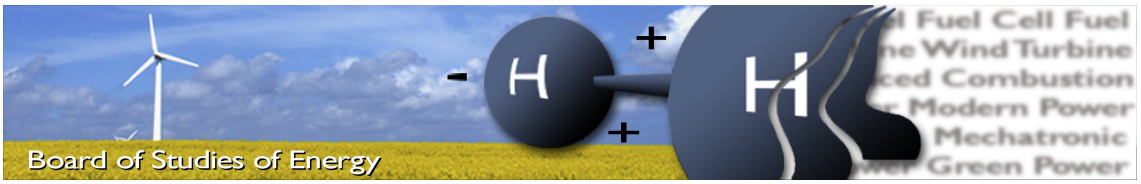
Data Refinement for Improving Dynamic Simulation of Wastewater Treatment Plants

Master's Thesis

Thermal Energy and Process Engineering

Simon Schraml

AALBORG UNIVERSITY
STUDENT REPORT



Title:	Data Refinement for Improving Dynamic Simulation of Wastewater Treatment Plants
Semester:	10th semester of Thermal Energy and Process Engineering
Semester theme:	Master's thesis
Project period:	01.02.2018 to 01.06.2018
ECTS:	30
Supervisor:	Chungen Yin, Department of Energy Technology, AAU
	Peter Aichinger, SYNECO tec GmbH
Project group:	TEPE4-1002

Simon Schraml

SYNOPSIS:

With the development of increasingly sophisticated models, dynamic simulation is continuing to gain popularity in design and improvement of wastewater treatment plants. However, this raises the demand for dynamic input data of sufficient quality, which is usually not available. Within this work a refinement model is developed based on the specific data situation given in the research project *ICAWER*. The model aims to resolve certain problems that are identified in available data within a detailed analysis. It facilitates the ad hoc refinement of plant data including induction of relevant dynamic phenomena in an adjustable manner for creating datasets at an hourly resolution. Demonstration of the individually developed algorithms in an exemplary case illustrates the changes induced in the data and produces realistically appearing time series. The application of the model is seen to enable a significant improvement in the realism of dynamic simulation results compared to the original data.

Pages, total:	69
Appendices:	1
Supplements:	Model files incl. add-in functions; dynamic simulation input data

By signing this document, each member of the group confirms that all group members have participated in the project work, and thereby all members are collectively liable for the contents of the project. Furthermore, all group members confirm that the project does not include plagiarism.

Preface

This master thesis is drawn up in collaboration with the research project *ICAWER*. The collaboration is made possible courtesy of Aalborg University and *Syneco tec GmbH*.

Reading guide

The layout of the report is designed for two-sided print.

Referencing is done based on the IEEE standard with modifications. In the text sources are indicated by numbers in square brackets, sorted by their order of appearance. Citations for single sentences are placed before the full stop. However, if a passage of multiple sentences refers to the same source, the citation is placed after the full stop and followed by a line break. Information on the respective source is found in the Bibliography at the end of the report.

For better readability and indication of the flow of thoughts the text is separated into paragraphs. Closely related adjacent paragraphs are merely separated by a line break. Paragraphs starting new topics which are not closely related to the previous one are preceded by a free line.

When cross-referenced in this work, sections, subsections and any parts of text on lower levels are generally either cross-referenced as *Section XY*, with *XY* denoting its respective numbering, or using clear naming of the respective part of text.

Units used for the display of variable values may vary and are used as most practical depending on the parameter and application. Units are always given along with numerical values.

Captions are placed directly below the respective figure and directly above the respective table.

In order to avoid excessive repetition, synonyms are used for frequent terms. Symbols and written expressions for different variables are used interchangeably. The nomenclature gives an overview of used abbreviations, variables and indices. Expressions which are merely given for information of the reader but not further used throughout the text are omitted in the nomenclature.

Software

Microsoft Excel 2016 along with *Microsoft Visual Basic for Applications 7.1* and *MathWorks MATLAB 2017b* are used for data analysis as well as development and implementation of the presented model. Dynamic simulations are conducted in *Dynamita SUMO 16*. Graphs are made in *MathWorks MATLAB 2017b*. Flow charts are constructed in *Microsoft Visio 2016*. *Adobe Illustrator CC* and *INKSCAPE 0.92* are used for the creation and refinement of various graphics.

Acknowledgements

I would like to genuinely thank both, my company supervisor Peter Aichinger as well as my university supervisor Chungen Yin for their patient guidance, sharing of knowledge and ubiquitous willingness to help.

A special thank you also to *Aalborg University* for what I experienced to be excellent education in an enjoyable environment and to Christoph Larch and *SYNECO tec GmbH* for making this highly interesting collaboration possible.

Last but not surely not least, my sincerest thank you and much love goes out to my amazing and supportive family, my caring girlfriend and all of my true friends. THANK YOU ALL for continuing to put up with me and all of my crap and for not only embellishing my life but really making it what it is.

Abstract

The aim of this work is to create a model for refinement of data documented at wastewater treatment plants in order to achieve increased realism in dynamic plant simulation. An analysis of real industrial scale data in combination with established high quality data from literature unveils the dynamics that appear on different timescales as well as details about the presence of errors and gaps within data obtained from the plants. Based on these findings, a model for ad hoc refinement of data for the specific situation given in the research project *ICAWER* is created and implemented. It enables systematic removal of definite errors, completion of fragmentary data and temporal interpolation down to hourly values including addition of relevant dynamics, all the while preserving statements made by the original data. Demonstration of the model in an exemplary case yields results that appear realistic. However, as dynamic features are adjusted by means of tunable parameters, the results and their quality rely on the experience and assessment of the model user. Dynamic simulation utilizing original plant data as well as intermediate and final model outputs as respective input data enables a comparison of the results and added value obtained courtesy of the individual model algorithms. A significant change in the simulation results is observed, which emphasizes the relevance of dynamics in the temporal progress of the different plant parameters. Overall, the realism and hence quality of the simulation are seen to increase considerably by virtue of application of the refinement model. While the general effect observed for increased input dynamics is expected to stay the same, absolute values are likely sensitive to adequate calibration of the refinement model.

Nomenclature

General abbreviations

ASM	Activated sludge model
BSM	Benchmark simulation model
CHP	Combined heat and power unit
COD	Chemical oxygen demand
CSTR	Continuously stirred tank reactor
ICAWER	Interregional Concept for Advanced Wastewater Energy Reclamation
MAX	Maximum
MIN	Minimum
NH_4-N	Ammonium nitrogen
RAS	Recycle activated sludge
TN	Total nitrogen
TP	Total phosphorus
VBA	Visual Basic for Applications
WAS	Waste activated sludge
WWTP	Wastewater treatment plant

Technical symbols and general variables

α	Significance level; polynomial coefficient
β	Polynomial coefficient
γ	Polynomial coefficient
δ	Polynomial coefficient
$\dot{m}$	Mass flow rate
μ	Relative mass flow rate
σ	Standard deviation
ϑ	Relative temperature
c	Mass concentration, general

n	Number of instances
p	P-value
$profile$	Characteristic diurnal profile
Q	Volumetric flow rate of influent
q	Relative volumetric influent flow rate
r	Pearson correlation coefficient
Rel	Relative deviation
T	Temperature
t	Duration
w	Placeholder for parameter used as weight (in weighted average)
x	General variable; regressor variable
y	General variable; response variable
z	General variable of interest

Indices

a	Annual
d	Day; daily
i	Consecutive number denoting instance
$interpol$	Interpolated value
k	Inspected timescale
l	Reference timescale
m	Month; monthly
p	Period
s	Substance of interest

ASM-variables (Chapter 4; all mass concentrations)

S_A	Fermentation products
S_F	Readily biodegradable organic substrates
S_I	Inert soluble organic material
S_{ALK}	Alkalinity
S_{NH4}	Dissolved ammonium nitrogen

S_{NO3}	Nitrate plus nitrite nitrogen
S_{O2}	Dissolved oxygen
S_{PO4}	Ortho-phosphates
X_H	Inert soluble organic material
X_I	Inert particulate organic material
X_S	Slowly biodegradable substrates
X_{AUT}	Nitrifying organisms
X_{PAO}	Phosphate-accumulating organisms
X_{PHA}	Poly-hydroxy-alkanoates
X_{PP}	Poly-phosphate
X_{TSS}	Total suspended solids

Contents

Contents

1 Introduction

"The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge."

—Daniel J. Boorstin, †2004

In the Digital Age vast amounts of data are collected and processed on a regular basis. However, relying on inadequate data can distort the picture of reality and lead to false conclusions. It should generally be "Fit for Use" [1] in its intended role which sets different requirements to the numerous aspects of data quality depending on the specific application. With growing computational power and establishment of more and more sophisticated methods, simulations are continuing to become increasingly popular for design and improvement of wastewater treatment plants (WWTPs) in research as well as in practice. They have become one of the major tools for design and dimensioning of these systems and provide safe ways of investigating risk and potential of measures to be taken as well save resources compared to extensive large scale experiments. However, the quality of data output from simulation is highly dependent on that of the input data. In modelling of WWTPs, this can often propose a significant problem, as suitable data for simulation input is mostly not available in practical applications.
The research project *Interregional Concept for Advanced Wastewater Energy Reclamation (ICAWER)*, as a part of the *Interreg* program supported by the *European Regional Development Fund*, aims to improve the energy efficiency of wastewater treatment plants through collaboration of universities, consultancy companies and wastewater associations in the Austria-Italy region [2, 3]. Within the project, dynamic simulation is used for plant analysis and improvement. This requires knowledge about the plant as well as the quantity and quality of the influent, which is the wastewater to be treated. However, the data, as it is obtained from the WWTPs, is unfit for use for these types of simulation.

The central task of this thesis is the development of data refinement algorithms for creating realistic influent data to improve the quality of dynamic WWTP simulation, considering specifically the situation in the research project *ICAWER*. Refinement, in the context of this work, describes the addition of the required complexity to the system without violating statements of the underlying data.

Within this chapter, some relevant background knowledge is firstly presented including terms and concepts from the field of wastewater treatment. In the next section, the problem is described in detail and the scope of the work is defined along with the planned approach. A literature review is then presented to outline proposed methods for tackling problems of similar nature.

1.1 Relevant conceptualities in wastewater treatment

This section informs the reader about important terms and principles utilized in the field of wastewater treatment to improve understanding of the matter at hand. Firstly, relevant concepts for wastewater characterization are presented. Next, an overview is given on how relevant contaminants are removed in a modern WWTP.

1.1.1 Wastewater characterization

Wastewater or sewage refers to water which is collected in the sewer from domestic or industrial areas for purification prior to release into the environment [4]. This treatment is necessary as the water, originating from sources such as private households, industry or surface runoff, contains different contaminants that are problematic for the ecosystem.

Contaminants are generally classified into different groups defined by similar properties relevant for their removal. Sum parameters therefore cumulatively include various underlying substances and are used most commonly for the description of pollutant amounts.
Originally, wastewater treatment was focused on the removal of organic substances, including chemical compounds such as fats, proteins, acids, alcohols or carbohydrates [4, 5]. The main parameter used to describe them as a whole is the **chemical oxygen demand (COD)**, specifying the amount of elementary oxygen theoretically necessary for a full oxidation of all present organics. It is determined analytically with the aid of potassium dichromate ($K_2Cr_2O_7$) as an oxidant. The COD indicates the total amount of organics independent of their nature and hence the total carbon amount to be handled by the system. [4, 5]
More detailed subdivision of organic matter is possible and may be utilized upon relevance. For practical plant operation a categorization into particulate and dissolved form is often used [4, 5]. While settling processes are designed on particulate (settleable) fractions, dissolved fractions are used for biological process design. Commonly used models in dynamic plant simulation further distinguish between different reaction rates of dissolved and particulate COD fractions [6].
Nitrogen is a nutrient for plants and microorganisms and can lead to eutrophication of bodies of water, finally leading to oxygen depletion or failing ecosystems [7]. It is present in wastewater in different (partly) oxidized or reduced forms, some of which are toxic to wildlife. In water ammonia (NH_4^+) and ammonium (NH_3) are present in equilibrium which is affected by prevalent ambient conditions. Along with nitrogen bound in organic substances they make up the vast majority of nitrogen present in raw sewage. All three are summarized under the term total kjeldahl nitrogen. Nitrate (NO_3^-) and nitrite (NO_2^-) emerge at intermediate stages of the modern purification process, where they are often recorded for process analysis and control. The **total amount of nitrogen (TN)** is the sum of nitrogen bound in all these different forms. Similar to organic matter, specific fractioning is considered for simulation purposes.
Phosphorus represents the third important class of contaminants. It is a nutrient as well and hence shall be removed due to abovementioned reasons. **Total phosphorus (TP)** includes phosphorus in ortho-phosphates as well as organically bound phosphorus which are the forms present in wastewater.

The explained sum parameters are often used in terms of concentrations or mass flow rates (also called specific loads or for simplicity further just loads) and documented as average values over certain time periods. Loads describe the total amount of a substance to be handled by the plant while concentrations are relevant for the speed at which reactions occur. Composite samples are collected and analysed manually or automatically.

The **volumetric flow rate of the plant influent**, further also referred to as inflow or influent flow rate, represents the hydraulic load put on a plant. It is usually documented as a total or average value over a certain period of time and determines the residence time in the basins, which act as reactors. Both, pollutant and hydraulic load are consequences of the behaviour of dischargers. Additionally, the influent flow rate is largely influenced by surface runoff resulting mainly from rain and meltwater. This , in turn, has an impact on the concentration of the pollutants in the wastewater, as the relationship between the three can be described as:

$$\dot{m}_s = Q \cdot c_s \tag{1.1}$$

where Q represents the volumetric flow rate of influent, $\dot{m}$ is mass flow, c is mass concentration and the indices s and j denote the substance and point of interest respectively. Influent flow rates are often measured with automatic measuring devices.

In addition to water flow rates as well as contaminant loads and concentrations, further chemical and physical properties can be used to describe wastewater. Reaction rates are generally dependent on **temperature** and in biochemical reactions it is decisive for the activity of microorganisms [5]. Temperatures are mostly measured as instantaneous values but often documented as time-weighted averages.

Other properties are not elaborated upon due to irrelevance for the present work.

1.1.2 Contaminant removal in wastewater treatment plants

Modern wastewater treatment combines knowledge from different fields, such as biology, chemistry, process and energy engineering and control engineering [4]. In order to rid the water from contaminants, a variety of different processes is used and combined into sophisticated and complex intertwined systems. The demands set for plant effluent concentrations vary throughout different regions and commonly used technologies and apparatuses do as well. This section aims to give an overview of the most important processes and their respective purpose in a common plant setup, as shown in Figure 1.1.

Mechanical Pretreatment happens prior to the main treatment in the WWTP, upstream of where influent measurements take place. Coarse impurities, sand and grease can cause problems such as clogging or excessive wear in pumps and other process apparatuses. Coarse matter is retained by a screen and removed from there. Grit and grease removal are mostly combined within one unit where sand settles to the bottom and induced circulation transports grease into a calm zone where it can float to the water surface and be taken out. [4, 5]

Sedimentation is an inexpensive and frequently used means of removal of settleable matter by the forces of gravity. When given enough time, the difference in densities causes the heavier solids to sink to the bottom of the respective unit, where they are withdrawn in

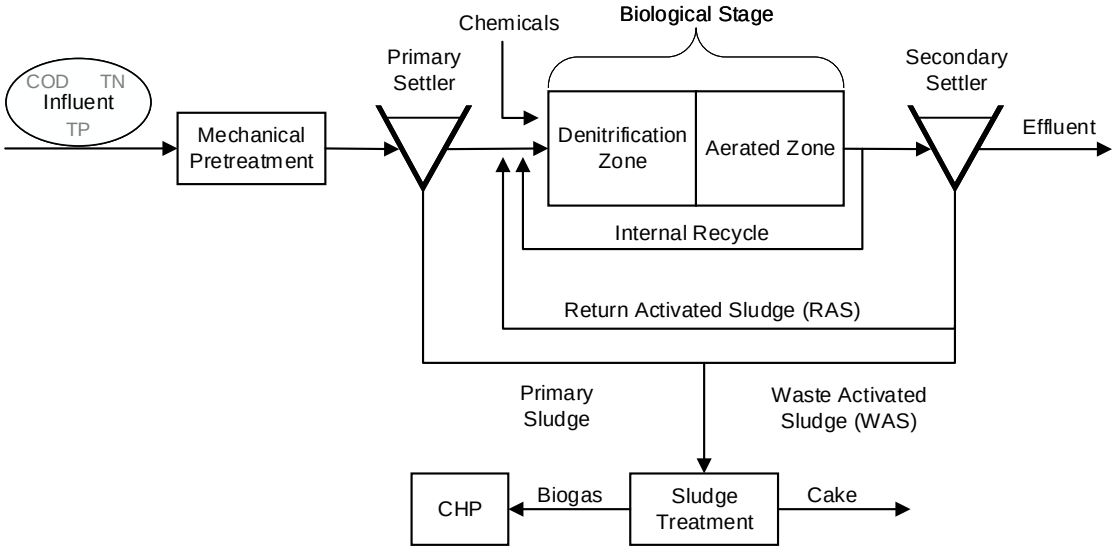


Figure 1.1. Schematic depiction of a typical modern wastewater treatment plant.

form of so called sludge. The **Primary Settler** can thus reduce the load on subsequent process components by removing **Primary Sludge** from the process.

The so called **Biological Stage** usually incorporates biological and chemical treatment. Biological treatment is used as the centrepiece of the purification process in the majority of modern WWTPs [4]. Here metabolic reprocessing of distinct types of microorganisms is utilized for transforming problematic constituents of the wastewater into acceptable end-products as well as converting dissolved into gaseous substances or settleable biomass to make them more readily removable. While using similar types of bacteria, biological treatment is generally divided into suspended growth processes, where microorganisms are kept in liquid suspension and biofilm processes, where these sit on some inert carrier material. [5]
The most widely used type of biological treatment in municipal wastewater treatment is a suspended growth process referred to as Activated Sludge Process. Here the microbial suspension is mixed with wastewater in an **Aerated Zone** [5]. In modern plants, where nitrogen removal is relevant, the classical Activated Sludge Process is extended by the implementation of an anoxic regime in the **Denitrification Zone** upstream thereof. This is to achieve different reactions for contaminant removal depending on the prevalent surrounding conditions:

- When elementary oxygen is present (referred to as oxic regime), oxic respiration of heterotrophic (meaning: use carbon from organic sources for cell synthesis) consumes COD by oxidation of organic compounds to CO_2 and H_2O. At the same time, dissolved COD is bound in form of biomass by the growth of the microorganisms, i.e. the COD is transformed to a settleable form. As the metabolism of these heterotrophic bacteria is very fast, this process leads to rapid depletion of O_2, fast removal of COD and large biomass growth.
  At the same time, two kinds of autotrophic bacteria oxidize ammonium to nitrite and subsequently to nitrate in a metabolistic process called nitrification. This, however, is much slower compared to the oxic respiration of the heterotrophic bacteria. COD

removal is hence the preferred reaction in oxic regimes and nitrification improves as COD gets depleted.
- When oxygen is only present in bound, but not in elementary form, one speaks of an anoxic regime. In these conditions the aforementioned heterotrophic bacteria substitute nitrite or nitrate for elementary oxygen as the oxidant in the degradation of organic substances. This reduces the nitrogen to elementary, gaseous form, in which it automatically exits the system.

Biomass growth generally involves a bonding of surrounding nitrogen and phosphorus as nutrients.
Precipitants and flocculants are used to enable and improve sedimentary removal from the wastewater. The addition of these **Chemicals** leads to precipitation of dissolved substances and agglomeration into larger particle structures. Phosphorus removal is very commonly enabled courtesy of the addition of iron or aluminium salts. [4]
An **Internal Recycle** is necessary to supply nitrate, as a product of the oxidation of ammonium, to the upstream Denitrification Zone. Sometimes an anaerobic zone is implemented further upstream of the entrance point of the Internal Recycle to achieve biological phosphorus elimination using specialized bacteria, which is not further described here [4, 5].

The **Secondary Settler** is used for the final clearance of the water from biomass and other particulate substances by sedimentation. To keep biomass levels in the Biological Stage at a constant level, **Return Activated Sludge (RAS)** is recycled, while **Waste Activated Sludge (WAS)** is removed from the process.

WAS and Primary Sludge undergo **Sludge Treatment**, usually including preliminary thickening, anaerobic stabilisation and dewatering, after which the resulting **Cake** has to be disposed of. The biogas emerging in the anaerobic digestion is mostly utilized in a **Combined Heat and Power Unit (CHP)** to increase the self sufficiency in terms of energy for the plant.

1.2 Problem analysis and scope of the work

WWTPs are complex systems in terms of process engineering as well as energy efficiency. Many factors, both internal and external can influence the operation of such a plant and hence impact important parameters from contaminant concentrations in the effluent to the energy usage of certain processes. Internal factors like used apparatuses or control strategies are chosen by the plant operator. External factors such as hydraulic or contaminant loads are presented by the influent as the main disturbance to the system. These are a result of environmental conditions like rainfall or meltwater and the behaviour of private households and industrial facilities discharging wastewater to the plant and can hence not be controlled by the plant operator in any way. Consequently, respective parameters commonly experience large variation over time. These variations are of great relevance for the operation of a plant, rendering the system highly dynamic.

Testing and implementing changes in the process in terms of equipment or operational strategy in reality is problematic. The complexity of the system as well as the fact that living organisms are used within the so called activated sludge processes renders it rather

delicate. The consequences of things going wrong can be severe since there is direct impact on the environment. When calibrated and used correctly, computer simulations can be powerful tools for predicting real life occurrences.
Models used for dynamic simulation of activated sludge processes describe the most important phenomena and their impact on pollutant degradation based on the kinetics of underlying biological reactions. A modelling approach, referred to as Activated Sludge Model (ASM) was originally proposed by Henze et al. within the introduction of their ASM1 [6]. Ever since, it has evolved further and grown more complex and different models have been combined into comprehensive deterministic numerical whole plant models [8]. Scientific research in the field is dealing with expanding and thus enhancing these models as well as application of the models for the development of technical advancements in WWTPs [9–11]. To have a standardized base for purposes like evaluation of control strategies, a benchmark, referred to as Benchmark Simulation Model No. 1 (BSM1), has been established and is widely accepted within research [12]. It fully defines the plant layout, used simulation models (including the ASM1 for the activated sludge process) as well as all influent parameters over a 14 day period in a detailed dynamic manner. More than 300 scientific publications related to the BSM1 were published by 2010. This shows the large demand of such a benchmark as well as the high relevance of dynamic influent data for WWTP simulation. This response of the scientific community has led to further revision and extension for long-term use. [13, 14]
Using a pre-defined benchmark is helpful for the comparability of methods developed by different research. In practical applications, however, when examinations are to be done for a specific problem, the plant setup as well as input data for influent parameters should represent the specific plant as best as possible.

Legal obligations require certain parameters to be recorded in WWTPs. Additionally, plant operators usually choose to document further supplementary variables which aid in monitoring and control. However, laboratory analyses are time demanding and automatic measuring apparatuses can be costly as well as inaccurate if not calibrated and maintained frequently. Measurement and documentation strategies are chosen by the plant operator as deemed adequate for the purpose at the specific plant and can hence vary among different WWTPs. There are many aspects to data quality and the demands to each of them highly depend on the given ambition rather than being universal [1]. When using influent measurements from a plant in different applications such as dynamic simulation, issues in terms of input data quality can arise. This is problematic, as simulation results are highly dependent on the input data.

Generally, the resolution used for simulation should enable a representation of relevant features in the dynamics of the data. For dynamic simulation in this field, input resolutions of hourly values are common and are in focus as the desired output. As reactions occur in basins of large volumes, the response time of the systems is large. Fluctuations occurring on smaller time scales are neglected, as they decrease simulation speed while often not considerably improving the outcome.

The most important parameters in the characterisation of the plant influent shall be refined for simulation purposes by the choice of appropriate methods and under comprehensive consideration of real process variables. Particularly, the specific situation given in the

research project *ICAWER* is of interest. Here, documentation of the main process variables for influent characterization can be obtained from a plant. These include the flow rate of influent, contaminant loads summarized by COD, TN and TP and the temperature of the influent, which are considered specifically relevant and are utilized for the development of refinement algorithms. Depending on the plant management, the values are generally recorded as daily values or as averages over longer time periods of multiple days or weeks up to a whole month, for simplicity further also referred to as period averages. The time span recorded values relate to is referred to as timescale in this work [1]. Consequently, the terms larger and smaller timescales refer to longer (e.g. one month compared to one day) and shorter (e.g. one hour compared to one day) related time spans respectively. Algorithms created for the data refinement shall enable ad hoc creation of hourly data for any specific plant in the research project, maximally utilizing the available knowledge about influent quality and quantity. Relevant dynamic features are to be implemented while maintaining statements made by the original data. Generally, a deterministic approach is desired, so that results can be reproduced exactly.

Several subproblems are imposed by this task:

- It is expected that erroneous values can occur in the data and hence distort the picture of reality, leading to unrealistic conditions. Clearly identifiable errors shall be eliminated.
- Gaps can be present in the data obtained from the plant. Complete datasets are required and gaps on the timescale of documentation need to be filled.
- Average values documented at daily to monthly timescales are deemed insufficient to accurately mimic relevant dynamics for simulation. Depending on the timescale of documentation, different steps might be required to reach the desired hourly outputs.

The approach to the problem is structured as following:
Within this chapter, a literature review helps in getting a first idea about approaches to similar problems as well as the thereby described dynamic phenomena found in the influent parameters. In order to build up understanding of the information, important features and problems with different data, real life industrial scale data available in the *ICAWER* project is analyzed on different timescales in Chapter 2. Using the gained knowledge, a data refinement model is then developed and implemented to provide a tool for systematically tackling the considered problems within Chapter 3, including demonstration of the model algorithms in exemplary cases. Original plant data as well as intermediate and final model outputs are finally used in a dynamic WWTP simulation in Chapter 4. Contemplation of simulation outputs relevant for plant operation and energy efficiency gives some insight into the added value provided by the refinement model.

1.3 Literature review

A study of relevant scientific literature shows the different efforts that have been taken for the generation of influent data in different situations. The most relevant approaches are outlined here along with the specific situation they are used for.

