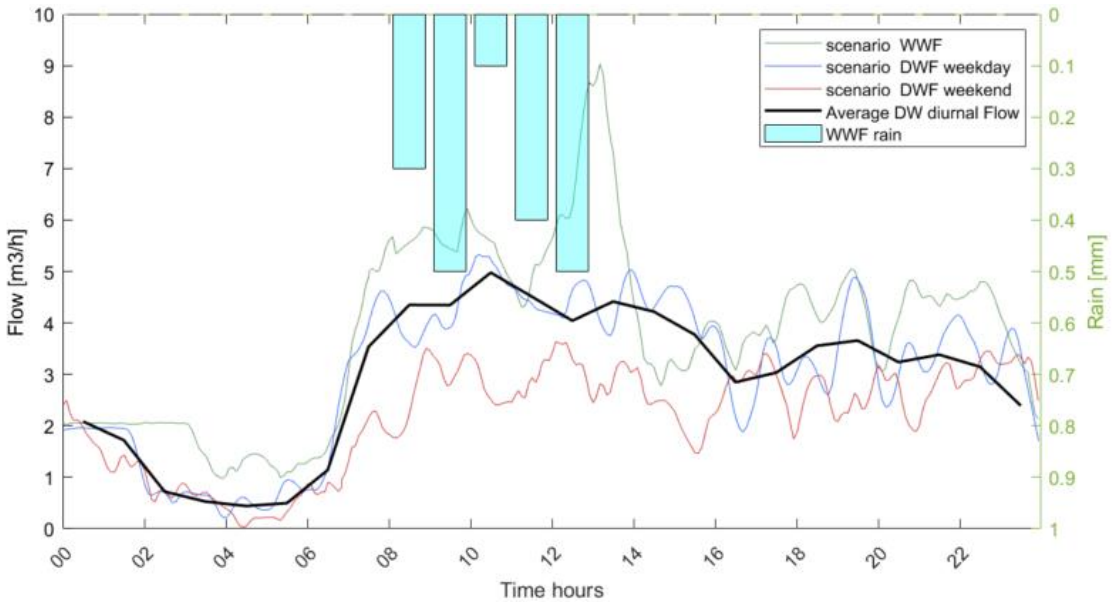


Figure 3-13 Results of three flowrate measurement campaigns using the method described in Section 3.

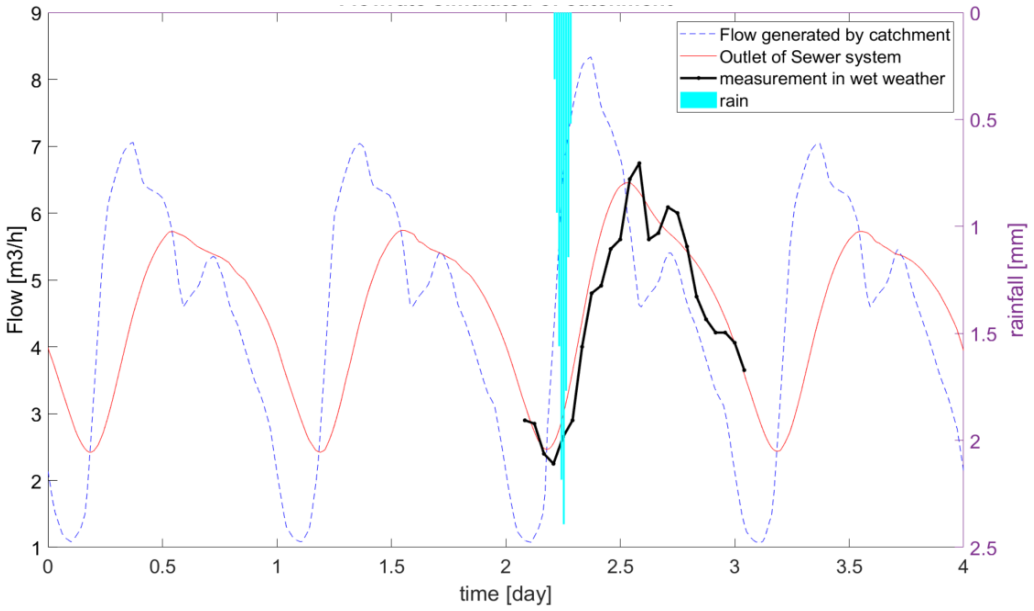


Figure 3-14 WEST simulation for validation of the catchment model by independent measurements of flow rate.

3.5.3 Pollutant characterization and fractionation

The characterization of the wastewater composition is carried out for dry weather in order to understand the diurnal water quality dynamics and for wet weather in order to demonstrate the influence of precipitation. The measurement campaigns were carried out as shown in Table 3-4.

Table 3-4 Summary of measurement campaign for wastewater pollutant characterization and fractionation,

Measurement campaign	Time	Weather	Season	Objectives	Application
1	October, 2018	DW	Autumn	TSS, total COD, soluble COD N,P K, NH_3-N, conductivity	Training phase
2	October, 2018	DW	Autumn	TSS/ VSS COD fractionation	Training phase
3	October, 2018	WW	Autumn	TSS/ VSS	Training phase
4	October, 2018	WW	Autumn	COD fractionation	Training phase
5	November, 2018	WW	Autumn	TSS, total COD, soluble COD N,P K, NH_3-N, conductivity	Training phase
6	February, 2019	DW	Winter	TSS, total COD, soluble COD N,P K, NH_3-N, conductivity	Validation phase

6	April, 2019	WW	Winter	TSS, total COD, soluble COD COD fractionation	Validation phase

First, the diurnal pattern of TSS and (total and soluble) COD concentrations during DWF are presented in Figure 3-15. The pink line represents the DWF for the October measurement campaign, corresponding to autumn and the blue line represents the DWF in February, corresponding to winter conditions.

In general, the diurnal activities can be observed from the lab results: there is a strong morning peak (around 8h), and a smaller peak in the evening (18-22h) and the concentrations stays low and stable at night (0-5h). This reflects the student activities. Noticeably, the concentration in winter is slightly lower than in autumn, possibly because the student population is higher during the autumn session.

By comparing the DWF and WWF, see WWF of Figure 3-15, the dilution effect can be observed. At the beginning of the rain event, the concentrations (10h) are still very close to the corresponding dry weather concentrations. However, in the middle of the storm (12h-14h), the concentrations decrease significantly. At the end of the storm, (after 4h the next day), the concentrations start to return to the DWF level.

Figure 3-16 describes the nutrient (total N and total P) concentration dynamics for DWF and WWF, (measurement campaign 1 and 5 in Table 3-4, with the methodology provided in section 3.3.2.3). It can be observed that the potassium, ammonia and conductivity follow the same pattern. Again, one can observe two peaks: one in the morning and one in the evening. And again, the concentration in autumn is slightly higher than in winter.

Figure 3-18 shows the TSS and VSS (volatile suspended solids) dynamics, and their ratio. VSS is analysed since it gives a rough approximation of the amount of organic matter present in the suspended fraction of wastewater. The figure demonstrates that the raw wastewater of the study site is highly organic (the ratio of VSS over TSS is nearly 85% - 95%), compared with the value in the ASM modelling guideline (VSS/TSS around 80%) (Rieger et al., 2012). The ratio is high during the day and decreases by the evening. The WWF results illustrate the impact of stormwater on the wastewater composition: the TSS and VSS concentrations are lower due to dilution. Also, the VSS/TSS-ratio decreases, which reflects that the stormwater introduces inorganic solids into the sewer system and resuspends material that had been in the sewer for a while, with its organic fraction gradually biodegraded.

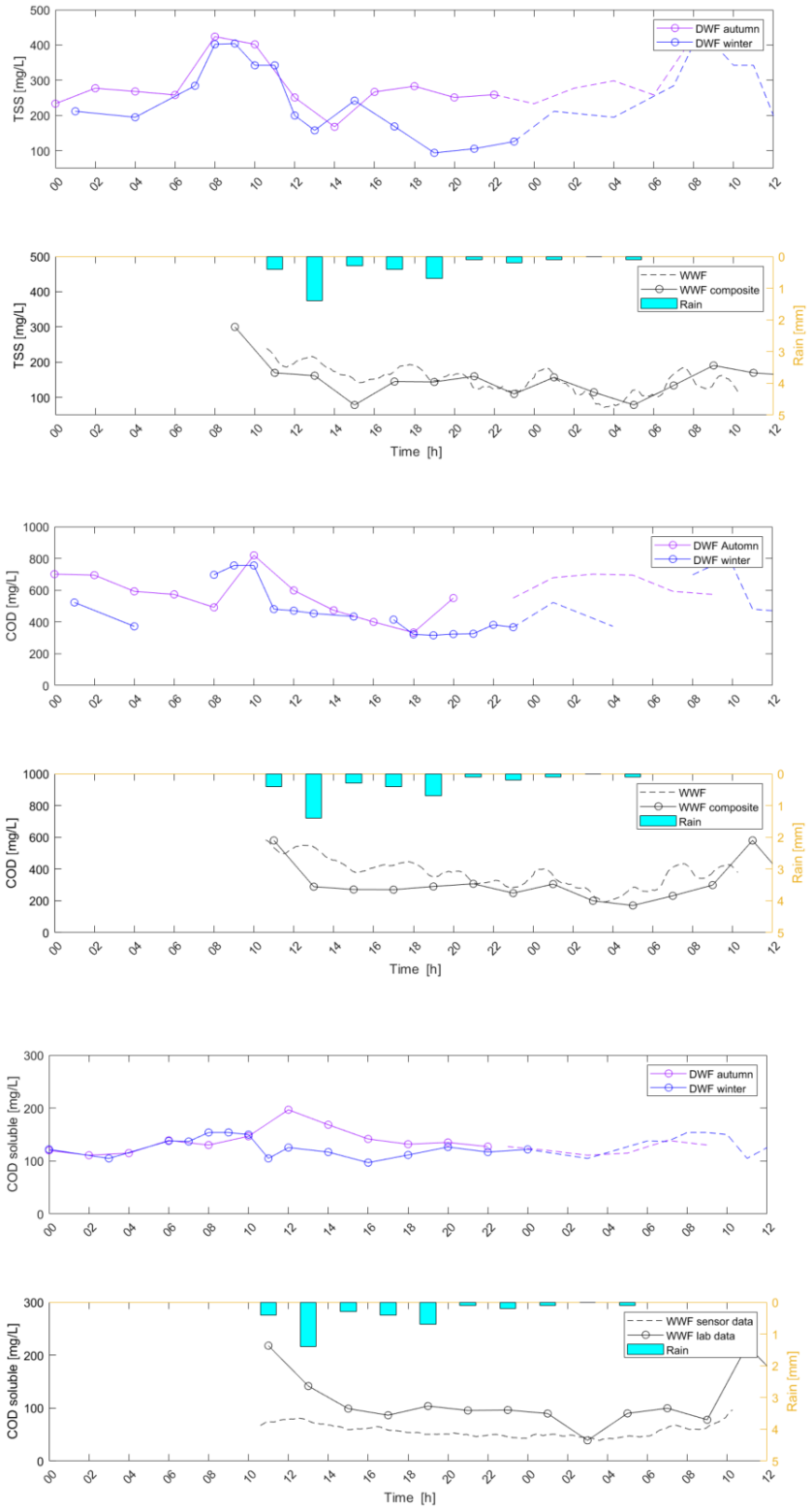


Figure 3-15 TSS, total COD and soluble COD diurnal pattern for DWF (measured in autumn and winter) and WWF at the pilEAUte facility. The circles represent the measured values. In the DWF plots, the dashed lines represent the pattern repeated from the previous day and in the WWF plots, the dashed lines represent the continuous sensor data.

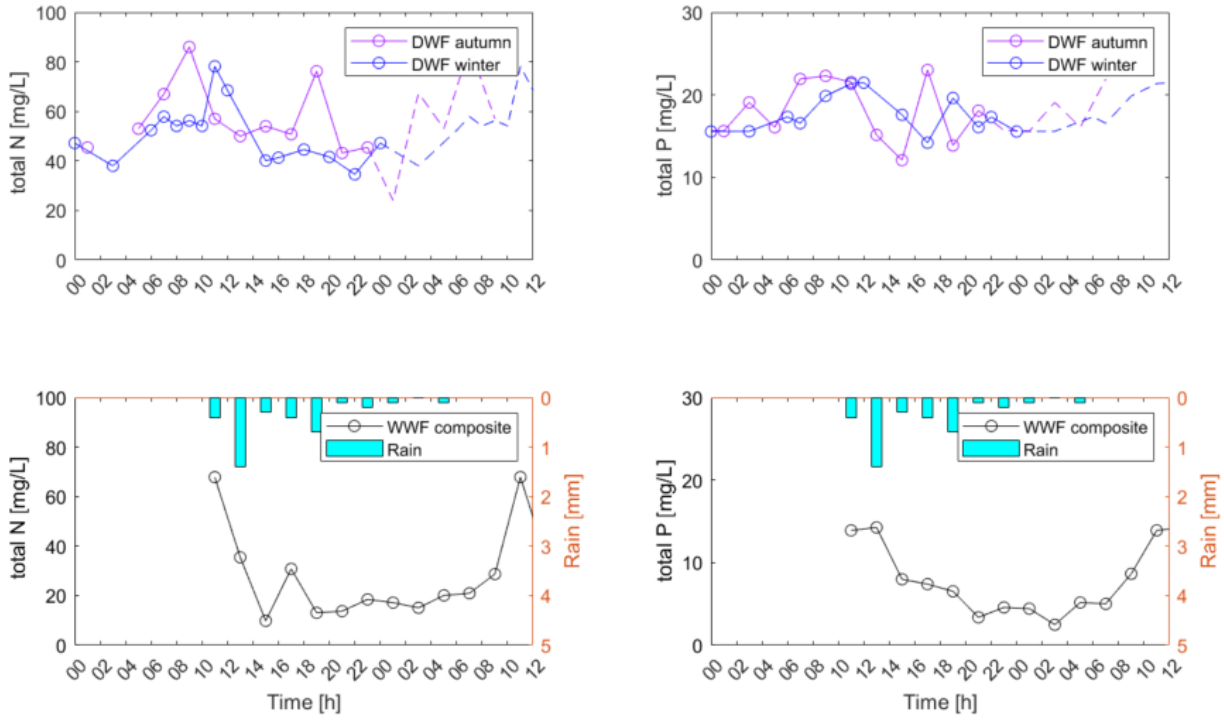


Figure 3-16 Diurnal pattern for the nutrient concentrations (total N and total P) for DWF (measured in autumn and winter) and WWF at the pilEAUte facility, the dashed lines represent the repeated DWF value of previous measurements.

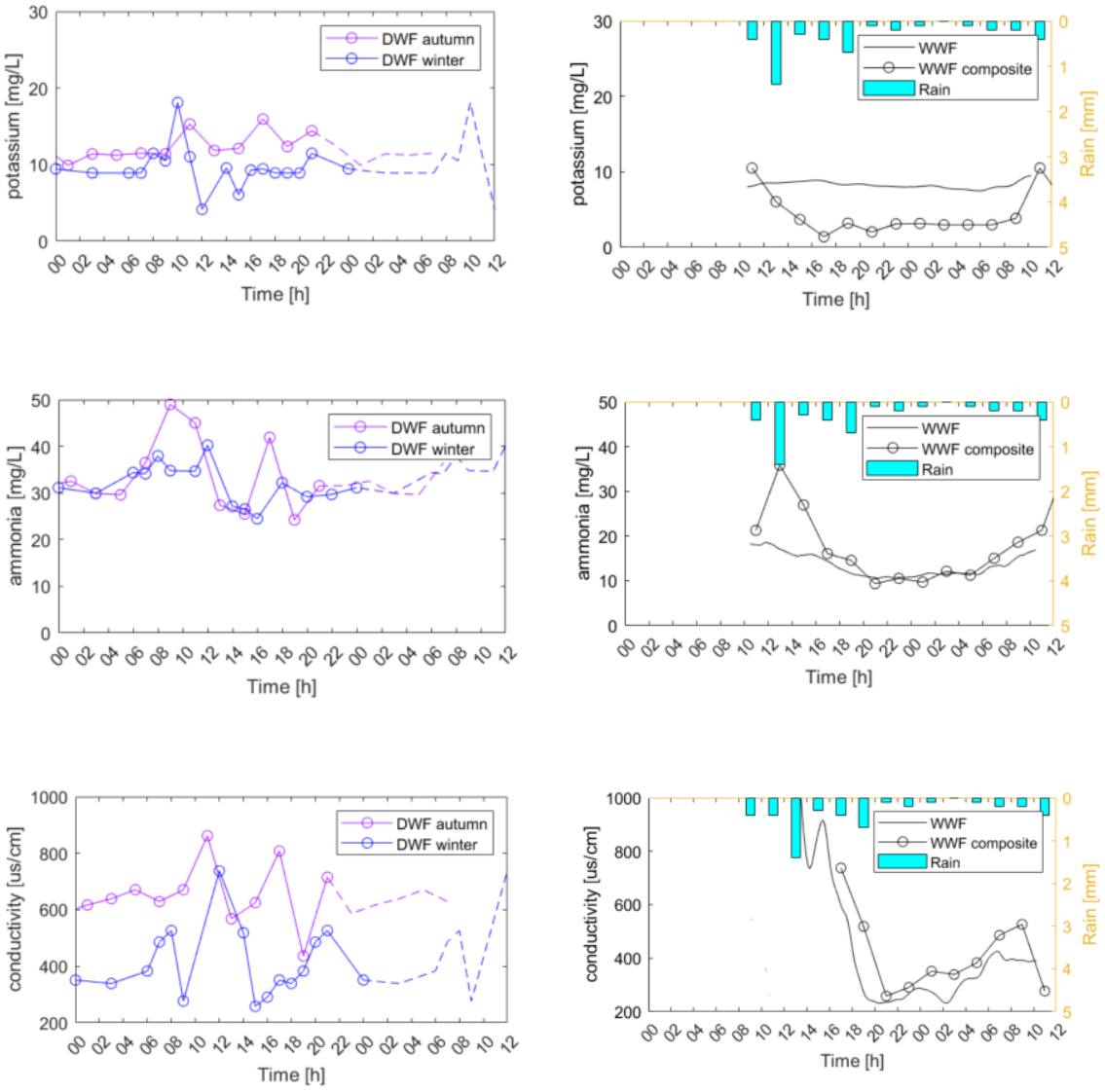


Figure 3-17 Diurnal pattern for potassium, ammonia and conductivity for DWF (measured in autumn and winter) and WWF at the pilEAUte facility. The circles represent the measured value of composite samples, the dashed line for DWF represent the repetition of previous measurements at the same time of the day, and the black line for WWF represents the continuous sensor data.

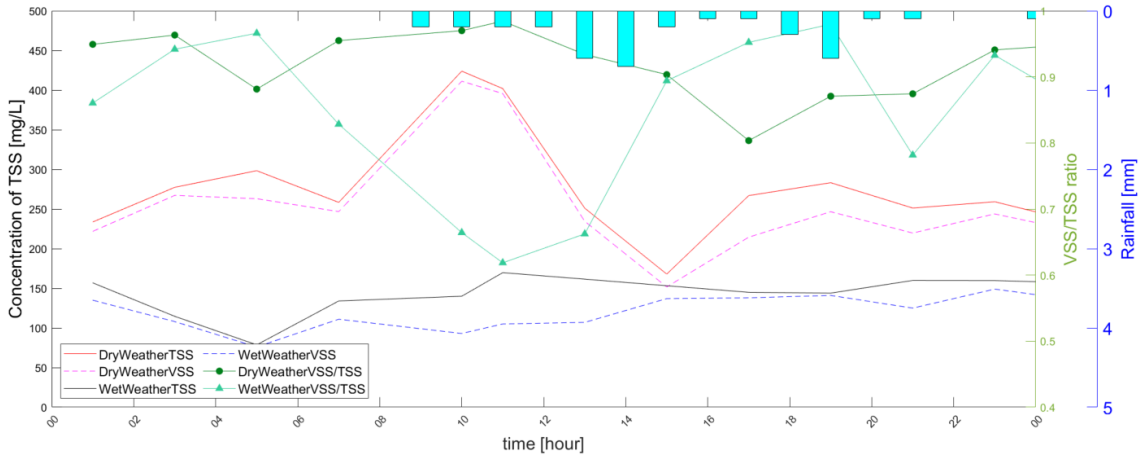


Figure 3-18 Diurnal pattern for TSS, VSS and the VSS/TSS ratio for DWF (measured in autumn and winter) and WWF at the pilEAUte facility

Figure 3-19 focuses on the COD fractions of the wastewater, showing the result of measurement campaign 4 in Table 3-4. Similar to other substances in the wastewater, the concentration of the different COD fractions varies with time. Subfigure (a) represents the COD fractions (left axis) and the soluble and total COD concentrations (right axis) during DWF. A peak COD concentration of the early morning can be observed, confirming the earlier characterization conclusion of Figure 3-15. Noticeably, the COD fractions vary considerably over time. The biodegradable fraction is higher in daytime and it decreases during the night. Consequently, the unbiodegradable COD fraction is increasing at night. Even though the dynamics of the COD fractions are not widely studied (see Chapter 1, section 1.1), it was reported before in literature that the biodegradable fraction is higher in daytime (Zawilski and Brzezinska, 2009).

The daily dynamics of the COD fractions is different from the diurnal COD concentration pattern. It is important to notice the variations of the S_S and X_S fractions, as they indicate the available substrates for heterotrophic growth and have a major influence on the dynamic behaviour of the activated sludge process, such as oxygen demand, exogenous supplemental carbon sources, etc. Thus, by applying dynamic fractionation of the influent data series, the design and process modelling can be improved, and the modelling can be more realistic. Moreover, it may be possible to better operate and control plants with the improved insights in wastewater composition. Further study is surely warranted.

The second subplot (Figure 3-19, b) reveals the influence of rain on the fractions during WWF. The increased inflow leads to the overall dilution of the COD concentration, and an increase of the unbiodegradable fraction. This confirms that the rain can resuspend deposited substances in the sewer and, thus, this modification of the influent composition will lead to a different wastewater treatment process performance.

The readily biodegradable concentration is divided into volatile fatty acids (VFA: such as such as pyruvate, acetate, propionate acid, etc.) and non-VFA components in ASM2 and ASM2d models for biological phosphorus removal. In addition, alkalinity levels are important for nitrification. The lack of alkalinity may result in incomplete nitrification and too low pH values in the treatment process. In order to extend the characterization, the VFA and alkalinity dynamic profile have also been determined by off-line titrimetric measurements (Ponzelli, 2019; Tohidi, 2019).

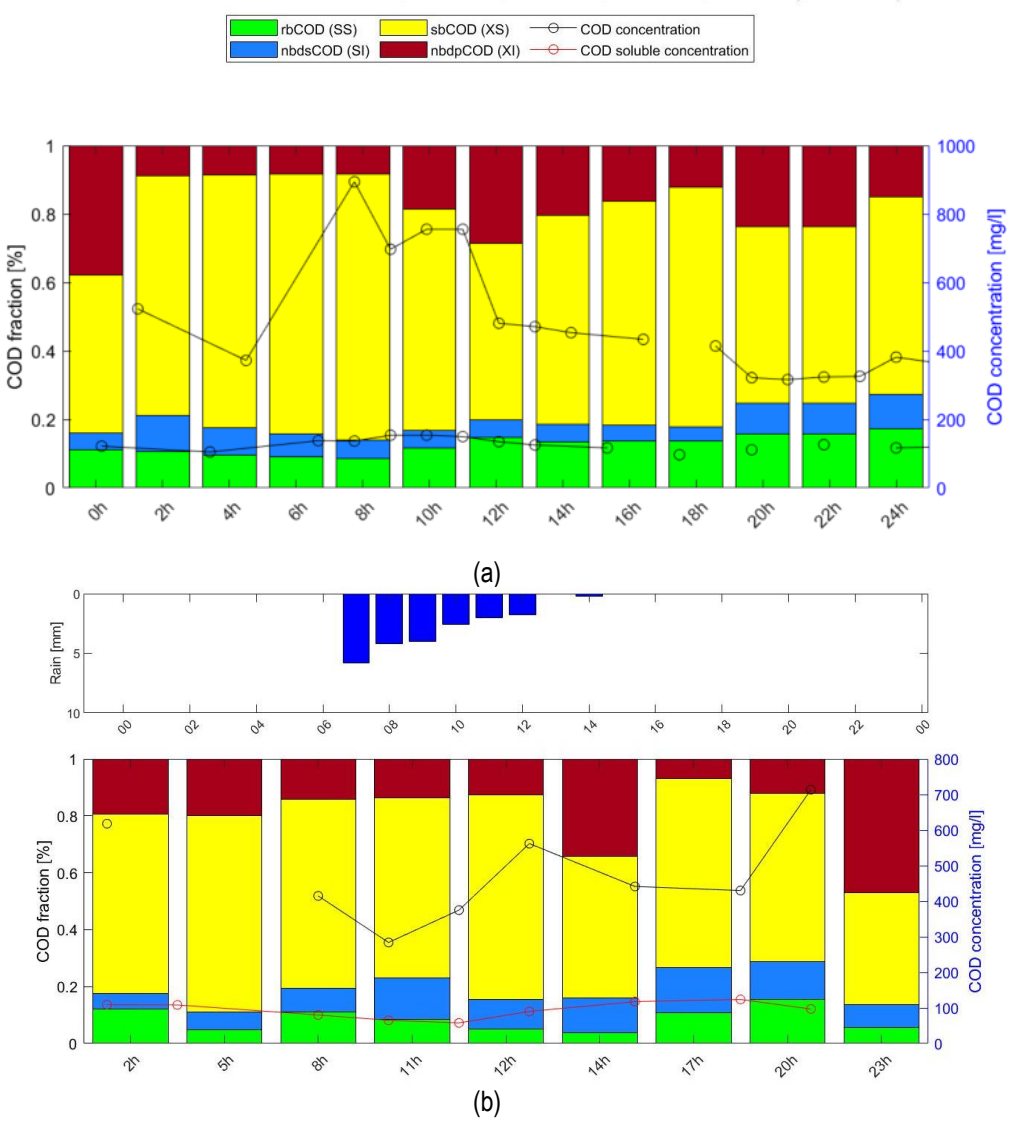


Figure 3-19 COD fractions for DWF (a) and WWF (b) at the pilEAUte facility

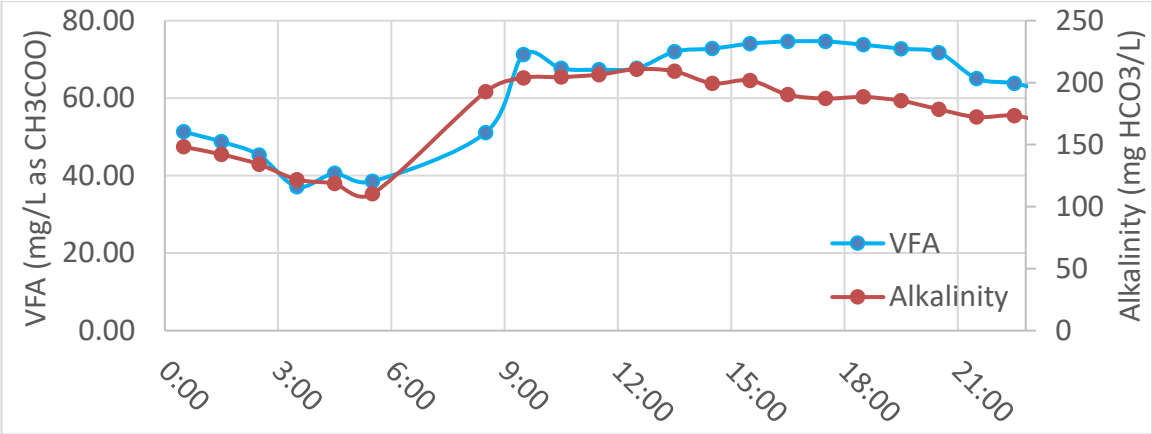


Figure 3-20 Diurnal pattern of VFA and alkalinity concentrations in the pilEAUte under DWF (Ponzelli, 2019; Tohidi, 2019)

Table 3-5 COD fractionation in the pilEAUte study compared to previous studies of municipal wastewater under DWF

Fraction for DW	**Ss**	**Si**	**Xs**	**Xi**	**Biodegradable fraction (Ss+Xs)**
Average result in pilEAUte study	12.5%	6.5%	59.7%	21.3%	60%-86%
(Henze et al., 2000)	25%	10%	40%	15%	65%
(Rössle and Pretorius, 2001)	8%-25%	4%-10%	50%-77%	7%-20%	75%-85%
(Zawilski and Brzezinska, 2009)	6-45%	4-9%	20-70%	4-69%	78%

To evaluate whether the wastewater at the pilEAUte plant is representative of municipal wastewater and to evaluate the quality of the measurement campaigns, typical ratios between pollutant fractions found in literature are compared with the case study result (Table 3-5 and Table 3-6).The results demonstrate that the majority of the influent pollutant fractions of this case study are fitting the typical range. Apparently, the influent composition varies according to different catchment characteristics, as found in other case studies (Choi et al., 2017; Rössle and Pretorius, 2001). Similarly, one can observe the influence of stormwater and other seasonal effects on the influent composition as well (see Figure 3-15-Figure 3-19).

Table 3-6 Typical ratios of municipal wastewater influent compared with the ratios in reference in (Brdjanovic, 2020).

Typical ratios	Case study	(Henze et al., 2008)	(Rieger et al., 2012b)
COD/TN	10.85	8-12	10.5
COD/BOD	2.2	2.0-2.5	2.06
COD/TP	30	35-45	62
COD/VSS	1.85	1.4-1.6	-
VSS/TSS	0.85	0.6-0.8	0.74
VFA*/COD	0.09	0.04-0.08	-

*VFA is mg/L as CH_3COOH

3.5.4 Pollutant modelling results

The pollutant model (Figure 3-8) was calibrated and validated with the series of measurement campaigns, shown in Table 3-3 and Table 3-4.. Figure 3-21 illustrates the modelling results for the validation scenario (measurement campaign 6 in Table 3-4). The model was calibrated for the DWF conditions (measurement campaign 1-3 in Table 3-3, and 1-4 in Table 3-4), for which lab measurements were collected (Figure 3-15-Figure 3-19). The validation scenario is firstly simulated by a steady simulation of 7 days with the DWF diurnal profile, followed by a dynamic simulation with the presence of rain. The first day in the simulation results is for DWF conditions, with the rain starting at 19h. The continuous line represents the simulation results, and the symbols represent the lab measurements.

The dilution effect can be observed clearly for both total COD and soluble COD. In the model, the component XI is washed off the impermeable areas of the catchment at the beginning of the rain event, and, thus, the XI concentration is less diluted than the other substrates. After the rain event, the concentrations return gradually to their DWF concentrations. The observed variation of the COD fractions in wet weather has suggested that the impact of the rain is different for each fraction (biodegradable, non biodegradable, particulate, soluble) because of the resuspension and transport of the solids accumulated in the sewer or washed off from the catchment.

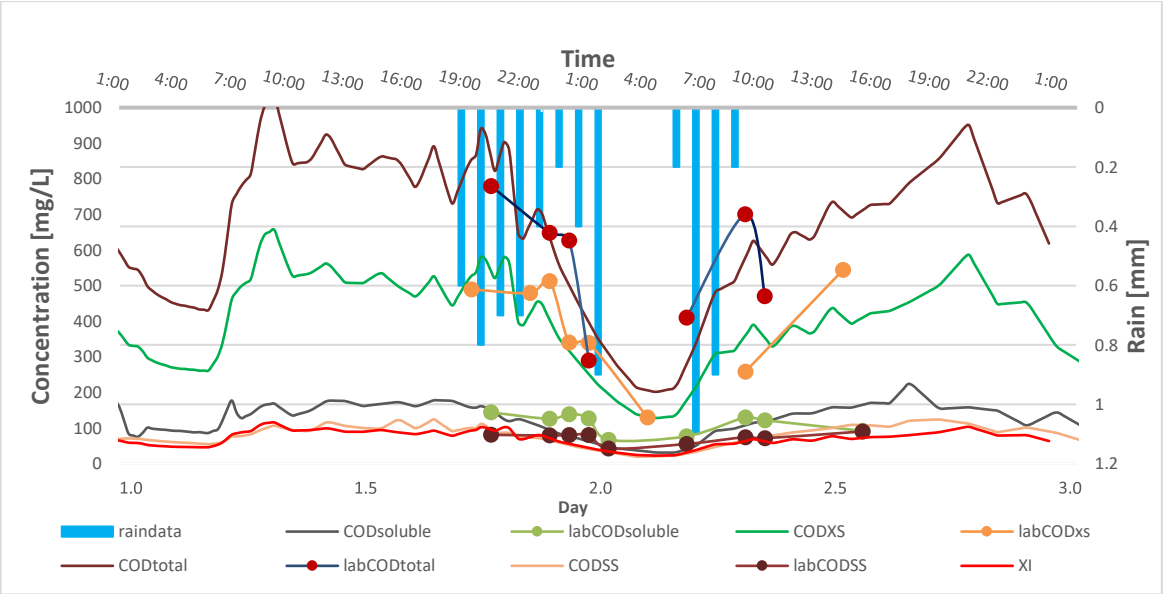


Figure 3-21 Modelling result for the validation set (dynamic simulation) of COD fractions measured under WWF conditions. The model has been initialized by steady simulation with the DWF diurnal profile. The points represent the lab measurements, and the lines represent the model simulation results for the different COD fractions. The rain started from 19h of first days and stopped at 10h of second day.

