Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Proxy measurement and strategic metering methodologies for resource transparency of industrial systems

Dominik Seiler

Department of Mechanical and Manufacturing Engineering
Trinity College Dublin
Ireland

A thesis submitted to the University of Dublin in partial
fulfilment of the requirements for the degree of

Doctor of Philosophy

August 2020

Supervisor:
Dr. Garret E. O’Donnell

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Dominik Seiler

Dominik Seiler

August 2020

This work was funded under the Energy Systems Integration Partnership Programme (ESIPP), which is financially supported by the Science Foundation Ireland under the SFI Strategic Partnership Programme (Grant Number SFI/15/SPP/E3125), industrial partners, and a philanthropic donation.

Summary

For the latest advancement of manufacturing systems, namely ‘Industry 4.0’ or Smart Manufacturing, large amounts of precise process data are required. However, the acquisition of sufficient data represents a challenge for many manufacturing plants. Developed metering frameworks from academia are often not applicable to real-world manufacturing processes due to monitoring gaps. These gaps result from missing and broken measuring devices of industrial systems and lead to an incomplete monitoring outcome that can prevent system optimisations. Therefore, this research focuses on the development of effective data gathering tools for complex industrial systems in order to provide a comprehensive process understanding. By considering industrial needs, the design and development of functional resource measurement methodologies and the study of predictive models with a reduced-order characteristic have been addressed to accelerate holistic data acquisition in industrial environments.

Following from the investigations of two purified water and one steam system in an industrial case facility, a comprehensive metering strategy for detailed information gathering of technical building services has been developed. Although these systems require large shares of resources, holistic analysis methods are not comprehensively addressed in the literature. The novel four phase metering methodology in this research fills this gap by identifying available data sources, abstracting the central process steps, and mapping the resource flows within technical building services. In the last phase, the absence of functional online meters is surrogated by a basic proxy metering strategy that enables the approximation of missing parameters by combining related data sources in a regression model. This holistic strategy represents an effective data acquisition approach with high adaptability to existing conditions and constraints that has not been reported in the literature to date.

The gathered process information from the holistic meter approach allowed the development of an automated calculation method for the total cost of technical building services. By focusing on the interactions of the identified resources and further added values, this cost index tracks the real value that is embedded in the analysed system. Compared to similar approaches in the literature, the cost calculation method in this research can be automated which enables the impact of system changes to be studied.

Based on the potential of basic proxy metering devices (PMDs) for estimating missing process data and the industrial need for simpler implementation strategies, the development of comprehensive PMD modelling methodologies has been successfully addressed in the second part of this research. For the common case of small datasets, the regression performance of PMD algorithms for diverse dataset regression complexities has been analysed. As well as the complexity characterisation of five small datasets from an experimental rig, the estimation accuracies of the most common linear and non-linear PMD algorithms have been studied, including the impact of training sample manipulations with bootstrap and artificial noise injection. The results of this comparison highlight which PMD algorithm and training sample manipulation technique is appropriate, depending on the regression complexity of the dataset and the number of training samples. With this novel methodology, industrial users can select a targeted and straightforward PMD development approach based on their specific regression complexity and use case.

Acknowledgments

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible without the help, guidance, and support of the following people.

Firstly, I would like to sincerely thank my supervisor Dr. Garret E. O'Donnell who gave me the opportunity to pursue this work and offered me constant support, advice, and encouragement throughout the past four years.

I would also like to thank the Energy Systems Integration Partnership Programme (SFI/15/SPP/E3125) for providing financial support and helpful inspirations throughout my PhD.

A special thank you goes to Dan Donovan from Merck Millipore in Carrigtwohill. With his help, a close collaboration with industry became possible which turned out to be very rewarding for this work and my own engineering development. Thanks must be also extended to the rest of the team in Merck Millipore, in particular to Barry, Ronan, Tim, Tom, Liam, Fintan, Brendan, and Derek. Through their constant help, they have contributed significantly to the success of this work.

I would also like to extend my gratitude to Gerry Byrne and Kevin Kelly from the Department of Mechanical and Manufacturing Engineering in Trinity College Dublin for their knowledge and technical assistant.

To all the postgrads in the Mechanical Department, past and present: Morten, Nicolas, Marios, Parth, Alessia, Federico, Agnieszka, Harry, Mark, Tom, Daniel, Cian, Luke, Jaakko, Jason, Maeve, and Elenora. Thank you for all the friendship, advice, lunchtime company, and laughs inside and outside of the office.

To my brother and sister, Hendrik and Lara, thank you for always being so helpful in numerous ways throughout the last four years and beyond.

To Michelle, who's love, care, patience, and companionship has been a major source of strength throughout this PhD. Thank you for always being by my side.

Finally, my deepest gratitude goes to my parents, Barbara and Klaus. Their belief in me and endless love has allowed me to follow my dreams. You have always been, and always will be, my greatest inspiration. This is for you.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

ANI	Artificial noise injection
BFD	Block flow diagram
BR	Boorstrap replications
CC	Pearson correlation coefficient
CPG	Production line in case facility
DCS	Distributed control system
ERP	Enterprise resource planning
HMI	Human-machine interface
HVAC	Heating, ventilation, and air conditioning
IC1	Production line in case facility
IC2	Production line in case facility
IEA	International Energy Agency
IoT	Internet of Things
ISO	International Organization for Standardization
IT	Information technology
KPI	Key Performance Indicator
MAD	Median absolute deviation
MES	Manufacturing execution system
MLP	Multi-layered perceptron
MLR	Multiple linear regression
MSE	Mean squared error
NIPALS	Nonlinear Iterative Partial Least Square
NN	Neural Network
OECD	Organisation for Economic Co-operation and Development
OEE	Overall equipment effectiveness
PCA	Principal Component Analysis
PID	Proportional-integral-derivative
P&ID	Piping and Instrumentation Diagram
PLC	Programmable logic controller
PLSR	Partial Least Square regression

PMD	Proxy meter device
PW	Purified water
RMSE	Root mean squared error
RTU	Remote terminal unit
SCADA	Supervisory control and data acquisition
SLR	Simple linear regression
SME	Small- to medium-sized enterprise
SMF	Production line in case facility
SRU	Solvent recovery unit
STD	Standard deviation
TBS	Technical building services
TOC	Total organic carbon
UV	Ultraviolet
WPD	Water purification devices

Measuring devices and components in process/ block flow diagrams

Measuring and control devices

First-Letter(s)		Succeeding-Letter	
F	Flow rate	C	Control
L	Level	I	Indicator (offline)
P	Pressure	T	Transmitter (online)
pH	pH-value		
T	Temperature		
XM	Pump		

Components

P	Pump
V	Valve

Chapter 1

Introduction

1.1 Energy and resource sustainability in the manufacturing sector

Since the first industrial revolution, the industrial sector has continued to be a key driver for welfare and prosperity within society, accounting for nearly one-quarter of all jobs globally to date [1]. Not only does industry provide employment, it is also decisive for global presence and economic strength of industrialised nations. Due to these significant global impacts, the industrial sector plays a leading role in technological, economical, political, and societal advancement. The downside of industrial processes is the direct and indirect contribution to the depletion of natural resources and the resulting environmental burdens that are affecting humanity and the eco-system [2]. According to the International Energy Agency (IEA) [3], the industrial sector accounted for 37% of global energy use in 2017, with an average increase of 1% per year for the period from 2010 until 2017. As depicted in Figure 1.1 [a], a further year-on-year energy consumption increase for the industrial sector was predicted until 2025.

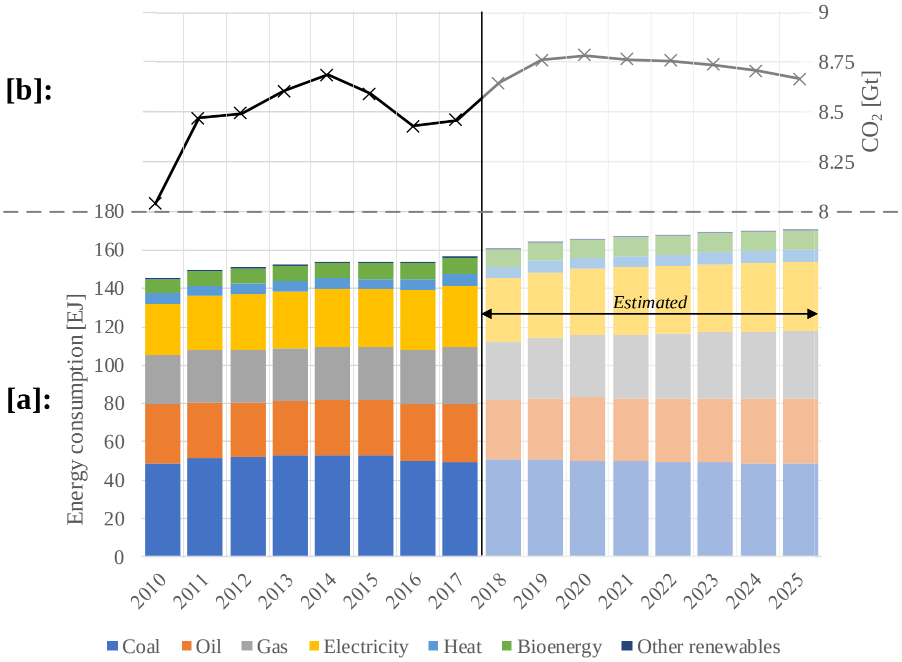


Figure 1.1: Global industrial energy consumption [a] and direct CO_2 emissions [b]; adapted from [3]

However, current developments in light of the Coronavirus pandemic have demonstrated that unforeseeable occurrences can have a major impact on the industrial energy consumption and production rates. A first IEA publication on the impact of the pandemic [4] indicates that the worldwide industrial production will decrease significantly 2020 which may lead to the largest year-on-year reduction of industrial energy demand to date. For example, the natural gas consumption of the industrial sector is expected to decrease by 5% compared to the year 2019. Nonetheless, the industrial production rates and energy consumptions are expected to return to pre-pandemic levels and increase further in the long term.

Besides energy, further resources, such as water, are required for industrial production. Industrial facilities require 19% of the global water demand [5]. On the emission side, 8.5 Gt of CO_2 emissions, equivalent to 24% of the global CO_2 output, came directly from industrial activities in 2017 [3], see Figure 1.1 [b]. The majority of these consumptions and emissions relate to the manufacturing sector, a subset of the industrial sector which involves the fabrication or assembly of raw materials and components into discrete finished products through the systematic division of labour [2]. Manufacturing processes are responsible for 90% of the industrial sectors energy consumption and 84% of the industrial sectors energy-related CO_2 emissions [6]. This high consumption of resources and the emission of pollutants has led to a rising awareness and need for sustainability within manufacturing companies. Stark et al. [2] show that the use of current industrial technologies to meet future economic growth expectations will result in a resource consumption that exceeds every accountable environmental, economic and social boundary. Sustainable manufacturing is seen as the practice of creating manufactured products by using processes that minimise negative environmental impacts, conserve energy and natural resources, are safe for employees, communities, and are economically sound [7]. Besides the scarcity of numerous natural resources and high emissions, the efforts towards sustainable manufacturing are pushed further by stricter governmental regulations and taxes that foster environmentally benign manufacturing and by an increasing demand from society for products with minimal environmental impact [8].

To save energy and resources in manufacturing, Herrmann et al. [9] pointed out that a reduction of production activities is not seen as a feasible solution. Instead, the improvement of energy and resource consumption in complex manufacturing processes is needed. The IEA have estimated that energy savings of up to 29% are realisable in the manufacturing sector if all proven technologies and best practice guidelines are implemented [10]. Further energy savings are achievable but will require global cooperation and innovation to develop highly sustainable production systems [11]. A fundamental requirement and first step towards optimising these systems is the measurement and processing of performance data [12], [13]. Process monitoring systems provide these measures by combining manufacturing system information and enabling process control.

1.2 Monitoring of manufacturing processes

A process monitoring system is a build-up of measuring instruments that utilise sensors as the primary elements for signal acquisition [14]. These signals are subject to analogue and digital conditioning and processing. According to Teti et al. [15], the aim of monitoring systems is to generate functional signals that can be correlated to process conditions. For manufacturing systems, the signals being acquired and their processing is dependent on the process itself and the desired output parameter(s), e.g. temperature, power, vibration etc. [14]. For the majority of modern monitoring instruments, the sensor output-signal is in an electric form which simplifies signal transmitting, processing, and storing. Once the signal is acquired, various process and data analysing steps are carried out to extract useful information. These steps can include filtering, amplification, and segmentation [15]. In many measuring instruments the required parameter can only be attained through an indirect measurement. In these cases, the actual quantity is inferred by a relation between sensor signals and physical variations in the form of a mathematical model, i.e. transfer function. This function may be of linear or nonlinear form.

To monitor and control a manufacturing system, a multitude of measuring instruments are used to collect process information. Typically, a process or facility-wide monitoring system is implemented that acquires data from the metering devices and offers data visualisation for operators through human-machine interfaces (HMIs) [16], displayed in Figure 1.2. From this information, the process state is determined by operators and/or unmanned decision making support systems [17]. If a process anomaly is detected, control actions are carried out to manipulate process variables in order to regulate or stop the process [18].

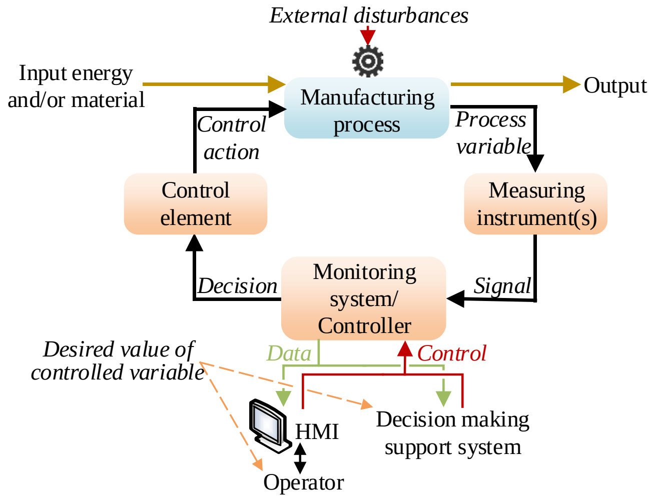


Figure 1.2: Process monitoring feedback-control system; adapted from [15], [19]

1.3 The increased need for process data

Modern industrial facilities require numerous measuring instruments for the monitoring and storing of process information. This represents an enormous increase in process data compared to the early days of manufacturing. As Tao et al. [20] described, the manufacturing data that was available during the handicraft and machine age was limited and the performance of work was mainly based on experience (Figure 1.3). This changed with the development and integration of information technology (IT) within manufacturing. Automated and computer-based systems were introduced, exposing the manufacturing sector to a more data-rich environment.

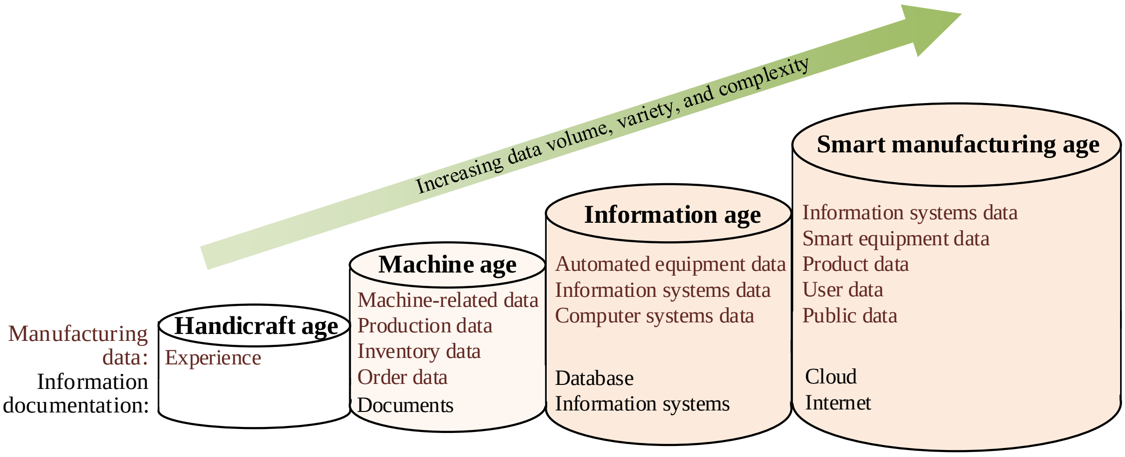


Figure 1.3: Evolution of data in manufacturing environments; adapted from [20]

In terms of the current development of manufacturing systems towards smart and interconnected systems, known as Industry 4.0 and Smart Manufacturing, a further exponential increase of process information is required to provide the data needs of these new advancements [21]. This increased amount of real-time process data aims to convert current manufacturing lines into intelligent process systems that yield positive impacts on many aspects of manufacturing [22]. These impacts include improved process control and flexibility to realise individual customer demands [23], increased production rates, and decreased energy and resource requirements [24]. While these opportunities are desirable for many manufacturers, several challenges exist when looking to implement currently available smart manufacturing concepts. The increased need for accurate and real-time process data is seen as one of these key challenges for upgrading manufacturing processes to intelligent systems, as reported by Kusiak [25]. As Gontarz et al. pointed out [21], this need leads to complex monitoring architectures and additional sensor requirements to enhance data collection, processing and storage in many manufacturing plants. However, this data requirement cannot easily be met, especially by small- to medium-sized enterprises (SMEs) whose capabilities tend to be more limited. These facilities are confronted with the trade-off between reliability, accuracy, and cost for monitoring systems. Ultimately, companies that do not gather sufficient data to enable smart manufacturing developments face the risk of losing competitiveness in the near future [25].

Several approaches have been developed within academia to help manufacturers meet the requirements for increased data acquisition. However, many of these concepts have limited or no application for manufacturing systems whose variables are not already accurately metered [26]. To close this gap, several authors [25]–[27] outlined a need for more process monitoring and data gathering frameworks that have evolved from the cooperation between industry professionals and researchers.

1.4 Research aims and objectives

With a focus on manufacturing processes, the goal of the present research is to demonstrate that basic metering tools are able to provide detailed insights into complex industrial system. The hypothesis of this research is that the use of reduced-order predictive models coupled with functional measurement strategies can enable advanced data gathering and comprehensive process understanding of resources in industrial systems.

The research objectives are:

1. Definition and development of a methodological framework for metering resource flows and consumptions in industrial processes. This framework must be applicable to various system setups and consider the restrictions imposed by the industrial process environment.
2. Characterisation of resource interactions within industrial systems to reveal the embedded value. This objective investigates the need of considering interrelationships of multiple resources within an industrial system.
3. Development of universal methodologies for the implementation of reduced-order models that aim to approximate data gaps in industrial processes. This objective completes the previously defined metering framework by providing easy to adopt algorithms and effective model development steps for the estimation of missing parameters.

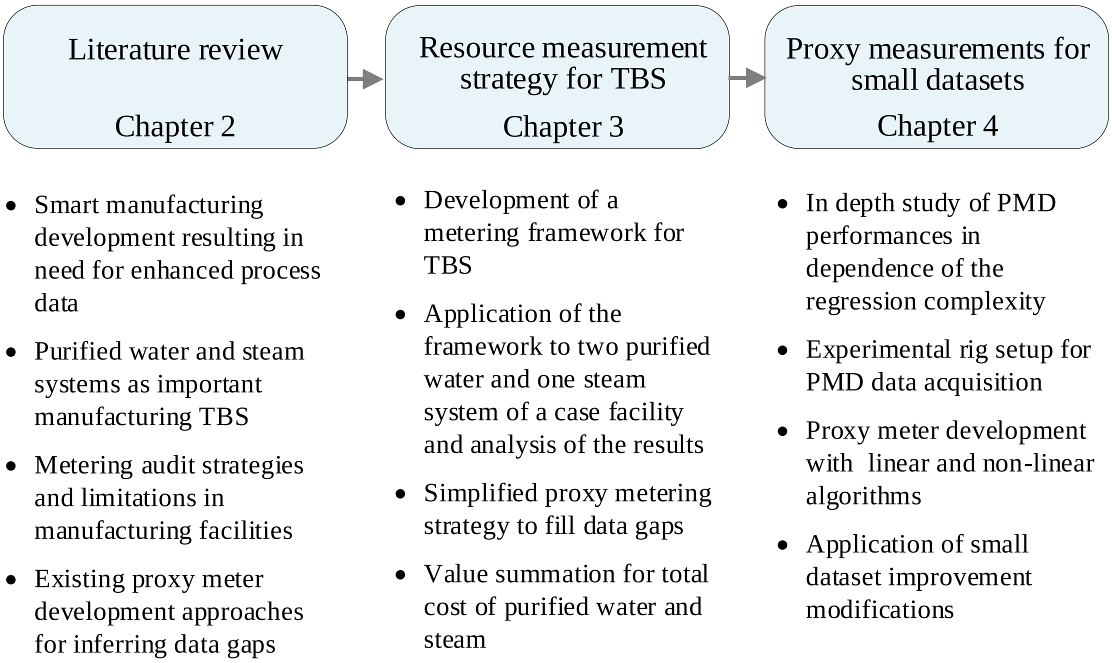
Contribution to the existing research

The first stage of this research focused on the generation and distribution of purified water and steam within a manufacturing facility. Purified water and steam were selected as two widely used technical building services (TBS) in manufacturing surroundings that use significant amounts of energy and water. Although these systems require large shares of resources, detailed metering guidelines are not comprehensively addressed in the literature. A metering strategy for these two subsystems was developed during an industrial case study in a large multinational life science company located in Ireland. By approaching the metering framework from an industrial perspective, this work bridges the gap between academic research and engineering application by considering industrial needs and constraints. Through the application of the developed metering strategy in the case facility, sufficient process data was collected to demonstrate the necessity to consider purified water and steam systems as interconnected energy and resource structures. This interconnection was demonstrated by an automated calculation methodology for the total cost of purified water and steam generation.

The second major component of this research focused on the development of virtual meters, i.e. proxy metering devices (PMDs), as a technique to estimate data gaps during industrial monitoring activities. This has been addressed due to the industrial demand and literature gap for simpler PMD implementation methods. Stemming from five diverse datasets generated from a constructed experimental rig, various PMDs were tested in order to develop the modelling framework presented within this work. This framework bears in mind the often limited size of industrial datasets and the needed effortlessness to be adaptable to diverse industrial scenarios. This novel methodology provides industrial users an effective development approach of PMD models with a reduced-order characteristic that has not been reported in the literature to date. In the context of this work, the term reduced-order model describes a model type that can be developed with low efforts and high transparency.

Thesis layout

A summary of the thesis layout is presented in the following figure.



Literature review

Chapter 2

- Smart manufacturing development resulting in need for enhanced process data
- Purified water and steam systems as important manufacturing TBS
- Metering audit strategies and limitations in manufacturing facilities
- Existing proxy meter development approaches for inferring data gaps

Resource measurement strategy for TBS

Chapter 3

- Development of a metering framework for TBS
- Application of the framework to two purified water and one steam system of a case facility and analysis of the results
- Simplified proxy metering strategy to fill data gaps
- Value summation for total cost of purified water and steam

Proxy measurements for small datasets

Chapter 4

- In depth study of PMD performances in dependence of the regression complexity
- Experimental rig setup for PMD data acquisition
- Proxy meter development with linear and non-linear algorithms
- Application of small dataset improvement modifications

Figure 1.4: Thesis chapter layout and methodology

Chapter 2

Literature review

2.1 Smart Manufacturing and Industry 4.0

The aim of Smart Manufacturing and Industry 4.0 is to integrate product systems both vertically and horizontally [23]. Vertical integration promises to revolutionise all production steps – from order processing and resource management through to product manufacturing and delivery [28]. Horizontal integration allows the linking of different companies within the value chain. The result is smart factories that provide higher flexibility in the fulfilment of individual customer requirements.

Smart Manufacturing and Industry 4.0 are similar concepts that have originated from global policymakers. While the terminologies differ, they share the same overarching vision of implementing digitalised and data-driven smart factories. Smart Manufacturing originates from the U.S. Department of Energy. The Smart Manufacturing Leadership Coalition [24], a coalition of companies, universities, manufacturing consortia, and consultants, defines Smart Manufacturing as "the intensified application of advanced intelligence systems to enable rapid manufacturing of new products, dynamic response to product demand, and real-time optimization of manufacturing production and supply chain networks". Industry 4.0 originated as a flagship program from the German government and has had an increasing impact on European policy. Industry 4.0 was first announced in 2011 and is seen as the fourth industrial revolution [29]. Three previous revolutions have led to significant changes in manufacturing surroundings (Figure 2.1): The first industrial revolution brought water and steam power to production areas, the second revolution led to mass production through assembly lines, and the third revolution increased the automation through use of robots and information technology (IT). The introduction of communication systems in manufacturing companies is seen as the fourth industrial revolution that will lead to interconnected machines and production processes that communicate with each other [28]. In contrast to Smart Manufacturing, according to Culot et al. [30], no agreed definition for Industry 4.0 currently exists. The authors outline that this lack of definition is due to the complexity and multidisciplinary nature of Industry 4.0 as well as the ongoing developments of the "announced revolution" [30].

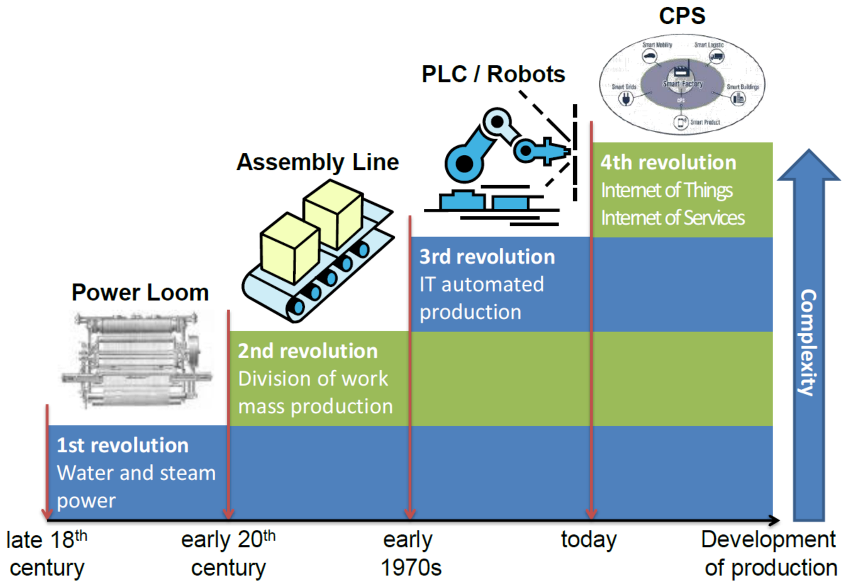


Figure 2.1: Four industrial revolutions in manufacturing systems, taken from [23],with PLC: Programmable logic controller, CPS: Cyber-Physical system

Both schemes (Smart Manufacturing and Industry 4.0) are expected to change current automated manufacturing systems towards intelligent manufacturing systems. In the following paragraphs, solely the term smart manufacturing is used in reference to the Smart Manufacturing and Industry 4.0 scheme.

Important principles that enable communication between machinery are the Internet of Things (IoT) alongside Cyber-Physical systems [31]. By using these, physical units are equipped with intelligent sensors, radio-frequency identifications, and microprocessors to allow automated data collection, data evaluation and processing, and data exchange with other units or systems. The product under manufacture is, therefore, able to control its own production process. The increased data volume is feeding enhanced process analytics as well, e.g. in the form of machine learning or the wider field of artificial intelligence. Besides, digital manufacturing is enabling more precise predictions and control through the simulation of the entire manufacturing process [32]. A general concept of a smart manufacturing enterprise is depicted in Figure 2.2 [33]. The system in Figure 2.2 is comprised of two basic layers; the manufacturing equipment layer and the cyber layer. Both are linked through an interface. Data is sent from the manufacturing equipment to the central cyber layer where it is analysed. Based on this data flow, production decisions are sent back from the cyberspace or made locally by the manufacturing equipment through use of artificial intelligence.

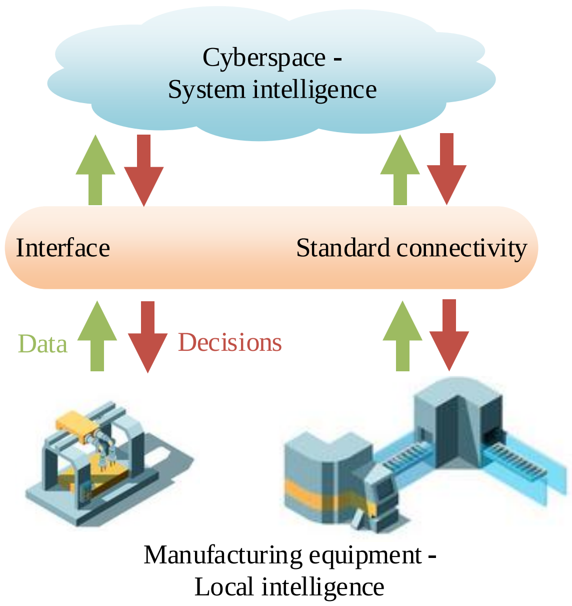


Figure 2.2: General concept of a smart manufacturing facility, adapted from [33]

The possibilities of intelligent manufacturing processes are extensive. Increased monitoring leads to the availability of all relevant information in real-time, shorter reaction times during failures, and optimal usage of resources in all process areas [34]. Due to the deep fusion between IT and manufacturing, an increased amount of process data has to be collected and analysed, as reported by Tao et al. [20]. This huge volume of data progressively elevates the degree of manufacturing smartness in the form of improved process control and flexibility. Furthermore, the increased digitalisation of manufacturing lines will be used by data-driven modelling approaches to further optimise production [33]. This will allow quicker responses to changing product requirements and raw material supply. Thereby, facilitating the fulfilment of individual customer demands with small production numbers or one-off items. In Figure 2.3 the relationship between variety and volume per model is displayed for the different manufacturing paradigms. Smaller volumes per model and increased variety are realisable with decoupled, fully flexible and highly integrated smart manufacturing lines [8], [35]. Together with a decrease in product defects, the increased variety in manufacturing leads to rising customer satisfaction. Besides the possibility of individual production units, the overall production rate can be increased in a smart manufacturing production line. It has been shown that productivity increases of up to 50% have been achieved in companies that introduced smart manufacturing [36]. The benefits of smart manufacturing processes are not solely limited to production lines. Kagermann et al. [29] expect that a wide range of

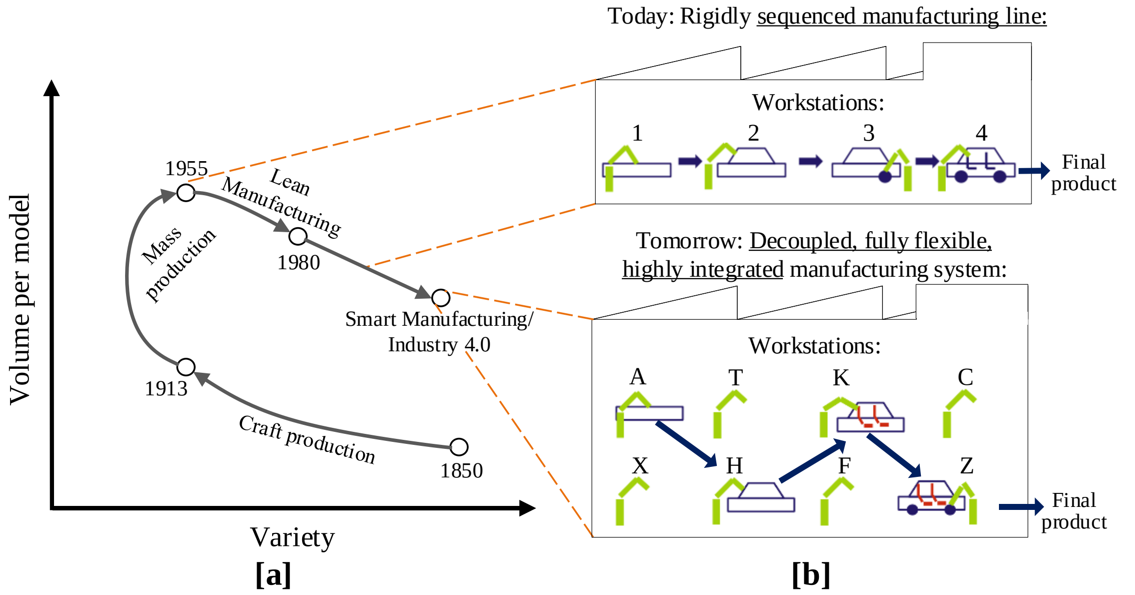


Figure 2.3: Volume per model to variety relationship in maufacturing paradigm [a], adapted from [8]; Manufacturing assembly paradigm [b], adapted from [35]

companies are going to offer innovative business models and services around smart manufacturing. High-quality machines and products will be needed to satisfy demand. Additionally, integrated system solutions will be required to interconnect the different machines and complex applications. These services will accelerate the introduction of smart production lines by manufacturing companies since they can receive coordinated systems and support. With reference to Germany, it is expected that approximately 100,000 new jobs will be created around the smart manufacturing sector within the next years [36].

2.1.1 Impediments and challenges of smart manufacturing

Despite all the advantages of Smart Manufacturing and Industry 4.0, some experts have reservations about the forecasted changes in manufacturing systems. Schütze et al. [37] view the upcoming transformation as being mainly politically driven to establish new business models. In a publication by Fuchs [38], the labelling of the upcoming changes as a technological revolution is criticised since the technological and economic developments have just started and smart manufacturing simply represents the continuing progress to achieve better knowledge and control of production systems. Furthermore, Fuchs [38] expects the digital automation to face an antagonism between profit and human interests: The capitalistic imperative to increase profit and reduce costs will result in increased technologically induced unemployment of workers and a loss of production control. This might go hand-in-hand with an advancement of capital concentration and monopolisation since the adoption of smart manufacturing approaches requires huge investments that cannot be afforded by many small and medium-sized enterprises (SMEs).

From a company viewpoint, there are numerous challenges when it comes to the practical implementation of smart manufacturing concepts. In many cases, companies have difficulties in adapting to the new technologies and finding the right specifications for their use [39]. As a prerequisite, smart manufacturing requires the collection of sufficient process data [40]. Yet companies often struggle even at this initial stage, not knowing what to measure or when [25]. Accordingly, manufacturers need support in terms of what type of data should be measured, with which sensor and where to install these sensors. Li and Kara [41] reported that there are not sufficient measuring concepts readily available to assist manufacturers. Besides the impediments to smart manufacturing outlined above, a recent survey between researchers and industrial professionals showed that the absence of practical metering guidelines also inhibits the move towards sustainable manufacturing [42]. Kusiak [43] stated that even when sufficient data can be collected from a process, many companies do not know how to interpret the gathered data and how to use it to identify process and product improvements. A big volume of data does not lead directly to knowledge gain, rather the correct analysis of this data is important to obtain hidden information. Experts see the manufacturing industry unprepared for the coming changes [44]. The following five necessities have been identified for the adoption of smart manufacturing philosophies [25]:

- Adopt strategies for data gathering and sensor placement
- Improve data collection, use and sharing concepts
- Design predictive models that show the feasibility and advantages of new machining technologies
- Study general predictive models that are able to handle uncertainties or data errors
- Connect factories and control processes through open interfaces and universal standards to guarantee a more dynamic and open manufacturing environment

To meet these requirements, closer collaboration between academia and industry is needed. In many cases, technological improvements are made in a lab-scale environment by academics without the consideration of how they can be integrated into an industrial environment [15], [25]. Future development of assessment frameworks for manufacturing processes should be evolved cooperatively between industry professionals and researchers, as reported by Bhanot et al. [27]. To enhance the collaboration between academia and industry, one possible solution could be online platforms where companies upload their problems to find suitable experts for solving them. Additionally, more governmental support for smart manufacturing is needed. This could be in the form of smart manufacturing working groups to bring together representatives from government, academic, and industry [22]. Another form could be the financial support of companies, i.e. by giving an investment tax credit for updating machinery or offering training courses for manufacturing staff [45].

If companies do not overcome the challenges in adopting smart manufacturing processes, they face the risk of falling behind the technological progress worldwide and might lose competitive ability [28].

2.1.2 Adoption concepts for smart manufacturing

This section outlines a number of approaches for adopting smart manufacturing concepts through the exploration of four example publications that focus on monitoring within smart manufacturing environments.

An integrated solution for sensor affixed machines, data acquisition modelling, and data analytics is presented by Balaji et al. [46]. The objective of this approach is to improve monitoring and to infer insights on the efficiency in the shop floor area. The presented framework is divided into different layers that include sensor measuring at shop floor level, progressing the signals through a data communication node, and transferring the data to analytic and visualisation tools. During data processing, harmful machinery situations are automatically detected and actions taken to prevent damages. The shop floor operators and managers are notified via optical or mobile information. The data analysis tool allows the comparison of different trends and calculates key performance indicators like the overall equipment effectiveness (OEE). While the framework by Balaji et al. seems useful to visualise and monitor relevant process parameters, it is similar to a variety of commercially available energy management software e.g. eSight Energy [47].

Li and Kara [41] presented an alternative approach to monitoring in a smart manufacturing surrounding by using wireless sensor networks. The concept enables the connection of multiple hardware and software platforms by first designing the system architecture and determining the selection criteria for each component. Similar to the previous approach by Balaji et al., the system layout is based on four layers that include data acquisition and transmission, data processing, cloud storage and analytics, and visualisation (Figure 2.4). For the selection of different components in each layer, the authors give decision support by summarising the relevant criteria for suitable products and services. A case study for metering and monitoring of temperature in an office environment shows a possible use case. This publication is an example of bridging the gap between academic research and easy-to-adopt methodologies for the industry. By explaining the system layout and providing decision support in every layer the architecture is valuable for manufacturing companies attempting to set up an advanced monitoring system.

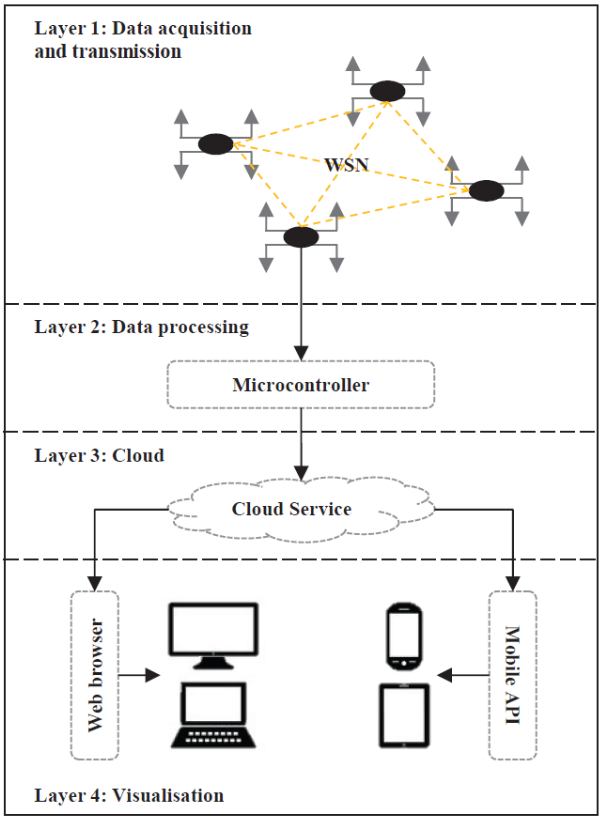


Figure 2.4: Wireless Sensor Networks architecture for manufacturing environments, taken from [41]

A further contribution to monitoring in smart manufacturing facilities is presented by O'Donovan et al. [22]. The authors illustrate a strategy for utilising industrial data analytics applied to equipment maintenance. The proposed methodology uses an open and extendable system architecture that interacts with smart devices on automation networks and offers analytics based on the process data. Stemming from the requirements of the architecture layout, the proposed system features high resilience to failure, dynamic scalability, and uses open standards to promote data accessibility. The methodology is built up of six stages, as indicated below and in Figure 2.5:

1. A cloud-based site management program acts as a central repository of metadata.
2. Ingestion software agents ingest data from maintenance equipment applications, such as sensors, and store it in the cloud when instructed by the management program.
3. The raw data in the cloud is buffered in a message queue until it is handled by data processing components.
4. The notification and handling of newly available datasets between the message queue and the data processing components is controlled by the subscription service.
5. Raw datasets are then processed by components to transfer the time series data into a suitable format for analysis.
6. Finally, the data access stage provides an open interface to the processed data to be analysable by suitable tools.

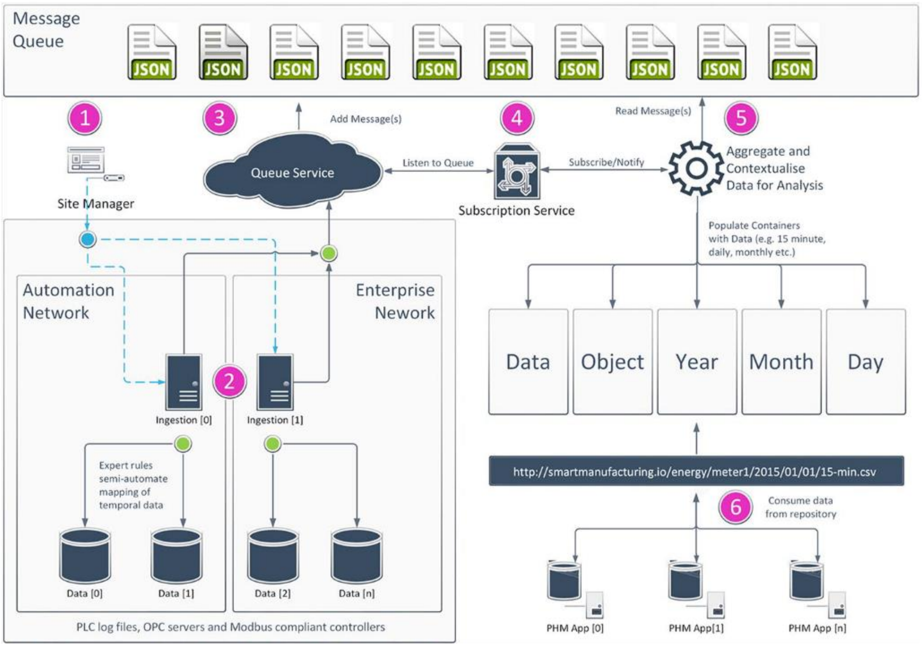

Message Queue
JSON
1
3
Add Message(s)
4
Read Message(s)
5
Site Manager
Queue Service
Listen to Queue
Subscription Service
Subscribe/Notify
Aggregate and Contextualise Data for Analysis
Populate Containers with Data (e.g. 15 minute, daily, monthly etc.)
Automation Network
Enterprise Nework
2
Ingestion [0]
Ingestion [1]
Data
Object
Year
Month
Day
Expert rules semi-automate mapping of temporal data
http://smartmanufacturing.io/energy/meter1/2015/01/01/15-min.csv
6
Consume data from repository
Data [0]
Data [1]
Data [2]
Data [n]
PLC log files, OPC servers and Modbus compliant controllers
PHM App [0]
PHM App[1]
PHM App [n]

Figure 2.5: System architecture for smart monitoring of equipment maintenance, taken from [22]

The methodology for setting up a smart monitoring system that is presented in the publication by O’Donovan et al. represents a further approach for transferring production facilities into a smart manufacturing system. Due to the structures that are used in the system architecture, it is particularly suitable for large sites that have to deal with large data volumes.

Yen et al. [48] presented another framework for smart monitoring and diagnosis of manufacturing systems. The idea behind the methodology is to provide software as a service to manufacturing facilities to support and standardise the monitoring and fault diagnosing process. The software contains different modules, such as a process knowledge database, data processing service, diagnostics, data storage management, and end-to-end security. Input by the manufacturer is needed in the form of sensor data and metadata. The raw data is then sent to the cloud where it is processed by different modules. A range of different tools are available for analysing of the received data, e.g. anomaly detection, pattern mining, and periodicity analysis. Additionally, image and video files can be analysed for fault detection. The used algorithms are first trained on a sample set of data to be able to detect abnormalities during the later monitoring task. The presented framework shows a possible solution for superior monitoring activities in a smart manufacturing environment. In particular, SMEs can benefit from this service as they often have limited experience / knowledge in analysing gathered data [49]. However, the authors state that the correct identification of abnormalities decreases in accuracy with increasing system complexity. Additionally, companies might hesitate to give away sensitive process data to external service providers due to data protection issues.