[1] As an example: Values on a daily timescale means that the values are documented as describing individual days (e.g. daily average flow rates). It does not necessarily mean that it is known for every single day.

In [15], the authors complete fragmentary experimental records obtained from a purification plant. They calculate average loads for the four seasons of a year from scarcely distributed daily measurements, disregarding the 5th and 95th percentile margins. A subsequent redistribution of this load based on a normal distribution leads to a complete dataset of daily values for the given year. Precipitation events are often differentiated into rain and storm events, where rain events typically refer to precipitation and increased resultant surface runoff over a full day or more, while storm events describe high intensity rainfall over a shorter period of time [12]. Both of these can lead to increased pollutant amounts due to runoff from impure, impervious surfaces as well flushing of deposited pollutants from the sewer system, referred to as first flush event [5, 16]. Many factors can influence whether and how much contaminant loads increase by virtue of increased wastewater amounts. First flush events are identified in [15] depending on amplitude and distribution of daily flow values. Factors are implemented with the contaminants to account for this phenomenon. The occurrence of repetitive patterns in influent parameters on an annual, weekly or daily basis is frequently mentioned in literature [17–19]. The authors of [15] claim that for the respective area of interest loads and flow rates are typically lower during the weekends, which is why they apply a weekend reduction factor on the respective days. The authors then overlay the data with a repetitive pattern for diurnal variation. No further information about the creation and features of the pattern are given, but it is used to obtain dynamic data during the course of a day based on daily averages. Overall, the approach found in [15] can be used for ad hoc generation of data when daily measurements are not available completely or to interpolate from daily values to smaller timescales. Figure 1.2 shows that the chosen methods produce a somewhat stepwise, rather than a continuous change from one season to the next and the originally known values are not included in the final data. Discontinuities are also produced at the transition between adjacent days. [15]

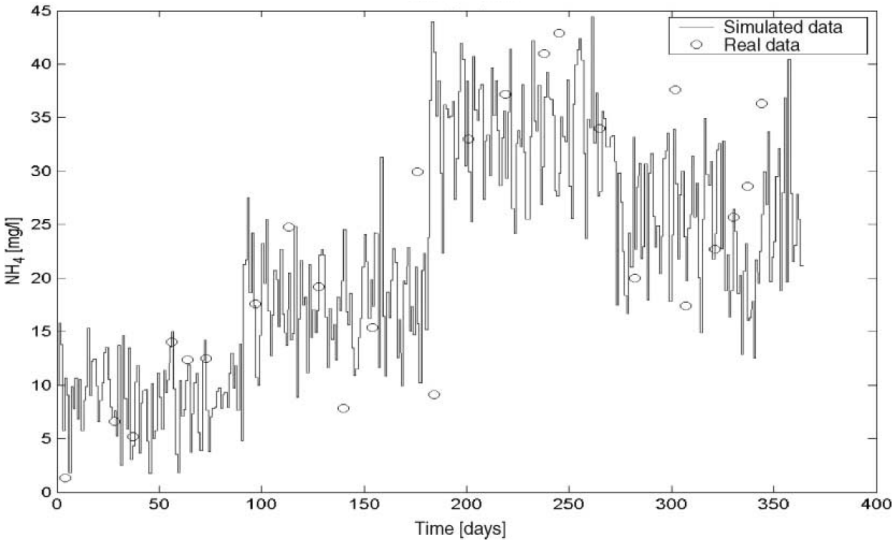


Figure 1.2. Measured and simulated data for ammonium concentrations over the period of one year as found in [15].

As diurnal variations in influent parameters have frequently been reported and characterized (see e.g. [20]), harmonic functions have been used in efforts to model and use these for the generation of more dynamic data. In [21], the flow rates and concentrations of

the influent are modelled as a sum of infiltration water, urine with flush water and other types of domestic wastewater. The periodic patterns occurring in the latter two kinds during the course of a day are described by second order Fourier series. Several form parameters necessary for a unique definition of the Fourier series are estimated by fitting the equations to measured data. When applied to varying mean values for subsequent days, the problem of stepwise changes between adjacent days found in [15] is expected. [21]
To avoid these discontinuities, [22] introduces a moving average, to which the diurnal pattern is applied. Though not specifically stated in the work, this is expected to alter the daily average values. Modelling of the pattern is done in a similar manner to [21], but the exact form of the used Fourier series differs. [22]

De Keyser et al. [18] propose an influent generator which makes use of a database of emission strings for primary pollutants collected from literature. These are overlaid with different options of daily, weekly and yearly patterns created from expert knowledge. The user can modify the catchment area by specifying different pollutant sources such as households, offices, or special dischargers like dentists, as these can have highly unique pollutant productions. Patterns can also be adjusted by the user upon desire. Stochasticity is introduced by overlay of the time series with white noise as well as the possibility to include randomly sampled pattern parameters. This tool enables a creation of influent data without any experimental measurements from the plant but merely based on knowledge about the catchment area. This can be very helpful in the dimensioning of completely new plants. However, in the prevalent case, certain plant data is available and should be used to the best possible extent. [18]

A different methodology is used for phenomenological approaches that try to model the wastewater characteristics depending on understanding of the underlying phenomena of wastewater generation. Here, the most sophisticated model to date is found in [16], which originates in efforts of extending the BSM1 for long term process evaluation. The proposed influent generator is separated into different parts, called model blocks, which account for various factors influencing the wastewater entering a plant. The annual average water amount produced by households is created based on information about the catchment area. The data is overlaid with diurnal, weekly and yearly patterns, which can be modified. This works similarly for the industry model block. A seasonal model block accounts for infiltration at different times of the year and is used together with a soil model and a rain generator to account for extra amounts of wastewater on of top of the dry weather profile. Summation of all contributing sources accounts for the final amounts of wastewater. For pollutants, household and industry block are used in a similar manner. The model also provides the possibility to convert the pollutants into the fractions used in the most popular ASM models based on fixed relationships, which is done similarly in the BSM1 (data and fractioning originally proposed by [23]). Several of the model blocks provide the possibility of including zero mean white noise with a tunable standard deviation to introduce some randomness into the model and reduce correlation between different variables. Modelling of first flush events includes calibrateable accumulation of pollutants up to a maximum amount and flushing depending on the triggering flow rate. Finally, the sewer block models the smoothing of the obtained profiles depending on the size of the sewer system. The model is deemed to be a versatile and sophisticated tool for influent data generation, taking into account a large variety of effects by different

influential factors. The model inputs do however not match the given situation. [16]

Recently, an empirical model based on a comprehensive study of observed water quality has been proposed in [24]. It uses detailed data about the influent quantity for a plant to model its respective quality by distinguishing between scenarios of dry weather and different phases of wet weather flow. Modelling different processes and superimposing them based on the identified scenario was reported to reasonably predict concentrations of different pollutants for dry as well as wet weather flow based on the hydraulic load. [24]

For introduction of uncertainty into the data generation, additional methods like the use of a time warping approach for disruption of regular profiles as well as models based on autocorrelation have been used. These are not discussed in detail at this point but can be found in [25, 26].
[17] critically reviews and comprehensively summarizes literature about analysing, completing and generating influent data for a number of situations. It provides an excellent source to be consulted as a first step upon further interest on the topic.

The literature review shows that several sophisticated approaches have been proposed for problems related to the nature of the prevalent situation. They describe several dynamic phenomena, are based on different ideas and their suitability is dependent on the specific circumstances. The used methods can partly be utilized and adapted when deemed suitable within the approach developed for the situation relevant in this work.

Preliminary Data Analysis 2

An analysis of available data shall help in understanding the information, dynamic features and problems contained within different data relevant for the prevalent situation. Another intention is to investigate whether phenomena described in literature can be found in the available large scale industrial data.

2.1 Methodology

All calculations in the preliminary data analysis are done in *Microsoft Excel 2016* or *MathWorks MATLAB 2017b*. The utilized data is available courtesy of the research project *ICAWER*. The time frame of plant data used for analysis of data gaps, errors and dynamics in monthly and daily averages is summarized in Table 2.1. While temperature is only included in the datasets of the plants Passeier, Tramin, Unteres Pustertal and Zirl, the other parameters are included in all named datasets. In case of the plants Sompunt,

Table 2.1. Time frame of data consulted for analysis on a monthly and daily basis.

WWTP	Monthly	Daily
Bozen	2016	-
Branzoll	2017	-
Brixen	2016	-
Passeier	2016	-
Pontives	2016	-
Sompunt	2016	2016
Tramin	2016	2016
Tschars	2015	-
Unteres Eisacktal	2016	-
Unteres Pustertal	2016	2016
Zirl	2015	2015

Tramin, Unteres Pustertal and Zirl, monthly values are obtained by averaging. Generally, when averaging is done in this work, time weighted averages are calculated for volumetric and mass flow rates while weighted averages with flow rates as the weight are computed in case of temperatures if not stated otherwise. Missing values from datasets are not considered. Weighted averages are calculated as:

$$\bar{z} = \frac{\sum(w_i \cdot z_i)}{\sum w_i} \tag{2.1}$$

where z_i and w_i denote the variable of interest and the weight for instance i respectively.

2.1.1 Data errors

While it is suspected that errors might be present in the data, the exact nature of these is unknown at this point. A careful inspection of suspicious values shall hence help in identifying definite types of errors in the data. Knowingly erroneous values are treated as if they were non-existent and are hence excluded in the further analysis. This is also referred to as complete case analysis [27].

2.1.2 Fragmentary data

It is relevant to understand where and in what way gaps occur in the data. Different patterns of missingness in multivariate data are described in literature [27, 28]. The data is inspected with respect to these patterns of missingness, as different methods for treating gaps in time series are suited only for certain problems [28]. This is relevant so the approach taken in the later development of the refinement model can be chosen in a way so that all prevalent gaps are treated in a suitable manner.

2.1.3 Correlation of parameters based on daily averages

From logical understanding of wastewater generation mechanisms it becomes clear that contaminants often originate from the same sources. Moreover, the amount of wastewater produced by different dischargers is expected to stand in context with pollutant loads and water temperature. Therefore, the presence of linear correlation between different parameters in available daily data is investigated. Correlation coefficients indicate the presence of linear relationships between the observed variables and can take values between -1 and 1, where values close to 1 indicate a strong positive relationship, values close to -1 indicate a strong negative relationship and values close to zero indicate no linear relationship. The null hypothesis claims that there is no linear relationship between chosen parameters. A t-test is conducted in *MathWorks MATLAB* according to standard methods [29]. The corresponding p-value, indicating the likelihood of seeing results as least as extreme as prevalent, is calculated and the null hypothesis is rejected if the p-value is lower than the significance level α, which is chosen to be 1 %. This is to examine the statistical significance of the result. Pearson's correlation coefficient r (see Equation 3.1) as well as p-values are calculated in *MathWorks MATLAB 2017b* which follows established methods from literature [30]. Correlation coefficients alone can be misleading as they might indicate linear relationship where it is not given [31]. Therefore, scatterplots are inspected visually to examine the statements made by the correlation coefficients.

2.1.4 Dynamics in monthly averages

The dynamics contained in monthly values are examined. The plant sizes of the considered WWTPs and hence total hydraulic and contaminant loads differ greatly. As temporal changes are relevant for this work rather than the absolute values, relative variations around mean values are contemplated for interpretation. This enables a better comparison of different plants. Curves of the respective variables are plotted to allow for visual inspection and educated interpretation.

New expressions are introduced to describe relative values for different timescales and parameters.
For the influent flow rate this is:

$$q_{k/l} = \frac{\bar{Q}_k}{\bar{Q}_l} \tag{2.2}$$

where the subscripts k and l indicate the inspected timescale and the reference timescale respectively. Specific timescales will be denoted by subscripts a for annual, m for monthly, d for daily and h for hourly. In words $q_{k/l}$ can be described as the average volumetric flow rate of influent on timescale k relative to the average volumetric flow rate of influent on timescale l. As an example, $q_{m/a}$ is the monthly average of the flow rate of influent relative to its respective annual average.

Similarly, for the loads:

$$\mu_{s,k/l} = \frac{\bar{\dot{m}}_{s,k}}{\bar{\dot{m}}_{s,l}} \tag{2.3}$$

In words $\mu_{s,k/l}$ can be described as the average mass flow rate of substance s on timescale k relative to the average mass flow rate of substance s on timescale l. As an example, $\mu_{COD,d/m}$ is the daily average of mass flow rate of COD relative to its respective monthly average.

Lastly, for temperature:

$$\vartheta_{k/l} = \frac{\bar{T}_{inf,k}}{\bar{T}_{inf,l}} \tag{2.4}$$

where T is the relative temperature in °C. In words $\vartheta_{k/l}$ can be described as the average influent temperature on timescale k relative to the average influent temperature on timescale l.

2.1.5 Dynamics in daily averages

Similar to the monthly values, dynamics in daily averages are examined contemplating relative variations. For the display of the curves, centered 31 day moving average values are included, considering 15 days each before and after the respective day. This is to illustrate a seasonal baseline. Special attention is paid to recognizing seasonal and weekly variations, which are frequently reported in literature [16–19].

2.1.6 Dynamics in hourly averages

Understanding the dynamics occurring on this timescale is relevant, as hourly data is the desired output in this project. However, as no hourly data is available from the plants for influent flow rates or contaminant loads, some substitute data is contemplated.
The BSM1 contains high quality input data in 15 minute intervals for Q as well as $\dot{m}_{COD}$ and $\dot{m}_{TN}$. Averaging allows for an inspection of the respective values on an hourly basis to show the dynamic features contained in this type of WWTP influent data.
Expert knowledge from the research project unveils that usually around 60 % of all nitrogen in the influent is bound in form of ammonium or ammonia, also summarized under the term ammonium nitrogen (NH_4-N). The parameter can hence give indications about the behaviour of the TN content. A high quality time series of the concentration of ammonium nitrogen c_{NH4-N} from the WWTP Tschars is available. However, influent

flow rates and hence loads are not included in the dataset. The provided data series of c_{NH4-N} is converted to hourly values by means of time weighted averaging and used for inspection.
Additionally, highly resolved temperature measurements available courtesy of the WWTP Zirl are converted to hourly values and investigated. Time weighted averaging has to be used, as no influent flow rates are available.

No data about phosphorus in the influent is currently available on this timescale. However, some conclusions are made from the interpretation of the phenomena found in other parameters.

2.2 Results and discussion

This section is to show the results of the data analysis, so a thorough understanding for the features and problems found in the different dynamic data can be created.

2.2.1 Data errors

No definite errors can be identified in daily influent flow rate measurements. The maximum peaks observed in the data are between 250 and 380 % for the different locations. All values are in a realistic range, but could also result from flawed measurements.

The inspection of suspiciously high, monotone or low values in the daily data for the parameters and plants specified in Section 2.1 reveals certain definite faults. For the WWTP Tramin some irregularities are found in the temperature. Here, measured values drop to zero in the middle of the year, stay at exactly zero for a while, and finally jump back to the range they had previously been in. These can be identified as erroneous with great certainty and likely indicate a malfunctioning measurement device.
For contaminant loads, an important fault was detected. Consultation of underlying concentrations reveals a profile that is stepwise constant in intervals of multiple days for the WWTP Tramin. Concentration measurements are done approximately once a week and, combined with the influent flow rate, used to calculate the daily load according to Equation 1.1. Until the next measurement, the last known concentration value is used for this calculation. However, this can lead to a severe misconception. When the measurement is conducted on a day with low influent flow rate and regular or above average pollutant loads, the resulting concentration is high. If the same total amount of pollutants is produced by dischargers the next day, but there is a lot of relatively unpolluted surface runoff, e.g. from rain or melting water, flow rates increase significantly, resulting in low concentrations in reality. However, upholding of the previous (high) concentration value and determination of the load using the high flow rate results in incorrectly large pollutant amounts unrepresentative of real life occurrences. In Figure 2.1, an example of this is shown for the COD loads and concentrations. The concentration measurement is conducted on June 22 and upheld for the five following days. Due to rising flow rates during this time (not shown in the Figure), the resulting loads increase nearly three-fold. The same phenomenon in reverse can also lead to underestimation of loads. When concentration measurements are done on days with a lot of highly diluted influent (e.g. by large amounts of rain without any significant flushing) this can lead to excessive declines

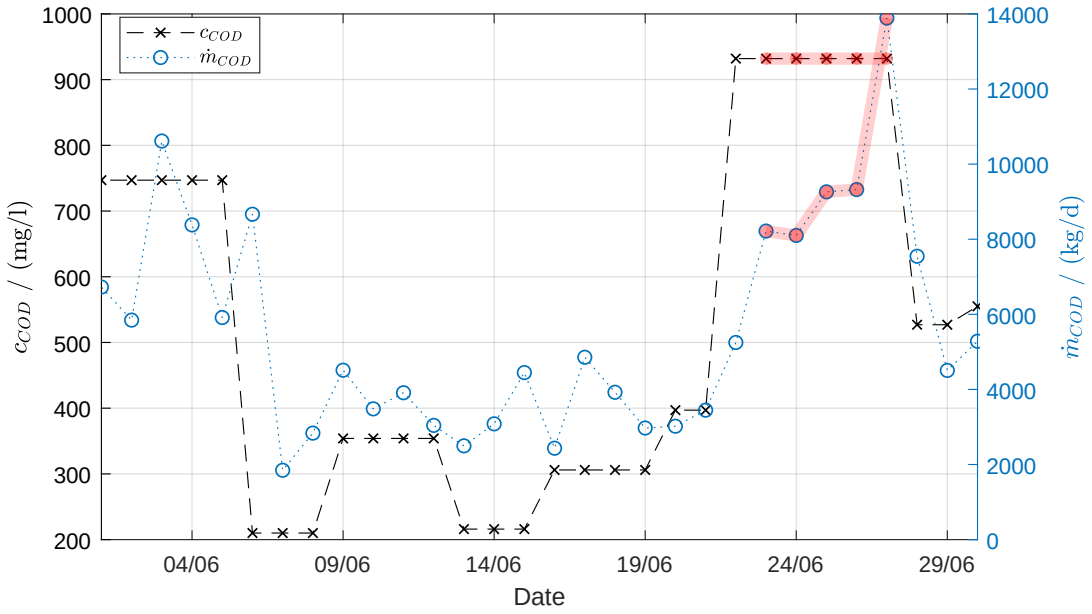


Figure 2.1. Upholding of outdated values for COD concentration leading to misleadingly high values in the respective load (area of attention marked in red) at the WWTP Tramin.

in loads when they are upheld and influent amounts decrease subsequently. Generally, an upheld concentration value and hence the computed load is considered bland in statement. While no definite errors could be found in the monthly data obtained directly from the plants, these values are calculated internally, and it is possible that the underlying data for this contains some of the identified problems. This exceeds the scope of this work and the monthly data is considered as adequate.

2.2.2 Fragmentary data

All the data obtained as monthly average values from the different plants is complete and does not have any gaps in it. On a daily basis, influent flow rates as well temperatures (when included in the dataset) are seen to only rarely contain short term gaps in an arbitrary manner, probably owed to failure of automatic measurement devices. Contaminant loads and concentrations are fragmentary for the plant Sompunt. These are only determined approximately once a week and the values in between are left as gaps. Expert knowledge obtained internally from the research project reveals that this is a measurement and documentation strategy found also in other plants. As clearly erroneous values are eliminated, this leaves additional gaps of different frequency and length.

Overall, observed patterns of missingness in the data vary between uni- and multivariate as well as monotone and arbitrary (based on the pattern descriptions found in [27, 28]). Length, frequency, and position of the gaps are subject to the individual observed plant and parameter and cannot be stated universally.

2.2.3 Correlation of parameters based on daily averages

Table 2.2 shows the minimum and maximum values (denoted MIN and MAX) obtained for the values of the Pearson correlation coefficient r for the parameters within a certain WWTP, using data from Tramin, Zirl, Unteres Pustertal and Sompunt. The correlation found for the different parameters is generally the highest between the loads of the different contaminants. This seems reasonable due to their similar origins. Correlation coefficients for inflow and pollutants are all positive, but vary widely. Temperature correlates the least with the other parameters and coefficients are seen to take both positive and negative values. The data can be seen as samples from the total population. The t-test and calculation of respective p-values shows that at a significance interval of 1 % there is sufficient evidence to reject the null hypothesis and hence conclude a linear relationship between loads of different contaminants as well as between influent flow rates and loads in all cases except the relationship of influent flow rate and TN load at the WWTP Zirl. Visual inspection of the scatterplots helps in approving the assumption of linear relationship for most of these cases. However, strong outliers are often seen in scatterplots of influent flow rate combined with pollutant loads and the linear correlation does not seem to hold for values of very high flow rates.

Table 2.2. Minimum and maximum values for correlation between different parameters within a WWTP based on daily averages from Tramin, Zirl, Unteres Pustertal and Sompunt

	Q		$\dot{m}_{COD}$		$\dot{m}_{TN}$		$\dot{m}_{TP}$	
	MIN	MAX	MIN	MAX	MIN	MAX	MIN	MAX
Q								
$\dot{m}_{COD}$	0.28	0.67						
$\dot{m}_{TN}$	0.06	0.80	0.60	0.92				
$\dot{m}_{TP}$	0.26	0.78	0.38	0.92	0.30	0.95		
T	0.10	0.32	-0.12	0.24	-0.40	0.20	-0.14	0.17

2.2.4 Dynamics in monthly averages

While parameters for all plants were analysed according to Section 2.1, only some datasets are selected for depiction and discussion as relevant. Figures 2.2, 2.3, 2.4 and 2.5 show progress of the monthly averages relative to the yearly average values for the influent flow rate and the load of COD, TN and TP, respectively. For all of these parameters it is seen that the monthly averages contain information about seasonal variation. Since the dischargers connected to the plants are disparate combinations of private households and industry, amplitude as well as pattern of the temporal changes vary between different plants. The general form of the curves for hydraulic and contaminant loads within a given plant seem to resemble one another. However, concentrations generally do not stay constant throughout the year. This can be seen for example at the plant Unteres Pustertal, where the magnitude of change of the influent flow rate is much lower compared to all the contaminant loads, indicating higher concentrations from December to February and more diluted wastewater in spring and autumn. This is most likely a result of seasonal operation from heavy duty dischargers in the surrounding industry. Moreover, the summer peaks vary between TP and the other contaminants indicating a shift in the pollutant amounts relative to each other throughout the season.

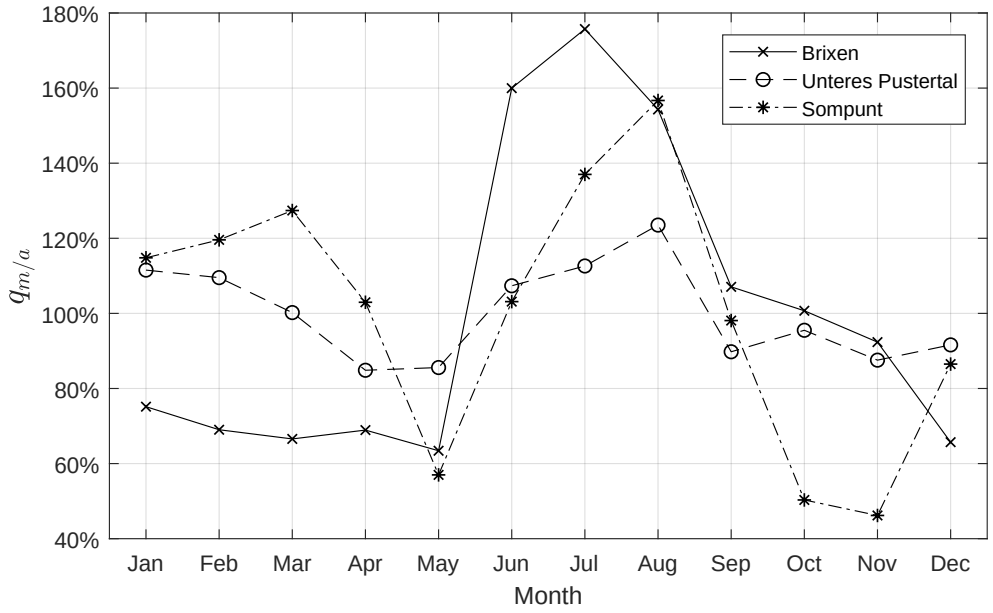


Figure 2.2. Monthly variation relative to annual mean for the volumetric flow rate of influent in 2016.

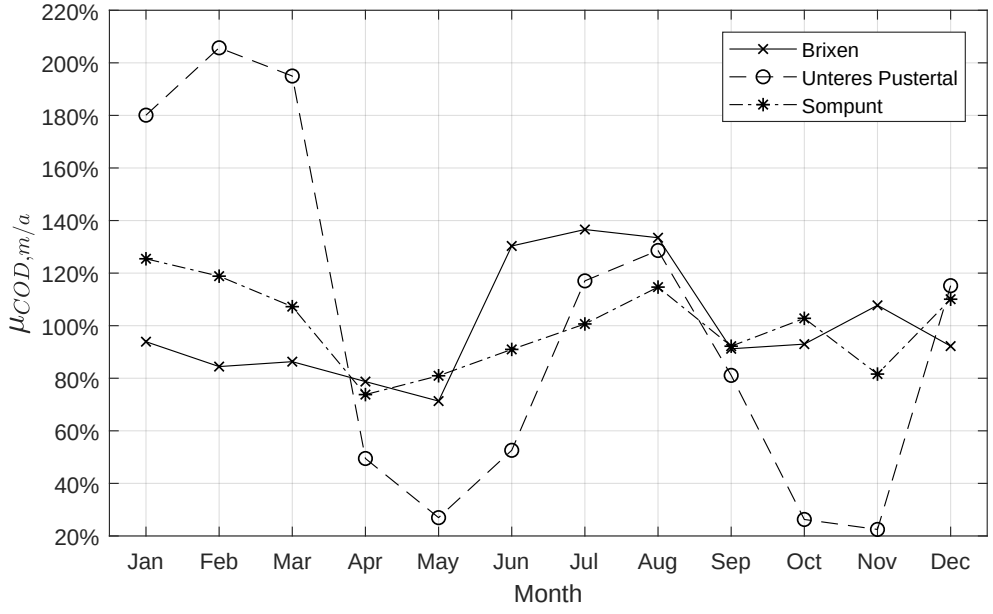


Figure 2.3. Monthly variation relative to annual mean for the COD load in 2016.