3.6 Conclusion and perspectives

The pollutant characterization results include the following major pollutants: TSS, COD_{total} and $COD_{soluble}$ (which are the most common parameters to characterize the organic wastewater quality), the nutrient pollutants: ammonia, organic nitrogen and phosphorus and, finally, the conductivity, which is an indicator of the salt content and can inform about the dilution effect during dry and wet weather conditions and also regarding the presence of de-icing agents. The data were collected both by installed sensors and multiple measurement campaigns. U user-friendly interactive dashboard was created to visualize the data in real-time by accessing the dat*EAU*base in which all data are compiled.

Various meteorological conditions have been monitored. Both dry weather and wet weather conditions were surveyed in detail to understand the daily pattern and the impact of rain, respectively. Besides, the difference between a weekday and a weekend day has been studied to understand the weekly pattern as influenced by the population's lifestyle. The biodegradability, COD and nutrient composition were investigated in detail, with a lot of emphasis on COD fractionation.

It is crucial to characterize the dynamics of the pollutant concentration and biodegradability because the wastewater treatment process and operation strategy are highly depending on the influent characteristics. The wastewater composition usually follows a diurnal pattern but is influenced by seasonal and weather conditions. Compared to typical values for domestic wastewater, the study site's wastewater is highly organic, and its biodegradability and TKN/COD ration are in the average range.

In the second part of this study, a conceptual modelling method was proposed for an urban combined sewer system at pilot scale. The proposed influent model was calibrated for the case of the pil*EAU*te facility, considering the DWF pattern and short sewer system effects. This influent model allows the evaluation of dynamic scenarios under different weather conditions. The modelling results showed that the model is able to generate dynamic, and representative data for the catchment by taking into account the rain event effects.

In conclusion, this study applied a protocol for wastewater characterization, developed a visualization tool for the influent dynamics and a conceptual model was calibrated for urban wastewater influent generation in a pilot scale catchment with combined sewer. Through this study, the influent characterization and modelling results could be prepared for wastewater treatment modelling tasks, generating database entries for digital water studies, including digital twin development.

Chapter 4. An essential tool for WRRF modeling: A realistic and complete influent generator for flow rate and water quality based on data-driven methods

This chapter has been published in Water Science & Technology:

Li, F., Vanrolleghem, P.A., 2022. An essential tool for WRRF modelling: a realistic and complete influent generator for flow rate and water quality based on data-driven methods. Water Sci. Technol. wst2022095.

4.1 Abstract

Modelling, automation, and control are widely used for water resource recovery facility (WRRF) optimization. An influent generator (IG) is a model, aiming to provide the flowrate and pollutant concentration dynamics at the inlet of a WRRF for a range of modelling applications. In this study, a data-driven IG model is proposed, based on routine data and weather information, without need for any additional data collection. The model is constructed by an artificial neural network (ANN) and completed with a multivariate regression for certain pollutants. The model is able to generate flowrate and quality (TSS, COD, and nutrients) at different time scales and resolutions (daily or hourly), depending on various user objectives. The model performance is analysed by a series of statistical criteria. It is shown that the model can generate a very reliable dataset for different model applications.

4.2 Résumé

La modélisation, l'automatisation et le contrôle sont largement utilisées dans le domaine d'optimisation de station de récupération des ressources de l'eau (StaRRE). Un générateur d'affluent est un modèle, visée sur fournir un profile dynamique du débit et de la concentration des polluants à l'entrée de StaRRE, pour les applications de modélisation. Dans cette étude, un nouveau générateur d'affluent est proposé, qui demande seulement les données routines et les informations météorologiques, sans avoir besoin de données supplémentaires collection. Ce modèle est construit par réseau de neurones artificiels (ANN) et complété par une régression multivariée pour générer des séries chronologiques des certains polluants. Le modèle est capable de générer des données de débit et de qualité (MES, DCO et nutriments) à différentes horizon temporel et résolutions temporel (quotidiennes ou horaires), en fonction des différents objectifs selon des utilisateurs. La performance du modèle est analysée par une série de critères statistiques. Il démontre que le modèle peut générer un ensemble de données très fiable pour différents modèles applications.

4.3 Introduction

Nowadays, modelling is widely used for water resource recovery facility (WRRF) design, upgrade and controller evaluation (Sweeney and Kabouris, 2013). However, because of a lack of adequate (dynamic) input datasets, most of the models are still used under steady state conditions. For example, for WRRF design, engineers usually make an initial sizing by using design guidelines and safety factors or some statistical evaluations to come up with the values for the design inputs, and the consequence of overly conservative safety factors is oversizing the WRRF (Talebizadeh, 2015). However, steady state simulations are not able to represent the temporal variability present in reality. Thus, a reliable dynamic influent generator (IG) is essential as key input for WRRF modelling (Gernaey et al., 2011; Martin and Vanrolleghem, 2014).

An Influent Generator (IG) is a model that generates realistic dynamic (or static) influent scenarios at the inlet of a WRRF. The outputs of an IG include flowrates and pollutant concentrations at different time scales (long-term, short-term) and resolutions (daily, hourly), according to different user objectives.

An IG model can be applied in very diverse domains, such as the integrated urban wastewater systems (IUWS) to globally improve the wastewater treatment process (Benedetti et al., 2013), and WRRF design and upgrade to face the growing amounts of wastewater produced by increasing urbanization and population. IGs can help to provide a fast and accurate estimation for WRRF designers and a credible input time series to anticipate the treatment performance under different operating conditions.

An IG model can also be used for influent database quality evaluation, optimization and completion (Martin and Vanrolleghem, 2014). Since data observation and collection become more and more important in wastewater management, an IG model enables quality evaluation of a dataset collected from online measurements in different ways, e.g. by identifying errors due to clogged sensors, detecting extreme measurement values (outliers), etc. An IG can also complete missing data (gap filling, i.e. temporary measurement failure) and interpolate a low frequency time-series into a denser time-series.

Different researchers (see below) have been developing IGs based on different modelling principles. The current IGs can be divided into two types: data-driven IGs and phenomenological IGs, also called ‘black’ box and ‘grey’ box, respectively. Data-driven IGs focus on finding the relation between their inputs and outputs, without any knowledge of the internals of the system. Therefore, the performance of data-driven IG depends on the provided dataset. Varying degrees of model complexity have been studied, such as harmonic function models, regression equations, and also artificial neural networks (ANN). The Fourier series IG based on harmonic function has been used to develop a simple and reliable generator of diurnal variations for dry weather influent data, and has been used in different studies (Langergraber et al., 2008; G Mannina et al., 2011). Ahnert et al. (2016) developed a

statistical method for the generation of a continuous time series of influent quality by only using routine data based on the Weibull-distribution. Troutman et al. (2017) developed an automated toolchain based on Gaussian processes to predict the dynamics of combined sewer system flow. Recently, machine learning is also being studied for flow forecasting over a short time horizon, with model structures such as artificial neural networks (ANN) (El-Din and Smith, 2002), multiple linear regression (Zhu and Anderson, 2016) or nonlinear autoregressive exogenous models (NARX) (Banihabib et al., 2019).

Unlike data-driven approaches, the phenomenological model is built by integrating some important processes, which influence the generation of the influent. This type of model is usually constructed around different submodels, such as dry weather generation, sewer system transport, soil model including infiltration of groundwater etc. (Gernaey et al., 2011). For example, the dry weather flowrate has a diurnal pattern with two peaks corresponding to the morning and evening urban activities, e.g. characterized by Butler et al., (1995). This diurnal pattern can reflect different lifestyles and can be described by a harmonic function. For storm flow, the dilution and first flush during a rain event can be integrated in the model (Bechmann et al., 1999; Jeroen Langeveld et al., 2017; Talebizadeh et al., 2016). Gernaey et al. (2011) developed and Flores-Alsina et al. (2014) extended a more conceptual phenomenological model, by focusing on flowrate scenarios and sub-models of the urban drainage system.

There are advantages and disadvantages in both types of models (Price and Vojinovic, 2011). Compared with a data-driven model, a phenomenological model contains more details of the influent generation process, which leads to an explicit result and good extrapolation power beyond the calibration range. Normally, such model consists of submodels with parameters related to physical processes. One of the shortcomings of such IG comes from the need of calibration, because of the series of parameters and submodels, it is difficult to calibrate all parameters needed, such as the catchment information, sewer system etc., which leads to less flexibility when applying to a different case study. In contrast, a data-driven model focuses only on the relation between input and output, instead of providing more understanding of the system. Thus, it is easier to calibrate and use a data-driven model. However, the high dependence on data demands a more complete dataset, in order to reach adequate model quality.

Currently, most influent generators in literature (either phenomenological or data-driven) suffer from the following two issues. The first issue is that it is difficult to balance the complexity and precision of the IG. Usually, a higher complexity leads to better performance, but it makes the modelling more time consuming and more intense in terms of calculation. The other issue is the ability to generate an adaptable influent profile, that can be easily adjusted to different user objectives. For example, the design process demands a long-term series but allows for low time resolution, while a process control application needs a higher time resolution profile.

To solve the issues with the available IG models, in this study, a data-driven IG is proposed that can generate dynamic WRRF influents of different time scales and resolutions. The model aims to generate reliable influent properties and more complete time series for the influent dynamics, including daily and hourly flowrate, daily and hourly total suspended solids (TSS), and daily chemical oxygen demand (COD) and nutrient concentrations for the case study at hand. With the available daily concentration results the IG is constructed such that it is able to obtain hourly concentration dynamics by applying an observed daily pattern or by developing a more detailed model in future studies.

The IG proposed in this work is intended to generate influent time series only based on weather information, without any further collection of flowrate or concentration data. This allows obtaining long-term simulation inputs and since it does not need a complicated dataset, it reduces the investment of labor and money for measurement campaigns. The proposed IG model is not built around a physical sub-model, resulting in lower modelling efforts than with a phenomenological model, so that it can provide a satisfactory compromise between model complexity and prediction quality.

4.4 Materials and methods

This section describes the case study and the preparation of the data set, followed by an overview of the modelling approach. First, the basic ANN model is presented followed by the stochastic process model that is added to increase the generated variability. Subsequently, the multivariate regression that allows calculating the nutrient time series is introduced. This section finishes with a definition of the criteria that will be used to assess the quality of the proposed IG.

4.4.1 Case studies and dataset preparation

The modelling approach was developed and tested on two urban catchments, Quebec City (Canada) (Tik and Vanrolleghem, 2017) and Bordeaux, Clos de Hilde (France) (J. M. Ledergerber et al., 2020). These two case studies are both combined sewer systems, with a similar number of population equivalents, 300 000 PE and 200 000 PE, respectively. However, the two catchments provide different extents (range, data frequency) of available data.

In this research, the Quebec City case study is developed around routine influent data that are collected regularly at the entry of the WRRF, including flowrate, COD, TSS. The nutrient concentrations (ammonia and phosphorus) are sampled and measured only once or twice per week since the plant is only required to remove organic matter. Data were available for 8 years (2011-2018). For the Bordeaux case study, which is also a carbon removal WRRF, hourly flowrate and TSS concentrations were available for three months in 2017 and 2018.

4.4.2 Modelling approach

The proposed IG was developed with MATLAB R2019b (www.mathworks.com) using the following machine learning methodology. In short, the dataflow through the model is according to the flowchart shown in Figure 4-1, detailed in the following sections. Figure 4-1 shows that the raw data series is pre-treated. The pre-treatment aims to remove outliers, which are defined as aberrant data, representing values or observations that differ significantly from other observations and cannot be expected to represent reality. During pre-treatment, the outliers are detected by a univariate method and the faulty data are replaced by estimated values obtained by the Gaussian kernel smoother (Alferes and Vanrolleghem, 2016).

Then the cleaned flowrate data is Fourier transformed to obtain a yearly-seasonal-daily pattern for the baseflow P(t). The input dataset of the ANN consists of this P(t) together with weather data (temperature T and rainfall intensity R, provided by the WRRF's pluviometry and treated at hourly frequency), creating input vectors with a number of lags τ, which represents the weather data in the previous τ time steps. A stochastic generator is subsequently applied on the ANN's output, to get a time series, which better mimics the reality in terms of variability (Q_{sim}). Finally, the nutrient concentrations ($C_{nutrisim}$) are obtained by applying a multivariate regression with input of the ANN's results and the weather data time series.

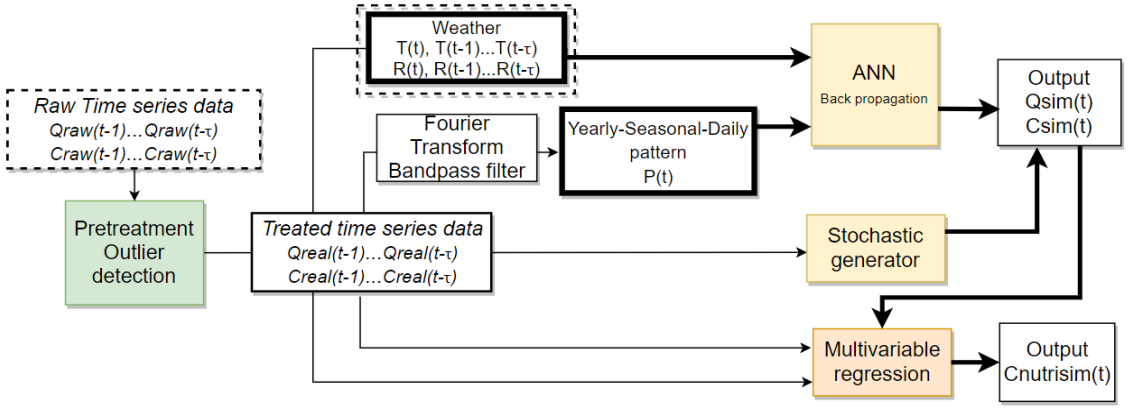


Figure 4-1 Description of the IG model and interaction between ANN, stochastic generator, and multivariate regression sub-models. The dashed line boxes represent the real data needed for training and the bold line boxes represent the model input for new influents

- ANN model

The artificial neural network takes inspiration from the biological learning process. As a powerful data-driven tool, it can simulate nonlinear systems and is increasingly used in water engineering, urban hydrology and catchment modelling (Fu et al., 2010; Maier and Dandy, 2000; Rajurkar et al., 2004).

Input data selection is important before using any ANN model. For a combined sewer system, the dry weather flow (DWF) component varies with the pattern of urban behaviour and is completed with hydrological processes (snowmelt, groundwater) and wet weather flow (WWF) also including direct storm water inflow, and rainfall-dependent infiltration and inflow (RDII) (Wright et al., 2001). Thus, the input of the ANN-based IG consists of a basic domestic flowrate pattern, weather information, including temperature and rainfall measurements with a number of lags τ time steps, to represent an internal auto-regression.

The DWF pattern is obtained by applying a Chebyshev bandpass filter after Fourier transform (Heideman et al., 1984). The Fourier transform enables converting the signal from the time domain into the frequency domain, and the Chebyshev bandpass filter allows only the signal with selected frequencies to pass (Schlichthärle, 2011). By removing high-frequency contributions such as measurement noise, the dry weather flow pattern can be obtained. In this study, the bandpass filter is focusing on extracting two major signals: the seasonal/yearly effect and the daily effect. To this end, the frequencies that were retained are: 4 $year^{-1}$ and 2 $year^{-1}$ for the seasonality and 1 day^{-1} and 2 day^{-1} for the diurnal phenomena.

As activation function in the ANN's hidden layers, the sigmoid function was selected to capture the non-linearities. The output layer uses a simple linear activation function. In order to minimize the loss function, the Levenberg–Marquardt backpropagation algorithm (Levenberg, 1944) was used for training, because it is fast and easy to use and has been a first-choice in supervised search algorithms. In this study, the performance evaluation is focused on measuring the residual between the model result and target value, therefore, the root mean square of errors (RMSE) was used, one of the most used loss functions in regression problems in ANN learning:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{oi} - y_{si})^2} \qquad 4\text{-}1$$

where, y_{oi} and y_{si} represent observed data and simulated data, respectively.

The dataset was divided into a training set, a cross-validation set and a test set. Based on the common basics of machine learning, the dataset is split with a ratio of 70%, 15% and 15%, respectively. To determine the architecture of the ANN (hyper-parameters: number of neurons and layers), a series of ANNs with different numbers of layers and neurons were trained. For each training, the performance of the model was recorded and compared. The number of layers and neurons giving the best performance for the cross-validation dataset, was

selected as final architecture. To avoid local minima during training, a series of iterations of the training procedure was applied. The final test of the ANN model was performed with the test set.

- Stochastic process

The time series model can be decomposed into a deterministic part and a stochastic part with an auto-correlated error term. After obtaining the ANN-only result (i.e. without the stochastic part), the residuals between the measurement and the IG output have been analysed. It was observed that the model output exhibited less variability than observed in the measurements. However, it is important to have a probability distribution of the generated time series similar to reality, i.e. this allows especially to better design WRRF parameters such as the expected load and hydraulic capacity and certain percentiles of their distribution etc.

Therefore, a stochastic process was added in order to optimize the IG model adequacy. The aim was to simulate a more random time series, with a statistical characterization that is more reflecting the distribution of the measured reality.

Theorical analysis shows that adding an auto-correlation model residual can improve model reliability (Villez et al., 2020). Therefore, a stochastic process was added that is modelled by a k-order random-walk based on the error between the ANN output and the measurements:

$$\boldsymbol{R_t = \sum_1^k \varphi_k * R_{t-k} + \varepsilon} \qquad \textbf{4-2}$$

where R_t is the error value at time t, corresponding to the difference between the ANN model and the actual data, φ_k are the model's k coefficients and ε is a white noise sequence with a Gaussian distribution and the standard derivation of the white noise sequence was pre-defined as 5% of the error between ANN output and the measurements, which turned out to lead to good results.

- Multivariate regression

Experiments show that nutrients (such as ammonia) are highly correlated with the flowrate and COD concentrations (Petersen et al., 2002). Thus, it is proposed that nutrient data can be generated by a multivariate regression using the COD and flowrate results of the ANN together with weather information as additional input variable. The adopted multivariate polynomial regression aims to calculate the nutrient concentration time series by using its relationship with other variables.

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{k1} & & {x_{11}}^2 & & {x_{k1}}^k \\ 1 & x_{12} & x_{22} & x_{k2} & \dots\dots & {x_{12}}^2 & \dots\dots & {x_{k2}}^k \\ \dots & \dots & \dots & \dots & & \dots & & \dots \\ 1 & x_{1m} & x_{2m} & x_{km} & & {x_{1m}}^2 & & {x_{km}}^k \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_{..} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{\dots} \end{bmatrix} \quad 4\text{-}3$$

where y_i are the target variables, x_{ij} are the multivariable data, k refers to the k[th]-order of the polynomial, θn are the regression coefficients, and ε is white noise. To avoid overfitting during higher order regression (when k increases), the Lasso (Least Absolute Shrinkage and Selection Operator) regression regularization method was used to determine the optimal model complexity and regression parameters (Tibshirani, 1996).

$$\theta^{lasso} = argmin\{\sum_{i=1}^{m}(y_i - \theta_0 - \sum_{j=1}^{n} x_{ij}\theta_j)^2 + \lambda \sum_{j=1}^{n}|\theta_j|\} \quad 4\text{-}4$$

where λ is a pre-specified parameter, determining the amount of regularization and θ_j are coefficients of regression. Thus, the Lasso method solves the minimization problem with respect to the MSE, under the condition of minimizing the model parameters.

4.4.3 Criteria and error analysis

The model was optimized for MSE during the training procedure. In addition, the model performance was also evaluated by the mean absolute percentage error (MAPE) and the Nash-Sutcliffe efficiency (NSE), expressed as

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Q_s^t - Q_o^t)^2}{\sum_{t=1}^{T}(Q_o^t - \overline{Q_o})^2} \quad 4\text{-}5$$

where, Q_s is the simulated data and Q_o is the observed data for the variable under consideration. The optimal NSE value equals one, corresponding to a perfect match of modelled and observed data, and zero indicates that the model is equally good as a mean value model. A negative value means that the model is performing worse than the mean value of the observations (Hauduc et al., 2015).

In addition, a frequency histogram of the generated time series, the probability density function (PDF) and the cumulative density function (CDF) were created to compare the statistical characteristics of the measured and the simulated time series. The statistical characterization allows evaluating whether the model output can well represent the real dataset. In addition, the Kullback-Leibler divergency (KL divergency) (Kullback and Leibler, 1951) was calculated, in order to measure the difference between the probability distributions of model output and measurements, which is expressed as

$$D_{kl}(Q||P) = \sum_{x\in\chi} P(x)\ln\left(\frac{Q(x)}{P(x)}\right) \qquad 4\text{-}6$$

where, P(x) and Q(x) represent the probability distribution of the model output and the real measurement, respectively, and P(x) > 0 and Q(x) > 0 for any x in χ, which is the distribution space.

4.5 Results and discussion

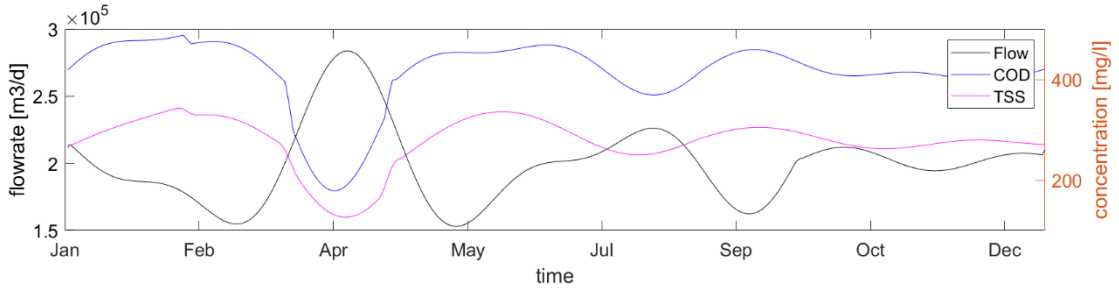
In this section, first, the result of each of the IG's sub-models is presented, including the daily and seasonal pattern, and the ANN modelling results of flowrate and pollutant concentrations for both Quebec City and Bordeaux. Next, attention is focused on the usefulness of adding a stochastic process to increase the variability of the generated time series. Subsequently, the result of the nutrient concentration calculations is shown. Then, the model performance is analyzed with two other criteria. Finally, this section ends with a detailed evaluation of the proposed IG from an overall perspective.

4.5.1 Model and submodel results

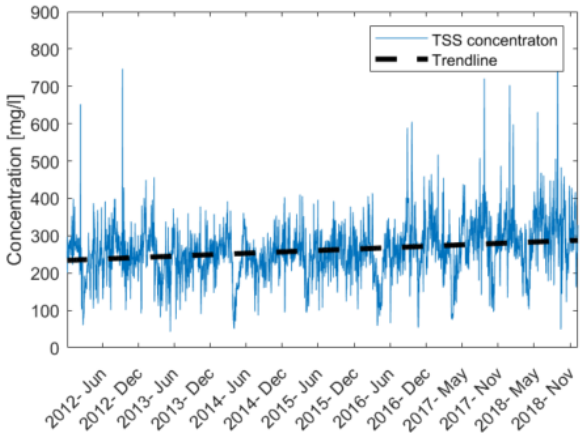
- ***Daily and seasonal pattern***

A Chebyshev bandpass filter (2nd order Chebyshev Type I) was applied after the Fourier transform of each time series to get the seasonal and daily DWF pattern of the urban activity. The Fourier transform allows to extract properties from the time series from the time-domain into the frequency domain, which is known as a powerful tool for extracting different periodic patterns from the time series. After the Fourier transform, the typical frequencies have been identified and a Chebyshev bandpass filter with the extracted frequencies were applied to obtain the pattern. Figure 4-2 shows the pattern for the Quebec City case study. The flowrate pattern exhibits a clear increase during the snowmelt from March to the end of April, while the COD and TSS patterns demonstrate dilution by the snowmelt. Other periodicities observed throughout the year could be explained by the variation of the groundwater level, which influences the infiltration into the sewer system, or the urban activities, such as the holiday periods, etc. The TSS daily concentration shows a slight trend from 2011 to 2018. The correlation coefficient is equal to 0.24, indicating a weak trendline. There are two explanations for this phenomenon: (i) an increase of the urbanization brings increasing TSS discharges and (ii) thanks to the

improvements made to the sewer system, the infiltration inflow has decreased, which leads to a higher TSS concentration.



(a)



(b)

Figure 4-2 (a)Yearly pattern for flowrate, COD and TSS based on daily data for the Quebec City case study of 2011-2018, (b) yearly trend in TSS concentration from 2011 to 2018

Similarly, the 2nd order Chebyshev Type I bandpass filter is applied for the Bordeaux case study, in order to extract the daily pattern. Figure 4-3 illustrates that the daily pattern has two peaks corresponding to the increased urban activity in the morning and evening, as reported elsewhere in the literature (Butler et al., 1995). However, the peak and form are different for weekday and weekend days, because of the different weekend behaviour.

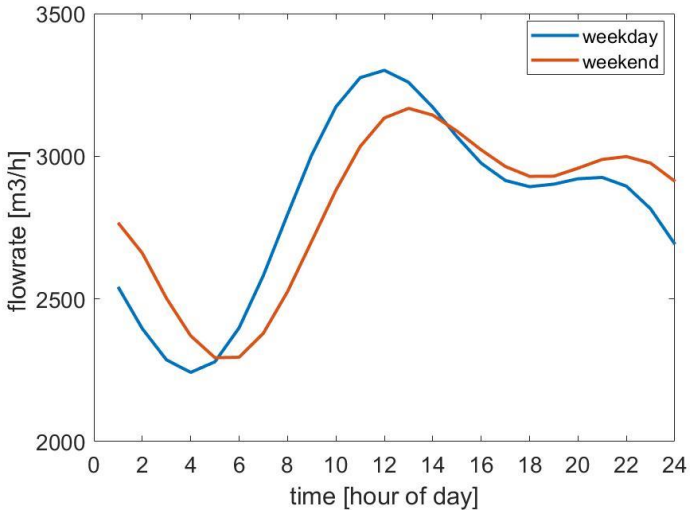


Figure 4-3 Daily pattern for flowrate in Bordeaux: hourly flowrate for weekday and weekend days based on the hourly data for May to August, 2018,

- ***ANN modelling result for Quebec City***

The Quebec City case study aimed at generating the flowrate, COD and TSS time series using an ANN model. The selected ANN has one input layer, one hidden layer and one output layer. The hidden layer contains four neurons for flowrate and five neurons for COD and TSS. Figure 4-4 shows the daily flowrate, TSS and COD concentration in the test set. The result clearly demonstrates the snowmelt effect at the end of winter, with an increase of the flowrate and dilution of TSS and COD. The impact of each rain event is adequately described by the IG model.

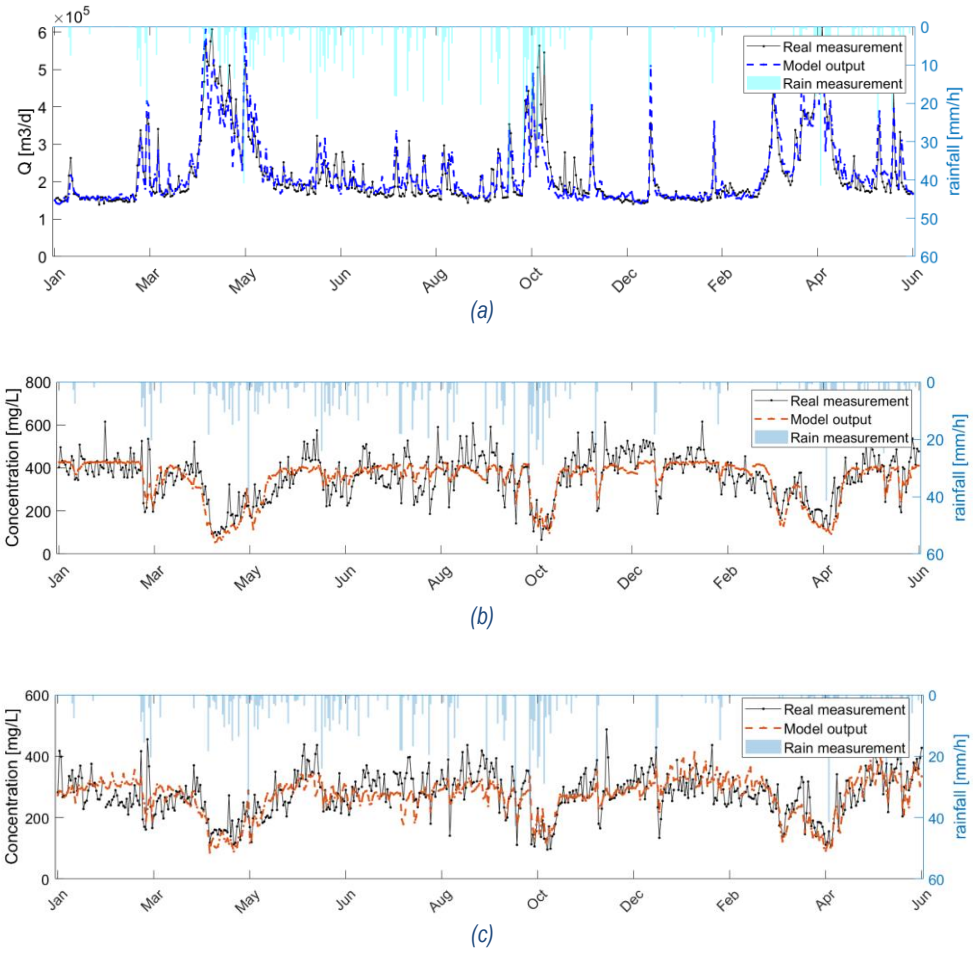


Figure 4-4 IG model output of the test set for the Quebec City case study: (a) daily flowrate, (b) COD concentration, (c) TSS concentration. Measurement were collected in 2017 and 2018.

- ***ANN modelling result for Bordeaux***

The ANN model also demonstrated good results for Bordeaux, as shown in Figure 4-5 for the hourly flowrate (a) and hourly TSS (b). The flowrate is generated by using a daily dry weather flow pattern and hourly rain data as input. The daily dry weather flow pattern profile and the increase of inflow by storm water can both be seen for each rain event. The TSS results demonstrate that the data-driven methodology is also able to describe water quality at high frequency. The TSS concentration follows a daily concentration pattern and it is diluted by the inflow of storm water.

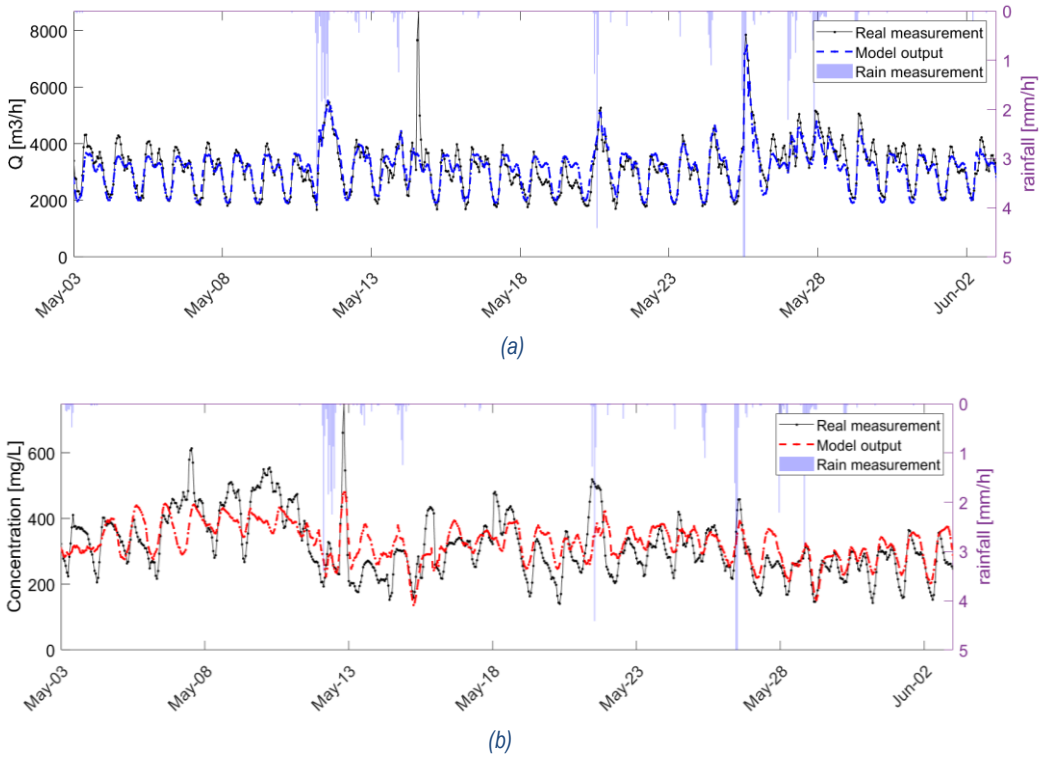


Figure 4-5 Hourly flowrate (a)and TSS concentration (b) generation by the IG model for the Bordeaux case study

The ability of the IG model to generate concentration time series can also be used for dataset gap filling, especially under wet weather conditions when sensor clogging occurs more frequently. Moreover, as shown in Figure 4-6, the model can be improved by also feeding it with daily average lab measurements as input. Although the model output is still less dynamic than what the sensor data exhibit, this extra input enables the model to better represent the variability of reality (by representing the differences of wastewater discharge on different days). For instance, thanks to this add-on, the lower concentrations on 13 and 14 June could be captured even though there was no wet weather effect. Finally, note that a longer time series will help to find a yearly or seasonal pattern so that the TSS concentration results would be further improved.