To conclude, an increasing number of monitoring concepts for smart manufacturing environments can be found in the literature. As demonstrated in the chosen case studies, the application of a suitable framework can improve industrial process data gathering and systematic analytics. Nevertheless, the application by industrial users has to prove if the available frameworks are also transferable to a variety of system configurations with different technical conditions.

2.1.3 Industrial state-of-the-art monitoring systems

The latest generation of commercially available monitoring systems for manufacturing plants already incorporates some of the technologies required for a smart manufacturing environment, e.g. these systems are capable of gathering large quantities of data in real time while providing advanced process control [50]. The implementation of these state-of-the-art monitoring systems can provide a high level of transparency and process understanding that is typically only available in research labs. As outlined by O'Driscoll and O'Donnell [51], a general plant-wide metering system for monitoring energy and resource flows is build up by three different levels of metering devices. At the facility level, meters are installed to monitor the energy and resource consumption of the entire facility. These meters are typically used to obtain aggregated utilisation data with low sampling rates and allow high-level analysis, such as the query of monthly bills. The second level enables the monitoring of process / value streams by applying metering devices with a higher sampling rate compared to the facility level measurements. These meters enable measurement of a process line's energy and resource consumption and provide detailed information that can be used to optimise production procedures towards higher sustainability or to facilitate benchmarking against other production areas. The third level of metering devices monitor the energy and resource consumption of a single machine / process step. Typically, metering devices at this level show high accuracy and resolution to facilitate performance optimisation, ascertain consumed resources per produced part, or to investigate equipment malfunctions.

The data from all metering devices is then gathered in a process or facility-wide monitoring system. There are two types of commercially available monitoring systems; supervisory control and data acquisition (SCADA) system and distributed control system (DCS). According to Karnouskos and Colombo [16], both systems enable process monitoring and control activities and are set up by a communication structure that can connect remote terminal units (RTUs), programmable logic controllers (PLCs), computers, and human-machine interfaces (HMIs). Although SCADA and DCS are composed of similar components for process monitoring and control, they differ in their intended use cases [52]. SCADA systems are orientated towards data acquisition, are mostly event-driven, and have been designed to monitor and control large process facilities. DCS are more process control orientated, run sequentially, and are more suitable for smaller-scale monitoring. Nonetheless, authors in the field see an increasing integration between SCADA and DCS in current and future developments [16], [53]. Both systems, in a separate or united form, will be a key component in a smart manufacturing environment. Future SCADA / DCS systems will have to cope with significantly higher amounts of diverse data and information in real-time, facilitate human-machine interactions from anywhere at any time, and provide improved scalability of the monitoring system [52].

The latest developments of commercially available SCADA / DCS systems already incorporate some aspects of future needs. Siemens AG released a DCS labelled *Siemens SIMATIC PCS neo* in 2019 [50]. The IT structure of this system is entirely web-based to allow immediate and secure access for monitoring and control of plants which can be distributed all over the globe. It incorporates a high degree of flexibility through a freely scalable architecture that permits plug-and-produce integration of automation packages. A SCADA system with comparable features has been released by Honeywell International Inc, labelled as *Honeywell Experion Elevate* [54]. As a cloud-enabled system, remote access is possible in addition to improved scalability and data analysing techniques. Similar systems have been developed by ABB [55] and Yokogawa [56].

The integrated features of the latest SCADA / DCS systems pave the way towards a new generation of process monitoring and control. Nonetheless, further developments of these systems are required to provide all the needs of smart manufacturing. Karnouskos and Colombo [16] envision that the next generation SCADA / DCS will be entirely cloud-based with a flat hierarchical structure that includes further information from other systems such as enterprise resource planning (ERP) and manufacturing execution system (MES), outlined in Figure 2.6. This will enable the collaboration of people, devices, and processes in the monitoring and control of manufacturing systems with an unprecedented scale of process performance assessment, evaluation of alternatives, and adjustment of resources.

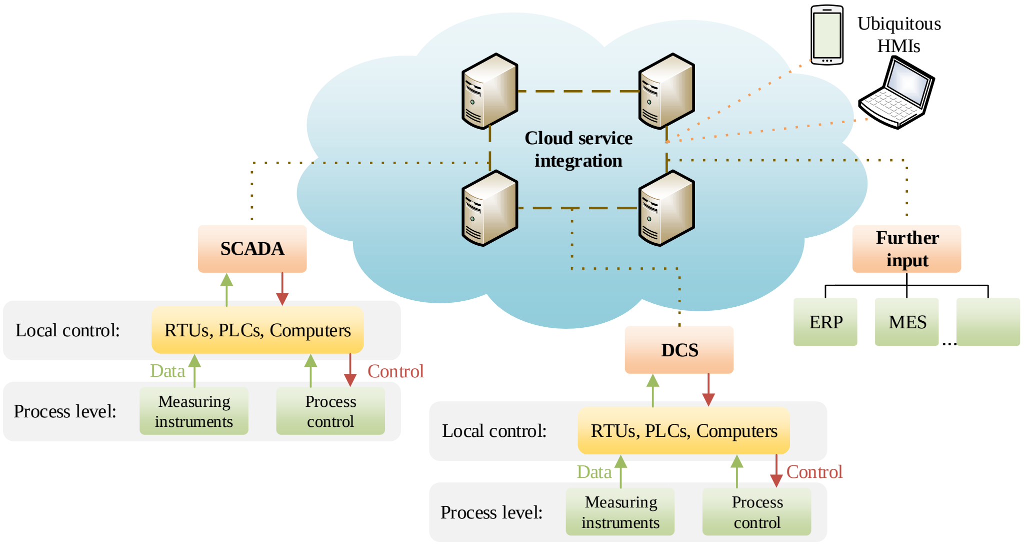


Figure 2.6: Vision of the next generation SCADA/DCS system, adapted from [16], [52]

2.2 Industrial resource metering

Industrial activities rely heavily on energy and natural resources, such as water, gas, and oil. This section underlines the importance of water and energy for manufacturing facilities by focusing on purified water (PW) and steam systems. Additionally, a holistic investigation method for water and energy is presented in form of the water-energy nexus. Finally, existing metering audit strategies for capturing the resource consumption of industrial systems are explained.

For clarification, the term ‘industry’ used throughout this research refers specifically to the manufacturing industry.

2.2.1 Water in manufacturing systems

The industrial sector is heavily dependent on water. The sector accounts for 19% of the globe's total water usage [5] which makes it the second highest consumer of water [57]. Water consumption by the industrial sector varies significantly within different regions. In Europe, the industrial sector consumes close to 40% of the total water extracted [58]. The current and estimated freshwater consumption of different sectors is depicted in Figure 2.7 [59]. These numbers account for water withdrawn from natural sources, such as lakes or groundwater, and do not include water reuse.

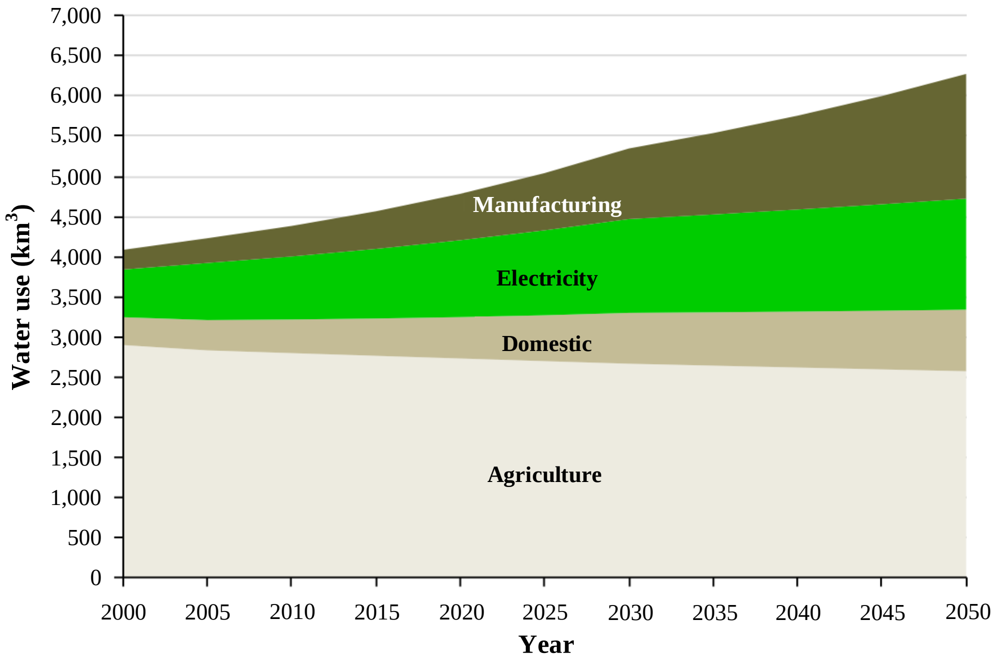


Figure 2.7: Estimated increase in freshwater demand by sector, adapted from [59]

A report from the Organisation for Economic Co-operation and Development (OECD) [59] highlighted that the demand for water of the manufacturing sector will increase further as more countries become industrialised and due to an increasing need for manufactured products. A water consumption increase of 400% is expected within the manufacturing sector on a global scale between 2000 and 2050, as shown in Figure 2.7. This is the highest growth compared to other sectors. The current usage and projected increase in water consumption shows the importance in raising awareness within manufacturing companies to place more importance on water efficiency. This is further driven by reasons that include [1], [60]; a rising water shortage in many countries due to climate change, stricter regulations on water use and drainage by governments worldwide, and competitive economic advantages due to the conservation of water and energy. According to Herrmann et al. [8], this need is being further strengthened from a growing consumer demand for products with minimal environmental impact.

The increasing importance to minimise water consumption in manufacturing systems is seen in the research field as well. A variety of different approaches have been published to decrease water consumption and increase the efficiency of manufacturing processes and technical building services (TBS) [61]–[63].

Hesselbach et al. [64] refer to TBS as equipment that is responsible for providing necessary utilities to the manufacturing processes to fulfil their designated process, e.g. steam and PW. This includes the generation, circuitry, and conditioning of the essential utilities. Accordingly, TBS ensure that production environment conditions are achieved and maintained [65]. The operation of TBS itself requires resources and energy, e.g. mains water, electricity, and natural gas. Figure 2.8 illustrates examples of TBS for a manufacturing environment.

In manufacturing systems, water is required at various locations. The majority of water is not directly used during production, but for TBS [66]. Therefore, the work presented within this thesis is focusing on two TBS that require large amounts of water in manufacturing facilities: PW and steam system. The following two sections outline these systems in more detail.

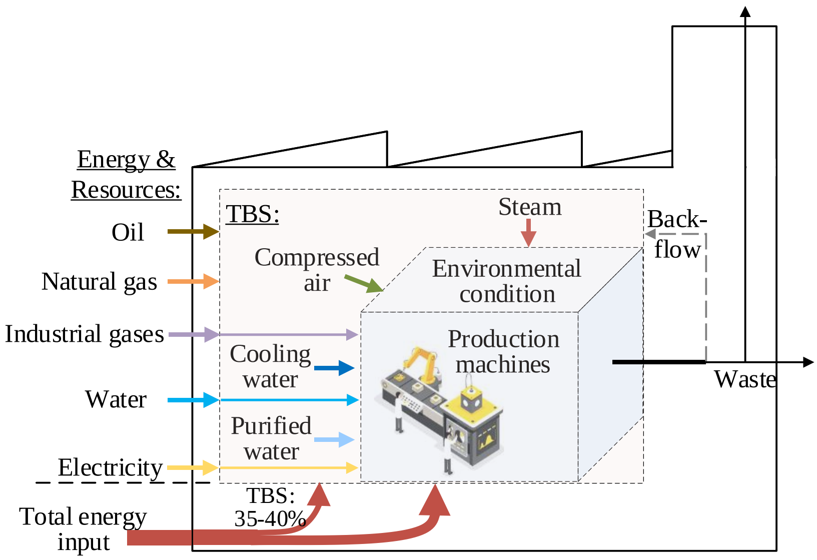


Figure 2.8: Example of interconnections between energy and resources, TBS, and manufacturing systems, data taken from [64]

Purified water systems

Water purification is a treatment process to guarantee high-quality water and is common for the manufacture of medical, healthcare and semiconductor products [67]. The term PW stands for a water quality category where almost all chemicals and contaminations have been removed. The requirements for the PW classification are subject to published standards that are regulated by regional organisations, i.e. by the European Pharmacopoeia Commission for member states of the European Union [68].

Depending on the source, water contains biological and inorganic matter in suspended, colloidal, and / or dissolved forms [69]. Membrane-based technologies remove particulate, organic, ionic and gaseous contaminants economically without the requirement of hazardous chemicals. Most of the industrial water purification processes are build up as hybrid systems that integrate one or more membrane processes in addition to chemical and electromagnetic radiation units. A typical layout may use microfiltration, ultrafiltration, and / or reverse osmosis as membrane processes in combination with water softeners and electromagnetic radiation, i.e. ultraviolet (UV) light. An example setup is shown in Figure 2.9. The characterisation of the single units in addition to the required operating conditions and maintenance is given in Table 2.1.

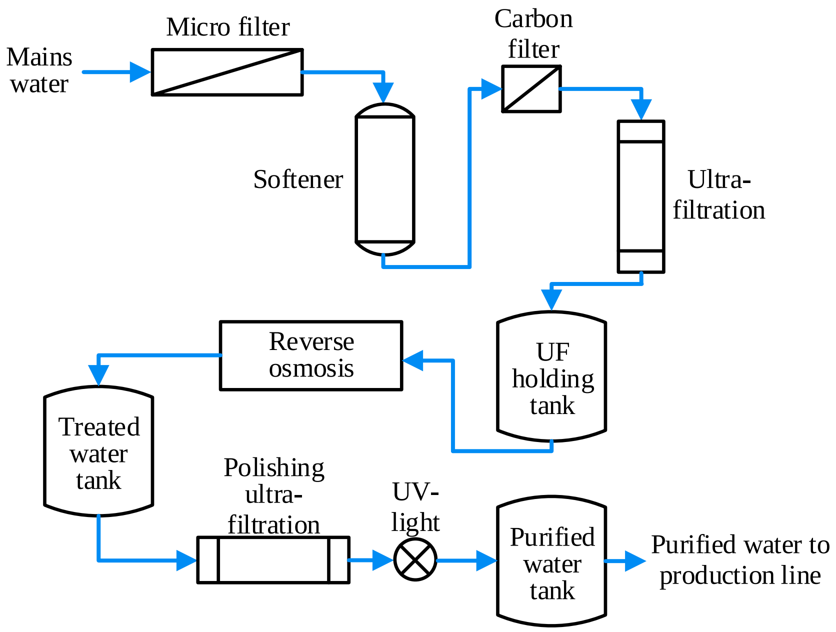


Figure 2.9: Common schematic of industrial PW systems, inspired by [70]

Table 2.1: Common water treatment units in PW systems, taken from [69], [71]

	Water treatment unit				
	Micro-filtration	Ultra-filtration	Reverse Osmosis	Softener	UV light
Process type	Physical: Membrane			Chemical	Electro-mag.
Treated impurities	Solids (100-1000 nm)	Solids (1-500 nm)	Solids (0.1-1 nm)	Minerals	Bacteria Virus TOC
Operation demand	p = 1-2 bar	p = 2-5 bar	p = 15-100bar	p = 1.5-6bar	elec. Energy
Maintenance + Replacement	Periodical sanitisation				
	Membranes			Backwash salt; Resin	UV lamp

TOC: Total organic carbon

Steam systems

Steam systems are part of many industrial plants. Within the manufacturing sector, the largest amount of steam is used for process heating, HVAC (heating, ventilation, and air conditioning) systems, and to drive machinery [72]. The common usage of steam is due to its many advantages such as high heat capacity, high heat transfer efficiency, easy distribution, non-toxicity, and ability to provide heat at a constant temperature, as outlined by Nieuwlaar et al. [73]. Although sizes of steam systems in industrial facilities vary greatly, the design of these systems generally follows an overall pattern as depicted in Figure 2.10. Pre-heated water is fed to the boiler, where it is heated to form steam [74]. The steam travels along the distribution pipes to various processes where the released heat from steam condensation is used. If lower steam pressures are required, pressure reduction valves are implemented before the steam using processes. The condensate is collected in a condensate tank and recirculated to the boiler. An implemented economiser preheats the boiler feed water by capturing waste heat from the burned flue gases. To make up steam and condensate losses, water is added to the condensate tank (known as makeup water). Chemical treatment is required for the makeup water to remove impurities. Since some impurities remain despite the chemical treatment, the boiler is drained periodically in a process known as blowdown. This prevents the agglomeration of the leftover impurities which can otherwise affect the boiler operation.

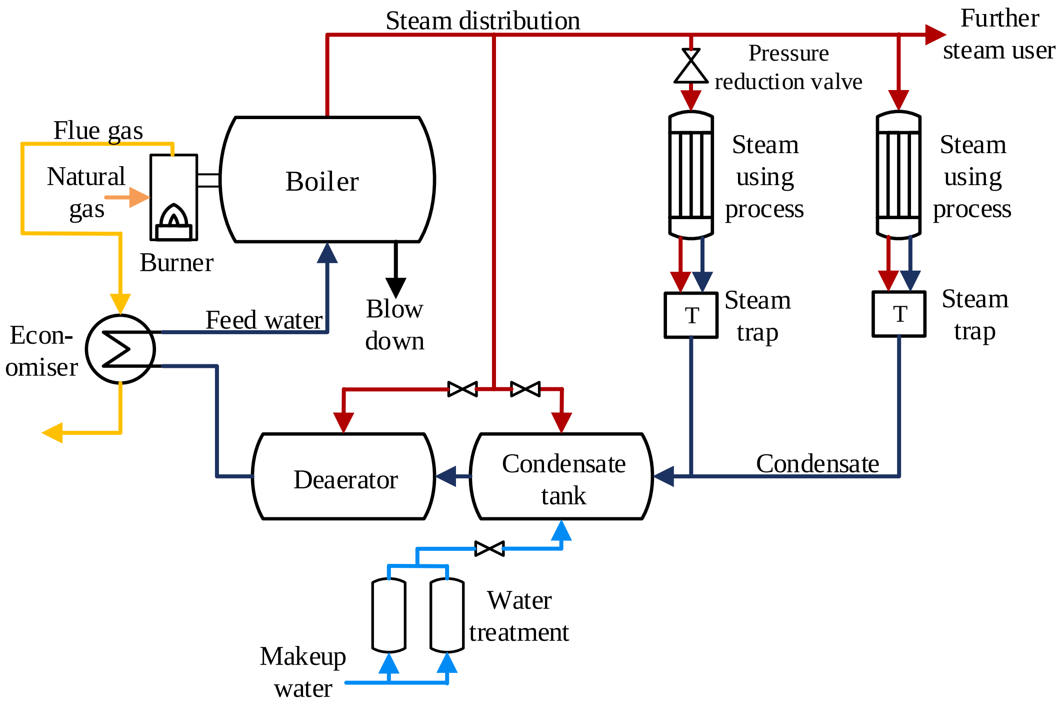


Figure 2.10: Common schematic of industrial steam systems, inspired by [74], [75]

The efficiency of steam systems

Boilers account for the vast majority of energy consumption within a steam system and hold a large share of the total energy consumption of a manufacturing plant. The energy efficiency for gas-fired boilers varies from 76% to 81 % (higher heating value) [76]. However, poor maintenance of the steam boilers can lead to an efficiency loss of up to 30% [74]. In general, periodical metering audits are seen as an important element to maintain the constant performance of the steam boiler. According to Barma et al. [77], this is not done on a routine basis in many places.

When looking at improving the efficiency of steam systems, many publications focus solely on the boiler due to its large amount of energy consumption [78], [79]. However, additional benefits and increased efficiency can be achieved when the entire system is considered. Hasanbeigi et al. [80] evaluated the possible efficiency improvements of steam systems that operate in industrial plants in China. Based on questioning experts and analysing the potential improvements and their cost-effectiveness, the authors concluded that 23% of coal energy could be saved in industrial steam systems in China. A key takeaway from this study is that existing steam systems offer a high potential for improved efficiency that is as yet unrealised. In a publication by Bhatt [81], the efficiency of various industrial steam systems was analysed. The author demonstrates that the highest inefficiencies are not coming from the steam boiler but from the heat losses in the steam distribution lines and unrecovered condensate. These segments of steam systems offer the highest scope for energy recovery and fuel savings. A single steam system of a potato starch production was studied by Bujak [79] with the goal of minimising energy losses. While calculating the energy losses of different steam system components, the author demonstrated that a significant efficiency improvement of the system could be achieved by preventing secondary evaporation through use of a pressurised condensate return. By applying this change to the analysed steam system, energy savings of up to 6% were achievable. Even greater were the savings on the makeup water where the introduction of the pressurised condensate tank reduced the added water required by up to 90%. In contrast to focusing on a single steam system, Therkelsen and McKane [82] looked at implementing the suggested efficiency improvements derived from more than 100 steam metering audits performed in US industrial facilities. The most common suggestions for improvements were to increase the amount of condensate return and to change the air-to-fuel ratio in the steam boiler. Re-auditing the plants after 24 months revealed that the majority of suggested improvements were not implemented. Cost concerns were cited as the primary reason for this; either the expenses were viewed as too high or the return-on-investment was too long.

In addition to large amounts of energy, steam systems require a supply of water. Although a vast amount of the condensate is typically reused in steam systems, the operation still demands substantial volumes of water. An analysis of the water usage of steam systems based on former steam metering audits in the U.S. has been presented by

Masanet and Walker [83]. From the results, the authors estimated that the amount of makeup water required to run U.S. steam systems is 376 million gallons per day, a similar amount to that required by the city of Los Angeles. In addition, a second publication by Walker et al. [84] modelled the water losses for different steam system components. The highest losses occurred due to direct steam injection, boiler blowdowns, secondary evaporation, and leaks. Although these approaches provide an insight into steam systems water usage, one major drawback is the lack of credible data and the resulting inaccuracies caused by estimations. According to Walker et al. [84], this is the reason for the small number of studies on the water use of steam systems. This absence leads to a high unawareness in the sector and excessive water consumption.

2.2.2 Water-energy nexus

To improve the water and energy consumption of manufacturing facilities, recent approaches are focusing on gaining a holistic understanding of industrial systems [61]. An important role herewith is the consideration of the water and energy relationship. According to the U.S. Department of Energy [85], the two resources are interdependent. All water system functions, apart from storage, are dependent on energy input, displayed in Figure 2.11.

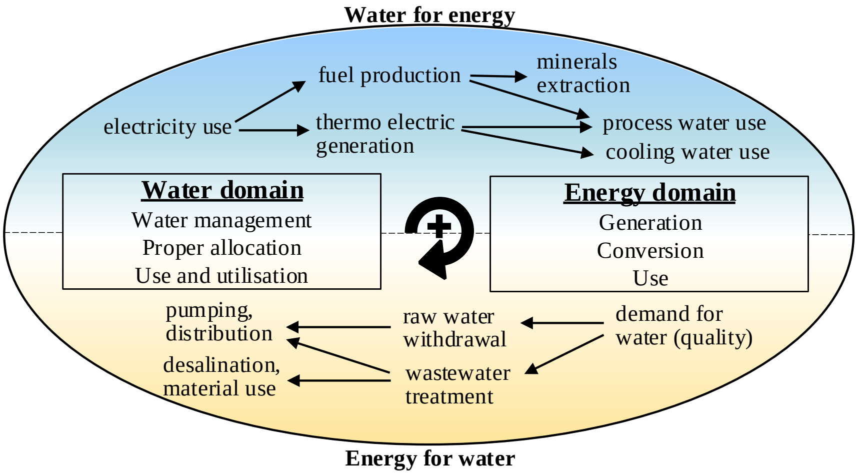


Figure 2.11: Water and energy relationship, adapted from [62]

Conversely, water is indispensable for energy production and transport needs. This relationship is referred to in the literature as the water-energy nexus. The importance of the linkage between water and energy is represented in the worldwide energy consumption: close to 7% of the energy produced worldwide is used by pumps for water distribution [86].

Historically, the two resources have been developed and optimised independently. Minimisation of energy has been a concern for manufacturing companies for many years because of energy's non-renewable status [87]. Until recently, many manufacturers were not concerned about the amount of water intake in their facility. Gleick [88] was the first to described the water-energy nexus for the industrial sector in 1993. Since then the topic has gained increasing attention and industrial relevance [89]. In various studies, it has been shown that industrial processes associated with water are one of the main contributors to the total energy demand of process chains [62]. To demonstrate this relationship, the impact of water conservation approaches on energy consumption has been displayed in a three-phase model by Novotny [90]. Figure 2.12 displays this model and highlights the required complexities and involved staff to achieve the water savings.

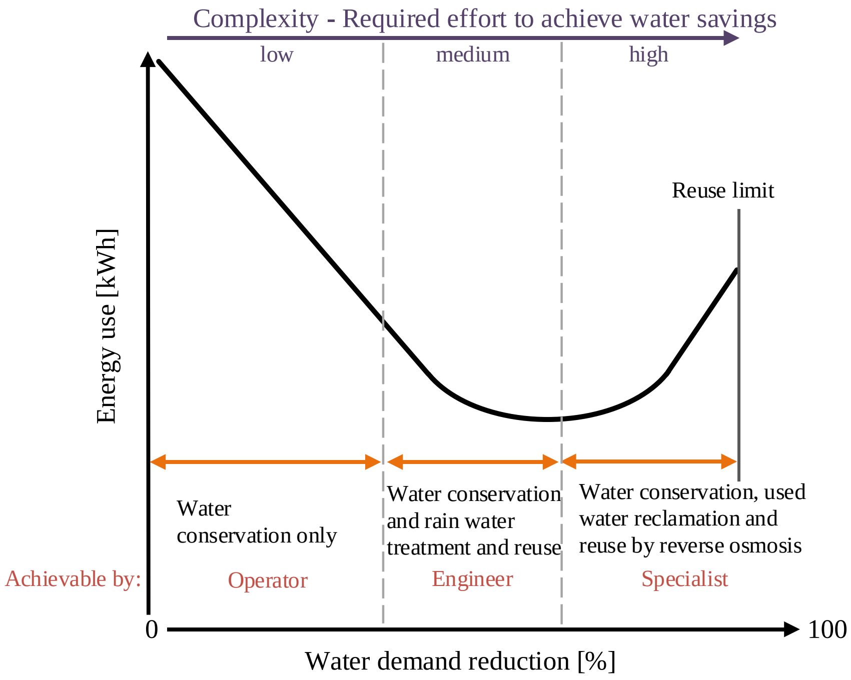

Complexity - Required effort to achieve water savings
low
medium
high
Reuse limit
Energy use [kWh]
Water conservation only
Water conservation and rain water treatment and reuse
Water conservation, used water reclamation and reuse by reverse osmosis
Achievable by:
Operator
Engineer
Specialist
0
100
Water demand reduction [%]

Figure 2.12: Three-phase model for water-energy demand reduction, adapted from [90]

Novotny suggests within the first phase of the model that the reduction of high water usage goes hand-in-hand with a proportional reduction in energy consumption. Further water demand reductions can be achieved in the second phase by using additional sources, such as rainwater. This leads to water conservations, although there will be less significant energy savings due to the need for additional treatment processes. In the third phase, advanced water treatment options, such as reverse osmosis units, have to be applied for an additional water demand reduction. As these technologies are energy-intensive, the overall energy demand increases.

Concerning the water-energy nexus in the manufacturing industry, a limited number of publications are available that explore this topic in a production setting, e.g. in [62], [89], [91]–[93]. Several approaches are based on modelling the water and energy relationship to pursue two objectives; decreasing water and energy consumption by considering the interactions between both resources and increasing the awareness of the water-energy nexus. The following section presents five recent publications that focus on water-energy simulations.

Modelling the water-energy nexus

Schlei-Peters et al. [91] used process functions and process factors to simulate the water and energy flow. For each process step, functions for process input and process output are determined and subsequently integrated into a dynamic system to create a material flow network. In contrast, two publications by Mousavi et al. [92], [93] identified that a more holistic approach is needed to improve the water and energy efficiencies in manufacturing facilities. By using a hierarchical classification with three levels (machine tool, process chain, and whole factory) similar to the one presented by Duflou et al. [1], the framework distinguishes five modules within these levels:

1. Process module
2. Steam generation module
3. Compressed air module
4. Production planning and control module
5. Integration and evaluation module (supervisory role in the simulation for coordinating activities, calculations and visualisation of the entire factory)

For validation of the introduced framework, a case study by Mousavi et al. [92] revealed the direct relationship between throughput, energy, and water consumption in a pharmaceutical plant. Different production parameters were considered for the concurrent minimisation of energy and water that resulted in resource savings of up to 6%.

A similar contribution to the simultaneous modelling of energy and water in a manufacturing plant was presented by Thiede et al. [62]. After identifying that most publications focus on one system-level only, the authors introduced five modules that are attached to three hierarchical levels. The modules exchange input and output data with each other, as depicted in Figure 2.13. To allow data exchange between the modules, the authors reviewed different coupling methods that are suitable for manufacturing environments, such as offline and direct coupling. The presented case study demonstrated the feasibility of the methodology by simulating the water and energy consumption of a manufacturing company. In the example, the base case scenario confirmed the importance of the water-energy nexus by showing that 69% of the total energy demand was used in the interaction of these two resources. Three improvement scenarios in the case study showed that energy and water demand can be reduced simultaneously. However, the results indicate amplifying and attenuating effects when several improvements are applied to the system.

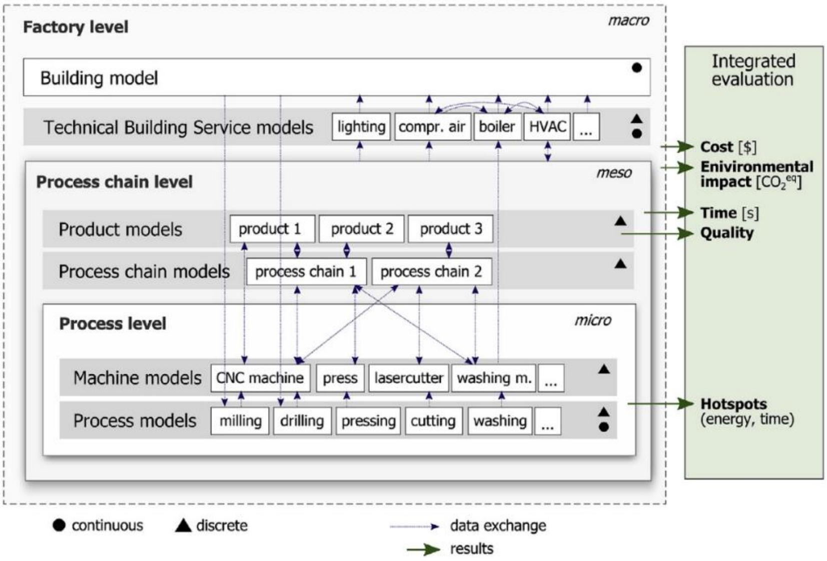


Figure 2.13: Framework for multi-level models to simulate the water-energy nexus, taken from [62]

By referring to the fact that transferable and systematic analysis methods are missing in modelling the water-energy nexus in manufacturing, Thiede et al. [89] made a second contribution to the research area. In contrast to the previous approach, the modules of the factory were situated on the same hierarchical level which simplified the coupling of the modules. Additionally, the transferability of the framework was improved by relying on common process components (pipes, pumps, heating, cooling, and wastewater treatment plant). In the selected case study, energy and water savings of up to 27% and 55% respectively were realisable in specific process scenarios.

In summary, a scant number of approaches for modelling the water-energy nexus of manufacturing systems can be found in the literature. Several authors, including Thiede et al. [62], Mousavi et al. [92], and Sachidananda and Rahimifard [94], point out that due to the importance of the topic and the savings that are possible through simulation, there is a need for more contributions in this area.

Critique on the water-energy nexus

While the water-energy nexus has developed into an important approach in the industrial sector, some researchers do not share the need for increased awareness. Williams et al. [95] argue that the enhanced focus on the water-energy nexus is mainly coming from an economical perspective with a view on business opportunity. While new technologies around the water-energy nexus promise continued growth, the nexus should lead to a fundamental change in thinking about the use of resources. Furthermore, the principle of integration within the nexus is seen as a panacea. Castree et al. [96] criticise this 'single concept of integrated knowledge' since this method is neither possible nor desirable for complex systems. Cairns and Krzywoszynska [97] see the nexus as a buzzword with the absence of a real definition and a strong belief by the users in the ultimate benefits of the approach. Buzzwords lead to narrow-minded views by influencing what is thinkable and doable. This often applies to academic research where publications follow existing policy approaches to secure funding and increase attention instead of being critical, challenging or innovative [98]. For the water-energy nexus, a risk exists that the significance of the term contributes to ‘creative rebranding’ of existing research with little novelty. The present dominance of the nexus in terms of business opportunities and knowledge integration is seen as worrying indications towards this narrow-minded trend [97]. To improve the current development of the water-energy nexus, Williams et al. [95] identified a need to integrate social sciences. Critical social concepts can reduce the dominance of the technocratic approaches by insisting on discrepancy and politicising the nexus. Likewise, Cairns and Krzywoszynska [97] recommend the integration of social science to question social and environmental justice and to support alternative practices and understanding.

2.2.3 Hidden costs of technical building services

The total cost of water

A further trend that incorporates the water-energy nexus is the evaluation of the hidden costs of water systems. The available approaches are aiming to raise awareness of the real cost of industrial water usage. From an economical perspective, water is known to be underpriced in many industrial environments [69]. A lack of understanding of the embodied costs for water processing and treatment exists in many places. According to Kurle et al. [99], this is due to the fact that water is often an invisible resource in manufacturing facilities, similar to electricity or compressed air. A knowledge gap exists about the functionality and the operation of many water systems [100], resulting in underestimated costs and over utilisation [60]. It is common practice to relate the water usage only to the water supply cost while neglecting the further resources and additional values that are necessary. However, water systems require more resources than water alone. Electrical energy is needed in various stages of the system as demonstrated by the previously outlined water-energy nexus. In addition to these resources, 'further added values' are necessary to operate the water system, as indicated in Table 2.1. Further added values are individual for every system, such as maintenance, drain flow treatment, or monitoring of water quality.

To date, the number of publications that address the total costs that result from the required resources and added values of industrial water is limited. Walsh et al. presented in three publications [100]–[102] a framework for addressing the total cost of water in an industrial facility. To raise awareness of the amount of water usage, the authors provide a structured guideline to analyse an industrial water system. First, major water flows are identified including water treatment and preparation steps. Second, every added variable is allocated a certain value that represents its cost per unit volume. An example of this value-added-table is shown in Table 2.2, where the shown variables represent the cost of the specific treatment per unit volume. The presented results illustrate that the total cost of the highest-value water flow inside a case facility is 14.05 times the amount of the raw water supply cost. This high value leads to a significant change when calculating the payback time of water optimisation approaches. Improvements to the water system would be profitable in a shorter time when referring to all water added values than just referring to the raw water supply cost [99].

Table 2.2: Example of a value-added table, adapted from [100]

	Value-added (€/m^3)						
Utility	Mains water or Overhead	Chemical treatment	Processing treatment	Energy consumed	Sampling/ Monitoring	System main-tenance	Equipment depre-ciation
Drinking water	M			E_d	S_d	SM_d	D_d
De-ionised water	M		P_{di}	E_{di}	S_{di}	SM_{di}	D_{di}
Chilled water	M	T_{ch}	P_{ch}	E_{ch}	S_{ch}	SM_{ch}	D_{ch}
Hot water	M	T_{hw}	P_{hw}	E_{hw}	S_{hw}	SM_{hw}	D_{hw}

A similar framework is shown by Leusder [103] which fixes on the expenses of an industrial water system. This work focuses on a water treatment system of a Singaporean brewery and illustrates how each individual driver affects the total cost of water. The main cost drivers in the case study were energy and capital investments. Compared to the entire operating costs, the total cost of water is within 0.4 - 0.5% which is negligibly low. The author concluded that the total cost of water offers the potential to function as an index parameter that allows easy comparison of industrial water systems and represents a facilities' dependence on water.

The publications by Walsh [100]–[102] and Leusder [103] can be seen as the first methodological contributions towards calculating the total cost of industrial water. Their work indicates that the total cost of industrial water is dependent on the required treatment steps and volumes. The total cost of industrial water can be further improved by providing a continuous cost calculation. This would represent an automated benchmark index that can show how the cost is changing over time and how system modifications influence this cost.

Hidden costs of steam systems

In parallel to water systems, resources and further added values are required for the operation of industrial steam systems as well. Steam is generated in a boiler by the combustion of energy carriers. Vast amounts of energy are required. In the USA up to 37% of the total fossil fuel consumption is used to produce steam [77]. Within the manufacturing sector, steam systems are the largest consumer of energy (single units). On a global scale, steam systems account for approximately 30% of the energy used in manufacturing facilities, as reported by Yang and Dixon [104].

Within Europe, the most common energy source for steam boilers whose size exceeds 1 MW_{th} was natural gas with a share of 70%, followed by oil (15%) and electricity (10%), according to a publication by the European Commission from 2016 [75].

In addition to thermal energy, varying amounts of water are required (makeup water) to replace steam or water losses and to produce directly injected steam [83]. Furthermore, electrical energy is needed in various stages of the system, e.g. to run pumps and boiler fans. Besides these resources, steam systems require further added values for operation, such as maintenance or makeup water treatment [84].

Similar to the total cost of water, the U.S. Department of Energy published a technical brief outlining a methodological approach for calculating the total cost of steam [105]. The department sees the evaluation of the total steam cost as a key parameter for the optimisation of steam systems which serves as an important driver for efficiency improvements as it visualises the real value that is embedded in steam [45]. For the calculation of the total steam cost, the required resources and further added values are incorporated. On the resource side, boiler fuel, feed water, and electricity are considered. As further added values, several expenses are added that include chemical treatment of the boiler feed water, system maintenance, and labour cost. When the methodology was applied in a case study, more than 90% of the total steam cost resulted from the boiler fuel, followed by electricity (3%), boiler feed water (2%), and wastewater treatment (<1%). As a conclusion, the authors point out that the steam generation rate can have a considerable influence on the steam cost. By considering resources and further added values for steam systems, this publication is an important approach for displaying the embedded value of steam. As a drawback, this method is designed for a single calculation of the steam cost by manual measurements. This impedes the monitoring of the cost over a longer period.

Internal and external costs

The added values to water and steam described in the previous section initiate the question of which costs industrial companies consider within their accounting. In general, costs for production facilities can be divided into internal and external costs [106]. Internal costs are directly monetisable for a product and include expenses such as resources, labour, or maintenance. In contrast, external costs, also referred to as environmental externalities, are effects that are not directly linked to the production or consumption of a good or service [107]. They represent the price of the business' impact on the environment and society [108]. In many cases, neither the market nor regulations assign external costs to the responsible companies, as depicted in Figure 2.14. The price for these costs is paid by society, e.g. by natural resources, losses in environmental quality, and / or global warming. Monetarising external costs is difficult since they are non-transparent [109]. However, with life cycle assessment a technique exists that strives to incorporate the environmental effects of products. Linke et al. [110] describe process life cycle assessment as a methodology that investigates the input and output streams of a product at different stages of it's life. In recent years, a shift in attitudes towards environmental protection increased the recognition of external costs that are revealed by life cycle assessments. Nonetheless, more effort is needed in recognising and reducing externalities since a failure to address these impacts today would lead to even higher costs for humanity in the future [111]. Besides higher costs, the inclusion of environmental externalities can lead to positive effects for companies. Regardless of regulatory mandatories, Shapiro [108] described that society and stakeholders appreciate businesses that consider and disclose environmental costs nowadays. Besides, international competitiveness can be increased by considering environmental impacts. New global regulations are often based on the highest environmental standards. Staying ahead of existing standards prevents companies from retrofitting designs and processes. However, many companies are still primarily focusing on minimising production costs and ignoring the external costs unless they are reflected in taxes [112].

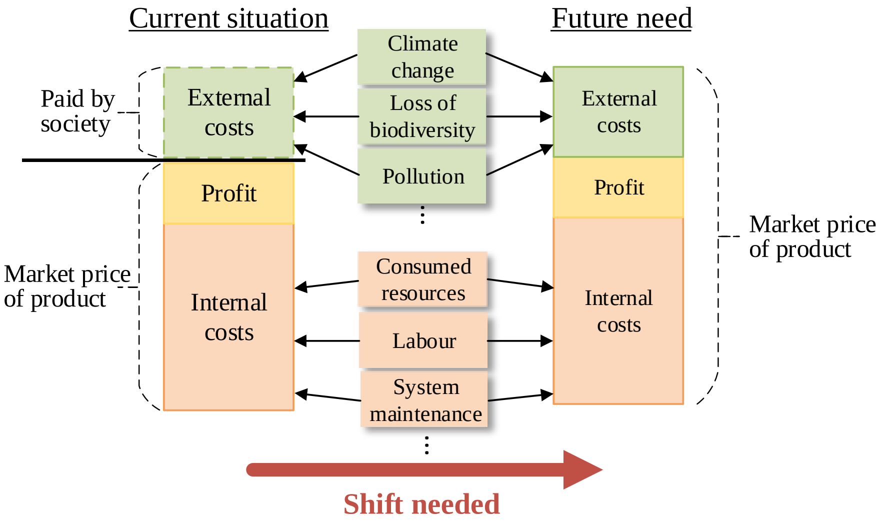


Figure 2.14: External vs. internal costs of a product

When considering the full cost of water in industrial surroundings, the external costs, such as any damage caused to return water, have to be included [113]. Similar to other resources, the monetarisation of the environmental impacts is difficult for water since the value of water is depending on the timing and reliability of water supplies. Rogers et al. [112] stated that many water prices nowadays do not even cover the full cost of supply nor do they take the external costs into account. This leads to inefficiencies and overuse of water in industrial facilities.

For steam systems, the fuel that is combusted in the boiler is the most required resource. In general, studies demonstrate that the combustion of natural gas, the most common energy carrier for steam boilers in the European Union, results in lower external costs compared to oil or coal [114]. Nonetheless, natural gas combustion still emits significant amounts of air pollutants that impact the surrounding society and environment in areas such as human health, crop yield, and climate change [111].

2.2.4 Benchmarking

Industrial water system benchmarks

To reveal the potential for optimisation within a water system, benchmarking measurements allow a simple comparison of water consumption within certain industrial sectors [115]. According to Hon [12], benchmarking is the continuous process of evaluating the performance of a system compared to the best in class performance of similar systems. For water systems, the detailed analysis of water intake in each process step gives comparative values of the water quantity used. Hereby, performance goals can be set and inefficient process steps can be identified and addressed in a subsequent optimisation approach.

To conduct a water benchmarking measurement, different approaches are available. In 2014 the International Organization for Standardization (ISO) presented a standardised framework for benchmarking the water consumption of a product during its lifetime in the form of the *Water Footprint* (ISO 14046) [116]. The concept has been applied by several authors in the form of life cycle analysis in manufacturing surroundings. Chen et al. [117] presented a methodology for assessing the water footprint of a machine tool. The authors concluded that the water footprint of a machine tool is dominated by the usage stage relative to the material for assembly, the manufacturing process, and the transport of the machine tool. Mousavi et al. [66] published a model to simulate the water footprint in manufacturing systems. The model is capable of taking both direct and indirect water usage into account and can differentiate between the different forms of virtual water. Figure 2.15 represents the used scheme for a generic manufacturing facility.