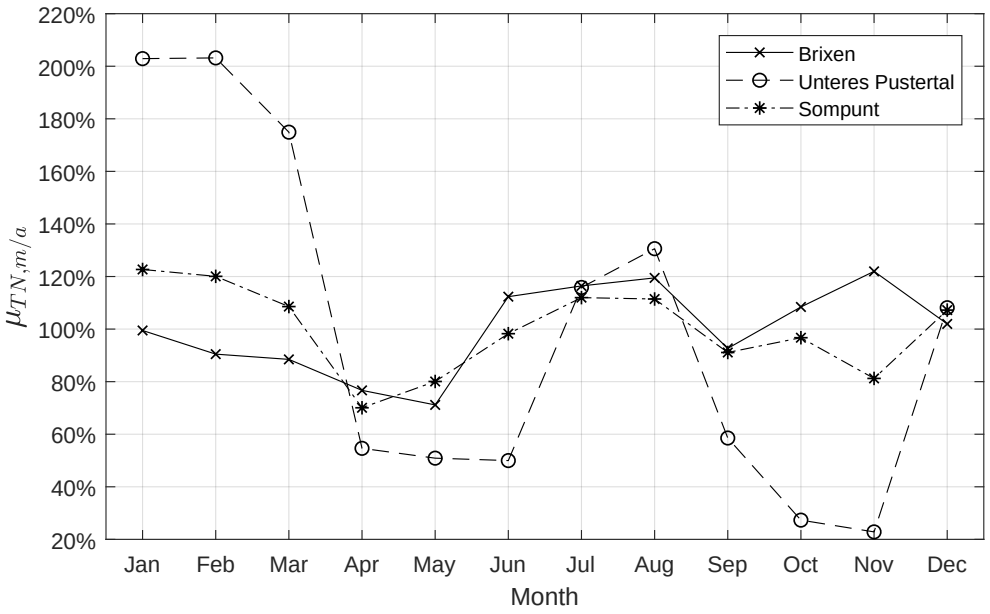


Figure 2.4. Monthly variation relative to annual mean for the TN load in 2016.

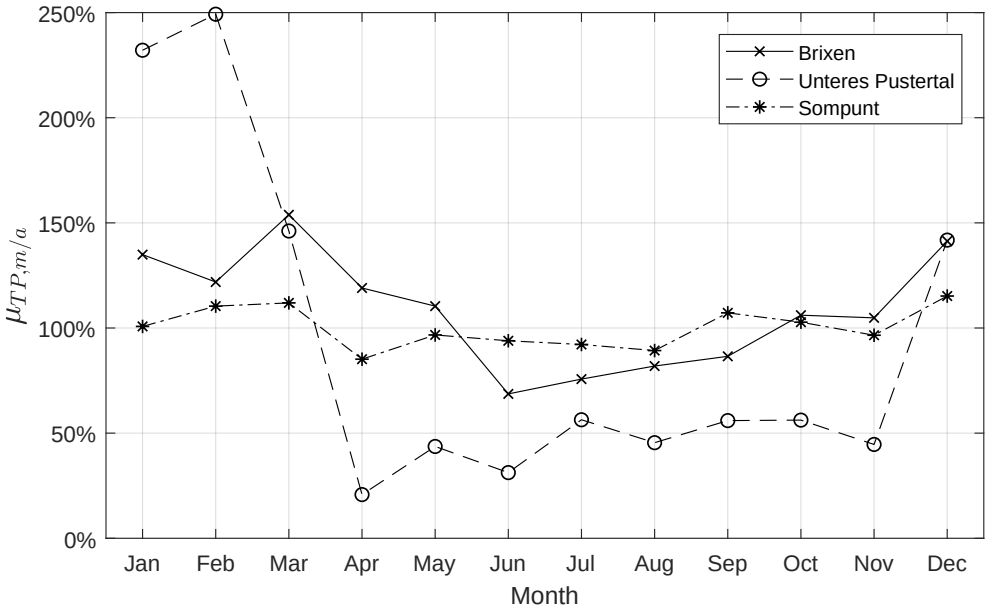


Figure 2.5. Monthly variation relative to annual mean for the TP load in 2016.

The depicted WWTPs are located in Southern Tyrol, which experiences a lot of tourism in winter, summer, or both, depending on the area. This is another factor contributing the differing shapes of the presented curves, which are highly individual for each plant.

In Figure 2.6, the temporal progress of monthly influent temperatures relative to the annual average is depicted for Passeier, Unteres Pustertal and Tramin. Among others, relevant factors contributing to it are likely the temperature of different produced wastewater, the time of flow from the discharger to the plant and very importantly ground and ambient temperature. It is seen that the general shape of the temperature variation is quite alike for all plants. However, the amplitude of variation changes.

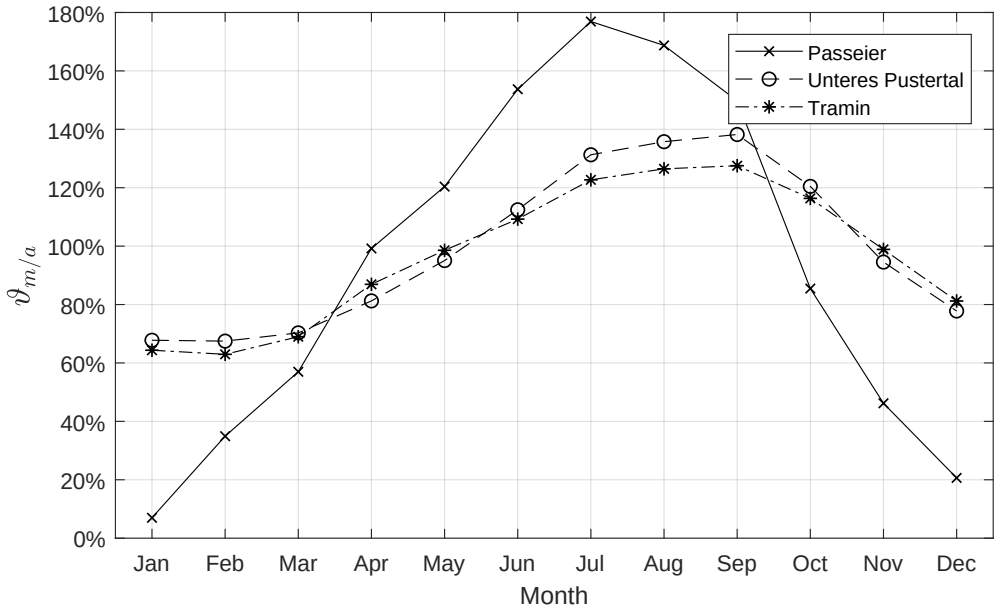


Figure 2.6. Monthly variation relative to annual mean for influent temperature in 2016.

2.2.5 Dynamics in daily averages

Daily values obtained from the plant Sompunt are selected from the analysed data (see section 2.1) to show an example of the observed features. Figure 2.7 and 2.8 depict the behaviour of the volumetric flow rate of influent and the load of COD relative to the annual mean. TN and TP are not shown separately. They are only measured approximately once a week at the plant and the detail of information in these is hence lower, but their behaviour generally was seen to be very similar to that of COD. Both, influent flow rates as well as contaminant loads demonstrate significant variation throughout the year. The flow rate of influent seems to follow some sort of baseline, as indicated by the 31 day symmetric moving average, with small fluctuations. The baseline changes due to the amount of dischargers and their behaviour. On top of that, a number of larger peaks with differing amplitude is seen and is likely a result from surface runoff. The frequency of these changes throughout the season with many seen in the summer months. In the region of South Tyrol, Italy, much of the precipitation in winter occurs as snowfall in higher altitudes. This results in a delayed increase in wastewater amounts influenced by the ambient temperature rather

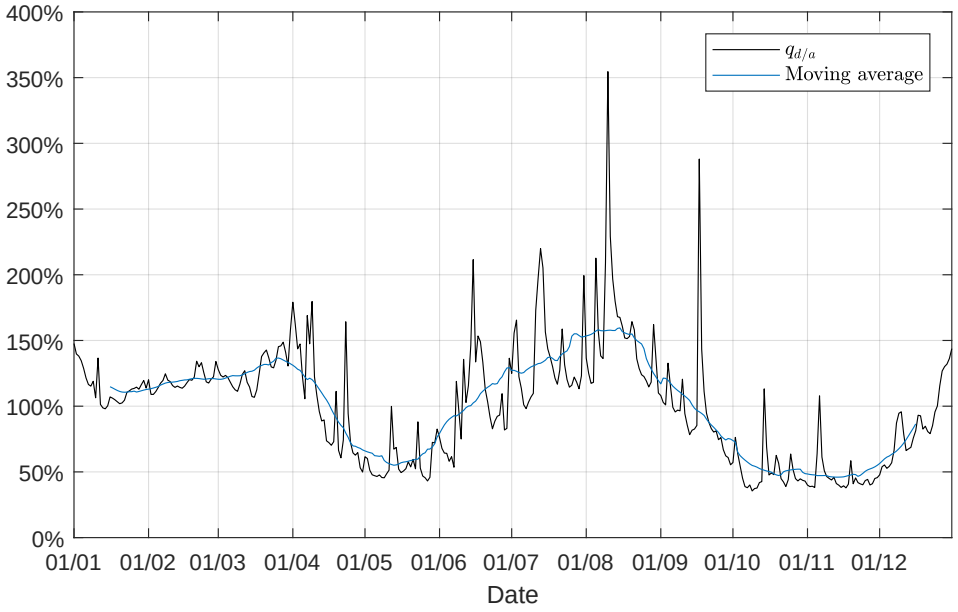


Figure 2.7. Daily variation relative to annual mean for the volumetric flow rate of influent along with a centered 31 day moving average at the WWTP Sompunt in 2016.

than having an immediate effect. The peaks contribute with an extra of up to around 200 % of the annual average on top of the baseline.

Contaminant loads express similar phenomena, including the appearance of some baseline, small fluctuations and large peaks. The combination of flow and COD load (as seen for mid February to mid July in Figure 2.9) is consulted for more insight into the origin of the varying loads. Up until the end of March, where the flow baseline is high, very large peaks are seen in the COD load but not in the influent flow rate. This could indicate that either some large scale industrial discharger is producing very highly concentrated waste water in certain occasions or it could originate from faulty measurement. Both of these options are seen as possible and are expected to occur in reality. TN and TP were not recorded at the days when the large peaks appeared, so they cannot be consulted for aid and no definite statement is made about their root cause. Coinciding peaks of influent flow rate and contaminant loads are seen more after mid March, though they are much lower compared to the previously seen ones. These could occur for example from public events on the weekends or from first flush events.

The basic described features are also found in the data of other plants: Influent flow rate as well as COD, TN and TP loads show fluctuation around a clearly recognizable baseline. This baseline, however, varies less throughout the different seasons compared to Sompunt. A clear difference between weekends and weekdays in terms of wastewater quantity and contaminant loads cannot be identified for the WWTP Sompunt, but a slight reduction in a range of up to 10 % compared to the preceding and succeeding weekdays is observed on weekends for some of the other plants.

Figure 2.10 shows the variation of the daily mean influent temperature relative to the annual mean for the WWTP Unteres Pustertal. The general shape of the baseline

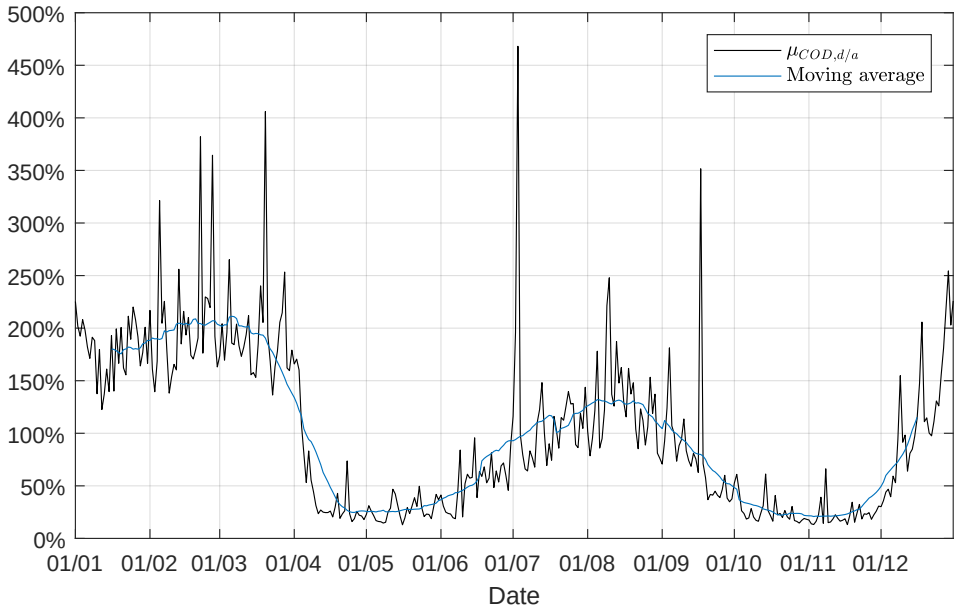


Figure 2.8. Daily variation relative to annual mean for the load of COD along with a centered 31 day moving average at the WWTP Sompunt in 2016.

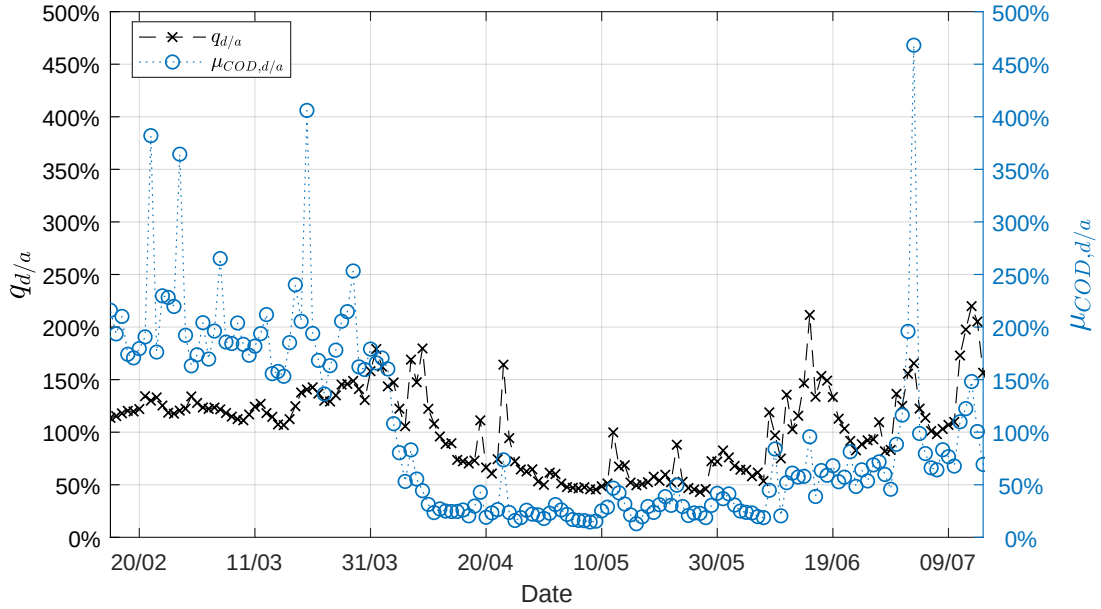


Figure 2.9. Daily variation relative to annual mean for influent flow rate and COD load at the WWTP Sompunt from mid February to mid July 2016.

indicated by the moving average looks very much alike for all observed plants. Moreover, the amplitude of fluctuations around this baseline as well as absolute temperature values are in similar ranges.

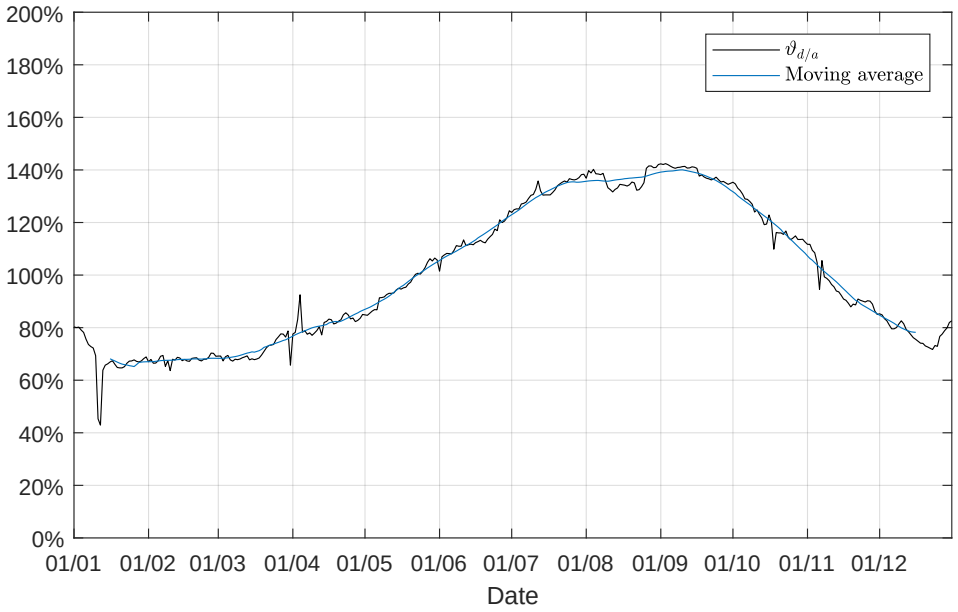


Figure 2.10. Daily variation relative to annual mean for the influent temperature along with a centered 31 day moving average at the WWTP Unteres Pustertal in 2016.

2.2.6 Dynamics in hourly averages

The examination of hourly data gives some insight into the dynamics that are incorporated in data of this temporal resolution. Figures 2.11, 2.12 and 2.13 show the influent flow rate, the load of COD and the load of TN over a period of 336 hours (14 days) for the three scenarios given in the BSM1. These are a scenario where all days are dry and no surface runoff occurs, one where prolonged rain occurs over two consecutive days in this period and one where storms lead to short term high intensity surface runoff in two occasions. It can be seen that loads and water flow rates experience a similar looking repetitive pattern on a daily basis. Wastewater amounts and contaminant loads are lower during nighttime, after which they rise to a peak in the morning. Another, slightly lower peak is seen in the evening. The amplitude of these daily variations differs slightly and lower peaks are seen on what is interpreted to be the weekend. For dry days, the relative variations are in a magnitude of about 50 % around the daily average and the highest value is mostly between two and three times higher compared to the lowest value observed on a specific day. This pattern is likely a result from the discharging behaviour of private households with increased wastewater production at times of cooking and using sanitary services. The prolonged rain is seen to increase the total influent flow rate over the concerned days while approximately preserving the general shape and hence the difference between peaks and low values within a given day. The two peaks seen in the storm scenario are seen to significantly alter this pattern and spikes can be observed in the influent flow rate where the increased runoff occurs.

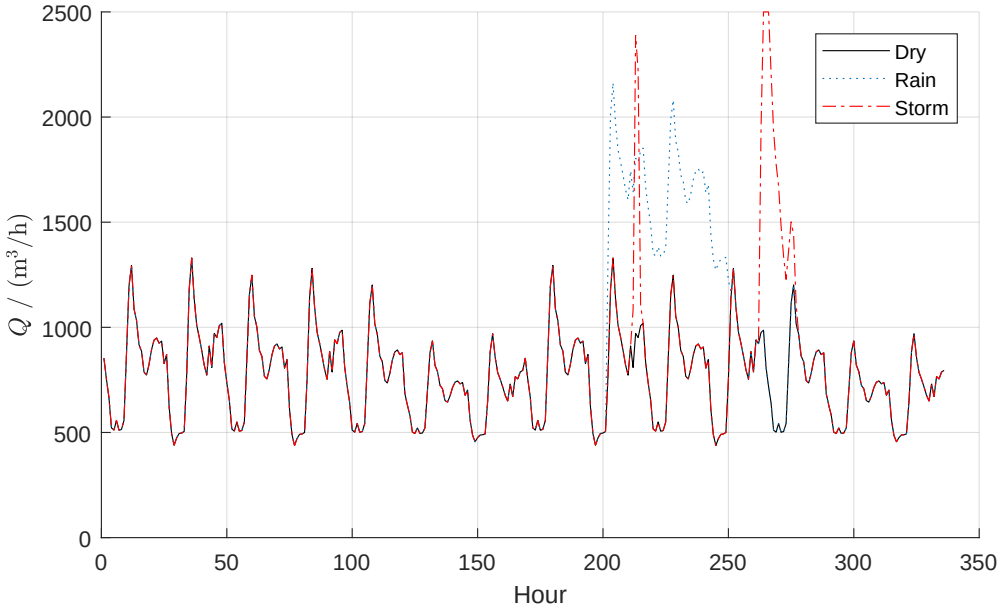


Figure 2.11. Volumetric influent flow rate in hourly averages from the BSM1 for three distinct scenarios.

In contrast to that, the load of COD (Figure 2.12) is behaving differently for the rain and storm scenarios. While being influenced by the prolonged rain only marginally, the first storm occurrence leads to a very high peak which is more than 450 % of the average load in this period. The spike from the second storm incident, though leading to even higher wastewater flow rates, is by far not as significant. These peaks are interpreted as a result of first flush events and hence do not only depend on the wastewater flow rate, but also on other factors such as the time span since the previous storm event. The relative variation on a normal, dry day is larger for the load of COD compared to the influent flow rate, with the load increasing on average more than five-fold for the highest compared to the lowest value of a day. This indicates reduced concentrations for the night.

The variation of the TN load (Figure 2.13) within regular days strongly resembles that of COD, while absolute numerical values are lower. It is also seen to be affected in a comparable manner by rain and storm, but the peaks in the loads resulting from storm are not as extreme as for COD. Moreover, comparison of the diurnal profile of $\dot{m}_{TN}$ and Q also indicates lower concentrations during the night.
TP loads are not available for an analysis on an hourly basis. Logical understanding of contaminant sources as well first flush events leads to the conclusion that the TP loads likely behave similar to TN and COD. The relative height of the peaks among the three parameters is subject to the composition of deposited (and hence subsequently flushed) impurities.

Figure 2.14 shows averaged values of measurements for c_{NH4-N} obtained from the WWTP Tschars. It clearly demonstrates characteristics reoccurring on a daily basis, similar to the afore examined loads and influent flow rate. Similar to the BSM1, concentrations are lowest during the night, implying a more extreme nocturnal decrease in the loads compared to the influent flow rate. The daily peak concentrations are seen to be more

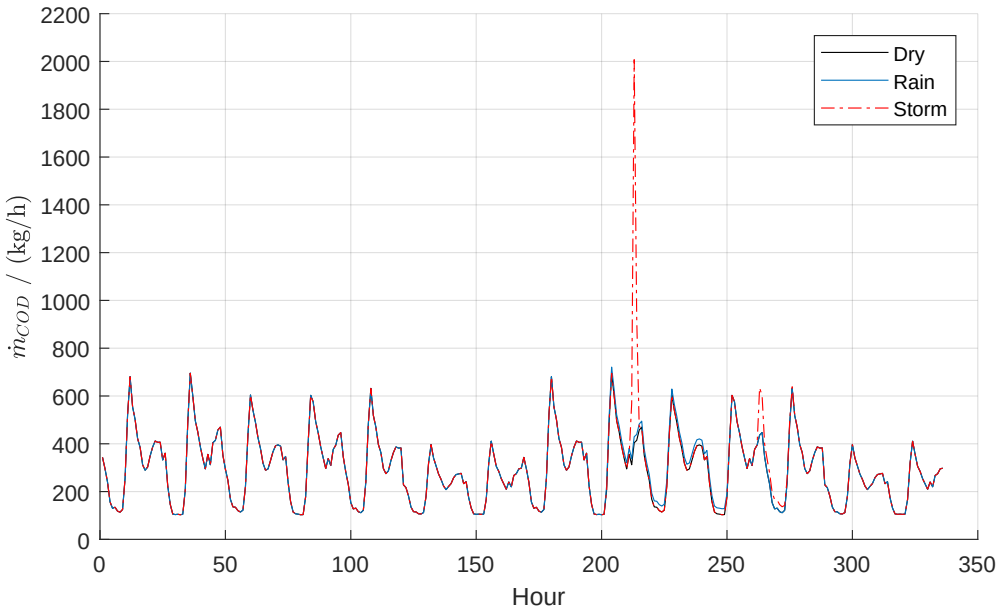


Figure 2.12. Load of COD in hourly averages from the BSM1 for three distinct scenarios.

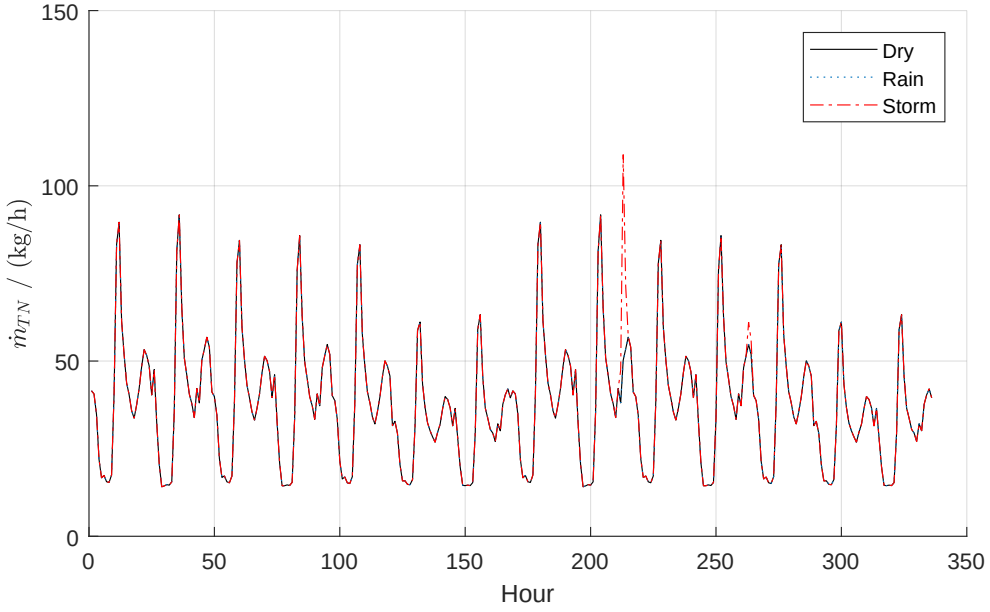


Figure 2.13. Load of TN in hourly averages from the BSM1 for three distinct scenarios.