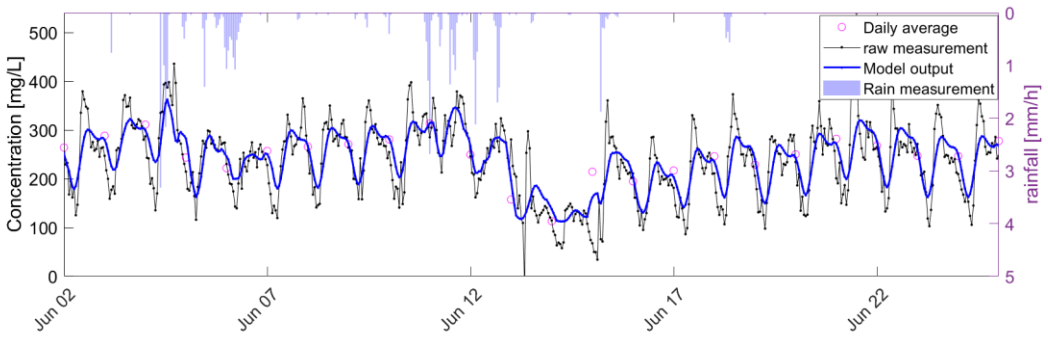


Figure 4-6 Hourly TSS generation improved by adding daily average measurements

- ***Better representing variability by adding a stochastic process***

As mentioned before in the Methods section Stochastic process, the obtained ANN model was extended with a random walk model after having analyzed the error between the ANN model output and the measurement data. As explained before, the residual between the IG-model output and the measured data is autocorrelated, because the input data to the ANN model do not include the target variable itself and the model didn't capture all the system dynamics. Therefore, to reconstruct this autocorrelation information and thus to better describe the stochastic properties of the system, the stochastic process is added, which allows the influent generator to statistically better approach the variability in the real measurements.

The random walk model is a special case of the autoregressive (AR) model. In order to define the order of the autoregression model, orders 1 to 5 time lags were evaluated for the Quebec City case study. As the autocorrelation coefficient plot and the MAE results of Figure 4-7 show, the best stochastic model is obtained at 4 time lags autoregression because this gave the best compromise between model complexity and precision.

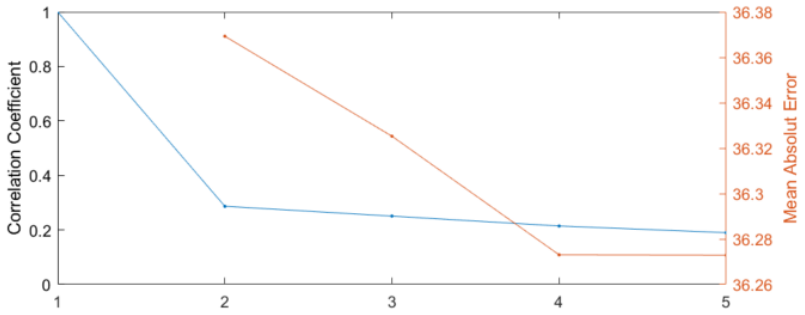


Figure 4-7 The autocorrelation coefficient plot and the MAE for each time lag for the Quebec City case study.

Figure 4-8 shows the results of COD and TSS generation with addition of the stochastic process. Compared with the ANN-only model, the stochastic process allows to better mimic the influent random variation. A more detailed analysis is provided in the performance analysis section.

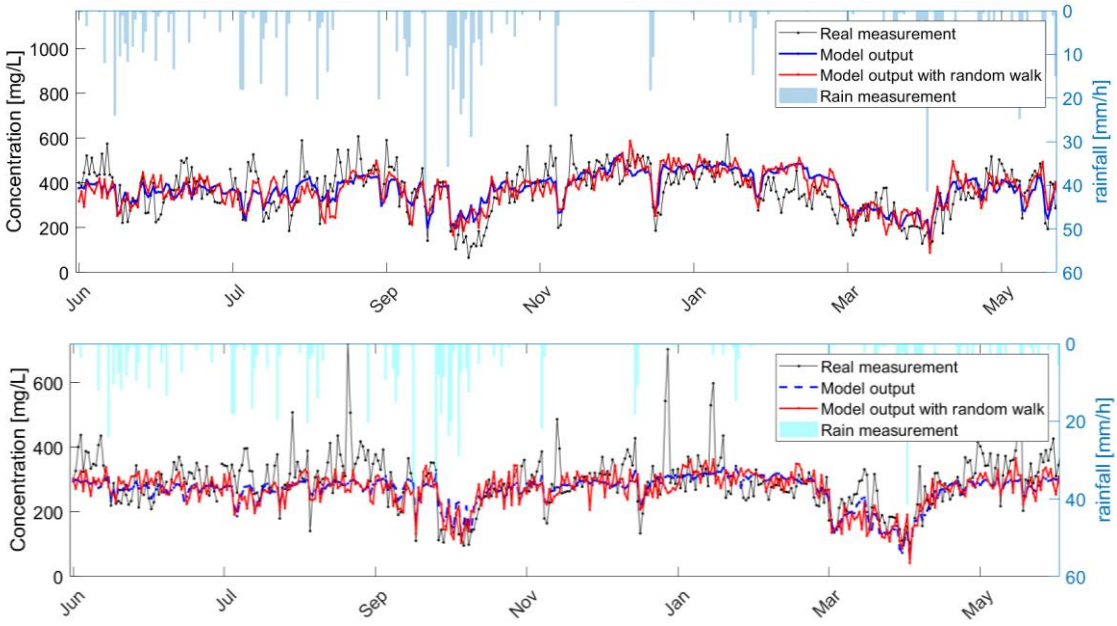


Figure 4-8 COD (top) and TSS (bottom) concentrations generated for the Quebec City case study by the ANN with and without stochastic process extension.

- ***Nutrient concentration generation by multivariate regression***

For the reasons explained before, in practice, the influent nutrient concentrations (ammonia and phosphorus) in Quebec City's COD removing plant are measured only once or twice a week. This makes it difficult to estimate the current treatment result from the limited nutrient data or to create an influent time series in view of upgrading

the plant for nutrient removal. Thanks to the high correlation between the nutrient concentrations and the flowrate, it was found that the concentrations could be generated quite well by multivariate regression, see equation (2). Based on the same principles of regression modelling as used in section 3.1.4, regression orders between 1 and 5 were tested and the best order of the regression was found to be 3, giving the best performance (RMSE) on both training set and validation set.

Figure 4-9 shows the model performance for the test set, demonstrating that the regression is able to adequately describe the ammonia concentrations on the basis of only flowrate and weather data.

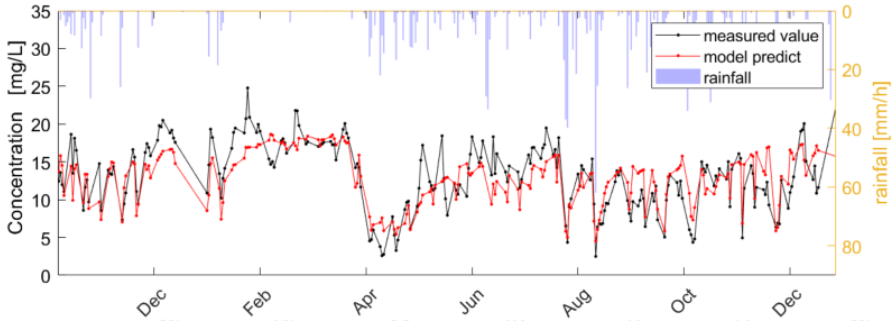


Figure 4-9 Ammonia concentration time series generation for Quebec City's WRRF by a 3rd order multivariate regression for the test set.

In the same way, Figure 4-10 presents the generated phosphorus concentration time series. Thus, the multivariate model enables generating a high frequency time series (daily simulation) from a low frequency measurement time series (weekly measurement).

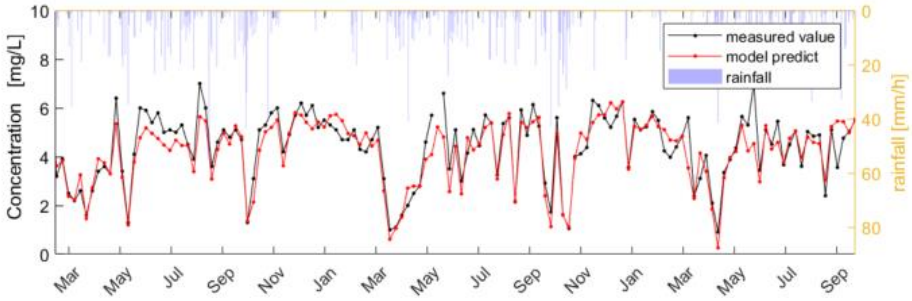


Figure 4-10 Phosphorus concentration time series generation for Quebec City's WRRF, by a 3rd order multivariate regression for the test set.

4.5.2 Model and submodel results

The performance of the final data-driven influent generation is summarized in Table 4-1. The criteria are all calculated based on the test sets. Table 4-1 demonstrates that the model is able to generate time series with an error of around 10% for flowrate and 13-20% for water quality. The NSE is around 0.5 to 0.7 indicating that the

model can match the observed dataset well. For water quality, the model is able to provide a good generation, too. However, it is worth remembering that the observed data consist of imperfect lab and sensor measurements and that these errors thus indirectly influence the quality of the water quality generation. Considering the high uncertainty and measurement errors of raw wastewater data (Bertrand-Krajewski et al., 2007; Montgomery and Sanders, 1986), the RMSE and MAPE obtained can be considered to be in the same order of magnitude as real-life measurements. These results thus indicate that the model performance is sufficiently good.

Table 4-1 Model performance for test set, evaluated by MAPE, RMSE and NSE

Variable	Average	RMSE	MAPE	NSE
Case study for Quebec City: daily data				
Flowrate	200 000 m^3/d	38 000	11.8%	0.74
COD	400 mg/L	52	19.8%	0.43
TSS	250mg/L	30	16.7%	0.59
Ammonia	12.5 mg/L	2.6	13.7%	0.68
Phosphorus	4.3 mg/L	0.8	13.4%	0.71
Case study for Bordeaux: hourly data				
Flowrate	12 000 m^3/h	1 200	13.5%	0.61
TSS	300mg/L	42	17.5%	0.70

Figure 4-11 compares the observed flow data with the model generated data in different ways. First the quantile-quantile (q-q) plot, see Figure 4-11 (a), shows the good correspondence of the generated data with the observation data (with coefficient of determination $r^2 = 0.88$). The CDF in Figure 4-11 (b) compares the statistical characteristics of the observed and generated dataset. The time series was also split into two subsets, one for winter (from January to May), and one for summer (from July to December) in order to demonstrate that the model's performance is not different for different seasons, see Figure 4-11 (c) and (d). Although the winter and summer flow distributions are different, i.e. the winter flow distribution is more dispersed than the summer one: in winter 85% of the flow rate is below $3.8*10^5$ m^3/d while in summer 85% of flow is below $2.4*10^5$ m^3/d, both the seasonal PDF and CDF analysis show that the distribution of the model outputs is similar to the real data distribution.

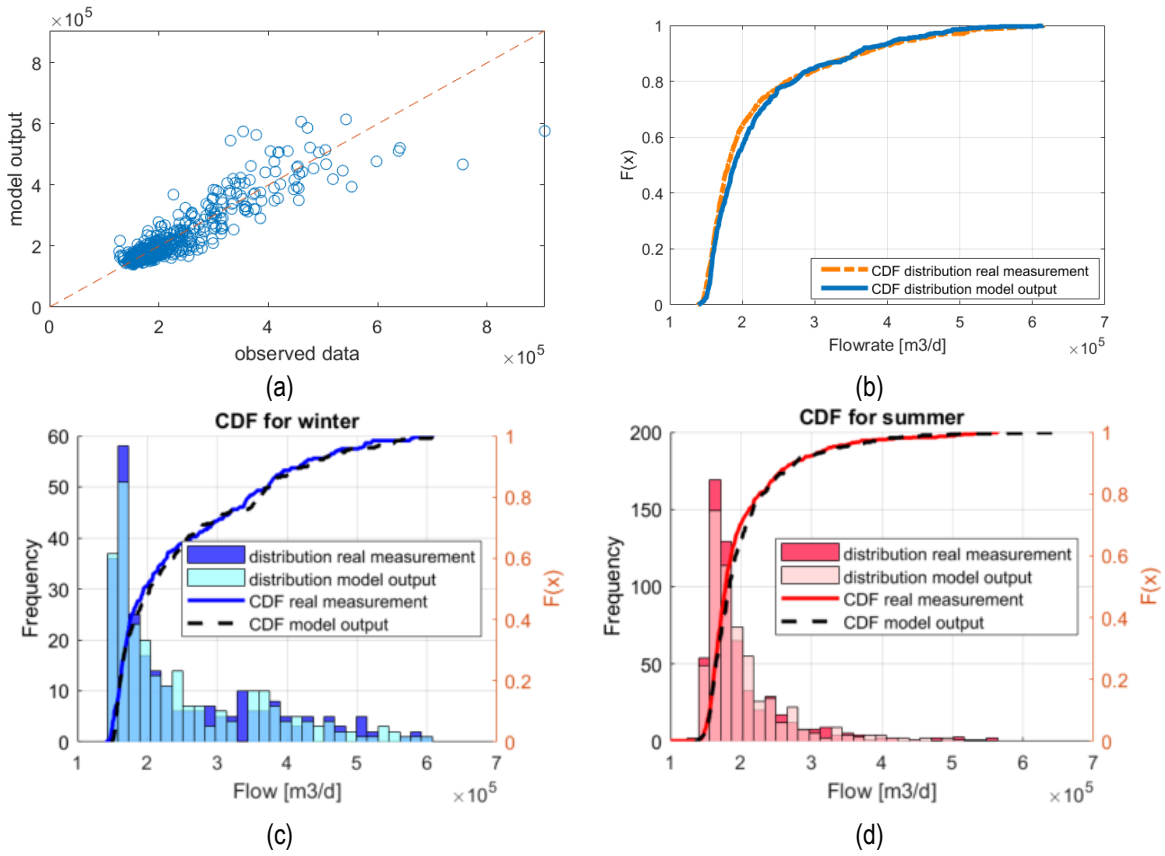


Figure 4-11 Model performance analysis for flow: (a) quantile-quantile plot for observed data and model output, (b) CDF for the complete test set. (c) PDF and CDF for flow data in winter and (d) in summer.

Figure 4-12 illustrates the COD concentration generation by the ANN model without (a) and with stochastic process extension (b). It can be concluded that the ANN model with the stochastic process can better mimic the observed variability. The stochastic process enables the model output distribution to be wider and closer to the distribution of the real measurement time series.

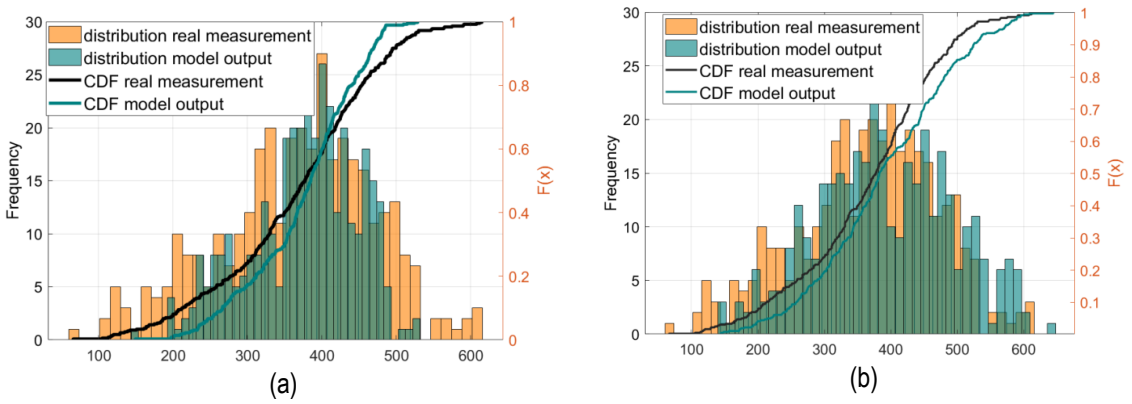


Figure 4-12 PDF and CDF of the COD concentrations for the Quebec City test set without stochastic process (a) and with stochastic process (b)

This improvement is confirmed by the KL divergence values Table 4-2. A smaller KL value represents a better similarity between two distributions: i.e. the KL divergence value with random walk is smaller than the KL value for the ANN model output without the extension, which indicates it is closer to the distribution of the observations.

Table 4-2 KL divergence calculation for COD and TSS concentrations generated with and without stochastic process extension for the Quebec City test set

KL divergence	COD	TSS
ANN model (a)	0.573	0.204
ANN model with stochastic process extension (b)	0.115	0.058

4.5.3 Discussion and evaluation

In general, the error analysis (Table 4-1) shows the high precision of the IG model outputs. Figure 4-11 demonstrates a good match between the model results and the measured data for the test set, which means the model can successfully generate flowrate and pollutant concentration time series. However, some discrepancies can be noticed for the hourly generation occurring at the beginning of wet weather flows (Figure 4-5). This might be caused by overflows generated in the sewer system or depression storage effects, which are not included in the model. On the other hand, high frequency water quality time series generation is based on sensor measurements, and it must be recognised that sensor clogging and anormalies often occur, especially during rain events (Bertrand-Krajewski et al., 2007).

The snowmelt period in the generated data is also shifted in time for some of the years that were studied (Figure 4-4 and Figure 4-5). This timing error is due to the model considering snowmelt water infiltration as an average amount and at an average time in the year. This issue may be dealt with by modifying the yearly pattern for different years, using for instance temperature data, which can indicate the time of the snowmelt period, i.e. a higher temperature may induce an earlier peak flow, by causing earlier melting, compared with average yearly behaivour.

The proposed IG model provides a tool for fast influent profile generation at different resolutions and it benefits not only the carbon removal but also the nutrient removal process. Moreover, instead of being a deterministic model, the proposed IG creates a stochastic process, which may be helpful for further WRRF modelling uncertainty and scenario analysis. Compared with available phenomenological models presented in the literature review, the advantage of using the proposed data-driven model is that there is no need to gather any information on the catchment, nor does it need calibration of physical parameters. Therefore, It is suitable for cachtments with unknown physical details.

Finally, to confirm its utility and advantages, the proposed model was compared with a very simple data-driven method: the generated TSS concentration was calculated simply as the average TSS load (calculated for the whole available data series), divided by measured flowrate data. This reflects the fact that a city typically generates a stable daily pollution load. The results of Appendix 1 show that the obtained ANN model is considerably more precise, using the same input and the same size of training set.

This simple data-driven model faces the same problem as any ANN model, i.e. it is risky to extrapolate if the new time series is too much different from the training set. To counter this limitation, a more complete database is required to make sure the model has been exposed to ‘more experiences’ and can thus capture more of the system properties, such as more diverse rain data, wider and higher temporal distribution of water quality measurements, etc..

4.6 Conclusion

This work developed an urban wastewater influent generator model for flowrate and pollutant concentration generation. The proposed IG model is data-driven and includes an ANN, a multivariate regression and a random walk process. Its performance was analyzed by different criteria: MAPE, NSE and statistical characteristics evaluating variability (KL_divergency). The IG model showed a high generation precision for the two case studies: 10% error for flowrate and 15-20% for the concentrations, which are comparable to the measurement errors of raw wastewater data (Bertrand-Krajewski et al., 2007; Montgomery and Sanders, 1986). Given the fact

that the performance of a data-driven model depends on the quality and coverage of the training data availability, a more complete dataset (in temporal or spatial sense) can help further improve the model performance.

The proposed IG is balanced in terms of modelling efforts and precision. Compared with a simple data-driven IG model based on average pollutant load and flow, the proposed IG demonstrated a higher accuracy. On the other hand, compared to phenomenological IG models, this model requires less modelling efforts, especially in calibrating the parameters for which prior expert knowledge is needed. Another significant improvement of the proposed IG is that stochastic variability is included in the model, which enables the generated time series to better represent the reality in terms of probability distribution of the generated data. Last but not least, the application of the proposed IG requires only routine WRRF data, thus not requiring any additional investments in data collection. Once the IG model is calibrated, the generation of new time series only requires weather data and is not relying on historical influent data. This ensures the prediction result will be stable over a long time horizon and will not accumulate errors over time.

In conclusion, the proposed IG generates a dynamic and complete data set (flowrate and pollutant concentrations, both organics and nutrients) with good performance, and it is able to generate a time series at different time resolutions (daily and hourly). Further research will be focusing on the optimization of the IG in order to better support WRRF modelling studies.

Chapter 5. An influent generator for WRRF design and operation based on a recurrent neural network with multi-objective optimization using a genetic algorithm

This chapter has been published in Water Science & Technology:

Li, F., Vanrolleghem, P.A., 2022. An influent generator for WRRF design and operation based on a recurrent neural network with multi-objective optimization using a genetic algorithm. Water Sci. Technol. 85, 1444–1453.

5.1 Abstract

Nowadays, modelling, automation and control are widely used for Water Resource Recovery Facilities (WRRF) upgrading and optimization. Influent generator (IG) models are used to provide relevant input time series for dynamic WRRF simulations in these applications. Current IG models found in literature are calibrated on the basis of a single performance criterion, such as the mean percentage error or the root mean square error. This results in the IG being adequate on average but with a lack of representativeness of, for instance, the observed temporal variability of the dataset. However, adequately capturing influent variability may be important for certain types of WRRF optimization, e.g., reaction to peak loads, control system performance evaluation, etc. Therefore, in this study, a data-driven IG model is developed based on the long short-term memory (LSTM) recurrent neural network and is optimized by a multi-objective genetic algorithm for both mean percentage error and variability. Hence, the influent generator model is able to generate a time series with a probability distribution that better represents reality, thus giving a better influent description for WRRF design and operation. To further increase the variability of the generated time series and in this way approximate the true variability better, the model is extended with a random walk process.

5.2 Résumé

Actuellement, la modélisation, l'automatisation et le contrôle sont largement utilisées pour la mise à niveau et l'optimisation de station de récupération des ressources de l'eau (StaRRE). Les modèles de générateur d'influence (GA) sont utilisés pour fournir des séries chronologiques d'entrées pour les simulations dynamiques de StaRRE sur ces applications. Les modèles GA actuels trouvés dans la littérature sont calibrés sur la base d'un critère de performance unique, tel que l'erreur moyenne en pourcentage ou l'erreur quadratique moyenne. Cela entraîne un GA adéquat en moyenne mais avec un manque de représentativité, par exemple, de la variabilité temporelle observée de l'ensemble de données. Cependant, la capture de la variabilité adéquate de

l'affluent peut être importante pour certains types d'optimisation StaRRE, par exemple, la réaction aux charges de pointe, l'évaluation des performances du système de contrôle, etc. Pour cette raison, dans cette étude, un modèle GA basé sur les données est développé basé sur le réseau de neurones récurrents à mémoire court-terme et long-terme (LSTM) et est optimisé par un algorithme génétique multi-objectifs pour le pourcentage d'erreur moyen et la variabilité. Par conséquent, le modèle de GA est capable de générer une série chronologique avec une distribution de probabilité qui représente mieux la réalité, donnant ainsi une meilleure description de l'affluent pour la conception et l'opération du StaRRE. Pour augmenter encore la variabilité des séries chronologiques générées et ainsi mieux se rapprocher de la vraie variabilité, le modèle est étendu avec un processus de marche aléatoire.

5.3 Introduction and background

Nowadays, modelling, automation and control are widely used for Water Resource Recovery Facilities (WRRF) upgrading and optimization. However, because of a lack of adequate input datasets, their full potential remains underexploited. For example, for WRRF design, engineers usually make initial sizing by using design guidelines based on average loads and safety factors (Talebizadeh et al., 2015). Also, controller performance can benefit from influent forecasting, e.g. with nitrogen load data for ammonia-based aeration control (Newhart et al., 2020). Thus, a reliable influent generator model is becoming increasingly necessary as input to WRRF modelling and process control studies (Gernaey et al., 2011; Martin and Vanrolleghem, 2014).

Many influent generators (IG) have already been presented in literature. They can be organized into two categories: phenomenological models (Flores-Alsina et al., 2014b; Jeroen Langeveld et al., 2017) and data-driven models (Ahnert et al., 2016; Borzooei et al., 2019; Li et al., 2020).

Compared with phenomenological models, data-driven models usually require less calculation efforts thanks to the lower model complexity, while still delivering a high-performance result. As a powerful data-driven tool, machine learning approaches have been increasingly used in water engineering thanks to the increasing data collection and data mining tool development (Corominas et al., 2018). Recently, Artificial Neural Networks (ANNs) have been studied for WWTP influent generation, especially for short-term flow forecasting and for pollutant concentration prediction (Aminabad et al., 2013; El-Din and Smith, 2002; Li et al., 2020; Ma et al., 2014; Shokry et al., 2018)

The first challenge of data-driven IG development is that despite the fact that ANNs have been proven to have an excellent ability for modelling water systems, it is necessary to first define an adequate ANN model architecture. Since the wastewater generation process of a sewer catchment is a complex nonlinear process

and it depends on historical information (e.g., previous rain events), the recurrent neural network (RNN) is expected to be the most appropriate architecture as it allows learning time-sequential behaviour thanks to its 'internal memory'. RNNs are found to be better than fully connected neural networks, such as multilayer perceptrons (MLP). The long short-term memory (LSTM) is one type of RNN that was proven to have excellent performance for time series modelling, especially when learning long-term dependencies (Hochreiter and Schmidhuber, 1997). It can therefore be expected to provide good performance for IG modelling.

The second challenge of influent generator development is that even though multiple influent generators have been presented in literature, the calibration process used is single objective optimization, in other words, IGs are calibrated with only one performance criterion in mind, typically defined as either the mean percentage error (MPE) or the root mean square error (RMSE). These criteria lead to IG that are adequate on average but they may not represent the observed variability well. In other words, without better consideration of the statistical distribution of the data, the model will not be able to fully capture the influent variability. Representing variability well is important to better define WRRF design parameters, such as the expected maximum load and hydraulic capacity and for scenarios analysis in which one studies the influent disturbance impact when evaluating operation under peak conditions, etc.

To solve these issues with existing IG models, a novel data-driven IG model is developed using the LSTM to describe wastewater quality. The model is calibrated by a multi-objective genetic algorithm (MoGA) (Fonseca and Fleming, 1999), which next to being able to handle multiple criteria also outperforms traditional gradient descent optimization tools in dealing with local minima, ensuring efficient optimization during IG development (Rojas, 1996).

The provided case study illustrates that such IG model is able to generate long-term concentration data for COD and TSS, based only on previous flowrate and weather data. It has good performance in terms of mean percentage error, while at the same time providing excellent similarity to the variability of the real dataset.

5.4 Case study description and data pre-treatment

In this study, the IG modelling approach was developed based on data collected in Quebec City, Canada (Tik and Vanrolleghem, 2017). The dataset is based on the same period as used in Chapter 4. The catchment is a combined sewer system, with a very variable flowrate and pollutant concentration under different weather conditions, such as storm water and particularly, the snowmelt at the end of winter. The Quebec City WRRF is a carbon removing treatment plant and its dataset includes weather information (rain and temperature), daily flowrate, and daily COD and TSS concentrations.

The influent flow rate and concentrations show recurring hourly and daily variations but are disturbed by rain events. Data pre-treatment was conducted according the procedure of Alferes et al., (2013), including outlier removal, data smoothing, and a univariate fault detection method. Then the normalization process is applied to scale the data in the range between 0 and 1.

In order to remove the effect of precipitation, signal noise, and obtain a typical long-term yearly dry weather pattern, a Chebyshev bandpass filter (2nd order, Type I) was selected to extract the signal at these specific frequencies from the data (Schlichthärle, 2011). This signal then can represent the seasonal effect and the dry weather flow corresponding to the urban activities.

5.5 Materials and methods

5.5.1 Fully connected ANN model

The MLP is a class of feedforward ANN where each neuron in one layer is fully connected to the next layer (El-Din et al., 2004; Raman and Sunilkumar, 1995; Zhang et al., 2019). It is one of the most used ANN architectures and is used in this study as the reference, see Figure 5-1(a). It is trained using the classic learning process only for minimal RMSE and using the gradient descent approach, i.e., the Levenberg–Marquardt backpropagation algorithm (Levenberg, 1944). The sigmoid function is selected as activation function in the hidden layer, in order to represent the nonlinearity of the influent generation process.

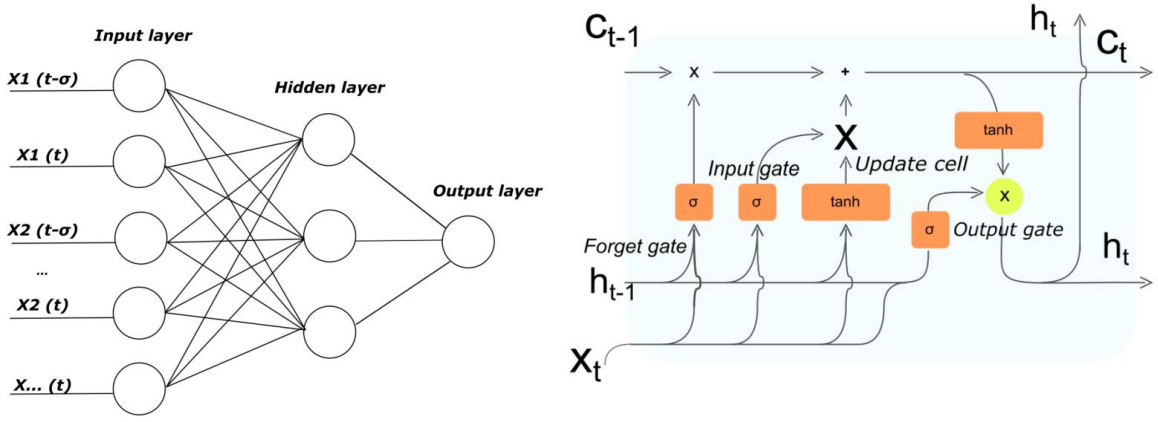


Figure 5-1 Architecture for (a) MLP fully connected ANN (b) LSTM architecture

5.5.2 LSTM

Despite the fact that the MLP can learn influent dynamics, the RNN is expected to be a more adequate model structure because it can better describe the pollutant concentration that is influenced by previous conditions. To handle the long-term dependency problem, the LSTM is selected to integrate this earlier information. Figure 5-1 (b) illustrates the architecture of LSTM with its gates structure, i.e., input, forget, and output gates. The long-term memory is stored and carried on by the cell state through time (Alex. Graves, 2012; Hochreiter, 1998). The input layers and architecture of the LSTM used in this study are presented in Figure 5-2.

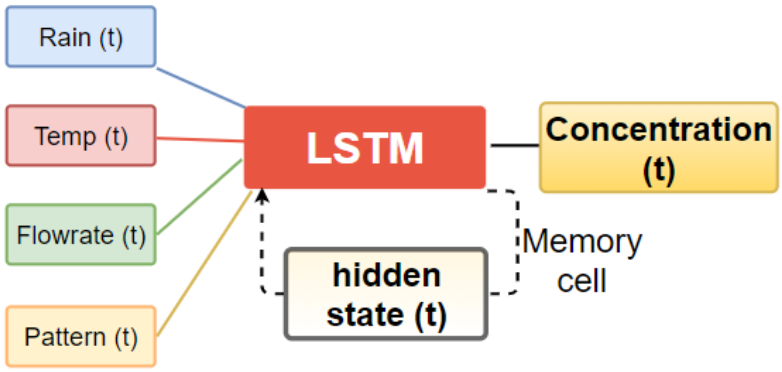


Figure 5-2 The architecture of the LSTM recurrent neural network used in this study for IG modelling

5.5.3 NSGA-II method

Genetic algorithms are popular optimization approaches inspired by the process of natural selection and have quickly became a popular evolutionary algorithm (Holland, 1992). It is playing an increasingly important role in machine learning and many other applications (Ercan and Goodall, 2016; Goldberg, 1989). Multi-objective optimization problems can be solved either by determining an entire set of optimal solutions on the Pareto front, or by combining the individual objective functions into a single one using, for instance, weights (Chiandussi et al., 2012; Konak et al., 2006). The NSGA-II method (non-dominated sorting genetic algorithm II) (Deb et al., 2002) is widely used in machine learning and starts to be used in the water engineering field to solve multi objective optimization problems (Ercan and Goodall, 2016; Wang et al., 2019; Yusoff et al., 2011).