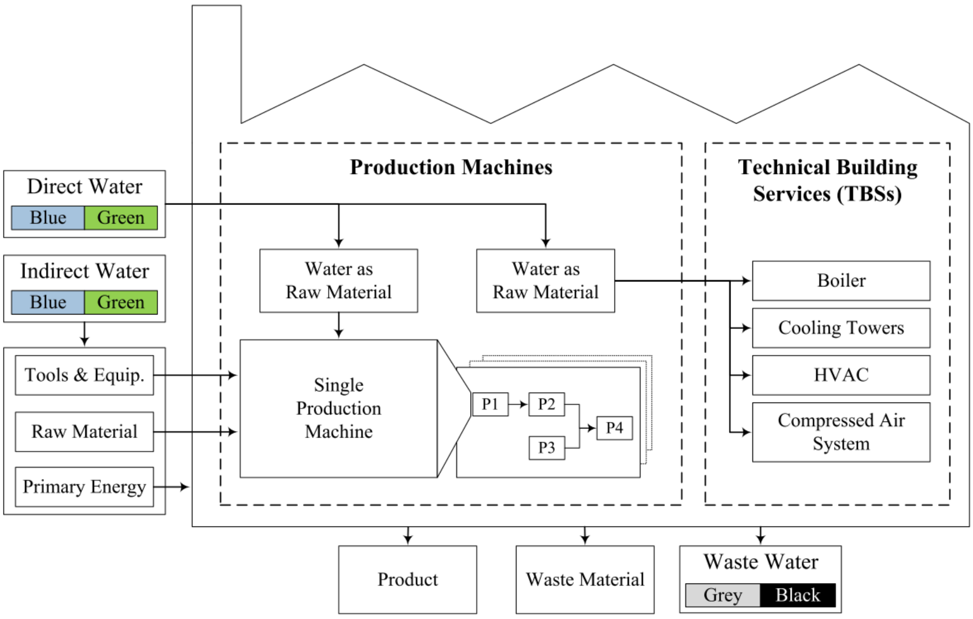


Figure 2.15: Schematic representation of an industrial water footprint, taken from [66]

In Figure 2.15, direct water stands for the amount of water that is directly used in the facility and indirect water encapsulates the total volume of water entering the factory through other inputs. The colours blue and green are referring to the origin of the used water; blue stands for fresh surface and groundwater, and green for water that is collected from rainfall. The colours grey and black stand for the volume of water needed to dilute pollutants to meet water quality standards, grey corresponds to mild pollution; black to heavy pollution. An application of the model in a pharmaceutical environment showed that only 22% of the water intake is converted to product, while the remainder is turned out as wastewater. A similar approach by Pham et al. [118] quantified water flows in industrial parks based on mass balances. Data is acquired through a questionnaire survey conducted in industrial parks in Vietnam. A water balance framework classified water flows in companies by withdrawal, effluent discharge, losses, use for production, and use for domestic. The results show the water consumption of nine different industry sectors and indicate that the textile and leather industry use and discharge by far the most water. The manufacturing sector is ranked fifth in terms of water intake and wastewater discharge. While the water usage characteristics are shown by the authors, one major drawback of this publication is that the water flows are not related to production outputs. It is not clear if high water consumption corresponds to large production rates or if they indicate inefficient water usage.

Besides the Water Footprint Standard, a variety of other benchmarks exist that are more specific to different manufacturing sectors. One example is the ‘Water Usage Effectiveness’ that measures the water used for cooling activities in data centres [119]. Another example of benchmarking water and energy usage within a manufacturing facility when considering the water-energy nexus was presented by Thiede et al. [89] in the form of the ‘Water-Energy intensity’. The benchmark describes the adopted energy content of water (in terms of thermal, kinetic, potential and chemical energy) compared to the total inserted amount of energy (embodied energy). The benchmark can be used for performance comparisons between factories or individual modules.

Industrial steam system benchmarks

For industrial steam systems, fewer benchmarks can be found in the literature compared to water systems. Up to now, no ISO standard exists for steam systems that is comparable to the Water Footprint presented in the previous section. The U.S. Department of Energy [120] advises calculating the fuel cost per generated amount of steam. According to the publication, this represents an effective way to assess the efficiency of steam systems. The parameter is dependent on the fuel type, fuel cost, boiler efficiency, feedwater temperature, and steam pressure. A similar ratio is presented by Bhatt [81] in the form of the steam to fuel ratio. This benchmark differs from the fuel cost per generated amount of steam parameter in that the cost of fuel for the boiler is not taken into account. Both parameters are an example of an effective way of characterising the performance of a steam generation system. However, they are limited in reflecting the performance of the entire steam system since the steam consumption side is not considered. To reflect the overall performance of a steam system, the total cost of steam parameter [105], outlined in section 2.2.3, is more suitable since it incorporates both steam generation and consumption.

2.2.5 Metering audits

A metering audit represents a cost-effective way to visualise resource flows and to identify inefficiencies by recording and quantifying the usage within a manufacturing facility [121]. For a detailed metering audit, all resources and further added values of the examined system must be recorded. The representation of the data in reports and graphs enables the identification of use patterns. For an enhanced visualisation of resource flows, Sankey diagrams represent a valuable tool by connecting related processes and indicating the quantitative information by line width [122]. Deviations from standard patterns can then be easily spotted and further investigated [87]. Benchmarks and Key Performance Indicators (KPIs) are useful during audits as reference points, for the control of activities, and to schedule maintenance procedures.

A key element for a successful metering audit is the gathering of sufficient process data from the examined system. According to Sachidananda and Rahimifard [94] three different methods exist for collecting data from an industrial process, summarised in Figure 2.16:

1. *Empirical methods*

Empirical methods are the most accurate and most common technique for data acquisition. Metering devices are installed in the process to measure the required parameter. Many meter technologies have evolved to cover a vast number of applications by directly or indirectly measuring the required process parameter. Common principals for measuring water flows are positive displacement flowmeters, variable area flowmeters, and electromagnetic flowmeters [123]. For steam flowrates, differential pressure and vortex flowmeters are widely used. Despite the high accuracy, empirical methods are costly, labour-intensive during meter installation and maintenance, and might require invasive access to the process.

2. *Theoretical methods*

By estimating flow rates and further parameters based on process knowledge and mathematical models (e.g. estimating the water volume in a tank by the tanks' dimension), theoretical methods are an alternative option in the absence of a physical measuring device. As a drawback, this method requires process knowledge and in some cases several assumptions which can result in inaccuracies.

3. *Statistical methods based on literature data*

The third method uses statistical calculations based on literature data (e.g. surveys for domestic water use). Their use in industrial systems is unusual and only applied if empirical or theoretical methods are not possible. Due to their generic character, they show high inaccuracies.

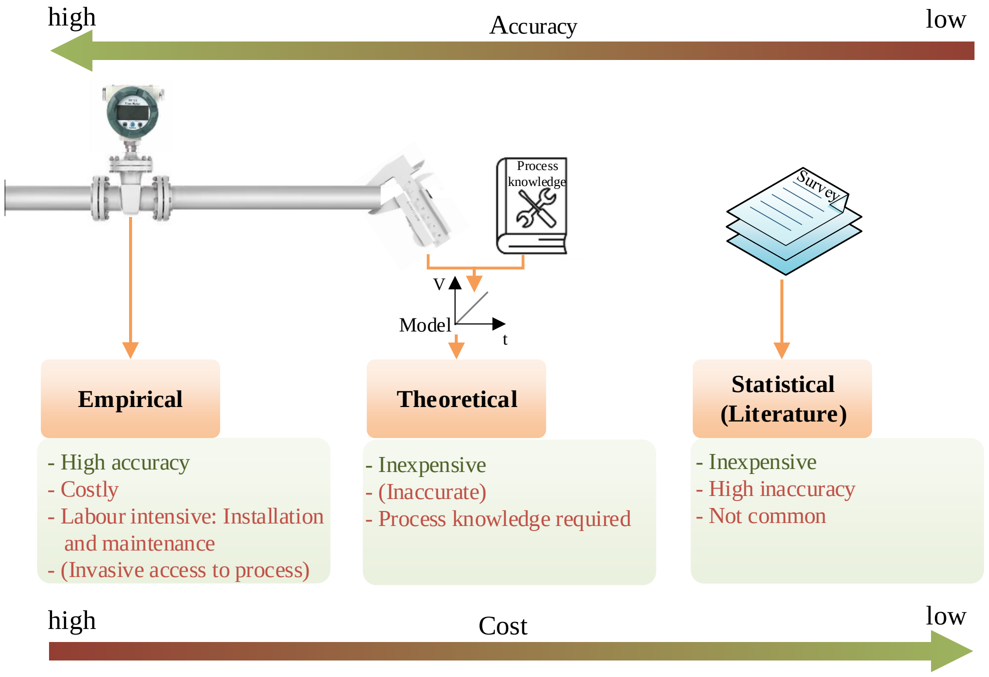


Figure 2.16: Data collection methods from industrial processes

In many industrial sites, the application of the three data acquisition methods is limited. This leads to a lack of data which is a barrier for metering audits, prevents process understanding, and system optimisation approaches.

Most manufacturing water systems are capable of measuring input and output flowrates at a factory level but do not provide an in-depth breakdown of water usage at process levels [94], [99]. Similar applies to steam systems, where detailed data about the steam generation and in particular steam consumption throughout an industrial plant are often missing. Several authors have identified this data gap as one of the key barriers to optimisation efforts for water systems [94], [99], [124], [125] and steam systems [45], [80]. Due to this knowledge gap operators do not know the potential for optimisation within their system or the corresponding cost saving, which leads to difficulties in identifying a starting point for resource minimisation efforts.

Even when enough metering devices are in place, continuous maintenance is necessary to guarantee accurate monitoring of the measured parameters and to replace broken instruments, as stated by Liu et al. [126]. Although new metering devices are becoming increasingly smart, e.g. with the ability for self-checks [127] or the correction of the instrument output when influenced by environmental variations, continuous maintenance is still required [14]. This maintenance is not done to the full extent in many industrial facilities due to the required effort and cost [128]. Missing and broken measuring devices prevent the accomplishment of metering audits and continuous monitoring of the process. Therefore, more practical metering strategies are needed from academia that take into account data gaps of industrial systems, as reported by Bunse et al. [26]. To date, research has been mainly focused around the efficiency and metering of energy [1], in particular of electrical energy, as outlined below. More attention is required to include all other resource streams of manufacturing processes [110].

For metering electrical energy consumption during production processes, several approaches have been published that take existing conditions in manufacturing sites into account. This can be achieved by different means including detailed methodologies for industrial implementation of an energy metering system [129], [130] and the analysis of consumed electricity of manufacturing machines during different modes [131]. Balogun and Mativenga [132] developed a three-phase model that provides a generic and robust estimation of the electricity consumption of a machining process. A more comprehensive metering strategy in the form of a two-step approach for life cycle analysis of machine equipment was presented by Kellens et al. [133], [134]. First, an initial screening, based on available process data and calculations, provides an insight into machining processes in terms of energy use, material, and possible system improvements. As second step, the gained knowledge is used to guide an in-depth metering approach to obtain accurate data and identify possible energy improvements. This is achieved by a time, power (energy), and consumables (materials, components, etc.) study. Within the scope definition of the in-depth approach, the authors outline the importance of defining clear system boundaries for the metering audit. These boundaries determine which unit processes and sub-processes (level of detail) are investigated, as depicted in Figure 2.17. The interaction with the surrounding technosphere, in terms of input and output flows to the selected system, have to be incorporated unless they do not significantly change the overall conclusion of the metering campaign. The functioning of the machining process is then isolated from the other elements of the general production system. The focus of the metering audit is on normal working conditions solely and neglects exceptional events. Nevertheless, the selection of suitable system boundaries can be challenging.

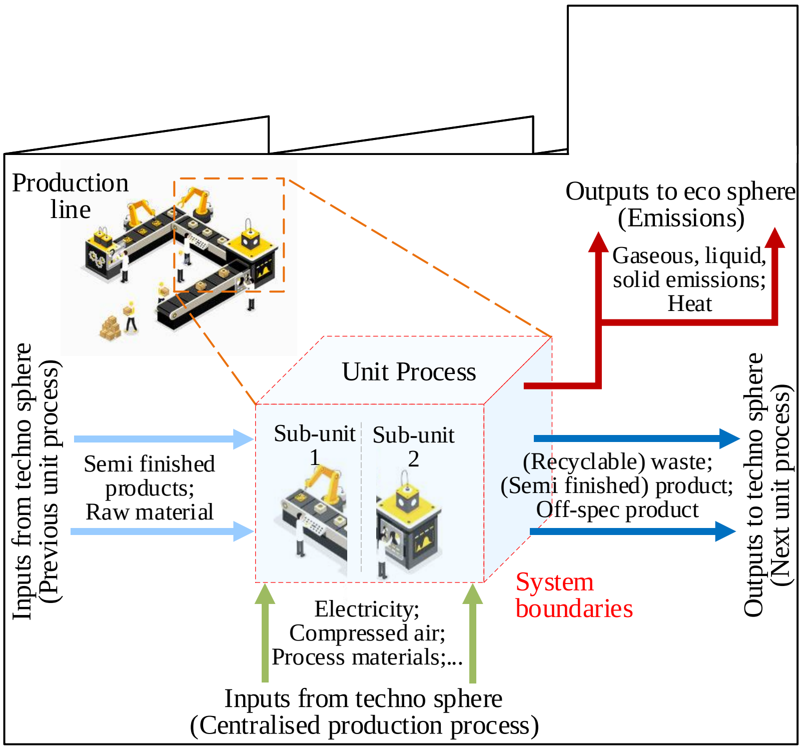

Production line
Outputs to eco sphere (Emissions)
Gaseous, liquid, solid emissions; Heat
Unit Process
Inputs from techno sphere (Previous unit process)
Semi finished products; Raw material
Sub-unit 1
Sub-unit 2
(Recyclable) waste; (Semi finished) product; Off-spec product
Outputs to techno sphere (Next unit process)
System boundaries
Electricity; Compressed air; Process materials;...
Inputs from techno sphere (Centralised production process)

Figure 2.17: System boundaries of a unit process, adapted from [133]

Rajemi et al. [135] demonstrated in their energy optimisation approach for machine tools the impact of different system boundaries on the optimum energy parameters. This study shows that a consensus for system boundaries is needed to avoid conflicting outcomes.

Similar to the life cycle analysis approach by Kellens et al. [133], [134] and the available methods for quantifying electrical energy in manufacturing processes, transferable metering frameworks are needed for industrial water and steam systems that overcome the challenges of limited numbers of installed metering devices [94].

Benefits of a metering audit

The application of a metering audit offers great advantages for water and steam systems and industrial processes in general. The gained data forms the foundation for deeper process understanding that can be used in a variety of procedures such as further process analysis or strategic investments at management level [136]. Furthermore, it can be used to identify the weak spots of a facility’s monitoring system. It has been shown that a well-operating monitoring system can lead to a system performance improvement of up to 30% due to adequate guidelines and increased awareness [21]. Another goal can be to identify the most suitable location for installing new metering devices into a system. Rao et al. [137] presented a strategy on how to use temporary meters during an audit to find the best position for the installation of permanent electrical meters. The location and frequency at which data is collected is a challenge, particularly for SMEs that cannot afford to install a huge number of permanent meters. The methodology presented by the authors suggests the collection of data from existing meters at the beginning of the audit in order to identify data gaps that can be filled by temporary submeters. After analysing the collected data of all meters, the need for permanent measuring devices is evaluated along with measuring location and frequency. A further benefit of a metering audit is pointed out by Sturman et al. [121]. Due to the depletion and shortages of many natural resources worldwide, stricter usage regulations are expected in the near future. Therefore, it is desirable to gain an increased knowledge about saving potentials within industrial systems through regular metering audits. This enables a company to react quickly to restrictions and supply shortages that may arise in the future.

2.3 Proxy measurements

Data gaps due to missing or broken measuring devices are a common problem in industrial metering systems, as outlined in section 2.2.5. To overcome this barrier, additional measuring devices can be installed. However, this results in high costs, increased labour-intensity during installation, and frequently invasive access to the process [138]. Due to these drawbacks, additional installation of measurement devices is often avoided if it is not crucial for the process.

An alternative to the use of physical measuring devices are proxy metering strategies. A proxy meter device (PMD) combines empirical and theoretical methods to estimate an operating parameter by one or more correlated and measurable system parameters in combination with statistical techniques [139]. A schematic example of the PMD method is shown in Figure 2.18.

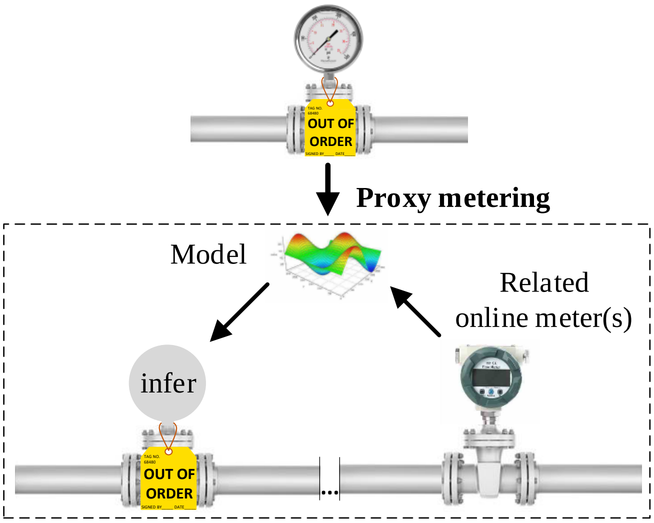


Figure 2.18: Schematic example of proxy metering

When using a PMD, no permanent physical meter device has to be integrated into the process to approximate the missing data. By relying on the existing infrastructure, PMDs follow the idea of using (relatively) easy accessible online data for the estimation of either difficult to measure, inaccessible, or offline process parameters [140]. Models dedicated to the estimation of other variables are known as proxy meters, soft sensors, or virtual sensors [126]. In this work, the term proxy meter device (PMD) is used instead of the frequently used term soft sensor. This choice has been made as most of the publications that include the term soft sensor represent a virtual meter device that comes from a mathematical perspective and focuses on ever advancing mathematical solutions, e.g. in

[141]–[145]. A PMD in the scope of this work is interpreted as a more basic form of virtual meters that comes from an engineering perspective and focuses on minimal modelling needs.

Fortuna et al. [138] and Doraiswami et al. [146] showed that PMDs offer promising advantages in comparison to physical measuring devices due to their dependence on existing infrastructure. These advantages include:

- Simplified process implementation: Integration of a PMD is possible while the process is running. There is no invasive access needed.
- Lower implementation cost: No further physical measuring device has to be purchased and implemented. Nowadays, the wide availability of digital monitoring systems often provides enough data for designing PMDs based on available data sources. However, depending on the complexity of the model, the development can be time intensive.
- Real-time data estimation: Data from offline measuring devices and hardware sensors with low sampling frequencies can be integrated into an online monitoring system by using PMDs.

Since the prediction of a PMD is based on a model in combination with physical metering devices, there is an error associated with using it as a predictor. According to Boiarkina et al. [147], this error is dependent on a variety of factors, e.g. the chosen model, the complexity of the underlying physical process, and the availability of related measuring devices. For the development of a PMD model, three categories exist that are dependent on the used approach. Figure 2.19 shows an alignment of the different models in comparison to their complexity and required process knowledge.

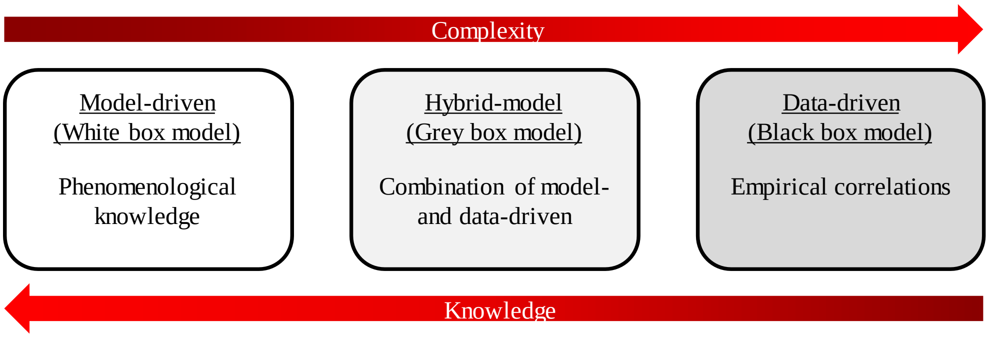


Figure 2.19: Modelling types for PMDs, inspired by [138]

The three different model types are explained in more detail below [144], [145]:

- Model-driven: Full phenomenological knowledge about the process is used for this type of model. This model is commonly derived from fundamental chemical or physical laws that describe the principles of the process. Deep knowledge of the process is needed which is not always available. In the literature, they are also known as white-box models.
- Data-driven: Empirical correlations are used for data-driven models. In contrast to the model-driven type, small knowledge of the process is needed. The process description is based on empirical observations. In the literature, data-driven models are also known as black-box models.
- Hybrid-models: This type of model combines model- and data-driven approaches. A combination of physical and empirical modelling is used to estimate the output of an unmeasured process. In the literature, hybrid-models are also known as grey-box models.

Nowadays, higher availability of process data has led to PMDs being predominantly modelled by data-driven models [148]. Since data-driven models are the most common type of PMDs in industrial applications to date, the following study in this research focuses on data-driven models only. When suitable data is available, data-driven models represent a straightforward method to assist plant operators in diagnostic, prognostic, and decision support.

2.3.1 Proxy meter development and maintenance

The general schematic for the development and operation of a data-driven PMD is depicted in Figure 2.20. The following paragraphs explain in detail the development of new PMDs for industrial applications and outline maintenance procedures required to ensure higher accuracy during operation.

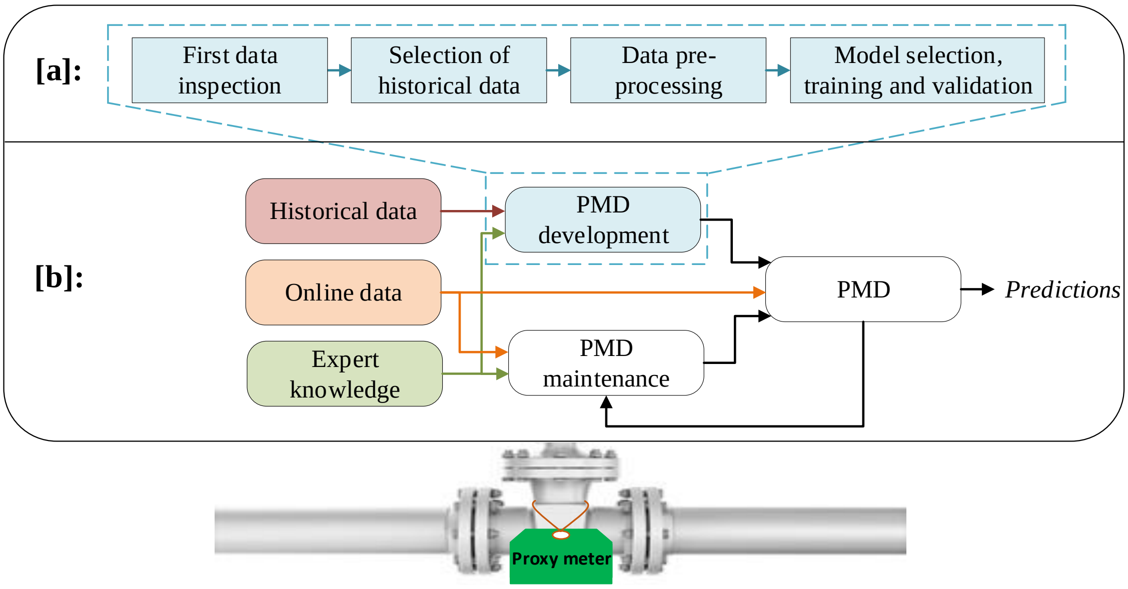


Figure 2.20: Proxy meter development and operation: PMD development method [a], Overall layout for the operation of PMDs [b], adapted from [145], [149]

A variety of approaches can be found in the literature that describe the development of PMDs in industrial environments [141], [145], [150]. Referring to Luttmann et al. [140], the modelling of a PMD consists of two main components; the selection of related measurement devices and the mathematical process. Once data gaps (target variable(s)) are identified, related online meters have to be found. Previously gained process knowledge assists in the selection of suitable, related measuring devices. In the case that there are no measuring devices associated with the target variable, an additional physical meter has to be integrated into the system. If related devices are available, the PMD development for data-driven models can start according to the generic framework shown in Figure 2.20 [a].

First data inspection

Within the first phase of this framework, an overview of the data structure is gained to identify preliminary issues within the data structure [145] e.g. locked variables that show a constant value. In addition, the initial data inspection provides insights into possible models that could be used in the later model selection phase.

Selection of historical data

Next, historical data records are selected from the related measurement devices for training and validation of the PMD. Fortuna et al. [138] demonstrated that a reduction of model complexity is achievable when only stationary parts of the data are used. Furthermore, only data from normal operating conditions should be considered within this phase, ignoring seldom occurring events.

Data pre-processing

Data pre-processing is a crucial phase in the development of a PMD. Although the pre-processing requires high efforts, it is of great importance as the developed models are only as good as the accuracy of the input data [151]. When data is collected from industrial applications it contains inconsistencies such as diverting sampling time, missing data, and outliers. In the data pre-processing step these problems are addressed to enable more effective modelling. In a case application of a PMD for estimating nitrogen oxide emissions in the automotive industry, Khachane et al. [152] demonstrated the benefits of including graphical tools in the data pre-processing stage as well. In this publication, scatter plots have proven to be a valuable tool to visually present the relation and inconsistencies between two variables.

Divergent sampling times of variables are common in industrial settings when different online monitoring systems are in place. Additionally, to reduce data size, some monitoring systems are set up to only record a new data sample if the parameter changes more than a pre-defined threshold. In such cases, the sampling time between each recorded reading varies. As stated by Souza et al. [141], the synchronisation of the selected variables is essential before their use in a PMD model. The most common approach to achieving this is the down-sampling method. This method excludes the values of the more frequently measured variable that do not have a corresponding value with their less frequent counterpart. However, this method can lead to inaccurate models due to information loss if the less frequent variable is scarcely sampled. As a more complex alternative, a finite impulse response model can be used. This method estimates values of the less frequent variable by using weighted sums of previous measurements in a window of finite length [149].

Missing values result from various causes, most commonly from hardware failures, abrasion effects, or data transmission problems. In general, there are two approaches to deal with missing data sections [145]. First, if the number of missing values is small, listwise deletion can be performed to remove the data gaps. Second, if larger sections of missing data exist, an alternative method for filling the missing values has to be applied to prevent information loss. A simple technique is to use the mean or median of the non-missing values for that variable (multiple imputations). The use of the maximum likelihood method represents a more complex approach to filling data gaps, whereby a model is assumed for the data distribution of the variable. It has been shown by Souza et al. [141] that significant improvement in the estimation of larger amounts of missing values is possible by using the maximum likelihood method instead of mean / median substitution.

The third common problem in collected datasets from industrial processes is data outliers. Lin et al. [150] considered values as outliers when they are not consistent with the majority of data and deviate significantly from normal values. Typical causes of outliers are sensor malfunction, communication errors, or sensor degradation. Outlier detection constitutes an essential step as the model may otherwise misrepresent the system. To start, obvious outliers that do not satisfy the technological limitations of physical meters can be easily identified and removed [148]. Based on statistical techniques, the simplest method for detecting an outlier in measurements is based on the 3σ rule (equation (2.1)) which assumes that the values follow a Gaussian distribution [153].

$$Outlier\ (x) = \mu(x) \pm 3\sigma(x) \tag{2.1}$$

where $\mu(x)$ is the mean value, 3 is the threshold, and $\sigma(x)$ is the standard deviation of the variable x. A data point will be labelled as an outlier if the value is three or more standard deviations from the mean value. When data is affected by many outliers, the 3σ rule is insufficient because the outliers inflate the variance estimation. A more robust version of the 3σ rule is the Hampel identifier [154], equation (2.2). Here, the mean value is replaced by the median value and the median absolute deviation (MAD) replaces the standard deviation.

$$MAD = 1.4826\, median\, \{|x_i - x^*|\} \tag{2.2}$$

where x^* is the median of the data sequence. The factor 1.4826 is chosen to make the expected value of the MAD scale equal to the standard deviation $\sigma(x)$ for Gaussian data sequences. Pearson [155] demonstrated that the use of the Hampel identifier within a moving window filter with a defined window size is favourable for identifying outliers in larger measurement samples.

As a necessity, the 3σ rule and the Hampel identifier assume that outliers follow a Gaussian normal distribution. However, for industrial measurements, this is not always the case. An example is the measurement of low absolute pressures. From physical limitations, measurements can never be negative. In these circumstances, the measurement distribution can be heavily skewed or restricted on one or two sides, which is known as truncated normal distribution. For these cases, the 3σ rule and Hampel identifier cannot be applied. Though, a logarithmic transformation allows changing the skewed data into or equal to a normal distribution [156]. For datasets with a variety of variables, univariate outlier detection approaches are limited in their function. In these cases, multivariate outlier detection approaches have to be considered that take the interaction among variables into account. Principal Component Analysis (PCA) and Partial Least Squares method are effective techniques that project and reduce the data dimension by finding an alternate set of coordinates [157]. For outlier detection within a PMD model, Fortuna et al. [138] used three statistical parameters, known as the Jolliffe parameters. These parameters are applied to the decomposed data from PCA and can identify outliers that do not conform to the correlation structure of the data or inflate the data variance. A more comprehensive description of the Partial Least Squares method is outlined in the subsequent section.

Model selection, training, and validation

Before a suitable model for a PMD is selected, a reduction in the number of input variables might be necessary when the number of available input variables is high. Fewer numbers of input variables simplify the modelling process due to fewer pre-processing needs, less computational demand, better interpretation, and a smaller risk of overfitting (for an explanation of overfitting please refer to section 2.3.3) [158]. Unsupervised and supervised variable selection methods can be applied to reduce the input dimension. As an unsupervised method, multivariate approaches like PCA are common to identify the variables with the highest variances. O’Driscoll et al. [159] presented a methodology to identify the variables with the highest influences for a machine tool by using pattern recognition in combination with PCA. A similar approach can be used within the modelling stage of a PMD when the number of input variables is large. As a supervised variable selection method, the choice of input variables is guided with the goal to attain the highest possible model accuracy [141]. To start, a subset of input variables is chosen. Kadlec and Gabrys [160] outlined in their review of industrial PMD development that process knowledge is crucial to facilitate the selection of input variables that have a high impact on the target variable. This subset is then used for training the PMD model. At a later stage, the trained model is validated with a suitable performance measure to judge the accuracy compared to other models that have been trained with different subsets. At the end of the PMD development, the subset with the highest model accuracy is selected. Andersen and Bro [158] outlined a similar variable strategy in their review of variable selection methods for regression algorithms, known as forward selection. In this method, the variable with the lowest predicted error is selected initially. Subsequently, the variable that shows the best performance in combination with the first variable is added to the model. This procedure is continued until the predictions no longer improve by adding new variables. To identify the relevance of single input variables in relation to the target variable, Andersen and Bro [158] recommend the use of the correlation to the target variable or the selectivity ratio. This ratio is defined by the explained variance divided by the residual variance and is calculated for every input variable. The explained variance is expressed by the Coefficient of Determination (R^2). For the residual variance, a linear regression model is first developed with the analysed input variable and the variance of the residuals are calculated subsequently. Higher correlations to the target variable and selectivity ratios indicate good predictive performance of the selected input variable.

A suitable PMD model is selected based on the pre-processed data and expert knowledge. The choice of the model depends on the application of the PMD, process layout, nature of data, and experience of the model developer. Kadlec et al. [145] point out that no unified theoretical approach exists for the selection of a PMD model. In the literature, Wolpert and Macready defined this limitation as the ‘no free lunch’ theorem [161]. The authors state that “if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems”. Thus, it is often required to use several approaches and base conclusions on a combination of analysis results [162]. For the model selection of a PMD, it is recommended to start with a simple model and gradually increase the complexity as long as significant improvements can be observed in the model’s performance [151]. Significant performance differences can be evaluated with the statistical t-test. A key influence on model performance is the complexity within the training and validation dataset. The relationship between regression algorithms and dataset complexity is outlined in more detail later on in this section. In general, there are two choices at the beginning of the model selection: a linear or a nonlinear model. For model simplicity, it is recommended to start with a linear model. According to the PMD developing experience of Fortuna et al. [138], linear models are still able to show acceptable performance in many industrial processes when non-linearity is only slightly present and processes are almost steady-state. If a linear model does not provide satisfactory results, the use of a nonlinear model can be considered. The most popular models for linear relationships in data-driven PMDs are the Multiple Linear regression (MLR) and Partial Least Squares regression (PLSR) models [149]. For nonlinear behaviours, Neural Networks (NN) and Fuzzy logic systems are frequently used. The functional principle of MLR, PLSR, and NN are outlined in further detail below. A combination of different models can be employed to improve functionality. This is adopted by the ensemble approach, whereby a set of base models is trained and their responses are combined to obtain a more accurate prediction compared to the use of a single model [163]. Successful implementations generate a diverse range of base models by varying initial conditions, manipulations of the training dataset, and / or using different learning algorithms. Moreover, the combination of these models is crucial. Common response combinations are achieved by using the mean value, a weighted average, or a NN according to Soares et al. [164].

In the training phase of a PMD, the model is tuned on the pre-processed dataset which includes all common operation states of the process. In most industrial processes, multiple operating modes exist [141]. Regularly training the model might be required if the system state changes frequently e.g. when the system is highly sensitive to external disturbances, such as changes in feedstock or ambient temperature. To evaluate the model's performance, it is tested on a new set of data. Therefore, it is common practice to split the historical dataset of the input variables into a training and validation set at the beginning of PMD development, known as holdout validation [165]. Once a model has been developed with the training set, it can be validated by using it on the validation set. The objective of this validation is to assess the performance of the trained model on new data samples. For performance metrics, the Mean Squared Error (MSE), the more intuitive Root Mean Squared Error (RMSE), or the Pearson Correlation Coefficient (CC) are generally used [145]. The MSE measures the average square distance between the predicted and the correct values [166], according to equation (2.3):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \tag{2.3}$$

where $\hat{f}(x_i)$ is the prediction that the trained model gives for the i-th observation, n is the number of all observations, and y is the predicted value. The MSE shows a small value if the predictions are very close to the true responses. It is a simple and common measure for learning algorithms. However, the interpretation of this parameter can be difficult due to the squaring. As an alternative, the RMSE is defined as the square root of the MSE. This simplifies the understanding of the error measure since it has the same unit as the original values [167]. In contrast to MSE and RMSE, the CC measures the correlation between two variables x and y [168], equation (2.4). The value of CC lies between -1 and 1. A value of 1 indicates that the data points lie on a straight line with a positive slope, with x and y increasing together. A value of -1 indicates a straight line with a negative slope. The CC is independent of the magnitude of the slope. A value near 0 implies no correlation between x and y.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.4}$$

where $\bar{x}$ is the mean of the x-values and $\bar{y}$ is the mean of the y-values. For PMD validation, the CC is a valuable measure of the correlation between actual and predicted target variable when the model is applied during validation. It has been used in several PMD use cases that can be found in the literature [169]–[171]. CC values closer to one imply high estimation capabilities [171]. Though, Maciel et al. [172] advise that some caution must be taken since correlation can also be observed at random and outliers can strongly influence the correlation result.

Besides the application of numerical performance evaluation, graphical tools can be very powerful for validation [138]. An example of the visual inspection of recorded and estimated values are scatter, residual distribution, and normal probability plots.

Dataset complexity evaluation

The complexity of datasets is decisive for the performance of different regression algorithms. To characterise data complexity, meta-learning approaches can be used. Thereby, model inputs and output(s) of a dataset are analysed to predict the performance of learning algorithms as shown by Brazidal et al. [173]. Meta-learning follows the aim to aid users in selecting a suitable predictive model for their application case. To date, meta-learning approaches for characterising datasets have been mainly focused on classification problems [172]. For the classification domain, measures exist that give recommendations for the selection of suitable algorithms based on the data complexity [174]. For regression problems, Lorena et al. [175] recently presented a set of measures that can be applied for regression problems of varying complexity. These measures are intended to characterise regression complexity from different perspectives by including the following categories; inputs correlation to the target variable, linearity, smoothness, and data geometry, topology, and density. Table 2.3 depicts and describes the regression complexity measures. In the last column of Table 2.3, the symbol ↑ denotes that higher values of the specific parameter indicate more complex regression datasets, whereas the symbol ↓ denotes the opposite. For further explanations and calculations of the shown measures, please refer to Appendix A .

Table 2.3: Dataset complexity measures for regression models, adapted from [175]

Object	Label	Measuring feature	Parameter relying on	Complexity
Correlation	C1	Maximum inputs correlation to target variable	Single inputs-target correlations	↓
	C2	Average inputs correlation to target variable	Average inputs-target correlation	↓
	C3	Collective input efficiency	Correlation + linear fit + residual error	↑
Linearity	L1	Mean absolute error	Multiple linear regression	↑
	L2	Residual variance	Multiple linear regression	↑
Smoothness	S1	Inputs distribution	Euclidean distance	↑
	S2	Error of nearest neighbour regressor	Nearest neighbour regression	↑
Geometry, topology, and density	L3	Non-linearity of linear regressor	Interpolation + linear fit	↑
	S3	Non-linearity of nearest neighbour regressor	Interpolation + nearest neighbour regression	↑
	T1	Average number of samples per dimension	Samples + input variables	↓

The application of the measures in a case study outlined the need to use several measures instead of a single parameter to get a better characterisation of regression complexity [175]. In addition, dataset normalisation between [0,1] is recommended before the complexity measures are applied (except for parameter *T1* in Table 2.3). This allows a simpler complexity comparison for different datasets. Variables in a dataset are normalised according to:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2.5}$$

Maintenance of PMDs

After the implementation and launch of a PMD within a monitoring system, Kadlec and Gabrys [160] reported that a gradual deterioration of the performance can be observed in most cases. The appearance of new operation states that have not been included in the historic dataset can cause lower performance. In addition, effects such as varying quality of input materials, abrasion of mechanical materials, and external environment changes can lead to data drifting over time. To maintain a high PMD accuracy during operation, Fortuna et al. [138] suggested the monitoring of the system inputs and to compare the PMD performance to suitable thresholds that are set by the operators. The maintenance procedure is based on three facets of information; online data, expert knowledge, and performance feedback. If the performance reaches unacceptable levels, the model has to be retrained or in the worst case rebuild from scratch. In addressing the issue of degrading PMD performance, most concepts in the literature do not provide any automated maintenance mechanism [145]. This leads to the requirement of manual quality control and maintenance of the PMD which can be a significant cost factor if process characteristics change promptly [176]. Nevertheless, some adaptive maintenance approaches have been published. Moving average windows are applied by several authors as a continuous self-update tool of the model [141]. The samples inside a time window are used to retrain / update the model, while samples outside of the window are discarded. Crucial for successful implementation of this approach is the chosen window size. A study of different window sizes has been presented by Kuncheva and Žliobait [177]. The authors conclude that a too small window can see the model adapting to noise while a too large window can limit the adaption capability. A different strategy to overcome data drifting is the use of sample weighting, as shown by Kadlec et al. [149]. The samples are weighted according to their age which causes a sample's importance to decrease with time. The model is periodically retrained with the weighted samples. Old samples are either discarded or tagged with low weights. Kaneko and Funatsu [178] presented a further approach for PMD model updating by considering time differences between objective and explanatory variables to compensate for data drifts and gradual changes over time.

While the number of publications that focus on automatic PMD maintenance is growing, some industrial engineers are hesitant to use these features. Kano and Nakagawa [179] reported concerns from engineers towards automatic adaptions as they might cause serious damage to the process. Therefore, automatic maintenance strategies have to be transparent and easily comprehensible to ensure industrial adoption.

Linear algorithms for PMDs

As stated above, the most common linear algorithms for PMDs are MLR and PLSR. In this subsection, these two regression models are explained in more detail. The main characteristics of these algorithms are summarised in Table 2.4.

Multiple Linear regression (MLR)

MLR is used to model the relationship between two or more input variables and one response variable by fitting a linear equation to the observed data [166]. It is an extension to simple linear regression to accommodate multiple predictors. The MLR model takes the general form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{2.6}$$

where x_p represents the p-th predictor with the according regression coefficient β_p. β_0 is the constant term of the model, ϵ represents the noise term / random error. The regression coefficients are estimated by the least square approach. This is achieved in mathematical software packages by a gradient descent algorithm that finds the optimal regression coefficients by minimising the sum of squared residuals.

Partial Least Squares regression (PLSR)

While MLR is an easy to adopt regression model that is suitable for a few input variables that are not significantly redundant (collinear), it shows unsatisfactory performance if the number of input variables is large or if the training and validation dataset is small. When limited data is available, it is common to derive a well-performing MLR model from the training dataset. However, the performance of the MLR model during validation decreases sharply. This is known as overfitting (for further explanation of overfitting please refer to section 2.3.3). As Tobias [180] proved on a spectrometric calibration example, Partial Least Squares regression (PLSR) represents a suitable alternative. This algorithm uses a smaller number of latent variables for the regression model compared to MLR, resulting in increased performance in the case of collinear, noisy, or small datasets. The principal of PLSR is based on finding latent variables that capture the underlying phenomena of the investigated data. This is well suited for industrial processes as many are data rich but information poor [181]. Figure 2.21 depicts the functioning of PLSR graphically. Latent variables (also referred as Partial Least Square components) are composed of linear combinations to maximise the covariance between input and target variables, as stated in a review for PMD development by Lahiri [151]. More precisely, the information of the dataset is distilled into a smaller number of orthogonal latent vectors for n samples according to:

$$X = SP^T + E \tag{2.7}$$

$$Y = UQ^T + F \tag{2.8}$$

where X is an $n \times N$ matrix of input variables, Y is an $n \times M$ matrix of target variables, S and U are $n \times p$ score vectors, P $(N \times p)$ and Q $(M \times p)$ represent matrices of loadings, E $(n \times N)$ and F $(n \times M)$ are the matrices of residuals. The matrices X and Y are decomposed to maximise the covariance between the score vectors in S and U [182]. This is achieved by rotating the loadings in matrices P and Q in order to maximise correlation between each set of X-scores (S matrix) and Y-scores (U matrix), depicted in Figure 2.21. The loadings are arranged to ensure that the first latent variable explains the largest variance of the target variable(s). When plotting the loadings for two latent variables, the relation of PLSR input variables can be analysed [158]. In this plot arrangement, input variables that are located close to each other indicate similarity. Additionally, relatively high loadings of a certain input variable imply high relevance to the target variable. The main aim of PLSR is to reduce the number of variables that are required for predicting the target variable(s). Therefore, only a few latent variables are used for modelling in most cases, defined in the PLSR training stage. Wold et al. [182] reported that this is commonly achieved by training PLSR models with different numbers of latent variables and validating the performance on a new set of data or by using cross-validation. Graphical support for this selection is a Y-variance plot which displays the explained variance of the target variable for different latent variables. Once the PLSR model has been trained on a historical dataset, predictions of the target variable(s) are performed by first extracting the X-scores S from the input variables with the X-loadings P, which have been predefined during the training stage of the model [180]. The Y-scores U are predicted through a previously trained MLR model and then the predefined Y-loadings Q are used to predict the target variable(s). A frequently used iterative method for PLSR in mathematical software is the NIPALS (Nonlinear Iterative Partial Least Squares) algorithm [182]. The algorithm has two main advantages; it can handle missing data and the calculation steps are easy to understand since it calculates the components sequentially, starting with the first component (direction of greatest variance).