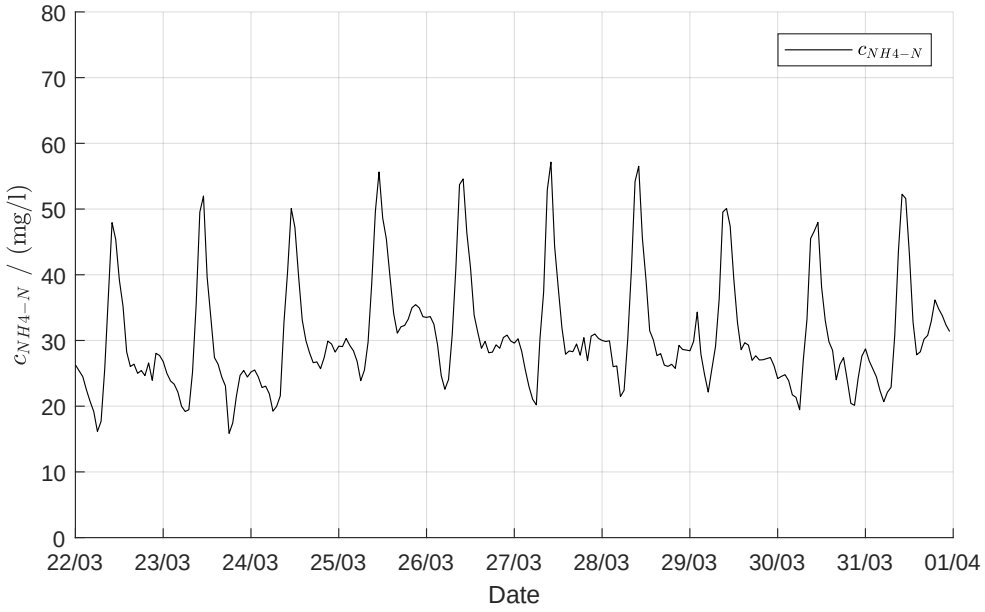


Figure 2.14. Concentration of ammonium nitrogen as measured in 2018 at the WWTP Tschars.

than double of the lowest daily value. However, this is suspected to be highly individual for different WWTPs. As ammonium makes up a large part of the influent nitrogen, this behaviour is suspected to be representable for the concentration of TN.

The top half of Figure 2.15 shows an excerpt of hourly influent temperatures from the WWTP Zirl recorded in October 2017. Similar to the afore inspected parameters, a repetitive pattern can be detected. The explosion in the bottom half of the figure indicates that low values are observed during the night and increasing temperatures during the day and a peak in the evening. Influent temperatures are seen to be slightly lower on the weekends compared to weekdays. The general shape of the variation being slightly altered compared to the other parameters could also result from the fact that data sources and hence respective ambient conditions differ. From practical understanding it is clear that the influent flow rate has a direct impact on the temperature of the water. While ambient and soil temperatures are mainly made responsible for the seasonal change observed in this parameter, the observed daily variation is probably mostly courtesy of different water dischargers. The high peaks in the evening hours possibly indicate increased usage of hot water for body hygiene in private households.

In summary, it should be noted that a repetitive pattern reappears on a daily basis for all examined variables. There is significant variation around the daily averages within the course of 24 hours, caused by the discharging behaviour of connected households and industry. The general appearance of the variation for the different parameters is similar, with peaks occurring simultaneously. Its intensity can change on the weekend due to different activity of private persons and businesses. The interpretation of the patterns appears to make good sense for wastewater from private households but one should be aware that the characteristics could be significantly altered by large scale industrial wastewater producers connected to a plant or personal habits related to wastewater production common in different areas and cultures. Occurrences of prolonged rain

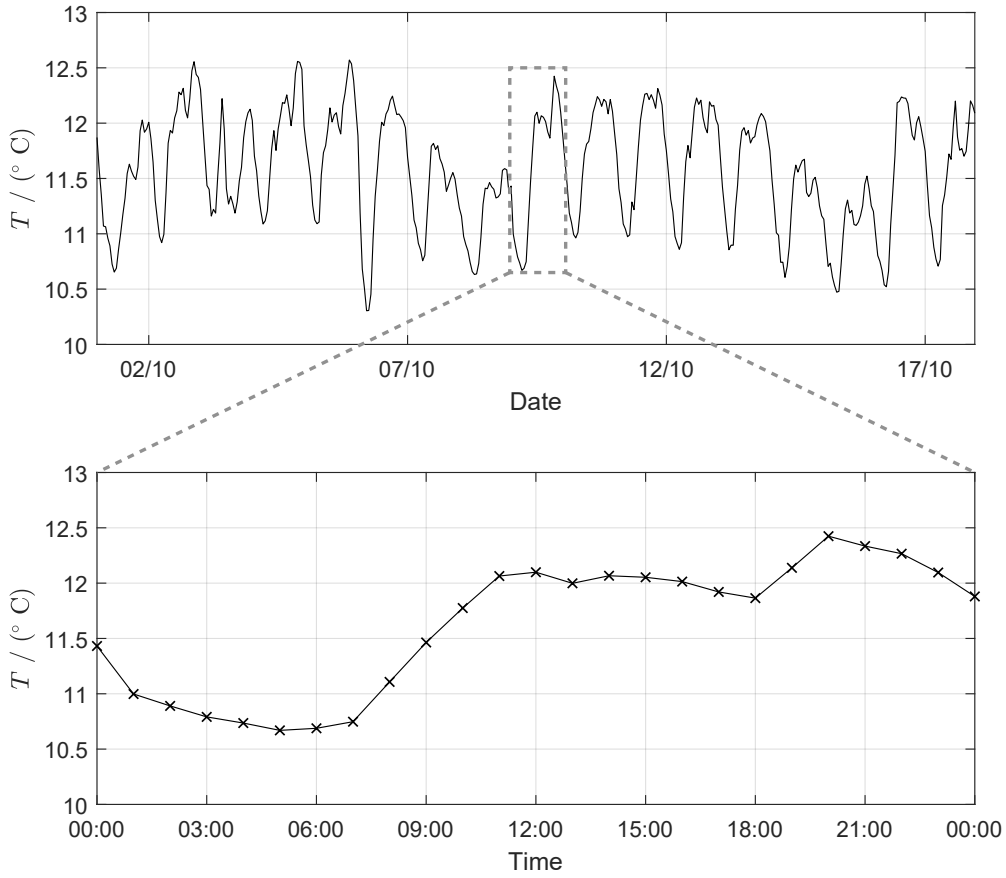


Figure 2.15. Influent temperatures recorded at the WWTP Zirl. Top: A period of 19 consecutive days. Bottom: Explosion showing one full day.

influence mostly the influent flow rate on a daily basis, hence diluting the wastewater, while storms can lead to irregular peaks interrupting the repetitive pattern for received water and pollutant amounts. The intensity of variation can vary for different parameters. It is expected that the variations become less intense for larger plants. When pipeline distances between discharger and treatment plant vary largely, differing times of travel for the water and hence results in an increasing distribution of wastewater caused by an instantaneous event [21, 32]. This is also indicated in the most commonly used guideline for plant dimensioning where static daily values are used in the calculations along with safety factors to account for daily peaks as suggested values to be adopted for these safety factors decrease with plant size [33].

Data Refinement Model 3

This chapter explains the data refinement model in detail. Chosen methods are outlined along with their implementation. The created model algorithms and resulting model files are discussed and application of the model in an exemplary case illustrates the inner workings of the individual procedures.

3.1 Methodology

With the knowledge gained from the data analysis in Chapter 2, methods are developed for dealing with different subproblems associated with the two distinct types of input data. Approaches found in literature (see Section 1.3) are utilized in parts and adapted when seen as suitable. The practical implementation of the model is explained and the model setup for exemplary demonstration is finally shown.

3.1.1 Creating a complete set of daily averages from daily input data including errors and gaps

As shown in the preliminary data analysis, daily plant data can suffer from misleading, erroneous values and is often fragmentary to a certain extent. Figure 3.1 shows Algorithm A, which is specifically designed to resolve these problems and produce a set of complete daily data. The different steps are explained in detail in the respective parts of the text.

3.1.1.1 Eliminating data errors

Two main types of errors could be definitely identified in the data in Section 2.2.1: Untrue zero values and false loads based on upheld rather than measured concentrations. Therefore it is important to consider how the concerned values can be identified systematically and how they should be treated subsequently. As no definite faults could be identified in monthly data, the approach here is focussed on recordings on a daily basis.

Identifying untrue zero values

The use of a computer program enables finding all numerical values of zero in the dataset. However, one needs to consider what additional criteria might be used to distinguish true from untrue values. Therefore, logical understanding of the physical nature of considered parameters is consulted.
With wastewater from great numbers of sources ending up in the sewer and finally in the WWTP, there is always a certain amount of wastewater and contained pollutants

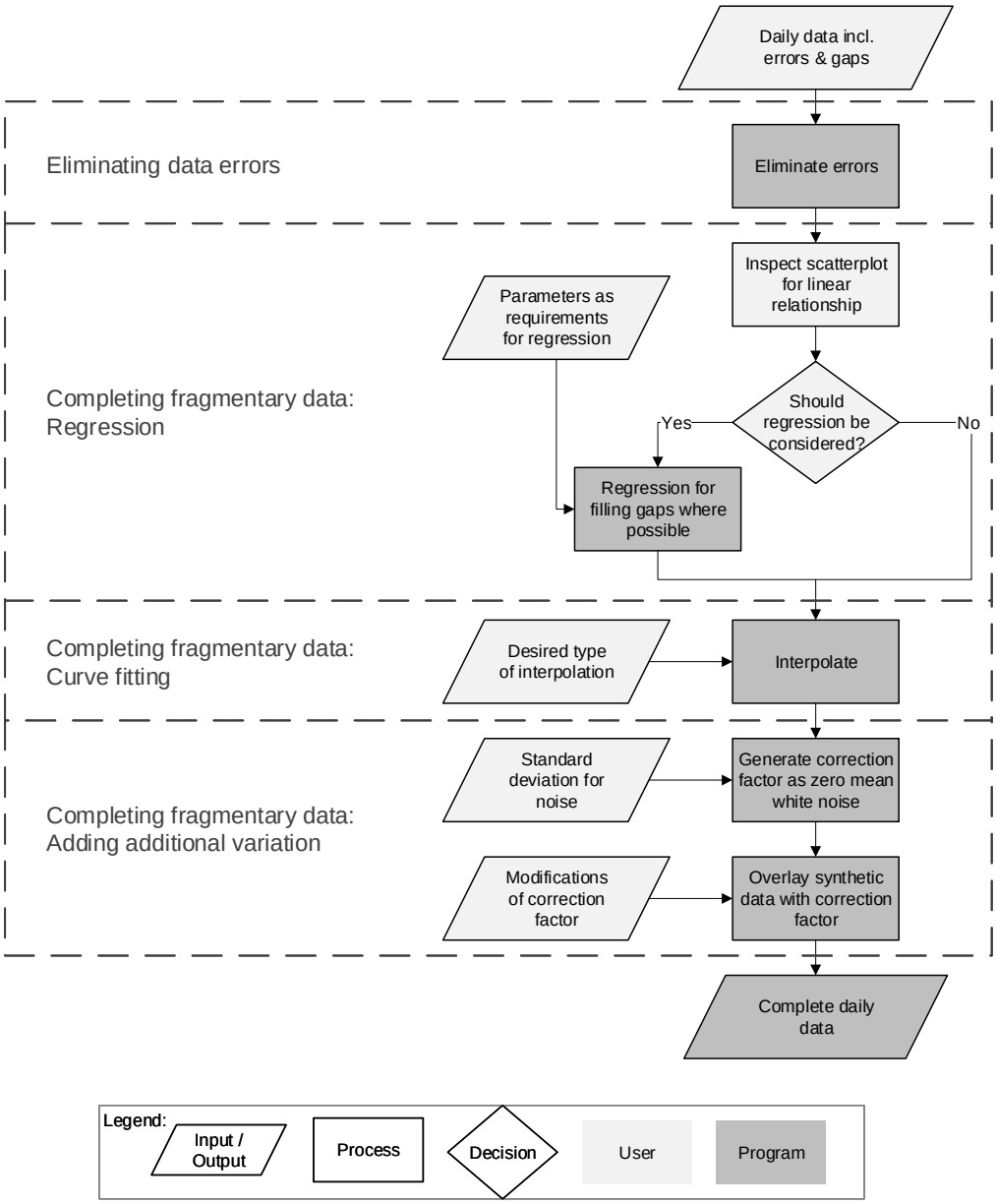


Figure 3.1. Algorithm A: Creating a complete set of daily averages from daily input data including errors and gaps. A detailed description of the procedures within the dashed boxes is found in the respective text section indicated on the left side of the graphic.

emerging. In some special cases influent wastewater is controlled and could be completely shut off before the plant. Assuming normal plant operation this is not considered. Therefore, neither influent flow rates nor contaminant loads should ever be equal to (or smaller than) zero for any given month, day, or hour. Moreover, wastewater temperatures have to be above 0 °C, as freezing of the water would be the logical consequence otherwise. This implies that numerical values of zero can be considered false in all cases and for all parameters of interest.

Identifying upheld concentration values

The upholding of concentration values can lead to significantly misleading loads, especially when influent amounts change a lot, as explained in detail in Section 2.2.1. Concentrations are calculated according to Equation 1.1 based on known loads and influent flow rates. Identical consecutive values in the series of concentrations can be identified in a computer program, exposing upheld values. Loads based on upheld rather than measured concentrations are seen as meaningless. They are all treated as definite errors by the model without consideration of the realism of the resulting load or any other error identification criteria.

Handling identified errors

After identifying an error it is relevant to consider how it shall be treated. Similarly to the approach used in the data analysis, all numerical values identified as erroneous are eliminated in the first instance, leaving gaps in the data. Subsequently, these gaps shall be treated along with all other gaps that might be present using the same methods (see Section 3.1.1.2).

3.1.1.2 Completing fragmentary data

As daily data was shown to often present with gaps of different nature, the completion of these time series is a relevant step within the refinement procedure. A review of [27, 28, 34–36] as relevant literature shows that fragmentary data is a commonly encountered problem when handling large amounts of data and that within the variety of possible solutions, no single one is always superior to others. The suitability of different methods changes depending on the specific considered case in terms of the behaviour of the underlying parameter as well as the properties of the data missingness.
There is a number of popular stochastic models based on the Box-Jenkins' approach (named after its inventors). Depending on relevant properties of a time series such as stationarity or seasonality, models of types such as Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) or Seasonal Autoregressive Integrated Moving Average (SARIMA) can be utilized. These have been shown to effectively analyse patterns in time series and predict values outside of the given ranges in a forecasting manner but require sufficient amounts of non-fragmentary data for model learning (calibration of model parameters). [27, 35, 36]
Based on the preliminary data analysis in this work, this cannot generally be assured. Many of the gaps are courtesy of the measuring strategy chosen by the plant operator and the patterns are hence repeated constantly. A long series of consecutive values for model learning is therefore not given for many datasets. In this work a different approach is taken for filling the gaps in daily plant data, utilizing two main methods in separate steps: Regression and curve fitting.

Regression

In regression, the value of a response variable is modelled based on the correlation with one or more predictor or regressor variables [29, 36]. Simple linear regression describes the linear relationship between one predictor and one output variable. Since the data

analysis showed good linear correlation between certain parameters, this type of regression is offered to the model user for filling gaps in time series based on available values from another parameter. Temperature generally showed poor correlation with the other variables and is consequently excluded for this step. The strength of correlation among the different contaminant loads and influent flow rate can vary and regression should never be blindly trusted [29, 31]. An examination of the scatterplot is thus considered to judge the validity of assuming linear relationship between the loads. The decision whether regression shall be utilized for influent flow rate and loads, for loads only, or not all, is made individually by the model user for a specific case. As additional criteria, a minimum threshold for the Pearson correlation coefficient r, a minimum number of required datapoints for model fitting n_{min} as well as a significance level α for hypothesis testing are set individually. The value of r can be calculated according to:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{3.1}$$

where x and y denote regressor and response variable, n denotes the number of samples (where both x and y are defined) and the subscript i denotes consecutive numbers representing any one instance from 1 to n. The fitting of the regression line is done by the ordinary least squares method, where the sum of the squared residuals (i.e. the difference between function value and observation) is minimized [29, 36]. The null hypothesis in hypothesis testing is chosen to be:

> *"There is no significant linear relation between variable x and y."*

Like in the preliminary data analysis a t-test is used for hypothesis testing according to standard methods from literature [29]. If the corresponding p-value p (the probability of achieving results at least as extreme as in the data if the null hypothesis were true) is lower than the set significance level, the null hypothesis is rejected and the alternative hypothesis, stating a significant linear relationship between the two chosen variables, is accepted. Hence, the following equations express the requirements used for utilizing regression for the estimation of missing values:

$$r \geq r_{min} \tag{3.2}$$

$$p \leq \alpha \tag{3.3}$$

$$n \geq n_{min} \tag{3.4}$$

When values are missing for one parameter y, the other parameters are considered as regressors x in the order of descending values of r. Upon satisfaction of the determined criteria, the respective fitted regression model is used to estimate missing values of parameter y. Thereby gaps in the time series of y can be filled wherever x is defined in the original data obtained from the WWTP. Upon the existence of gaps in the time series of y after this, the next parameter is considered as regressor x accordingly. Repeating this procedure for all fragmentary data finally results in a new dataset, where none, some or all of the original data gaps are filled.

Curve fitting

The remaining gaps are treated in a different manner. Numerical analysis methods such as estimating functions to describe the data based on available measurements and subsequently using these to establish missing values present deterministic, simple to understand approaches for filling gaps in time series [35]. Fitting polynomials to large numbers of points can be problematic and should be done with care. Low degree polynomials often do not follow the data well enough and high degree polynomials can show excessive oscillation under certain conditions (referred to as Runge's phenomenon [37]). Different algorithms are available for polynomial curve fitting. Generally, the function which is fit to the data does not necessarily include the data points, depending on the degree of polynomial, number of points, and fitting algorithm chosen.
Cubic spline functions are piecewise polynomial functions of third degree that can be used to connect consecutive points. They produce smooth, continuous curves and can be used to describe complex data series while staying locally simple. Moreover, the function passes through the specified knots, which enables a conservation of measured data. These properties are deemed to make them suitable for the given purpose. For n data points, the spline $S(x)$ is [38]:

$$S(x) = \begin{cases} C_1(x), & x_0 \leq x \leq x1 \\ C_i(x), & x_{i-1} \leq x \leq x_i \\ C_n(x), & x_{n-1} \leq x \leq x_n \end{cases} \tag{3.5}$$

where C represents a cubic function of the variable x. It takes the form:

$$C_i(x) = \alpha_i x^3 + \beta_i x^2 + \gamma_i x + \delta \tag{3.6}$$

where α, β, γ and δ denote the polynomial coefficients. In order to solve for the $4n$ coefficients, an equal number of conditions is required. The requirement of the spline being exact at the data points delivers $2n$ conditions. Furthermore, adjacent cubic functions shall have matching first and second derivatives at the point of intersection which assures smoothness of the function and adds $2(n\text{-}1)$ conditions. Finally, specifying boundary conditions delivers the last two equations. The so called natural boundary condition defines a second derivative of zero at the end points x_0 and x_n. [38]
Among different alternatives, another end condition is the Lagrange condition, where the first derivative at the end points is equal to the first derivative of the cubic polynomial connecting the respective end point with its three nearest neighbours [39]. The splines are commonly named after their end conditions.
A problem with these types of splines is that over- and undershoots produced by the function can lead to unrealistic results such as negative values in some cases. Piecewise cubic hermite polynomials are uniquely specified by the function values and derivatives at the ends of the respective interval. If the derivatives are not given, they have to be estimated. A popular algorithm used for this is the Fritsch-Carlson method which preserves monotonicity from data [40]. The resulting piecewise cubic hermite interpolant (or hermite cubic spline) lacks the property of matching second derivatives at the points of intersection of adjacent polynomials and is hence generally not as smooth as a cubic spline satisfying the conditions above, but the fact that it is monotonicity preserving can make it superior in some cases.
Another simple monotonicity preserving approach for data interpolation is linear

interpolation, where the interpolants are straight lines drawn between each two neighbouring points.
Figure 3.2 depicts the different behaviour of a linear interpolant, lagrange cubic spline, a natural cubic spline, and a hermite cubic spline on a set of randomly created datapoints. It is seen that local maxima and minima in the data are kept as such by the hermite spline

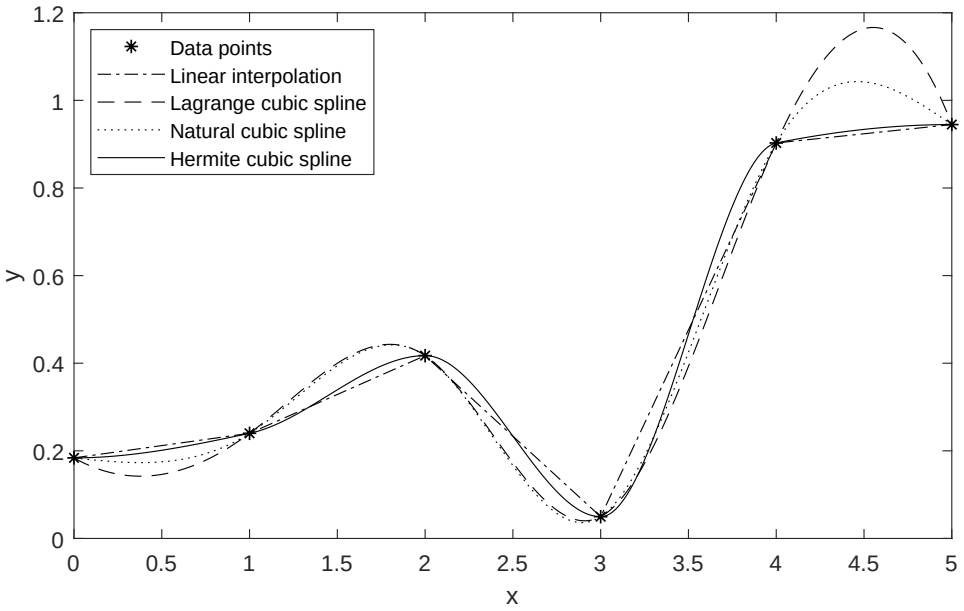


Figure 3.2. Comparison of the behaviour of different types of interpolation on a set of randomly created data.

and the linear interpolant, but not by the lagrange and natural spline. Which of the three solutions is the most realistic cannot be stated without knowledge about real underlying data. None of these interpolants is generally superior to the others in terms of producing the best results. While some might be more smooth and visually pleasing than others, it does not necessarily mean that the results are closer to reality. Linear interpolation and hermite cubic spline as monotonicity preserving as well as lagrange and natural spline as non monotonicity preserving methods are proposed as alternatives to each other and a choice shall be made individually by the model user.

Adding additional variation

The most advanced approaches for influent generation found in literature (see Section 1.3) include the addition of zero mean white noise to synthetically created data in order to reduce correlation between variables and introduce some randomness. This method shall be adapted here. The data created for a response variable in regression has a perfect correlation with the respective regressor and the curves produced by linear or spline interpolation methods follow the interpolant function perfectly, while the data in the preliminary analysis was described as experiencing some randomly looking fluctuations around a continuous baseline (see Section 2.2.5). Random numbers with a mean of zero and a defined standard deviation σ can be created in computer programs. In order to make the series of random numbers reproducible (i.e. the same numbers will always be created for the same standard deviation), the seed used for the (pseudo-)random number

generator is specified to a fixed number. Even though a non-deterministic solution is not in focus of this work, the option to randomize the seed for the number generation is implemented and can later be chosen upon desire. The created random numbers are used as a correction factor, numerically expressing the deviation relative to the computed value for any given day. Overlay of the synthetically created data with this factor leads to a new daily value. The values from the original dataset are not altered. The user can manually replace the random numbers, which might be desirable for example if he has additional information about certain states or if he wants to include more extreme peaks [1]. The final output of this procedure is a complete set of values on a daily basis.

3.1.2 Creating daily data from period averages

Often data is not documented at the plants on a daily basis but as averages over longer periods. It has been shown in this work as well as in literature, that dynamics occur seasonally, on a day to day basis, and as characteristic diurnal patterns. As monthly values are the most coarse resolution of input data considered, seasonal variation is already included in the data. Period averages are refined to a complete set of daily data firstly, utilizing Algorithm B, which is shown in Figure 3.3. The methods are described in detail in the respectively indicated parts of the text.

3.1.2.1 Integral preserving interpolation

The statements made by the original data shall be preserved, i.e. the mean of all daily values created synthetically for a certain period have to equal the originally recorded value. This is the same as to say that the integral over the respective period has to be preserved. A method found in literature uses sampling from a normal distribution with the recorded average as the mean which satisfies this demand [15]. However, as shown in Section 1.3, this produces clearly visible jumps between the respective periods. Modern mathematical computer software usually include different interpolation algorithms, such as polynomial, spline or linear interpolation (discussed in Section 3.1.1.2), but no algorithm enabling a specification of the integral or average value could be found. Therefore a workaround including integration, interpolation of the integral and subsequent consideration of the derivative of the calculated integrand is utilized.

The value of the integral of z, the variable of interest at the end of a given period p is calculated as:

$$\int_{p} z = \bar{z}_p \cdot t_p + \int_{p-1} z \tag{3.7}$$

where t is the duration of the respective period and $p-1$ denotes the preceding period. For the first period of consideration, the second term of this sum is always zero, i.e. the integral at zero is set to zero. The values of the integral at the ends of all periods are monotonically increasing. Interpolation is then performed on the integral. The average

[1] Consider the following example: Fragmentary daily data is available as obtained from the plant. Additionally, when transmitting the data, the plant operator mentions that an industrial client had to discharge a large amount of highly impurified water due to an emergency over three days in July. Influent concentrations were not measured in the relevant period. The person conducting the simulations might want to include additional peaks in the respective period in his data depending on the purpose of the simulation.