Figure 5-3 displays the flowchart of NSGA. An initial population is generated randomly. The non-dominated sorting fronts are ranked by the evaluation of their fitness. In this research, the crowding distance and tournament

selection are used for the selection process (Miller and Goldberg, 1996) and adaptive mutation and crossover ratios are applied to improve the search efficiency (Hassanat et al., 2019; Srinivas and Patnaik, 1994).

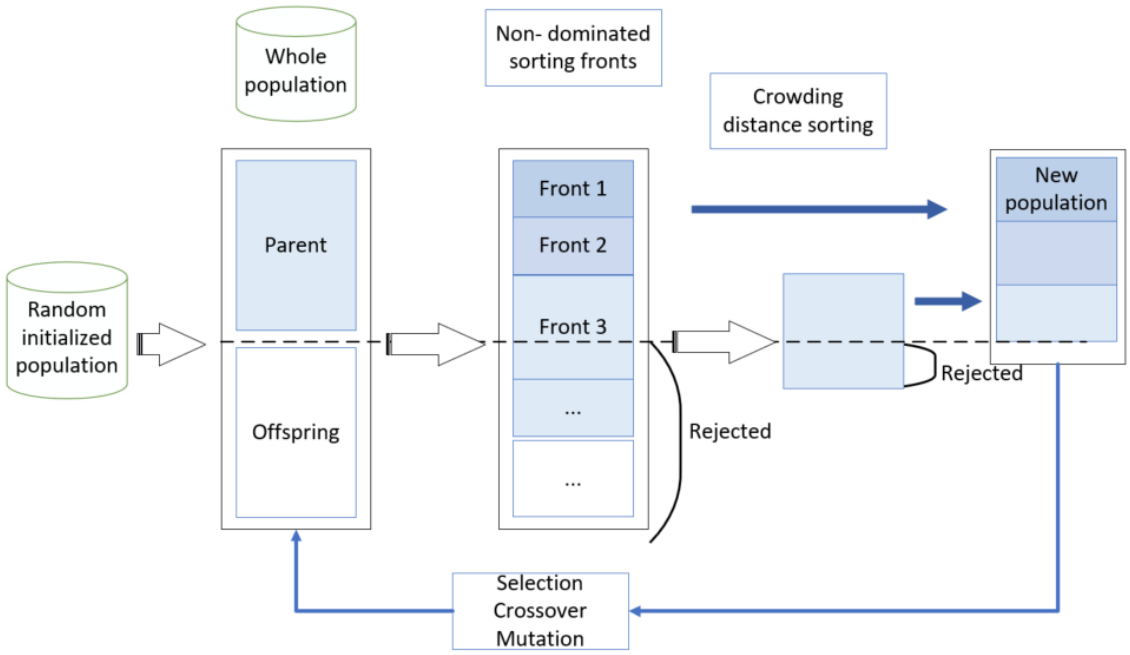


Figure 5-3 NSGA-II flowchart

Different quantitative performance criteria aim to measure how well a model simulation fits the available observations. A diversity of criteria has been studied and compared for engineering and wastewater applications (Chiandussi et al., 2012; Hauduc et al., 2015). In this study, the aim is to obtain an IG which has a high precision and, at the same time, a variability similar to the one of the observed datasets. Thus, the following two criteria were chosen to represent the desired IG performance: the mean absolute percentage error (MAPE), which is a common criterion for time series, and the $KL_{divergence}$ (Kullback and Leibler, 1951), which measures the difference between two probability distributions:

$$MAPE = \frac{1}{n}\sum \left|\frac{y_s - y_o}{y_s}\right| * 100\% \qquad 5\text{-}1$$

$$KL_{divergence}(P\|Q) = -\sum P(x) * \log\left(\frac{Q(x)}{P(x)}\right) \qquad 5\text{-}2$$

where, y_s is the simulated and y_o is the observed data, $P(x)$ is the probability distribution of the observed data and $Q(x)$ is the distribution of the simulated data.

5.5.4 Additional random walk

As shown in the results section, even though the IG model was optimized by NSGA-II, the result for variability was felt insufficient and it was tried to improve the model structure by increasing the variability it could generate. In general, a time series model is composed of a deterministic part, a stochastic part and an auto-correlated error term (Reichert and Schuwirth, 2012; Villez et al., 2020). Therefore, by adding an error term, in this case a random walk process, to the calibrated GA model, it is expected to improve the data series variability and eliminate the autocorrelated error now present because of the insufficient inclusion of short-term dynamics. By doing this, the IG simulation is expected to achieve statistical properties closer to those of the observed data.

The stochastic process is modelled by a k-order random walk model:

$$R_t = \sum_1^k \varphi_k * R_{t-k} + \varepsilon \qquad 5\text{-}3$$

5.6 Results and Discussion

5.6.1 Result Results of LSTM-NSGA-II

The IG model developed for the case study at hand generates results for daily TSS and COD concentrations. In order to keep the efficiency of the LSTM model, the input includes 4 previous days of data regarding weather (rain and temperature), wastewater flow and seasonal pattern of COD, obtained after data pre-treatment. Indeed, the auto-correlation analysis shows that lags up to 4 days back had a significant autocorrelation.

The final model consists of two LSTM hidden layers and one output layer. Based on previous experience, the following settings of the NSGA-II algorithm were adopted: in order to find the global optimal solution, a population of 2000 individuals was initialized at the beginning of the NSGA-II optimization process, and 25 search iterations were performed. The initial crossover rate and mutation rate were defined as 0.75 and 0.15, respectively.

The results of LSTM-NSGA-II are compared with a classical full-connected neural network with sigmoid function in the hidden layer, trained with back-propagation (ANN-BP). Figure 5-4 illustrates the Pareto front of the last iteration of the NSGA optimization for generating COD concentrations as an example. The green points represent a series of optimal non-dominated solutions in which the MAPE cannot be improved without sacrificing the KL divergence or vice versa. For visual comparison, the blue triangle represents the result of the ANN-BP. The final solution (red point) was chosen by balancing the two performance criteria and is discussed below.

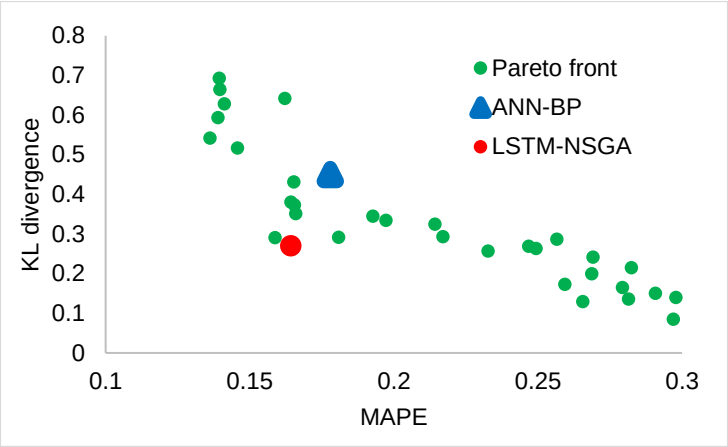


Figure 5-4 NSGA-II optimal Pareto front solutions: the red point is the solution chosen for further analysis and the blue triangle represents the ANN-BP result.

As shown in Figure 5-5, the results for the test set illustrate the good performance of the simulated time series, see Table 5-1. The dilution effect during wet weather caused by stormwater inflow and subsequent rain-induced infiltration can be observed. It can also be noticed that the LSTM-NSGA-II result is more variable than the time series generated by the ANN-BP model, which confirms that the LSTM is better to describe influent variability, albeit still insufficient, as discussed below. In addition, by analysing the probability density function (PDF) and the cumulative density function (CDF), the LSTM-results are very similar to reality for both COD and TSS concentrations (see Figure 5-6).

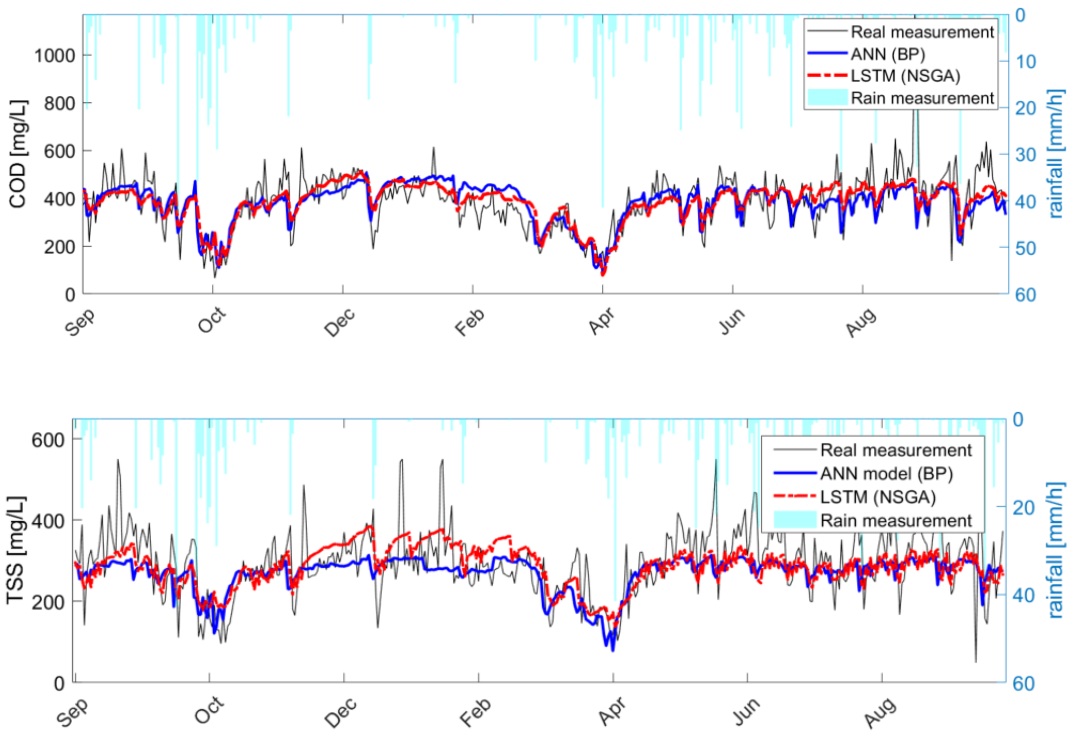
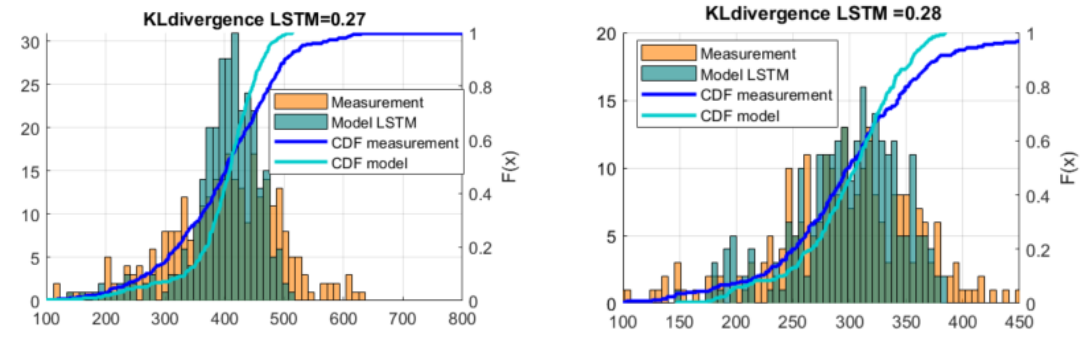


Figure 5-5 COD concentrations (top) and TSS concentrations (bottom) generated by LSTM (red line) and ANN-BP (blue line), by using an input data set only including weather and flowrate data.



COD [mg/L] TSS [mg/L]

Figure 5-6 PDF and CDF result for LSTM-generated COD (left) and TSS (right), the KL divergence values are given.

This model can not only be applied to generate time series of organic pollution (COD, TSS) but also for nitrogen species (such as ammonia, see Figure 5-7). This can be very beneficial for database gap filling, considering that in carbon removing plants such as the one under study, the ammonia concentration is usually not measured daily. The influent time series generated in this way allows enhancing the evaluation of future nitrogen removal performance, and can also contribute to modelling nutrient recovery, etc.

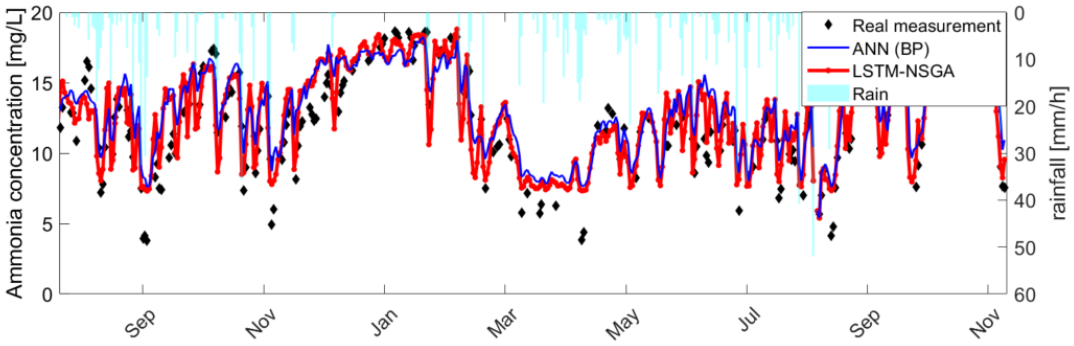


Figure 5-7 Ammonia concentration time series generated by LSTM-NSGA using occasional ammonia measurements at a carbon-removing plant.

Finally, the performance of the COD, TSS and ammonia concentration time series for the test set (one solution, the red point, chosen from the Pareto front of Figure 5-4) is summarized in Table 5-1 and compared to the full-connected back-propagation ANN model optimized by gradient descent (green triangle in Figure 5-4).

Table 5-1 Table 5-2 COD concentrations (top) and TSS concentrations (bottom) generated by LSTM (red line) and ANN-BP (blue line), by using an input data set only including weather and flowrate data.

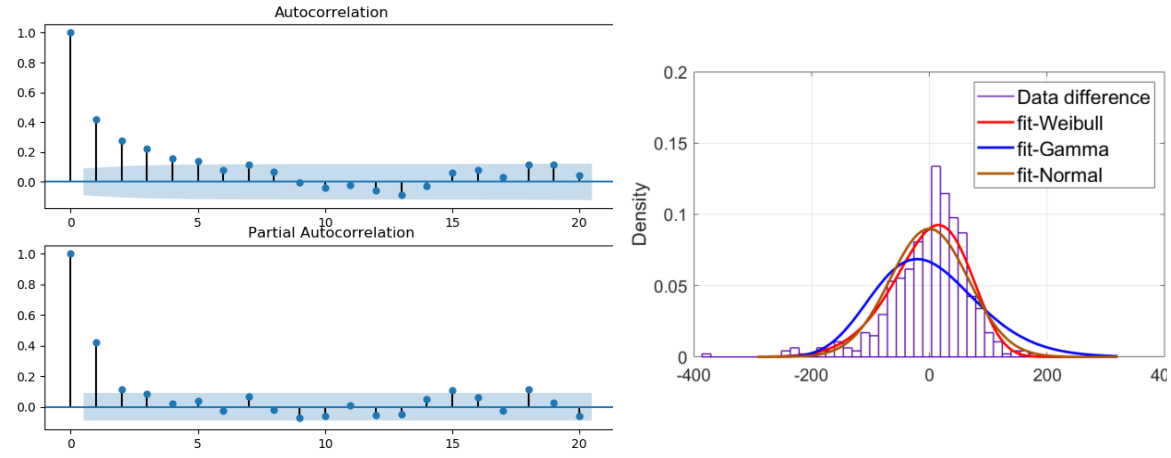
	LSTM optimized by NSGA-II		**ANN-BP**	
Criteria	MAPE	$KL_{divergence}$	MAPE	$KL_{divergence}$
TSS [mg/l]	18.5%	0.28	17.7%	0.49
COD [mg/l]	16.4%	0.27	17.8%	0.45
Ammonia [mg/l]	13.7%	0.22	15.1%	0.36

In general, the LSTM model performs better in terms of MAPE and especially KL_divergency. Thanks to cell memory, the LSTM is able to learn the impact of previous phenomena, such as the rainfall derived infiltration

and inflow (RDII). By ensuring the model quality in term of MAPE, while at the same time pursuing agreement on variability, the NSGA-II allows obtaining a distribution of model results similar to the real distribution, which makes the generated influent time series to have more variability than the one generated by the ANN-BP model. This improvement thus better describes the observed randomness of the concentration time series, which is important regarding to WRRF design and evaluation of WRRF control systems, for instance.

5.6.2 Results of benefits of adding a random walk process to the IG

As mentioned in the section additional random walk, it was tried to improve the IG by reconstructing the difference between the model output and the observations by extending the deterministic model with a random walk process. The autocorrelation and partial autocorrelation analysis (Figure 5-8a) suggest that an order equal to 4 should be selected for the k-order random walk model. In order to represent the noise term, different distributions were compared (see Figure 5-8b), and a Weibull distribution was selected as it gave the distribution most similar to the observed error distribution.



(a) (b)

Figure 5-8 (a) Autocorrelation and partial autocorrelation analysis for the error between the deterministic LSTM model output and the measured data, with the blue zone indicating the significant autocorrelation limit. (b) Fit of different distribution models to the error PDF.

After the autoregressive error reconstruction, Figure 5-9 illustrates the test set COD concentration results obtained with the model extended with a random walk process. The grey random walk band is the confidence

interval for the COD concentrations obtained by running 200 Monte Carlo simulations with different random sequences. The red line is one example chosen among these 200 simulations. The $KL_{divergence}$ for this example decreased from 0.28 to 0.14, while MAPE increased slightly (from 16.4% to 17.5%) because of the randomness of the random walk process. It is important to highlight that the random walk process aims to re-establish the variability of the pollutants, thus representing the stochastic process in the real time series. In other words, the objective of the random walk is not to synchronize the time series with the actual measured one and, therefore, the timing of the variations is not the most essential criterion for the model selection.

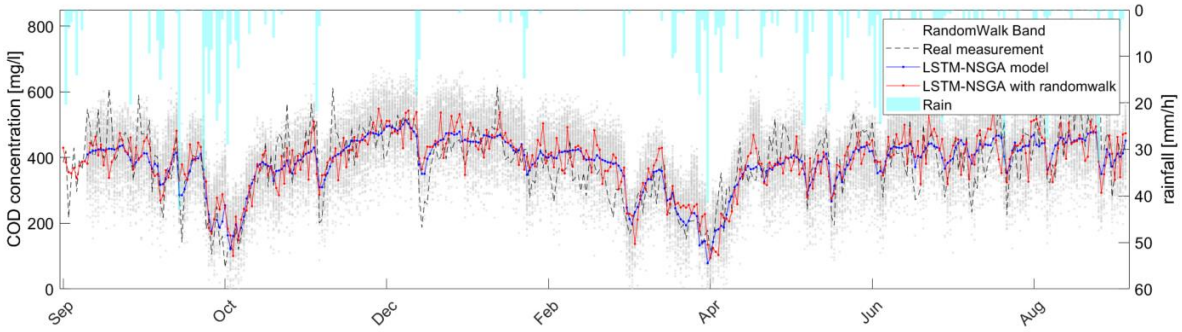


Figure 5-9 COD concentration generation with the IG model extended with a random walk process. The grey band represents the confidence interval of this stochastic model generated with 200 Monte Carlo simulations.

The model exhibits a good result for the ammonia concentration time series as well, see Figure 5-10. The $KL_{divergence}$ decreased from 0.22 to 0.17 with a slight increase in MAPE (from 13.7% to 14.26%). This approach improves the ammonia concentration profile generation and database gap filling: even though the real measurement is not performed on a daily basis, the model can generate an interval for the missing data.

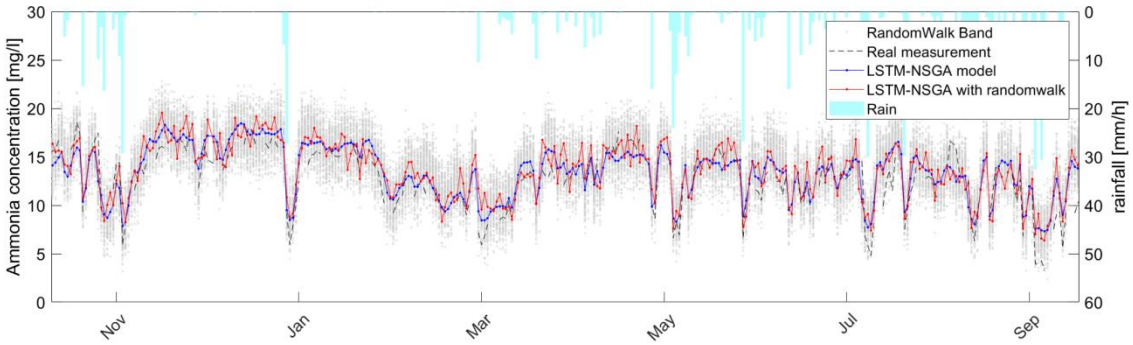


Figure 5-10 Ammonia concentration reconstruction with random walk process: the grey band represents the confidence interval generated using 200 Monte Carlo simulations.

5.7 Conclusion

The ability to properly describe the WRRF influent variability is very crucial for a number of WRRF modelling tasks. This paper has proposed an IG model based on the LSTM Recurrent Neural Network architecture to generate daily concentration data for COD, TSS and ammonia nitrogen. Compared with a plain ANN, it results in considerably better model performance in terms of influent generation when considering both accuracy and variability simultaneously. The latter feature of the new IG-model is pursued by using a multi-objective optimization, *in casu*, the NSGA-II algorithm.

A further improvement of the variability correspondence of the generated time series was achieved by adding a random walk process to the deterministic core LSTM model. The generated time series can now achieve a distribution very similar to the one observed, and therefore, it will help deciding on WRRF design parameters, especially for those related to peak hydraulic capacity and major influent disturbances, etc.

The proposed IG model was validated and tested on an actual case study, and it achieves excellent performance in both MAPE and time series distribution criteria. With the addition of a random walk process, it is able to generate a time series with a probability distribution that is even more similar to the observed reality, and thus gives better influent description for WRRF design and operation decision-making. It also allows for gap filling of incomplete databases, for instance for nitrogen species that are sparsely measured at carbon only removing plants as studied in this work.

Chapter 6. Including snowmelt in influent generation for cold climate WRRFs: Comparison of data-driven and phenomenological approaches.

This chapter has been published in Water Science & Technology:

Li, F., Vanrolleghem, P.A., 2022. Including snowmelt in influent generation for cold climate WRRFs: comparison of data-driven and phenomenological approaches. Environ. Sci.: Water Res. Technol., 2022.

6.1 Abstract

Influent generation models are developed to provide the influent disturbance at the inlet of a WRRF. A reliable influent model is important for WRRF design, upgrade and different digital twin studies. In this work, a data-driven methodology is proposed to create an influent generator (IG) model, which describes the influent flow and water temperature dynamics under the impact of snowmelt under cold climate conditions. The model structure applied was the long short-term memory (LSTM) artificial neural network with residual connection. The final result of influent generation for a Canadian case study is compared with a previously proposed phenomenological model. The performance is evaluated by different performance criteria and the results revealed that the LSTM approach has a better performance than the phenomenological model in terms of accuracy. In conclusion, the proposed model can successfully reproduce the influent dynamics of a combined sewer system's wastewater generation with snowmelt infiltration impacts.

6.2 Résumé

Des modèles de génération d'affluent sont développés pour fournir la perturbation des affluents à l'entrée d'une station d'épuration (StaRRE). Un modèle d'affluent fiable est important pour la conception, la mise à niveau et les différentes études de jumeaux numériques StaRRE. Dans ce travail, une méthodologie basée sur les données est proposée pour créer un modèle de générateur d'affluent, qui décrit la dynamique de débit de l'affluent et de la température de l'eau sous l'impact de la fonte des neiges dans des conditions climatiques froides. La structure de modèle appliquée était le réseau de neurones artificiels à mémoire longue et à court terme (LSTM) avec connexion résiduelle. Le résultat final de la génération d'affluents pour une étude de cas à Canada, est comparé avec un modèle phénoménologique proposé précédemment. La performance est évaluée par différents critères de performance et les résultats ont révélé que l'approche LSTM a une meilleure performance que le modèle phénoménologique en termes de précision. En conclusion, le modèle proposé peut

reproduire avec succès la dynamique des affluents de la production d'eaux usées d'un système d'égouts unitaires avec des impacts d'infiltration de la fonte des neiges.

6.3 Introduction

A reliable influent model is essential to run realistic Water Resource Recovery Facility (WRRF) simulations in digital water studies(Martin and Vanrolleghem, 2014). Such influent generator (IG) models are approached in two ways: phenomenological models and data-driven models. Different phenomenological models have been created and applied for flow and pollutant fate(Coutu et al., 2016; Ort, 2006). One of the most widely used models is the phenomenological dynamic influent pollutant disturbance scenario generator (DIPDSG)(Gernaey et al., 2011) developed for the benchmark simulation models (BSM). This phenomenological model consists of different components: the diurnal load pattern (including weekend effect), an infiltration block and sewer transport system blocks, etc. The model has been applied to different case studies(Flores-Alsina et al., 2012b). Other phenomenological IGs were applied for the integrated modelling of a variety of pollutants(Lindblom et al., 2006) .

Regarding the data-driven approach, weather-based IGs have been studied based on data mining methodologies(Li et al., 2020). Borzooei et al.(Borzooei et al., 2019) studied the impact of weather on influent wastewater flow and characteristics. In addition, different data mining methods have been applied for short-term predictions of influent and pollutant dynamics(Banihabib et al., 2019; Corominas et al., 2018; Zhu and Anderson, 2019). Therein, the artificial neural networks (ANN) are one of the most commonly used approaches, not only for forecasting the WRRF influent disturbance in view of improving wastewater treatment operation and real-time control(Kriger and Tzoneva, 2007; Wei and Kusiak, 2015; Zhou et al., 2019), but also for the simulation of rainfall-runoff and solids transport in sewers during storm events(Gong et al., 1996).

So far, research on IGs has not widely tackled WRRF operated under cold climate. Different research demonstrates that the snowmelt period at the end of winter causes an important increase in flowrate and a significant change in influent temperature and pollutant concentrations(Wang et al., 2017). These important changes challenge a WRRF's operation and design as well as its treatment performance(Alisawi, 2020; Di Trapani et al., 2013; Gullicks and Cleasby, 1990; Pishgar et al., 2021). The extreme flow peak and lower temperature affect the hydraulic and treatment capacity, respectively(Stachowiak, 2007). Indeed, the sudden decrease of temperature will influence the bioactivities, such as nitrogen removal efficiency during treatment(Plósz et al., 2009). Whereas the impacts of snow on WRRF influents have been characterized(Wang et al., 2017), very limited research has been conducted to model these wastewater dynamics under snowmelt.

As an example, a multiple linear regression model was built for short-term pollutant concentration prediction during the winter period(X. Wang et al., 2019). However, research gaps are still remaining with respect to the relationship between snowmelt and WRRF influent characteristics.

Current urban snowmelt models are mainly of two types(Debele et al., 2010; Moghadas et al., 2016): the temperature-index method and the energy budget approach. The temperature-index considers that the temperature is the basic driving force in the snowmelt processes. In contrast, the energy budget considers all the incoming, outgoing, and stored energies. Positive net incoming energy will lead to snowmelt, but the relation between snowmelt and infiltration into the sewer is not completely elucidated yet.

To adapt the influent profiles generated by the BSM influent generator for cold regions, some modifications were made by adding a snow-reservoir block, among others(Flores-Alsina et al., 2014a; Saagi et al., 2018). The snowmelt process is described as follows: the snow is melting leading to a snow depth change. If the temperature is higher than zero, the snow is considered to melt and is transformed into precipitation at low intensity. The delay effect is represented by using a transfer function. In the meanwhile, the water temperature is affected by the mixing of the household wastewater with snowmelt water, which leads to a decrease of the temperature compared with ambient temperature. However, it is felt necessary to evaluate whether these simplified modifications are precise enough for fully describing the effect of snowmelt on influent generation.

In order to better represent the snowmelt influence and improve the WRRF influent dynamics model, in this study, a data-driven methodology is proposed to create an influent generator based only on weather information and snow depth data. The model describes the influent and water temperature dynamics under the impact of snowmelt for cold climate regions. The model is created by using the long short-term memory (LSTM) with residual connection architecture. The performance is evaluated using different criteria and compared with a calibrated BSM phenomenological influent generator. Finally, the benefits of using the data-driven model are discussed and demonstrated with a scenario analysis.

6.4 Methodology

6.4.1 Case study and preliminary data treatment

The modelling approach was developed using data collected in Quebec City, located in Canada. The catchment is a combined sewer system, with very variable flowrates in different weather conditions, such as stormwater and snowmelt infiltration.

As the fifth snowiest city in the world, the average annual snowfall amount in Quebec City reaches around 300cm from December to May. The Quebec East WRRF is receiving domestic wastewater, industrial wastewater, and the infiltration of storm and ground water from a combined sewer system, with an average flowrate generated by around 300 000 PE (around $0.7*10^4$ m^3/h)(Tik and Vanrolleghem, 2017). This flowrate increases up to $2.5*10^4$ m^3/h at the end of winter, due to snowmelt and infiltration.

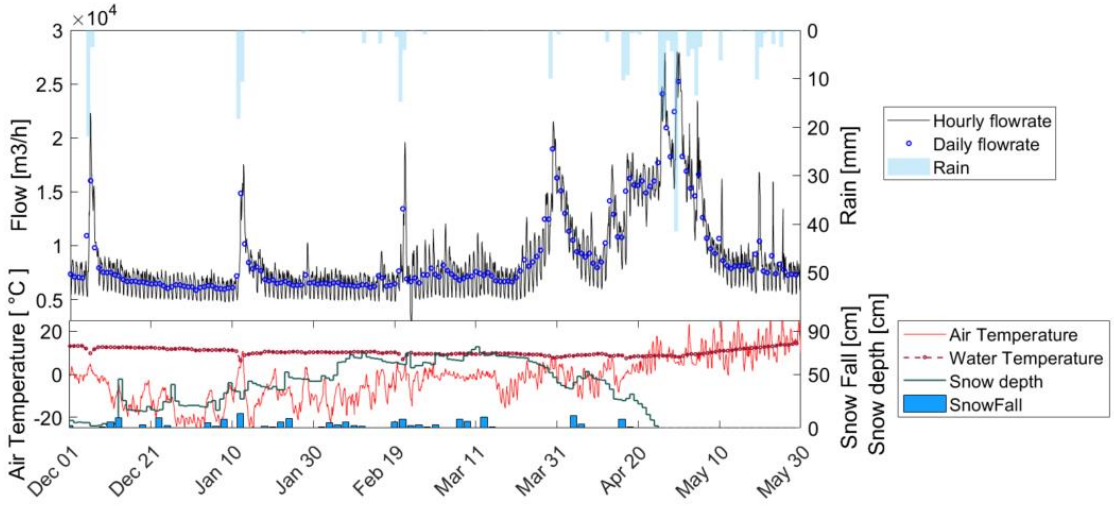


Figure 6-1 Flowrate and temperature influenced by snowfall and precipitation in winter, Quebec City

Focusing on the influence of snow on the wastewater influent generation, this study is based on a 5-year dataset of the winter period (December to May) (2014-2018). The hourly flowrate and hourly wastewater temperature will be generated according to air temperature, snow depth measured by an acoustic distance sensor, and rain gauge data, which are collected by weather stations operated by the meteorological service of Canada. In order to balance the data requirements and the model complexity, two weather stations around WRRF were selected to represent the catchment (Beauport, and Québec City).