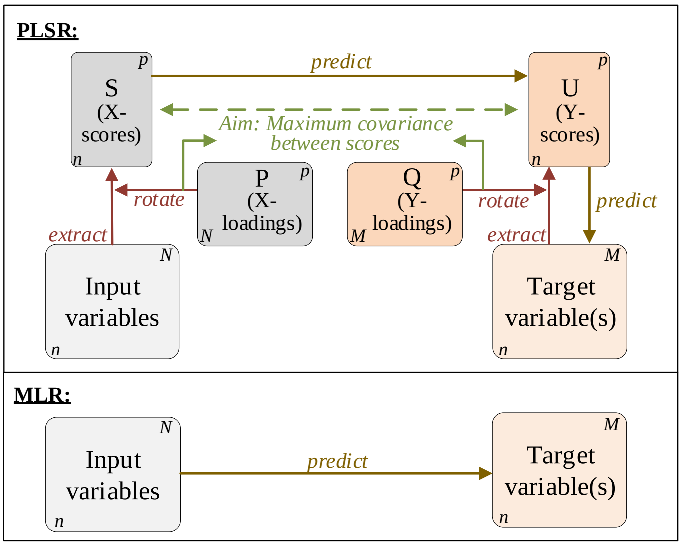


Figure 2.21: Schematic depiction of PLSR and MLR, inspired from [180], [183]

Neural Networks for PMDs

In a PMD developing review by Kadlec et al. [149], the authors identified Neural Networks (NNs) as the most commonly applied non-linear PMD algorithm. This section explains the functioning of NNs, the main characteristics are shown in Table 2.4.

Lahiri [151] sees NNs, also referred to as Artificial Neural Networks, as an attractive tool for non-linear process modelling since prior knowledge about the relationship between the process parameters is not required. NNs represent a computer modelling approach that learns through iterations. This means that the models can adapt to changing environments and are capable of dealing with uncertainties and noisy data.

Due to its model-free approximations, NNs have gained increasing popularity lately within manufacturing applications and beyond.

In a review of sensor signal processing strategies, Dornfeld and Lee [184] describe the structure of NNs as a build-up of a collection of simple, interconnected nodes which operate in parallel and store information in the node connections. This structure is inspired by the layout and functionality of the human brain. The central idea is based on extracting linear combinations of the inputs as derived features and using these to model the target of nonlinear functions [163]. A multilayer perception (MLP) NN is most widely used for industrial cases. This NN typically consists of three layers, as reported by Lahiri [151] and shown in Figure 2.22.

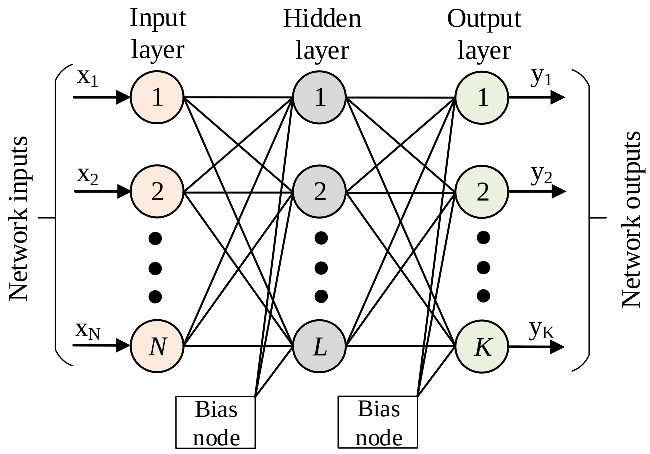


Figure 2.22: Neural Network architecture, adapted from [151]

These layers include an input layer consisting of the network inputs (input nodes), one hidden layer which is a build-up of a variable number of neurons (hidden layer nodes), and an output layer (output nodes). This structure can be used for approximating nonlinear relationships between input variables and target variable(s). Each node of the input layer is connected to the neurons of the hidden layer by using weighted connections. Likewise, all neurons of the hidden layer are connected to the output layer nodes with corresponding weights. The number of input nodes (*N* in Figure 2.22) equals the number of model input variables. Similarly, the number of output nodes (*K* in Figure 2.22) corresponds to the number of target variables, which for regression problems typically equals one. In contrast, the number of hidden neurons (*L* in Figure 2.22) is adjustable and defined by factors, such as desired generalisation and approximation capabilities. Furthermore, a bias is associated with every hidden and output layer node. In an MLP structure, this bias shows a fixed value of +1 and provides further adjustable weights for model fitting. The nodes of the hidden and output layer receive inputs from the previous layer and multiply them by their corresponding weights [185]. The resulting values are summed up and an activation function is applied to produce an output. As an example, the following equation shows the calculation of a neuron in the hidden layer:

$$x_i^{hidden} = g^{hidden}\left(\sum_j \omega_{j,i}^{hidden} x_j^{in}\right) \tag{2.9}$$

where g^{hidden} is the transfer function, x_j^{in} is the j-th variable of the input sample, $\omega_{j,i}^{hidden}$is the weight between the j-th input node and i-th hidden neuron, and x_i^{hidden} is the output of the i-th hidden neuron. Non-linearity is introduced to the model structure by the activation function, which are commonly sigmoid functions with output values between [0,1]. The following equation is an example of the sigmoid activation function of a neuron in the hidden layer:

$$g^{hidden} = \frac{1}{1 + \exp(-x)} \tag{2.10}$$

While most of the applications of NNs are based on nonlinear behaviours, Dornfeld and Lee [184] point out that linear NNs are realisable by setting the activation function to a linear function.

During an iterative procedure known as network training, the weights are adjusted to accurately generate outputs according to the model inputs within a training dataset. This training involves the minimisation of an error function that may show several local minima. Therefore, several training runs are necessary to find the global or the lowest local minima. The generic approach for finding the minima is by gradient descent, which in the case of NNs is most commonly achieved through use of a back-propagation algorithm [185]. Here, all weights are adjusted by forward and backward movements over the network to minimise the following error function E:

$$\omega^{new} = \omega^{old} + \Delta\omega \qquad \text{with } \Delta\omega = -\eta\frac{\partial E}{\partial \omega} \tag{2.11}$$

$$E = \frac{1}{2}\sum_{i=1}^{n}\left(x_i^{out} - y_i\right)^2 \tag{2.12}$$

where η is the learning rate of the algorithm, ω are the weights between the nodes, and E is the quadratic error calculated of the training samples. For the application in industrial systems, Yu and Wilamowski [186] see the Levenberg-Marquardt algorithm as a popular variation of the back-propagation algorithm. The authors in [186] view the Levenberg-Marquardt algorithm as one of the most efficient training algorithms, particularly for small size NNs. Once the model is trained, it can be applied with the fixed weights to generate outputs from new input samples.

The selection of several parameters has to be included during the training of a NN, such as the number of hidden layers, number of hidden neurons, activation function, initial values of weights, and the number of optimisation cycles. Hastie et al. [163] advise that the need for these selections and the nonconvex optimisation problem in non-linear NN applications can cause several issues during model training and validation. These issues include inconstant performances of several trained NNs due to the selection of initial weights and the possibility of various local minima of the error function. A further concern is the choice of suitable numbers of hidden neurons. When too few hidden neurons are selected, the model might not have enough flexibility to capture the nonlinearity of the dataset. However, too many hidden neurons can result in too many adjustable weights that lead to overfitting of the model. In particular, for small training datasets, NNs are prone to overfitting.

The following table summarises the main characteristics of MLR, PLSR, and NN:

Table 2.4: Main characteristics of MLR, PLSR, and NN, adapted from [151], [187]

	Multiple Linear Regression (MLR)	Partial Least Square Regression (PLSR)	Neural Network (NN)
Method	Linear	Linear	Non-linear / (linear)
Loss of training error	Least-squares	Least-squares	Least-squares
Solution	Unique	Unique	Local minima possible
Recommended training samples size	Medium	Low	High
Required meta-parameters	-	Number of latent variables	Number of hidden layers, Number of hidden neurons
Required additional parameters	-	-	Activation function, Initial value of weights, Optimisation cycles

2.3.2 Application areas of PMDs

Nowadays, the wide availability of digital manufacturing systems allows the application of PMDs in many industrial areas. Fortuna et al. [138] and Liu et al. [126] see the benefits of a PMD in four application areas:

- Back-up for physical meters: Some physical meters are installed in very harsh working environments. These conditions can frequently cause faults that require regular hardware substitutions. To overcome the down period required for physical meter maintenance / replacement, a PMD can substitute unavailable measurement instruments to avoid plant performance degradation and rising costs.
- Reducing measuring hardware requirements: PMDs are commonly implemented instead of a permanent measuring hardware device. In many scenarios, this solution is seen as a source of possible budget savings. For development of the PMD, measurements of the target variable have to be available for model training and validation. Therefore, the use of a temporary physical meter may be necessary, e.g. in the form of a clamp-on meter for flow rate measurements. When PMDs replace a physical meter permanently, increased attention should be paid to the model validation. High model performances are required in these cases due to the lack of any redundancy. Periodical PMD maintenance and retuning is crucial to guarantee consistent accuracy.
- Real-time estimation for offline meters and delayed measurements: A further application area for PMDs is the real-time estimation of system variables from offline / local physical meters or when the measurement is performed at a later stage. Time delays for measurements can be significant due to the execution time of the metering instrument, e.g. in gas chromatographs, or if high costs are associated with the measurement and therefore performed infrequently. In this application area, a PMD can infer system parameters in real-time to give operators enhanced process insights. Previous offline and delayed measurements can be used for the development and maintenance of the PMD without the necessity for a temporary installation of a physical meter.
- Fault detection and predictive maintenance: Fault detection and diagnosis are part of modern industrial control systems. Advanced techniques are used in control systems to perform early detection of faults and to provide decision support for maintenance and repair. This field has gained increased attention recently and is

known in the literature as predictive maintenance [188]. Liu et al. [189] see PMDs as an effective tool for predictive maintenance as they can combine physical measurement signals to detect process conditions and faults at an early stage.

PMDs are well suited for use in industrial processes that include water and steam systems. Concerning the replacement of physical flow meters, PMDs are commonly used in the oil and gas industry where the use of physical meters in wells is expensive and problematic (commonly referred to as a virtual flow meter in the oil and gas sector). The use of PMDs for metering multiphase flows (gas, oil, and water) from wells has been shown by Wang et al. [190]. Pressure and temperatures are used within the model in combination with mass and energy balances to estimate the mass flow rate of the target value. To verify the PMD model, it is compared to a physical flow meter. The results demonstrate that the PMD has lower accuracy (maximum deviation is 10%), but greater reliability, higher availability, and lower lifecycle costs. A similar but more complex PMD for measuring multiphase flows has been presented by Al-Qutami et al. [143]. The authors use available temperature and pressure measurements in combination with NNs to model the virtual flow. During validation, the developed PMD resulted in less than 10% deviation to the actual well flow data.

Several authors have considered PMDs within the water sector. Mass balances together with lift height and motor frequency of pumps were used by Äijälä and Lumley [191] to implement a variety of PMDs for flow metering in a wastewater treatment plant. Several PMDs were developed as a backup to physical meter devices. The examined wastewater treatment plant in the case study was experiencing failures with their installed physical metering devices when unexpected high flowrates occurred, e.g. after heavy rainfalls. An alarm was triggered if a deviation between the outputs of a physical meter and an equivalent PMD remained over a defined time period. Figure 2.23 shows an example of the measured and calculated (via PMD) bypass flow that includes a period of increasing deviation, marked by the red rectangle. In the real use scenario, the PMD took over control during the high deviation until the physical meter returned to normal.

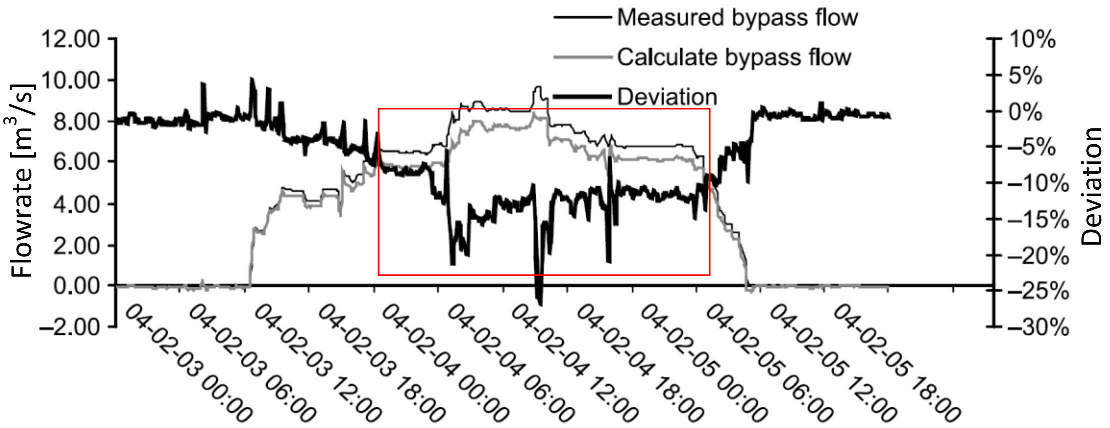


Figure 2.23: Flowrate deviation between measurement and estimation (by PMD), adapted from [191]

A further example on the use of a PMD in the water sector is presented by Juntunen et al. [192]. In this paper, the turbidity in drinking water was estimated based on dynamic and static linear regression. Due to the high amount of available input variables for the model in the presented case study, the most significant variables were identified by the forward selection method. With the reduced subset of input variables, different PMDs were developed that estimated the turbidity of drinking water with CCs of up to 0.86.

For industrial steam systems, the application of PMDs that can be found in the literature is limited. Xie et al. [193] developed a PMD for online measurement of steam quality in once-through steam generators. The presented model is based on first principles and encompasses an online updating mechanism that reduces the error between predicted and observed values. In two case studies from an oil sand extraction facility, the developed PMD achieved high prediction accuracy while showing a low computational cost. However, the modelling of the PMD requires deep process knowledge and is time-consuming due to the first principle approach. A further publication is presented by Shakil et al. [142] and focuses on inferring nitrogen oxide and oxygen emissions from steam boilers. In this approach, PCA is applied to reduce the number of temperature measurements. Subsequently, NNs are trained with the acquired principal components. Due to long measurement delays within the system, a modified NN that incorporated the system dynamics showed the highest performance.

2.3.3 Small datasets and proxy metering

In many industrial sites, only small datasets are available for modelling PMDs from related measuring devices. Fortuna et al. [171] reported that dataset limitations from industrial applications are often due to sampling cost, the requirement of a temporary physical meter, or low sampling rates, e.g. measured offline through laboratory analysis. Small datasets represent a challenge for use in PMDs as standard identification, modelling, and validation techniques have to be adapted [194]. This section outlines the required adaptions for PMD development on small datasets.

In the context of proxy metering strategies, Tsai and Li [195] classified small datasets as datasets that do not provide enough information to regression methods for a stable learning accuracy. Compared to big data, several authors have reported on the advantages of small datasets which include easier handling [196], higher uncertainty awareness by the user [197], and simplified visualisation capabilities [198]. Nevertheless, Di Bella et al. [199] pointed out that the main concern for small data usage in PMD development is the overfitting risk during model training. Overfitting describes the phenomena of a model that adjusts to specific random features of the training data that have no causal relation to the target function [164]. This results in high performance on the training dataset while the performance on unseen data during validation becomes worse. Additionally, small datasets might not include all run modes or operation stages of the process. The performance of a PMD model can be low when estimating unseen process run modes that have not been included during the training phase. As a rule of thumb, Babyak [200] suggested the inclusion of at least 10 to 15 observations per predictor variable in order to minimise the risk of overfitting in regression models. To avoid overfitting, different strategies can be applied to a small dataset before the data is used for model training. These are explained in further detail in the next subsection. Aside from the model training, the PMD validation can be challenging if the small dataset has to be used for model training and testing. Suitable validation techniques for these cases are *K*-fold cross-validation and leave-one-out-cross-validation, clarified in further detail below.

Small dataset manipulation techniques

A technique that aims to increase data diversity and improves the generalisation capabilities within a small dataset before being used for model training is referred to as bootstrap [171]. Since in practical applications dataset sizes are often limited, original data populations are unknown. To overcome this limitation, the bootstrap method replicates an original dataset x number of times by drawing random samples with replacement from the original dataset [168], Figure 2.24. Due to the replacement of the samples, the original dataset is not regenerated each time. Instead, datasets with a random fraction of duplicated original samples are created. James et al. [166] demonstrated that the bootstrap approach can be used to effectively estimate the true population of a dataset by replicating a limited number of the data points. The new distribution that is created with the resampling is referred to as bootstrap distribution and includes a larger number of samples than the original dataset. This increased sample number can be used for PMD model training in order to reduce the risk of overfitting, as shown by Napoli et al. [170]. However, the improvement is depending on the original dataset size and how well the chosen samples represent the underlying distribution [168]. In the case of very small datasets, the samples might not represent the underlying distribution correctly. When replicating these samples by the bootstrap method, the resulting output datasets represent the wrong distribution.

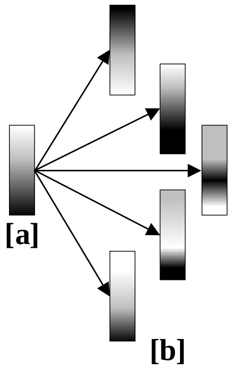


Figure 2.24: Bootstrap resampling technique: [a] original dataset; [b] resampled datasets, adapted from [201]

In addition to bootstrap, artificial noise injection (ANI) is another common methodology to reduce overfitting in small sample regression models. Andrijić et al. [202] added artificial data to the training samples of a PMD model which increased the diversity in the small dataset and proved that ANI can be an effective approach to improve learning accuracy and model robustness. Generally speaking, noise is seen as a drawback for model-based applications. Nevertheless, recent publications suggest that the addition of noise with zero-mean can have a positive effect on the generalisation performance of trained models [170], [202]. In the literature, two different methods for introducing ANI into a PMD training set are outlined: Soares et al. [164] injected noise with a Gaussian fixed-variance, σ^2, independent of the signal amplitude whereas Di Bella et al. [169] added the noise with a Gaussian fixed-variance, σ^2, dependent on the signal amplitude of the selected variable. The mean of the noise in both methods is set to zero. In general, the level of added noise has to be small enough to not significantly affect the relationship between input and target variables, as demonstrated by Fortuna et al. [171]. Though, if the noise level is too low it may not introduce enough diversity into datasets. For fixed-variance noise addition, a variance, σ^2, of 0.02 is generally considered as a suitable choice [164], [169], [171]. Besides its variance, the location of the injected noise is important as noise can be added to the input variables, to the target variables, or to both. The impact of noise location will be discussed in the subsequent section.

Small dataset validation

K-fold cross-validation and leave-one-out-cross-validation are two techniques used to optimise the use of small datasets during training and validation [203]. For *K*-fold cross-validation, the training set is randomly split into *K* exclusive subsets (the folds) of approximately similar size. The model training is performed on the *K*-1 folds while the remaining fold is used for validation [204]. This process is repeated, each time with a different partition, until all folds have been left out once for validation. The performance for the overall model is calculated by averaging the outcome of the *K* fold models. As a special case of *K*-fold cross-validation, leave-one-out-cross-validation leaves one single sample out for validation while the other samples are being used for model training. This process is repeated, until all samples have been left out once for validation. Arlot and Celisse stated [205] that leave-one-out-cross-validation is equivalent to *K*-fold cross-validation where the number of *K*-folds is equal to the number of samples.

Applications of small datasets in PMDs

Only a limited number of publications are focusing on using PMDs on small datasets in the literature. In publications by authors Napoli and Xibilia [170], Fortuna et al. [171], Caponetto et al. [206], and Di Bella et al. [169], [199], a variety of PMDs were developed based on small datasets from a thermal cracking unit. In the approach presented by Napoli and Xibilia [170], a small dataset was used for building a PMD model by applying NNs. For pre-modelling manipulation, bootstrap replications (BR) and ANI (noise added to the model target) were applied. Several NNs were trained and the best performing networks were aggregated by an ensemble approach (for further details on model ensemble please refer to section 2.3.1). The results showed that the applied pre- (BR and ANI) and post- (ensemble) modifications led to a performance improvement of up to 20% compared to a NN that was trained on the original small dataset. In a similar publication by Fortuna et al. [171], the influence of injected noise was analysed in further detail. The authors trained NNs on small datasets and varied the variance of the added noise before the data was used for model training. As a result, ANI improved the performance of trained NNs. A variance, σ^2, of 0.02 showed the best outcome, similar to the suggested variance by Caponetto et al. [206]. The position of the noise addition (input, target, or input and target variables) did not significantly affect the improvement. Furthermore, the dependency of the added noise on the signal amplitude did not change the results significantly. In contrast, Soares et al. [164] showed that the target variable is the best location for noise addition to a training dataset for PMDs. The performance of BR and ANI combined was better than without any dataset manipulation technique, but shown to be worse than ANI by itself. Taken together, the diverse results for optimal locations of ANI suggest that the impact of dataset modification techniques is dependent on additional factors like model complexity, sample size, and process behaviour. The impact of small sample sizes on NNs was studied by Tsai and Li [195]. The authors varied the number of samples that were used for NNs training between 5-35. The accuracy of the model increased greatly with each incremental change when using a smaller number of training samples. However, this incremental increase declined to a more moderate level for higher sample numbers. Similar to the previous publications, the application of bootstrap resampling increased the NN performance, especially for NNs trained on very small sample sizes (less than 10 samples). This result corroborates the findings reported by Andrijić et al. [202]. Besides the effect of bootstrapping for NNs, the authors analysed the performance of three different PMD algorithms on small datasets. MLR and multivariable adaptive regression splines (a particular form of step wise regression) models resulted in similarly satisfactory results while a NN achieved even better performance. In the presented case study, the small dataset replication by bootstrapping did not improve the performance of the linear model but had a positive influence on the multivariable adaptive regression splines and NN.

In summary, the limited number of publications that consider the validity of PMDs with small datasets demonstrate that sufficient accuracy can be achieved for the use of PMDs as industrial meters even in the processes with limited sample sizes.

However, most of the shown applications are only focusing on using NNs as regression algorithms.

2.3.4 Current challenges for PMDs

To date, the development of a PMD is a time and cost-intensive task for many industrial sites which limits the usage of these models, as reported by Kadlec and Gabrys [160]. In particular, for cases where the application of a PMD is only temporary, such as metering audits, the high development efforts needed do not justify its implementation. This corresponds to the outcome of a current status report on soft sensing by Luttmann et al. [140] where the need for simpler PMD implementation methods and increased knowledge transfer between academia and industrial facilities on PMD development were identified. Most of the existing approaches for PMDs in the literature are coming from a mathematical perspective with the focus on ever advancing mathematical solutions. However, many engineers and operators do not wish to implement complex statistical models in industrial systems, as these methods are not transparent [179]. Therefore, the lack of simplified PMD development approaches impedes an enhanced PMD application in industry. The status report by Luttmann et al. [140] advises that a key step towards simpler PMD development strategies would be to benchmark model performances from a variety of industrial use cases. This would assist future users in the selection of a suitable model for their particular case of application. Important hereby is to include meta-learning approaches in order to categorise the required model complexity (an explanation of meta-learning is shown in section 2.3.1). Another major issue up to now is the iterative execution of the PMD development to accommodate the knock on effect of a decision at the different development stages [160]. According to Kadlec et al. [145], after optimising one part of the model, the influence on the other parts has to be checked and might include retuning of the effected parts. Depending on the process knowledge and the developer expertise, a lot of effort might be required until a satisfactory model is built. One possible solution to this problem is an automated selection of the most appropriate algorithm from a pool of available algorithms within the PMD model. This leads to a model complexity increase but can transfer parts of the model development burden from the developer to the model. For automating the selection of a suitable model, meta-learning approaches are crucial along with standardised model development and implementation methods.

Standardised development is seen as a general need for simplifying and accelerating PMD usage in industry [140].

While the problem of small dataset sizes is common in industrial applications [187], the number of publications that address this issue during PMD development is limited. Furthermore, the majority of the available publications for small dataset PMDs are solely using NNs. The limited number of studies and the focus on NNs impede the development of PMDs on small datasets and represent a barrier for widespread adoption in industrial processes. This corresponds to the general view that small datasets are an important future research area for learning algorithms [207].

2.4 Summary

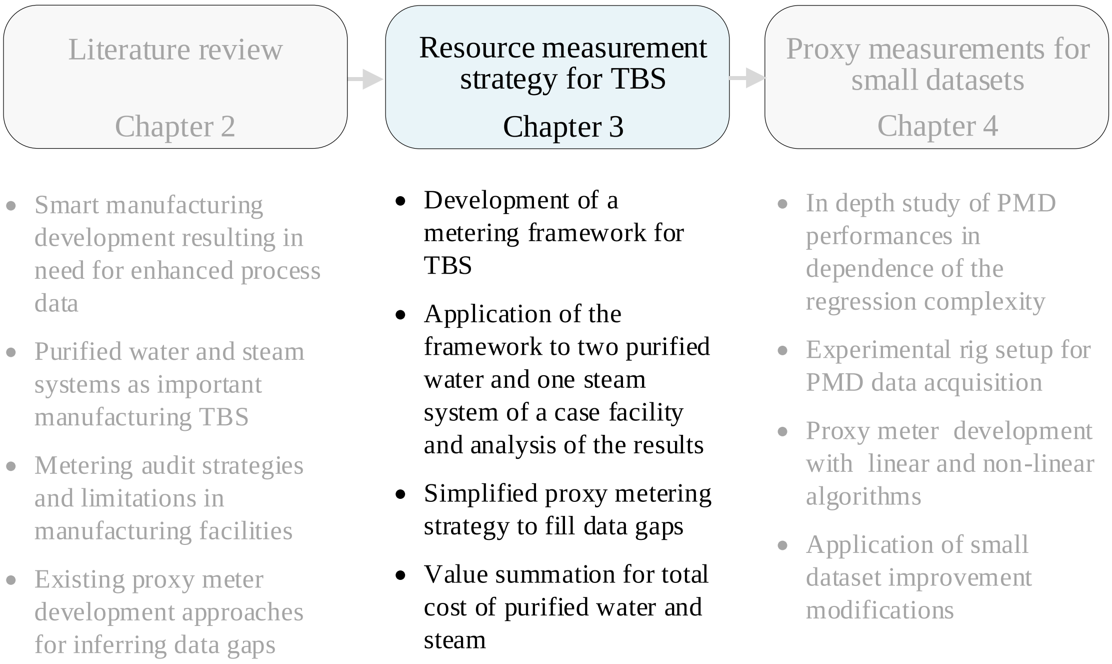
The current advancements of manufacturing processes, coined as ‘Industry 4.0’ or Smart Manufacturing, aim to integrate production systems for increased flexibility, customisation, and process control. An overlooked fact in many smart manufacturing studies is that converting existing industrial plants with often limited IT structures so that they can take advantage of smarter process solutions is not a trivial task. Kusiak [25] indicated that one of the issues for smart manufacturing solutions in industrial systems is a lack of sufficient process data. A research opportunity is present within the development of holistic and transferable monitoring frameworks that are adaptable to real-world processes where the possibility of accurate metering of every parameter is not given [94].

The importance of this need can be demonstrated within purified water and steam systems as relevant technical building services of manufacturing facilities. Although these systems require large shares of water and energy within an industrial facility, partly due to the water-energy nexus, detailed metering strategies are not comprehensively addressed in the literature. According to Walker et al. [84], this absence of in-depth process knowledge leads to a high unawareness resulting in excessive water and energy consumption. Authors, such as Thiede et al. [62], are calling for more research contributions that demonstrate the relationship between water and energy within industrial systems.

To complete the data acquisition from industrial processes, proxy metering devices (PMDs) enable the estimation of parameters that are not attained by a physical meter. A PMD acts as an inferential estimator by combining available process data in a regression model. The application of this type of virtual sensor requires significant effort since several development steps are needed, including data pre-processing, model selection, and validation. Due to this effort and the focus in the literature of ever advancing mathematical PMD solutions, the number of applied PMDs in industrial places is limited to date. Luttmann et al. [140] pointed out that more simplified implementation strategies for PMDs are needed for industrial application. In addition, small dataset sizes, which are common in the industry, represent a further constraint for current PMD modelling and require the application of adjusted development strategies [171]. A research opportunity is available in the presentation of targeted and straightforward PMD development steps for industrial applications that can account for the additional requirements and complexity of small datasets.

Chapter 3

Resource measurement strategy for technical building services



Literature review

Chapter 2

- Smart manufacturing development resulting in need for enhanced process data
- Purified water and steam systems as important manufacturing TBS
- Metering audit strategies and limitations in manufacturing facilities
- Existing proxy meter development approaches for inferring data gaps

Resource measurement strategy for TBS

Chapter 3

- Development of a metering framework for TBS
- Application of the framework to two purified water and one steam system of a case facility and analysis of the results
- Simplified proxy metering strategy to fill data gaps
- Value summation for total cost of purified water and steam

Proxy measurements for small datasets

Chapter 4

- In depth study of PMD performances in dependence of the regression complexity
- Experimental rig setup for PMD data acquisition
- Proxy meter development with linear and non-linear algorithms
- Application of small dataset improvement modifications

3.1 Case study system description

The objective of the industrial case study was to develop a metering framework for complex technical building services (TBS) from an industrial perspective. Three TBS, two purified water (PW) and one steam system have been selected for the in-depth study and framework development. The selected systems were part of a life science company located in the south of Ireland. The facility was one of the biggest single water consumers in the country, with an average water intake of 700 m^3/day. A large proportion of this incoming water was running through purification systems to remove undesirable contaminations from water. Once purified, the water was used in several production processes, mostly for solvent extracting and cleaning purposes. The remaining smaller proportion of the incoming water was used for cooling towers, steam boilers, and facility utilities (toilets, showers, canteen, and drinking water). A water assessment report from

2015 analysed the water quantities that have been used in the company for a one year period (11/2014 – 11/2015). Figure 3.1 [a] and [b] show the result of the report in the form of a proportional water usage distribution. As pointed out in Figure 3.1 [a], more than two-thirds of the total mains water intake was purified and subsequently used by the production. The second graph [b] in Figure 3.1 divides the proportional distribution of the PW into the single production lines. In the case facility, every production line had its own water purification system. IC2 and IC1, the two largest production lines in the case facility, used 71% of the total PW and 49% of the total mains water. Due to their large water intake, these two PW systems have been selected for an in-depth study based on the metering methodology which is introduced in section 3.2.

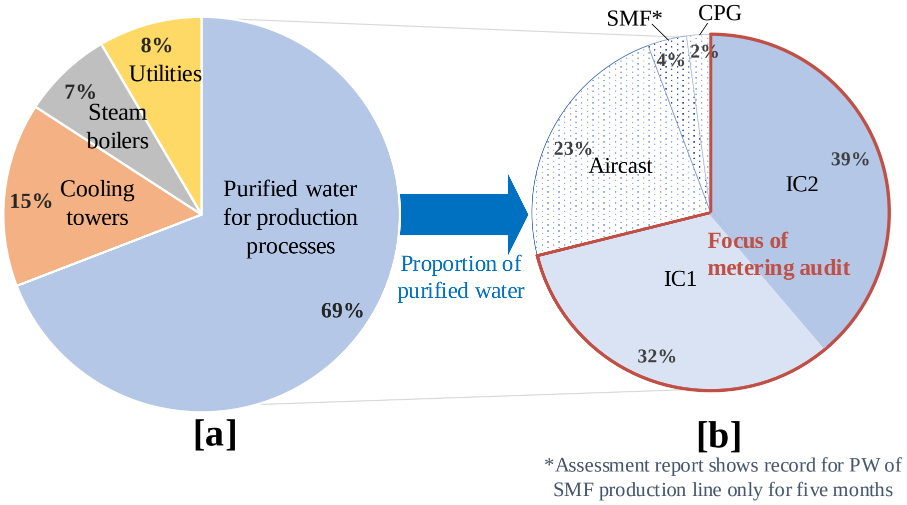


Figure 3.1: Proportions of water and natural gas usage in industrial case facility: mains water intake [a], PW for production processes [b]
SMF, CPG, and Aircast were further production lines within the case facility

In addition to mains water, the facility consumed 11,870 m^3 (126 MWh) of natural gas on a daily average in 2016. A majority of this gas was consumed to generate steam. Due to the high energy consumption, the later presented TBS metering methodology in section 3.2 has been applied additionally to the facility's steam system.

3.1.1 Purified water system layout in case facility

Water purification process

Due to process requirements and governmental restrictions the incoming mains water, supplied from the municipal utilities, had to be purified before usage in the production lines of IC2 and IC1. As outlined in the literature review in section 2.2.1, purification is a water treatment process that is applied in life science and pharmaceutical facilities to ensure high-quality water. In the case facility, the water purification process for IC2 and IC1 consisted of a series of water purification devices (WPDs) that are listed below.

1. A preliminary filter with a mesh size of 25 µm to remove coarse particles.
2. Two water softeners in parallel mode to lower the hardness of incoming mains water (calcium, magnesium, and other metal cations).
3. An activated carbon filter to remove chemicals, particularly organic compounds.
4. Reverse osmosis unit with semipermeable membranes to remove nanoparticles and ions from water.

5a. First ultrafiltration unit (blending-ultrafiltration) to remove small particles.

5b. Second ultrafiltration unit (polishing-ultrafiltration) as an additional membrane filtration process.

Figure 3.2 depicts the process layout of the water purification system with the referring labels for IC2. The process layout for IC1 differs slightly to the IC2 layout and is shown in the appendix in Figure D.1.

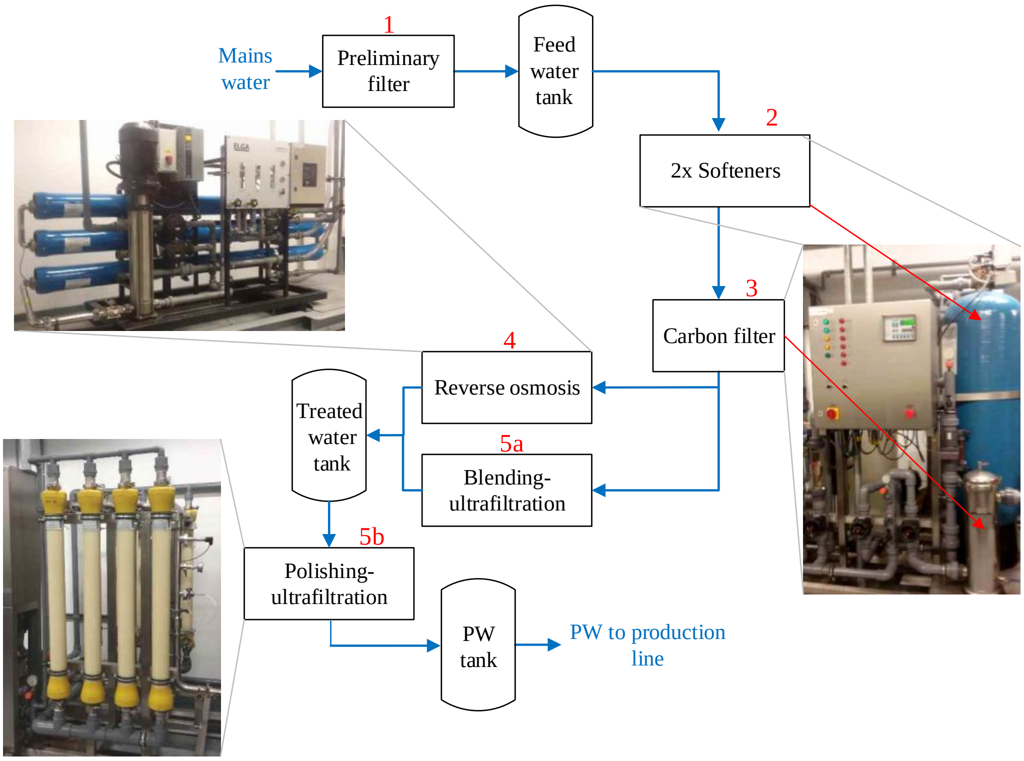


Figure 3.2: IC2 WPDs in case facility

In addition to the WPDs, different buffer tanks were installed within each system which served as intermediate storage and allowed periodic run intervals. Due to product-specific requirements, the blending-ultrafiltration unit and the reverse osmosis unit were set up in parallel.

Within each system, metering devices were installed to measure flowrates and further parameters, such as temperature and conductivity. Figure 3.3 illustrates the block flow diagram (BFD) with all WPDs, and the main flow (FI) and level indicators (LI) including their location for IC2 (appendix for IC1: Figure D.2). All flow meters were offline devices apart from one online flow transmitter (FT1), which recorded the passed volume of mains water and was linked to the company's supervisory control and data acquisition (SCADA) system. Three level meters were used for the monitoring of the tanks filling levels for IC2, two offline level indicators (LI1 and LI2) and one online level transmitter (LT3). Furthermore, temperature sensors and water quality meters were installed at different locations throughout the system.

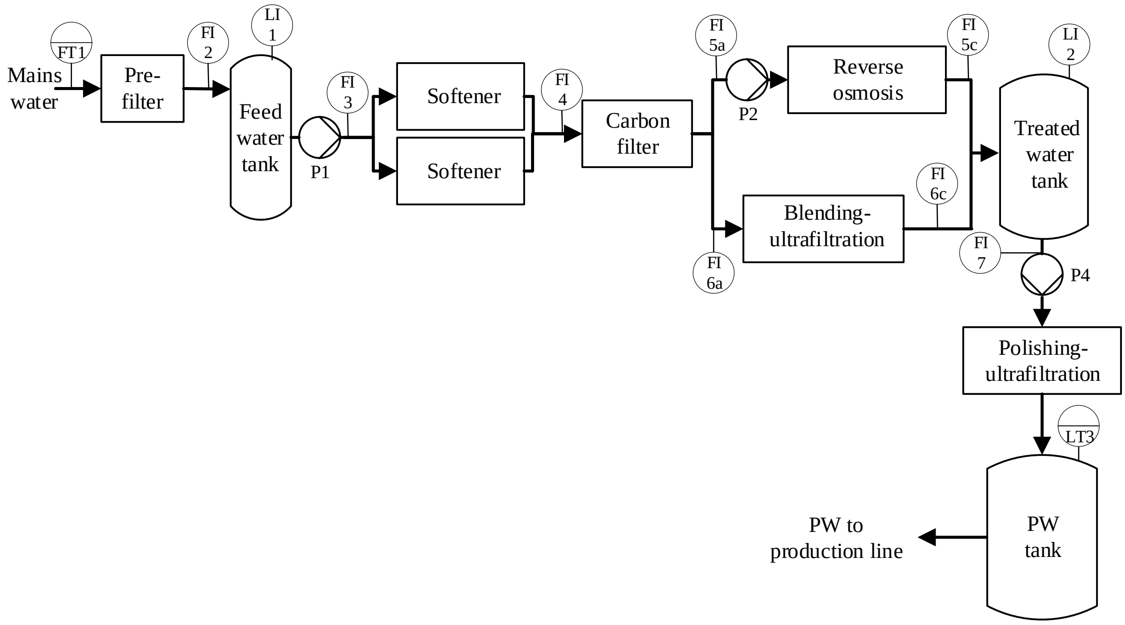


Figure 3.3: BFD of IC2 water purification process including flow and level indicators / transmitters

Purified water usage in industrial case facility

Subsequently to the PW generation in IC2 and IC1, the PW was distributed and used by machines in the production line. The PW storage tank functioned as a linkage between PW generation and PW distribution system. Figure 3.4 visualises the IC2 PW distribution system with all PW intake locations and the installed flow and level indicators (appendix for IC1: Figure D.3). The PW was flowing in a loop, while the unused PW was returned to the PW tank. This layout guaranteed constant PW movement to prevent biological contamination.

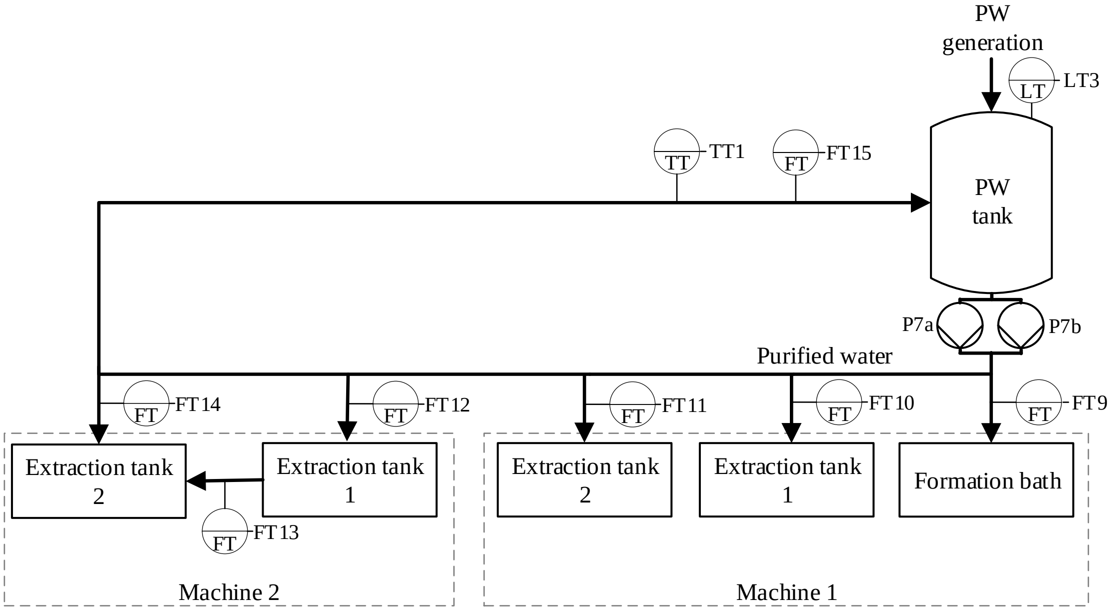


Figure 3.4: BFD of IC2 PW distribution including flow and level transmitters

The installed measuring devices within the PW distribution for IC2 were all online meters and attached to the company’s SCADA system. For IC1, most of the flow indicators were analogue meters, not attached to the SCADA-system (Figure D.3 in the appendix for IC1).

3.1.2 Steam system layout in case facility

The steam generation system of the case facility is depicted in Figure 3.5. The majority of the steam was generated by steam boiler 1. Steam boiler 2 and 3 were mainly kept in standby mode as a backup to boiler 1. Once the steam was used in a variety of usage processes throughout the plant, the condensate was returned and collected in a feedwater tank. The recirculation of the condensate to boiler 1 involved a condensate and deaerator tank in which the feedwater was pre-heated and dissolved gases were removed. For increased system efficiency, an economiser was implemented for further pre-heating of the boiler 1 feedwater by capturing waste heat from burned flue gases. Steam and condensate losses were complemented with the addition of softened mains water to the condensate tank. For system monitoring and control, a variety of temperature, pressure, and flow rate indicators as well as transmitters were implemented throughout the system, as shown in Figure 3.5.