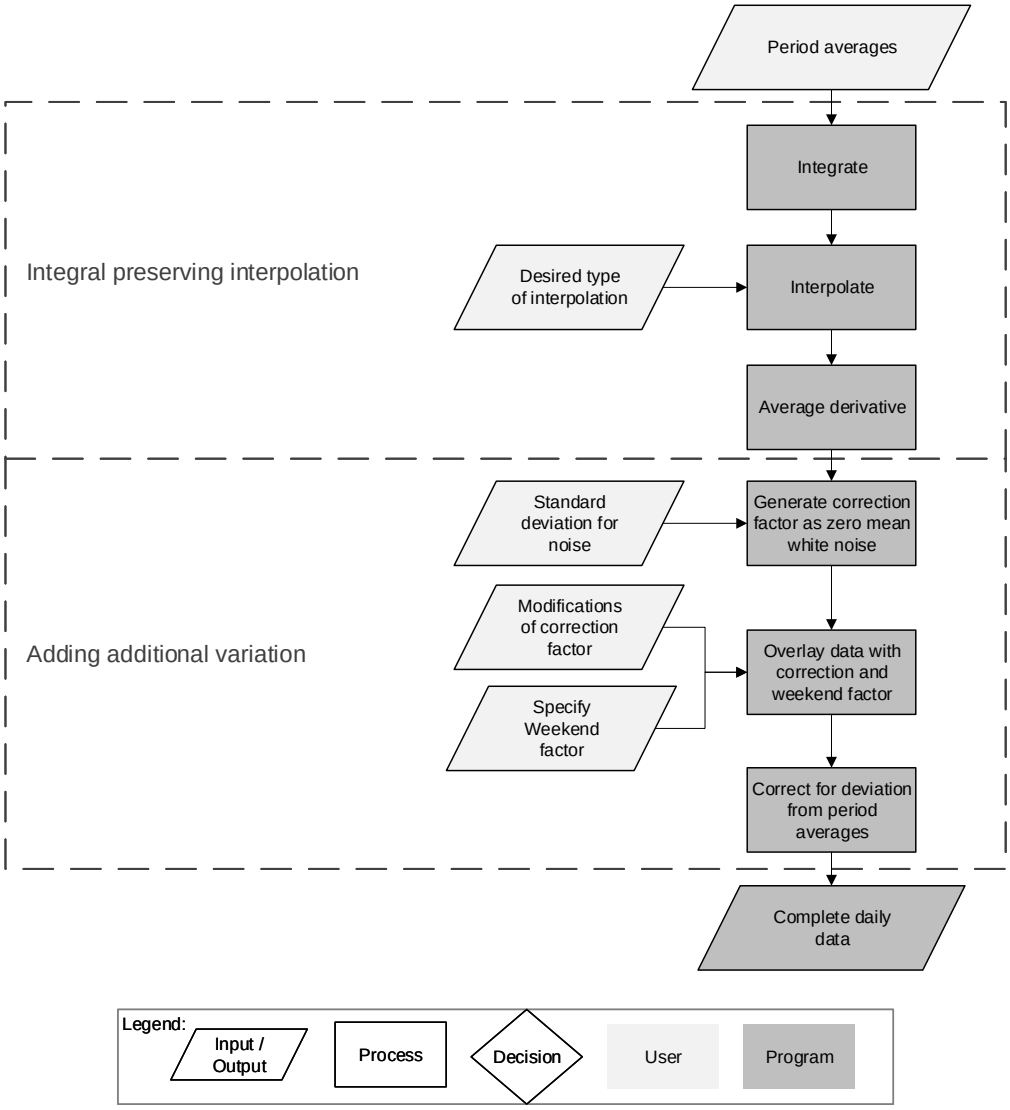


Figure 3.3. Algorithm B: Creating a complete set of daily data from period averages. A detailed description of the procedures within the dashed boxes is found in the respective text section indicated on the left side of the graphic.

derivative of the interpolant for any desired time is the average of the process variable z for that time. It is obtained by computing for the slope of line connecting the function values of the interpolant at the beginning and end of the desired time interval (here: at the beginning and end of a day). Linear interpolation would lead to a constant derivative and is hence unsuitable. Cubic splines with defined natural or lagrange end conditions have matching second derivatives at the points of intersection of the individual polynomials. This means that the curve for the variable of interest, obtained as the derivative of the interpolant, has a continuous derivative itself, leading to a smoothly looking function. For cubic hermite splines second derivatives are not specified. While the derivative is still a continuous function, the slope thereof can be subject to sudden changes, leading to a less smoothly looking function (which does not necessarily mean that it is less realistic). However, its monotonicity preserving property is highly desirable. If the monotonicity of the integral is not preserved by the interpolant, unrealistic negative values in the derivative (i.e. the curve for the variable of interest) are the logical consequence. The different splines

are offered as alternatives and have to be chosen for each specific case.

3.1.2.2 Adding additional variation

Random variation is introduced into the smoothly looking curves for more realistic patterns. Zero mean random numbers with a defined standard deviation σ are created. They are then used as a correction factor describing the relative deviation from the computed value. With a finite count of random numbers being created, the mean is not exactly zero for any given period which means that the averages are thereby altered. Additionally, a weekend factor is introduced as a tunable parameter. It is used to adapt the computed value by the defined relative amount so that changing wastewater production on the weekends can be accounted for. By summing up the absolute changes in the daily values introduced by this step, the alteration of the period average $\bar{z}'_p$ can be calculated:

$$\bar{z}'_p = \sum_{i=1}^{n_p} Rel_{d,i} \cdot \bar{z}_{d,interpol,i} \tag{3.8}$$

where Rel_d is the relative deviation from the previously interpolated daily mean $\bar{z}_{d,interpol}$ and the subscript i represents each instance from 1 to n_p, the number of days in the respective period. Rel_d is the sum of the randomly created correction factor and the weekend factor.
In order to keep the period mean to the original value, the additive inverse of $\bar{z}'_p$ is distributed evenly to all days in a respective period and added, which leads to a final daily mean value $\bar{z}_d$ of:

$$\bar{z}_d = \bar{z}_{d,interpol} + Rel_d \cdot \bar{z}_{d,interpol} + \frac{-\bar{z}'_p}{n_p} \tag{3.9}$$

To keep the approach deterministic, the seed used for the (pseudo-) random number creation is kept to a fixed value. If the values shall be modified, e.g. to include known rain peaks or to incorporate first flush events, the random number representing the relative deviation to the mean obtained from the interpolation procedure can simply be replaced manually upon desire, as the subsequent steps assure a preservation of the given means.

3.1.3 Creating hourly data from a complete set of daily values

Daily values obtained as the outputs of Algorithms A or B are finally refined to achieve the desired final output of hourly data. Methods in literature have proposed the overlay of daily averages with diurnal patterns but unwanted artifacts such as discontinuities between adjacent days or alteration of daily averages are observed (see Section 1.3) [21, 22]. Therefore Algorithm C is designed for creating hourly data from daily inputs including the incorporation of a diurnal profile as well as additional variation while utilizing additional methods for preservation given averages as well as avoiding unwanted discontinuities. The algorithm is outlined in the flow chart in Figure 3.4 and a detailed description is found in the respectively indicated sections of the following text.

3.1.3.1 Integral preserving interpolation

In order to perform integral preserving interpolation, successive integration, spline interpolation and computation of the average derivative is utilized. This is done in

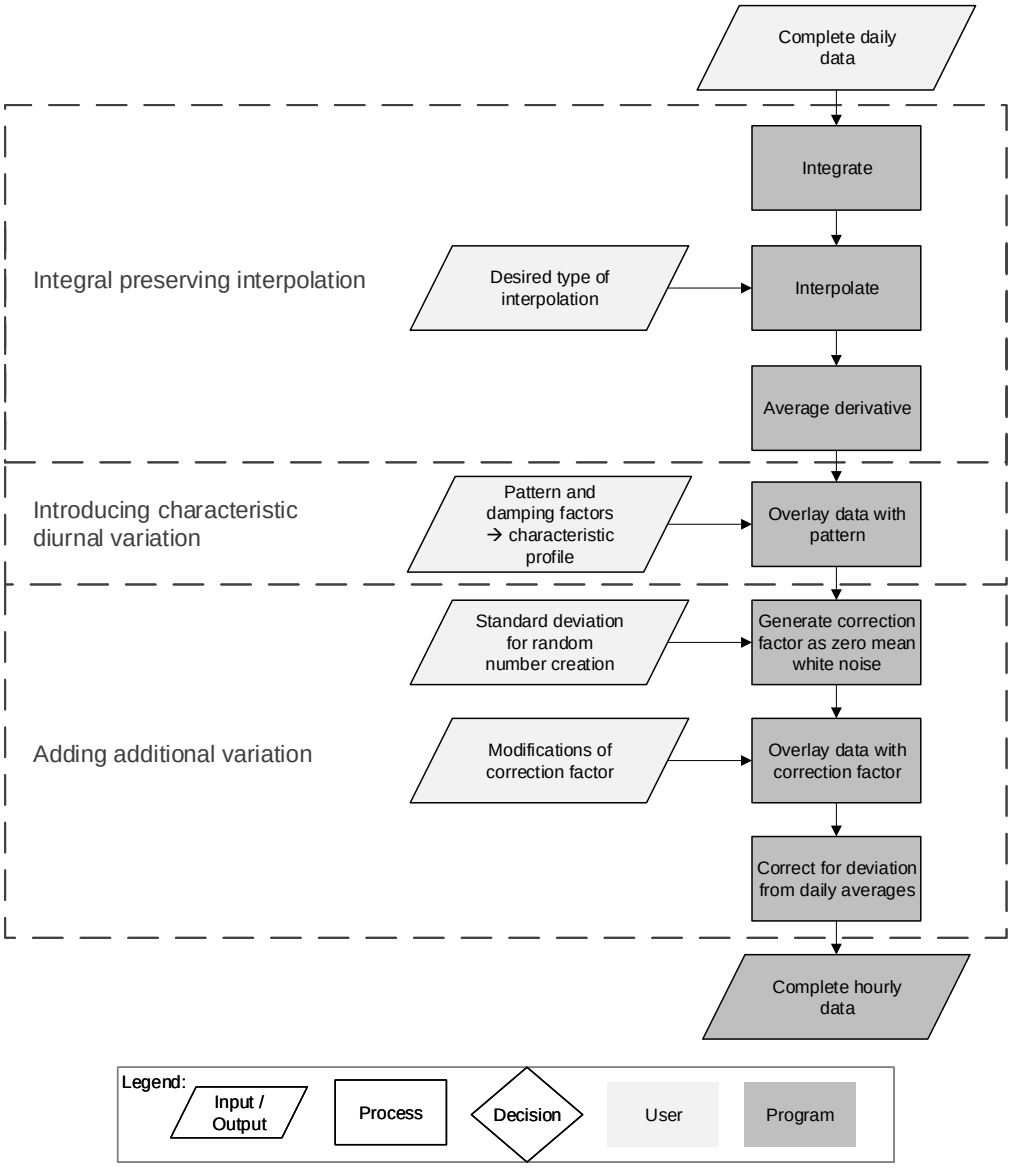


Figure 3.4. Algorithm C: Creating hourly data from a complete set of daily values. A detailed description of the procedures within the dashed boxes is found in the respective text section indicated on the left side of the graphic.

the same manner as when interpolating from period averages to daily averages (see Section 3.1.2) results in a continuous set of values on an hourly basis that maintains the daily average [2]. This is used as the base for the incorporation of a characteristic diurnal variation.

3.1.3.2 Introducing characteristic diurnal variation

Due to the habits of dischargers connected to a WWTP, a characteristic pattern in wastewater quantity and quality is commonly seen to reoccur on a day to day basis. Diurnal patterns are used in the most sophisticated data refining methods found in literature and have been shown in the preliminary data analysis. With households often

[2]Note that the period inputs from Section 3.1.2 are replaced by daily inputs and the desired output is now hourly instead of daily.

being the main source of wastewater, the general shape of these variations is expected to be similar for many WWTPs within regions of similar cultural habits. Moreover, the similarity of flow and load patterns has been shown in literature [20]. However, the patterns are deemed to generally vary between plants in two ways:

- Due to the effects of differently sized sewer systems, the amplitude of the variations changes for plants with similar types of clients but different capacities while the general diurnal progress stays the same.
- As the types of dischargers connected to plants can vary, large industrial wastewater producers can significantly influence incoming water and pollutant amounts and the common daily practices in the population can change depending on the culture. These factors can alter the general appearance of the patterns.

Patterns are stored as a series of 24 values with a mean of zero, representing the relative characteristic deviation from the daily mean for the hours of a day. They are created with the aid of a set of hourly values over the course of at least one full day. Calculation of the arithmetic mean, dividing each hourly value by this mean and subtracting 100 % from the result produces the value to be stored in the pattern for the respective hour. If hourly data is available for more than one day (which is desirable to increase the reliability of the resulting pattern), the averages of all values for the respective hour is utilized. If hourly values over the course of one day cannot easily be obtained for a plant of interest, standard patterns can be reused and adapted as seen most appropriate. For this purpose, some standard patterns for municipal wastewater are created from the dynamics found in the influent flow rate in [21] and the dry weather scenario in the BSM1 [12] in the above stated manner. The patterns can be adapted through division by a tunable damping factor to account for different plant sizes. This adaption and repetition of the pattern for all days of interest results in the characteristic diurnal profile for a specific parameter of a plant. The previously interpolated values are overlaid with the profile.

3.1.3.3 Adding additional variation

Finally, some additional variation is introduced, similarly to how it was done for daily values in Section 3.1.2.2. This is to introduce some randomness and deviations from the daily characteristics. A pseudo-random number generator is used with a defined seed to create a reproducible set of numbers for zero mean white noise with tunable standard deviation σ. These numbers can manually be replaced upon desire in order to account for storm occurrences with high intensity rainfall lasting less than a whole day. Therefore, $\bar{z}'_d$, the alterations to defined daily averages have to be accounted for. They are calculated as:

$$\bar{z}'_d = \sum_{i=1}^{24} Rel_{h,i} \cdot \bar{z}_{h,interpol,i} \tag{3.10}$$

where Rel_h is the relative deviation from the previously interpolated hourly mean $\bar{z}_{h,interpol}$ represented by the randomly created (or manually modified) correction factor and the subscript i represents each instance from 1 to 24, the hours of each day.
In analogy to Section 3.1.2.2, the additive inverse of $\bar{z}'_d$ is distributed evenly to the hours of the respective day to restore the correct daily averages. The final value for a parameter

$\bar{z}_h$ can be expressed as:

$$\bar{z}_h = \bar{z}_{h,interpol} \cdot (1 + profile + Rel_h) + \frac{-\bar{z}'_d}{24} \tag{3.11}$$

Where $profile$ refers to the value of the characteristic profile of the specific plant and parameter applicable for the respective hour.

3.1.4 Model implementation

Computer software is used for the practical implementation of the defined methods. The main platform is chosen to be *Microsoft Excel 2016*. It comes with a wide variety of pre-defined functions for different types of applications. User-defined functions can be included through code written in *Microsoft Visual Basic for Applications 7.1* (*VBA*), which is built into the program. *Microsoft Excel* uses cells to store text or numerical values. A cell value can be obtained by either manually entering it into the cell, by calling combinations of pre-defined and user-defined functions in the cell, or by macros written in *VBA* code. This way, data can be input, processed and output in Excel. Input fields are marked accordingly, while fields that contain calculations and hence display outputs should not be changed by the user and are hence proteted against modification.

Two separate model files are created for the two situations of inputs given as daily or as period averages. Several charts, such as scatterplots and correlation matrices for regression analysis and charts showing step by step development of the data by the successive methods are included and updated with each calculation. This is to aid the user in decision making, tracing possible problems, as well as checking credibility. Additional plausibility checks are implemented for monitoring purposes.
Whenever a cell value is changed, all other cells depending on it are automatically recalculated using the standard settings in *Microsoft Excel 2016*. This is not desired here, as it can slow down the handling of the model due to large numbers of calculations. Therefore, *VBA* code is embedded to automatically change these settings when opening the respective files, so recalculation is only done upon user demand. This is reversed and standard settings are restored when the files are closed.
For the conduction of regression analysis as well as for the different types of spline interpolation programs written in *MathWorks MATLAB 2017b* are utilized. It is superior in terms of speed and ease of use for these applications. To minimize effort and complexity for the user, the programs are deployed as Excel add-ins with the aid of the *Library Compiler* app included in *MathWorks MATLAB Compiler*. This necessitates the installation of the freely available *MathWorks MATLAB Runtime 9.3*, but the user does not need to have MATLAB installed. The function can then be used like a regular Excel function by calling it from a cell.
For the generation of random noise, *VBA*'s pseudo-random number generator is utilized. Macros are implemented for automatically creating the needed amount of random numbers with the desired seed and writing it to the dedicated cells in the Excel worksheet. The macros can be called from buttons implemented in the worksheet, representing the user interface of the model. The seeds used for the random number generator are -1, -2, -3, -4 and -5 for influent flow rate, loads of COD, TN and TP as well as temperature respectively. This is to ensure that the created random numbers differ among the parameters.

3.1.5 Model demonstration

In order to demonstrate the changes induced in the data by the created model, it is applied to available real life data. The three main algorithms are examined with two different datasets used as a starting point. Values of tunable parameters merely serve model presentation. Here, they are chosen in a range that is deemed to give somewhat realistic results using expert knowledge from the research project rather than being tuned towards a specific plant. The calibration of these parameters in the future poses the main challenge in the refinement for the specific model user and is based on his understanding of wastewater generation mechanisms generally and specifically for the respective plants as well as additional information he might obtain. Further discussion of model calibration by a future user as well as the expected sensibility to these parameters is found along with the discussion of the results in Section 3.2.

3.1.5.1 Algorithm A

Algorithm A is applied to a set of daily data as obtained from the WWTP Tramin. Relevant inputs are specified in Table 3.1. The calculated white noise is used as such as the correction factor and no values are manually adjusted.

Table 3.1. Inputs for refinement model demonstration: Algorithm A.

Dataset	Tramin 2016
Regression	
Regression analysis considered?	yes
Include Q in regression analysis ?	no
r_{min}	0.6
α	1 %
n_{min}	20
Curve fitting	
Type of interpolation for Q	hermite spline
Type of interpolation for $\dot{m}_{COD}$	hermite spline
Type of interpolation for $\dot{m}_{TN}$	hermite spline
Type of interpolation for $\dot{m}_{TP}$	hermite spline
Type of interpolation for T	hermite spline
Additional variation	
σ of white noise for Q	10 %
σ of white noise for $\dot{m}_{COD}$	17 %
σ of white noise for $\dot{m}_{TN}$	15 %
σ of white noise for $\dot{m}_{TP}$	20 %
σ of white noise for T	5 %

3.1.5.2 Algorithm B

For an illustration of the inner workings of Algorithm B, period averages are converted to daily values. Used model inputs are listed in Table 3.2. Ten additional peaks shall be implemented in the daily data for influent flow rates and loads. The computed random numbers are replaced by +50 % and +100 % on ten randomly chosen days throughout the year (+50 % on days 68, 127, 154, 230, 268 and +100 % on days 25, 69, 234, 243, 281).

Table 3.2. Inputs for refinement model demonstration: Algorithm B.

Dataset	Zirl 2015
Type of interpolation	hermite spline
Additional variation	
σ of white noise for Q	10%
σ of white noise for $\dot{m}_{COD}$	7%
σ of white noise for $\dot{m}_{TN}$	6%
σ of white noise for $\dot{m}_{TP}$	8%
σ of white noise for T	5%
Weekend factor Q	-5%
Weekend factor $\dot{m}_{COD}$	-5%
Weekend factor $\dot{m}_{TN}$	-5%
Weekend factor $\dot{m}_{TP}$	-5%
Weekend factor T	-2%

3.1.5.3 Algorithm C

The changes in data induced by Algorithm C are examined utilizing the output of Algorithm B from the previous section. For one of the days where rain peaks were added

Table 3.3. Inputs for refinement demonstration: Algorithm C.

Type of interpolation	hermite spline
Diurnal variation	
Used pattern	Standard pattern created from [21]
Damping factor Q	3.5
Damping factor $\dot{m}_{COD}$	1.5
Damping factor $\dot{m}_{TN}$	2
Damping factor $\dot{m}_{TP}$	2
Damping factor T	4
Additional variation	
σ of white noise for Q	2%
σ of white noise for $\dot{m}_{COD}$	2%
σ of white noise for $\dot{m}_{TN}$	3%
σ of white noise for $\dot{m}_{TP}$	4%
σ of white noise for T	2%

in the application of Algorithm C, a storm event shall be realized. The peak in the diurnal variation of the influent flow rate is included in the early evening by manually adjusting the value of the correction factor to 10 %, 30 %, 160 %, 60 % and 20 % for hours 16 to 20 on day 127. To include a first flush event triggered by the storm, peaks are also induced into the loads by adjusting the correction factor to 20 %, 40 %, 200 %, 80 % and 40 % for the respective time.

3.2 Results and discussion

This section is separated into two parts. The first part discusses the model algorithms and their application based on experiences from the development and understanding of the used methods. It shows some of expected possibilities and limitations of the model. In the second part, the results of an exemplary case are shown. They give insight into the inner workings of the model and the stepwise development of the data. Moreover, they are used to examine whether realistic data can be created as well as to investigate and back some of the statements drawn up in the first part.

3.2.1 Algorithms and model files

The combination of Algorithms A, B and C into final model algorithms enables data refinement for purposes of dynamic simulation in a systematic manner for an arbitrary plant in the *ICAWER* project.

Two final model algorithms can be differentiated for the distinct situations of plant data obtained on a daily timescale, possibly including gaps or errors of the identified types, or plant data obtained as period averages. The flow chart in Figure 3.5 describes these final model algorithms. The difference between the two scenarios lies in the steps necessary for

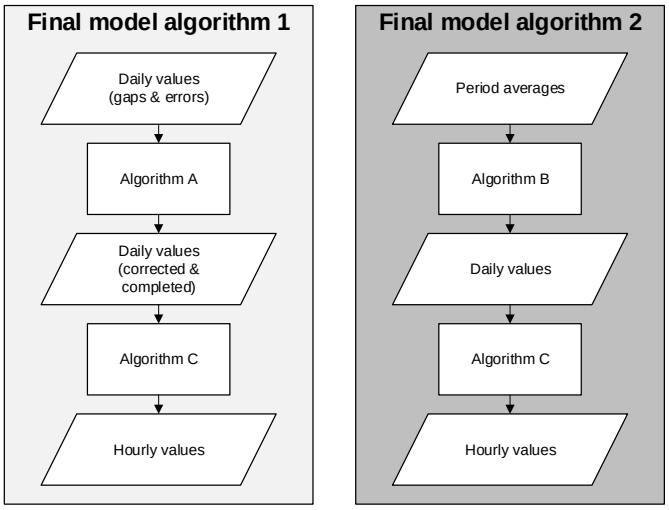


Figure 3.5. Final refinement model algorithms for the two distinct cases of input data.

obtaining a complete set of daily data. Application of Algorithm A to daily data resolves the identified types of data errors and fills gaps in the time series, while Algorithm B deals with the problems associated with the refinement of monthly values to daily values. From there on, Algorithm C is applied and treatment is the same in both cases.

As indicated in Figures 3.1, 3.3 and 3.4, apart from the obtained plant data, further inputs such as tunable parameters utilized throughout the different methods are relevant. Adequate values for these parameters cannot be set universally but depend on the specific case. Based upon his own judgement it is up to the experienced user to choose whether certain features shall be implemented or not. The more specific knowledge is obtained about prevalent conditions relevant for plant operation, the more accurately tunable parameters used throughout the different procedures can be determined and the more

representative induced dynamics can become for a specific case. As additional information might take different forms such as verbal statements by plant operators or indications from recorded climate data, no guideline for adequate parameter calibration is stated here. Nevertheless, the model results are expected to be very sensitive to the choice of these parameters and an experienced and skilled user is seen crucial.

The data analysis showed similar appearance of diurnal patterns throughout the different variables and literature indicates that the characteristics of municipal wastewater quantity and quality are somewhat similar among different areas [20, 41, 42]. Therefore, the utilization of standard patterns included in the tool after adaption with damping factors is deemed reasonable for plants with a majority of municipal wastewater. For industrial wastewater however, the diurnal variation can be significantly altered and should be adapted individually.

One has to be aware that the data errors corrected by the model only include those types that were identified in the preliminary analysis of various plant data. While more values in prevalent data might be flawed to different degrees and for different underlying reasons, only errors that can be clearly identified as such are corrected. It is difficult to make a definite, generalized statement about the accuracy of the methods used for treating fragmentary data, as the true behaviour during times with missing data is simply unknown. The same goes for the procedures used for the interpolation to smaller timescales. Nevertheless, it is important to consider that the goal of this work is not to create an exact replica of real life occurrences for a certain state in the past. It is rather to maximally utilize the knowledge obtained from plant data and refine it in a way to make it suitable for simulation. This includes making it more realistic by synthetically implementing dynamics resulting from relevant phenomena that are suspected to occur in real life.

First flush events triggered by storms are not modelled directly, which could present a weakness of the chosen approach. However, this is based on a conscious decision which was made considering the expected added value and complexity. The phenomenon could be modelled and included using tunable parameters. Its behaviour is highly dependent on each individual sewer system among other factors and accurate calibration would require detailed individually relevant data, which is usually not available. With an increasing number of implemented features, the complexity increases significantly and without accurate determination of adequate values for required variables, the added value is deemed limited. If desired, the user can still manually implement first flush events by modification of the initially randomly created correction factors found in Algorithms A, B and C at his own discretion.

Inadequate combination of tunable parameters and chosen methods for the individual steps can lead to unrealistic values at different steps in the process. Moreover, a similar problem can occur in the gap filling process under certain circumstances. The MATLAB function utilized for the curve fitting process allows the user to not only compute for values between known time series data (interpolation) but also values that lie outside of that range (extrapolation). In the latter case the polynomial determined for the adjacent interval is extended into the relevant range. When daily values are not given at the very beginning and end of the year, this extrapolation can lead to unrealistic values outside of

the range of available input data in certain cases. It is therefore essential for the user to monitor the process steps, judge the credibility with his expert knowledge and adjust the chosen inputs if necessary.