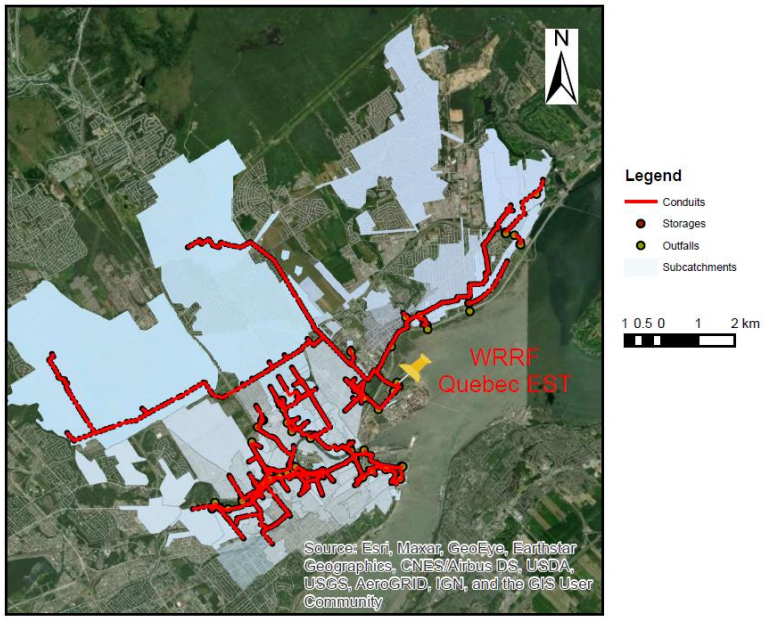


Figure 6-2 The catchment of the East WRRF in Quebec City, the red line represents the main lines of the combined sewer system, the pin represents the WRRF location. Source: Esri, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

The data pre-treatment applied aimed to eliminate and replace outlier data. The outliers were detected and replaced by the univariate data quality analysis of Alferes and Vanrolleghem, (2016). The outlier data are detected by defining acceptable thresholds to data features and to the residuals standard deviation (RSD). Then the faulty data are replaced by the smoothed data obtained by the kernel smoother.

Data normalization for the LSTM development involves adjusting the different data to a common scale. Min-Max normalization is used to scale the data between 0 and 1. The dataset is divided into three sub-sets: (i) the training set (70%) to estimate the weights in the model, (ii) the cross-validation set (15%) to determine the hyperparameters of the model and (iii) the test set (15%), which consists of unseen data that is used for the final evaluation of the model's performance.

To develop the phenomenological model, both the training and validation data sets were used to build the phenomenological model, and then the same test set was used for validation of this model. This different dataset split compared to data-based modelling is due to the fact that the phenomenological model is based on mass-balance and other physical equations, which have a fixed model structure. In contrast, the data-based model benefits from more freedom on its model structure, but this requires selection of the best structure, which is what the cross-validation set is used for in order to get the best hyperparameters for the model.

$$x_{norm} = \frac{x - min}{max - min} \qquad 6\text{-}1$$

Figure 6-3 represents the procedure for the data pre-treatment. The treatment was applied for both flowrate and water temperature data.

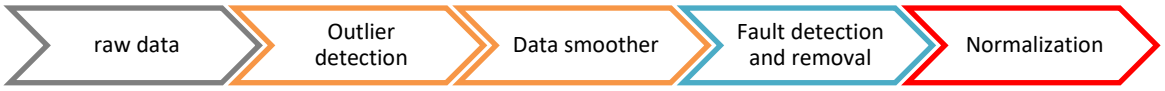


Figure 6-3 Flowchart for data pre-treatment

6.4.2 LSTM and residual connection

The WRRF influent is strongly influenced by previous conditions especially during the long period of snowmelt infiltration. Thus, a recurrent neural network is expected to be an adequate model structure. Recurrent neural networks (RNN) are a class of ANN specialized to learn temporal dynamic behaviour. The simple RNN suffers from short-term memory, which means that if a sequence is long, it is hard to carry information from earlier time steps all the way to the later ones. This will make that important information from the beginning is left out. To solve this issue, the LSTM is selected as an alternative model architecture.

The LSTM is a type of recurrent neural network, which was proven to have good performance on time series(Hochreiter and Schmidhuber, 1997; Van Houdt et al., 2020). The LSTM is widely used in different water science domains related to time series analysis, such as in hydrology for groundwater table prediction(Zhang et al., 2018), for drinking water quality prediction, and for wastewater treatment plant operation(Pisa et al., 2019) etc. Thanks to the LSTM memory unit, vanishing or exploding gradients, caused by the backpropagation through time in vanilla RNN training, which hampers learning of long-term dependency in data sequences, can be avoided in the long-term learning process (Bynagari, 2020; Alex Graves, 2012).

The LSTM neural network has an internal memory that can learn long-term dependencies of sequential data. The LSTM unit (Figure 6-4) consists of an internal memory cell and three gates: forget gate(f_t), input gate (i_t)and output gate (O_t). The forget gate decides which information can be forgotten from the previous state. The input gate receives the input data (X_t) at the current timestep, and together with the previous hidden state(h_{t-1}), it will decide what new information will be stored in the new cell state (C_t). Then the output gate highlights what information should be going to the next state, also known as output (h_t). Thanks to this gate structure, the relevant

information can pass through the long chain of sequences, without suffering from the problem of vanishing or exploding gradients.

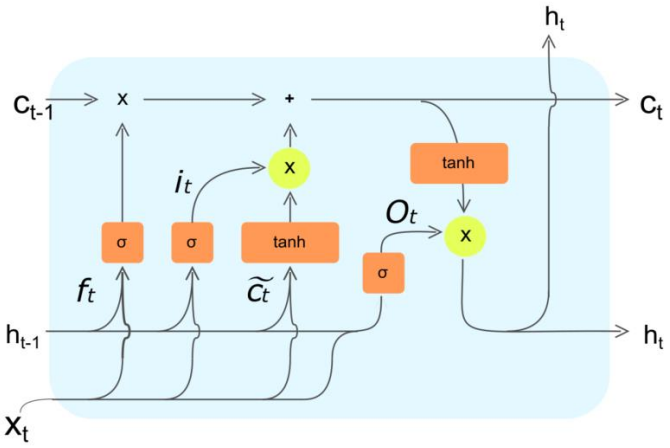


Figure 6-4 LSTM unit with gates structure: three gates and internal memory cell (more explanation, see text)

Figure 6-5 illustrates the architecture of the so-called LSTM with residual network, which provides a shortcut path between adjacent layers and has been applied to build the IG. Some deep learning studies have shown that such residual networks are easier to train because the skip connection architecture provides more efficient training when using multiple LSTM layers and can ensure the performance of the neural network will not degrade as the network's depth is increased(He et al., 2016; Pohlen et al., 2017; Wu et al., 2017). In other words, in theory, the network's performance is expected to be better with increasing number of layers, however, experience also shows that the accuracy (training error) starts to degrade when extra layers continue to be added beyond a certain critical number of layers. The residual connection can overcome this performance degradation problem, thus guaranteeing the performance of this approach of deep learning(He et al., 2016; Veit et al., 2016).

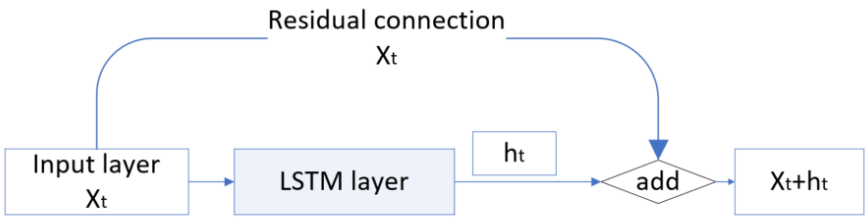


Figure 6-5 LSTM with residual connection

An overview of the proposed IG model is shown in Figure 6-6. The architecture consists of LSTM layers with residual network. The input data set includes two weeks of historical daily weather information and snow depth change during snowmelt, as well as recent 24 hour of hourly weather data. The daily household wastewater pattern is described by a harmonic function (second order Fourier series), which represents the diurnal variations of flow and load(Langergraber et al., 2008; Giorgio Mannina et al., 2011). The output is the hourly flowrate. In order to minimize the modelling efforts while considering the limited amount of data, the number of layers was defined as follows: the network consists of two blocks of LSTM layers with residual connection (one for the low frequency data input branch and another for the high frequency input branch), followed by one LSTM layer and one output dense layer. In terms of the number of units for each layer, the two residual connected LSTM layers consist of 4 units, which were predefined to minimize the modelling efforts, while the number of units in the last LSTM hidden layer will be decided on the basis of the cross-validation. Once the flow simulation result is available, the water temperature can be obtained by the same modelling approach.

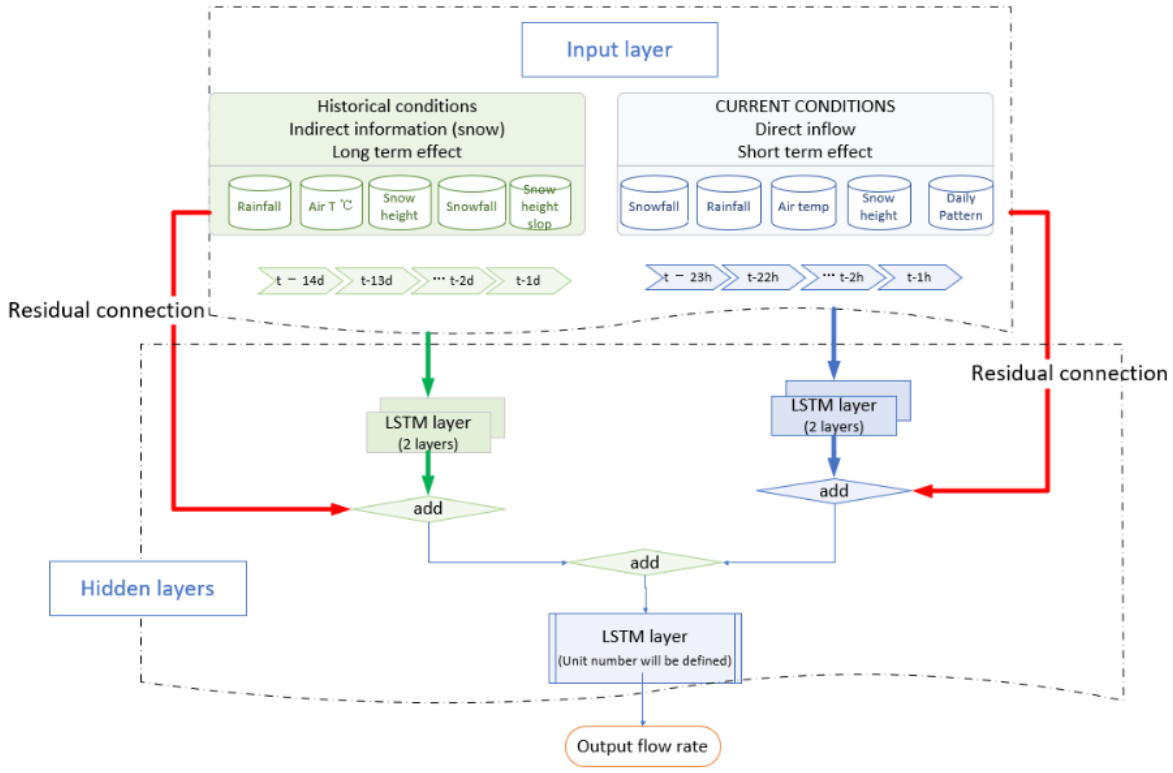


Figure 6-6 LSTM neural network model with residual connection architecture (the residual connection is represented by the red arrow). A similar approach is adopted for Temperature.

The parameter optimization used for training is based on gradient descent using the adaptive moment estimation algorithm (Kingma and Ba, 2015). The cost function used was the mean square error (MSE), penalized by the regularization parameters to avoid overfitting:

$$Cost\ function = \frac{1}{2m}[\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2] \quad 6\text{-}2$$

where $\hat{y}_i$is the generated data and y_i is the actual observation time series, λ is the penalty term between 0 to 1 and θ_j are the values of the weights. As the formula shows, the cost function enables finding the best balance between model complexity and accuracy. During the training procedure, the early stop technique is used to avoid overfitting. In other words, if the cross-validation is not reducing further, the training will stop after a certain number of epochs.

6.4.3 Modified BSM influent generator model

In order to compare and validate the machine learning model performance, the result was compared with the widely used BSM influent generator model. The original model was developed by Gernaey et al.(Gernaey et al., 2011) with modifications to adapt it to different case studies(Flores-Alsina et al., 2014a; Saagi et al., 2018). In order to adapt and apply the model to this snow-impacted case study, instead of using the existing models as such, the sub-model 'Rain generation and temperature sub-model' was modified using inspiration of the above-mentioned models. The layout of the original model is shown in Figure 6-7.

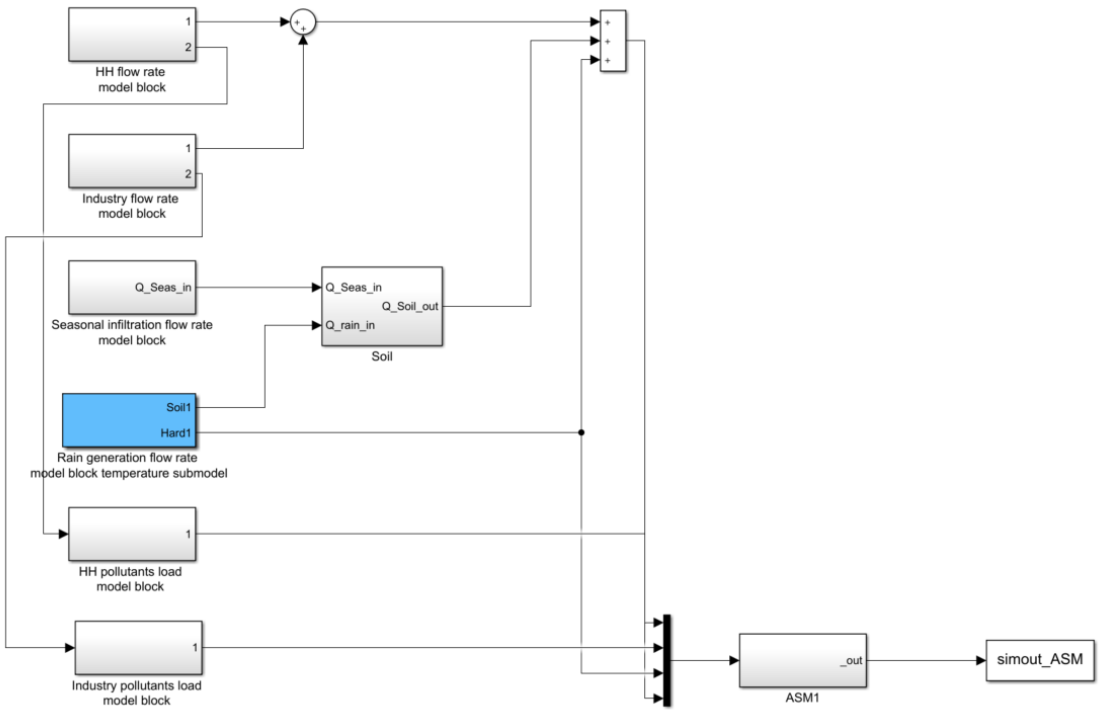


Figure 6-7 Layout of the BSM influent generator model (Gernaey et al., 2005)

The transport of snowmelt into the sewer system is a complex process. In the current IG, the snowmelt process is considered only when the ambient temperature is higher than zero. However, in practice, because of the effect of snow-melting agents, solar radiation or heat from the soil, etc., the snowmelt can also take place when the ambient temperature is below zero. This phenomenon can be described as shown in Figure 6-8: snowmelt can take place as long as the snow depth is positive and can even occur when the ambient temperature is negative up to a certain threshold, which represents a potential range of snow melting conditions thanks to the solar radiation or heat from the soil. If the snow depth temporal gradient is negative (snow depth is decreasing), the snowmelt rate is equal to this gradient. However, if the snow depth is increasing, snowmelt can still occur, and the snow is transferred to infiltration flow at a rate proportional to the air temperature and the snow depth.

$$\begin{aligned} if\ h_{(t+\Delta t)} > h_t: &\qquad flow = \frac{h_{(t+\Delta t)} - h_t}{\Delta t} \\ else\ if\ T_t > T_{threshold}: &\qquad flow = \frac{(T_t - T_{threshold}) * h_t}{k_{flowrate}} \\ else: &\qquad flow = 0 \end{aligned} \tag{6-3}$$

where h_t and T_t represent the snow depth and air temperature at time t, respectively, the $k_{flowrate}$ is a coefficient of transformation between snow height and temperature into flowrate with $T_{threshold}$ the temperature threshold below which infiltration will not occur.

To represent the infiltration delay between the snow depth decrease and the wastewater influent flow increase, a transfer function is added. The snowmelt water is split in two: the runoff on the impervious area leads to direct discharge to the combined sewer system, and the runoff from pervious areas will first go to the soil model and can then contribute to indirect infiltration to the sewers, with a calibrated delay.

The water temperature is also influenced by snowmelt infiltration. The process can be explained by the temperature sub-model presented in Figure 6-8. The water temperature generator consists of three parts: the temperature pattern, the ambient temperature influence and the infiltration correction factor. The decrease of temperature can be calculated proportionally to the amount of infiltration flow into the sewer. Two different gain parameters are applied in order to distinguish infiltration caused by rain or snowmelt.

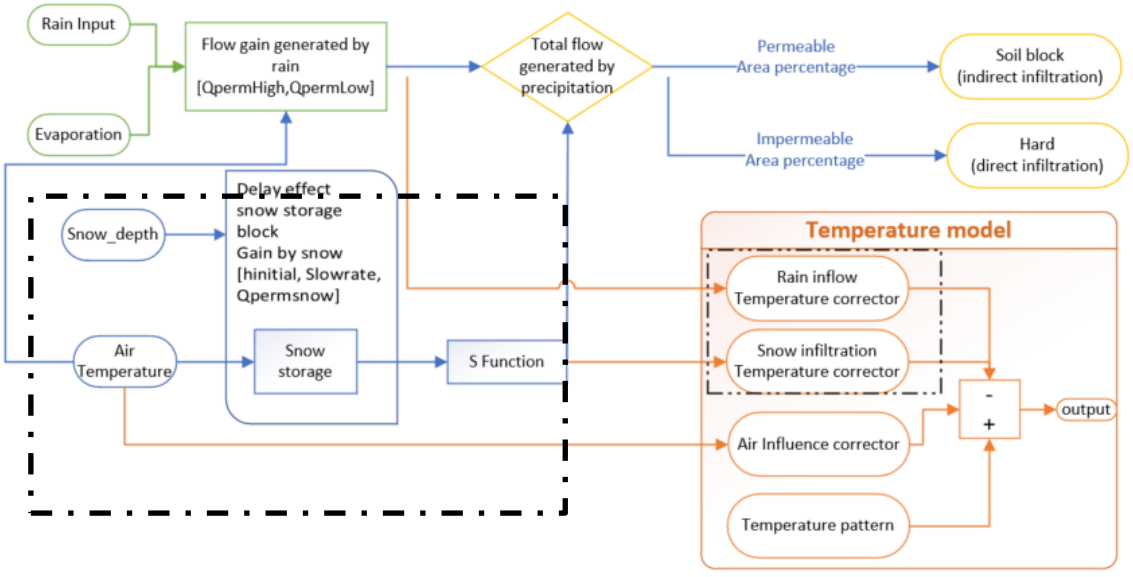


Figure 6-8 Layout of the modified rain generator model block and temperature sub-model of the phenomenological BSM influent generator. The two black dashed rectangles are two major modifications to handle the snowmelt.

The modified phenomenological model requires the input data set to include hourly weather information, i.e. precipitation, air temperature, and snow depth corresponding to the different inputs in Figure 6-8, as well as daily and seasonal patterns. Even though the original model provides default values for some of the parameters, a number of parameters remain to be calibrated for each case study.

The calibration performed included both manual and auto-calibration parts. The following parameters were calibrated manually: those related to the household diurnal profile and the industrial wastewater contribution, and also those of the seasonal soil infiltration pattern, which is described by two sine wave signals:

$$Soil\_infiltration(t) = a_k sin\left(\frac{2\Pi t}{365d} + phase_a\right)) + (b_k sin\left(\frac{2\Pi t}{182d} + phase_b\right) + bias \qquad 6\text{-}4$$

The other parameters, related to the snow storage block, S-function, and temperature model, were calibrated automatically by using the Levenberg-Marquardt parameter estimation method by minimizing the RMSE objective function(Levenberg, 1944). To minimize the possibility of local minima, the automatic calibration was started from different initial estimates of the model parameters. First, the flowrate was calibrated. Then, the second automatic calibration was started for the temperature sub-model in Figure 6-8, including the two temperature gains related to rainfall and snowmelt respectively. The air temperature correction gain was calibrated at the same time.

6.5 Results and discussion

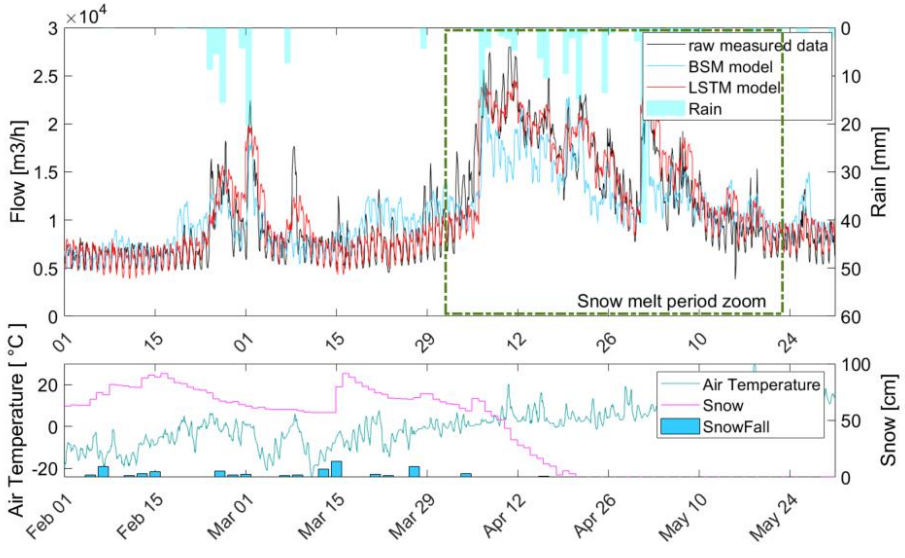
6.5.1 Qualitative comparison of the different models on test set simulation

The results of the different models will be analysed and evaluated for the test set, which is unseen during the learning process. Figure 6-9 shows that the data-driven LSTM model can successfully simulate the flowrate and water temperature variation in winter. The final LSTM network for flowrate generation consists of two blocks of LSTM layers (one for the low frequency data input branch and another for the high frequency input branch) with residual connection followed by one LSTM layer with sigmoid function for three gates and a tangent function to update the cell state. The output layer is a dense layer with linear function at the end. The two residual connected LSTM layers consist of 4 units, a number which was predefined to properly establish the model architecture. The number of units (8) in the last LSTM hidden layer was selected as the best performing in both training and cross-validation sets (Table 6-1).

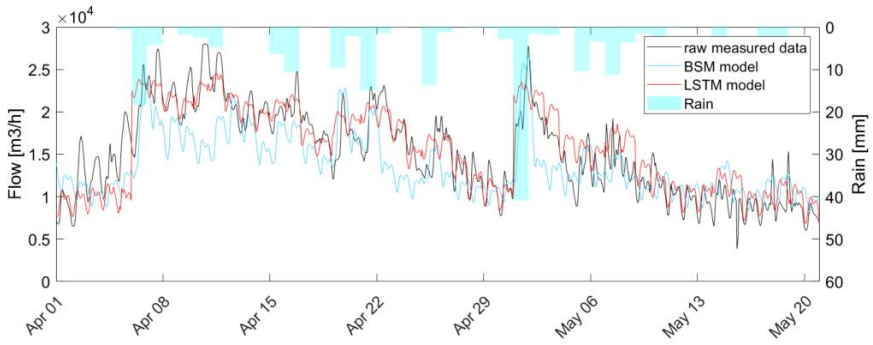
Table 6-1 Loss function result for different number of units in the last LSTM layer in the neural network model (architecture see figure 6) for the training and cross-validation set.

LSTM unit Number	Training Loss	Cross-validation Loss
5	0.0329	0.0311
6	0.0325	0.0326
7	0.0316	0.0303
8	0.0298	0.0287
9	0.0327	0.0295
10	0.0325	0.0332
11	0.0311	0.0317
12	0.0319	0.0313
13	0.0297	0.0318
14	0.0292	0.0337
15	0.0311	0.0346

Figure 6-9 shows the flowrate generation result for the test set for the winter period from February to the end of May 2017 and the bottom subfigure displays a zoom for the snowmelt period. The flowrate result demonstrates that there is a high dependency on the snowmelt and temperature. Indeed, the snow starts to accumulate from November until March when the ambient temperature starts to increase. It can be noticed that the snow can melt even when the ambient temperature is below 0 degrees because of snow-melting agents and rain effects.



(a)

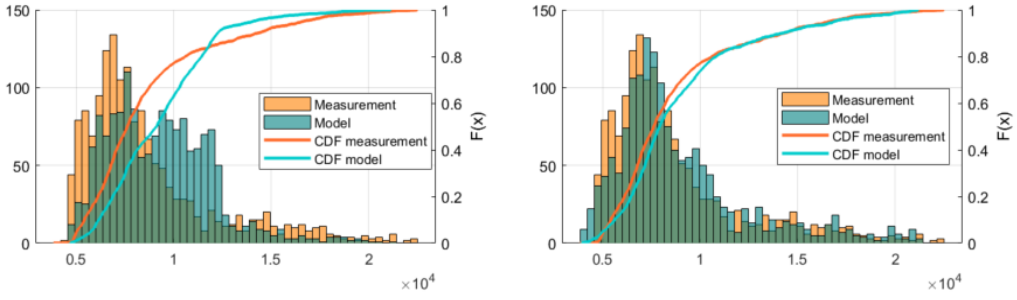


(b)

Figure 6-9 (a) Flowrate generation based on snow depth, snowfall, rain and air temperature: the black line represents the observed data, the blue line represents the BSM simulation result, and the red line represents the LSTM simulation result. Blue bars represent

There are two major flow peaks in the test set: the first, smaller one occurs in the beginning of March and the second peak, which is the maximum inflow, starts at the end of April and finishes in May. Such peaks happen almost each year with several weeks in between according to the particular snowmelt situation. In fact, as winter comes to an end, warmer and colder conditions start to alternate, with rain and snow fronts coming through. Indeed, the contributes to direct infiltration, with flow increases also depending on the precipitation intensities. The proposed LSTM-model can represent these quite complex phenomena.

In order to analyse the model performance in terms of the agreement between the statistical distribution of the measured and simulated data, probability density functions (PDF) and cumulative density functions (CDF) were evaluated, as shown in Figure 6-10. The plot demonstrates that the distribution of the LSTM modelling results is closer to the one of the measured data than the BSM influent generator's results.



(a) BSM phenomenological model

(b) LSTM data-driven model

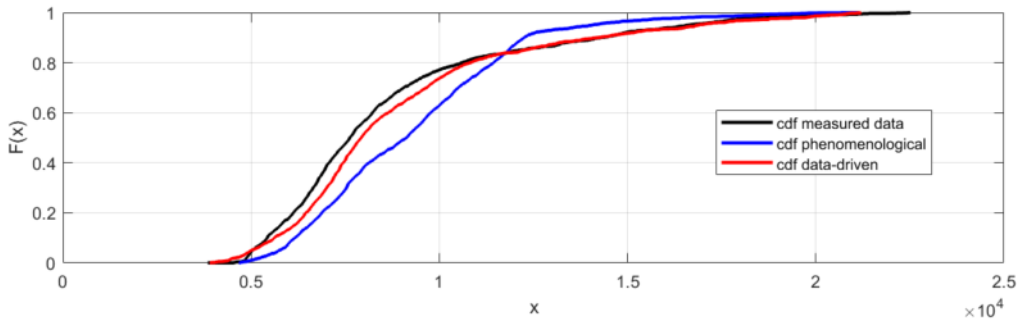


Figure 6-10 PDF and CDF comparison for flowrate generation by the phenomenological and data-driven model respectively.

Figure 6-11 exhibits the wastewater temperature result of both the BSM and LSTM model for the test set. It is shown that the influent temperature of the wastewater is influenced a lot by the snowmelt water infiltration. It can be observed that from the end of March till April, the ambient temperature is increasing while the water temperature is decreasing because of the large amount of snowmelt. Temperature decreases will strongly affect the nitrification rate (Plósz et al., 2009) and will thus influence effluent quality predictions. This phenomenon should thus be considered with more attention in view of system operation or process design in cold climate regions because of the bioactivity change in cold region WRRFs. Figure 6-12 exhibits the PDF and CDF for the temperature modelling results. The difference between the phenomenological and LSTM model is quite dramatic.

As an example of better performance, the LSTM model reacts significantly faster than the BSM model to the rain event of 12 April. This is probably because the LSTM model is more sensitive to precipitation in winter, while the BSM is more restricted in dealing with these occurrences. This property can also be observed for the rain event on March 29th, when the BSM water temperature simulation decreases below the actual water temperatures. Even though the BSM model performed very well with the training dataset (results not shown), it does not as well for the testing dataset. For instance, the BSM model predicts lower water temperatures between March 01 to 15 which is very different from the measurements and the data-driven model prediction. This deviation is mainly due to the air temperature influence corrector, the temperature pattern and the evaporation (see Figure 6-8), which are based on a harmonic function calibrated on the training dataset. However, the seasonality of this set is very different from the test set. This issue could be solved by improving the calibration by shifting the sine harmonic function.

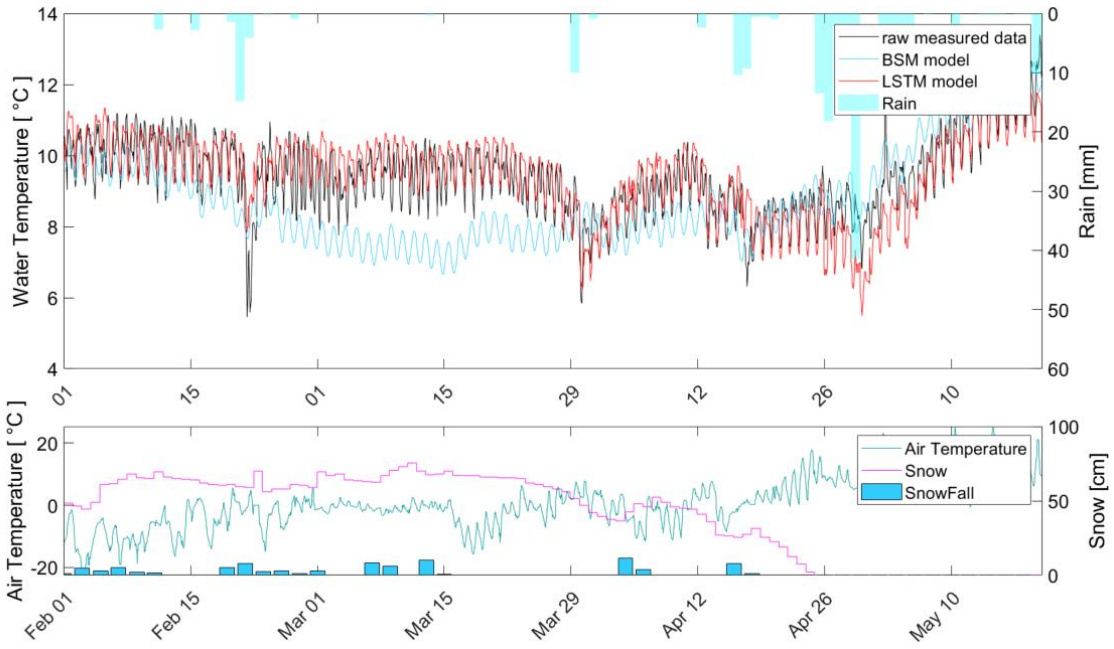
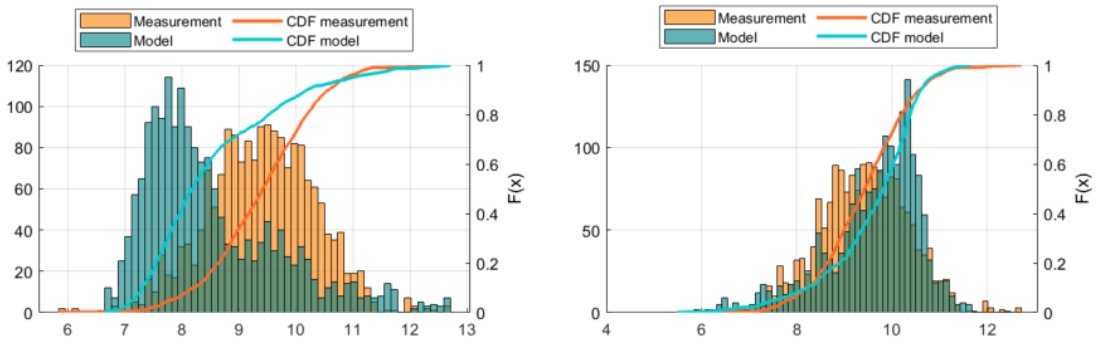


Figure 6-11 Influent temperature generation for the test set covering February to May 2018.