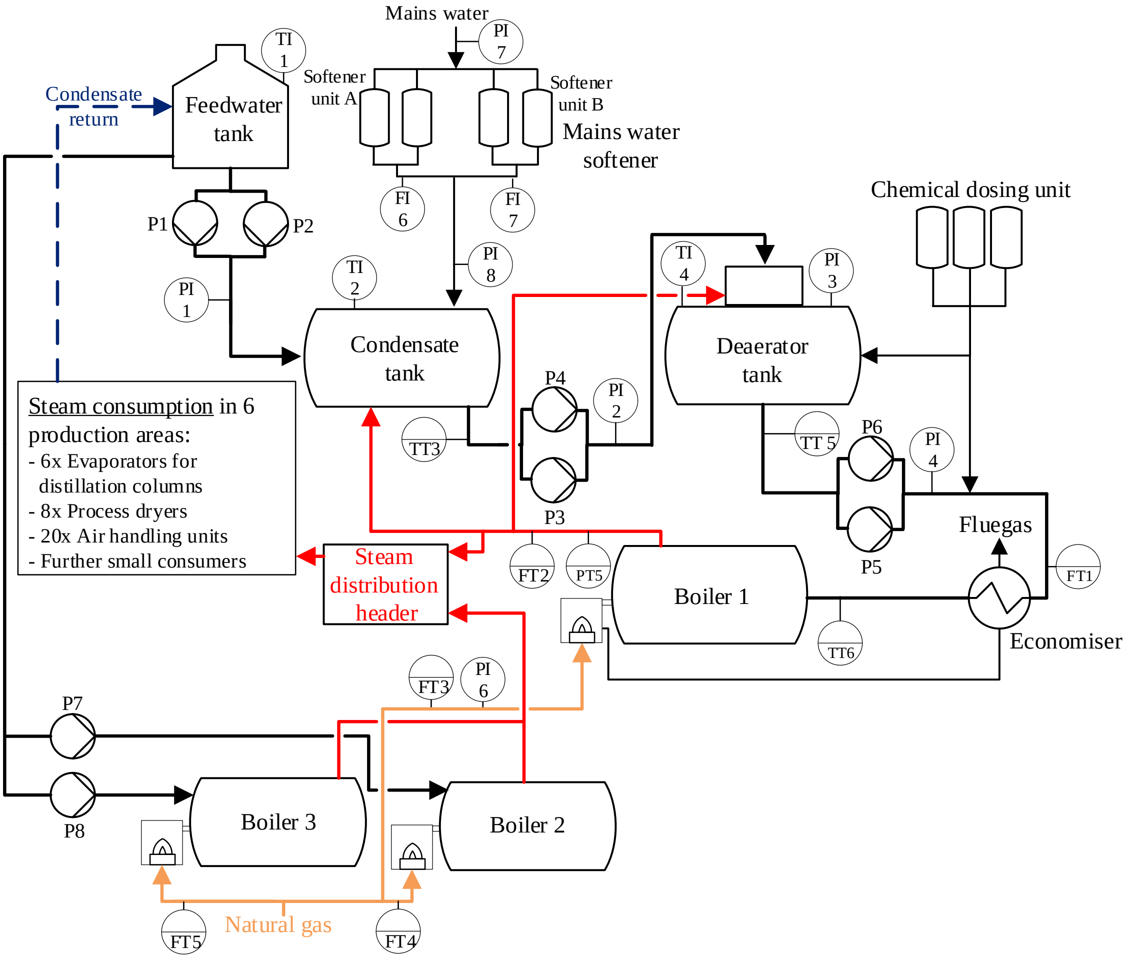
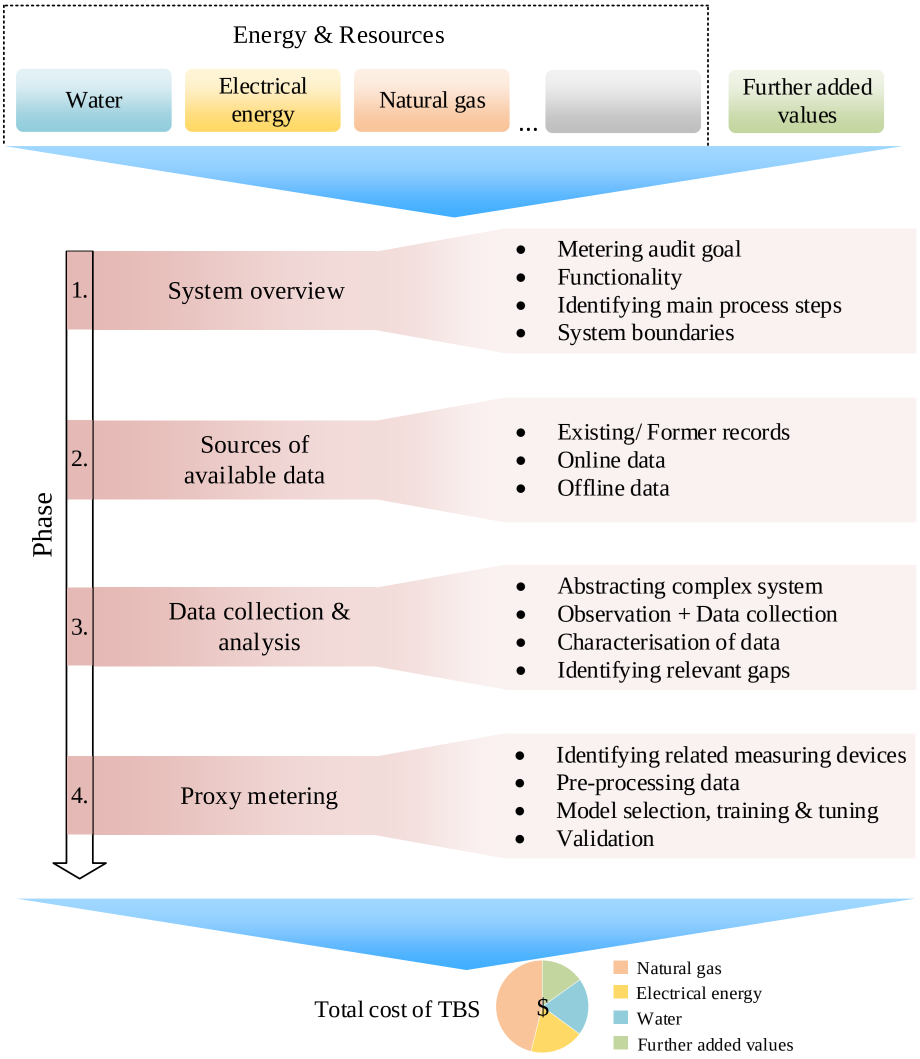


Figure 3.5: BFD of the steam generation system in case facility

The generated steam was distributed throughout the facility and used within six production areas. A variety of steam usage processes existed within the six production areas, whereby main steam consumers were evaporators for distillation columns, process dryers, and air handling units.

3.2 Methodological metering framework

The proposed framework for TBS metering campaigns has been developed during the investigations of the PW and steam systems in the case facility. This approach was chosen to consider circumstances and constraints of real industrial processes and to ensure the feasibility of the developed steps to a variety of industrial systems. The segmentation of the single metering phases is based on the in-depth approach for process life-cycle inventories introduced by Kellens et al. [133], [134]. In the first phase of the presented methodology by Kellens et al. a goal and scope definition sets system boundaries and identifies functional units, Figure 2.17. The next phase consists of a detailed energy, consumables, and emission study of manufacturing equipment. In correspondence to this metering strategy, this research proposes a metering framework for industrial TBS systems, subdivided into four steps as depicted in Figure 3.6.



- Metering audit goal
- Functionality
- Identifying main process steps
- System boundaries

- Existing/ Former records
- Online data
- Offline data

- Abstracting complex system
- Observation + Data collection
- Characterisation of data
- Identifying relevant gaps

- Identifying related measuring devices
- Pre-processing data
- Model selection, training & tuning
- Validation

Figure 3.6: Structured framework for TBS metering audits

Phase 1: System overview

The first phase of the framework for industrial TBS metering audits is designed to obtain a general overview of the selected system. As a first step, a goal of the metering audit has to be defined, e.g. to identify energy saving potentials or to acquire data to analyse the system from a life cycle perspective. At the end of the audit, the results determine if the goal has been reached or if the scope of the metering campaign needs further expansion.

Every TBS system needs electrical energy and other resources, for example, water and natural gas. In addition, further added values are indispensable for operation, which are individual for every system. Within the first phase of the methodology, all of these input and output flows for the TBS system have to be analysed. This first system screening leads to an increased understanding of the process which is further enhanced by the use of Piping and Instrumentation Diagrams (P&IDs) in connection with operating staff. The process knowledge gained enables to identify the main process steps and system architecture. The operating phase of the TBS system is then isolated by applying suitable system boundaries, disregarding the further processes that are attached.

Phase 2: Available data sources

Within the second phase, available data sources are identified. Three different data sources are typically available in a manufacturing context: existing data from former records (e.g. former metering audits or economic data from utility bills), constantly measured data that is available through an online monitoring system (for example SCADA), and data from offline measuring devices that are not connected to the online monitoring system. All three sources are used for the data collection and analysis in the following phase. To be functional in the metering audit, all recorded or measured parameters must be quantified with their specific unit. Furthermore, the location and functionality of installed metering devices have to be checked to identify if all relevant data can be recorded or if data gaps exist.

Phase 3: Data collection and analysis

To ensure an efficient data collection from complex TBS in the third phase of the metering framework, the complex system is abstracted in a simplified version where only relevant process steps and flows are included as depicted in Figure 3.7. The identification of relevant process steps and flows is based on the former system understanding of phase one. For information capture, a flow diagram of the abstracted system is complied.

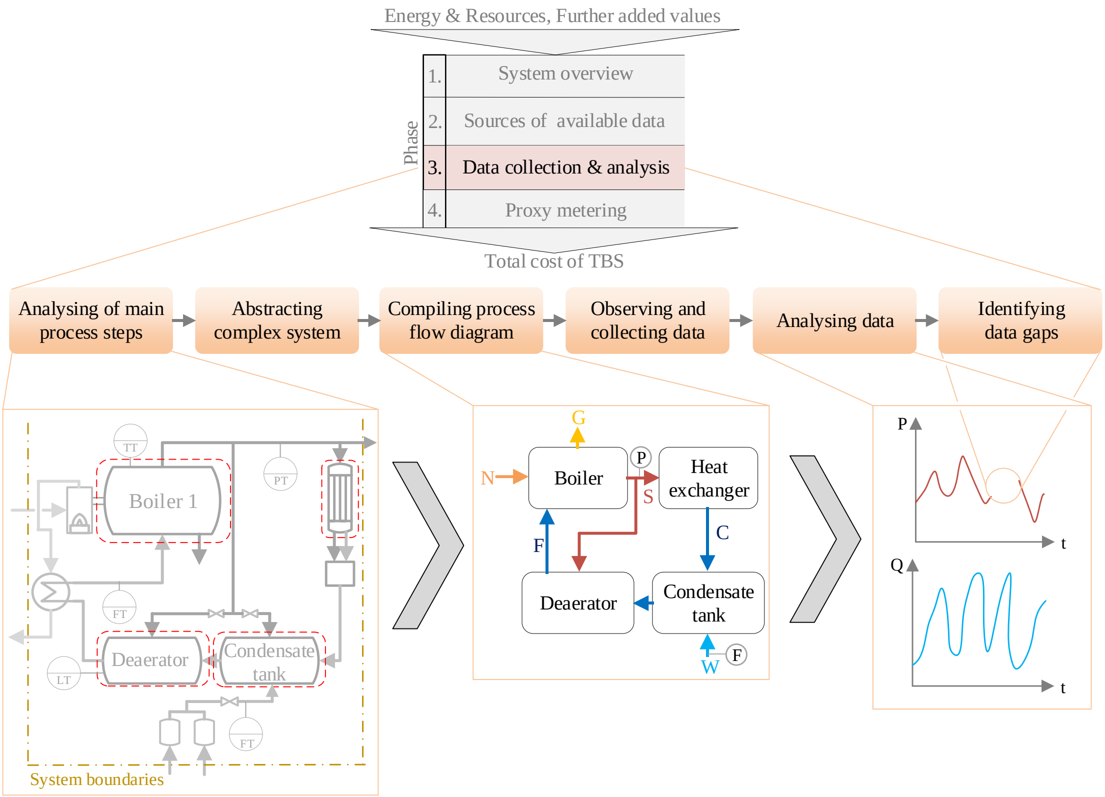

Energy & Resources, Further added values
Phase
1. System overview
2. Sources of available data
3. Data collection & analysis
4. Proxy metering
Total cost of TBS
Analysing of main process steps
Abstracting complex system
Compiling process flow diagram
Observing and collecting data
Analysing data
Identifying data gaps
TT
PT
Boiler 1
FT
LT
Deaerator
Condensate tank
FT
System boundaries
G
N
Boiler
P
S
Heat exchanger
F
C
Deaerator
Condensate tank
W
F
P
t
Q
t

Figure 3.7: Metering framework phase 3: Data collection and analysis

Subsequently, the system is observed over a time period to identify different run modes and run times of the devices, while taking readings of all available metering devices. The duration of the observation typically ranges between 1 to 10 weeks and depends on the system dynamic, number of different run modes, and system size. To decrease the required observation time for complex systems, only working conditions with high shares of the covered period are considered. In general, the focus during the data collection phase is on normal working conditions, rare events are not recorded within the metering audit.

During and after the observation of the TBS, the gathered data is analysed and characterised. This demonstrates whether the dynamic behaviour of single parameters is adequately recorded and enables the identification and calculation of Key Performance Indicators (KPIs). The direction of the analysis depends on the goal of the metering audit which has been set in phase one. Relevant data gaps in the recordings can be addressed by an expansion of the metering scope. However, if the data gaps result from missing, broken, or only offline available measuring devices, proxy metering can close this gap in the next phase, described in further detail in section 3.3.

3.2.1 Application to the purified water systems of the case facility

The metering strategy, as described above, was applied to the IC2 and IC1 water systems in the industrial case facility to gain a detailed water map for each system. The systems required water and electrical energy as main resources, natural gas was not consumed.

Phase 1: Purified water system overview

This metering campaign aimed to analyse the performance of the system, to identify saving potentials, and to collect data for the subsequent calculation of the total cost for water purification. Following from the metering framework in Figure 3.6, technical and economical requirements were initially checked. System boundaries were introduced to isolate the PW generation and distribution at the production machines, as shown in Figure 3.8. The functionalities of the processes were analysed in consistency with P&IDs and consultation with operating staff. Former water bills were used as an economic basis and showed the dynamic of the water intake in the past.

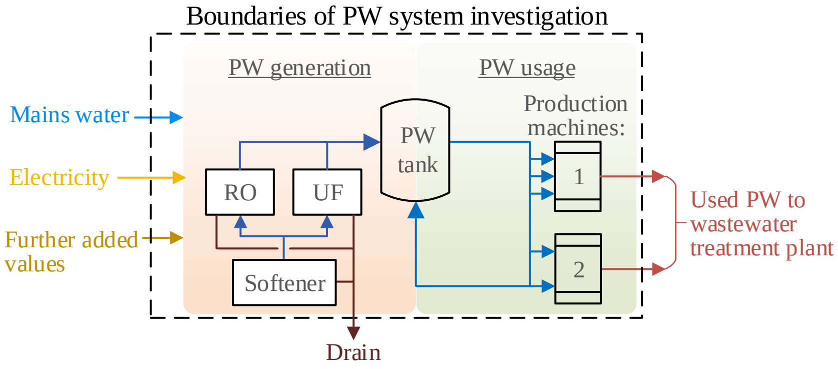


Figure 3.8: Boundaries of PW system investigation

Phase 2: Sources of available data for the purified water systems

Several existing data sources were available for the water systems: As mentioned above, a previous water audit report from 2015 identified the amount of mains water that was used in every process line. In addition, the main parameters within each water system were recorded daily by the facility's staff. These recordings have been used to analyse the dynamic of the parameters and their correlation to different ambient conditions. Performance analyses of each unit in the PW generation systems were carried out every 6 months. The results of these investigations have been archived in reports. Besides these existing data records, measuring devices within the systems have been used for data gathering. Online meters were available through the company's SCADA system. Additionally, offline flow meters were manually observed and demonstrated the current value of the measured parameter.

Phase 3: Data collection within the purified water systems

The data collection phase was carried out for the PW generation systems and the subsequent usage in the production lines of IC2 and IC1. The procedure for collecting the data followed the pattern of Figure 3.7. The process familiarisation led to a more complete process layout of the PW generation and subsequent usage including all flows, for IC2 shown below in Figure 3.9 and for IC1 in Figure D.4 in the appendix. In the first step of phase three, the main WPDs and water flows were analysed. The WPDs, apart from the polishing-ultrafiltration unit, were running occasionally depending on the filling level in the PW tank. The polishing-ultrafiltration unit was running constantly and the permeate was looped back to the treated water tank, once the other devices were in standby mode. The softener and blending-ultrafiltration units showed an additional run mode, in the form of a backwash, which was activated depending on the passed volume or run time (Figure 3.9, Figure D.4). During normal operation of the blending-ultrafiltration within IC2, the drainage flow was reused in the feed water tank. In IC1, the drainage flow of the polishing-ultrafiltration unit was reused in the feed water tank.

The used PW from the production line was sent as drainage flow to the wastewater treatment plant. Furthermore, the reverse osmosis and polishing-ultrafiltration unit were sending a fraction of the incoming water with a high concentration of salinity to drain. In IC2, part of the reverse osmosis drainage flow was recycled in a second, recovery reverse osmosis unit. The permeate of this unit was reused in the facility's cooling towers to save mains water. The recovery reverse osmosis unit was not a relevant step for generating PW, however, it was taken into account for the electrical energy consumption since the unit was attached to the single energy meter of the WPDs.

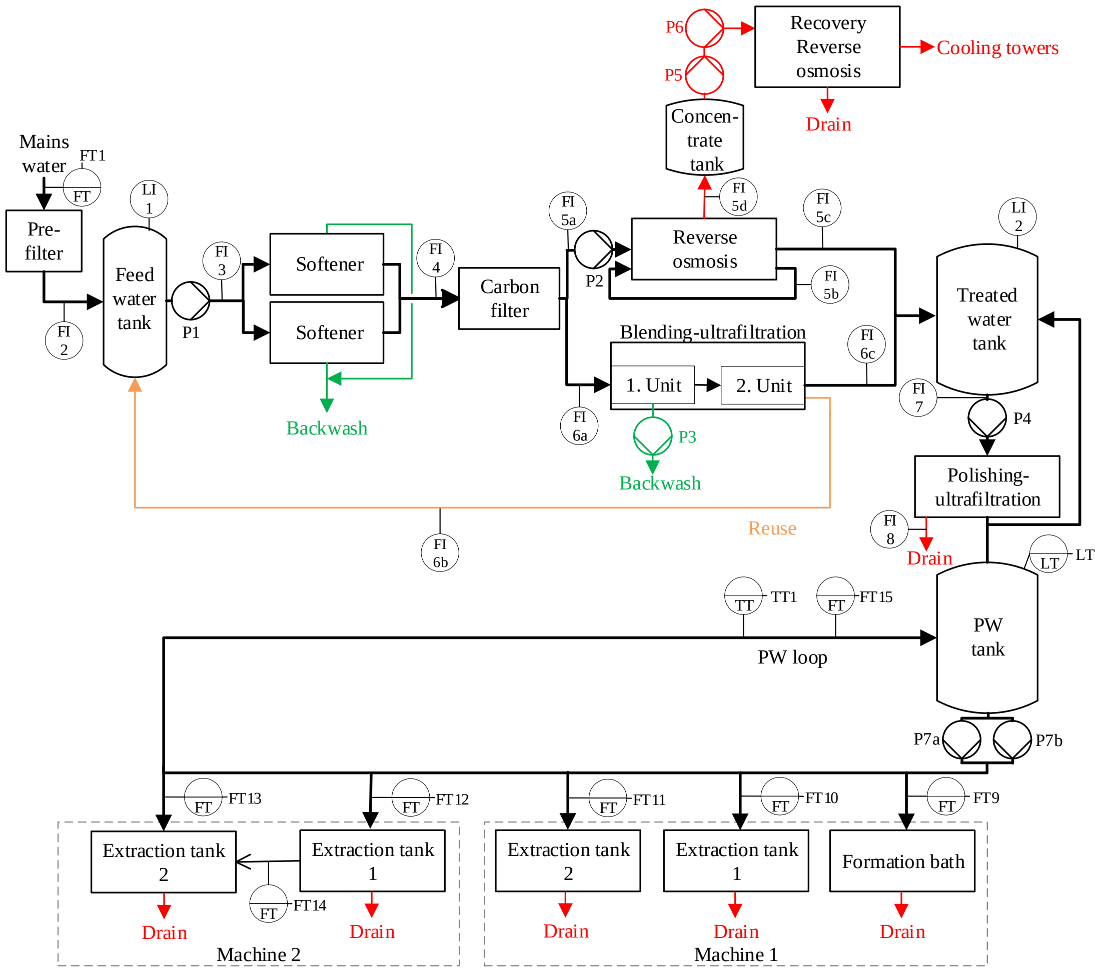


Figure 3.9: BFD of IC2 water system including sub-flows and measuring instruments

To accelerate the data collection process, abstracted versions of the complex water systems were developed, including the main process steps and water flows. Figure 3.10 depicts this abstracted BFD including electricity consumers and further added values for IC2, Figure D.5 for IC1. This step ensured less monitoring was necessary while all the relevant process steps of the water system were still taken into account.

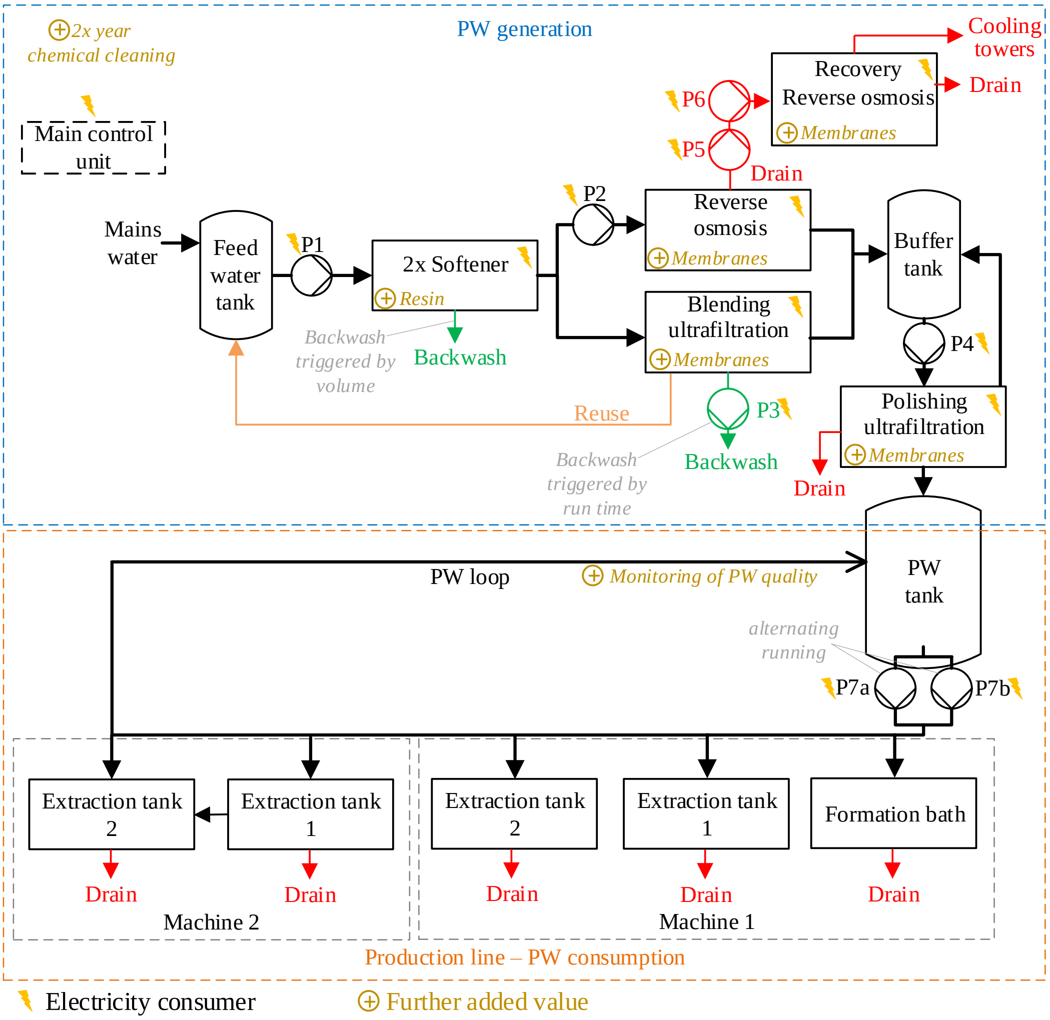


Figure 3.10: Abstracted BFD for IC2 water system including electrical energy consumers and further added values

The process control element within the PW systems controlled the standby and running times for the process units (except the polishing-ultrafiltration unit which was running constantly). Figure 3.11 depicts the control layout for the IC2 PW generation. Once the filling level in the PW tank was decreasing below 60%, the units were switched on until a filling level of 90% was reached in the tank. Furthermore, one of the two softeners went into backwash mode as soon as 134.5 m^3 of treated water, a design specific threshold, was reached. The other softener kept running. During this backwash time, the overall PW production decreased slightly due to the missing softening volume. The blending-ultrafiltration unit entered a backwash mode when a running time of 120 minutes was reached. The backwash of this unit lasted for about one minute, the other devices kept running during this time and the overall PW production decreased about 25% while the backwash was active. For IC1, the start and stop of the blending-ultrafiltration unit were controlled by the conductivity in the PW tank. When the conductivity was rising above 75 µS/m^3 the unit stopped until the value dropped below this threshold again (appendix for IC1: Figure D.6).

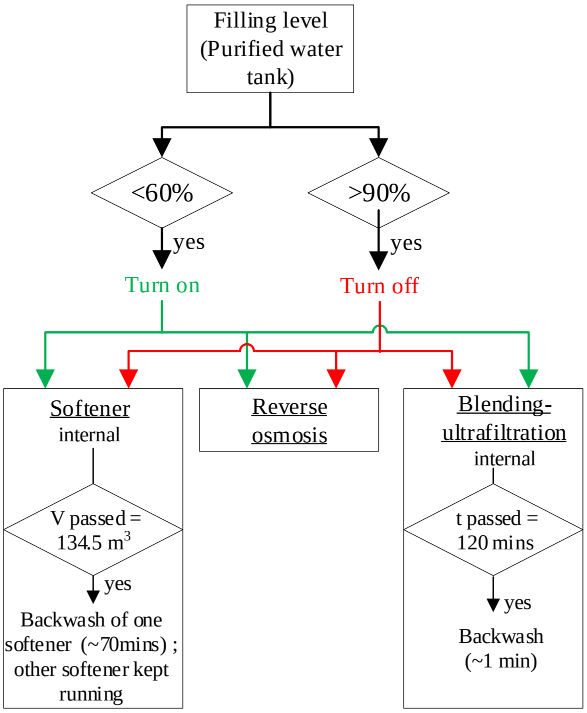


Figure 3.11: Process control layout for IC2 water system

Besides the water flows and process control, the electrical energy consumptions of the water systems were monitored. The largest component of electrical energy was consumed by pumps. Membrane filtering units, such as reverse osmosis, require high pressure flows for their operation, as outlined in Table 2.1. To provide the required pressure, eight pumps of different sizes were implemented in the IC2 PW generation and distribution system, Figure 3.10. In addition to these pumps, the energy was consumed by further smaller devices, such as control units and switch boxes. For IC2, the electrical power of the PW generation system was monitored by one local offline metering device. Additionally, a further offline meter device measured the power that was supplied to the two pumps in the PW loop.

In the IC1 water system, seven pumps have been installed, the location of these pumps is shown in the appendix in Figure D.5. The consumption of energy for the entire IC1 water system (PW generation and distribution) was recorded by an online metering device with a sample rate of 15 minutes. The energy consumed by the blending-ultrafiltration device was not attached to this meter. No energy meter was available for this device.

Results achieved through the implemented metering framework

Results for the IC2 water system

The IC2 water system was observed and data collected over a period of three months (February to April 2017). During this monitoring period, one product campaign was running in the production line (labelled as campaign A). The system was constructed for a second product campaign, which can lead to a varying PW intake of machines 1 and 2. Due to a limited time in the industrial case facility, the monitoring of the second campaign was not possible. The top part of Table 3.1 presents the results in terms of average flow rates and WPDs run times of the PW generation system.

Table 3.1: Water flow measurement results for IC2 water system, Campaign A, with standard deviation values in blue and estimations for unrecorded flows in orange, three month observation period

IC2 PW generation				
Mains water intake [m^3/h]: 6.9 ±*1.1*				
	Permeate [m^3/h]	Drain / Reuse in feed water tank* [m^3/h]	Backwash [m^3/h]	Run time per hour [min]
Softener	6.7 ±*0.7*	-	0.3	31 ±*2.4*
Blending-ultrafiltration	0.9 ±*0.1*	0.1* ±*0.0*	0.1	31 ±*2.4*
Reverse osmosis	3.2 ±*0.3*	2.2 ±*0.2*	-	31 ±*2.4*
Polishing-ultrafiltration	3.3 (9.0 incl. loop)	0.8 ±*0.0*	-	60 ±*0.0*

IC2 PW intake at the production line			
	Formation bath [m^3/h]	Extraction tank 1 [m^3/h]	Extraction tank 2 [m^3/h]
Machine 1	0.5 ±*0.2*	0.6 ±*0.1*	1.4 ±*0.1*
Machine 2	-	0.8 ±*0.5*	0.1 ±*0.0*

In the top part of Table 3.1, three flow rates (backwash of softener and blending-ultrafiltration; permeate of polishing-ultrafiltration) had to be estimated based on former reports due to missing flow meters. The calculated standard deviations indicate the dynamic of the single flows. For all water flows, it could be observed that the dynamic was small once the WPDs were running. However, due to the occasional run characteristics of most WPDs, the varying run time increased the dynamic of the flows per hour. The run time depended on the PW usage at the production line. The metering results of this PW usage at machine 1 and 2 is presented in the second part of Table 3.1. During normal operating conditions, the PW intake of machine 1 and 2 showed small fluctuations due to the regulation via PID-controllers. Increased dynamic behaviour was seen when production was not running normally e.g. for change over or cleaning purposes. To visualise the obtained results, Figure 3.12 depicts the average flow rates of the IC2 PW generation system.

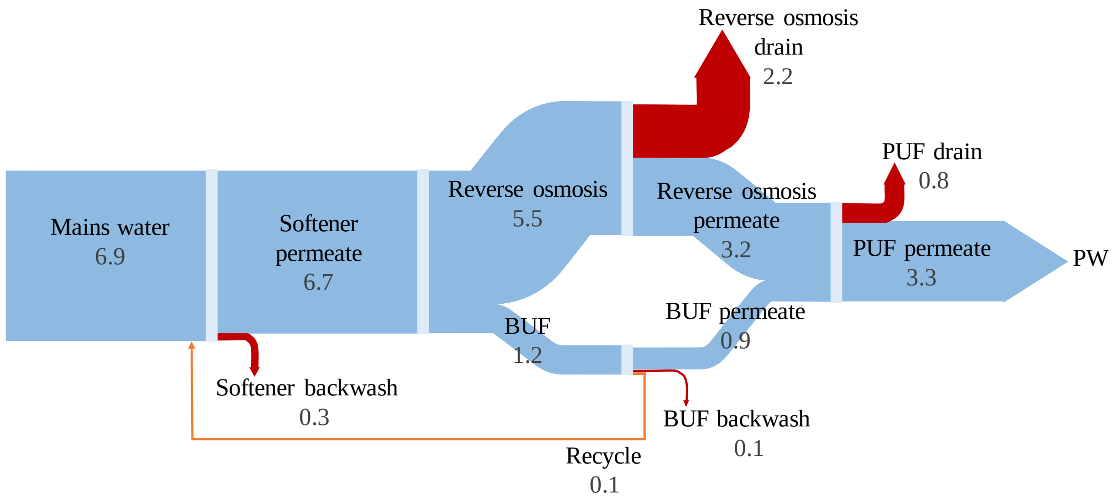


Figure 3.12: Sankey diagram for average flow rates of IC2 PW generation in m^3/h, with BUF: Blending-ultrafiltration, PUF: Polishing-ultrafiltration, due to rounded values the sum may not equal 100%

The results of the electrical power metering of the IC2 PW system are depicted in Table 3.2. The supplied electrical power varied depending on the run mode of the WPDs. While the polishing-ultrafiltration unit and small consumers (control units and switchboards) were running constantly with a steady power supply, other units demonstrated two different power levels depending on running or in standby mode. The electrical power was monitored by taking into account which WPDs were running at the recording time. Due to the rare occasion of the softeners backwash, no power metering was recorded for this specific run mode. The second part of Table 3.2 presents the energy consumption of the PW loop at the production line. The consistent flow rate in the loop was provided by two pumps that were running in an alternating mode.

With the average flow rates and run times of the WPDs from Table 3.1, the entire IC2 water system required an average power of 16.9 kW during the observation period of three months. When the recovery reverse osmosis unit was excluded, the average power decreased to 13.1 kW.

Table 3.2: Electrical power measurement results for IC2 water system; with standard deviation values in blue, three month observation period

IC2 PW generation					
	Polishing-ultrafiltration	Softener + reverse osmosis + blending ultrafiltration	Recovery reverse osmosis	Small consumers (control units, switchboards)	Electrical power [kW]
Run mode	✓	✓	✓	✓	18.2 ±0.2
	✓	✓	X	✓	10.9 ±0.5
	✓	X	✓	✓	11.2 ±0.2
	✓	X	X	✓	3.3 ±0.3

IC2 PW intake at the production line			
	Pump 7a	Pump 7b	Electrical power [kW]
Run mode	✓	X	5.7 ±0.1
	X	✓	6.0 ±0.1

Results for IC1 water system

Equivalent to IC2, the IC1 PW system was observed for a period of 3 months (February to April 2017). During this time one product campaign was running in the production line (labelled as campaign B). The results for the IC1 water system are shown in Table 3.3. Due to the continuous failure of the mains water flow meter throughout the observation period, the mains water intake for IC1 had to be estimated. Similar to the IC2 water system, the flow rates showed small fluctuations once the WPDs were running. Though, this dynamic increased due to the occasional run and standby characteristics of most WPDs, influenced by the PW intake at the production line. In contrast to the IC2 PW distribution, the PW intake of machine 1 and 2 was set manually by operators in the IC1 production line. For visualisation, Figure 3.13 presents the average flows as a Sankey diagram.

Table 3.3: Water flow measurement results for IC1 water system, Campaign B, with standard deviation values in blue and estimations for unrecorded flows in orange, three month observation period

IC1 PW generation				
Mains water intake [m³/h]: 4.1				
	Permeate [m³/h]	Drain / Reuse in feed water tank* [m³/h]	Backwash [m³/h]	Run time per hour [min]
Softener	3.8 ±0.4	-	0.2 ±0.0	43 ±4.1
Blending-ultrafiltration	0.8 ±0.1	0.2 ±0.0	0.1	43 ±4.1
Reverse osmosis	2.1 ±0.2	0.7 ±0.1	-	31 ±3.0
Polishing-ultrafiltration	2.8 (5.5 incl. loop)	0.3* ±0.0	-	60 ±0.0

IC1 PW intake at the production line			
	Formation bath / Extraction tank 3 [m³/h]	Extraction tank 1 [m³/h]	Extraction tank 2 [m³/h]
Machine 1	0.1 ±0.0	0.3 ±0.1	0.9 ±0.1
Machine 2	0.4 ±0.2	0.8 ±0.3	0.3 ±0.1

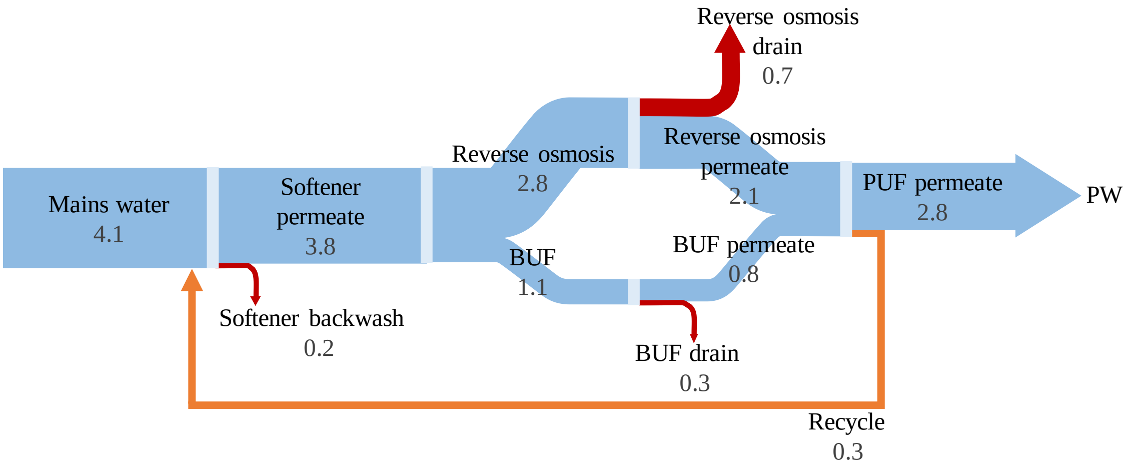


Figure 3.13: Sankey diagram for average flow rates of IC1 PW generation in m^3/h, with BUF: Blending-ultrafiltration, PUF: Polishing-ultrafiltration, due to rounded values the sum may not equal 100%

The consumed electrical energy for the IC1 water system is presented in Table 3.4. In accordance with the results for IC2, the electrical energy has been monitored for different run modes of the WPDs. Due to the rare occasion of the softeners backwash, no power metering was recorded for this specific run mode. As a metering device for the blending-ultrafiltration unit was missing, the energy consumption had to be estimated for this device based on the manufacturers' technical datasheet.

With the average flow rates and run times of the WPDs from Table 3.3, the entire IC1 water system showed an average power of 13.6 kW during the observation period of three months.

Table 3.4: Electrical power measurement results for IC1 water system, with standard deviation values in blue and estimations for unrecorded values in green, three month observation period

	IC1 PW generation and usage at the production line					
	Polishing-ultrafiltration	Reverse osmosis	Blending-ultrafiltration	PW pump	Small consumers (Control units, switchboards)	Electrical power [kW]
Run mode	✓	✓	✓	✓	✓	19.6 *±1.3*
	✓	X	✓	✓	✓	9.1 *±0.8*
	✓	X	X	✓	✓	6.1 *±0.8*

With the obtained data from the IC2 and IC1 PW systems, performance indexes were calculated. Table 3.5 shows KPIs that are used by the case facility for system characterisation and benchmarking. The significant difference of the PW generation efficiencies for IC2 and IC1 are discussed in section 3.2.3.

Table 3.5: KPIs for IC2 and IC1 PW system, Campaign A for IC2 and campaign B for IC1, three month observation period

KPI	IC2	IC1
PW generation efficiency (PW / Mains water intake)	48 %	68 %
Backwash ratio (Backwash / Mains water intake)	8 %	6 %
Drainage ratio (Drain flow / Mains water intake)	43 %	20 %
Electrical energy per m^3 PW	5.1/4.0* kWh/m^3	5.0 kWh/m^3

*: Electrical energy per m^3 PW calculated excluding the electricity requirement of the recovery reverse osmosis unit

3.2.2 Application to the steam system of the case facility

Additionally to the PW systems, the metering strategy has been applied to the steam system of the case facility. A major focus during this investigation was the steam generation of boiler 1, since this device generated the vast majority of steam and accounted for a high proportion of the facility's natural gas consumption.

Phase 1: Steam system overview

This metering campaign aimed to acquire data to show the interconnection between different resources for steam generation systems and to identify the main steam consuming production areas. The boundaries of the metering campaign were set to only include the steam generation, distribution, and condensate return system, as presented in Figure 3.14. The further attached processes were excluded. The functionality of the system was studied with P&IDs and the help of operating staff.

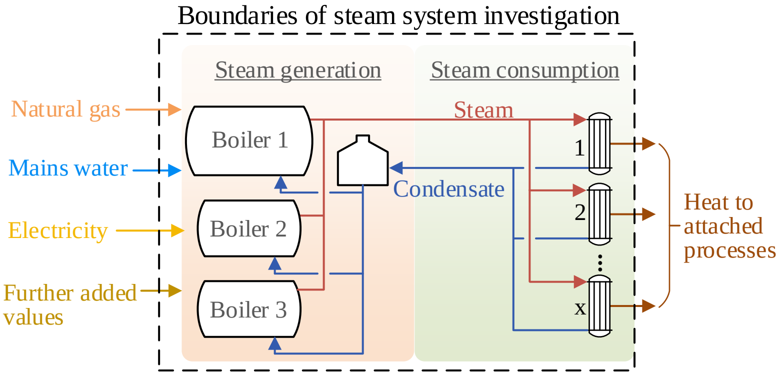


Figure 3.14: Boundaries of steam system investigation

Phase 2: Sources of available data for the steam system

Similar to the PW systems, several existing data sources were available for the steam system: A former steam metering audit has been conducted in 2014 and focused on the steam boilers and steam consumption. Several offline and online meters (linked to the companies SCADA-system) were distributed throughout the system. Figure 3.5 depicts these metering devices within the steam generation process.

Phase 3: Data collection within the steam system

The steam system was observed and data collected to characterise the performance and dynamics of different devices and flows. In the first step of phase three, the main devices in the steam generation system and the main steam consumers were analysed. An abstracted BFD of the system was compiled and is presented in Figure 3.15.

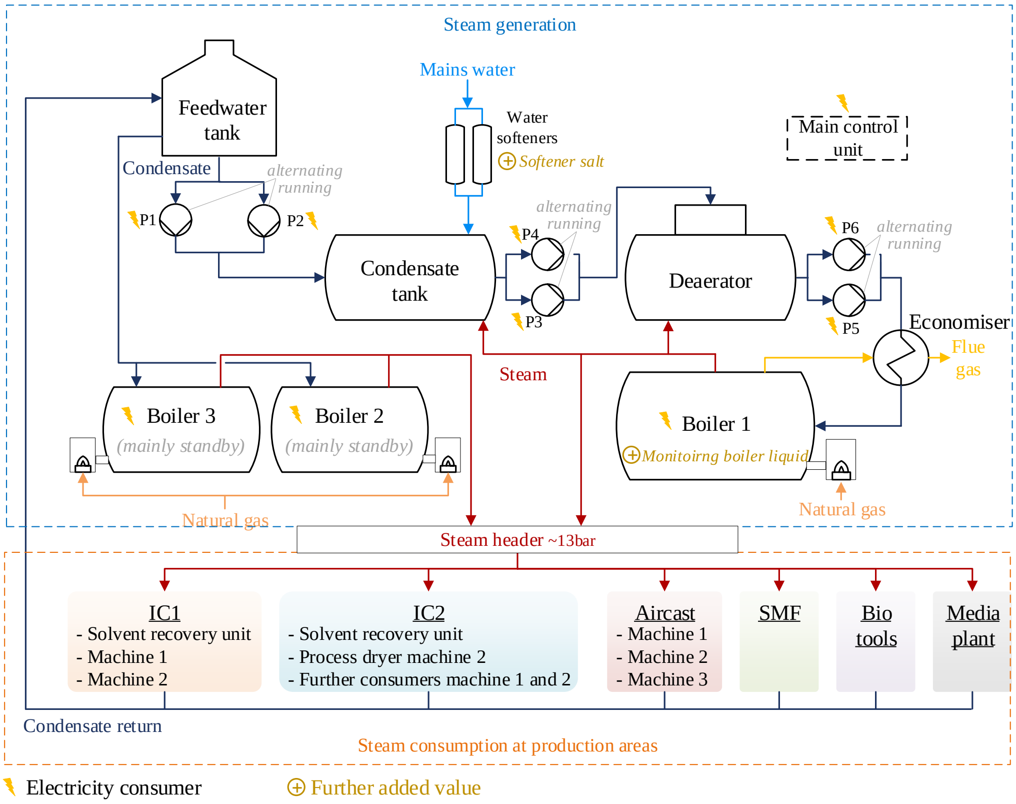


Figure 3.15: Abstracted BFD for the steam system, including electrical energy consumers and further added values

During the observation period, steam boiler 1 was running continuously, consuming the largest proportion of natural gas and generating the vast majority of steam. The amount of consumed natural gas and generated steam was monitored through online flow meters. Boilers 2 and 3 were kept in hot standby mode, generating small amounts of steam. Similar to boiler 1, online flow meters were in place to measure the consumed natural gas and generated steam for boiler 2 and 3. However, comparisons between the consumed amount of steam that was shown in the companies SCADA system and the locally displayed values of the flow meters revealed that deviations were present for boiler 2 and 3. To overcome these inconsistencies, local readings of the accumulated amount of steam that was generated, have been used in the subsequent results section. In addition, the company’s facility team was informed about the misleading information displayed in their SCADA system. To make up for steam and condensate losses, softened mains water was added to the condensate tank. This water addition was controlled by the filling level of the condensate tank and activated as soon as the filling level dropped under a fixed threshold. No online meter was in place to measure the intake of mains water, therefore, it had to be estimated with a PMD (further described in section 3.3.2). Furthermore, electrical energy was consumed at various stages throughout the steam generation unit, mainly for pumps and boiler fans. Six pumps were installed in the system, whereby only three pumps were running at the same time due to alternating run modes. Only the status of the pumps was monitored via the companies SCADA system. Therefore, the electricity consumption of the steam generation system had to be estimated based on the states of the single units and the electricity consumption from component manuals. Besides these resources, further added values were required in the form of salt for the water softeners and periodical monitoring of the boiler liquid.