The created algorithms provide possible solutions for resolving the identified data issues for two different cases of input data. As the methods for the individual steps were chosen so that statements made by the originally given values are always preserved (except if the values are identified as erroneous), the same is expected for the final procedure and will be examined for the specific example in the following section. The application of these algorithms is not specifically limited to the research project, but could be used in other applications where the situation in terms of available data and desired output is similar. Upon adaption of the utilized parameters, this is likely not limited to the field of WWTP simulation but possibly other fields with similar problems. The resulting Excel model files along with the created add-ins based on MATLAB functions are specifically tailored for the prevalent situation and are seen as toolboxes for efficiently refining WWTP data for dynamic simulation input in the future. They can be found in attachment to the report.

3.2.2 Model demonstration

3.2.2.1 Algorithm A

For daily inputs, the changes in the data induced by Algorithm A are examined, essentially consisting of error elimination and the filling of data gaps by the specified means. COD load and influent temperature are shown here as representative process parameters, focussing on a three month period from August to September. Figure 3.6 displays a comparison of the datasets of $\dot{m}_{COD}$ and T that are obtained from the plant with the model outputs after the initial step of error elimination for a three month period. It

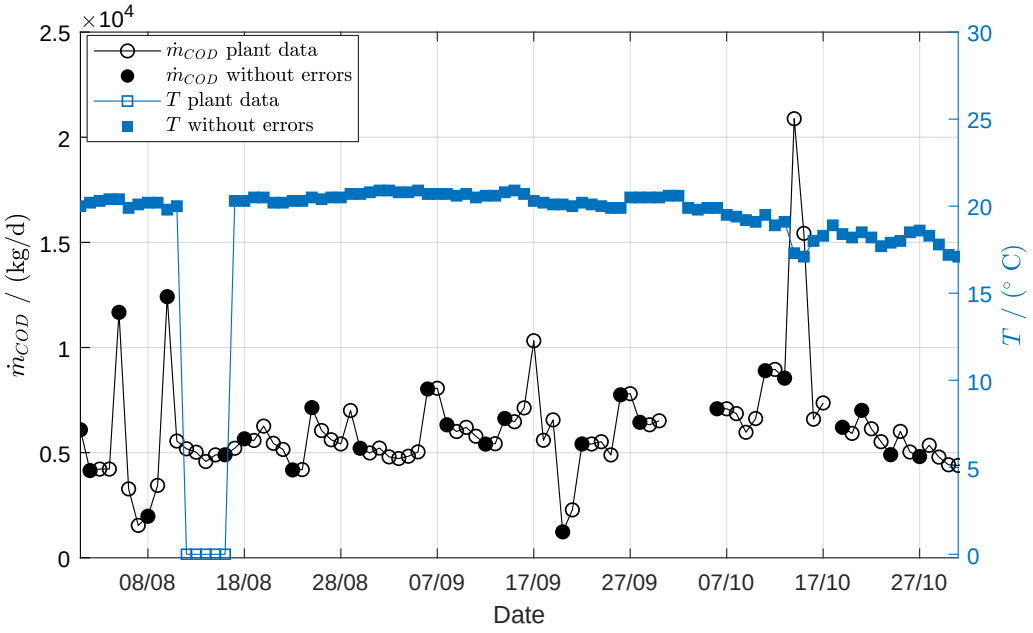


Figure 3.6. Error elimination in daily data of COD load and influent temperature.

shows that among many other values, the largest peak in organic load occurring in the

observed period was a result of upheld concentration measurements. All loads based on an upheld concentration value rather than a measured one are automatically excluded from the dataset. This expresses a consequent removal of what in this context is seen as a systematic fault in the documentation strategy. The model also deletes all zero values identified in the time series of the influent temperature.
The subsequent treatment differs slightly among the different parameters. The specified model inputs demand for regression to be utilized only among the different loads when the specified criteria are met. The prevalent plant data however mostly contains gaps that are univariate within these parameters. Thus, within the three month period only one missing datapoint (on 19/10) for the COD load is determined through regression. This can be observed in Figure 3.7, showing the stepwise development of the final daily values for the COD load from fragmentary data. Intermediate model outputs indicate

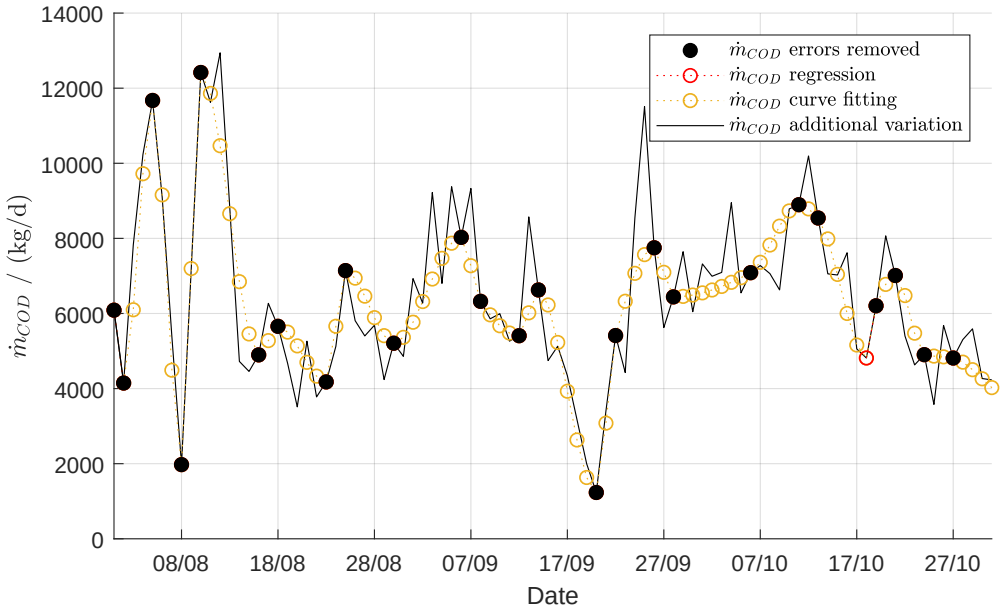


Figure 3.7. Stepwise development of final daily values from fragmentary data about the COD load.

that the value is computed based on the TN load. The value of r for these parameters is 0.61. The amount of data which can be obtained through regression is highly dependent on the specific plant data and the specified criteria. T is generally excluded from this step of the refinement process, so no data changes are observed here. Subsequently, all parameters are treated similarly. The stepwise development of the temperature data plotted in Figure 3.8 shows that the gaps arising from the previously eliminated zero values are filled based on the chosen curve fitting method. The remaining gaps in the COD load are filled in a like manner. The data here shows the monotonicity preserving property of the hermite spline nicely, as it does not produce any over- and undershoots. As a last step, the implementation of the additional variation contributes significantly to making the synthetically created data look more realistic. The final established values for T blend in with the rest of the data and the COD load shows a more naturally appearing fluctuation compared to the curve fitting result. With respect to the possible problem arising through curve fitting for the necessary extrapolation of values outside the given

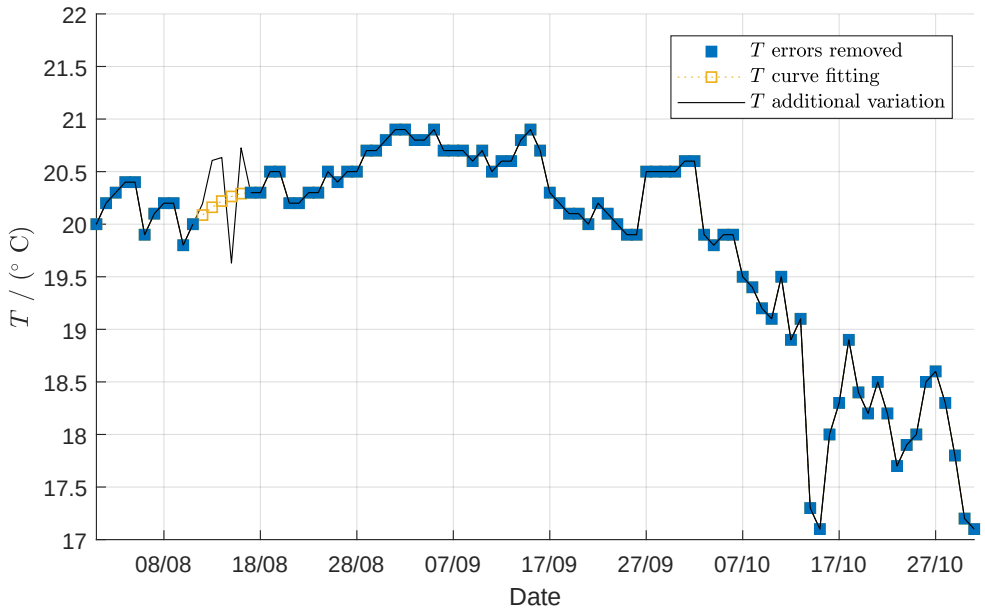


Figure 3.8. Stepwise development of final daily values from fragmentary data about the influent temperature.

time series mentioned in the previous section, it is seen that input data is given on both the first and final days of the year and hence no unrealistic values appear. Except for the eliminated errors all original datapoints are seen to be left unchanged by the refinement methods and are hence included in the final daily data. As indicated in Table 3.4, a change in annual average values is induced as slight increases or decreases are observed for all parameters. After checking for credibility by the experienced model user, the data is ready for further processing as specified by the final model algorithms.

Table 3.4. Changes in annual averages induced by Algorithm A.

Parameter Unit	$\bar{Q}_a$ m³/d	$\bar{\dot{m}}_{COD,a}$ kg/d	$\bar{\dot{m}}_{TN,a}$ kg/d	$\bar{\dot{m}}_{TP,a}$ kg/d	$\bar{T}_a$ °C
Original data	7931.27	5397.27	436.26	61.56	15.32
Errors removed	7933.91	5284.01	437.56	62.60	15.53
Regression	7933.91	5291.84	438.11	62.60	15.53
Curve fitting	7945.73	5313.31	432.01	62.59	15.60
Additional variation	7942.28	5402.08	431.57	63.66	15.60

3.2.2.2 Algorithm B

Merely Q and $\dot{m}_{COD}$ are selected for visual presentation here, but the shown results can be seen representative for all parameters.
The plots in Figures 3.9 and 3.10 indicate the stepwise development of daily data from the monthly average inputs within Algorithm B. The chosen hermite splines assure the avoidance of unrealistic negative values for both parameters. The interpolant of the integral is visually smooth, while the resulting derivative curve can experience sudden changes in its slope (seen e.g. between May and June in Figure 3.9). Random numbers

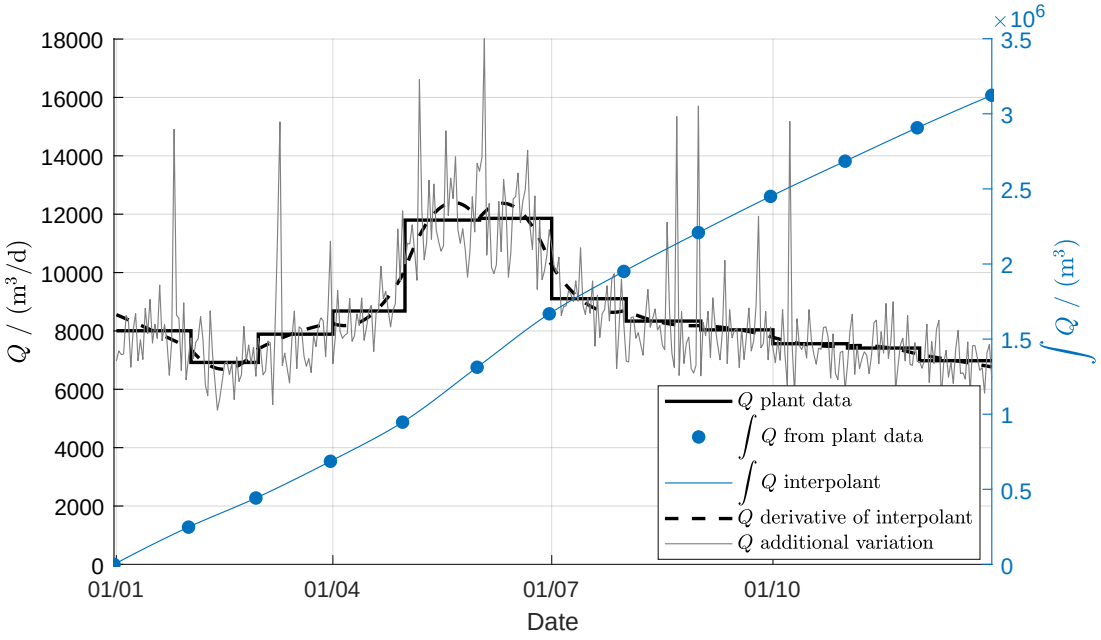


Figure 3.9. Stepwise development of daily values of influent flow rate from period averages as inputs.

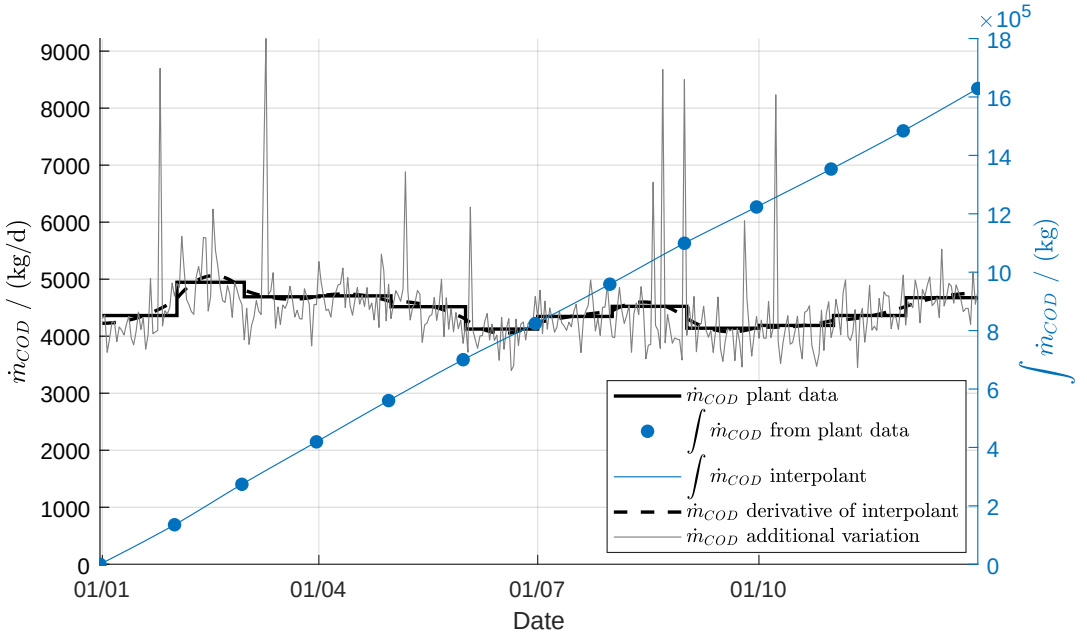


Figure 3.10. Stepwise development of daily values of COD load from period averages as inputs.

created for the correction factor create fluctuations similar to those seen in the data analysis. Weekends are not clearly identifiable at first glance, as the chosen weekend reduction factor is smaller than the majority of the induced random fluctuations. The peaks added by manual replacement of the created random numbers for correction factors are distinctly visible in the final data. Auxiliary calculations implemented in the model compute for the annual average based on the original as well as the final data. This unveils that the annual averages of 8556.82 m^3/d and 4463.00 kg/d for Q and $\dot{m}_{COD}$ respectively are indeed preserved throughout the process. Moreover, average values of all individual periods are successfully preserved for all parameters.

3.2.2.3 Algorithm C

After obtaining a complete set of daily averages as shown in the previous section, these are subsequently converted to hourly values. This unveils data changes made by the steps of computation which are summarized in the final model algorithm under Algorithm C. The same algorithm is also used subsequently to Algorithm A when using daily input data, which is why the here shown results are relevant for both considered scenarios of utilized input data. Figure 3.11 shows how the data is developed in a stepwise manner for a selected subset of twelve days using Q as a representative parameter. The creation and usage of

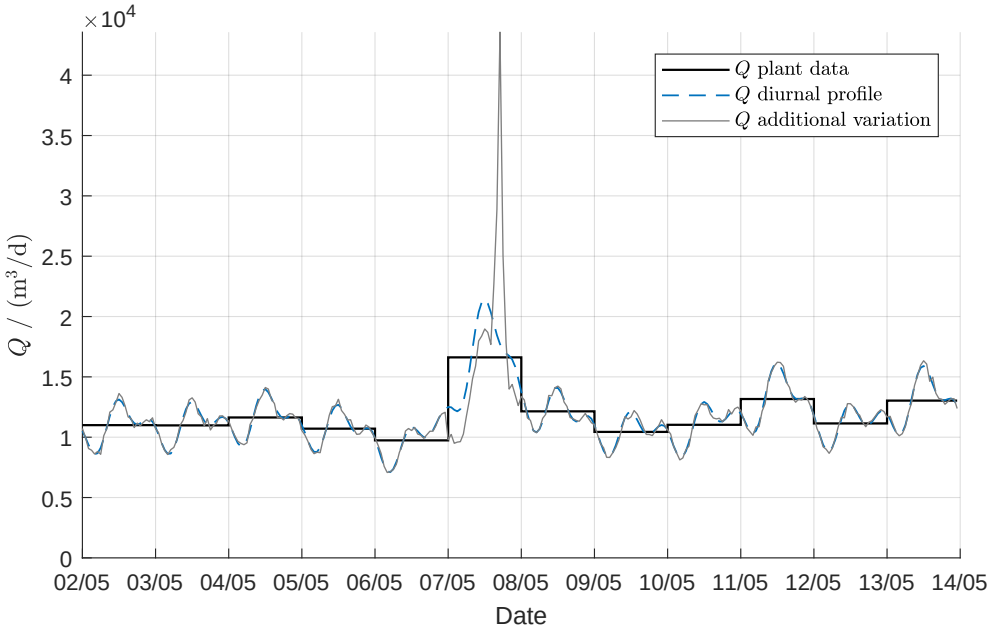


Figure 3.11. Stepwise development of hourly values from daily averages.

the integral is similar Algorithm B and not shown in the graphic. The absolute variation induced by the characteristic profile is seen to vary with the daily average value. It does make sense for this variation to scale with daily averages induced by private or industrial wastewater producers, as these are the main contributor to the emergence of reappearing profiles. If a daily value is significantly altered by phenomena such as surface runoff and first flush events, the variation induced by dischargers in reality would be influenced by this. For these cases scaling diurnal changes with less volatile parameters such as monthly averages or defining a characteristic absolute variation rather than a relative one could possibly give a more realistic result. However, the underlying root causes to varying daily

values are likely manifold in most cases and cannot be identified merely from recorded or synthetically created data. Investigation of the best solution as an overall compromise or implementation of a differentiation between changes in daily values originating from various root causes could possibly help in improving the solution in the future.
The observed profile can be generally stated as realistic for a generic plant with a majority of wastewater from households in the cultural area of Western Europe, but additional knowledge about dischargers and their behaviour or highly resolved measurements would be necessary to know whether the profile adequately represents this exact plant.

Figure 3.12 shows the final hourly data for influent flow rate and contaminant loads over a twelve day period. Though the same characteristic pattern is used, the different damping

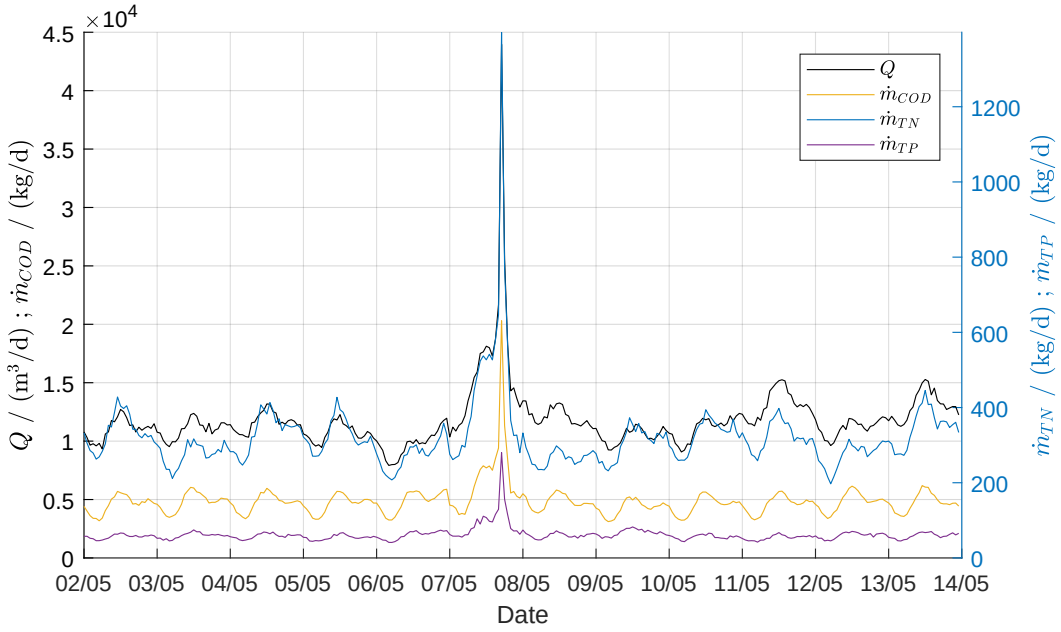


Figure 3.12. Hourly values for volumetric influent flow rate and contaminant loads as final model output.

factors create variations of different intensity for the respective parameters rather than just assuming congruent variational profiles. The irregularities induced by adding white noise as the correction factor once again aid in generating more variational, realistically appearing data. The manual adjustment of the correction factors is seen to successfully create a storm peak in the influent flow rate as well as a first flush event indicated by the peaks in the contaminant loads.
Contemplating the model results for the temperature, shown for the respective period in Figure 3.13, the temperature is unaffected by the first flush event, as the correction factor was left unchanged. Since the damping factor used for the daily pattern is the largest for T, the relative daily variation is the lowest and the diurnal profile is disturbed by the noise the most.

The resulting contaminant concentrations shown in Figure 3.14 vary in a similar way as was observed for c_{NH4-N} in high quality industrial scale data within the data analysis. However, one should be aware that this is a result of the chosen diurnal patterns and damping factors. Among others, these are important aspects influencing the appearance

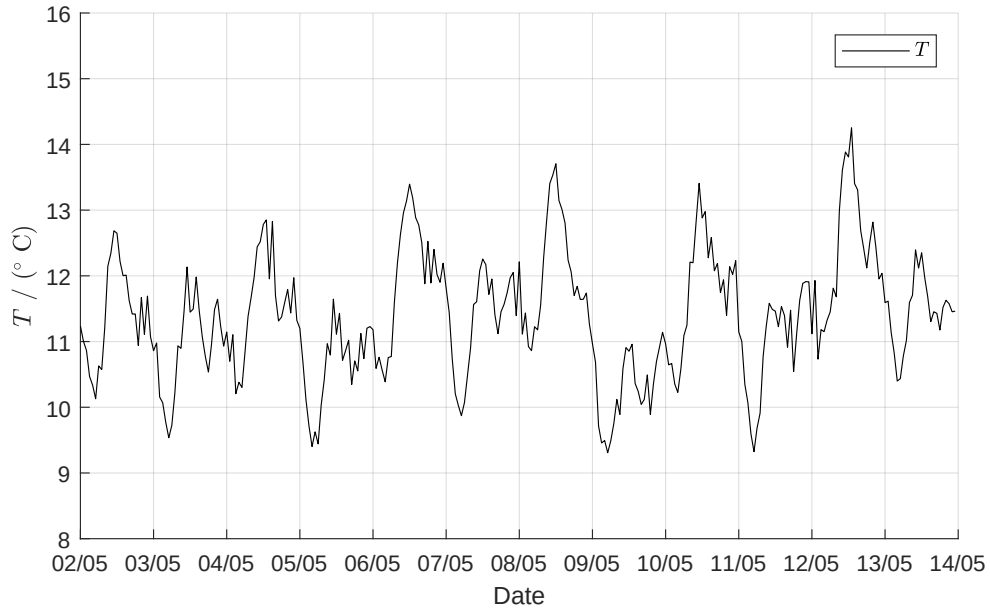


Figure 3.13. Hourly values for influent temperature as final model output.

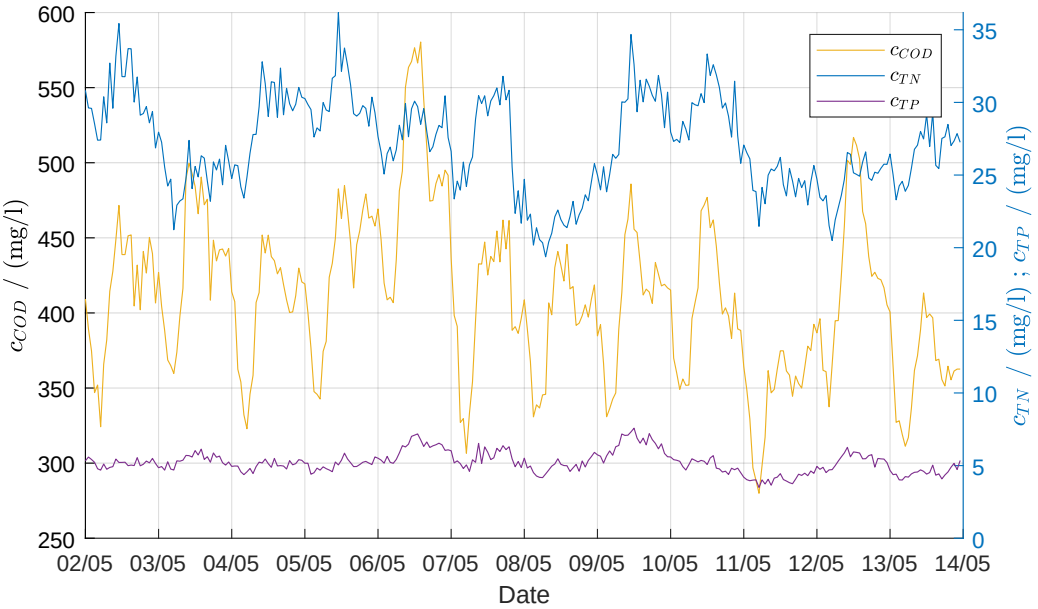


Figure 3.14. Hourly contaminant concentrations resulting from synthetically created hourly data.

of the result. This leaves a lot of room for adjustment for the model user but also shows the relevance of model calibration.