(a) BSM phenomenological model *(b) LSTM data-driven model*

Figure 6-12 PDF and CDF analysis of predicted influent temperature. (a) and (b) represent the BSM model and LSTM model results, respectively

6.5.2 Quantitative comparison and analysis of the different models

In this study, the performance of the different IG models was evaluated by the following quantitative criteria: mean absolute percentage error (MAPE), Nash-Sutcliffe Efficiency coefficient (NSE) and the Kullback-Leibler divergence (KL divergence). The MAPE is a commonly used metric for machine learning and the NSE is commonly used in mechanistic model performance assessment.(Hauduc et al., 2015) The KL divergence is an

indicator of the distance between two probability distributions(Kullback and Leibler, 1951). A smaller KL-divergence between two statistical distributions means that the statistical distributions of the two time series compared are closer to each other. The results for these different criteria are presented in Table 6-2. By comparing these common statistical indices, the model results can be evaluated from different perspectives.

$$MAPE = \frac{1}{n}\sum\left|\frac{\hat{y}_i - y_i}{y_i}\right| * 100\%$$
$$Nash - Sutcliffe = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad 6\text{-}5$$
$$KL_{divergence}(P\|Q) = -\sum P(x) * \log\left(\frac{Q(x)}{P(x)}\right)$$

where $\hat{y}_i$ is the simulated and y_i is the observed data, $P(x)$ is the probability distribution of the observed data and $Q(x)$ is the distribution of the simulated data.

Table 6-2 Summary of model performance

	LSTM with residual connection	**BSM_IG model**
MAPE flowrate	14.6%	24.7%
MAPE temperature	7.2%	9.1%
KL-Divergency flowrate	0.18	0.27
KL-Divergency temperature	0.11	0.41
NSE flowrate	0.81	0.66
NSE temperature	0.68	0.40
Relative calibration efforts	Fast: 2-3 days	Slow: more than a week
Training set data needed	6 – 7 k (hourly data~=300days)	less

In this study, the proposed LSTM model outperformed the modified BSM phenomenological for influent generation. Two possible explanations are proposed. First, the snowmelt sub-model that was modified in this work is not sufficiently complex to represent the real system. The snowmelt-induced infiltration probably exhibits more complex nonlinear behaviour and this may be more difficult to achieve by the conceptual model that was originally included in the BSM IG. Thanks to the LSTM architecture, and the residual connection, the data-driven model is able to 'remember' a very long-term effect of the snow, with inputs of a week or more before still influencing current conditions. However, in the BSM model, the snow-melting rate only depends on the current condition. Thus, the LSTM permits a better simulation than the BSM model, including such important features as at the peak in flowrate, where the LSTM can keep increasing whereas the BSM result already decreases before reality does. On the other hand, the decrease of snow depth can be caused not only by melt, but also by

the compaction of snow, or due to snow evacuation by wind. Additionally, the model didn't consider that the density of snow varies between 0.1 and 0.5 g/cm^3, which explains why the phenomenological model is expected to be biased due to its over-simplified model of this complex process. In contrast, the LSTM model can simulate this behaviour by providing it with a diversity of related data (temperature, precipitation etc.) thanks to its 'black box' characteristics and nonlinear learning ability.

The second proposed explanation of the lower performance is that the calibration performed may not have reached the global minimum. Even though the model was trained by both manual and automatic calibration steps and initiated from different initial estimates, the limited number of calibration iterations performed and the lack of data may have caused that the best parameters for the BSM IG model were not found.

Regarding the needed calculation times for simulation and the efforts for calibration, the machine learning model is more automated, faster to run and easier to train even though one must recognize that the computation time of the data-driven model increases as the number of layers and neurons increases.

On the contrary, it should be acknowledged that the calibration of a mechanistic or a phenomenological model in wastewater systems is hard to automate and is therefore often expert-based and thus strongly depends on the modeller(Rieger et al., 2012b). Therefore, the calibration efforts can vary between different experts performing the calibration and the result reported here may thus be biased. Still, although a well-trained expert may require less calibration time, both manual and automatic calibration efforts will still be required.

Finally, it is worth mentioning that phenomenological models are more interpretable than data-driven models thanks to their physical parameters that can be adjusted according to engineering experience or direct measurement. Instead, it is more difficult to interpret data-driven models, and they always need sufficient data to achieve good calibration. It is important to highlight that, in absence of data sets, the BSM phenomenological model can still offer valuable insights of a WRRF influent, even though it may be limited in quality. Indeed, such prior knowledge-based model allows for transferability to other case studies and get a rough result, but data and considerable calibration efforts will still be required in order to obtain an adequate model for each new case study.

Overall, the authors believe that the proposed LSTM modelling approach can better balance accuracy and modelling efforts and provide good simulation results for flowrate and temperature of urban wastewater influents under cold climate conditions.

6.6 Conclusion and perspectives

In this study, an original data-driven model is proposed based on the LSTM with residual connection architecture. The results presented in this paper demonstrate that the machine learning method provides good performance for WRRF influent data modelling and simulation. It enables a simple, effective and accurate simulation for WRRF influent flowrate under the influence of snowmelt. This result will contribute to solve issues when modelling WRRFs for operation and water management in cold regions.

The proposed LSTM model input is only based on weather information (rain, snow depth, ambient temperature). Once the model is calibrated, the flowrate can be generated for a very long timescale. Compared with the popular BSM phenomenological model, the machine learning method has a better performance in terms of MAPE and the agreement between the probability distributions of simulated and measured time series. Also, in contrast to the phenomenological IG, it does not ask the modeller to manually calibrate some of the physical parameters, thus reducing the modelling and calibration efforts.

In conclusion, the proposed LSTM model can generate flow rate and water temperature dynamics under the influence of snowmelt and storm water at the end of winter, with better prediction performance than the BSM phenomenological model. The model is able to generate a time series with a probability distribution that is more similar to reality, allowing for a good influent description for WRRF design, and operational decision-making for cold climate conditions. This model can also contribute to digital twin applications, characterized by a bidirectional connection between the model and live data, allowing for automatic dynamic updating of the influent profile, and thus for future wastewater treatment automation modelling studies(Torfs et al., 2022) .However, considering the drawbacks of data-driven models, further research will concentrate on creating a hybrid model that takes advantage from both prior knowledge (extrapolation power) and data-driven modelling (improved accuracy and easier calibration). The validation of the proposed LSTM-based IG approach should also be conducted for other case studies in order to evaluate the potential of its transferability and extensibility.

6.7 Special acknowledgement

The sharing of the BSM Influent generator, created at the Electrical Engineering and Automation (IEA), Lund University, Lund, Sweden, by Dr Krist V. Gernaey, Dr Xavier Flores-Alsina, Dr Ramesh Saagi, Dr Ulf Jeppsson et al. is gratefully acknowledged.

Chapter 7. Data-driven influent generator for WRRF database gap filling and model-based control evaluation

This chapter is in preparation:

Data-driven influent generator for WRRF database gap filling and model-based control evaluation. In preparation, submitted for proceeding IWA World Water Congress & Exhibition 2022

7.1 Abstract

Dynamic and high frequency wastewater influent data are required for water resource recovery facility (WRRF) modelling and optimization. It is advantageous for practise if influent data generation can be based on routine data collected at high frequency (e.g., filling the gaps between daily measurements data) and can provide a longer forecasting time horizon. In this study, a data-driven influent generator model is proposed and tested. The model is able to generate high frequency pollutant concentration time series from low frequency routine measurements. In addition, influent prediction has been tested for different time horizons and good performance was demonstrated up to 4 hour prediction provided weather information is available for the forecast horizon. This generator can be widely applied to WRRF modelling and real-time control system evaluation.

7.2 Résumé

Des données dynamiques et à haute fréquence sur les affluents d'eaux usées sont nécessaires pour la modélisation et l'optimisation des Station de récupération des ressources de l'eau (StaRRE). Il est avantageux pour la pratique si la génération de données d'affluent peut être basée sur des données de routine collectées à haute fréquence (par exemple, combler les lacunes entre les données de mesures quotidiennes) et peut fournir un horizon temporel de prévision plus long. Dans cette étude, un modèle de générateur d'influent basé sur les données est proposé et testé. Le modèle est capable de générer des séries chronologiques de concentrations de polluants à haute fréquence à partir de mesures de routine à basse fréquence. De plus, la prévision des affluents a été testée pour différents horizons temporels et de bonnes performances ont été démontrées jusqu'à 4 heures de prévision à condition que les informations météorologiques soient disponibles pour l'horizon de prévision. Ce générateur peut être largement appliqué à la modélisation StaRRE et à l'évaluation du système de contrôle en temps réel.

7.3 Introduction and background

Nowadays, the growing amount of data from in-situ sensors installed in wastewater systems is increasingly found useful to automatically identify abnormal behaviour and ensure high data quality for modelling and automatic control for WRRF optimization. Modelling is playing an important role in automation and optimization of WRRF (Sweeney and Kabouris, 2015). However, the use of many models is limited by a lack of adequate input datasets, poor quality of influent data or low time resolution data and so on. For instance, for WRRF design, engineers usually make initial sizing by using design guidelines based on average loads and safety factors, instead of a dynamic input time series (Talebizadeh et al., 2016). As another example, the AvN (ammonia versus nitrite/nitrate) controller is important in view of preparing the effluent of a nitrogen-removing WRRF for a subsequent Anammox (anaerobic ammonium oxidation) process for nitrogen removal. Influent ammonia forecasting can be expected to improve controller performance, but a high-quality and real-time ammonia prediction is not always available. Good quality influent data in terms of precision and predictive ability is therefore becoming crucial for more advanced WRRF modelling and process control development.

Heretofore, a large number of influent generators (Gernaey et al., 2011a; Martin and Vanrolleghem, 2014) and influent predictive models have been presented in literature. They can be divided into two categories: phenomenological models (i.e. mechanistic, knowledge based) (Flores-Alsina et al., 2014b; J Langeveld et al., 2017) or data-driven models (i.e. black box, data based) (Borzooei et al., 2019; Li et al., 2020). Machine learning and deep learning approaches are increasingly applied in the water domain, soft sensors are more and more developed for process monitoring, trials are made with unstaffed WWTP, methods are proposed for anomaly detection in data time series, and algorithms for short-term predictions are implemented, etc (Gopakumar et al., 2018; Russo et al., 2021; Schneider et al., 2020; Shokry et al., 2018)

Even though data-driven influent prediction models have already been presented in literature, the following challenges are still remaining:

The first challenge is that influent databases are often not complete or continuous in time. For instance, the TSS concentration can be measured by a relatively simple turbidity sensor, whereas the COD or ammonia concentration is usually measured in the inlet only on a daily basis. However, for modelling purposes, the COD and ammonia concentrations are expected to be available at a higher frequency, allowing process control and better water treatment quality management.

The second challenge is the need for long-term horizon prediction for process control. The long-term prediction is usually required to predict important variables over a time horizon corresponding to the hydraulic retention time in the bioreactor (typically 4-8 hours), in order to anticipate control actions, such as a change of DO set

point. The fact is that many proposals have been made with one step ahead prediction or short-term forecasting (El-Din and Smith, 2002; Ma et al., 2014; Zhu and Anderson, 2016). However, it is easily understood that pursuing longer time horizon prediction will enhance controller performance.

Therefore, in this research, an influent prediction model has been developed, which aims to create a higher frequency time series and a longer time horizon prediction, by using a limited dataset, as typically available in practice from routine measurements.

To solve the issues with available influent generation and predictive models, in this study, an advanced data-driven model is proposed that can generate dynamic WRRF influents at different time resolutions (hourly, 20 min) and different time horizons (one timestep, multi timestep) in order to adapt to different user demands. The model aims to generate a realistic dynamic profile for the influent pollutants which are not continuously measured. This model will be favourable to reduce the monitoring cost and to increase the quality and completeness of the influent dataset.

7.4 Materials and Methods

7.4.1 Qualitative comparison of the different models on test set simulation

In this study, the modelling approach was developed using data collected from the pil*EAU*te, a pilot-scale WRRF located at Université Laval, in Québec, see Figure 3-1. The online data collection system dat*EAU*base (Plana et al., 2019) stored data of the influent concentrations and the pil*EAU*te bioreactor. The pilot catchment receives the wastewater from a student dormitory and two kindergartens through a small combined sewer system, with a very variable pollutant concentration under different weather conditions. The influent shows recurring hourly and daily variations but is disturbed by rain events.

Data pre-treatment was conducted according the procedure shown in Figure 7-1, including outlier removal, data smoothing and normalization, and faulty data detection (Alferes et al., 2013).

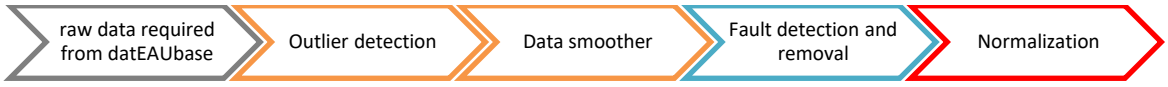


Figure 7-1 Data pre-treatment procedure, including outlier removal, data smoothing and normalization, and faulty data detection (Alferes et al., 2013).

It is known that the rain events and ambient temperature have an important influence on the influent of a WRRF. The weather information is obtained by an open weather station located at the university campus and is applied into the model input. Besides, the daily pollutant pattern under dry weather flow conditions, which represents the behaviour of the people served by the sewer system. is extracted from the data using a Chebyshev bandpass filter (Schlichthärle, 2011), see Figure 7-2.

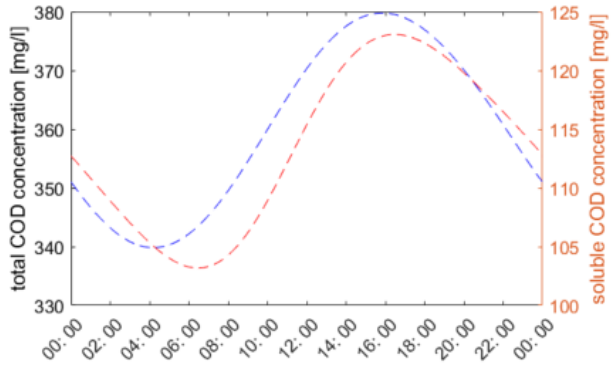


Figure 7-2 Total COD and soluble COD concentration for the training phase, diurnal pattern extracted from the online data obtained with a spectro::lyser (s::can, Vienna, Austria)

7.4.2 LSTM model and residual connection

Recurrent neural networks (RNN) are a category of neural networks that is widely used for time series modelling problems. The Long Short Term Memory (LSTM) model is a special structured RNN, composed of different so-called gates. It has been demonstrated that LSTM is very powerful for water quantity and quality prediction for different timescales, such as water demand predictions, anomaly detection and water quality generation etc. (Li and Vanrolleghem, 2021; Mu et al., 2020).

The proposed model is developed based on the Python environment with the machine learning platform TensorFlow and Keras library (Chollet, 2015). The model architecture is a hybrid with a dense layer (fully-connected) and a LSTM layer with different numbers of neurons for the different layers. A residual connection has been inserted between layers. In addition, the residual connection architecture is adopted to ensure that, while increasing network depth, the performance of the neural network will not degrade, see Figure 7-3. This residual connection architecture provides more efficient training and has been successfully applied in different deep learning studies (He et al., 2016; Wu et al., 2017).

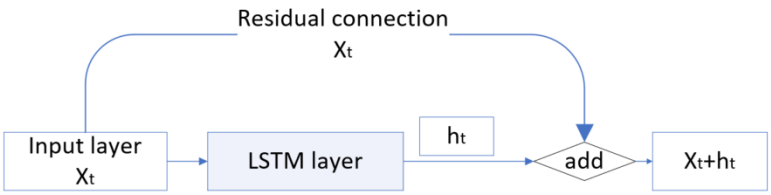


Figure 7-3 LSTM with residual connection

During the training process, the dataset was split into three sets: a training set, a cross validation (CV) set and a test set. The training and CV set are shuffled and contain 80% and 20% of the data, respectively. The test set is an unseen time series data, which is different from the training and CV sets.

Figure 7-4 represents the input time series for the different variables with different frequencies and Figure 7-5 represents the architecture of the retained LSTM and residual connection principle used in this study. This architecture was inspired by the work reported in the methodology section 6.4 of Chapter 6. Part of the hyperparameters has been pre-defined (the type of layers and number of layers) in order to establish the architecture and minimize the modelling efforts (for details, see section 6.4.2). The other part of the hyperparameters (number of neurons, the activation functions, the learning rate) has been trained using the root mean square error as objective function and by performing a grid search in order to decide the optimal architecture based on the CV set.

Time series	…	t-3d	…	t-2d-15min	t-2d	t-1d-30min	t-1d-15min	t-1d	…	t-90min	t-75min	t-60min	t-45min	t-30min	t-15min	t
Conductivity																
TSS																
Pattern																
Weather																
COD mean		t-3d		t-2d				t-1d	t (daily average)							
TSS mean		t-3d		t-2d				t-1d	t							
CODs mean		t-3d		t-2d				t-1d	t							
COD output									…				**t-45**	**t-30**	**t-15**	**t**
CODs output													**t-45**	**t-30**	**t-15**	**t**

Figure 7-4 Input variables time series with frequencies of 15min, 1 day respectively

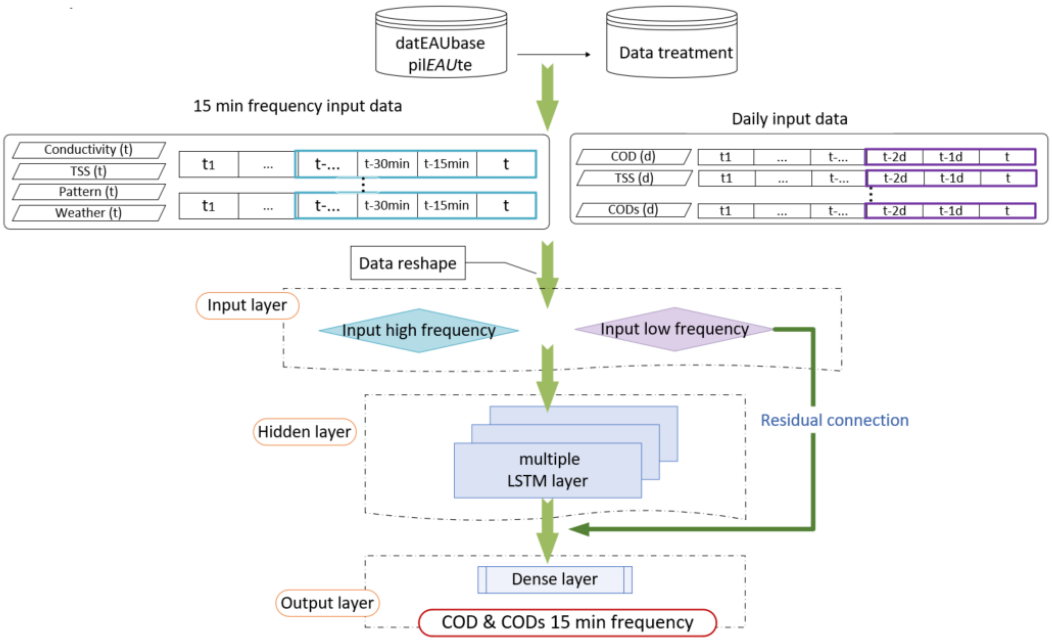


Figure 7-5 Modelling architecture based on LSTM and a residual connection, details see section 6.4.2.

7.4.3 *NARX RNN model*

The multilayer perceptron (MLP) is a class of feedforward ANN where each neuron in a layer is fully connected to the next layer. It is one of the most used ANN architectures and is used in this study as the reference. The NARX (nonlinear autoregressive network with exogenous inputs) is a particular recurrent dynamic network based on MLP, which has already been used in different studies (Banihabib et al., 2019; Boussaada et al., 2018). The defining equation is:

$$\hat{y}(t+1) = F\begin{pmatrix}\hat{y}(t), \hat{y}(t-1), \dots \hat{y}(t-\tau) \\ x(t), x(t-1), \dots x(t-\tau)\end{pmatrix} \qquad 7\text{-}1$$

where F is the function of the neural network, $\hat{y}(t+1)$ is the output of the NARX at time *t* for time *t+1* (predicted value), $x(t)$ represent the exogenous inputs of the NARX, and lag τ the time delay.

In this research, a NARX model is trained in order to compare it with the LSTM model for multi timestep forecasting performance. The NARX is fully connected with a sigmoid function in the hidden layer, in order to represent the nonlinearity.

7.4.4 Performance indicators

Four different indicators given in Table 7-1 Performance indicators for model result evaluation and mathematical were calculated to evaluate the modelling results on the test set: the relative root mean square error (relative RMSE), the mean absolute percentage error (MAPE), the Nash-Sutcliffe efficiency coefficient (NSE) and the Index of Agreement, where C_t is the measured value at time t and $\hat{C}_t$ is the modelled value at time t and $\bar{C}$ is the mean of the measured values, N is the total number of forecasted values in the test set.

Table 7-1 Performance indicators for model result evaluation and mathematical formula

Indicator	Mathematical formula
RMSE	$\frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(C_t - \hat{C}_t)^2}}{\bar{C}} * 100\%$
MAPE	$\frac{100}{N}\sum_{t=1}^{N}\frac{(C_t - \hat{C}_t)}{C_t}$
NSE	$1 - \frac{\sum_{t=1}^{N}(C_t - \hat{C}_t)^2}{\sum_{t=1}^{N}(C_t - \bar{C})^2}$
Index of Agreement	$1 - \frac{\sum_{t=1}^{N}(C_t - \hat{C}_t)^2}{\sum_{t=1}^{N}(\lvert C_t - \bar{C}\rvert + \lvert\hat{C}_t - \bar{C}\rvert)^2}$

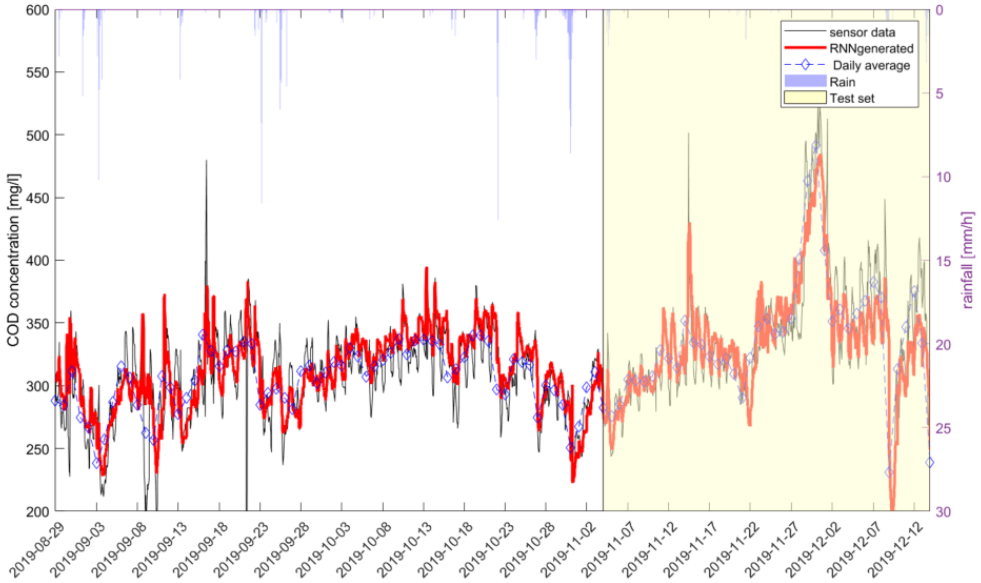
7.5 Results and Discussion

7.5.1 Results for high resolution water quality generation

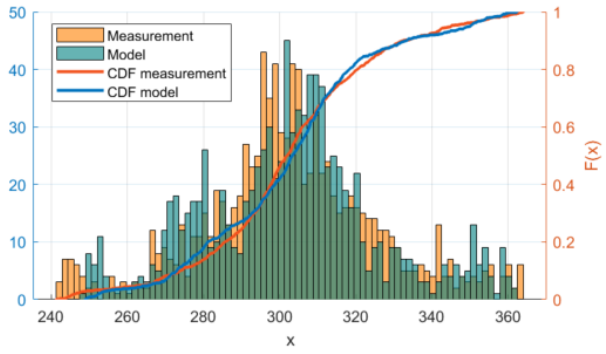
As discussed above, the first objective of this study was to generate a high frequency time series from low frequency measurements. In reality, the total COD and soluble COD are usually measured daily while hourly TSS and conductivity data, as well as weather data (air temperature and rain intensity) can be more easily obtained at low equipment cost. As in many modelling processes for optimization and control, an influent time series with a 15 min (quarter hour) timestep is desired.

A total COD time series at a 15 min frequency was pursued by the trained model. The sensor data is represented by black line and the model output data is represented by the red line and the diamond markers represent the daily average of COD data, see Figure 7-6. The result shows a very good match between the observed data and the simulated data in both training set and test set.

The Cumulative Distribution Function (CDF) and Probability Density Function (PDF) of Figure 7-6 demonstrate that the statistical distributions of the model results and the observed data are very close, which means that the model can capture the variability of reality.



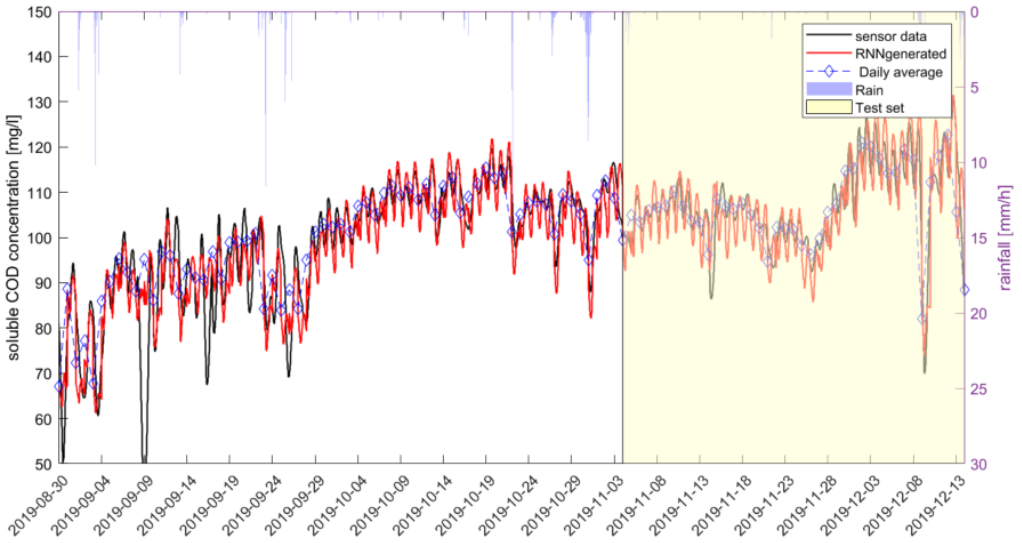
(a)



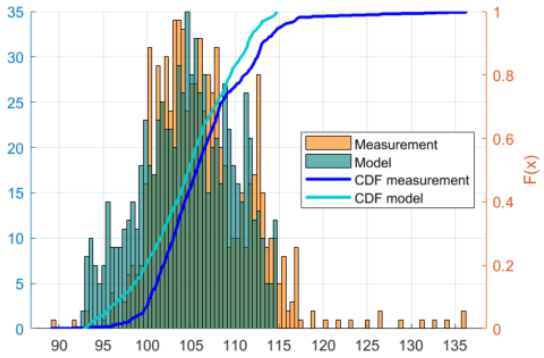
(b)

Figure 7-6 Total COD concentration simulation results (red line) compared to the training and test set (black line) (a), and the PDF and CDF for the test set (b)

The model is also trained for generating soluble COD concentrations, shown in Figure 7-7. The result again shows good performance, and the concentration dynamics have been adequately represented. A difference can be observed in the drastically lower sensor values around 09-10 September, but the model result didn't present the same behaviour. The explanation is the following: during the cleaning process that was conducted that day, the sensor was providing outlier data and recovered after maintenance.



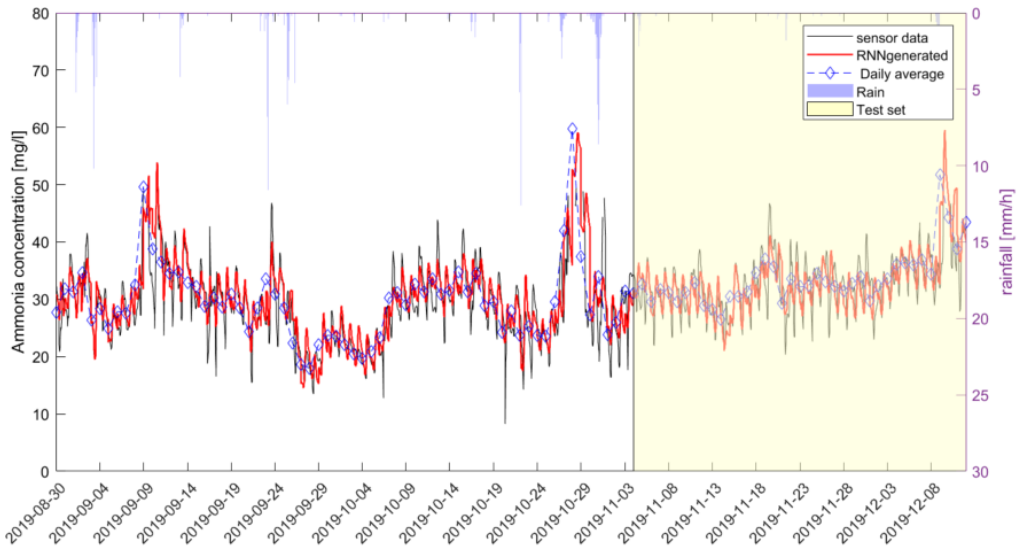
(a)



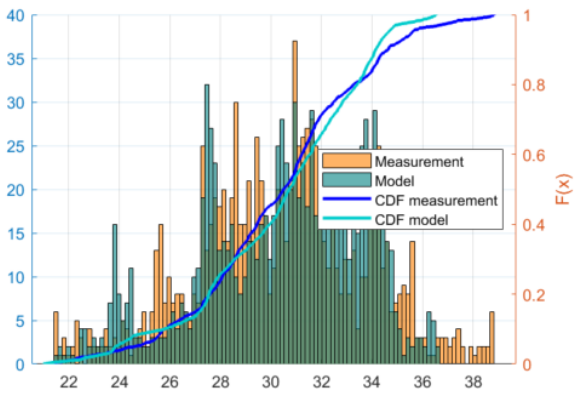
(b)

Figure 7-7 Soluble COD concentration simulation results (red line) compared to the training and test data (black line) (a), and the PDF and CDF for the test set (b)

As a main source of nutrients in wastewater, having ammonia concentration data at high frequency can be crucial for WRRF process optimization and resource recovery: Figure 7-8 shows the LSTM model results for 15min ammonia concentration predictions. During two weekends (07-08 Sep and 26 -27 Oct 2019) the sensor clogged so that the data was drifting with the value recovering after Monday's maintenance. Comparing with the sensor data with many peak value and outlier data, the model results are smoother.



(a)



(b)

Figure 7-8 Ammonia concentration simulation results (red line) compared to the training and test set data (black line) (a), and the PDF and CDF for the test set (b)

As mentioned in the section Performance indicators, the model performance was evaluated using the indicators presented in Table 7-2. The result shows that the predicted total COD and soluble COD concentrations are obtained with a very good accuracy. Despite the fact that the ammonia concentration accuracy is slightly lower, the performance is still very high. The reason for the lower performance may be that the correlation between the ammonia and TSS, which is one of the input variables, is more complex, which leads to a more difficult training

of the model. Another reason may be that in this case study, flowrate data are not available. This is expected to enhance the model prediction performance.

Table 7-2 Model performance evaluation

	Total COD	Soluble COD	Ammonia
Relative RMSE	6.07%	5.56%	8.06%
MAPE	4.42%	5.74%	7.89%
NSE	0.60	0.64	0.61
Index of Agreement	0.89	0.91	0.87

7.5.2 Results for multi time-step forecasting

The second useful application of an influent generator is for multi timestep forecasting. First, the result for ammonia concentration forecasting for one timestep (1h) ahead is presented.

In Figure 7-9, the results of the MLP-NARX and residual connected LSTM models are compared on the test set. It is shown for all performance criteria that the residual connected LSTM is better for the one hour ahead prediction, see Table 7-3.