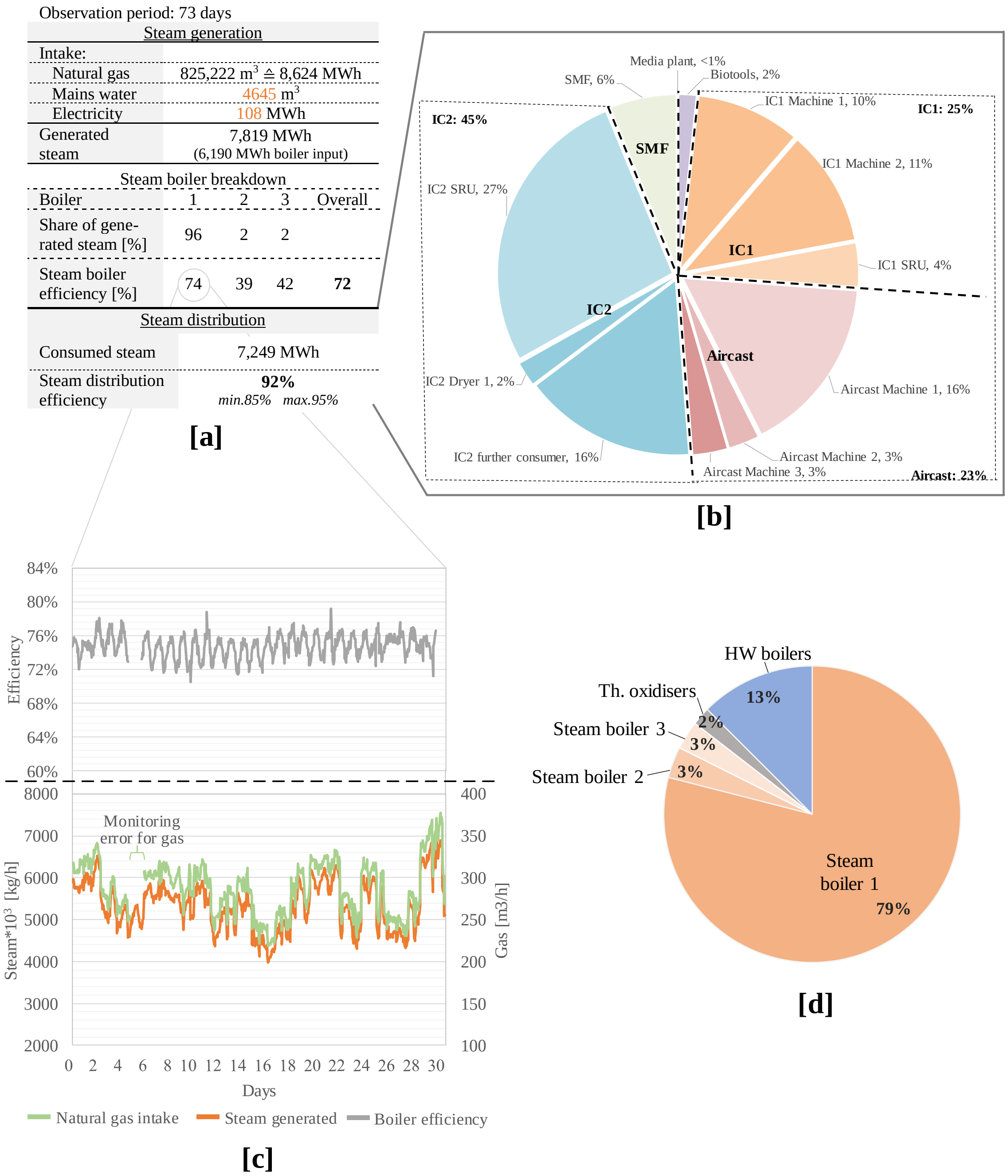
The generated steam was consumed in six production areas. The main consuming units were distillation columns in solvent recovery units, process dryer units, and air handling units. Within most steam consumption processes, the steam was condensed and the condensate returned to the feedwater tank. Only in one distillation column, steam was directly injected into the column body. To measure the consumed amount of steam by the different facility areas, online flow meters and local pressure indicators were available. Similar to boiler 2 and 3, comparisons between the consumed amount of steam that was shown in the company's SCADA system and the locally displayed values for some of the flow meters resulted in deviations. Therefore, local readings of the accumulated amount of steam for these devices have been used for the subsequent results. In addition, some steam pressure indicators were broken. For these devices, the pressure was estimated based on archived pressure levels and surrounding pressure indictors.

Results from the implemented metering framework

The steam system was observed and data collected over a period of 10 weeks. Figure 3.16 presents the results for steam generation and steam consumption. The table shown in Figure 3.16 [a] represents the resource and energy intake of the steam generation system and the efficiency of the steam system. On the steam generation side, on average 471 m^3/h of natural gas, 2.7 m^3/h of mains water, and 62 kWh of electricity were consumed per hour. Boiler 1 generated the vast majority of steam (96%) during the observation time. Boiler 2 and 3 remained mainly in standby mode, contributing 2% each to the steam generation. The overall fuel to steam efficiency, calculated according to equation (3.1) (steam enthalpy divided by supplied natural gas energy) was 72%.

$$\eta_{steam\_generation} = \sum_{boiler\_i=1}^{3} \frac{\sum_{t_1}^{t_2} m_{i,steam\_generated} * \Delta h}{\sum_{t_1}^{t_2} V_{i,gas} * \ CV} \tag{3.1}$$

$$\eta_{steam\_distribution} = \sum_{t_1}^{t_2} \frac{m_{p,steam\_consumed} \ * \ h_{p,saturated}}{m_{i,steam\_generated} * h_{i,saturated}} \tag{3.2}$$



Observation period: 73 days

Steam generation				
Intake:				
Natural gas	825,222 m^3 ≙ 8,624 MWh			
Mains water	4645 m^3			
Electricity	108 MWh			
Generated steam	7,819 MWh (6,190 MWh boiler input)			
Steam boiler breakdown				
Boiler	1	2	3	Overall
Share of generated steam [%]	96	2	2	
Steam boiler efficiency [%]	74	39	42	72
Steam distribution				
Consumed steam	7,249 MWh			
Steam distribution efficiency	**92%** *min.85% max.95%*			

Figure 3.16: Results of the steam metering audit in the case facility: table of steam generation and steam consumption efficiency (estimations for not measured values in orange) [a], steam consumption of different production areas [b], Boiler 1 performance analysis [c], natural gas consumption distribution [d]

The second part of the table in Figure 3.16 [a] demonstrates the efficiency of the steam distribution system, on average 92%. This value is calculated according to equation (3.2), where the generated and consumed amount of steam is first multiplied with the saturated steam enthalpy at the corresponding pressure level and subsequently divided by each other. The average value of 92% is slightly higher than efficiency values that can be found for steam distribution lines in the literature. Bhatt [81] suggested an efficiency of up to 90% for steam distribution lines. One reason for the calculated efficiency of 92% is the assumption of saturated steam at the consumer side. Due to missing temperature measurement devices, it could not be verified if the steam was saturated at each consumption point.

Figure 3.16 [b] provides a division of the steam at the consumer side during the observation period. The largest single consumer was a set of distillation columns inside the IC2 solvent recovery unit (SRU) that accounted on average for 27% of the total steam consumption, followed by the Aircast machine 1. A more detailed analysis of the natural gas consumption, steam generation and resulting fuel to steam efficiency for boiler 1 is shown in Figure 3.16 [c] for a time extract of 30 days. The calculated efficiency shows only small fluctuations during this period with a standard deviation of 2%. Steam generation and the corresponding natural gas consumption show higher fluctuations with standard deviations of 602 kg/h and 32 m^3 respectively. Figure 3.16 [d] depicts the natural gas distribution of the case facility for the entire observation period of 10 weeks. On average, 13,145 m^3 of natural gas were consumed per day. The steam system accounted for 85% of the total gas consumption, boiler 1 showed the highest consumption with 79%. Besides the steam system, this graph displays further gas consumers in the form of two thermal oxidisers and two hot water (HW) boilers.

As an additional representation of the results, Figure 3.17 presents the results of the steam metering audit in the case facility in the form of a Sankey diagram. This figure also includes the distribution of natural gas as well as some of the results from the metering audit of the PW systems IC2 and IC1 that were presented in section 3.2.1.

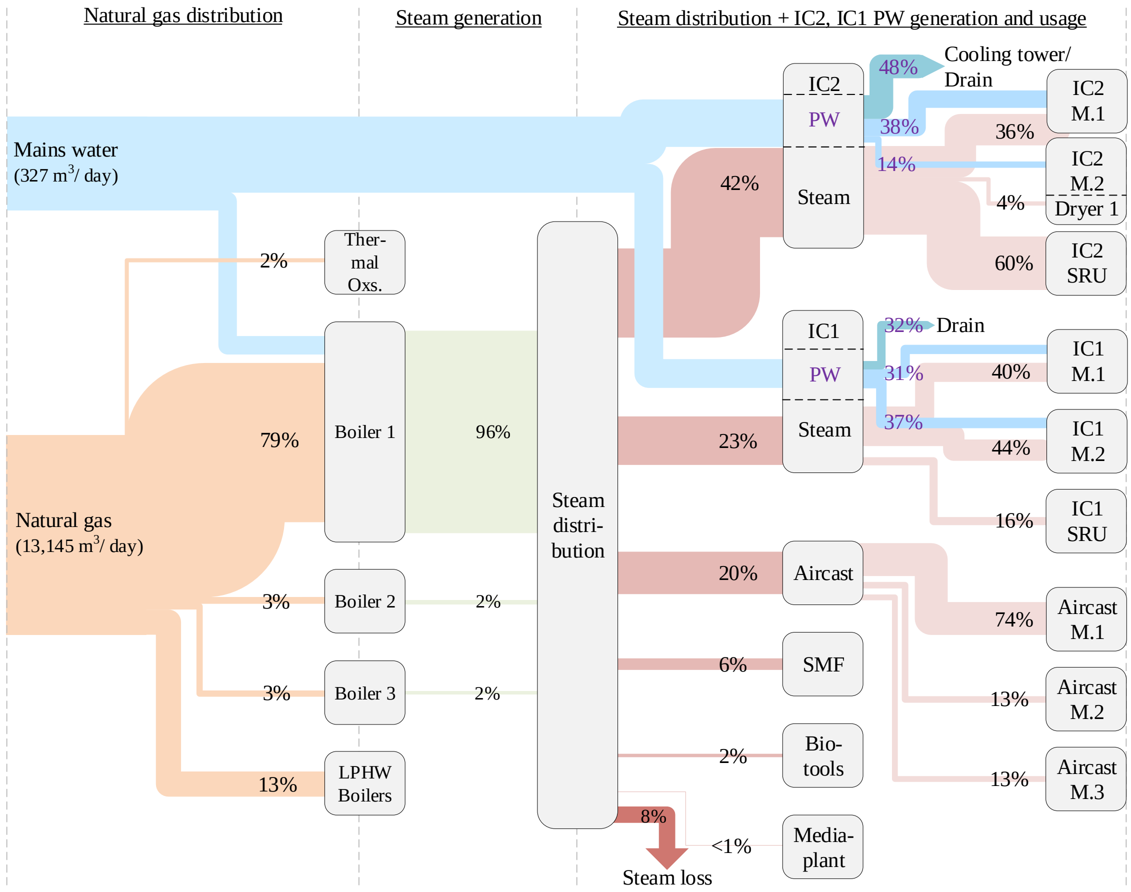


Figure 3.17: Sankey diagram for natural gas and steam distribution of case facility, including PW systems of IC2 and IC1, values for mains water and natural gas intake based on average values of metering campaign, due to rounded values the sum may not equal 100%

3.2.3 Discussion of metering framework application

The application of the metering methodology within the PW and steam systems of the case facility revealed the main energy and resource flows. Due to a high number of offline measuring devices in the analysed systems, the implementation of the metering campaign required repeated readings of the local measuring devices. This was a time-intensive procedure that inhibited a quick data gathering and limited the analysis of system dynamics.

Nonetheless, the results of the metering audits led to a deeper insight into the operation which was not possible before the implementation of this audit due to a lack of information. For the PW systems, the KPIs indicated a higher efficiency in the PW generation of IC1 compared to IC2. Based on the gathered data, it could be identified that this difference was mainly due to divergent performances of the reverse osmosis and polishing-ultrafiltration units. This result was used by the operators as the first approach for possible water-saving opportunities. Further optimisation possibilities were identified, most of them demonstrated an easy to adapt characteristic resulting in a quick and inexpensive implementation. One example was to review if the PW intake could be turned off in the production machines during a production stop. The PW metering campaign revealed that PW was frequently used, although both machines in the production line were stopped for several hours. An example is shown in Figure 3.18, where approximately 45 m^3 of PW could have been saved in the presented time frame of 48 hours within the IC2 production line if the intake would have been turned off. Furthermore, Figure D.7 in the appendix reinforces the assumption of steady PW intake within the IC2 production line independent of the production activity of machine 1 and 2.

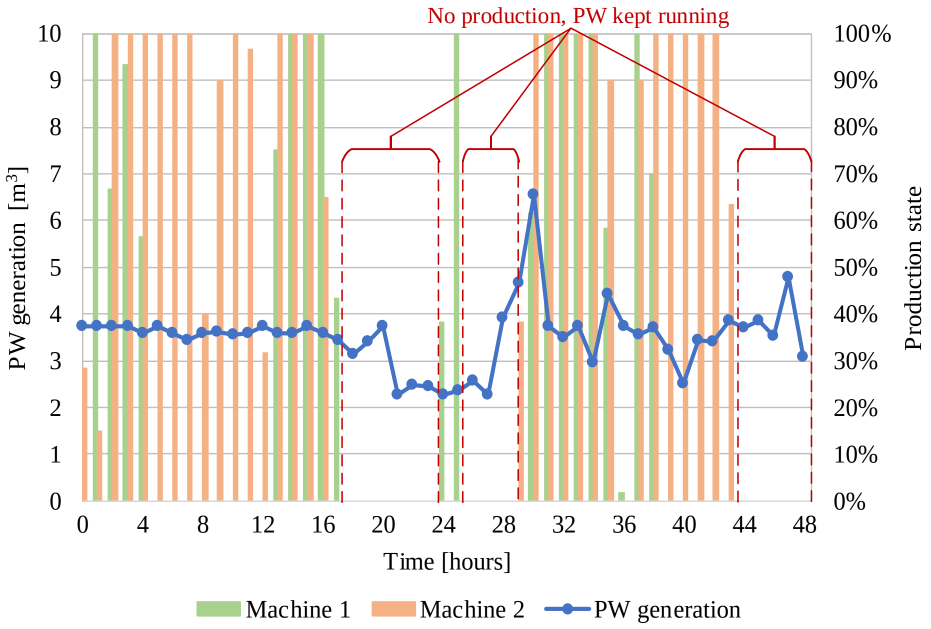


Figure 3.18: Example of IC2 PW saving possibility during production stop of machines 1 and 2

Likewise, the analysis of the data from the steam metering campaign revealed energy saving potentials in the form of steam losses and system reconfigurations. As an example, the receiver vessel of the blowdown from boiler 1 showed a continuous steam discharge. Operators were informed to discuss if the high blowdown rate of boiler 1 is necessary and if the lost steam could be condensed and reused as feedwater in the boiler.

Additionally, the data from both systems can be used for early identification of malfunctions by comparing flow rates and device performances to expected values that were gained during these metering campaigns. Further benefits are the possibility of risk analysis studies. As an example, the operators of the PW systems are now able to assess the maximum production run time that would be left in case one of the PW generation systems would break down.

In terms of metering system limitations, the question arises relating to how detailed the metering campaign has to be. The case studies revealed that not every flow and power supply within the systems was measured by a metering device. Furthermore, not every available meter was working correctly or attached to the online monitoring system. As a consequence, the presented metering approach is suitable for basic estimations over a longer time interval. To allow an overall and automated metering system that enables a more precise study of system dynamics, these data gaps have to be filled. Proxy metering can be a possible method for filling these gaps effortlessly and inexpensively. To improve the accuracy and to allow temporary insights of missing parameters, proxy models have been applied to the estimated values of the IC2 water system and the steam system, presented in the next section. The use of a proxy model for the IC1 water system was not possible due to too many missing and broken online metering devices.

3.3 Simplified proxy metering approach

The fourth phase of the developed metering framework (Figure 3.6) addresses data gaps by the implementation of a PMD. If the data collection of the previous phase three did not show any relevant data gaps, this phase can be neglected. This section presents a simplified proxy metering approach that can be realised with minor efforts within a metering audit. A more comprehensive approach for PMDs is presented in chapter 4.

The proposed methodology for implementing a PMD as part of a metering audit is based on the general steps for proxy meter development that were introduced in chapter 2.3. Figure 3.19 illustrates the methodology for proxy meter development in the scope of this research.

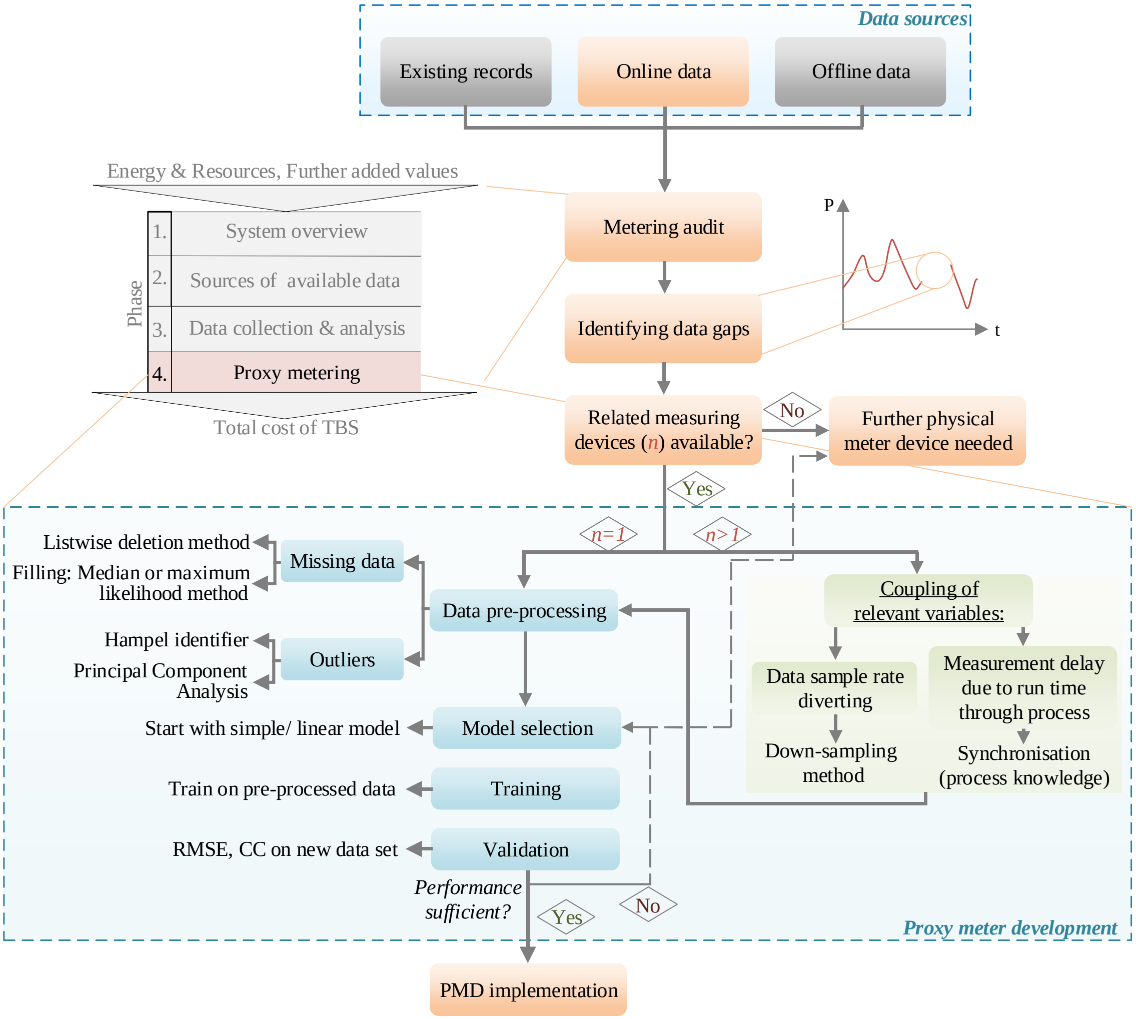

Data sources
Existing records
Online data
Offline data
Energy & Resources, Further added values
Phase
1. System overview
2. Sources of available data
3. Data collection & analysis
4. Proxy metering
Total cost of TBS
Metering audit
Identifying data gaps
P
t
Related measuring devices (n) available?
No
Further physical meter device needed
Yes
n=1
n>1
Listwise deletion method
Filling: Median or maximum likelihood method
Missing data
Hampel identifier
Principal Component Analysis
Outliers
Data pre-processing
Coupling of relevant variables:
Data sample rate diverting
Down-sampling method
Measurement delay due to run time through process
Synchronisation (process knowledge)
Start with simple/ linear model
Model selection
Train on pre-processed data
Training
RMSE, CC on new data set
Validation
Performance sufficient?
Yes
No
Proxy meter development
PMD implementation

Figure 3.19: Proxy meter development framework as part of a metering audit

Pre-development steps

According to the previously introduced framework for metering audits for TBS (Figure 3.6), three different data sources are combined: existing data records, online available data, and offline data. For the implementation of a PMD, only online data sources are considered.

Once data gaps (target variables) are detected in the third phase of the metering audit, related online meters are identified as possible PMDs. The selection of suitable related metering devices is based on the process knowledge that was gained from the previous steps of the TBS metering audit. For the case that no online measuring device relates to the target variable, an additional physical meter has to be implemented in the system. If related devices are identified, the data from these devices is pre-processed in the following phase.

Data pre-processing

When data from related measuring devices are used for a proxy meter model, it has to be guaranteed that no inconsistencies are present within these datasets. Therefore, diverting sampling time, missing data, and outliers are addressed in the pre-processing step.

Missing data is handled either by the listwise deletion method for small missing sections or by filling these gaps with the median or maximum likelihood estimation. Data outliers are identified either by the univariate Hampel identifier or by the multivariate Principal Component Analysis.

If only one variable is selected as being relevant, the pre-processed data can be directly used within a model to estimate the target variable.

If more than one variable is selected, further pre-processing steps are necessary. These include the examination for diverting sampling time between the devices and if necessary their adjustment using the down-sampling method. Furthermore, diverging positions of related meter devices can result in measurement delays due to the required run time through the process. Commonly, these measurement delays vary. As an example, water treatment devices within a PW system run occasionally in many setups, depending on filling levels of buffer tanks. These measurement delays are compensated by synchronisation based on process knowledge. Depending on the process, these issues can lead to an insufficient performance of the PMD which is only solved by installing a further physical meter.

PMD training, tuning, and validation

Subsequently, a model for the PMD is chosen with regard to the process layout and the number of related measuring devices. For cases with a high number of related measuring devices, multivariate approaches like Principal Component Analysis can be used to identify the variables with the highest variance. As previously outlined in the literature section 2.3, Kadlec et al. [145] reported that no unified theoretical approach is available for the selection of a PMD model. Since the focus of this research is on simplified models of PMDs, it is recommended to start with a simple model type. If sufficient process knowledge is available and the variation of the target variable can be attributed to a small number of measurable variables, a linear regression model can be used initially. Once a model is selected, it is trained and tuned based on the pre-processed dataset. The validation on a new set of data shows the performance of the chosen model. To evaluate the accuracy, two validation parameters are applied: the Root Mean Squared Error (RMSE) and the Correlation Coefficient (CC).

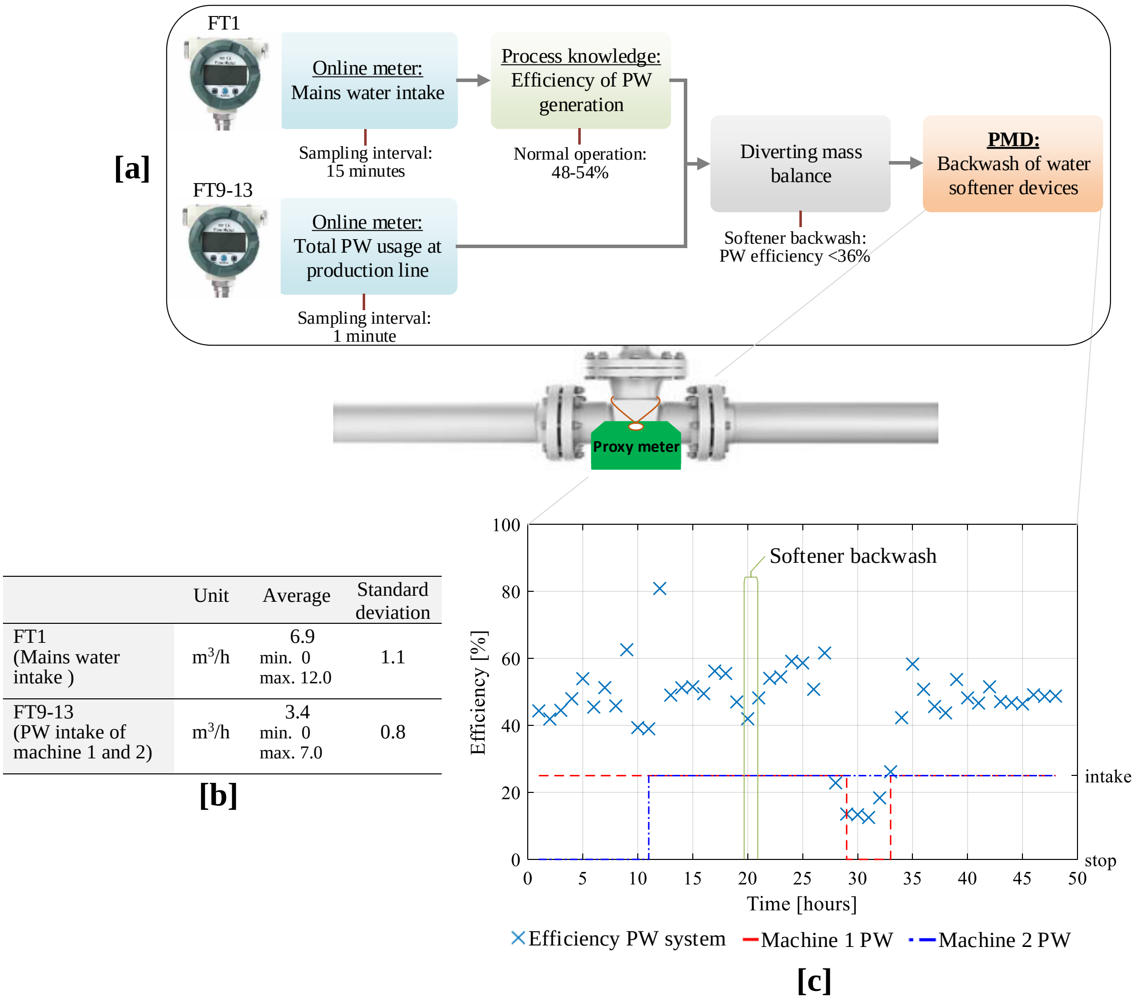
3.3.1 Application to one purified water system of the case facility

The simplified proxy metering approach has been applied to the IC2 water system of the industrial case facility. As shown in Table 3.1 in section 3.2.1, three flowrates could not be measured due to missing metering devices. To increase the accuracy of the estimated values that are shown in Table 3.1, two PMDs have been developed. The first PMD was developed to detect the backwash mode of the water softener devices. Additionally, a second PMD was developed to track the run time of the WPDs and to infer the flow between the polishing-ultrafiltration unit and the PW tank.

The implementation procedure for the two PMDs followed the methodology outlined above and depicted in Figure 3.19. For the two PMDs, the same pre-processing steps have been applied to the datasets of the related measuring devices. Small sections of data were missing due to recording errors of the SCADA system, captured as non-numerical data points. These error values were identified by the models and removed by the listwise deletion method. Furthermore, data outliers were present, many of them could be clearly identified as they were outside the sensors measuring range. The Hampel identifier was applied in the models for identifying and deleting outliers.

First PMD: Backwash of water softeners

For regenerative purposes, one of the two water softener devices in the IC2 PW system was entering a backwash mode after the treatment of a certain volume of water. A PMD was developed to detect this backwash mode, Figure 3.20 [a].



	Unit	Average	Standard deviation
FT1 (Mains water intake)	m^3/h	6.9 min. 0 max. 12.0	1.1
FT9-13 (PW intake of machine 1 and 2)	m^3/h	3.4 min. 0 max. 7.0	0.8

Figure 3.20: PMD for detection of water softener mode in IC2 water system: development methodology [a], characteristics of PMD training data [b], example of PW generation efficiency in relation to PW intake at machine 1 and 2 [c]

Within the preceding metering audit, the mains water intake of the PW system (FT1 in Figure 3.9) and the sum of PW usage at the production line (FT9-13 in Figure 3.9) were identified as suitable related measuring devices. These online devices were attached to two parallel running SCADA-systems in the case facility. FT1 showed data records with a timestamp of 15 minutes in between each record, whereas recordings from FT9-13 were available with a timestamp of 1 minute. To achieve synchronisation, the down-sampling was applied, composing the data from the more frequent variables (FT9-13) to the 15 minute time interval of the less frequent variable FT1. The characteristics of the PMD model input data is displayed in the table shown in Figure 3.20 [b]. The model to detect when one of the softeners was in backwash mode dependent on the diverting efficiency of the PW system:

$$\eta_{PW\_system} = \frac{FT1}{\sum_{i=9}^{13} FT(i)} \tag{3.3}$$

During normal operation, 48-54% of the mains water was converted into PW, equation (3.3). This efficiency margin resulted from performance changes of the reverse osmosis unit due to different mains water temperatures during the year. When one softener entered the backwash mode, the efficiency decreased below 36%. This efficiency deviation was used as an indicator within the PMD model.

Results for the first PMD for the IC2 water system

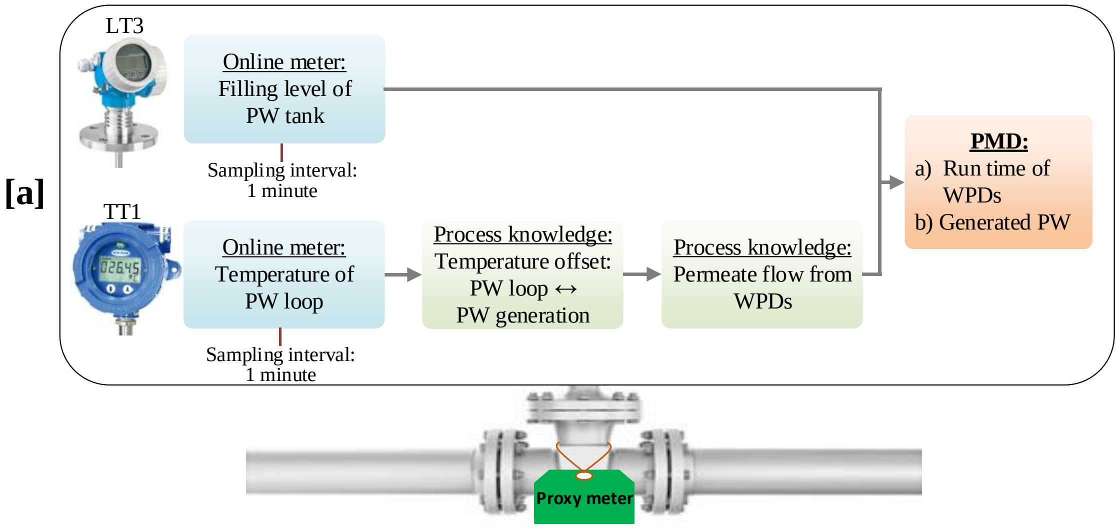
For the detection of the softener backwash, the PMD showed insufficient performance during the model training and tuning phase. The efficiency of the PW generation system, which was used for the detection of the backwash mode, showed a dynamic behaviour. Four main rationales were identified that led to changing efficiencies:

- Due to the position of the related measuring devices, at the mains water intake in the PW generation system (FT1) and the PW intake in machines 1 and 2 at the production line (FT9-13), a time offset of up to 35 minutes emerged. Additionally, this time offset varied, depending on the filling level of the integrated buffer tanks.
- The intake of PW at the production line changed frequently due to the halting of production machines, run mode changes, cleaning purposes, and oscillating behaviour of flow controllers. Due to the time offset between the related measuring devices, these changes were not immediately visible at the mains water intake (FT1).
- The smallest sampling time of 15 minutes did not allow a timely exact identification of the backwash beginning and end, the softener backwash lasted for 70 minutes.
- Within the production line, operators could take out PW from taps for manually cleaning operations. The usage of these taps was not monitored by a measuring device.

Figure 3.20 [c] depicts an example of the efficiency of the IC2 PW system (equation (3.3)) for 48 hours. After approximately 20 hours, one of the water softeners entered the backwash mode. According to the graph in Figure 3.20 [c], a clearly lower PW efficiency is not distinguishable during the backwash mode. After 27 hours, the performance decreased significantly for six hours while no PW was used in machine 1. A possible explanation for the low efficiency during the specific time frame is the usage of PW from un-monitored taps in the production line for cleaning purposes by operators.

Second PMD: Run time of WPDs and flow rate to PW tank

A second PMD was developed to track the run time of the WPDs and to infer the PW flow between the polishing-ultrafiltration unit and the PW tank, Figure 3.21 [a]. As related measuring devices, the level transmitter of the PW tank (LT3 in Figure 3.9) and the temperature transmitter of the PW loop (TT1 in Figure 3.9) were selected. Both devices were available through the facility’s SCADA system with a sampling interval of 1 minute. The characteristics of these two model input data sources are displayed in the table shown in Figure 3.21 [b].



	Unit	Average	Standard deviation
LT3 (Level change during PW addition)	%	29.1 min. 27.0 max. 30.5	1.2
TT1 (Temperature of PW loop)	°C	12.6 min. 9.7 max. 18.6	1.3

[b]

Figure 3.21: PMD development to estimate IC2 WPDs run time and generated PW: PMD development methodology [a], characteristics of PMD training data [b]

Low and high-level setpoints controlled the filling of the PW tank. Within the model of the PMD, a data filter was applied to detect the start time and duration of this filling process based on the available data from LT3. The WPDs (except the polishing-ultrafiltration unit) ran with a time delay to the filling process due to an additional buffer tank, equation (3.4).

$$\Delta t_{WPDs\,run\,time} = \left(t_{LT3,full} + t_{end\_delay}\right) - \left(t_{LT3,empty} + t_{start\_delay}\right) \qquad (3.4)$$

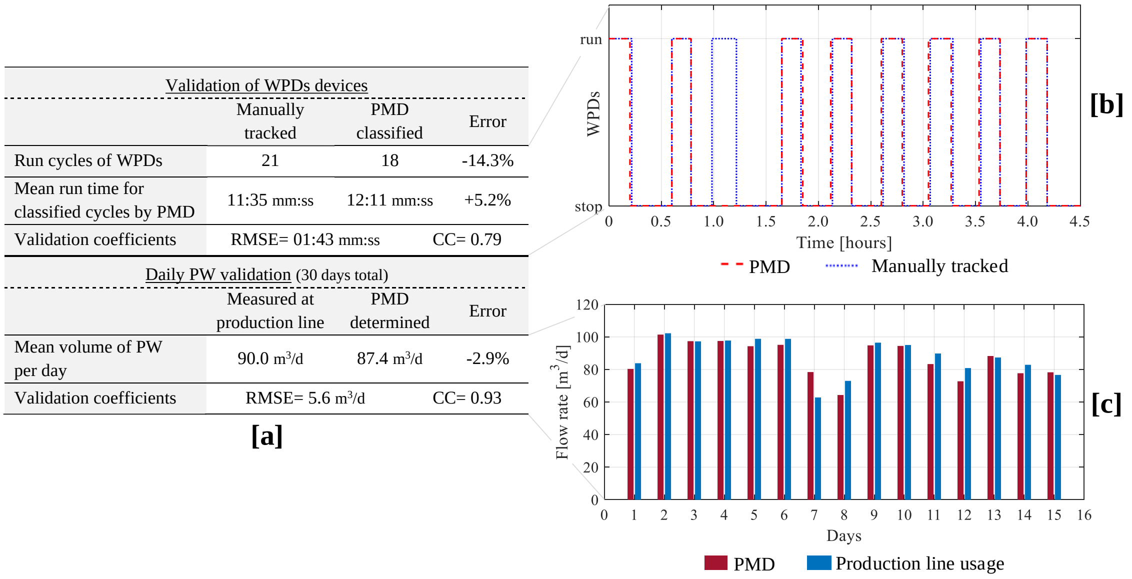
The flow rate to the PW tank was determined based on the devices run time, equation (3.5). The two ultrafiltration units (BUF and PUF in equation (3.5)) showed a consistent permeate flow rate while running. For the reverse osmosis (RO) unit in equation (3.5), the permeate flow rate depended on the water temperature. Since no online temperature transmitter was available within the PW generation system, the PW loop temperature TT1 was applied to estimate the reverse osmosis permeate flow.

$$Q_{perm.\ to\,PW\,tank} = \Delta t_{WPDs\,run\,time} * \left(Q_{RO\_perm.} + Q_{BUF\_perm.}\right) - Q_{PUF\_drain} \qquad (3.5)$$
$$with\ \ Q_{RO\_perm.} = f(TT1)$$

Results for the second PMD for the IC2 water system

After developing and tuning the model for the second PMD, the device was validated on a new set of data. To verify the estimated run time of the WPDs, a dataset of 21 run cycles was applied. The limited number of validation cycles resulted from the requirement of manually tracking the run time in the case facility. The model identified 18 run cycles correctly, as shown in the top part of the table in Figure 3.22 [a] and the graph of Figure 3.22 [b]. Some of the run cycles were not identified due to intermediate run modes that appeared when no PW was consumed at the production line for a period longer than 30 minutes. Overall, the PMD showed a RMSE of 1 minute 43 seconds and a CC of 0.79 for the 21 run cycles.

The PMD’s calculated flow rate between the polishing ultrafiltration unit and the PW tank was validated by comparing the daily volume to the PW intake at the production line (FT9-13 in Figure 3.9). A dataset for 30 days was applied. Due to the limitation of available meters, PW that was manually taken out by operators at the production line was not considered in the validation measures. The validation results are depicted in the lower part of the table in Figure 3.22 [a] and the graph of Figure 3.22 [c]. The PMD demonstrated a high accuracy with a RMSE of 5.6 m^3/d and a CC of 0.93.



Validation of WPDs devices			
	Manually tracked	PMD classified	Error
Run cycles of WPDs	21	18	-14.3%
Mean run time for classified cycles by PMD	11:35 mm:ss	12:11 mm:ss	+5.2%
Validation coefficients	RMSE= 01:43 mm:ss		CC= 0.79
Daily PW validation (30 days total)			
	Measured at production line	PMD determined	Error
Mean volume of PW per day	90.0 m³/d	87.4 m³/d	-2.9%
Validation coefficients	RMSE= 5.6 m³/d		CC= 0.93

Figure 3.22: Validation results for second PMD within the IC2 water system: table of validation parameters [a], extract of run times inferred by PMD and manually tracked [b], extract of daily estimated PW flow rate by PMD and measured at the production line [c]

3.3.2 Application to the steam system of the case facility

Additionally to the two PMDs for the IC2 water system, one further PMD was developed for the estimation of the added mains water within the facility's steam system. The missing data of the added mains water was identified as being relevant for an automated ongoing total cost of steam calculation in the subsequent step.

Mains water was added to the condensate tank to makeup system losses of steam and returning condensate, Figure 3.23. The amount of added mains water was regulated by a local control valve linked to the condensate tank filling level (LC-1). To reduce the hardness of the added mains water, it was first passed through two water softener units, set up in parallel. Each unit consisted of two softener columns that were running in alternating modes, switching over once a specified volume of water had been softened. Within each water softener unit, an offline flow meter measured the volume of softened water (FI6 and FI7 in Figure 3.23).

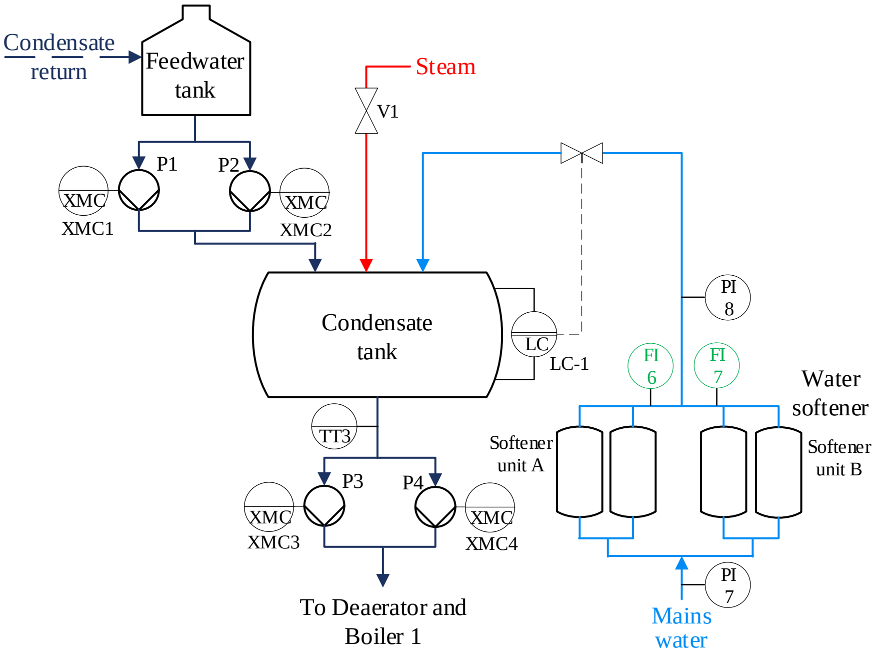
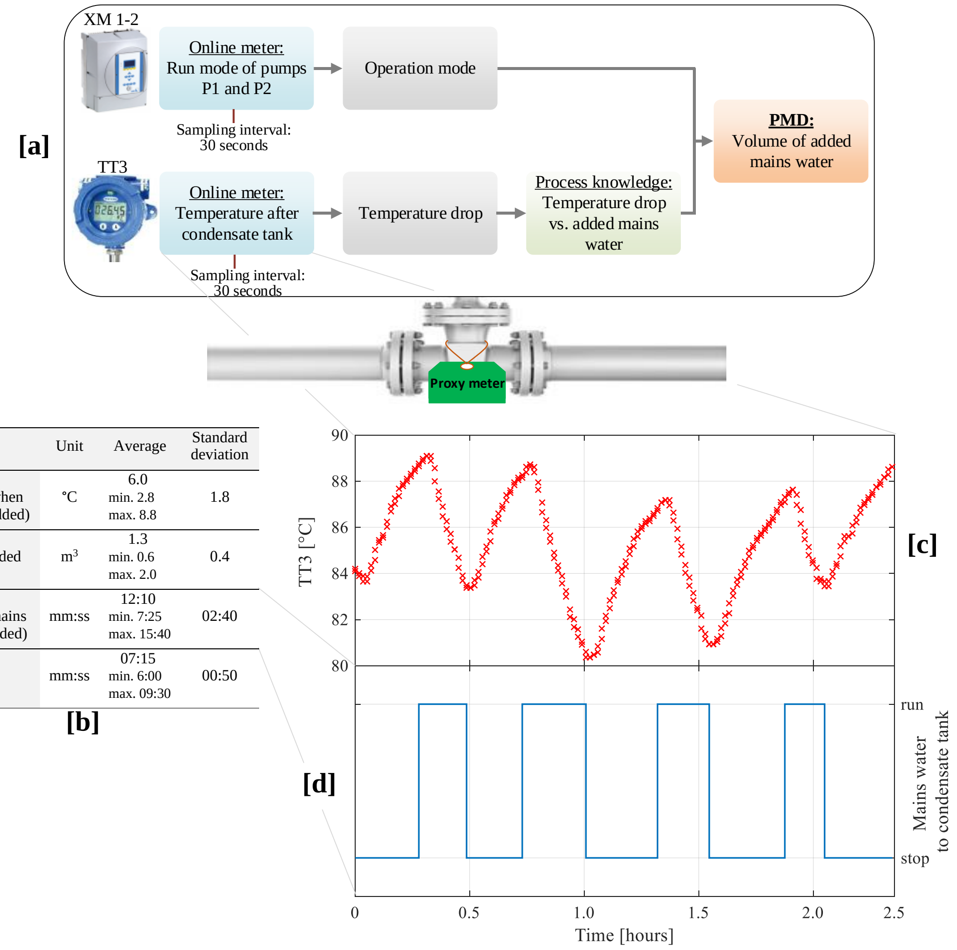


Figure 3.23: Process layout for condensate tank of case facility’s steam system including relevant meter devices and target variable of developed PMD in green

For the PMD development, the flow temperature after the condensate tank (TT3 in Figure 3.23) and the run modes of the feedwater transfer pumps P1 and P2 (XMC1 and XMC2 in Figure 3.23) were determined as suitable related online measuring devices for the estimation of the added mains water. The data for these devices was available through the facility’s SCADA system with a sampling interval of 30 seconds. Similar to the developed PMDs for the IC2 PW system, data gaps were addressed by the listwise deletion method and data outliers identified and deleted via the Hampel identifier. In addition to the available online data, locally conducted recordings of the offline available target variables (FI6 and FI7) were integrated into the model. On average, the adding cycle of mains water was activated every 31 minutes and 30 seconds with a cycle time of 12 minutes and 10 seconds and an added water volume of 1.30 m^3. The characteristics of the available input data for the model are displayed in the table shown in Figure 3.24 [b]. As a key relation, a linear model was developed that used the temperature drop measured after the condensate tank (TT3) that appeared once mains water was added, Figure 3.24 [c] and [d]. A data filter detected and quantified these temperature drops in the recorded data from TT3, Figure 3.24 [a]. The run mode signal of the feedwater transfer pumps (P1 and P2) provided an additional parameter to indicate if feedwater was added during an identified temperature drop.