Overall, the hourly values as desired final model outputs appear realistic and physically credible. They demonstrate the most important features observed in the analysis of industrial scale data as well as widely accepted synthetic data (BSM1). Auxiliary calculations show that annual and period averages of influent flow rate, contaminant loads and temperature stated by the original data as well as synthetically created daily averages, are perpetuated. The results indicate large influence of the tunable parameters on the output quality. It is up to the model user to make an appropriate choice for their values depending on his understanding of wastewater creation generally as well as prevalent conditions at the specific plant of interest. Thus a skilled and experienced model user paired with auxiliary information to aid in tuning these parameters is expected to significantly improve the results compared to an arbitrary or unqualified calibration. Nevertheless, dynamic hourly data can be created even from minimum input.

Simulation 4

The intended function of use for the created model is refinement of available influent data for dynamic simulation purposes. It is thus relevant to consider how the induced manipulation of data affects the outcomes of such a simulation, which is examined in this chapter.

4.1 Methodology

All dynamic simulations considered in this chapter are conducted in *Dynamita SUMO 16*, a modern wastewater process simulator. The program has a number of widely used process models implemented, including those of the ASM family.
In the activated sludge process the relevance of phosphorus is mainly limited to being a necessary nutrient for biomass growth, thus being able to inhibit it when depleted. However, the ASM1 does not consider nutrient limitation for biomass growth [6, 8]. Therefore, the ASM2d, one of its successors and the second most used model after the ASM1, is chosen [43]. Details about the inner workings of the model are not discussed here but can be found in [8] or [44].

4.1.1 ASM-fractionation

The different ASM models each use a specific set of state variables. The specification of these variables is often referred to as fractionation, as different fractions of commonly used contaminant classes have to be defined. Generally, a differentiation between particulate and dissolved form, denoted by X and S respectively, is made. The available process parameters need to be converted into the pollutant fractions recognized by the model. Methods for the fractionation vary among different literature and depend on the specific situation of available data and the utilized model. Protocols for the conduction of WWTP simulation suggest different possibilities for determining the fractionation of COD and TN based on additional measurements [45–48]. A typical composition is given along with the presentation of the ASM2d in [44]. The fractionation which was used in creating the BSM1 input data is presented in [23] and [16] describes a fractionation method based on soluble and particulate COD, ammonium, TKN and phosphate.

The following text illustrates how the ASM2d variables are determined from the available parameters. Names of input state variables are emphasized in bold:
The **influent flow rate** Q as well as the **temperature** T are both among the refined plant variables and can be directly used as inputs in the ASM2d. All of the remaining process variables describe amounts of different substances and are to be specified as concentrations in the influent. Therefore available loads of total COD, TN and TP are firstly converted

to concentrations based on Equation 1.1. The amount of **total suspended solids** X_{TSS} is approximated by assuming a fixed ratio of 0.58 g(TSS)/g(COD). The ASM model uses S_{NH4} for the concentration of **dissolved ammonium plus ammonia nitrogen**. This is the same as c_{NH4-N}, which has previously been considered in this report and its value is assumed to be 60 % of the computed TN. These relationships are obtained as a typical value from internal knowledge of the research project. The particulate COD is calculated based on X_{TSS}, as seen in [23]. The dissolved COD hence makes up the difference between total and particulate COD. The **inert soluble organic material** S_I, **fermentation products** S_A and **readily biodegradable organic substrates** S_F are then calculated according to [16]. **Inert particulate organic material** X_I, **slowly biodegradable substrates** X_S and **heterotrophic organisms** X_H, which make up the particulate COD, are computed based on [23]. All phosphorus is assumed to be present in the form of **ortho-phosphates** S_{PO4}. The **alkalinity of the wastewater** S_{ALK}, which is relevant for the buffering capacity, is set to a constant value of $7\,\mathrm{mol}(HCO_3^-)/\mathrm{m}^3$. As suggested by the authors of the ASM2d, several concentrations are set to zero in the influent [8, 44]. These include:

- the **nitrifying organisms** X_{AUT},
- the **phosphate-accumulating organisms** X_{PAO},
- **poly-phosphate** X_{PP} as a cell-internal storage product of phosphate accumulating organisms,
- **poly-hydroxy-alkanoates** X_{PHA} as a cell-internal storage product of phosphate accumulating organisms,
- **metal-hydroxides** X_{MeOH}
- **metal-phosphate** X_{MeP}
- **dissolved oxygen** S_{O2} and
- **nitrate plus nitrite nitrogen** S_{NO3}.

4.1.2 General simulation setup

A plant setup needs to be defined in detail in order to be able to start a simulation. As the main purpose of this chapter is the inspection of the impact of the refinement model on dynamic simulation rather than inspection of a specific real life case, the standardized plant setup from the BSM1 is used and adapted to suit the purpose. The general plant layout is shown in Figure 4.1 [12]. It consists of a Biological Stage and a downstream Settler. The Biological Stage is divided into five compartments, modelled as continuously stirred tank reactors (CSTR). They are denoted as CSTR 1 - 5 in the figure. The first two compartments are not aerated and hence provide a zone for denitrification. The aeration in compartments three and four is kept at a constant air flow, defined by a specified value for the volumetric mass transfer coefficient k_La of 10 h^{-1}. For calibration of the air flow constant inputs based on annual averages calculated from the respective original plant data are used and the air flow is varied until the specified value for k_La is reached. For the last compartment a fixed S_{O2} of 2 mg/l is specified. The air demand is hence automatically adjusted according to that, basically imitating an ideal controller. The settler model specified in [12] is replaced by a volume-less clarifier with a fixed effluent solids content of 10 $\mathrm{g/m}^3$ as the main focus shall be the impact on biokinetics, but not on physical settling processes. Additionally, linear scaling is used to adjust the plant size.

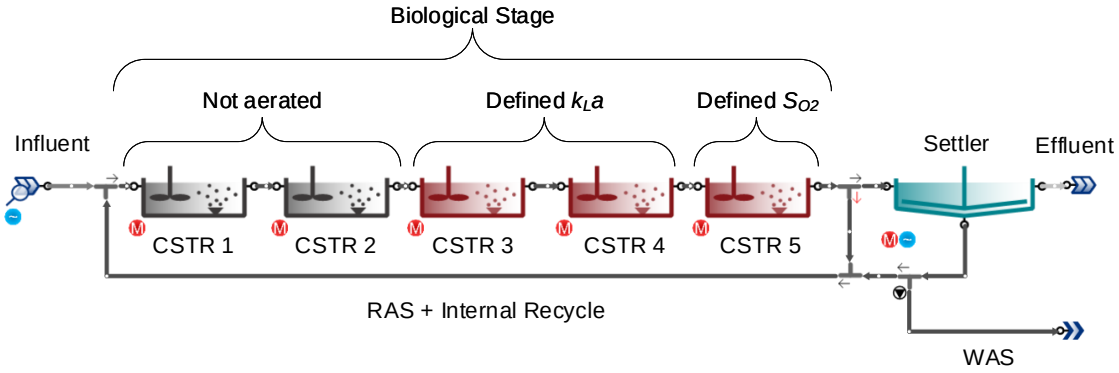


Figure 4.1. Plant layout used for simulation as shown in *Dynamita SUMO 16*. Description of components added manually. Layout obtained from [12].

Therefore, the ratio of the average COD load based on the respective plant data and the average COD load of all three BSM1 scenarios is calculated. The volumes of the different compartments of the Biological Stage as well as the amount of WAS specified in [12] are then multiplied with this factor, thus adapting the capacity to the prevalent load. RAS and Internal Recycle are controlled depending on the influent flow rate. Similarly to the BSM1, they are set to equal one and five times the amount of volumetric influent flow rate respectively and the values for biological parameters are adopted from there [12].

In order to get the plant to the initial state, a 100 day simulation using annual averages is conducted. This provides the starting point for the respective dynamic simulation. Simulation results are always specified to be output in hourly values.

4.1.3 Influent input data

The effects of the individual refinement algorithms on the simulation results shall be shown. Therefore outputs produced by the refinement model at relevant stages of the refinement process are used as simulation input. The effects of algorithms A, B and C are examined separately. Data from the model demonstration in the previous chapter is utilized for coherence. Time series of utilized inputs are found in separate files as supplements to this report.

4.1.3.1 Algorithm A

The simulation software needs to be provided with input values for each time step. State variables of the influent are kept constant until a new value is specified. This means that if any values in the time series of the loads are deleted by the refinement model due to upheld concentrations, an error of the same type is then recreated in the simulation software if the resulting fragmentary data is used. The effects of error elimination and gap filling summarized under Algorithm A are therefore considered as a whole. Original daily data from the WWTP Tramin in 2016 is used as the reference simulation input. It is compared to the complete daily data including all implemented data features which are obtained as intermediate model results after the application of the refinement algorithm according to the model demonstration in the previous chapter (refinement model setup is found in Section 3.1.5.1, respective results in Section 3.2.2.1). The plant setup is calibrated utilizing

the annual averages listed in Table 4.1. The resulting initial plant state as specified by

Table 4.1. Input for calibration of plant setup WWTP Tramin.

Parameter	Q	$\dot{m}_{COD}$	$\dot{m}_{TN}$	$\dot{m}_{TP}$	T
Unit	m^3/d	kg/d	kg/d	kg/d	°C
Value	7931.27	5397.27	436.26	61.56	15.69

the values of the state variables in the five CSTR units is found in the appendix. The parameter of S_{NH4} is consulted for evaluation of the results.

4.1.3.2 Algorithm B

Period averages as initial inputs and daily averages as final outputs of Algorithm B shall be examined in simulation context. Monthly averages of the WWTP Zirl in 2015 are used as reference inputs and simulation results are compared with those from daily input values obtained after refinement according to Section 3.1.5.2. The calibration of the plant setup is done utilizing the annual averages obtained from the WWTP data stated in Table 4.2. This results in an initial state of the plant for starting the simulation as given in the

Table 4.2. Input for calibration of plant setup WWTP Zirl.

Parameter	Q	$\dot{m}_{COD}$	$\dot{m}_{TN}$	$\dot{m}_{TP}$	T
Unit	m^3/d	kg/d	kg/d	kg/d	°C
Value	8556.82	4463.00	303.37	58.66	12.58

appendix. Added value obtained by this part of the refinement model is demonstrated and discussed by virtue of the parameters of S_{NH4} and the amount of biomass withdrawn in WAS.

4.1.3.3 Algorithm C

Finally, the relevance of interpolation to an hourly timescale including implementation of all features specified in Algorithm C of the refinement model is considered. Synthetically created daily and hourly data for the WWTP Zirl obtained from the model demonstration (see Sections 3.1.5.2 and 3.1.5.3) are hence used as inputs for the dynamic simulation which enables a comparison of the respective simulation results. As the data corresponds to the same WWTP as in the previous section, the plant calibration remains the same. Apart from S_{NH4}, relevant influence of the refinement model on simulation output is shown by means of the air flow in compartment 5 of the Biological Stage, as this is representative for aeration control strategies and energy use at the plant. Selected statistical data is computed to quantify the variation in the air flow $\dot{V}_{air}$, which is important for blower capacity considerations. This includes the time weighted average and minimum and maximum values. Additionally, Δ is calculated as a measure variation in the air flow:

$$\Delta = \frac{\sum \mid \dot{V}_{air,i} - \dot{V}_{air,i-1} \mid}{365 \cdot 24} \tag{4.1}$$

Results from monthly input values are included in the observation of air flow.

4.2 Results and discussion

The results of the exemplary dynamic simulations are presented and discussed here to show the impact of the refinement and discuss differences and added value compared to usage of the original data.

As mentioned in the discussion of the model algorithms in Section 3.2.1, the output of the refinement model is highly responsive to the choice of tunable parameters and interpolation methods. Logically, there is deemed to be significant impact on the consequently obtained simulation results. The dominant effects observed for added dynamics in simulation input are expected to be generally valid. However, exact obtained values are likely rather sensitive to the choice of refinement model calibration. Moreover, when investigating a specific real life case, the plant setup should be adapted accordingly and values of biological parameters in the simulation should be validated and adjusted according to established simulation protocols [45–48]. The key takeaway of the results shown in this section are hence not absolute values, but rather the general effects induced in simulation results by previous application of the refinement model compared to the original data as well as the inherent improvements and possibilities.

4.2.1 Algorithm A

Not only total hydraulic and contaminant loads in a given time span, but also their respective temporal distribution over time is highly relevant for plant operation and resulting purity of the effluent. Similarly, the temporal progress of influent temperature influences biological activity. Both, individual values in the time series as well as total average values (and thus total amounts for wastewater and pollutants) were seen to be altered by the elimination of unequivocally identified errors and subsequent completion of fragmentary data in the previous chapter. This shows considerable impact on the simulation results.
In most countries there are legal regulations for maximum effluent concentrations of different pollutants such as COD, TN, NH_4-N and TP. In Italy, these are 100 mg/l, 15 mg/l, 8 mg/l and 2 mg/l respectively [49]. The number of days where permits are violated is specifically relevant. This refers to the average daily concentration being above the legal limit. COD levels are usually rather unproblematic due to its fast degradation. This is also found in the results of both, original, as well as refined data. Effluent COD concentrations are around 42 mg/l for the whole year with only little variation in both cases. The remaining COD consists mainly of inert soluble parts, the concentration of which was specified constant with a value of 30 mg/l in the influent. The effluent concentration of ammonium nitrogen is usually the most problematic parameter in malfunctioning or (temporarily) overloaded WWTPs. This is because of the slow rate of the nitrification process induced by autotrophic bacteria (see Section 1.1). However, ammonia is poisonous to fish and other wildlife even in small doses, which is why its proper removal is essential. Effluent concentrations for ammonium nitrogen differ significantly when using the different inputs, as shown in Figure 4.2. It can be seen that application of the refinement model can both, decrease, as well as increase the peaks in the effluent concentrations compared to the original daily plant data, though the latter is seen much more rarely. In many cases the upholding of concentration values has lead to excessively

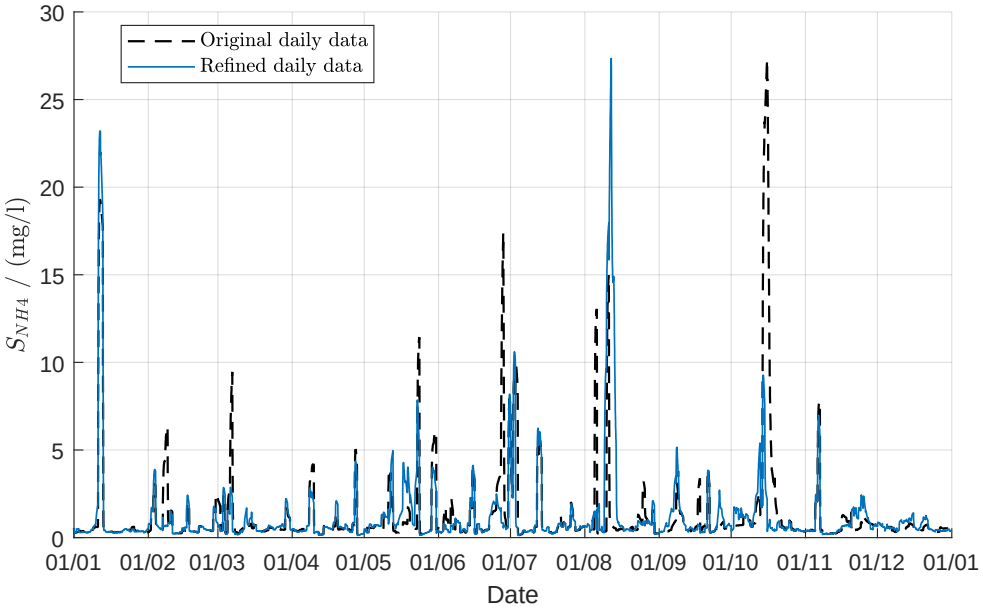


Figure 4.2. Effluent concentrations for dissolved ammonium nitrogen in the examination of refinement model Algorithm A in simulation context.

high influent loads (explained in detail in Section 2.2.1). The elimination of those influent loads and replacement according to the chosen methods is seen to remove or significantly lower many of the larger peaks in S_{NH4} of the effluent. In two particular cases, however, concentration peaks are actually amplified. Consultation of the respective influent loads shows that in these cases concentrations were measured during times of high influent flow rates representing truly present peak loads. However, upholding of these concentrations and combination with subsequently decreasing influent flow rates results in a quick decrease after the peak. The refinement model therefore deletes loads based on upheld concentrations and replaces them either by means of regression based on other parameters or via curve fitting, often resulting in a smoother decline and hence increased total contaminant amount delivered by a peak load, ultimately leading to higher effluent concentrations when ammonium degradation is already insufficient. While the original value might in some cases be close to reality, the consequent removal of the systematic fault is expected to overall increase the realism of the solution.

The total days of ammonium nitrogen permit violations is reduced from 45 to 44 in the prevalent case. While the change in this number is the essential phenomenon to be paid attention to in this case, this is excessively often and would be highly problematic in reality. It likely originates from the fact that due to the insufficient data resolution peaks in influent contaminants always last at least a full day, leading to high amounts of total influent contaminants delivered by such a peak. However, calibration of model parameters for both models was purely empirical for model demonstration and absolute values are hence of minor interest.

While the application of Algorithm A in this case leads to a relative reduction of around 1 % in the total amount of TN delivered by the influent over the course of the year, the total amount of NH_4-N contained in the effluent as computed by the simulation is reduced by nearly 14 %. This has a considerable effect on oxygen uptake in the receiving waters.

4.2.2 Algorithm B

Utilizing the specified period and daily averages as inputs for the simulation demonstrates the importance of the added dynamics for realistic simulation outcomes. Once again effluent NH_4-N concentrations are considered as a critical parameter for indication of appropriate plant function. Examination of the graphs illustrating the effluent concentrations as simulation results for the respective inputs in Figure 4.3 shows considerable differences between the two cases. The usage of monthly input values results

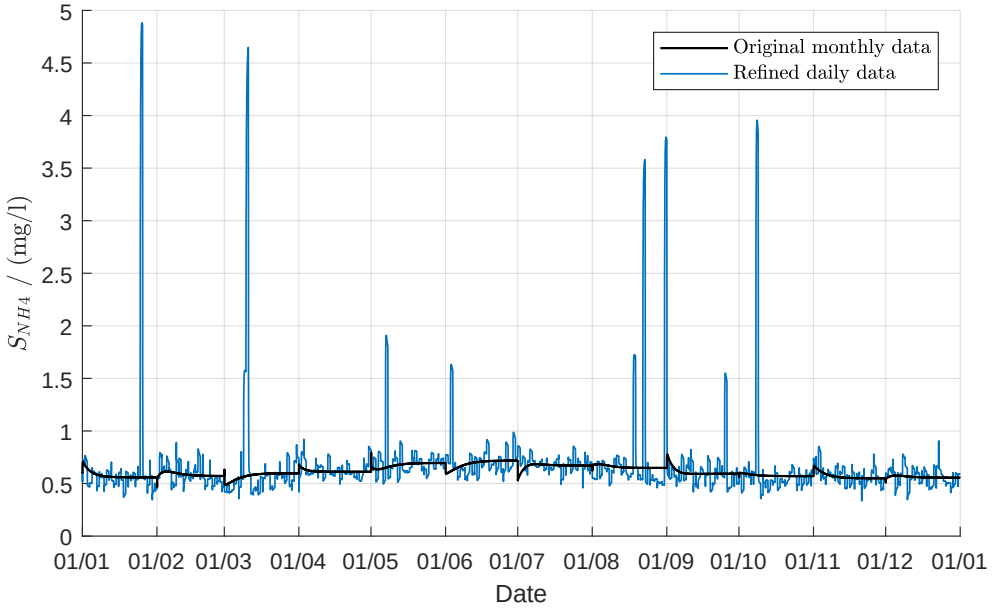


Figure 4.3. Effluent concentrations for dissolved ammonium nitrogen in the examination of refinement model Algorithm B in simulation context.

in constantly low values throughout the year, suggesting unproblematic plant operation with zero days of effluent permit violations. Slight changes are seen to occur at the beginning of each month with new input values, after which the system seems to quickly move towards a steady state with nearly constant effluent concentrations. However, realistic peak loads induced by the refinement model are seen to constitute a significant challenge for a sufficiently thorough purification in the plant and effluent concentrations indicate maximal utilization and even overload of available plant capacities. The number of days of permit violations is seen to increase from zero to five. The conclusions that can be drawn from the different results hence vary drastically which can be of great relevance for adequate decision making.

Though the refinement procedure in Algorithm B does not change the total amount of contaminants in the influent for any given period or the whole year, the resulting amounts leaving the plant in the effluent can differ significantly, with a 15 % increase in total ammonium nitrogen amounts for the observed case. This once again emphasizes that not only total contaminant amounts to be degraded, but very importantly the temporal distribution of these, is of pertinence for the WWTP. When distributed evenly across long time periods, large pollutant amounts can be processed better compared to an uneven distribution. Peak loads lead to problems in effluent concentrations as the delivered substances cannot be degraded sufficiently within the given residence time in the basins.

Biomass growth is dependent on a number of factors such as the availability of nutrients, prevalent temperatures and concentrations or the type of microorganisms. As explained in Section 1.1, growing biomass is removed as WAS to keep the ratio of food to microorganisms in the biological stage at a suitable level. The sludge is usually dewatered and subsequently stabilized anaerobically in a mesophile digestion process after being withdrawn from the wastewater. This leads to emergence of biogas which can then be utilized to gain electric and heating energy. For the examination of processes downstream of the sludge removal, the dynamics and total amounts of sludge production are highly relevant. Moreover, the amount of dry matter removed as sludge is relevant as this ultimately has to be disposed of, causing substantial costs. Figure 4.4 shows the amount of removed biomass as mass flow of solids in WAS $\dot{m}_{TSS,WAS}$ for monthly and daily input data. Similarly to the effluent concentrations, the sludge production indicates a

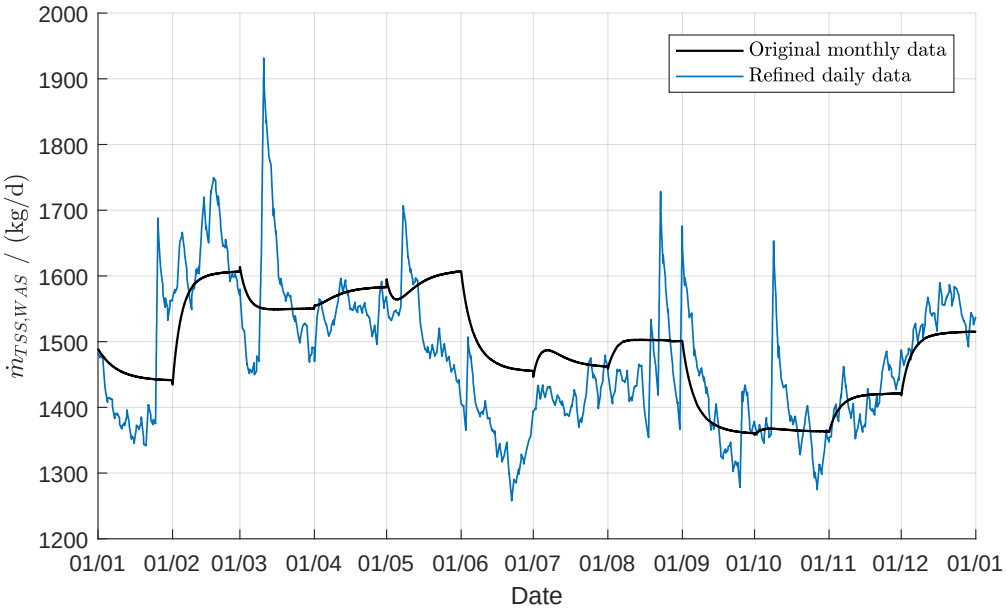


Figure 4.4. Amount of biomass removed in WAS in the examination of refinement model Algorithm B in simulation context.

significant change in operational state only at the beginning of each month after which a steady state seems to be gradually approached when using the period inputs. The use of daily data shows highly dynamic sludge production, proposing variational load on the biological system in the digester. Higher dynamics could require better buffer storage for the sludge depending on the capacity utilization of the digester or it might be relevant for questions evolving around appropriate digester feed strategies. The total annual WAS amount, which influences the electricity production of the WWTP is also seen to differ between both input scenarios, though a difference of only 1 % is observed in this particular case.

Control strategies become particularly important in several areas of plant operation such as basin aeration. This is discussed along with the evaluation of Algorithm C in simulation context in the following section.

4.2.3 Algorithm C

Added dynamics from a daily to an hourly timescale are expected to be relevant for simulation outcomes as refinement methods from literature frequently utilize overlay with diurnal patterns (see Section 1.3). In terms of effluent concentrations, the general course of S_{NH4-N} is somewhat similar to that produced from daily data with absolute values in comparable ranges, as can be observed in Figure 4.5. However, the repetitive diurnal

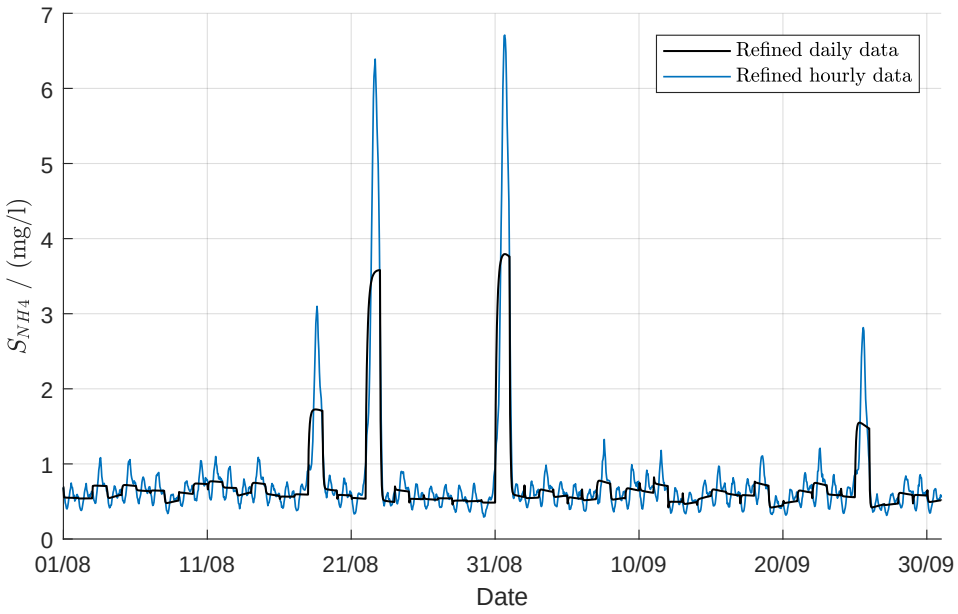


Figure 4.5. Two month excerpt of effluent concentrations for dissolved ammonium nitrogen in the examination of refinement model Algorithm C in simulation context.

profile induced by the refinement model is seen to be carried over to effluent concentrations, most importantly further amplifying effluent peaks and changing their appearance. The number of days of permit violations increase from five to six and the total annual amount of ammonium nitrogen in the effluent is increased by 11 % compared to daily (nearly 28 % compared to monthly) input data.