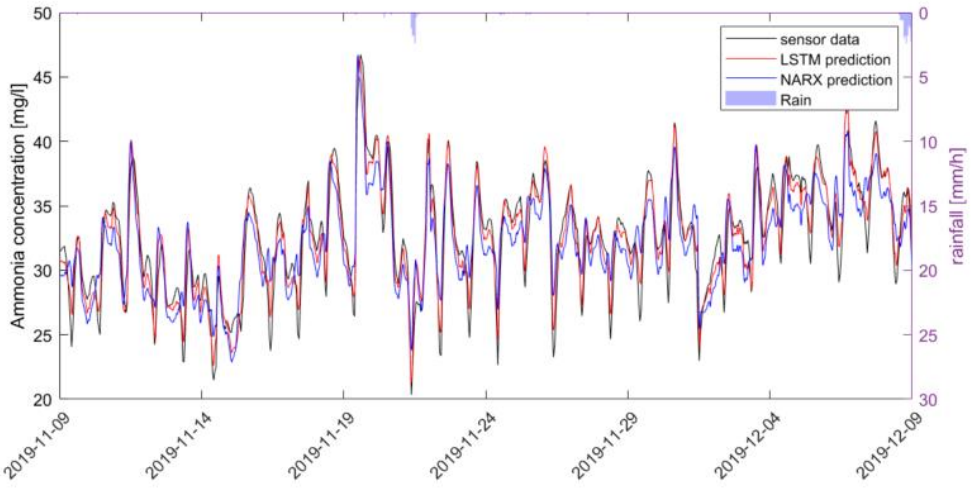


Figure 7-9 Result of MLP-NARX and residual connected LSTM for one timestep prediction of MLP-NARX and residual connected LSTM for one timestep prediction

Table 7-3 Result analysis for one timestep prediction comparing LSTM and NARX-ML models

	LSTM with residual connection	NARX-MLP
Relative RMSE	3.03%	6.50%
MAPE	3.52%	7.06%
NSE	0.93	0.71
Index of Agreement	0.98	0.91

Then, because modelling studies may not only require a one timestep but also a multi timestep prediction, the model has also been trained for multiple timesteps (from one hour ahead to six hours ahead) in order to evaluate whether a good balance between accuracy and prediction horizon could be maintained. In this study, instead of using a recursive strategy in which the prediction error may accumulate, a "many-to-many model" was trained:

$$prediction(t+4)\ldots prediction(t+1) = model(obs(t-1), obs(t-2), \ldots, obs(t-n))$$

Table 7-4 Input variables time series with frequencies of 15min (blue), and hourly output time series (red).

Time series	…	t-24h	….	t-4h	t-3h	t-2h	t-1h	t	t+1h	t+2h	t+3h	t+4h
Conductivity												
TSS												
Pattern												
Weather												
COD output									**t+1h**	**t+2h**	**t+3h**	**t+4h**

The first scenario of Figure 7-10 shows the model result without exogenous input of forecasted weather information (forecasted rain and air temperature). Table 7-5 shows that the performance decreases with increasing number of timesteps. A four-hours ahead time horizon appears feasible, allowing future applications.

However, as shown in Figure 7-10, the model is not capable of long-term prediction (see the rain event of 27 November) if the rain that happens within the prediction time horizon is not considered in the input. This is logical, because the model did not get any information that may be inducing the concentration change. However, once the model gets the historical rain data, it is capable to capture the concentration dynamics again.

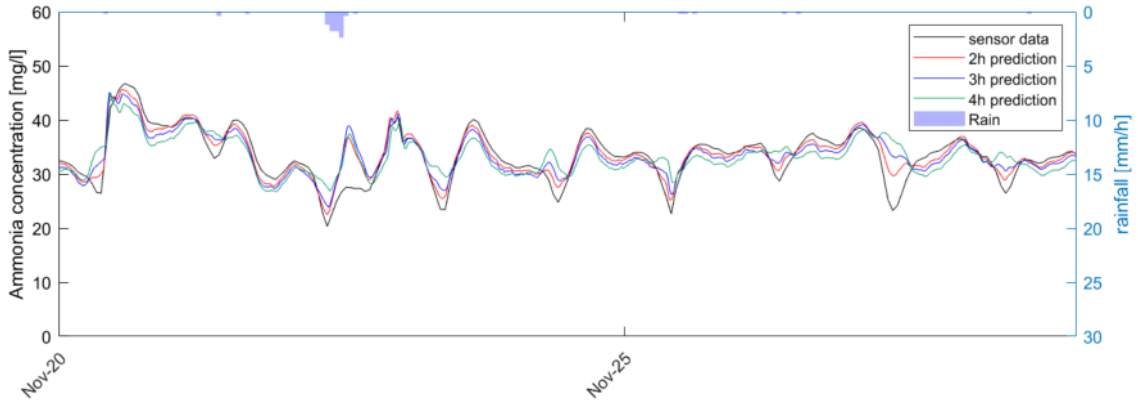


Figure 7-10 Multi-timestep prediction of ammonia without future weather information input

Table 7-5 Performance analysis of multi timestep prediction of ammonia without future weather information input

	Multi timestep prediction 2h	Multi timestep prediction 3h	Multi timestep prediction 4h
Relative RMSE	3.80%	7.22%	9.87%
MAPE	4.99%	5.86%	8.36%

The second scenario consisted of training the model with the exogenous input of future weather information. Indeed, the weather can usually be forecasted over a horizon of a few hours and it has an important effect on the influent profile change. Forecasted weather information (rain and air temperature) for 3h and 4h was applied.

$$Prediction\ (t+4) = model\ (\mathrm{obs(t-1)}, \mathrm{obs(t-2)}, \ldots, \mathrm{obs(t-n)}, \mathrm{exogenous(t+1)}, \mathrm{exogenous(t+2)}..\mathrm{exogenous(t+4)})$$

Table 7-6 Models for multi timestep prediction: Input variables time series with frequencies of 15min (blue), and hourly output time series (red).

Time series	…	t-24h	….	t-4h	t-3h	t-2h	t-1h	t	t+1h	t+2h	t+3h	t+4h
Conductivity												
TSS												
Pattern												
Weather												
COD output												**t+4h**

Figure 7-11 shows the results for 3 and 4h ammonia predictions on the test set. The improvement compared to the results in Figure 7-10 can be noticed, i.e. the model reacts more quickly and accurately during and after the rain event. The criteria reported in Table 7-7 also indicate that the model with exogenous input outperforms the previous model.

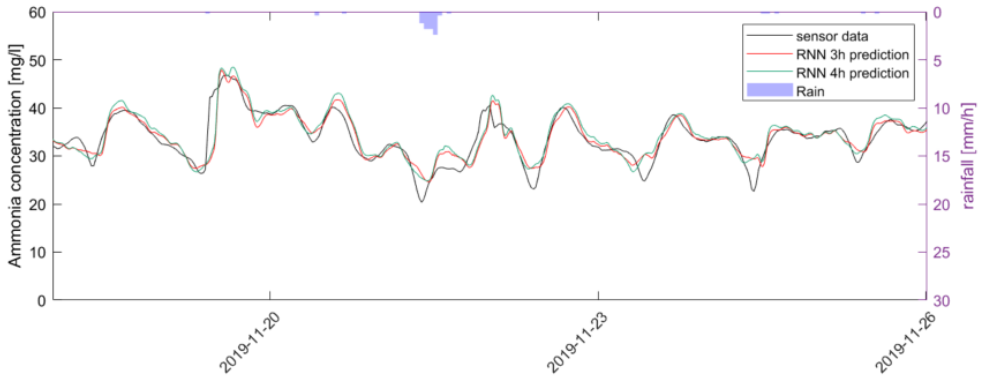


Figure 7-11 Multi timestep prediction of ammonia concentrations with exogenous input of weather predictions over 3 and 4h respectively.

Table 7-7 Performance analysis of the multi timestep prediction of ammonia with exogenous input of 3 and 4h weather prediction respectively.

	Multi timestep prediction 3h	Multi timestep prediction 4h
Relative RMSE	6.45%	7.28%
MAPE	4.89%	5.75%

7.6 Conclusion

A data-driven methodology was proposed and successfully applied to generate influent pollutant concentration time series, dealing with two main challenges in WRRF influent generation practice. The first model can generate a higher frequency time series (hourly) than the frequency of available concentration data (daily), thus allowing WRRF modelling and simulation based only on low frequency routine influent measurements. In other words, thanks to the data-mining approach, a more detailed and more dynamic influent profile can be generated by using the model, without requiring complex mechanistic modelling and calibration efforts and without investments in additional data collection.

The second achievement of this study is the multi timestep forecasting of the ammonia concentration. It was shown on the case study that the model can predict the concentrations four hours ahead. For the two objectives of this work, it was demonstrated that the LSTM with residual connection architecture is able to predict not only one timestep ahead but also multi timestep. Exogenous input of weather information significantly enhances the model forecasting capacity and improves the prediction precision.

Conclusions and Perspectives

Conclusions

In this PhD project, the characterization of influents for a variety of model applications has been studied and different influent generator models have been developed, built, and applied for generating WRRF influent time series, for flowrate and pollutant concentrations with different time resolutions and time horizons.

First, at source urban wastewater has been characterized in order to advance the understanding of influent dynamics, and the experimental results have been represented by a phenomenological model for the small catchment under study. Then, a data-driven influent generator (IG) method was developed based on machine learning approaches, with different neural network architectures, including plain ANN, RNN, LSTM, NARX, using input data from different information sources. The approach adopted is based on routine WRRF data, and can thus be applied as a very practical tool in both research and engineering practice.

To better illustrate the feasibility and the performance of the proposed IG model, it was applied to different case studies and compared with conventional methods (statistical regression models and the BSM phenomenological model). Furthermore, the model has been optimized for different objectives, regarding both the generated time series' precision and the correspondence with the observed variability. The IG have also been applied to two case studies (Québec, QC, Canada and Bordeaux, France) in order to confirm the adaptability of the data-driven model and to demonstrate that it is a pragmatic model. The IG was found to be able to deal with different influent issues.

Overall, the proposed data-driven method has been successfully developed and applied for different types of objectives:

- The availability of dynamic influent data has always been a major bottleneck for the application of the ASM model to evaluate the design and operational strategies of WRRF. The proposed influent generator models that use machine learning methods, are based on available historical data, which overcomes the difficulties of data collection investments.

- The nonlinearity and complexity of the process of wastewater generation in a catchment and the transport in the sewer system make that the phenomenological or mechanistic modelling of influent dynamics is not trivial and may require several time-consuming and expensive experiments. The proposed data-driven IG models avoid these investments of effort spent on experimentation and modelling, and therefore, enable applications in engineering practice without being hampered by the limited availability of time or financial resources.

- The input data is selected based on prior engineering knowledge and time series analysis (i.e., the lag operator, the analysis of parameters influencing the urban wastewater, e.g. rainfall, precipitation, air temperature for snowmelt, etc.). This enhances the combined use of specialist knowledge and the power of big data, so that the modelling becomes more efficient.

- By comparing the outputs of the developed IGs with the observed data and by evaluating the performance based on a suite of criteria, it was shown that the proposed IG is capable of generating influent time series that not only have good performance on the average metric but also possess the same statistical properties as the observation, with variability of particular interest given its importance in evaluation of process control performance or the compliance assessment by exceedance of effluent criteria.

- Given the importance of influent variability on the performance of WWTPs, one of the most innovative outcomes of this research work is the multi-objective optimization for the dynamic generation of WRRF influent wastewater quality, which allows to simultaneously optimize the precision and statistical properties of the generated time series using the Kullback-Leibler (KL) divergency model performance metric.

- To better apply the IG models and to advance the research objectives, different modifications and optimizations of the data-driven models have been studied. The research was conducted for different case studies with different objectives, including (i) a pilot scale WRRF, pilEAUte, in Quebec City, QC, Canada, (ii) the East WRRF, also in Quebec City, QC, Canada and (iii) the WRRF of Bordeaux, France. From this diversity of applications, it can be expected that the developed procedures can be generally applied and easily transferred to other case studies. Moreover, the flexibility of the proposed tool allows users to easily incorporate city growth and climate change into the generated influent time series.

- The contribution of snowmelt to wastewater flow and composition for cold climate WRRFs has been studied for the first time at this level of detail. The proposed data-driven LSTM-IG has been compared to the BSM phenomenological model with snowmelt module whose rain generator and temperature blocks were slightly modified in this PhD study. The result showed that the IG is able to handle the changes in influent characteristics under the influence of snowmelt and stormwater for cold climate WRRFs. Only snow depth data and air temperature data are needed to feed this model under winter conditions.

- The flexibility of the proposed IG also allows users to easily adapt the IG model for other applications, with particular objectives: (i) for influents with both long-term time series needs or high frequency data

needs; (ii) for improving influent database management and for data mining: the IG can be used to complete missing data (gap filling) or to interpolate a low frequency time series into a denser time series; (iii) by incorporating continuous online data, the data-driven IG models was also shown able to perform influent prediction over different time horizons (one time step or multi time step forecasting) and at different frequencies (hourly or 15min). Such predictions are essential for feedforward process controllers that rely on information on disturbances to anticipate their negative impacts and perform anticipatory action.

In conclusion, the aim of this PhD research was to advance the field of WRRF influent data generation, analysis and data mining to the benefit of urban wastewater quantity and quality modelling. This was accomplished on the one hand by achieving a better understanding of the influent dynamics and by designing and correctly developing data-driven models.

On the other hand, from a practical standpoint, the proposed IGs overcome barriers of IG model application in real industrial context by considering data availability, influent variability, catchment properties and model complexity and parameter uncertainty, thus making the developed IG model more accessible and more flexible.

Perspectives

The developed IG model allows estimating and forecasting the flowrate and pollutant concentrations in view of model-based WRRF design, upgrade, control strategy development and, ultimately, digital twin modelling. The conducted studies provide a considerable amount of possibilities to be extended, for the IG model to be applied in future studies.

First, as a crucial input to any WRRF digital twin, it is strongly suggested to include the proposed data-driven IG model as input feed for WRRF models. Thanks to the adopted non-deterministic modelling approach, i.e. the IG models with incorporated random walk process, it becomes possible to test uncertainty of WRRF model results in terms of uncertainty on the WRRF's configuration and sizing, its operation performance and costing, as well as in process control applications. The IG model can also play a useful role in the frame of Integrated Urban Wastewater System (IUWS) modelling, where a full evaluation of the design and operational scenarios of WWTPs can be implemented within a context of computing-intensive integrated modelling. In the IUWS framework, instead of developing the sewer system model with relatively complex mechanistic models, the data-driven IG could replace this complex model, which always requires considerable modelling efforts to build. This would allow not only evaluating the impact of rain or snowmelt events on the WRRF, but also the WRRF's effluent quality impact on the receiving water.

A second extension should focus on including experimental results with the developed IG model. To broaden the IG's compatibility with the ASM family, it would be beneficial to include fractionation (not only carbon but also nitrogen and phosphate) dynamics as well as salts such potassium that play a role in resource recovery models. This would allow designing facilities that close the nutrient cycle by, for example, developing fertilizers from wastewater, promoting wastewater as a recoverable resource.

To complete this research project and advance the understanding of the effect of sewer characteristics and industrial discharges, it would definitely be invaluable to carry out fractionation experiments on full scale plants. With respect to the Quebec City case study, further measurements could be made to improve the fractionation and high frequency information on the water quality at the inlet of the WRRF in Quebec City. These experiments would make it possible to verify and validate the proposed model and to expand the IG's utilisations.

Even though the IGs have achieved a good result for influent generation, the challenge of interpretability of the IG and data requirements remain, the natural drawback of any data-driven model. Considering these drawbacks of data-driven models, creating a hybrid model taking advantage of both prior knowledge (extrapolation power) and data-driven modelling (improved accuracy and easier calibration) is promising. Therefore, it is recommended to create a hybrid model by integrating mechanistic (phenomenological) and data-driven models. The proposed LSTM-based IG approach should also be validated in other case studies in order to evaluate its transferability and extensibility. Many hybrid model examples have demonstrated the potential of combining knowledge-based and data based-models based on different architectures (serial configuration or parallel configurations) (Anderson et al., 2000; Lee et al., 2002; Thompson and Kramer, 1994; Villez et al., 2019). It would thus be worthwhile to create a hybrid IG model in order to gain the benefits and overcome drawbacks of each part, i.e. reducing the data requirements, increasing the interpretability, reducing the risk of overfitting, introducing expert knowledge and so on.

The active learning of WRRF will be playing a more and more important role. Engineers may opt to use different influent profiles to train and improve the learning performance under different operation strategies and different wastewater treatment processes. With this, there could be a significant advancement in the use of IA techniques in the wastewater treatment filed.

Generally, it can be concluded that, although considerable progress has been made in the field of influent data mining, further developments are required that should not only be limited to improving the IG's performance and the data-driven methodology itself, but should also evaluate new potential applications by coupling the IG within a broader framework (digital twin, hybrid models, integrated models). Further studies should encourage the attention for these two crucial aspects.

References

Achleitner, S., Möderl, M., Rauch, W., 2007. CITY DRAIN © – An open source approach for simulation of integrated urban drainage systems. Environ. Model. Softw. 22, 1184–1195.

Aguado, D., Frank, B., Juan Antoinio, B., Kris, V., Maria victoria, R., Oscar, S., Queralt, P., 2021. Digital Water: The value of meta-data for water resource recovery facilities, IWA publishing, London, UK.

Ahnert, M., Marx, C., Krebs, P., Kuehn, V., 2016. A black-box model for generation of site-specific WWTP influent quality data based on plant routine data. Water Sci. Technol. 74, 2978–2986.

Alferes, J., Tik, S., Copp, J., Vanrolleghem, P.A., 2013. Advanced monitoring of water systems using in situ measurement stations: data validation and fault detection. Water Sci. Technol. 68, 1022–1030.

Alferes, J., Vanrolleghem, P.A., 2016. Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. AI Commun. 29, 701–709.

Alfiya, Y., Dubowski, Y., Friedler, E., 2018. Diurnal patterns of micropollutants concentrations in domestic greywater. Urban Water J. 15, 399–406.

Alisawi, H.A.O., 2020. Performance of wastewater treatment during variable temperature. Appl. Water Sci. 10, 1–6.

Almeida, M.C., Butler, D., Friedler, E., 1999. At-source domestic wastewater quality. Urban Water 1, 49–55.

Aminabad, M., Maleki, A., Hadi, M., Shahmoradi, B., 2013. Application of Artificial Neural Network (ANN) for the prediction of water treatment plant influent characteristics. J. Adv. Environ. Heal. Res. 1, 89–100.

Anderson, J.S., McAvoy, T.J., Hao, O.J., 2000. Use of hybrid models in wastewater systems. Ind. Eng. Chem. Res. 39, 1694–1704.

Bach, P.M., Rauch, W., Mikkelsen, P.S., McCarthy, D.T., Deletic, A., 2014. A critical review of integrated urban water modelling - Urban drainage and beyond. Environ. Model. Softw. TA - TT - 54, 88–107.

Bailey, O., Hofman, J.A.M.H., Arnot, T.C., Kapelan, Z., Blokker, M., Vreeburg, J., 2018. Developing a stochastic sewer input model to support sewer design under water conservation measures, in: 11th International Conference on Urban Drainage Modelling (UDM). Palermo Italy, September 23-26 2018. pp. 74–78.

Bailey, O., Zlatanovic, L., van der Hoek, J.P., Kapelan, Z., Blokker, M., Arnot, T., Hofman, J., 2020. A stochastic model to predict flow, nutrient and temperature changes in a sewer under water conservation scenarios. Water 12, 1187.

Banihabib, M., Bandari, R., Peralta, R., 2019. Auto-regressive neural-network models for long lead-time forecasting of daily flow. Water Resour. Manag. 33, 159–172.

Bartosz, S., Kiczko, A., Studzinski, J., Daabek, L., 2018. Hydrodynamic and probabilistic modelling of storm overflow discharges. J. Hydroinformatics 20, 1100–1110.

Bechmann, H., Nielsen, M.K., Madsen, H., Kjølstad Poulsen, N., 1999. Grey-box modelling of pollutant loads from a sewer system. Urban Water 1, 71–78.

Belia, E., Benedetti, L., Johnson, B., Murthy, S., Neumann, M.B., Vanrolleghem, P.A., Weijers, S., 2021. Uncertainty in Wastewater Treatment Design and Operation: Addressing Current Practices and Future Directions, Scientific and Technical Report No. 21, IWA Publishing, London

Benedetti, L., Langeveld, J., Comeau, A., Corominas, L., Daigger, G., Martin, C., Mikkelsen, P.S., Vezzaro, L., Weijers, S., Vanrolleghem, P.A., 2013. Modelling and monitoring of integrated urban wastewater systems: Review on status and perspectives. Water Sci. Technol. 68, 1203–1215.

Béraud, B., Steyer, J.-P., Lemoine, C., Gernaey, K.V., Latrille, E., 2007. Model-based generation of continuous influent data from daily mean measurements available at industrial scale, in: 3rd International IWA Conference on Automation in Water Quality Monitoring. Gent, Belgium, September 5-7, 2007.

Bertrand-Krajewski, J.L., Winkler, S., Saracevic, E., Torres, A., Schaar, H., 2007. Comparison of and uncertainties in raw sewage COD measurements by laboratory techniques and field UV-visible spectrometry. Water Sci. Technol. 56, 17–25.

Borsányi, P., Benedetti, L., Dirckx, G., De Keyser, W., Muschalla, D., Solvi, A.-M., Vandenberghe, V., Weyand, M., Vanrolleghem, P.A., 2008. Modelling real-time control options on virtual sewer systems. J. Environ. Eng. Sci. 7, 395–410.

Borzooei, S., Teegavarapu, R., Abolfathi, S., Amerlinck, Y., Nopens, I., Zanetti, M.C., 2019. Data mining application in assessment of weather-based influent scenarios for a WWTP: Getting the most out of plant historical data. Water. Air. Soil Pollut. 230, 5.

Bourgeois, W., Burgess, J.E., Stuetz, R.M., 2001. On-line monitoring of wastewater quality: A review. J. Chem. Technol. Biotechnol. 76, 337–348.

Boussaada, Z., Curea, O., Remaci, A., Camblong, H., Bellaaj, N.M., 2018. A nonlinear autoregressive exogenous (NARX) neural network model for the prediction of the daily direct solar radiation. Energies 11.

Brdjanovic, G.C.G.A.E.M.C.M. van L.D. (Ed.), 2020. Wastewater characteristics, in: Biological Wastewater Treatment: Principles, Modeling and Design. IWA Publishing, London, UK. pp. 77–110.

Breinholt, A., Thordarson, F.Ö., Møller, J.K., Grum, M., Mikkelsen, P.S., Madsen, H., 2011. Grey-box modelling of flow in sewer systems with state-dependent diffusion. Environmetrics 22, 946–961.

Brombach, H., Weiss, G., Fuchs, S., 2005. A new database on urban runoff pollution: comparison of separate and combined sewer systems. Water Sci. Technol. 51, 119–128.

Brouckaert, C., Mhlanga, F., 2013. Characterisation of wastewater for modelling of wastewater treatment plants receiving industrial effluent. Water S.A 39, 403–407.

Butler, D., Friedler, E., Gatt, K., 1995. Characterising the quantity and quality of domestic wastewater inflows. Water Sci. Technol. 31, 13–24.

Bynagari, N.B., 2020. The difficulty of learning long-term dependencies with gradient flow in recurrent nets. Eng. Int. 8, 127–138.

Carstensen, J., Nielsen, M., Strandbaek, H., 1998. Prediction of hydraulic load for urban storm control of a municipal WWT plant. WATER Sci. Technol. 37, 363–370.

Chiandussi, G., Codegone, M., Ferrero, S., Varesio, F.E., 2012. Comparison of multi-objective optimization methodologies for engineering applications. Comput. Math. with Appl. 63, 912–942.

Choi, Y.Y., Baek, S.R., Kim, J.I., Choi, J.W., Hur, J., Lee, T.U., Park, C.J., Lee, B.J., 2017. Characteristics and biodegradability of wastewater organic matter in municipal wastewater treatment plants collecting domestic wastewater and industrial discharge. Water 2017, 9, 409.

Chollet, F., 2015. Keras, GitHub. https://github.com/keras-team/keras.

Cierkens, K., Plano, S., Benedetti, L., Weijers, S., De Jonge, J., Nopens, I., 2012. Impact of influent data frequency and model structure on the quality of WWTP model calibration and uncertainty. Water Sci. Technol. 65, 233–242.

Corominas, L., Flores-Alsina, X., Muschalla, D., Neumann, M.B., Vanrolleghem, P.A., 2011. Verification of WWTP design guidelines with activated sludge process models, in: WEFTEC2010, 137–146.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., Poch, M., 2018. Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environ. Model. Softw. 106, 89–103.

Corominas, L., Rieger, L., Takács, I., Ekama, G., Hauduc, H., Vanrolleghem, P.A., Oehmen, A., Gernaey, K. V., Van Loosdrecht, M.C.M., Comeau, Y., 2010. New framework for standardized notation in wastewater treatment modelling. Water Sci. Technol. 61, 841–857.

Coutu, S., Del Giudice, D., Rossi, L., Barry, D.A., 2012. Parsimonious hydrological modeling of urban sewer and river catchments. J. Hydrol. 464–465, 477–484.

Coutu, S., Pouchon, T., Queloz, P., Vernaz, N., 2016. Integrated stochastic modeling of pharmaceuticals in sewage networks. Stoch. Environ. Res. Risk Assess. 30, 1087–1097.

De Keyser, W., Gevaert, V., Verdonck, F., De Baets, B., Benedetti, L., 2010. An emission time series generator for pollutant release modelling in urban areas. Environ. Model. Softw. 25, 554–561.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. TA - TT - 6, 182–197.

Debele, B., Srinivasan, R., Gosain, A.K., 2010. Comparison of process-based and temperature-index snowmelt modeling in SWAT. Water Resour. Manag. 24, 1065–1088.

Delli Compagni, R., Polesel, F., von Borries, K.J.F., Zhang, Z., Turolla, A., Antonelli, M., Vezzaro, L., 2019. Modelling micropollutant fate in sewer systems - A new systematic approach to support conceptual model construction based on in-sewer hydraulic retention time. J. Environ. Manag. TA - TT - 246, 141–149.

Demchenko, Y., Grosso, P., Laat, C. de, Membrey, P., 2013. Addressing big data issues in Scientific Data Infrastructure, in: 2013 International Conference on Collaboration Technologies and Systems (CTS). July 2013, San Diego, CA, USA, pp. 48–55.

Dent, S., Wright, L., Mosley, C., Housen, V., 2000. Continuous simulation vs. design storms comparison with wet weather flow prediction methods, in: Water Environment Federation. Rochester, NY, May 2000., pp. 373–392.

Devisscher, M., Ciacci, G., Fé, L., Benedetti, L., Bixio, D., Thoeye, C., De Gueldre, G., Marsili-Libelli, S., Vanrolleghem, P.A., 2006. Estimating costs and benefits of advanced control for wastewater treatment plants--the MAgIC methodology. Water Sci. Technol. 53, 215.

Di Trapani, D., Christensson, M., Torregrossa, M., Viviani, G., Ødegaard, H., 2013. Performance of a hybrid activated sludge/biofilm process for wastewater treatment in a cold climate region: Influence of operating conditions. Biochem. Eng. J. 77, 214–219.

Dixon, A., Butler, D., Fewkes, A., Robinson, M., 2000. Measurement and modelling of quality changes in stored untreated grey water. Urban Water 1, 293–306.

Dobbelaere, M.R., Plehiers, P.P., Van de Vijver, R., Stevens, C. V., Van Geem, K.M., 2021. Machine learning in chemical engineering: Strengths, weaknesses, opportunities, and threats. Engineering 7, 1201–1211.

Dürrenmatt, D., 2011. Data mining and data-driven modeling approaches to support wastewater treatment plant operation. PhD thesis, ETH Zurich, Switzerland.

Dürrenmatt, D.J.Ô., Gujer, W., 2012. Data-driven modeling approaches to support wastewater treatment plant operation. Environ. Model. Softw. 30, 47–56.

Eerikäinen, S., Haimi, H., Mikola, A., Vahala, R., 2020. Data analytics in control and operation of municipal wastewater treatment plants: Qualitative analysis of needs and barriers. Water Sci. Technol. 82, 2681–2690.

El-Din, A.G., Smith, D.W., 2002. A neural network model to predict the wastewater inflow incorporating rainfall events. Water Res. 36, 1115–1126.

El-Din, A.G., Smith, D.W., El-Din, M.G., 2004. Application of artificial neural networks in wastewater treatment. J. Environ. Eng. Sci. 3, S81–S95.

Elliott, A.H., Trowsdale, S.A., 2007. A review of models for low impact urban stormwater drainage. Environ. Model. Softw. 22, 394–405.

Ercan, M.B., Goodall, J.L., 2016. Design and implementation of a general software library for using NSGA-II with SWAT for multi-objective model calibration. Environ. Model. Softw. 84, 112–120.

Eriksson, E., Andersen, H.R., Madsen, T.S., Ledin, A., 2009. Greywater pollution variability and loadings. Ecol. Eng. 35, 661–669.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Mag. 17, 37–53.

Flores-Alsina, X., Corominas, L., Neumann, M.B., Vanrolleghem, P.A., 2012a. Assessing the use of activated sludge process design guidelines in wastewater treatment plant projects: A methodology based on global sensitivity analysis. Environ. Model. Softw. 38, 50–58.

Flores-Alsina, X., Gernaey, K. V., Jeppsson, U., 2012b. Global sensitivity analysis of the BSM2 dynamic influent disturbance scenario generator. Water Sci. Technol. 65, 1912–1922.

Flores-Alsina, X., Saagi, R., Lindblom, E., Thirsing, C., Thornberg, D., Gernaey, K. V., Jeppsson, U., 2014a. Calibration and validation of a phenomenological influent pollutant disturbance scenario generator using full-scale data. Water Res. 51, 172–185.

Flores-Alsina, X., Saagi, R., Lindblom, E., Thirsing, C., Thornberg, D., Gernaey, K. V, Jeppsson, U., 2014b. Calibration and validation of a phenomenological influent pollutant disturbance scenario generator using full-scale data. Water Res. 51, 172–185.

Fonseca, C., Fleming, P., 1999. Genetic algorithms for multiobjective optimization: Formulation discussion and generalization, in: 5th International Conference on Genetic Algorithms. San Francisco, United States, 1999, pp. 416–423.

Friedler, E., Butler, D., 1996. Quantifying the inherent uncertainty in the quantity and quality of domestic wastewater. Water Sci. Technol. 33, 65–75.

Fu, G., Makropoulos, C., Butler, D., 2010. Simulation of urban wastewater systems using artificial neural

networks: embedding urban areas in integrated catchment modelling. J. Hydroinformatics 12, 140–149.

García, J.T., Espín-Leal, P., Vigueras-Rodriguez, A., Castillo, L.G., Carrillo, J.M., Martínez-Solano, P.D., Nevado-Santos, S., 2017. Urban runoff characteristics in combined sewer overflows (CSOs): Analysis of storm events in Southeastern Spain. Water 9.

Gernaey, K. V., Flores-Alsina, X., Rosen, C., Benedetti, L., Jeppsson, U., 2011. Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. Environ. Model. Softw. 26, 1255–1267.

Gernaey, K. V, Rosén, C., Jeppsson, U., 2005. BSM2: A Model for Dynamic Influent Data Generation. Technical Report, Lund University of Technology, Lund, Sweden.

Ginestet, P., Maisonnier, A., Spérandio, M., 2002. Wastewater COD characterization: Biodegradability of physico-chemical fractions. Water Sci. Technol. 45, 89.

Goldberg, D.E., 1989. Genetic algorithms in search, optimization and machine learning, 1st ed. Addison-Wesley Longman Publishing Co., Inc., MA, United States.

Gómez-Llanos, E., Matías-Sánchez, A., Durán-Barroso, P., 2020. Wastewater treatment plant assessment by quantifying the carbon and water footprint. Water (Switzerland) 12, 1–16.

Gong, N., Ding, X., Denoeux, T., Bertrand-Krajewski, J.L., Clément, M., 1996. STORMNET: A connectionist model for dynamic management of wastewater treatment plants during storm events. Water Sci. Technol. 33, 247–256.

Gopakumar, V., Tiwari, S., Rahman, I., 2018. A deep learning based data driven soft sensor for bioprocesses. Biochem. Eng. J. 136, 28–39.

Graves, Alex., 2012. Supervised Sequence Labelling with Recurrent Neural Networks, Volume 385. ed, Studies in Computational Intelligence. Springer, Berlin, Heidelberg, Germany.

Graves, Alex, 2012. Long Short-Term Memory, in: Graves, A. (Ed.), Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin, Heidelberg, pp. 37–45.

Grievson, O., 2020. Digital water: the role of instrumentation in digital transformation. IWA white paper, International Water Association, London, UK.

Gujer, W., 2008. Systems Analysis for Water Technology. Springer, Berlin, Heidelberg, Germany.

Gullicks, H.A., Cleasby, J.L., 1990. Cold-climate nitrifying biofilters: Design and operation considerations. Res. J. Water Pollut. Control Fed. 62, 50–57.

Guo, L., Tik, S., Ledergerber, J.M., Santoro, D., Elbeshbishy, E., Vanrolleghem, P.A., 2019. Conceptualizing the sewage collection system for integrated sewer-WWTP modelling and optimization. J. Hydrol. 573, 710–716.