[a]
XM 1-2
Online meter:
Run mode of pumps P1 and P2
Operation mode
Sampling interval: 30 seconds
PMD:
Volume of added mains water
TT3
Online meter:
Temperature after condensate tank
Temperature drop
Process knowledge:
Temperature drop vs. added mains water
Sampling interval: 30 seconds
Proxy meter
[b]
[c]
[d]
TT3 [°C]
90
88
86
84
82
80
run
stop
Mains water to condensate tank
0
0.5
1.0
1.5
2.0
2.5
Time [hours]

	Unit	Average	Standard deviation
TT3 (Temp. drop when mains water added)	°C	6.0 min. 2.8 max. 8.8	1.8
FI6 / FI7 (Volume of added mains water)	m^3	1.3 min. 0.6 max. 2.0	0.4
LCV1 (Duration of mains water being added)	mm:ss	12:10 min. 7:25 max. 15:40	02:40
P1 / P2 (Average run times)	mm:ss	07:15 min. 6:00 max. 09:30	00:50

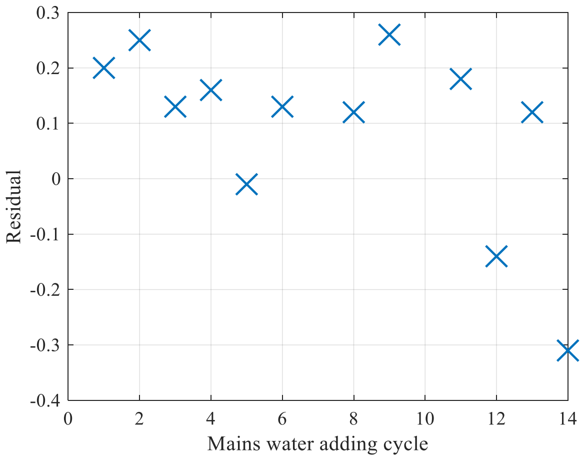
Figure 3.24: Steam system PMD development: PMD development methodology [a], characteristics of PMD training data [b], temperature after condensate tank (TT3) [c], mains water addition to condensate tank [d]

Results for the steam system PMD

For validation, the developed PMD for the estimation of the added mains water has been applied to a new set of data to analyse the performance. During testing, a dataset of 14 mains water adding cycles was analysed. The limited number of validation cycles resulted from the requirement of manually tracking the mains water adding cycles in the case facility. The model identified 12 mains water adding cycles correctly, as depicted in the table shown in Figure 3.25 [a]. The PMD's calculated amount of added mains water to the condensate tank was verified by comparing the volume to locally conducted recordings for the 14 cycles. On average, the inferred amount by the PMD was accurate within 15% of the added volume of mains water, Figure 3.25 [a]. For the validation cycles, the PMD showed a RMSE of 0.30 m^3 and a CC of 0.80. Furthermore, the performance of the PMD is depicted in Figure 3.25 [b] in the form of the resulting residuals for every cycle, i.e. estimated amount of added water by PMD minus locally recorded amount. Most of the residuals that are shown in the graph have a value greater than 0. This deviation is attributed to varying condensate tank temperatures that resulted from the amount of injected steam (V1 in Figure 3.23) and fluctuating condensate return flows. Figure D.8 in the appendix shows schematically the average condensate tank temperatures for a period of 14 days. Varying condensate tank temperatures influenced the performance of the PMD since the model is based on the temperature drop after the condensate tank that appeared once mains water was added.

	Manually tracked	PMD classified	Error
Mains water adding cycles	14	12	-14.3%
Mean volume of added mains water per cycle	1.02 m^3	1.17 m^3	+14.7%
Total volume of added mains water during validation	14.26 m^3	14.05 m^3	+5.4%
Validation coefficients	RMSE= 0.30 m^3		CC= 0.80

[a]



[b]

Figure 3.25: Validation results for the steam system PMD: table of validation parameters [a], residuals for the estimated mains water addition [b]

3.3.3 Discussion of simplified proxy metering application

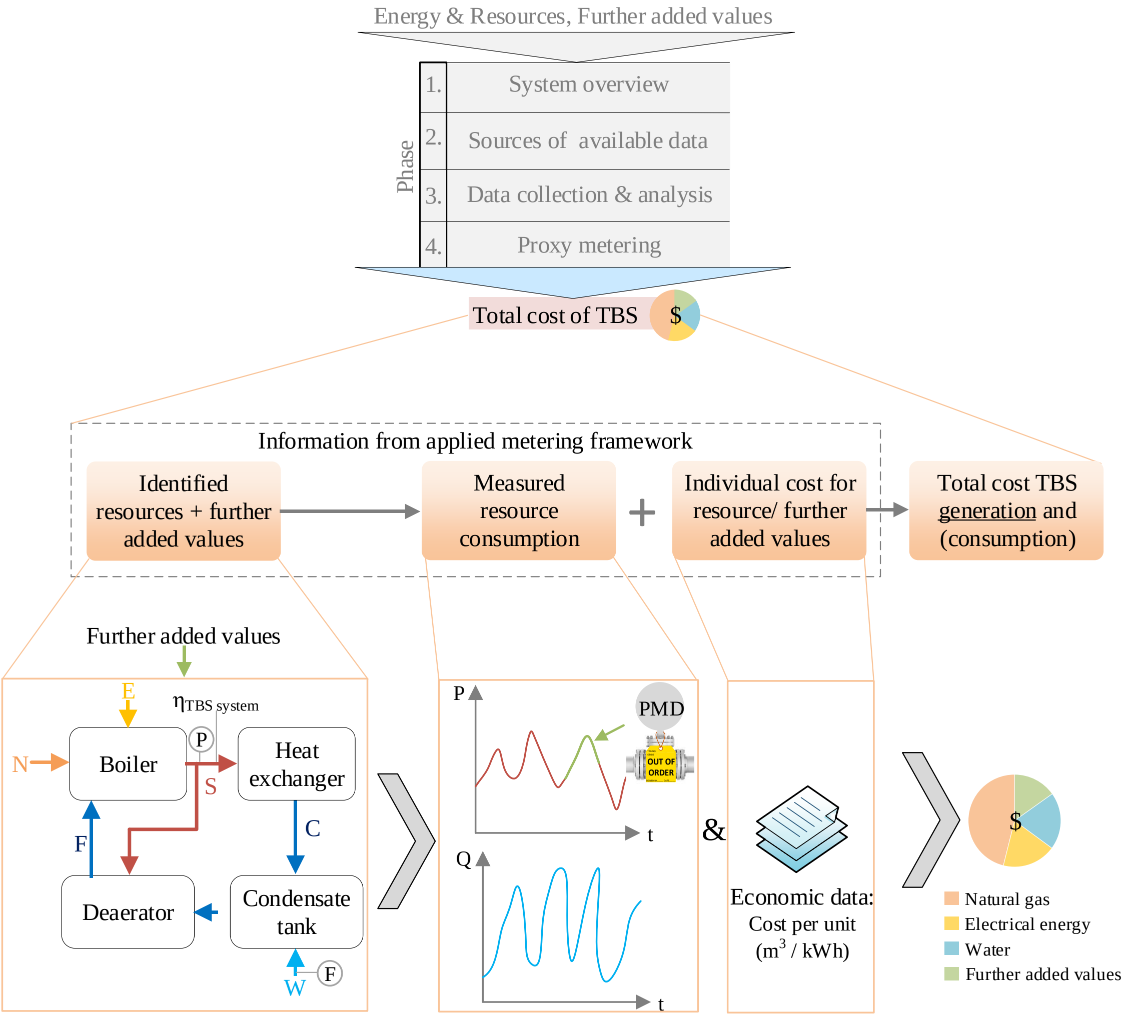
The application of the proxy meter development framework in the three case scenarios led to diverging results. Within the first application, the backwash mode of the water softener could not be tracked with sufficient accuracy during the PMD training phase. This case demonstrates the importance of suitable sample time for the identification of events in the PMD target as well as the influence of large time offsets between related measuring devices, due to the run time through the process. For similar cases, the inclusion of further measuring devices related to the target variable and a more complex model type could increase the performance. However, since no other related online measuring device was available in the analysed system, the PMD could not be further improved in reliability and a further physical meter would be needed to track the backwash mode.

The second PMD that was developed for the IC2 PW system and the PMD for the steam system demonstrated better performances with CC of 0.79/0.93 and 0.80 respectively during validation. These accuracies lack reliability for the monitoring of a critical parameter, however, they are acceptable for the quantification process of a metering audit. The models were developed in a short time due to the previously gained process knowledge and the first-order characteristics of the model. The required development effort of the model can increase for more complex systems where a first-order approach is not sufficient. The required training and validation datasets for the two PMDs had to be recorded manually during the metering campaigns. Therefore, the size of these datasets was small. In particular for more complex regression algorithms, the dataset limitation can result in insufficient model accuracies. Nonetheless, different training sample modifications are possible for the PMD development with small datasets which aim to increase the accuracy. Since this is not comprehensively addressed in the literature, as outlined in the literature review in section 2.3, further investigations of the impact of small datasets on different regression algorithms for PMD development and possible sample modifications are examined in chapter 4.

In some industrial applications of PMDs, no measurements of the target variable are present from the system under investigation. This absence leads to failure of PMD development and the need to implement a physical meter device if no target data can be acquired through an alternate method. An alternative way of collecting data from the target variable for PMD model training and validation can be the implementation of a temporary meter device, for example in the form of an ultrasonic flowmeter attached to the pipe exterior. Furthermore, an alternative PMD validation method can be applied to enhance the confidence of the model for cases of inadequate availability of target variable data. An example of an alternative validation method is to check for states of the target variable when the outcome is clear, for instance when a system is shut down. Likewise, the use of a temporary meter device and alternative validation methods are options for PMD development when insufficient performances result from small datasets.

3.4 Methodology to calculate the total cost of TBS

This section presents a calculation to investigate the total costs related to TBS in industrial facilities, as depicted in Figure 3.26. It is not directly part of the previously presented TBS metering methodology. However, it is based on the results of the metering audit and demonstrates the value of the gathered data. The total cost calculation of a TBS is a method to express the interconnection of different resources and further added values which can be used to monitor their impact. In the context of this work, the term total cost considers the variable costs of a TBS such as the required resources and further added values but does not include the fixed costs such as the equipment depreciations or labour costs.



Energy & Resources, Further added values
Phase
1. System overview
2. Sources of available data
3. Data collection & analysis
4. Proxy metering
Total cost of TBS
$
Information from applied metering framework
Identified resources + further added values
Measured resource consumption
+
Individual cost for resource/ further added values
Total cost TBS generation and (consumption)
Further added values
E
$\eta_{TBS\ system}$
N
Boiler
P
Heat exchanger
S
F
C
Deaerator
Condensate tank
W
F
P
PMD
OUT OF ORDER
t
Q
t
&
Economic data:
Cost per unit
(m^3 / kWh)
$
Natural gas
Electrical energy
Water
Further added values

Figure 3.26: Methodology for calculating the total cost of TBS

The cost of all required resources and further added values are combined for the calculation of the total TBS generation cost. The individual resource cost is provided from economic data (for instance utility bills) which were gathered during phase two of the metering methodology. This cost needs to be specified per unit to be attributed to the measured consumptions, i.e. for natural gas and water per cubic meter (c_{gas}; c_{water}) and for electrical energy per kilowatt-hour ($c_{elec.}$). This allows the multiplication of the measured resource consumptions with the equivalent cost for a certain time interval ($t_2 - t_1$), as presented in equation (3.6). Subsequently, the sum is divided by the total mass / volume of TBS that was generated during the same period. The following equation displays an example for the calculation of natural gas, water, and electricity as consumed resources of a TBS system:

$$c_{resources} = \frac{\int_{t_1}^{t_2} V_{gas} * c_{gas} + \int_{t_1}^{t_2} V_{water} * c_{water} + \int_{t_1}^{t_2} E_{elec.} * c_{elec.}}{\int_{t_1}^{t_2} m_{TBS} \; or \; V_{TBS}} \tag{3.6}$$

Additionally, an equivalent cost is allocated to the further added values that have been identified during the previous metering campaign. This cost is assigned for every mass / volume unit of generated TBS, as presented in equation (3.7). Hence, the cost of the single added value ($C_{added\,value,i}$) for a certain time frame ($t_2 - t_1$) is selected and divided by the total mass / volume of generated TBS in the same period.

$$c_{added\,value,i} = \frac{\left[C_{added\,value,i}\right]_{t_1}^{t_2}}{\int_{t_1}^{t_2} m_{TBS} \; or \; V_{TBS}} \tag{3.7}$$

Finally, the cost for the measured resources and the sum of all further added values (n) is added up, as seen in equation (3.8). This results in the total cost that is required for generating one mass / volume unit of TBS in the analysed system.

$$c_{TBS\_generation} = c_{resources} + \sum_{i=1}^{n} c_{added\,value,i} \tag{3.8}$$

For some TBS distribution systems, significant losses can be observed during TBS transmission to different consumer locations. One example are the steam systems, where it is common that part of the steam is lost due to steam leaks and condensation in the distribution pipes. This can result in higher costs of the TBS on the consumer side. If sufficient data has been collected during the metering campaign and the distribution losses appear to be significant, the consumption cost of the TBS distribution system can be calculated by taking the distribution efficiency into account, according to the following equation:

$$c_{TBS\_consumption} = \frac{c_{TBS\_generation}}{\eta_{TBS\_distribution}} \tag{3.9}$$

The total TBS cost calculation can be automated in case the data for resource consumption and TBS generation is based on online measuring devices and / or PMDs. When an automated calculation procedure is implemented, the TBS system has to be observed regularly to verify that the list of further added values and their corresponding costs, as well as the resource costs, are up to date. The advantages of an automated TBS cost calculation process are diverse. This cost accumulation can be used as an index for tracking system performances over a certain period and for identifying run modes that result in high costs. Also, the cost index can function as a measure of impact for system changes within the TBS generation and it can be used as a benchmark for comparisons to similar systems.

3.4.1 Application to one purified water system of the case facility

The methodology for the total cost calculation has been applied to the IC2 PW system of the industrial case facility. The TBS metering campaign from section 3.2 and the developed PMD from section 3.3 were applied as data sources.

As the main resources for the IC2 PW generation system, mains water and electrical energy were used. The facility’s water and electricity cost per cubic meter / kilowatt hour have been identified from former bills during the economic data gathering, as displayed in Table 3.6. The mains water intake for the system was measured by FT1 (Figure 3.9). For the determination of the WPDs run time, the developed PMD from section 3.3.1 has been applied due to a lack of online measuring devices. Also, this PMD allowed to estimate the amount of PW that was generated. To calculate the electrical energy consumption in the PW generation of IC2, the estimated run time of the WPDs was used in addition to the measured energy consumption during the water metering campaign. By applying the PMD to detect how long the units that ran occasionally (softener, reverse-osmosis, blending-ultrafiltration, and recovery reverse-osmosis) are active per hour, the electrical energy consumption was determined more accurately than by using an overall average value.

On the further added values side, system maintenance, chemical cleaning, and monitoring of PW quality have been identified as significant cost parameters for the operation of the system during the metering campaign. The cost for each expense is displayed in Table 3.6.

Table 3.6: Resource and further added value costs for the IC2 PW system within the case facility

Resource	Cost
Mains water ($c_{m.water}$)	2.13 €/m^3
Electrical energy ($c_{elec.}$)	0.10 €/kWh
Further added values	Cost per year
Maintenance of softeners	2,667 €/year
Maintenance of membrane units	24,067 €/year
Sanitisation + Chemicals	10,000 €/year
Monitoring of PW quality	49,494 €/year
Total cost further added values ($C_{added\ values,year}$)	86,228 €/year

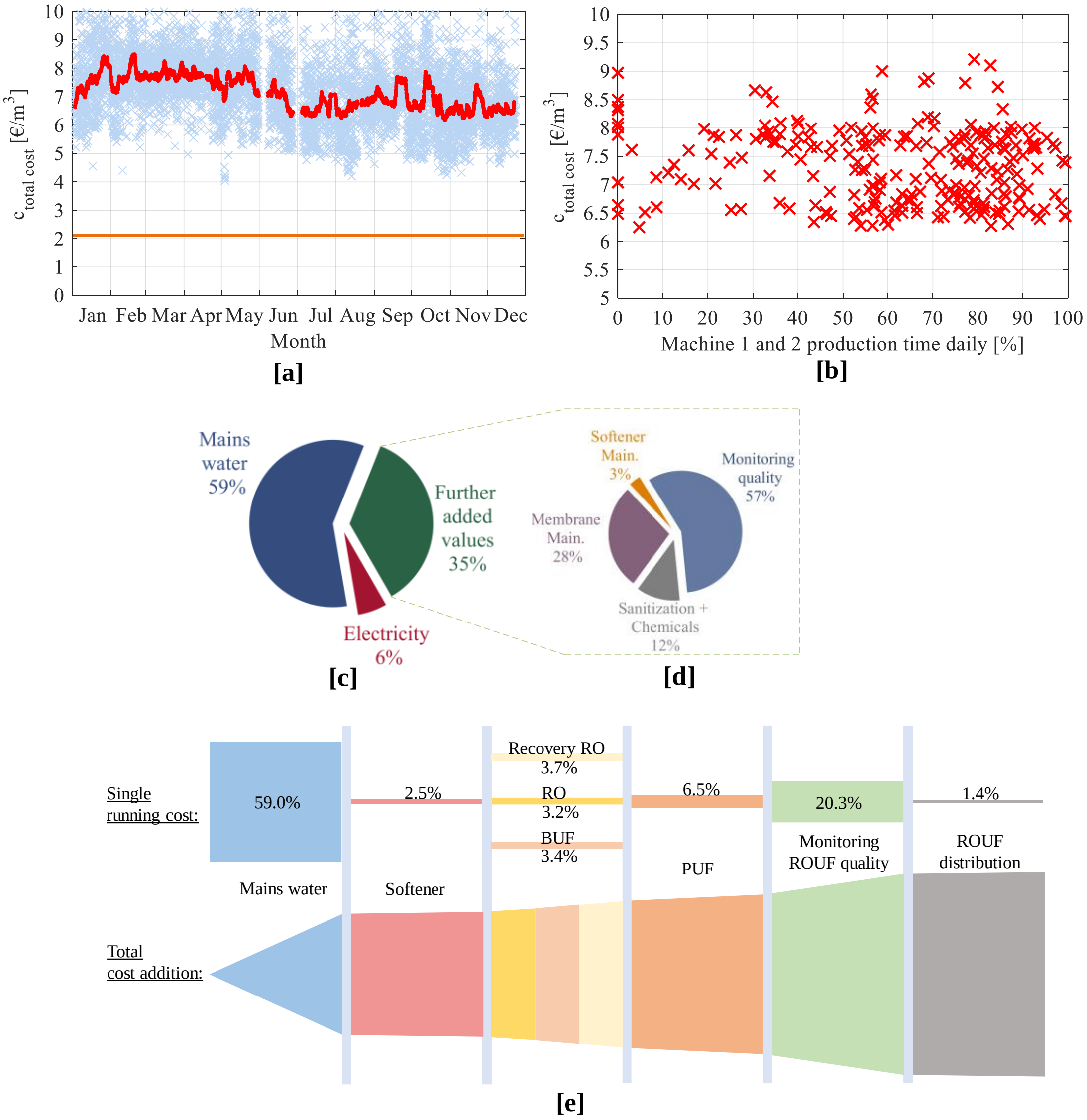
With the gathered data, the total cost of PW was calculated according to the following equation:

$$c_{PW\ total\ cost} = \frac{\int_{t_1}^{t_2} F\dot{T}1 * c_{m.water} + \int_{t_1}^{t_2} \left(\bar{E}_{S\_units}\right) * c_{elec.} + \frac{C_{added\ values,year}}{Q_{\ PW,year}}}{\int_{t_1}^{t_2} Q_{PW}} \qquad (3.10)$$

where $(t_2 - t_1)$ represents a selected time interval, $\bar{E}_{S\_units}$ is the locally recorded electricity consumption of the purification system depending on the states of the single units. The PMD from section 3.3.1 infers the unit's states as well as the yearly amount of generated PW $Q_{PW,year}$. The costs for mains water ($c_{m.water}$), electricity ($c_{elec.}$) and further added values per year ($C_{added\ values,year}$) are displayed in Table 3.6.

Results for the total cost of PW

The total cost of PW within the IC2 system has been analysed for one year. Figure 3.27 [a] depicts the resulting cost, calculated per hour. On average, the cost for generating one cubic meter of PW was 7.14 €/m^3, which is 3.4 times higher than the company’s mains water supply cost of 2.13 €/m^3.



[a]

[b]

[c]

[d]

[e]

Figure 3.27: Results of the total cost analysis for the IC2 PW system: total cost of PW water hourly including moving average filter (red) and mains water supply cost (orange) [a], daily PW cost depending on machine 1 and 2 productivity [b], PW cost allocation [c], cost allocation of further added values [d], cost allocation for individual PW generation steps [e]

A significant difference in Figure 3.27 [a] is the total PW cost between the first half (January until June) and the second half (July to December) of the chosen year. On average, the cost for the first half was 7.39 €/m^3 whereas the second half showed an average cost of 6.77 €/m^3. This difference is attributed to a membrane replacement at the first reverse osmosis unit in June. The new membranes increased the efficiency of this unit and therefore reduced the amount of required mains water. In addition, the temperature of the water influenced the permeate flow of the reverse osmosis unit.

In Figure 3.27 [b] the daily cost of PW generation of machine 1 and 2 is displayed in dependency on the daily production activity. From this graph, no trend is observable between PW cost and low or high production rates.

When considering the entire year, more than half of the total costs resulted from the mains water intake (4.22 €/m^3), followed by the costs for further added values (2.53 €/m^3), and electricity (0.41 €/m^3). Figure 3.27 [c] displays this cost distribution. The major share of the further added value cost (57%) emanated from the monitoring of the PW quality, Figure 3.27 [d]. Daily conducted laboratory analyses were performed to ensure the generated PW met predefined standards. Figure 3.27 [e] illustrates the cost of the different process steps involved in the generation of PW.

As an additional result representation, Table 3.7 shows the average total cost for generating 1 cubic meter of PW in the IC2 water system every month and the weighted average for the entire year (days per month considered). The listed value-added factor is a parameter that indicates the additional cost in relation to the mains water supply cost and is calculated according to the following equation:

$$Value\ added\ factor = \frac{c_{\ total\ cost\ PW}}{c_{m.water}} \tag{3.11}$$

Table 3.7: Total cost of PW and value-added factor per month for IC2 water system

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	Weighted average
Average total cost PW[€/m^3]	7.62	7.71	7.79	7.71	7.54	6.98	6.60	6.67	7.07	6.78	6.65	6.56	7.14
Value-added factor	3.6	3.6	3.7	3.6	3.5	3.3	3.1	3.1	3.3	3.2	3.1	3.1	3.4

3.4.2 Application to the steam system of the case facility

The methodology of the total cost approach for TBS has been applied additionally to the steam system of the case facility. Based on the collected data from the steam metering campaign and the developed PMD for the estimation of the added mains water, the total steam generation cost of boiler 1 was analysed. Boiler 2 and 3 were not taken into account due to their low share of the total amount of generated steam. The previously conducted data gathering identified the required economic data, in the form of natural gas, water, electricity, and further added value costs, as presented in Table 3.8.

Table 3.8: Resource and further added value costs for case facility's steam system

Resource	Cost
Natural gas (c_{gas})	0.31 €/m^3
Mains water $(c_{m.water})$	2.13 €/m^3
Electrical energy $(c_{elec.})$	0.10 €/kWh
Further added values	Cost per month
Water softener salt	144 €/month
Monitoring of boiler liquid	373 €/month
Total cost further added values $(C_{added\ values,month})$	517 €/month

With the gathered data, the total cost of steam generation for boiler 1 was calculated according to the following equation:

$$c_{steam\ gen.\ total\ cost} =$$

$$= \frac{\int_{t_1}^{t_2} V_{gas} * c_{gas} + \int_{t_1}^{t_2} PMD_{m.water} * c_{m.water} + \int_{t_1}^{t_2} \left(\bar{E}_{S\_units}\right) * c_{elec.} + \frac{C_{added\ values,month}}{m_{steam\ gen.,month}}}{\int_{t_1}^{t_2} \dot{m}_{steam\ gen.}} \qquad (3.12)$$

where $(t_2 - t_1)$ is a chosen time interval, $PMD_{m.water}$ is the inferred volume of added mains water by the PMD, and $\bar{E}_{S\_units}$ is the electrical energy required by the single consumers in the steam generation system. This electricity consumption was estimated based on the operation states of the single units which were available through the facility's SCADA system.

Results for the total cost of steam

The total cost of steam generation within the case facility was analysed for 35 days. Due to a limited availability of data from the SCADA system, the analysing period of the total cost of steam had to be limited to this time frame. The specific cost per tonne of generated steam was calculated on an hourly basis. For the available time, the total steam generation cost was 26.48 €/t on average, Figure 3.28 [a]. The main driver in this cost accumulation was natural gas at 24.28 €/t, followed by electrical energy at 1.07 €/t, and mains water at 0.98 €/t. Further added values showed the lowest contribution to the total cost at 0.15 €/t. The cost distribution for the single resources and further added values as well as the single process steps are depicted in Figure 3.28 [c] and [d].

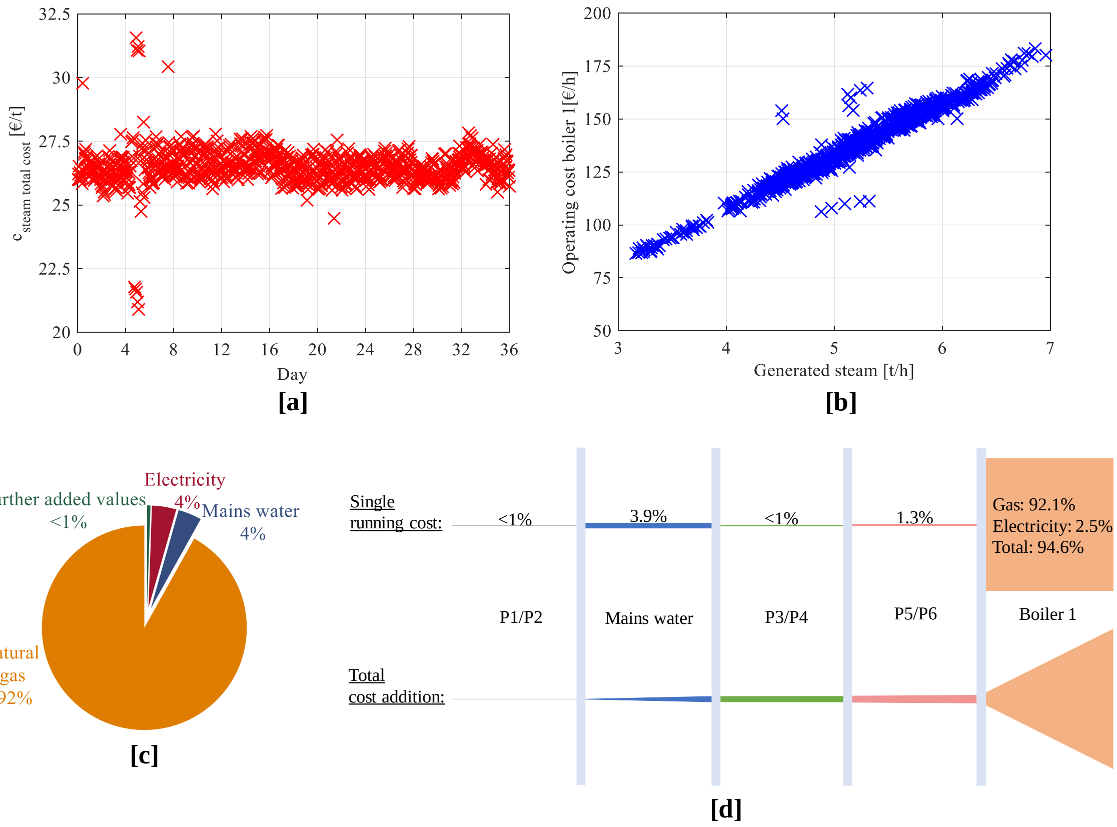


Figure 3.28: Results of the total cost of steam analysis in the case facility: total cost of steam generation hourly [a], operating cost for boiler 1 [b], cost allocation for the generation of steam [c], cost allocation for individual steam generation steps [d]

Figure 3.28 [b] reveals the relationship between the amount of generated steam per hour and the resulting operation cost for boiler 1. This graph illustrates that the operation costs were increasing linearly with higher rates of steam generation within the interval between 3 and 7 t/h of generated steam.

By taking into account the previously analysed steam distribution efficiency (92%; Figure 3.16 [a]), the steam consumption cost was calculated according to equation (3.9). This cost was on average 28.78 €/t for the analysed steam system.

3.4.3 Discussion of the total cost of TBS methodology

The results from the two applications of the total cost methodology for TBS in the case facility revealed how each individual driver affected the total cost of PW / steam generation. The presented cost accumulation underlines the real value that is embedded in TBS by focusing on the interaction of multiple resources. In particular for the PW system, the results showed that the added value was on average 3.4 times higher than the mains water supply cost. Walsh et al. [100] showed that a value-added factor for PW can be up to 14 times higher than the mains water supply cost when more PW generation steps are involved and the mains water is supplied with lower costs.

For the analysed steam system, natural gas was by far the main cost driver, followed by electrical energy. These cost allocations and their magnitudes are in accordance with values for steam generation systems given by Masanet and Walker [83] and Hasanbeigi et al. [80].

Compared to similar total cost models that can be found in the literature [100], [103], [105], the developed methodology in this research has the novelty that it is based on online / PMD meters. This allows an automated benchmark with the possibility of small step analyses and live tracking. The automation enables the tracking of system changes and displays their impact on the total cost, as seen in the PW system in the form of membrane replacements in the first reverse osmosis unit (Figure 3.27 [a]).

The gained data and insights are used by the facility to justify improvements of their systems by providing a shorter payback time based on the now available total cost. Moreover, the revealed total cost enhances the awareness of the value within the generated PW / steam system and led to an increased effort to realise saving opportunities. When striving for system savings, the total cost benchmark offers the potential to function as a KPI and can be part of a companywide energy and resource management system.

The accuracy of the presented total cost methodology can be limited, mainly when many further added values are involved or when each value has to be estimated due to limited data availability. Nonetheless, the purpose of this methodology is to raise awareness of the real value that is involved in TBS as well as to express the interconnection of different resources rather than the calculation of an accurate total cost.

3.5 Summary

This chapter focused on the development and application of a holistic metering strategy for detailed information gathering of TBS. A fundamental requirement for this task was to take existing conditions and constraints of industrial environments into account [94]. To ensure applicability for industrial users, the assessment framework has been developed and applied with the study of two PW and one steam system in a case facility.

To start a TBS investigation, the first two phases of the presented metering framework focus on a system familiarisation, identification of available data sources, and abstraction of the main process steps. The third phase involves a targeted collection and analysis of process data towards a previously set metering goal. By mapping the required resources and studying the performance of single devices of the three TBS systems in the case facility, operators gained an increased system understanding including energy and resource-saving opportunities. Although the high number of offline and missing measuring devices in these case systems led to increased metering efforts and limited the study of system dynamics.

The solution to the absence of functional online meters is represented in the fourth phase of the metering framework by the integration of a simplified proxy metering strategy. Data from related online measuring devices are first pre-processed and subsequently applied in a regression model. The higher inaccuracies of these models compared to a physical meter device are acceptable due to the benefits in the form of cost and time savings. Within the case applications of the PW and steam systems, three PMDs have been developed to infer missing process parameters. During the model training stage for one of the PMDs, inadequate sampling times and diverse positions of related measuring devices resulted in poor accuracy. The other two PMDs were developed with sufficient accuracy and improved the system understanding within the metering campaign. A constraint for all three PMD approaches were the small modelling datasets which resulted from the necessity of manual data acquisition.

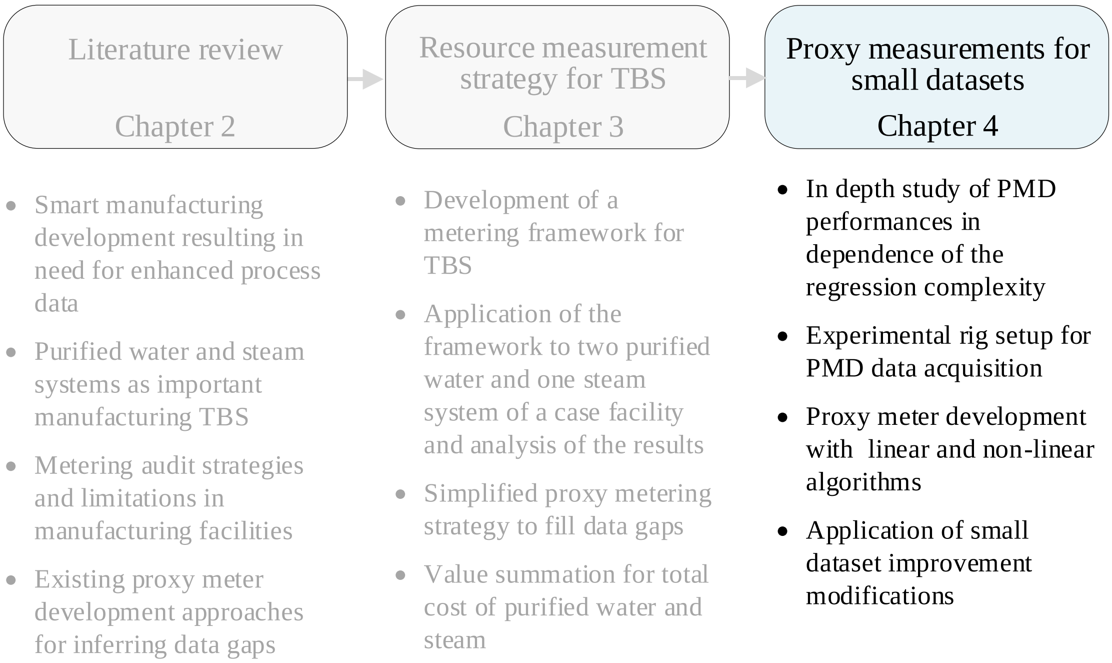
The information emanating from the metering framework can then be utilised to calculate the total cost of TBS. This cost index is based on the required resources and further added values. The total cost demonstrates the interaction of multiple resources and visualises the value that is embedded in the investigated system. In comparison to similar approaches that can be found in the literature [100], [103], [105], the presented cost calculation methodology in this research relies on online measuring devices and PMDs and can therefore be automated. This novelty enables the tracking of the effectiveness of system modifications, as seen in the case of the investigated PW system where a membrane change in one of the units reduced the PW generation cost by approximately 9%.

Overall, the developed metering framework introduces a data acquisition and characterisation strategy for TBS that has not been reported to this extent in the literature to date. This research accelerates the move towards sustainable TBS by unfolding energy and resource-saving opportunities and showing their magnitudes in the form of the total TBS cost.

To further improve data visibility when functional online measuring devices are absent, more applicable guidelines from the research are needed for PMDs that have to be developed on small datasets. An in-depth study of PMD behaviour depending on the complexity of small datasets and concluding modelling methodologies are analysed in chapter 4.

Chapter 4

Proxy measurements in small dataset scenarios


Literature review
Chapter 2
Resource measurement strategy for TBS
Chapter 3
Proxy measurements for small datasets
Chapter 4

- Smart manufacturing development resulting in need for enhanced process data
- Purified water and steam systems as important manufacturing TBS
- Metering audit strategies and limitations in manufacturing facilities
- Existing proxy meter development approaches for inferring data gaps

- Development of a metering framework for TBS
- Application of the framework to two purified water and one steam system of a case facility and analysis of the results
- Simplified proxy metering strategy to fill data gaps
- Value summation for total cost of purified water and steam

- In depth study of PMD performances in dependence of the regression complexity
- Experimental rig setup for PMD data acquisition
- Proxy meter development with linear and non-linear algorithms
- Application of small dataset improvement modifications

The research in this chapter presents an analysis of three different proxy meter device (PMD) algorithms that were applied on five small datasets. These datasets have been acquired from an experimental rig. As an initial step of the testing methodology, a set of complexity measures was applied to classify the regression complexity of each dataset. Subsequently, the datasets were used in PMD regression models and the estimation was studied. For further improvement, the influences of pre-modelling modifications in the form of bootstrap replications (BR) and artificial noise injection (ANI) was investigated. In total, the presented analysis provides a comprehensive PMD development strategy for industrial users that is applicable to small datasets and suitable to individual process complexity. This straightforward approach presents necessary guidance to enhance the use of PMDs in industrial facilities [140].

4.1 Experimental rig

To investigate the performance of PMDs, an experimental rig was developed that enabled the collection of small datasets with varying complexity. The experimental rig process layout, depicted in Figure 4.1, was designed by following the arrangement of the condensate tank setup of the previously analysed case facility’s steam system in section 3.3.2 (Figure 3.23).

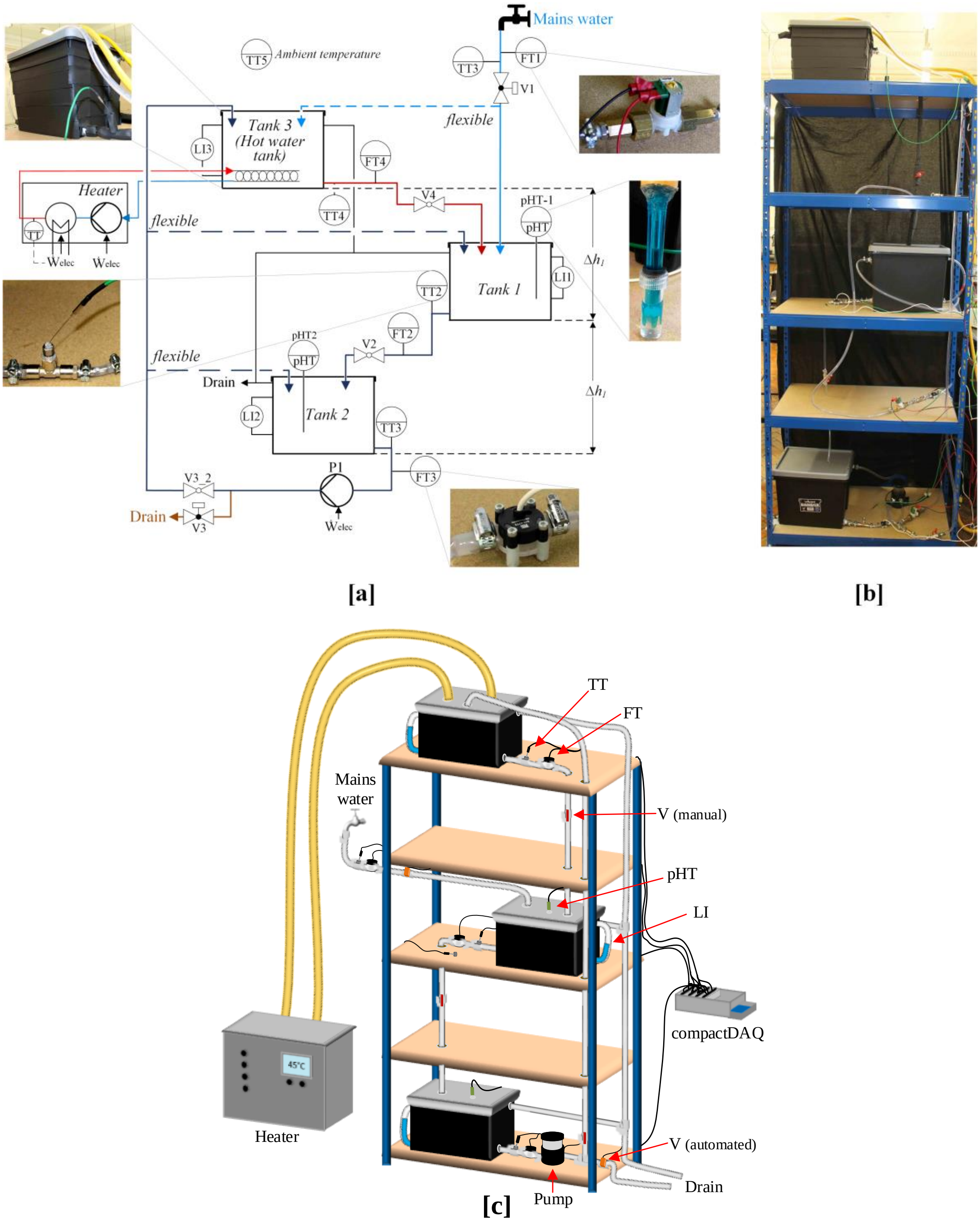


Figure 4.1: Experimental rig layout for PMD development: process layout [a], photo of the entire experimental rig [b], 3D model of experimental rig (layout for setup A) [c]

A high degree of flexibility to enable different process setups was important for the experimental rig design. Water was circulated between three tanks within the experimental rig, which were located with a height difference (Δh_1 in Figure 4.1 [a]) of 105 cm between each other to reduce the number of required pumps. Flexible tubes with an internal diameter of 1 cm were used for the tank connections. The three tanks were similar in size, with an internal volume of 24 litre and an internal cross-sectional area of 28.5 x 23 cm. Mains water could be added to tank 1 or 3 and excessive water was drained at various locations. In tank 3, a heating coil was implemented to raise the water temperature. This coil was connected to an external heating device by an independent water loop. The water flows within the experimental rig were set by manual ball valves (V2; V3_2; V4) and two automated solenoid valves (V1 and V3). Throughout the process, different measuring devices were installed that recorded temperatures (TT; type K-thermocouples), flow rates (FT), and pH value (pHT) with a sample rate of 1 Hz. These measurement devices, as well as the solenoid valves V1 and V3, were connected to a compactDAQ (cDAQ-9178) from National Instruments, a data acquisition and control platform. The compactDAQ was operated by the National Instrument programming language LabVIEW 2018. As additional measuring devices, three analogue bypass level indicators (LI) showed the filling level of tanks 1, 2, and 3. These devices indicated the filling levels of each tank in proportional to the tank height. The external heating device which was connected to the heating coil of tank 3 had its own control system with the option to set constant water temperatures at the heater output. Tank 1 represents the link of the experimental rig design to the earlier mentioned condensate tank setup of the case facility’s steam system. In this tank, water flows can be added with different temperatures and flow rates. The adjusting temperature within the tank is measured at the tank exit. This setup is similar to the functionality of the condensate tank within the analysed steam system, shown in Figure 3.23.