Control strategies are among the most relevant topics in WWTP operation and are one of the main areas of scientific research within the field. As the aeration in the basins of the Biological Stage is usually the largest contributor to the total energy usage in a plant, it is one of the areas of focus. Strategies here vary widely among different plants. One popular solution is the adjustment of air flow to reach a desired content of dissolved oxygen in the basin, as was specified for compartment 5 in the simulation setup. The prevalent simulation uses a perfect controller, always adjusting the air flow exactly to keep S_{O2} levels constant. Figure 4.6 depicts the air flow necessary for keeping this level of dissolved oxygen as proposed by the simulation for monthly, daily and hourly input values. The variation present in the parameter differs drastically for the different cases. The graph in Figure 4.7 unveils that the air demand calculated based on hourly values experiences some diurnal variation resembling the patterns used in input refinement while it stays largely constant during the day for input data with lower resolution. The upper curve in Figure 4.8, showing the power intake over the course of a full day as recorded at

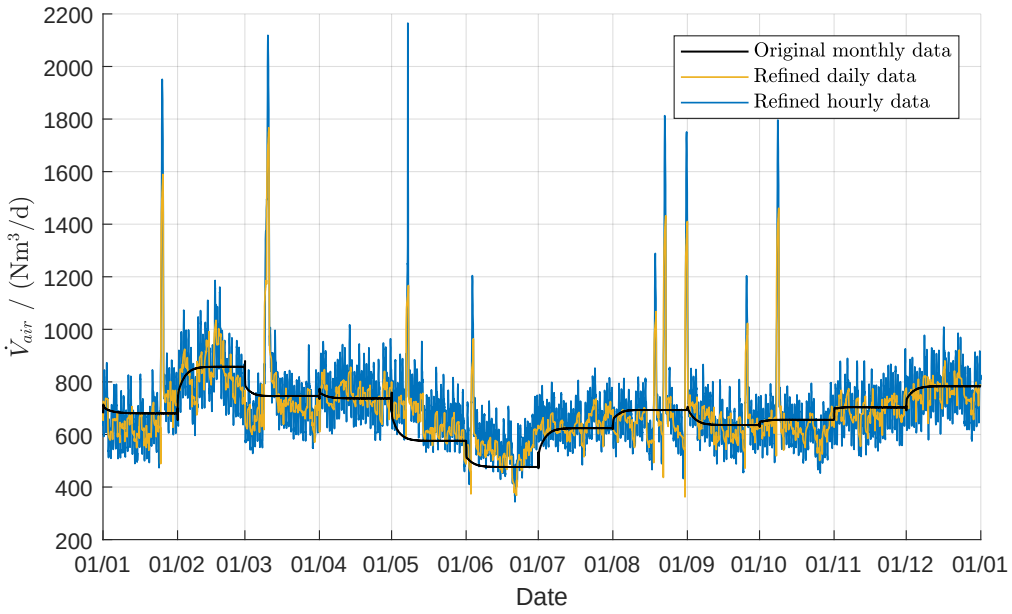


Figure 4.6. Air flow in compartment 5 of the Biological Stage over the course the year 2015 as seen in the examination of refinement model Algorithm C in simulation context.

the WWTP Mittelvinschgau in Italy in 2016, experiences a diurnal variation. Since power intake scales with air supply rates, this clearly speaks for the increased realism of results from hourly inputs compared to monthly or daily data.

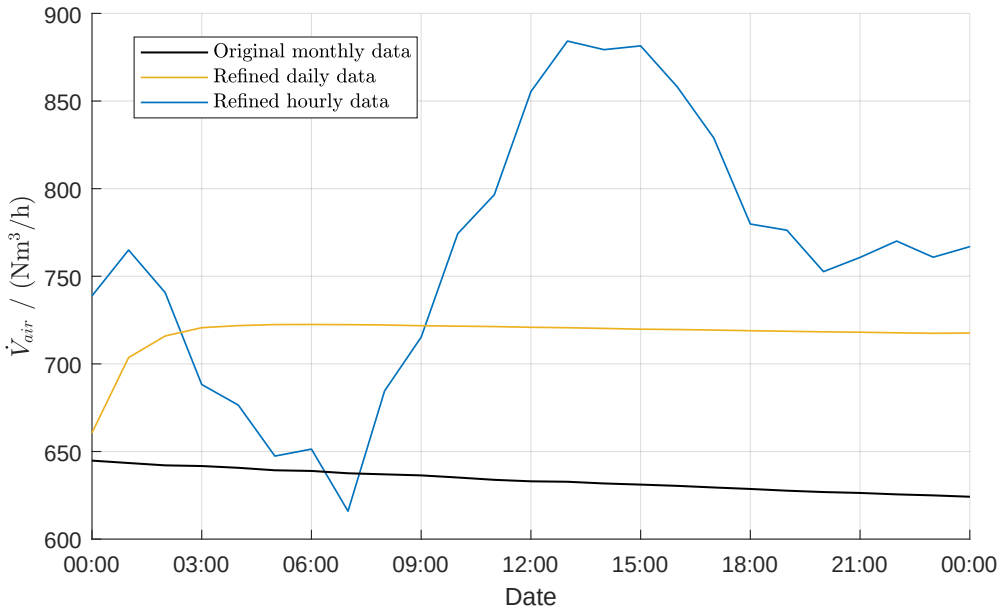


Figure 4.7. Air flow in compartment 5 of the Biological Stage on 02/05/2015 in the examination of refinement model Algorithm C in simulation context.

Controllers cannot be chosen universally, as the demands to the controller are highly dependent on the system and respective dynamics of inputs and disturbances. The dynamics in the ideal air flow are seen to change significantly depending on the input data,

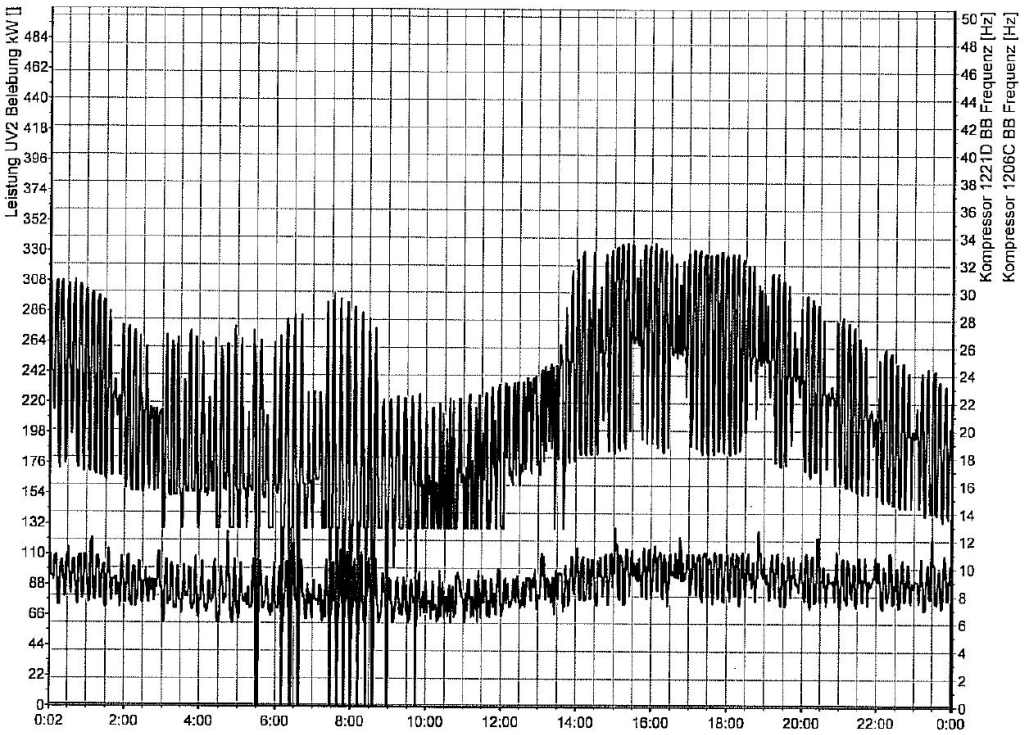


Figure 4.8. Original graphic obtained from the WWTP Mittelvinschgau. Right y-axis and bottom curve irrelevant. Top curve showing curve of power intake for aeration. Left y-axis indicates power intake, x-axis indicates time. Date of recording: 21/08/2016.

which would set entirely different demands for a desired controller. Table 4.3 summarizes selected statistical data about $\dot{V}_{air}$. It shows that Δ, representing the average change

Table 4.3. Statistical data about the air demand calculated from monthly, daily and hourly data in the evaluation of Algorithm C in simulation context.

Used input	Original monthly data	Refined daily data	Refined hourly data
Unit	Nm^3/h	Nm^3/h	Nm^3/h
Average	679.7	691.1	716.1
Minimum	471.4	362.3	343.7
Maximum	879.5	1767.1	2165.0
Δ	0.2	5.9	23.4

in air demand from one hour to the next and hence the variability of the parameter, increases more than hundredfold when using hourly compared to monthly input data and nearly four times when using hourly compared to daily inputs, which is important for setting controller demands. The annual average and hence total air demand, which translates almost directly into energy usage, increases with the data resolution. Moreover, the difference between maximum and minimum air demand increases drastically. The maximum air flow computed from monthly values is exceeded 417 times annually according to daily data and 700 times according to the results from hourly data. The maximum air flow computed from daily values is exceeded 35 times according to the results obtained from hourly data. The simulation results from hourly data amplify the importance of factors such as wide spectrum of operation and fast reaction of the blower to changes in air demand. The efficiency of a blower is highly dependent on the operating point.

Dimensioning and calibration of an aeration system based on simulation results from inadequate inputs can lead to significant problems in terms of pollutant degradation due to insufficient aeration capacity or waste of energy because of unfavourable blower operating points.

Conclusion and Outlook 5

Several efforts for creating data for WWTP simulation have been made in literature. Their applicability depends on the specific situation, as available information and purpose of the simulation can vary in practice, posing different demands to the respective data creation or refinement procedure. The approach proposed in this work is specially designed for the specific situation given in the research project *ICAWER*. The unique combination of methods inspired by state-of-the-art literature as well as approaches that are novel in the field is designed to enable creation of dynamic hourly data for simulation purposes using a systematic and reproducible approach, all the while preserving statements made by the originally given data. For both scenarios of input data the application of the refinement model brings considerable improvements in the realism of the respective simulation results. Inadequately static data does not produce reasonable simulation outcome and is hence unsuitable for inspections of the water purification process or any other downstream procedures such as sludge treatment. In contrast to that, the refinement model increases realism and unlocks potential for the simulation when used in examination of process issues, control, design or dimensioning.

It is important to understand that the information loss occurring in the measurement and documentation of plant data is irreversible. The model does not aim to create an exact replica of the time series of a parameter occurring in reality. Instead, certainly identified errors shall be removed and replaced by what is expected to be realistic based on available data and expert knowledge of the examining engineer. Furthermore, temporal interpolation to an hourly resolution is facilitated and addition of dynamics can be adjusted by the model user as desired. This enables great freedom and adaptability but the quality of the model output as well as thereof obtained simulation results is expected to rely heavily on adequate calibration of tunable parameters and hence on a skilled engineer as well as his understanding of wastewater generation and the specific plant of interest.

The implementation of the created algorithms in software enables their efficient execution in the future. The application of the developed approach could possibly be extended to other parameters or even completely different fields of expertise where similar situations are given.

Future efforts could focus on further enhancing some of the model details such as a differentiation between changes in the parameters induced by dischargers or ambient factors (e.g. surface runoff) and hence more differentiated adjustment of diurnal variations. As extrapolation in the filling of data gaps in Algorithm A might sometimes be necessary depending on the input data and can possibly lead to unrealistic values, a solution for this should be investigated and implemented. Moreover, the development of calibration

protocols for model parameters based on information that is frequently available in addition to the treated parameters (such as measurements of solids content or climate data) could help the less experienced user in creating valid dynamic data as well as standardizing the entire procedure. The model currently does not take future predictions into account. Implementation of predictive models for available data and subsequent application of the refinement model or vice versa could be investigated to extend its use. A method for dynamic fractionation into parameters of the ASM family could be developed and implemented in the model files.

Appendix A

This appendix gives additional information on the simulation setup. Tables A.1 and A.2 indicate the starting points of the dynamic simulations using data from the WWTPs Tramin and Zirl respectively. Units are displayed as output by *Dynamita SUMO 16*.

Table A.1. Initial plant state for the WWTP Tramin. Units as as output by *Dynamita SUMO 16*. This serves as the starting point for the simulation conducted in the evaluation of Algorithm A.

Symbol	CSTR1	CSTR2	CSTR3	CSTR4	CSTR5	Unit
S_{O2}	0.00	0.00	1.79	3.21	2.00	- g COD.m-3
S_F	1.43	0.74	0.71	0.44	0.38	g COD/m3
S_A	1.67	10.99	0.39	0.05	0.02	g COD/m3
S_{NH4}	5.11	5.84	3.17	1.37	0.55	g N/m3
S_{NO3}	0.48	0.01	2.14	4.10	5.07	g N/m3
S_{PO4}	9.37	9.62	9.50	9.54	9.62	g P/m3
S_I	30.00	30.00	30.00	30.00	30.00	g COD/m3
S_{ALK}	0.01	0.01	0.01	0.01	0.00	kmol HCO3-.m-3
S_{N2}	34.72	35.18	35.58	35.75	35.99	g N/m3
X_I	1695.33	1696.49	1698.06	1699.62	1701.19	g COD/m3
X_S	118.18	116.42	96.43	80.07	67.67	g COD/m3
X_H	2167.85	2158.54	2171.01	2174.95	2176.08	g COD/m3
X_{PAO}	3.37	3.37	3.39	3.41	3.41	g COD/m3
X_{PP}	2.74	2.70	2.74	2.77	2.78	g P/m3
X_{PHA}	0.03	0.13	0.07	0.03	0.01	g COD/m3
X_{AUT}	62.50	62.37	62.81	63.15	63.28	g N/m3
X_{TSS}	3285.41	3276.38	3274.31	3267.15	3260.17	g TSS/m3
X_{MeOH}	0.00	0.00	0.00	0.00	0.00	g TSS/m3
X_{MeP}	0.00	0.00	0.00	0.00	0.00	g TSS/m3
T	15.64	15.64	15.64	15.64	15.64	°C

Table A.3 shows operational parameters in the unit CSTR5 for all simulations. These are suggested default values and are relevant for the necessary amount of air computed by the program to reach the specified S_{O2} of 2 mg/l. Note that the format of the variables and units is not adapted to this report but specified as reported by the program.

Table A.2. Initial plant state for the WWTP Zirl. Units as as output by *Dynamita SUMO 16*. This serves as the starting point for the simulations conducted in the evaluation of Algorithms B and C.

Symbol	CSTR	CSTR2	CSTR3	CSTR4	CSTR5	Unit
S_{O2}	0.00	0.00	2.39	3.88	2.00	- g COD.m-3
S_F	1.38	0.73	0.72	0.46	0.40	g COD/m3
S_A	1.74	8.46	0.39	0.06	0.03	g COD/m3
S_{NH4}	3.78	4.31	2.45	1.23	0.61	g N/m3
S_{NO3}	0.41	0.01	1.56	2.92	3.64	g N/m3
S_{PO4}	8.07	8.23	8.17	8.21	8.27	g P/m3
S_I	30.00	30.00	30.00	30.00	30.00	g COD/m3
S_{ALK}	0.01	0.01	0.01	0.01	0.01	kmol HCO3-.m-3
S_{N2}	24.40	24.79	25.00	25.11	25.29	g N/m3
X_I	1692.74	1693.63	1694.81	1696.00	1697.19	g COD/m3
X_S	109.34	108.24	92.91	79.90	69.59	g COD/m3
X_H	2105.66	2098.74	2108.30	2111.73	2113.17	g COD/m3
X_{PAO}	1.19	1.19	1.20	1.20	1.20	g COD/m3
X_{PP}	0.97	0.96	0.97	0.98	0.98	g P/m3
X_{PHA}	0.01	0.04	0.02	0.01	0.00	g COD/m3
X_{AUT}	55.85	55.76	56.06	56.30	56.40	g N/m3
X_{TSS}	3208.89	3202.40	3200.72	3195.17	3189.72	g TSS/m3
X_{MeOH}	0.00	0.00	0.00	0.00	0.00	g TSS/m3
X_{MeP}	0.00	0.00	0.00	0.00	0.00	g TSS/m3
T	12.60	12.60	12.60	12.60	12.60	°C

Table A.3. Operational parameters for CSTR5 in all simulations. Variables and units as output by *Dynamita SUMO 16*.

Symbol	Value	Unit
HRT	0.02	d
hdiff	4.80	m
SSOTE	6.00	%/m
alpha	70.00	%
Tair	15.00	*decC*
pair	102325.00	Pa
Beta	71.04	%
F	80.00	%
hsea	200.00	m
Lair	0.01	K/m
tR_air	10.00	s
GO2_air_inp	20.95	%v/v

Bibliography

[1] T. N. Herzog, F. J. Scheuren, and W. E. Winkler. *Data Quality and Record Linkage Techniques*. Springer Science+Business Media, LLC, 2007.

[2] Autonome Provinz Bozen Südtirol. *Interreg V-A Italy-Austria 2014-2020*. 2016. URL: http://www.interreg.net/en/programme.asp.

[3] SYNECO Group GmbH. *ICAWER: Interregional Concept for Advanced Wastewater Reclamation*. 2017. URL: https://sites.google.com/syneco-group.com/icawer/deutsch.

[4] W. Gujer. *Siedlungswasserwirtschaft*. 3rd ed. Springer Berlin Heidelberg, 2006.

[5] Metcalf & Eddy, Inc. et al. *Wastewater engineering: treatment and reuse*. McGraw Hill, 2003.

[6] M. Henze et al. *Activated Sludge Model No. 1*. IAWPRC, 1987.

[7] D. W. Schindler and J. R. Vallentyne. *The Algal Bowl: Overfertilization of the World's Freshwaters and Estuaries*. University of Alberta Press, 2008.

[8] M. Henze et al. *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA Publishing, 2000.

[9] G. S. Ostace, V. M. Cristea, and P. Ş. Agachi. "Cost reduction of the wastewater treatment plant operation by MPC based on modified ASM1 with two-step nitrification/denitrification model". In: *Computers & Chemical Engineering* 35.11 (2011), pp. 2469–2479.

[10] F. Gao et al. "Modeling and simulation of a biological process for treating different COD:N ratio wastewater using an extended ASM1 model". In: *Chemical Engineering Journal* 332 (2018), pp. 671–681.

[11] Z. Zhu, R. Wang, and Y. Li. "Evaluation of the control strategy for aeration energy reduction in a nutrient removing wastewater treatment plant based on the coupling of ASM1 to an aeration model". In: *Biochemical Engineering Journal* 124 (2017), pp. 44–53.

[12] J. Alex et al. *Benchmark simulation model no. 1 (BSM1): Report by the IWA Taskgroup on Benchmarking of Control Strategies for WWTPs*. 2008.

[13] I. Nopens et al. "Benchmark Simulation Model No 2: Finalisation of plant layout and default control strategy". In: *Water Science and Technology* 62.9 (2010), pp. 1967–1974.

[14] U. Jeppsson et al. "Benchmark simulation models, quo vadis?" In: *Water Science and Technology* 68.1 (2013), pp. 1–15.

[15] M. Devisscher et al. "Estimating costs and benefits of advanced control for wastewater treatment plants - the MAgIC methodology". In: *Water Science and Technology* 53.4-5 (2006), pp. 215–223.

[16] K. V. Gernaey et al. “Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach”. In: *Environmental Modelling and Software* 26.11 (2011), pp. 1255–1267.

[17] C. Martin and P. A. Vanrolleghem. “Analysing, completing, and generating influent data for WWTP modelling: A critical review”. In: *Environmental Modelling and Software* 60 (2014), pp. 188–201.

[18] W. De Keyser et al. “An emission time series generator for pollutant release modelling in urban areas”. In: *Environmental Modelling and Software* 25.4 (2010), pp. 554–561.

[19] X. Flores-Alsina et al. “Calibration and validation of a phenomenological influent pollutant disturbance scenario generator using full-scale data”. In: *Water Research* 51 (2014), pp. 172–185.

[20] M. Almeida, D. Butler, and E. Friedler. “At-source domestic wastewater quality”. In: *Urban Water* 1.1 (1999), pp. 49–55.

[21] G. Langergraber et al. “Generation of diurnal variation for influent data for dynamic simulation”. In: *Water Science and Technology* 57.9 (2008), pp. 1483–1486.

[22] G. Mannina et al. “A practical protocol for calibration of nutrient removal wastewater treatment models”. In: *Journal of Hydroinformatics* 13.4 (2011), p. 575.

[23] H. Vanhooren and K. Nguyen. “Development of a simulation protocol for evaluation of respirometry-based control strategies”. In: *Report University of Gent and University of Ottawa* (1996).

[24] J. Langeveld et al. “Empirical sewer water quality model for generating influent data for WWTP modelling”. In: *Water (Switzerland)* 9.7 (2017), pp. 1–18.

[25] G. Gins et al. “Data Alignment Via Dynamic Time Warping as a Prerequisite for Batch-End Quality Prediction”. In: *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining. ICDM 2006. Lecture Notes in Computer Science* 4065 (2006). Ed. by P. Perner, pp. 506–510.

[26] C. Martin et al. “ARMA models for uncertainty assessment of time series data: application to Galindo-Bilbao WWTP”. In: *Proceedings of the Seventh International IWA Symposium on Systems Analysis and Integrated Assessment in Water Management, Washington DC, USA, 7th–9th May*. 2007.

[27] M. Soley-Bori. *Dealing with missing data: Key assumptions and methods for applied analysis*. Technical Report. Boston University: School of Public Health, Department of Health Policy & Management, 2013.

[28] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2014.

[29] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2012.

[30] MathWorks, Inc. *Documentation: corrcoef*. URL: https://se.mathworks.com/help/matlab/ref/corrcoef.html.

[31] D. M. Diez, C. D. Barr, and M. Cetinkaya-Rundel. *OpenIntro Statistics*. 3rd ed. 2015.

[32] C. Ort et al. "Modelling stochastic load variations in sewer systems". In: *Water Science and Technology* 52.5 (2005), pp. 113–122.

[33] German Association for Water, Wastewater and Waste. *Standard ATV-DVWK-A 131E: Dimensioning of Single-Stage Activated Sludge Plants*. 2000.

[34] A. Sorjamaa. "Methodologies for Time Series Prediction and Missing Value Imputation". PhD thesis. Aalto University School of Science and Technology, Espoo, Finland, 2010.

[35] D. S. Fung. "Methods for the Estimation of Missing Values in Time Series". MA thesis. Edith Cowan University, Perth, Western Australia, 2006.

[36] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer Science+Business Media, LLC, 2006.

[37] C. de Boor. *A Practical Guide to Splines*. Springer, 2001.

[38] T. Young and M. J. Mohlenkamp. *An Introduction to MATLAB® Programming and Numerical Methods for Engineers*. 6th ed. Department of Mathematics, Ohio University, 2015.

[39] MathWorks, Inc. *Documentation: csape*. URL: https://se.mathworks.com/help/matlab/ref/corrcoef.html.

[40] F. N. Fritsch and R. E. Carlson. "Monotone piecewise cubic interpolation". In: *SIAM Journal on Numerical Analysis* 17.2 (1980), pp. 238–246.

[41] D. Butler. "The influence of dwelling occupancy and day of the week on domestic appliance wastewater discharges". In: *Building and Environment* 28.1 (1993), pp. 73–79.

[42] D. Butler and K. Gatt. "Synthesising dry weather flow input hydrographs: a Maltese case study". In: *Water Science and Technology* 34.3-4 (1996), pp. 55–62.

[43] H. Hauduc et al. "Activated sludge modelling in practice: An international survey". In: *Water Science and Technology* 60.8 (2009), pp. 1943–1951.

[44] M. Henze et al. "Activated Sludge Model No. 2d". In: *IAWPRC Scientific and Technical Report No. 1*. Vol. 39. 1. 1987, pp. 165–176.

[45] J. Hulsbeek et al. "A practical protocol for dynamic modelling of activated sludge systems". In: *Water Science and Technology* 45.6 (2002), pp. 127–136.

[46] P. A. Vanrolleghem et al. "A comprehensive model calibration procedure for activated sludge models". In: *Proceedings of the Water Environment Federation* 2003.9 (2003), pp. 210–237.

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

[48] G. Langergraber et al. "A guideline for simulation studies of wastewater treatment plants". In: *Water Science and Technology* 50.7 (2004), pp. 131–138.

[49] Autonome Provinz Bozen Südtirol. *Landesgesetz vom 18. Juni 2002, Nr. 8, Bestimmungen über die Gewässer, Anhang A: Emissionsgrenzwerte für Kläranlagen für kommunales Abwasser mit einer Leistung bis 2000 EW*. 2002.