Hadjimichael, A., Comas, J., Corominas, L., 2016. Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector. AI Commun. 29, 747–756.

Han, H.G., Zhang, H.J., Liu, Z., Qiao, J.F., 2020. Data-driven decision-making for wastewater treatment process. Control Eng. Pract. 96.

Hassanat, A., Almohammadi, K., Alkafaween, E., Abunawas, E., Hammouri, A., Prasath, V.B.S., 2019. Choosing mutation and crossover ratios for genetic algorithms—A review with a new dynamic approach. Information

10, 390–402.

Hauduc, H., Neumann, M.B., Muschalla, D., Gamerith, V., Gillot, S., Vanrolleghem, P.A., 2015. Efficiency criteria for environmental model quality assessment: A review and its application to wastewater treatment. Environ. Model. Softw. 68, 196–204.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA pp. 770–778.

Heideman, M., Johnson, D., Burrus, C.S., 1984. Gauss and the history of the Fast Fourier Transform. IEEE Signal Process. Mag. 1, 14–21.

Henze, M., Gujer, W., Mino, T., van Loosdrecht, M., 2000. Activated Sludge Models ASM1, ASM2, ASM2D, ASM3, IWA Publishing, London, UK.

Henze, M., van Loosdrecht, M., Ekama, G., Brdjanovic, D., 2008. Biological Wastewater Treatment: Principles, Modeling and Design. IWA Publishing, London, UK.

Hocaoglu, S.M., Insel, G., Cokgor, E.U., Baban, A., Orhon, D., 2010. COD fractionation and biodegradation kinetics of segregated domestic wastewater: Black and grey water fractions. J. Chem. Technol. Biotechnol. 85, 1241–1249.

Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertainty, Fuzziness Knowlege-Based Syst. 6, 107–116.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780.

Holland, J.H., 1992. Adaptation in Natural and Artificial Systems : An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, MIT Press. Cambridge, Mass.

Hug, T., Maurer, M., 2012. Stochastic modeling to identify requirements for centralized monitoring of distributed wastewater treatment. Water Sci. Technol. 65, 1067–1075.

Hutagalung, A., 1967. Bayesian estimation for decision-making of return periods of CSO volumes in sewer system management. Angew. Chemie Int. Ed. 6(11), 951–952. 5–24.

Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven Soft Sensors in the process industry. Comput. Chem. Eng. 33, 795–814.

Khalil, B., Adamowski, J., Abdin, A., Elsaadi, A., 2019. A statistical approach for the estimation of water quality characteristics of ungauged streams/watersheds under stationary conditions. J. Hydrol. 569, 106–116.

Khalil, B., Ouarda, T.B.M.J., St-Hilaire, A., 2011. Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. J. Hydrol. 405, 277–287.

Kingma, D.P., Ba, J.L., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.2015, San Diego, pp. 1–15.

Konak, A., Coit, D.W., Smith, A.E., 2006. Multi-objective optimization using genetic algorithms: A tutorial. Reliab. Eng. Syst. Saf. 91, 992–1007.

Kriger, C., Tzoneva, R., 2007. A neural network model for control of wastewater treatment processes. In: IFAC Proceedings Volumes. pp. 981–986., 7th IFAC Symposium on Nonlinear Control Systems, 2007, Pretoria, South Africa.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. TA - TT - 22, 79–86.

Langergraber, G., Alex, J., Weissenbacher, N., Woerner, D., Ahnert, M., Frehmann, T., Halft, N., Hobus, I., Plattes, M., Spering, V., Winkler, S., 2008. Generation of diurnal variation for influent data for dynamic simulation. Water Sci. Technol. 57, 1483–1486.

Langeveld, Jeroen, Van Daal, P., Schilperoort, R., Nopens, I., Flameling, T., Weijers, S., 2017. Empirical sewer water quality model for generating influent data for WWTP modelling. Water 9, 1–18.

Le, N.D., France, X., Pontvianne, S., Poirot, H., Leclerc, J.P., Pons, M.N., 2017. Daily wastewater pollutant dynamics with respect to catchment population structure. Urban Water J. 14, 1016–1022.

Ledergerber, J. M., Maruéjouls, T., Vanrolleghem, P.A., 2020. No-regret selection of effective control handles for integrated urban wastewater systems management under parameter and input uncertainty. Water Sci. Technol. 81, 1749–1756.

Ledergerber, J.M., Pieper, L., Binet, G., Comeau, A., Maruéjouls, T., Muschalla, D., Vanrolleghem, P.A., 2019. An efficient and structured procedure to develop conceptual catchment and sewer models from their detailed counterparts. Water 11, 1–19.

Lee, D.S., Jeon, C.O., Park, J.M., Chang, K.S., 2002. Hybrid neural network modeling of a full-scale industrial wastewater treatment process. Biotechnol. Bioeng. 78, 670–682.

Lee, D.S., Vanrolleghem, P.A., Jong, M.P., 2005. Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. J. Biotechnol. 115, 317–328.

Lepot, M., Aubin, J.B., Bertrand-Krajewski, J.L., 2013. Accuracy of different sensors for the estimation of pollutant concentrations (total suspended solids, total and dissolved chemical oxygen demand) in wastewater and stormwater. Water Sci. Technol. 68, 462–471.

Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares SQUARES. Q. Appl. Math. 2, 164–168.

Li, F., Amaral, A., Vanrolleghem, P.A., 2020. An essential tool for WRRF modelling: A realistic and complete influent generator for flow rate and water quality based on machine learning, in: 93rd Water Environment Federation Technical Exhibition and Conference 2020, WEFTEC 2020. Water Environment Federation, pp. 303–309.

Li, F., Vanrolleghem, P.A., 2021. An influent generator for WRRF design and operation based on a recurrent neural network using a genetic algorithm for multi-objective optimization, in: 7th IWA Water Resource Recovery Modelling Seminar (WRRmode2021). 2021, Switzerland.

Lindblom, E., Gernaey, K. V, Henze, M., Mikkelsen, P.S., 2006. Integrated modelling of two xenobiotic organic compounds. Water Sci. Technol. 54, 213–221.

Lotfi, K., Bonakdari, H., Ebtehaj, I., Mjalli, F.S., Zeynoddin, M., Delatolla, R., Gharabaghi, B., 2019. Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology. J. Environ. Manage. 240, 463–474.

Ma, S., Zeng, S., Dong, X., Chen, J., Olsson, G., 2014. Short-term prediction of influent flow rate and ammonia concentration in municipal wastewater treatment plants. Front. Environ. Sci. Eng. 8, 128–136.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model. Softw. 15, 101–124.

Malik, O.A., Hsu, A., Johnson, L.A., de Sherbinin, A., 2015. A global indicator of wastewater treatment to inform the Sustainable Development Goals (SDGs). Environ. Sci. Policy 48, 172–185.

Man, Y., Hu, Y., Ren, J., 2019. Forecasting COD load in municipal sewage based on ARMA and VAR algorithms. Resour. Conserv. Recycl. 144, 56–64.

Mannina, G, Cosenza, A., Vanrolleghem, P.A., Viviani, G., 2011. A practical protocol for calibration of nutrient removal wastewater treatment models. J. Hydroinformatics 13, 575–595.

Manny, L., Duygan, M., Fischer, M., Rieckermann, J., 2021. Barriers to the digital transformation of infrastructure sectors, Policy Sciences. Springer, Berlin, Heidelberg, Germany.

Martin, C., Vanrolleghem, P.A., 2014. Analysing, completing, and generating influent data for WWTP modelling: A critical review. Environ. Model. Softw. 60, 188–201.

Mehmood, H., Mukkavilli, S.K., Weber, I., Koshio, A., Meechaiya, C., Piman, T., Mubea, K., Tortajada, C., Mahadeo, K., Liao, D., 2020. Strategic Foresight to Applications of Artificial Intelligence to Achieve Water-related Sustainable Development Goals. UNU INWEH Rep. Ser..

Meirlaen, J., Huyghebaert, B., Sforzi, F., Benedetti, L., Vanrolleghem, P., 2001. Fast, simultaneous simulation of the integrated urban wastewater system using mechanistic surrogate models. Water Sci. Technol. 43, 301–309.

Metcalf & Eddy, I., Tchobanoglous, G., Burton, F., Tsuchihashi, R., Stensel, H.D., 2014. Wastewater Engineering: Treatment and Resource Recovery, Fifth edit. ed. McGraw-Hill Education, New York, NY.

Mhlanga, F.T., Brouckaert, C.J., 2013. Characterisation of wastewater for modelling of wastewater treatment plants receiving industrial effluent. Water SA, 39, 403-407.

Michelbach, S., 1995. Origin, resuspension and settling characteristics of solids transported in combined sewage. Water Sci. Technol. 31, 69–76.

Miller, B.L., Goldberg, D.E., 1996. Genetic algorithms, selection schemes, and the varying effects of noise. Evol. Comput. 4, 113–131.

Mitchell, V.G., Duncan, H., Inman, M., Rahilly, M., Stewart, J., Vieritz, A., Holt, P., Grant, A., Fletcher, T., Coleman, J.R., Maheepala, S., Sharma, A., Deletic, A., Breen, P., 2007. Integrated urban water modelling - Past, present and future. Rainwater Urban Des. 806.

Moghadas, S., Gustafsson, A.M., Muthanna, T.M., Marsalek, J., Viklander, M., 2016. Review of models and procedures for modelling urban snowmelt. Urban Water J. 13, 396–411.

Montes, C., Kapelan, Z., Saldarriaga, J., 2021. Predicting non-deposition sediment transport in sewer pipes using random forest. Water Res. 189, 116639.

Montgomery, R.H., Sanders, T.G., 1986. Uncertainty in Water Quality Data, in: El-Shaarawi, A.H., Kwiatkowski, R.E.B.T.-D. in W.S. (Eds.), Statistical Aspects of Water Quality Monitoring. Elsevier, Amsterdam, The Netherlands, pp. 17–29.

Montserrat, A., Bosch, L., Kiser, M.A., Poch, M., Corominas, L., 2015. Using data from monitoring combined sewer overflows to assess, improve, and maintain combined sewer systems. Sci. Total Environ. 505, 1053–1061.

Mu, L., Zheng, F., Tao, R., Zhang, Q., Kapelan, Z., 2020. Hourly and daily urban water demand predictions using

a Long Short-Term Memory based model. J. Water Resour. Plan. Manag. TA - TT - 146, 1–11.

Muharemi, F., Logofătu, D., Leon, F., Logofătu, D., Leon, F., 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. J. Inf. Telecommun. 3, 294–307.

Mullapudi, A., Lewis, M.J., Gruden, C.L., Kerkez, B., 2020. Deep reinforcement learning for the real time control of stormwater systems. Adv. Water Resour. TA - TT - 140.

Murat Hocaoglu, S., Insel, G., Ubay Cokgor, E., Baban, A., Orhon, D., 2010. COD fractionation and biodegradation kinetics of segregated domestic wastewater: black and grey water fractions. J. Chem. Technol. Biotechnol. 85, 1241–1249.

Nasrin, T., Sharma, A., Muttil, N., 2017. Impact of short duration intense rainfall events on sanitary sewer network performance. Water 9, 225.

Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y., 2019. Data-driven performance analyses of wastewater treatment plants: A review. Water Res. 157, 498–513.

Newhart, K.B., Marks, C.A., Rauch-Williams, T., Cath, T.Y., Hering, A.S., 2020. Hybrid statistical-machine learning ammonia forecasting in continuous activated sludge treatment for improved process control. J. Water Process Eng. 37, 101389.

Nielsen, P.H., Raunkjær, K., Norsker, N.H., Jensen, N.A., Hvitved-Jacobsen, T., 1992. Transformation of wastewater in sewer systems – A review. Water Sci. Technol. 25, 17–31.

Novotny, V., Zheng, S., 1989. Rainfall-runoff transfer function by ARMA modeling. J. Hydraul. Eng. 115, 1386–1400.

Ort, C., 2006. Short-term dynamics of micropollutants in sewer systems. PhD thesis, ETH Zurich, Switzerland.

Ort, C., Schaffner, C., Giger, W., Gujer, W., 2005. Modeling stochastic load variations in sewer systems. Water Sci. Technol. 52, 113.

Patry, B., 2020. Suivi, compréhension et modélisation d'une technologie à biofilm pour l'augmentation de la capacité des étangs aérés. PhD thesis, Université Laval, Québec, QC, Canada.

Pasztor, I., Thury, P., Pulai, J., 2009. Chemical oxygen demand fractions of municipal wastewater. Int. J. Environ. Sci. Technol. 6, 51–56.

Pedersen, A.N., Borup, M., Brink-Kjær, A., Christiansen, L.E., Mikkelsen, P.S., 2021. Living and prototyping digital twins for urban water systems: Towards multi-purpose value creation using models and sensors. Water (Switzerland) 13, 592.

Petersen, B., Gernaey, K., Henze, M., Vanrolleghem, P.A., 2002. Evaluation of an ASM1 model calibration procedure on a municipal-industrial wastewater treatment plant. J. Hydroinformatics 4, 15–38.

Pisa, I., Santin, I., Morell, A., Vicario, J.L., Vilanova, R., 2019. LSTM based wastewater treatment plants operation strategies for effluent quality improvement. IEEE Access 7, 1.

Pishgar, R., Morin, D., Young, S.J., Schwartz, J., Chu, A., 2021. Characterization of domestic wastewater released from 'green' households and field study of the performance of onsite septic tanks retrofitted into aerobic bioreactors in cold climate. Sci. Total Environ. 755, 142446.

Plana, Q., Alferes, J., Fuks, K., Kraft, T., Maruéjouls, T., Torfs, E., Vanrolleghem, P.A., 2019. Towards a water quality database for raw and validated data with emphasis on structured metadata. Water Qual. Res. J.

Canada 54, 1–9.

Plósz, B.G., Liltved, H., Ratnaweera, H., 2009. Climate change impacts on activated sludge wastewater treatment: A case study from Norway. Water Sci. Technol. 60, 533–541.

Pohlen, T., Hermans, A., Mathias, M., Leibe, B., 2017. Full-resolution residual networks for semantic segmentation in street scenes, in: 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 3309–3318, 21-26 July 2016, Honolulu, Hawaii.

Ponzelli, M., 2019. Volatile Fatty Acids Recovery in a Reactive Primary Clarifier: A Pilot Case Study. MSc Thesis, Ryerson University, Toronto, Canada.

Price, R.K., Vojinovic, Z., 2011. Urban Hydroinformatics: Data, Models and Decision Support for Integrated Urban Water Management. IWA Publishing, London, UK.

Qiao, J., Zhang, W., 2018. Dynamic multi-objective optimization control for wastewater treatment process. Neural Comput. Appl. 29, 1261–1271.

Quiza, R., López-Armas, O., Davim, J.P., 2012. Hybrid modeling and optimization of manufacturing: combining artificial intelligence and finite element method. Springer, Berlin, Heidelberg, Germany.

Rajurkar, M.P., Kothyari, U.C., Chaube, U.C., 2004. Modeling of the daily rainfall-runoff relationship with artificial neural network. J. Hydrol. 285, 96–113.

Raman, H., Sunilkumar, N., 1995. Multivariate modelling of water resources time series using artificial neural networks. Hydrol. Sci. J. 40, 145–163.

Rammal, M., 2016. Comparison of different scenarios of suspended solids production in a combined sewer system using an adapted hydrodynamic model. PhD thesis, Université Paris-Est, France.

Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the environmental modelling process - A framework and guidance. Environ. Model. Softw. 22, 1543–1556.

Regmi, P., Miller, M., Jimenez, J., Stewart, H., Johnson, B., Amerlinck, Y., Volcke, E.I.P., Arnell, M., García, P.J., Maere, T., Torfs, E., Vanrolleghem, P.A., Miletić, I., Rieger, L., Schraa, O., Samstag, R., Santoro, D., Snowling, S., Takács, I., 2019. The future of WRRF modelling - Outlook and challenges. Water Sci. Technol. 79, 3–14

Reichert, P., Schuwirth, N., 2012. Linking statistical bias description to multiobjective model calibration. Water Resour. Res. 48, 1–20.

Rieger, L., Gillot, S., Langergraber G., Ohtsuki T., Shaw, A., Takacs, I., Winkler, S., 2012. Guidelines for Using Activated Sludge Models. IWA Publishing, London, UK.

Rieger, L., Thomann, M., Gujer, W., Siegrist, H., 2005. Quantifying the uncertainty of on-line sensors at WWTPs during field operation. Water Res. 39, 5162–5174.

Rieger, L., Vanrolleghem, P.A., 2008. monEAU: A platform for water quality monitoring networks. Water Sci. Technol. 57, 1079–1086.

Rodriguez, J.P., Mcintyre, N., Diaz-Granados, M., Achleitner, S., Hochedlinger, M., Maksimovic, C., 2013. Generating time-series of dry weather loads to sewers. Environ. Model. Softw. 43, 133–143.

Roeleveld, P.J., van Loosdrecht, M.C.M., 2002. Experience with guidelines for wastewater characterisation in The Netherlands. Water Sci. Technol. 45, 77.

Rojas, R., 1996. Chapter 17 Genetic Algorithms, in: Rojas, R. (Ed.), Neural Networks: A Systematic Introduction. Springer, Berlin, Heidelberg, Germany, pp. 429–450.

Rossi, L., Krejci, V., Rauch, W., Kreikenbaum, S., Fankhauser, R., Gujer, W., 2005. Stochastic modeling of total suspended solids (TSS) in urban areas during rain events. Water Res. 39, 4188–4196.

Rössle, W.H., Pretorius, W.A., 2001. A review of characterisation requirements for in-line prefermenters paper 1: Wastewater characterisation. Water SA 27, 405–412.

Rossman, L.A., 2015. Storm Water Management Model User's Manual Version 5.1. National Risk Management Research Laboratory-Office of Research and Development: Cincinnati, OH, USA.

Rousseau, D., Verdonck, F., Moerman, O., Carrette, R., Thoeye, C., Meirlaen, J., Vanrolleghem, P.A., 2001. Development of a risk assessment based technique for design/retrofitting of WWTPs. Water Sci. Technol. 43, 287–294.

Russo, S., Besmer, M.D., Blumensaat, F., Bouffard, D., Disch, A., Hammes, F., Hess, A., Lürig, M., Matthews, B., Minaudo, C., Morgenroth, E., Tran-Khac, V., Villez, K., 2021. The value of human data annotation for machine learning based anomaly detection in environmental systems. Water Res. 206, 117695.

Russo, S., Lürig, M., Hao, W., Matthews, B., Villez, K., 2020. Active learning for anomaly detection in environmental data. Environ. Model. Softw. 134.

Saagi, R., 2017. Benchmark Simulation Model for Integrated Urban Wastewater Systems: Model Development and Control Strategy Evaluation. PhD thesis, Lund University, Sweden.

Saagi, R., Lindblom, E., Grundestam, C., Andersson, S., Åmand, L., Jeppsson, U., 2018. Model-based evaluation of a full-scale wastewater treatment plant for future influent and operational scenarios, in: 11th IWA World Water Congress and Exhibition (WWC&E2018). September, 2018, Tokyo, Japan.

Schilperoort, R., 2011. Monitoring as a Tool for the Assessment of Wastewater Quality Dynamics. PhD thesis, Delft University of Technology, Delft, The Netherlands

Schlichthärle, D., 2011. Analog Filters - Digital Filters: Basics and Design. Springer, Berlin, Heidelberg, Germany.

Schlütter, F., Mark, O., 2003. Dynamic modelling of pollutants from CSOs. Water Sci. Technol. 47, 149–156.

Schneider, M.Y., Carbajal, J.P., Furrer, V., Sterkele, B., Maurer, M., Villez, K., 2019. Beyond signal quality: The value of unmaintained pH, dissolved oxygen, and oxidation-reduction potential sensors for remote performance monitoring of on-site sequencing batch reactors. Water Res. 161, 639–651.

Schneider, M.Y., Furrer, V., Sprenger, E., Carbajal, J.P., Villez, K., Maurer, M., 2020. Benchmarking soft sensors for remote monitoring of on-site wastewater treatment plants. Environ. Sci. Technol. 54, 10840–10849.

Seggelke, K., Löwe, R., Beeneken, T., Fuchs, L., 2013. Implementation of an integrated real-time control system of sewer system and waste water treatment plant in the city of Wilhelmshaven. Urban Water J. 10, 330–341.

Shokry, A., Vicente, P., Escudero, G., Pérez-Moya, M., Graells, M., Espuña, A., 2018. Data-driven soft-sensors for online monitoring of batch processes with different initial conditions. Comput. Chem. Eng. 118, 159–179.

Solon, K., Flores-Alsina, X., Gernaey, K. V., Jeppsson, U., 2015. Effects of influent fractionation, kinetics,

stoichiometry and mass transfer on CH4, H2and CO2 production for (plant-wide) modeling of anaerobic digesters. Water Sci. Technol. 71, 870–877.

Solon, K., Volcke, E.I.P., Spérandio, M., Van Loosdrecht, M.C.M., 2019. Resource recovery and wastewater treatment modelling. Environ. Sci.: Water Res. Technol. 5, 631–642.

Solvi, A.M., Benedetti, L., V., V., G., S., S., P., Weidenhaupt, A., Vanrolleghem, P.A., 2006. Construction and calibration of an integrated model for catchment, sewer, treatment plant and river, in: 7th International Conference on Hydroinformatics (HIC 2006). September 2006, Nice, France.

Souza, F.A.A., Araújo, R., Mendes, J., 2016. Review of soft sensor methods for regression applications. Chemom. Intell. Lab. Syst. 152, 69–79.

Srinivas, M., Patnaik, L.M., 1994. Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Trans. Syst. Man Cybern. 24, 656–667.

Stachowiak, A., 2007. Identifying the true hydraulic capacity of a wastewater treatment plant, in: 13th International Conference on Cold Regions Engineering, July 2006, Orono, Maine, USA p. 21.

Sweeney, M., Kabouris, J., 2013. Modeling, instrumentation, automation, and optimization of wastewater treatment facilities. Water Environ. Res. 85, 1322–1338.

Sweeney, M.W., Kabouris, J.C., 2015. Modeling, instrumentation, automation, and optimization of water resource recovery facilities. Water Environ. Res. 87, 1178–1195.

Talebizadeh, M., 2015. Probabilistic Design of Wastewater Treatment Plants. PhD thesis. Université Laval, Quebec, QC, Canada.

Talebizadeh, M., Belia, E., Vanrolleghem, P.A., 2016. Influent generator for probabilistic modeling of nutrient removal wastewater treatment plants. Environ. Model. Softw. 77, 32–49.

Talebizadeh, M., Belia, E., Vanrolleghem, P.A., 2015. Probabilistic design of wastewater treatment plants, in: 88th Annual WEF Technical Exhibition and Conference, WEFTEC 2015. Chicago, IL, USA, September 26-30 2015.

Therrien, J.-D., 2017. Utilisation d'un respiromètre pour le suivi et la modélisation d'une station de récupération des ressources en eau. MSc.thesis. Université Laval, Quebec, QC, Canada.

Therrien, J.-D., Nicolaï, N., Vanrolleghem, P.A., 2020. A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. Water Sci. Technol. 1–22.

Thompson, M.L., Kramer, M.A., 1994. Modeling chemical processes using prior knowledge and neural networks. AICHE J. 40, 1328–1340.

Thorndahl, S., 2009. Stochastic long term modelling of a drainage system with estimation of return period uncertainty. Water Sci. Technol. 59, 2331–2339.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288.

Tik, S., Vanrolleghem, P.A., 2017. Chemically enhancing primary clarifiers: Model-based development of a dosing controller and full-scale implementation. Water Sci. Technol. 75, 1185–1193.

Tohidi, M., 2019. Titrimetric Monitoring of Chemical Equilibrium and pH Dynamics in a Pilot-scale Water Resource Recovery Facility using PHREEQC and Buffer Capacity Modelling.. MSc thesis. Université Laval, Quebec, QC, Canada

Torfs, E., Nicolaï, N., Daneshgar, S., Copp, J.B., Haimi, H., Ikumi, D., Johnson, B., Plosz, B., Snowling, S., Townley, lloyd R., Valverde-Pérez, B., Vanrolleghem, P.A., Vezzaro, L., Nopens, I., 2022. The transition of WRRF models to digital twin applications. Water Sci. Technol.

Troutman, S.C., Schambach, N., Love, N.G., Kerkez, B., 2017. An automated toolchain for the data-driven and dynamical modeling of combined sewer systems. Water Res. 126, 88–100.

Vallabhaneni, S., Chan, C.C., Burgess, E.H., 2007. Computer Tools for Sanitary Sewer System Capacity Analysis and Planning. U.S. Environmental Protection Agency, Office of Research and Development: Washington, D.C, ,USA.

Van Houdt, G., Mosquera, C., Nápoles, G., 2020. A review on the long short-term memory model. Artif. Intell. Rev. 53, 5929–5955.

van Luijtelaar, H., Rebergen, E., 1997. Guidelines for hydrodynamic calculations on urban drainage in The Netherlands: Backgrounds and examples. Water Sci. Technol. 36, 253–258.

Vanhooren, H., Meirlaen, J., Amerlinck, Y., Claeys, F., Vangheluwe, H., Vanrolleghem, P.A., 2003a. WEST: modelling biological wastewater treatment. J. Hydroinformatics 5, 27–50.

Vanrolleghem, P.A., Kong, Z., Rombouts, G., Verstraete, W., 1994. An on-line respirographic biosensor for the characterization of load and toxicity of wastewaters. J. Chem. Technol. Biotechnol. 321–333.

Vanrolleghem, P.A., Lee, D.S., 2003. On-line monitoring equipment for wastewater treatment processes: State of the art. Water Sci. Technol. 47, 1–34.

Veit, A., Wilber, M., Belongie, S., 2016. Residual networks behave like ensembles of relatively shallow networks. Adv. Neural Inf. Process. Syst. 550–558.

Veldkamp, R.G., Wiggers, J.B.M., 1997. A statistical approach to pollutant emissions from combined sewer systems. Water Sci. Technol. 36, 95–100.

Vezzaro, L., Benedetti, L., Gevaert, V., De Keyser, W., Verdonck, F., De Baets, B., Nopens, I., Cloutier, F., Vanrolleghem, P.A., Mikkelsen, P.S., 2014. A model library for dynamic transport and fate of micropollutants in integrated urban wastewater and stormwater systems. Environ. Model. Softw. - 53, 98–111.

Vezzaro, L., Eriksson, E., Ledin, A., Mikkelsen, P., 2010. Dynamic stormwater treatment unit model for micropollutants (STUMP) based on substance inherent properties. Water Sci. Technol. 62, 622–629.

Vezzaro, L., Pedersen, J.W., Larsen, L.H., Thirsing, C., Duus, L.B., Mikkelsen, P.S., 2020. Evaluating the performance of a simple phenomenological model for online forecasting of ammonium concentrations at WWTP inlets. Water Sci. Technol. 81, 109–120.

Villez, K., Billeter, J., Bonvin, D., 2019. Incremental parameter estimation under rank-deficient measurement conditions. Processes 7, 75.

Villez, K., Del Giudice, D., Neumann, M.B., Rieckermann, J., 2020. Accounting for erroneous model structures in biokinetic process models. Reliab. Eng. Syst. Saf. 203, 107075.

Vrecko, D., Gernaey, K. V, Rosen, C., Jeppsson, U., 2006. Benchmark Simulation Model No 2 in Matlab-Simulink: Towards plant-wide WWTP control strategy evaluation. Water Sci. Technol. 54,6 65–72..

Wang, Q., Wang, L., Huang, W., Wang, Z., Liu, S., Savić, D.A., 2019. Parameterization of NSGA-II for the optimal

design of water distribution systems. Water 11, 971.

Wang, X., Kvaal, K., Ratnaweera, H., 2019. Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant. J. Process Control 77, 1–6.

Wang, X., Kvaal, K., Ratnaweera, H., 2017. Characterization of influent wastewater with periodic variation and snow melting effect in cold climate area. Comput. Chem. Eng. 106, 202–211.

Wei, X., Kusiak, A., 2015. Short-term prediction of influent flow in wastewater treatment plant. Stoch. Environ. Res. Risk Assess. 29, 241–249.

Wright, L.T., Dent, S., Mosley, C., Kadota, P., Djebbar, Y., 2001. Comparing rainfall dependent inflow and infiltration simulation methods. J. Water Manag. Model. 9, 235–257.

Wu, S., Zhong, S., Liu, Y., 2017. Deep residual learning for image steganalysis. Multimed. Tools Appl. 1–17.

Yang, C., Daigger, G.T., Belia, E., Kerkez, B., 2020. Extracting useful signals from flawed sensor data: Developing hybrid data-driven approaches with physical factors. Water Res. 185, 116282.

Yusoff, Y., Ngadiman, M.S., Zain, A.M., 2011. Overview of NSGA-II for optimizing machining process parameters. Procedia Eng. 15, 3978–3983.

Zawilski, M., Brzezinska, A., 2009. Variability of COD and TKN fractions of combined wastewater. Polish J. Environ. Stud. 18, 501–505.

Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J., 2018. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. J. Hydrol. 561, 918–929.

Zhang, Q., Li, Z., Snowling, S., Siam, A., El-Dakhakhni, W., 2019. Predictive models for wastewater flow forecasting based on time series analysis and artificial neural network. Water Sci. Technol. 80, 243–253.

Zhou, P., Li, Z., Snowling, S., Goel, R., Zhang, Q., 2019. Short-term wastewater influent prediction based on random forests and multi-layer perceptron. J. Environ. Informatics Lett. 1, 87–93.

Zhu, J., Anderson, P.R., 2019. Performance evaluation of the ISMLR package for predicting the next day's influent wastewater flowrate at Kirie WRP. Water Sci. Technol. 80, 695–706.

Zhu, J., Anderson, P.R., 2016. Assessment of a soft sensor approach for determining influent conditions at the MWRDGC Calumet WRP. J. Environ. Eng. 142, 04016023.

Appendix 1

The TSS concentration generated by the ANN model is compared to an average load model that calculates the TSS concentration as follows:

$$Concentration\left[\frac{mg}{l}\right] = \frac{average\ load\left[\frac{kg}{d}\right]}{flowrate\left[\frac{m^3}{d}\right]} * 10^3$$

The model uses three years of TSS load data (Figure A.1) and leads to a model which generates a TSS concentration time series (Figure A.2) with the following performance criteria: 25% for MAPE and 0.24 for NSE.

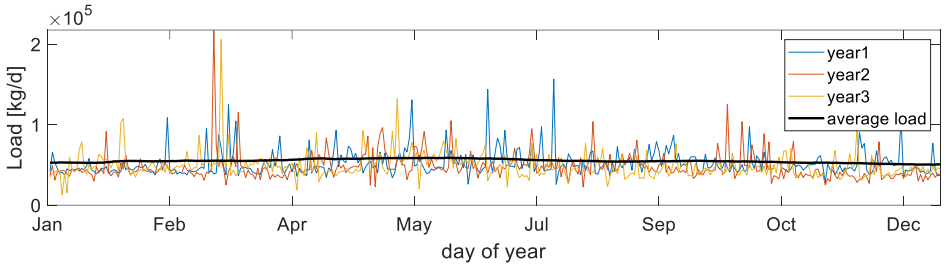


Figure A.1 TSS loads calculated for three years of daily TSS – Flow data at the Quebec City WRRF.

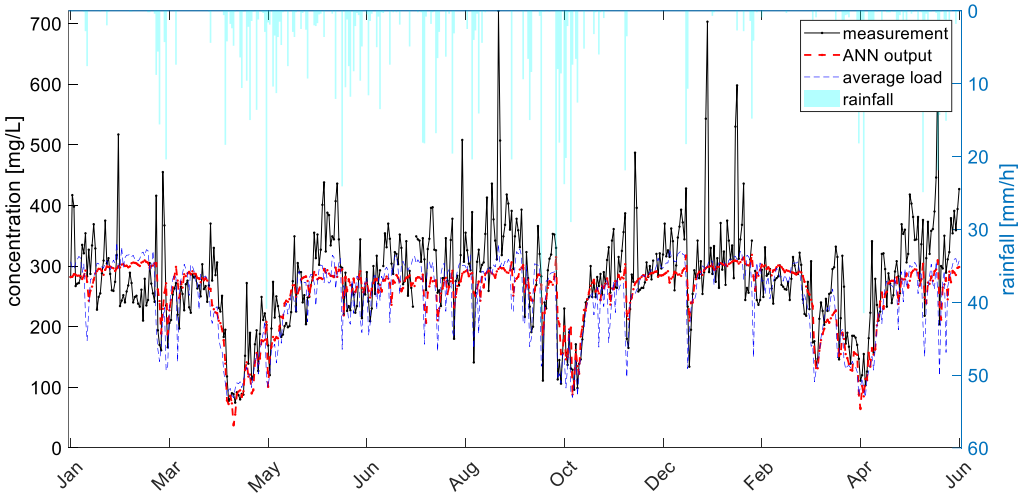


Figure A.2. Comparison of ANN generated and average load model generated TSS concentrations for 18 months of Quebec City case study data.