Calibration of measuring devices

The implemented measuring devices were calibrated according to their operating conditions in the experimental rig. A detailed description of the calibration and validation procedure for the single measuring devices is shown in the appendix section B. The following table displays the outcome of the calibration process in the form of uncertainty parameters which were obtained during the subsequent device validation.

Table 4.1: Calibration results for the measuring devices of the experimental rig

Measuring device	Calibration range	Measuring principle	Uncertainty parameters	
			Max. deviation	Mean deviation
TT1	18-78°C	K-type thermocouple	0.13 °C	0.06 °C
TT2			0.19 °C	0.13 °C
TT3			0.15 °C	0.07 °C
TT4			0.19 °C	0.15 °C
TT5			0.09 °C	0.05 °C
FT1	0.5-2.0 l/min	Hall effect flow sensor	0.03 l/min	0.01 l/min
FT2			0.04 l/min	0.02 l/min
FT3			0.03 l/min	0.01 l/min
FT4			0.04 l/min	0.02 l/min
pHT1	pH 2.5-10.5	Glass and reference electrode	pH 0.09	pH 0.04
pHT2			pH 0.11	pH 0.07

Limitations of the experimental rig

Although the experimental rig has been designed according to the condensate tank setup of the previously analysed case facility's steam system in section 3.3.2, the small scale of the rig can impact the PMD performance study. The small dimensions of the pipes and tanks of the experimental rig and the abstracted design led to divergent flow behaviours and measurement phenomena compared to larger scale real world industrial processes. This limits the transferability of the PMD development methodology which is outlined in the following sections. Nonetheless, the advantage of the experimental rig compared to a real world industrial system, is a higher level of process control and flexibility: Various measuring cycles that are outlined in the following section could be repeated several times in a short time period which would not have been possible in most real world industrial processes due to ongoing manufacturing activities. In addition, the implemented measuring devices, with a sample rate of 1 Hz, allowed a detailed PMD performance study. In many industrial setups, a similar study extent would not have been possible due to missing online measuring devices and / or low sample rates, as seen for the developed PMDs of the case facility in chapter 3.3.

In summary, the experimental rig enables a comprehensive PMD performance study that would not have been possible in the same time period within an industrial setup. However, the developed PMD methodology needs to be applied to a real world industrial process to prove its validity for larger scale systems.

4.2 Proxy meter developments

Based on the built experimental rig, five PMDs were developed to approximate one of the measured process parameters. In a real industrial system, this PMD arrangement would represent a back-up for physical meters and can be used to identify abnormal deviations between physical and virtual meters, similar to the PMD use case presented by Äijälä and Lumley [191] in section 2.3.2. For the developed experimental rig, the physical meters allowed an in-depth performance comparison between PMD models and actual measured parameters by providing the required data. This data enabled the training and validation of different PMD regression models and allowed further study of the influence of small dataset manipulation techniques.

4.2.1 PMD experimental rig setups

Depending on the flow arrangement of the experimental rig, different target and input variable relationships were selected for each of the five PMDs, labelled by the letters A to E. Table 4.2 depicts an overview of the target and input variables for each PMD setup. In addition, the following list presents a summary of each PMD setup and Figure 4.2 presents the respective flow path within the experimental rig.

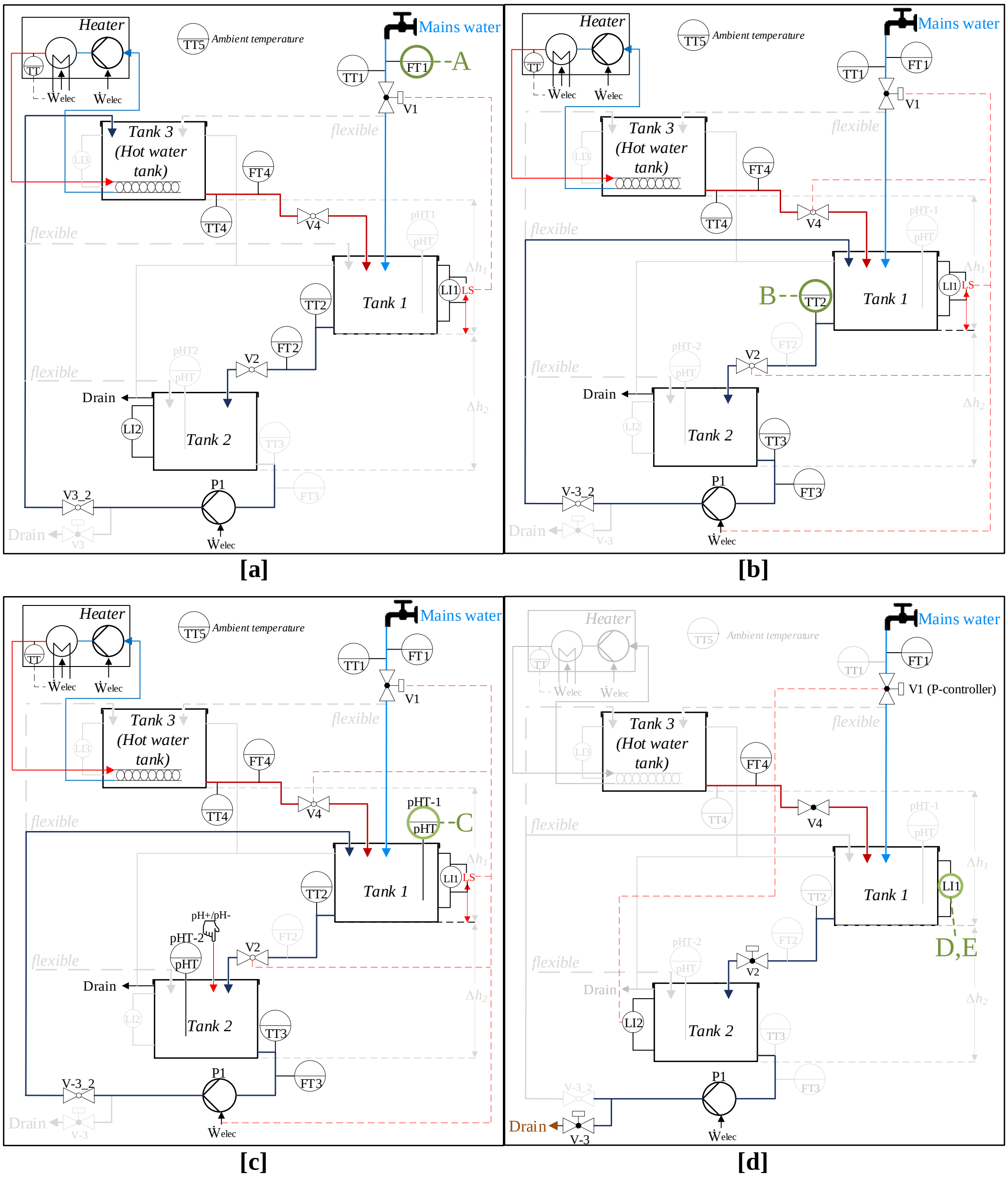
- Setup A: Water was looped between tanks 1, 2, and 3. The external heater was active to increase the water temperature within tank 3. Mains water was added occasionally to tank 1, start and stop of the mains water addition was controlled by the filling level of tank 1. The flow rates of the incoming and exiting flows to tank 1 and the external heater temperature were varied in between the mains water addition cycles. The amount of added mains water (FT1) during one cycle was selected as the target variable for the PMD in setup A. This experimental rig setup is similar to the conditions of the developed PMD within the case facility’s steam system, shown in section 3.3.2. Figure 4.2 [a] displays the flow path within the experimental rig for setup A.
- Setup B: The empty tank 1 was filled with water from tanks 2, 3, and the mains. The adding flows showed diverse temperatures and flow rates and were varied in between the filling cycles. The temperature of tank 1 (TT2) that resulted after each filling cycle was chosen as the target variable for the PMD in setup B. Figure 4.2 [b] displays the flow path within the experimental rig for setup B.

- Setup C: Similar to setup B, water flows with different flow rates and temperatures were mixed in tank 1. The pH-Value (pHT1) of tank 1 that resulted after mixing of these flows has been selected as the target variable for PMD setup C. According to Barron et al. [208], pH-values decrease with an increase in the liquid temperature. To reinforce the change of pH-values in tank 1, sodium carbonate (increase of pH value) or sodium bisulfate (decrease of pH value) was added to tank 2 and subsequently pumped to tank 1 where it was mixed with mains water and water from tank 3. Figure 4.2 [c] displays the flow path within the experimental rig for setup C.
- Setup D: The filling level of tank 1 (LI1) after a running time of 10 minutes (t_predict in Table 4.2) was selected as the target variable of the PMD in setup D. In this setup, water was added to tank 1 from tank 3 and the mains. To increase the model complexity, the mains water flow (V1) was controlled by using a 2-position controller with a set target filling level for tank 2. In addition, the water flow rate from tank 3 (FT4) was changed once at a random time point during the data recording (t_change_FT4 in Table 4.2). The filling levels of tank 1 and tank 2 were randomly selected at the start. In contrast to the previous setups, the dataset for this process scenario has been simulated based on physical relations. A more detailed explanation of this simulation is presented below and in the appendix section C. Figure 4.2 [d] displays the flow path within the experimental rig for setup D.
- Setup E: Setup E is similar to setup D, except that the water flow from tank 3 (FT4) was continuously decreasing from a randomly chosen filling height at the beginning. As in setup D, the filling level of tank 1 (LI1) after a running time of 10 minutes (t_predict in Table 4.2) was selected as the target variable of PMD setup E. The dataset has been simulated based on physical relations, a more detailed explanation of the simulation is presented below and in the appendix in section C. Figure 4.2 [d] displays the flow path within the experimental rig for setup E.

Table 4.2: Input and target variables for PMDs developed from the experimental rig

		PMD-setup				
		A	B	C	D	E
Target variable		FT1 (added mains water to tank 1)	TT2 (temp. in tank 1 after mixing flows)	pHT1 (pH value in tank 1 after mixing flows)	LI1 (filling level in tank 1 at time t_predict)	LI1 (filling level in tank 1 at time t_predict)
	median	4.3 litres	33.2 °C	pH 7.8	0.17 m	0.22 m
	Std.	1.1 litres	3.9 °C	pH 0.7	0.04 m	0.04 m
	min.	2.8 litres	25.3 °C	pH 6.4	0.09 m	0.13 m
	max.	7.3 litres	44.5 °C	pH 9.6	0.26 m	0.29 m
Input variables		FT2	FT1	FT1	FT4_1*	LI1*
		FT4	FT3	FT3	FT4_2 (after t_change_FT4)	LI2*
		TT2 (dT dropping + t of dT dropping)	FT4	FT4	LI1*	LI3*
		TT1	TT1	TT1	LI2*	FT3
		TT4	TT2*	TT3	t_change_FT4	
		TT5	TT3	TT4		
			TT4	TT5		
			TT5	pHT2*		
				pHT1*		
Total # of input variables		7	8	9	5	4
Note					Based on simulation	Based on simulation

*: measured value at t_start



[a]

[b]

[c]

[d]

Figure 4.2: Process layout for experimental rig for PMDs cases A-E; including target variables (green) and relevant input variables: setup A [a], setup B [b], setup C [c], setup D and E [d]

Setups D and E

As mentioned above, the datasets for the PMD setups D and E have been acquired through a simulation model which represented parts of the experimental rig. This simulation aimed to acquire data with a more complex regression structure compared to the datasets of PMD setups A-C. To realise this complexity improvement, the filling level of tank 1 (LI1) has been identified as a suitable target variable. The simulation was chosen since the implemented level indicators within the experimental rig (LI1, LI2, and LI3) were offline devices and could not be connected to the compactDAQ data acquisition system. In the model, tank 1 is filled by inputs from tank 3 (FT4/LI3) and the mains water and the tank 1 output is filling tank 2. To increase the setup complexity, the filling level of tank 2 was used as a target parameter of a 2-position controller that regulated the amount of mains water addition with the solenoid valve V1. For setups D and E, a fixed tank 2 filling level (LI2) of 0.2 m was used as a target for the 2-position controller. Within setup D, the added water from tank 3 to tank 1 (FT4) was set to two constant flow rates which changed at a random time point (t_change_FT4) during the simulation time of 10 minutes. For setup E, the added water from tank 3 to tank 1 was continuously decreasing according to the filling level of tank 3 which was set to a random start level at the beginning of the simulation.

As a modelling environment, MATLAB Simulink has been used. All parameters of the simulation, such as the maximum and minimum tank filling levels and flowrates, matched the conditions of the real experimental rig. The simulation of setup D and E was running for a duration of 10 minutes and the filling level of tank 1 (LI1) at the end of this time was used as the target value for the PMD estimation. Further detail in the form of the model derivation and validation is presented in the appendix section C. Figure 4.3 depicts example simulation results of the tank filling levels and flow rates, in [a] for setup D and in [b] for setup E. The simulations were conducted 80 times to generate the required datasets for PMD model training and validation.

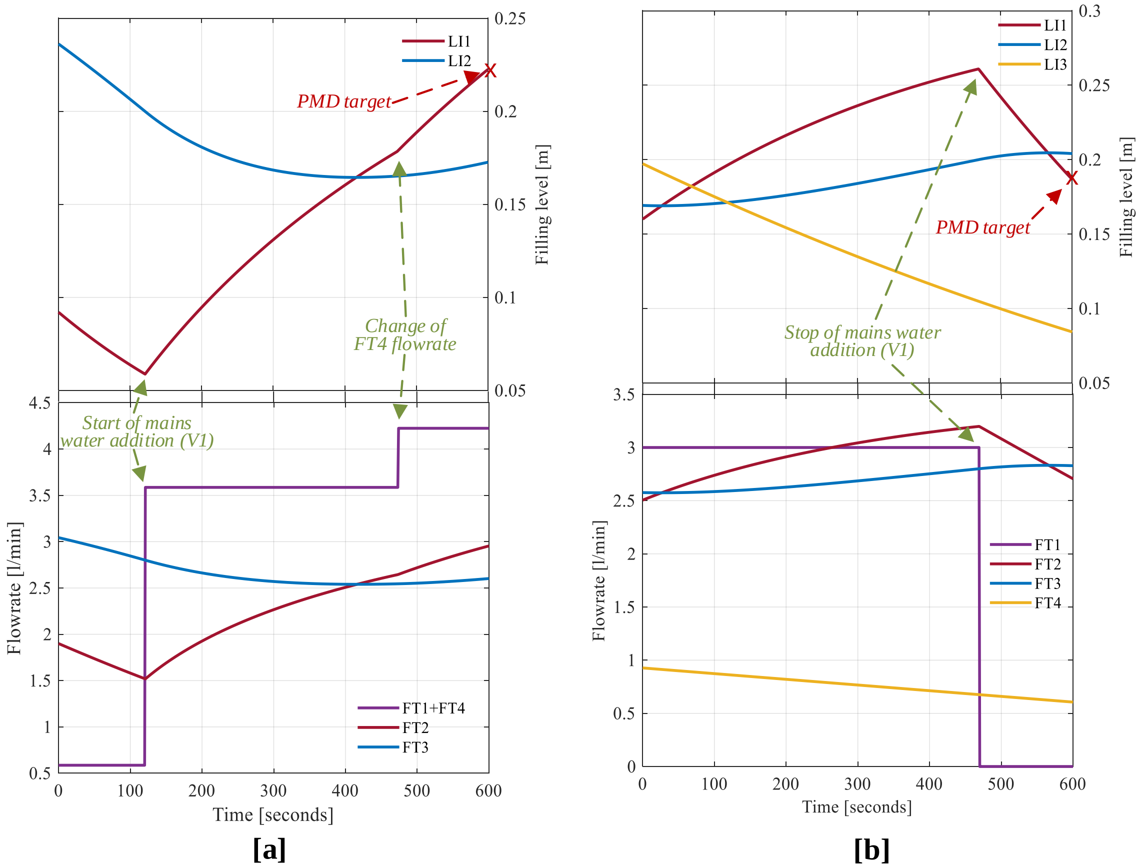


Figure 4.3: Example of simulated experimental rig filling levels and flow rates for setups D [a] and E [b]

4.2.2 Data pre-processing, model training and validation

For each of the five developed PMDs, three different regression algorithms were applied: Multiple linear regression (MLR), Partial Least Square regression (PLSR), and Neural Network (NN). According to Kadlec et al. [149], these algorithms are most frequently used for linear and non-linear PMD development (further details in section 2.3.1). The characteristics of MLR, PLSR, and NN were presented in Table 2.4.

The overall procedure for PMD development with experimental rig data is shown in Figure 4.4. The entire code for data pre-processing, model training and validation that is presented in the experimental results section, was developed with the MATLAB 2018 software environment. Initially, data from the input and target variables were collected from the experimental rig with the compactDAQ for setups A, B, and C and pre-processed according to the outlined PMD model developing steps in section 3.3. As only small data sections were missing, the listwise deletion was applied to remove samples containing data gaps. Data outliers were detected by the Hampel identifier. The entire data was gathered by the compactDAQ with a sampling time of 1 Hz for each sensor. Therefore, the diversity of sampling times between measured variables did not occur. Each set of data records for the setups A, B, and C included multiple cycles of the measured phenomena, e.g. mains water additions to tank 1 for setup A. Therefore, data filters were used after the data pre-processing to extract the relevant values of the input and target variables that are displayed in Table 4.2. One cycle of the measured phenomena represented one sample for the PMD development. Between 85 and 100 samples were recorded for setups A, B, and C. For setups D and E, the required data was provided by the simulation model. To have an equal number of samples for each setup, the dataset size was limited to 80 samples (referred as original dataset below).

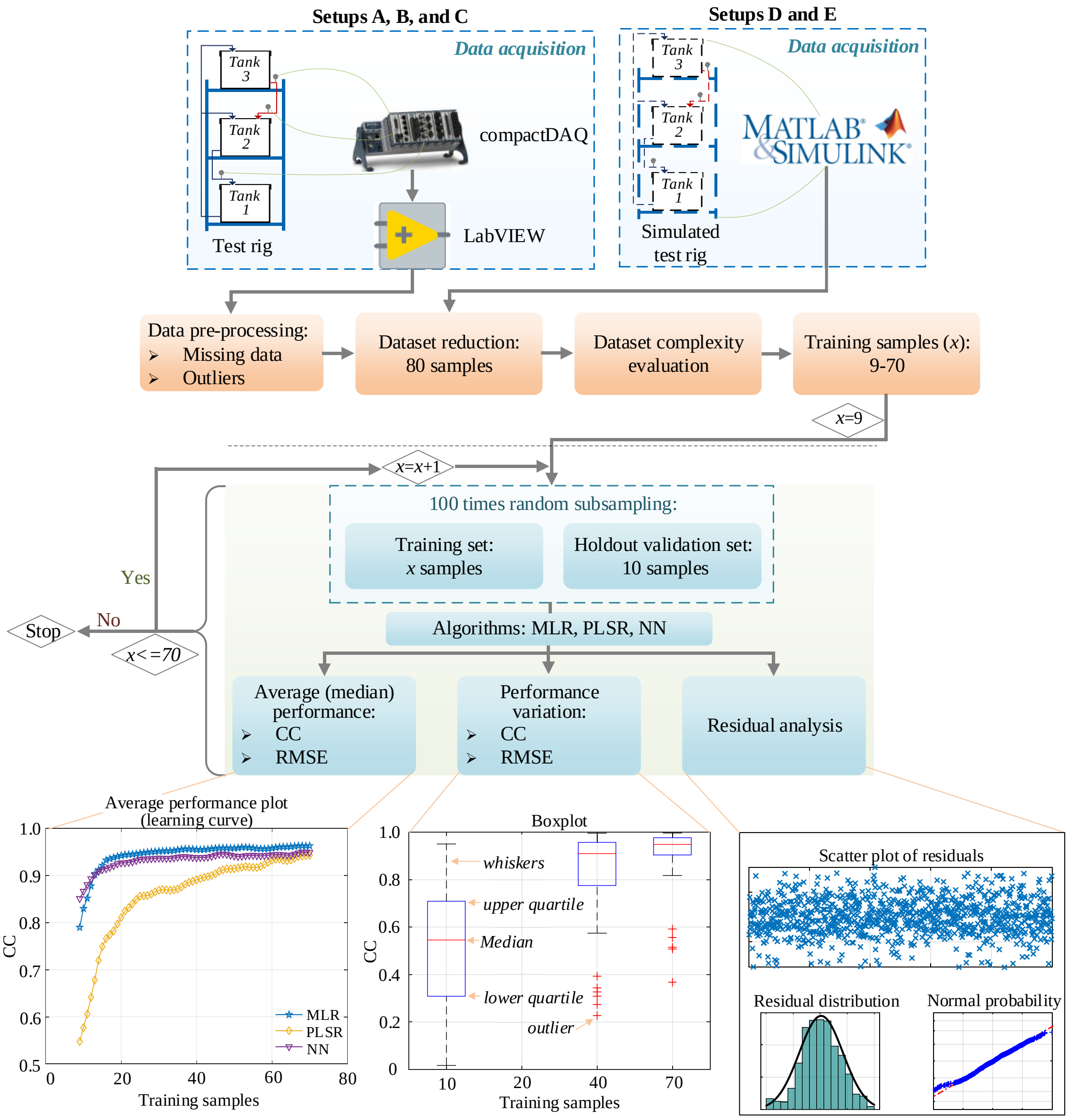


Figure 4.4: PMD development methodology for acquired datasets from the experimental rig

Subsequently, each dataset was analysed in terms of dataset complexity. Therefore, the complexity measures for regression models that were introduced by Lorena et al. [175] have been applied, described in further detail in the literature section 2.3.1 and in the appendix section A. For the complexity evaluation, the datasets were normalised between [0,1] to allow an easier complexity comparison between multiple datasets.

For PMD model training, the number of training samples varied between 9 and 70. These samples were randomly selected from the original dataset (80 samples) for each setup (also referred as random subsampling). For validation, the developed MATLAB code selected 10 random samples that were not included in the model training set. This form of validation is referred to as holdout validation in the literature [165]. Holdout validation was suitable in this scenario as enough validation samples were left out of the training set. As a further benefit, the computational demand for holdout validation is lower compared to K-fold cross-validation which decreased the model validation time. This was an important aspect as the random selection of training samples, model training and validation was repeated 100 times for every training sample size to introduce diversity. This resampling procedure is suggested in the literature by several authors, e.g. Japkowicz and Shah [209] and James et al. [166]. The set number of 100 repetitions is reported in the publication by Martens and Dardenne [210] as an adequate number for sufficient precision when dealing with small dataset regression models.

In addition to the training and validation with the original datasets, bootstrap replications (BR) and artificial noise injection (ANI) were applied to the training data and the influence on the model performance was analysed for the regression algorithms. For bootstrap, the effect of different BR of the original small training datasets in terms of model performance was studied. For ANI, the influence of different noise variances, σ^2, on the model performance was analysed. The noise was added either to the input, target, or input and target variables of the training samples. In addition, BR and ANI were applied in combination. Here, the original training dataset was first resampled by BR x number of times. Artificial noise was then injected into the BR.

To evaluate the estimation capabilities of the trained PMDs for the 10 left-over validation samples, the developed MATLAB code calculated the RMSE and CC between the measured / simulated target variable and the estimated target variable by the PMD. In the results section below, the average (median) RMSE and CC from the 100 repetitions for each algorithm and training sample size are shown for validation. The increasing number of training samples allowed to study the effect of training sample size towards the model performance and resulted in a learning curve [165]. Besides the average performance, the performance variation of the 100 training repetitions was analysed and are presented below. Similar to the procedure in the publication by Tange et al. [187], boxplots are used to display the performance variation for the regression algorithms for selected training sample sizes. Boxplots are valuable tools for the comparison of model performances and display the following information [211] (additionally pointed out in the boxplot in Figure 4.4):

- Horizontal line inside the boxes: median value
- Upper / lower edge of the box: upper / lower quartile
- Upper / lower end of the whisker: 1.5-times interquartile range
- Points above / below the whiskers: outliers

Furthermore, the residuals between the estimated target variable by the PMD and measured / simulated target variable were analysed to study the model fit. As shown in the Figure 4.4, the scatter plot of the residuals was used to analyse if a residual pattern is visible from the fitted values. In two addition plots, the residual distribution was studied: With the help of a superimposed normal density function, the residuals were examined for being distributed normally. Any departure from normality was also made visible with the help of a normal probability plot.

Significance tests to identify performance differences between regression algorithms or between different training sample sizes of the same algorithm were not conducted. Several authors [209], [212], [213] reported that resampling procedures of datasets violate the assumption of sample independence which is vital for ordinary significance tests such as paired t-test and ANOVA. Dietterich [212] demonstrated that for paired t-tests, the violation of sample independence due to resampled data leads to a high probability of incorrectly detecting a difference when no difference exists (type one error). According to Japkowicz and Shah [209], this leads to a gravely affected t-test result with limited validity. To overcome this limitation, modified significance tests have been suggested in the literature; the corrected resampled t-test [214], and the 5x2 cross-validated t-test [212]. Though, these tests cannot be applied to test different training sample sizes of the same algorithm nor for very small sample sizes. Due to these limitations of significance tests and in accordance to publications with similar testing procedures of regression algorithms [171], [187], [202], significance tests were not implemented in the result sections.

In general, the research shown in this chapter is not focusing on demonstrating a significant difference in algorithm performances in a specific use case. Instead, this research aims to provide general guidance for the selection of a suitable modelling strategy for PMDs in small dataset scenarios for industrial users.

4.2.3 Input variable reduction and model parameters

For the PMD setups A-E, the model input variables were selected based on process knowledge. To increase the performance of the trained algorithms, in particular for smaller datasets, it is important to reduce the number of input variables. As outlined by Andersen and Bro [158], fewer numbers of input variables decrease the risk of overfitting and require less computational demand. In particular, for MLR and NN, the reduction of input variables can have a considerable effect on model performance. However, if too few input variables are selected, the model performance decreases. In order to identify a suitable subset of input variables for setups A-E, the single input variables were analysed in relation to their relevance towards the target variable. This identification was based on the variable selection strategies that were recommended by Andersen and Bro [158] (section 2.3.1): Input variable correlation to target variable and the selectivity ratio. Based on the procedure of forward selection, the subset with the highest performance was identified.

Compared to MLR and NN, the reduction of input variables is less significant for PLSR since the algorithm is based on latent variables (further explained in section 2.3.1). Nonetheless, the number of latent variables for the most accurate PLSR performance must be identified. For the PMD setups below, various numbers of latent variables were tested during PLSR training and validation. Subsequently, the number of latent variables that resulted in the most accurate PLSR performance was selected.

A similar approach was applied to the training of NNs. All NNs for setups A-E were developed with the MATLAB Neural Network Toolbox. An example of a NN layout architecture for the developed PMDs in setups A-E is presented in Figure 4.5.

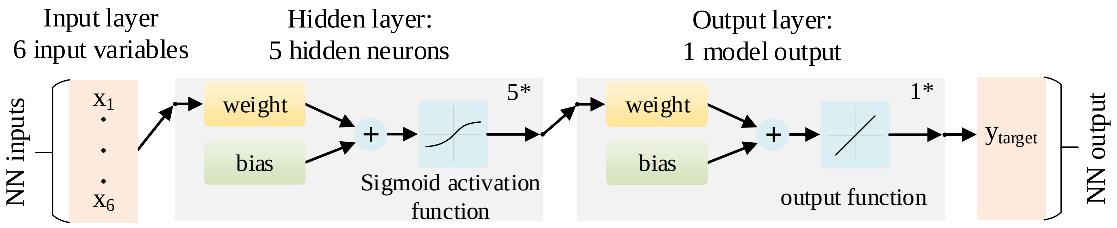


Figure 4.5: Example of developed NN architecture for PMDs setups A-E

According to the outlined literature in section 2.3.1, several model parameters of NNs have to be set during the model training stage. To keep the structure of the trained NNs for the setups A-E small, one hidden layer has been applied. As activation function for the neurons, the sigmoid function was used. The NNs were trained by the Levenberg-Marquardt algorithm since this algorithm provides a high training efficiency for small size NNs according to Yu and Wilamowski [186]. The initial value of weights for the connections between the neurons was chosen randomly. The maximum number of optimisation cycles was set to 1000. However, the number of optimisation cycles was usually much smaller than 1000, since the used NN training algorithm incorporated an automated stop function, which was activated when the internal error parameter did not decrease any further for six iterations in a row. To identify the optimum number of hidden neurons within the hidden layer, several NNs were trained with varying numbers of hidden neurons for each setup. The number of hidden neurons that resulted in the highest performance was then selected. This selection approach for the number of hidden neurons has been applied in various publications in the literature [138], [163], [201], [202].

4.3 Experimental results

The goal of this investigation was to analyse the performance of different regression algorithms on small datasets with varying complexity for use as a PMD. This section outlines the results that were obtained for the five PMDs that were developed based on the experimental rig system described in Figure 4.1. First, the results of the dataset complexity evaluation for setups A-E is shown. The subsequent section presents the results for the PMDs that were developed on the original datasets. Next, the impact of input variable reduction on the model performances is analysed for each setup. Finally, the effect of training sample modification with BR and ANI is studied for the PMD algorithms that resulted in the highest performance of each setup. The results of the PMD performance studies are discussed separately in each subsection.

4.3.1 Dataset complexity evaluation for setups A-E

To assess the regression complexity of the available datasets, the measures from Table 2.3 were used. Table 4.3 presents the results for the setups A-E. The upwards arrow ↑ denotes that higher values of the specific parameter indicate more complex regression problems, whereas the downwards arrow ↓ denotes the opposite. The measures have been applied to the entire, original datasets (80 samples). As a graphical illustration, Figure 4.6 depicts the result of the complexity measures for setup A-E in the form of a radar chart. In this chart, the coloured lines represent the setups and each vertex is related to a different regression complexity measure.

Table 4.3: Regression complexity measures for experimental PMD setups, values for entire datasets (80 samples)

	Correlation			Linearity		Smoothness		Geometry, topology, and density		
	C1↓	C2↓	C3↑	L1↑	L2↑	S1↑	S2↑	L3↑	S3↑	T1↓
Setup A	0.81	0.44	0.08	0.06	0.01	0.62	0.02	0.004	0.005	11.4
Setup B	0.82	0.29	0.09	0.04	0.01	0.71	0.01	0.002	0.003	10.0
Setup C	0.95	0.27	0.08	0.06	0.01	0.69	0.01	0.004	0.001	8.9
Setup D	0.61	0.33	0.34	0.11	0.02	0.76	0.02	0.012	0.004	16.0
Setup E	0.52	0.23	0.60	0.18	0.05	0.64	0.03	0.031	0.021	20.0

Complexity measure explanation (according to Table 2.3 and appendix A)		
	Measuring feature	Parameter relying on
C1	Maximum inputs correlation to target variable	Single inputs-target correlations
C2	Average inputs correlation to target variable	Average inputs-target correlation
C3	Collective input efficiency	Correlation + linear fit + residual error
L1	Mean absolute error	Multiple linear regression
L2	Residual variance	Multiple linear regression
S1	Inputs distribution	Euclidean distance
S2	Error of nearest neighbour regressor	Nearest neighbour regression
L3	Non-linearity of linear regressor	Interpolation + linear fit
S3	Non-linearity of nearest neighbour regressor	Interpolation + nearest neighbour regression
T1	Average number of samples per dimension	Samples + input variables

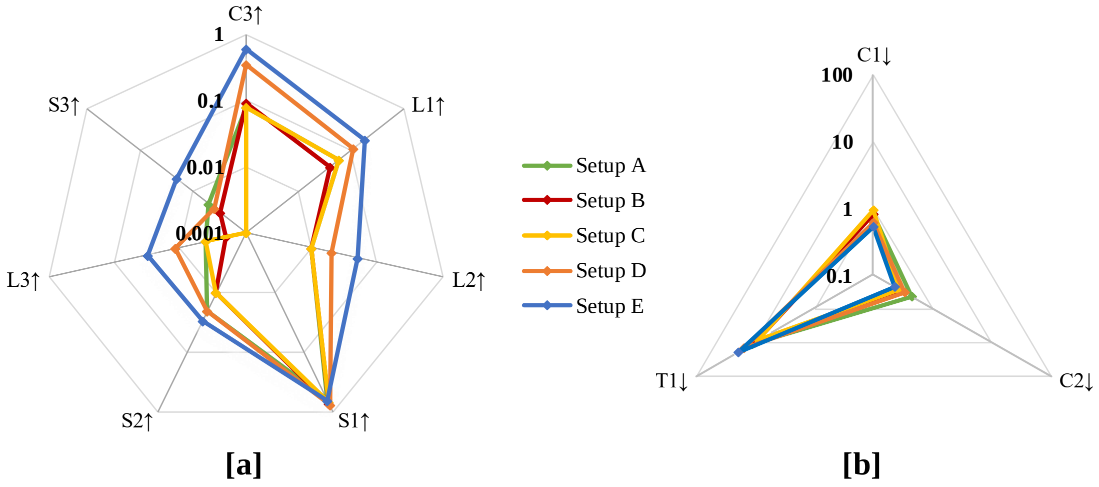


Figure 4.6: Radar charts of regression complexity measures for setups A-E, logarithmic scaling: complexity measures indicating increasing complexity by higher values [a], complexity measures indicating increasing complexity by lower values [b]

From Table 4.3 and the graphs in Figure 4.6 it can be observed that PMD-setups A, B, and C are showing similar results in all four complexity areas: correlation (C1, C2, and C3), linearity (L1 and L2), smoothness (S1 and S2), and geometry, topology, density (L3, S3, T1). In Figure 4.6 [a] and [b], setups A, B, and C cannot be clearly distinguished from one another. When compared with examples given by Lorena et al. [175], the values for setups A, B, and C indicate simple regression problems. In particular, the high values for the parameter of maximum input correlation to the target variable C1 and the low values of the linearity measures L1 and L2 imply that linear PMD algorithms can be suitable to estimate the target variable with sufficient accuracy. PMD-setup D differs in most of the measurements from the previous three. Lower correlation and higher values for linearity, smoothness, and geometry values (except for S3 and T1) suggest a more complex regression problem. In Figure 4.6 [a], setup D can be clearly distinguished from the other setups, except for the parameter of the input distribution S1 and the non-linearity of the nearest neighbour regressor S3. For PMD-setup E, the values of the measures are even higher than the values of PMD-setup D. This implies that the dataset of PMD-setup E shows the highest complexity of all five setups. The low value of the measure for maximum input correlation to the target variable C1 and the high values of the linearity complexity measures L1 and L2 indicate that linear PMD algorithms might fail to estimate the target variable accurately. Similar to setup D, setup E can be clearly distinguished in the radar chart of Figure 4.6 [a] from the other setups, except for the parameter of the input distribution S1.

4.3.2 PMDs trained on original datasets with all input variables

Setup A

The developed PMD in setup A estimates the mains water that was added to tank 1. According to Table 4.2, seven input variables were identified from the process knowledge as being relevant to the target variable FT1. In this section, the three PMD algorithms were trained and validated on the original, pre-processed dataset without any modifications. The training of MLR, PLSR, and NN included the variation of sample numbers between 9 and 70. For each training set size, 100 repetitions were conducted with randomly drawn samples. Figure 4.7 and Figure 4.8 present the results for setup A. In Figure 4.7, [a] and [b] show the average performance for the increasing sample number for the three algorithms in terms of CC and RMSE. For all figures that show the average performance within this experimental results section, moving average filters were applied to improve the visual representation of the performance over several training samples. Furthermore, these plots show various results for NNs and PLSR. NNs were trained with different numbers of hidden neurons in one hidden layer. The number of applied hidden neurons is shown in the performance figures by the attached number in the legend, e.g. NN_5 represents the performance of a NN that was trained with five hidden neurons. The figures below display the NN with the number of hidden neurons that resulted in the highest performance as well as one further NN with a different number of hidden neurons to allow a comparison. For each setup, the performance of NNs with a further variety of hidden neurons is presented in the appendix, for setup A in Figure D.9. This figure arrangement was used to identify the optimum number of hidden neurons for the trained NNs, for setup A, NNs with two to four hidden neurons resulted in the highest performance. To decrease the NN algorithm complexity, the lowest number of hidden neurons was chosen for similar performing NNs, therefore, for setup A two hidden neurons were selected as the NN with the highest performance.

For the shown PLSR algorithms in the performance plots below, different numbers of latent variables are shown that resulted in high performances. Similar to the NNs, the attached number behind PLSR in the legend indicates the number of latent variables for the specific PLSR algorithm, e.g. PLSR_5 represents the performance of a PLSR algorithm that was trained with five latent variables.

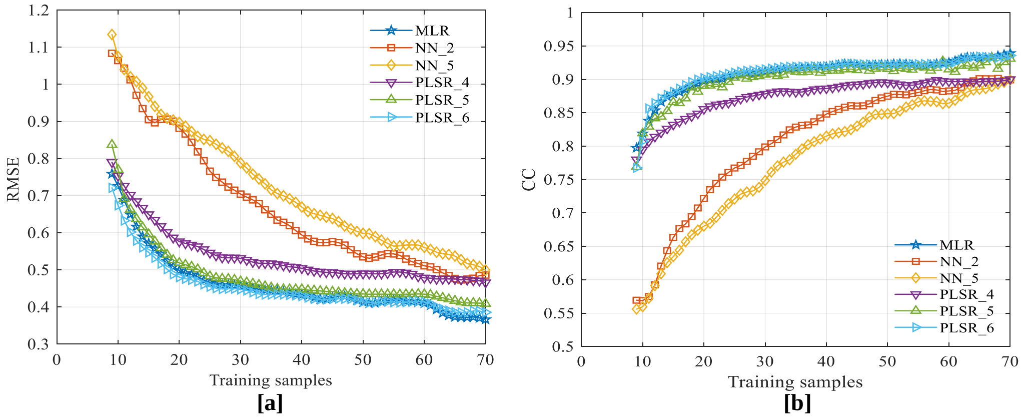


Figure 4.7: Setup A: average regression performance, all input variables: RMSE (Root mean squared error) [a] and CC (Correlation coefficient) [b] performance for different regression algorithms with increasing training samples

For setup A, the presented results in Figure 4.7 reveal that the average performance of MLR and PLSR is higher than the NN with the highest accuracy (two hidden neurons: NN_2) over the entire training sample range, in particular for small training samples. The almost constant performance of MLR and PLSR for training samples larger than approximately 30 suggests that the models have approached their minimal model error. This contrasts with the performance of NNs, where the performance is increasing over the range of training samples, in particular for smaller training samples. This suggests that more than 70 training samples might result in a further NN performance increase. For PLSR, the performance with the highest accuracy is obtained for six latent variables. Between MLR and PLSR_6, a slightly lower RMSE and higher CC can be seen for PLSR_6 when very small training samples are applied.

In addition to the average performance, Figure 4.8 [a], [b], and [c] present the RMSE performance variation of the three algorithms for the 100 training and validation repetitions. The corresponding CC performance variation for the same algorithms are displayed in the appendix in Figure D.9 for setup A. The boxplots in Figure 4.8 [a], [b], and [c] display the performance variation for training sample sizes 10, 20, 40, and 70 for MLR and the best-performing algorithms of PLSR (PLSR_6) and NNs (NN_2). In addition, Figure 4.8 [d] presents an example of the residuals that were obtained from an MLR model which was trained with 40 training samples.

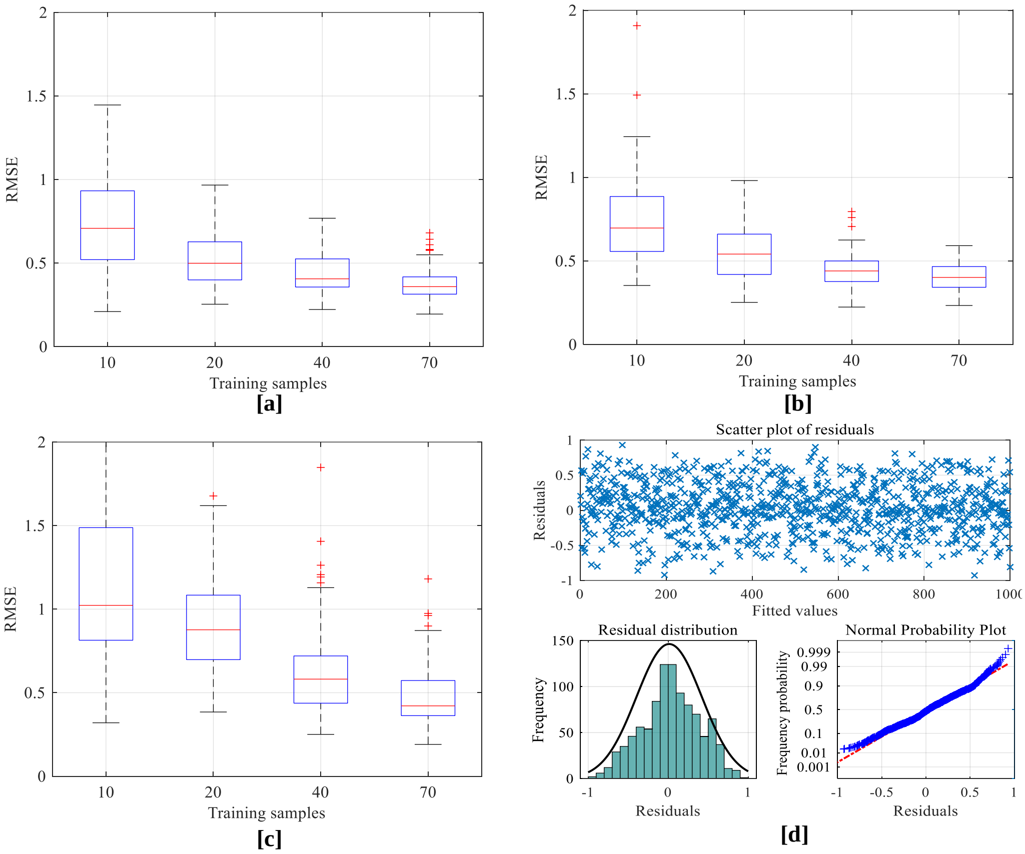


Figure 4.8: Setup A: RMSE variation and residual plots, all input variables: box plots for performance variations of algorithms MLR [a], PLSR_6 [b], and NN_2 [c], example of residual analysis for MLR algorithm trained with 40 samples [d]

The boxplots in Figure 4.8 reveal that all three algorithms show a similar influence of training sample size on the model performance: With the increasing number of training samples, the performance variation decreases. For NN_2, the variation of the performance is higher throughout the displayed number of training samples compared to MLR and PLSR_6. Similar performance variations can be seen for MLR and PLSR_6.

The plots in Figure 4.8 [d] demonstrate that the residuals for an MLR algorithm trained with 40 samples are randomly distributed along with the fitted values. No pattern can be observed. The two lower plots in [d] show that the residuals are approximately normally distributed.

Setup B

The developed PMD in setup B estimates the water temperature of tank 1 after the mixing of several flows. According to Table 4.2, eight input variables were identified from the process knowledge as being relevant to the target variable TT2. Similar to the results of setup A, the average performance over an increasing number of training samples for MLR, PLSR, and NNs are presented for setup B in Figure 4.9 and the RMSE performance variation and residual plots are displayed in Figure 4.10. The corresponding CC performance variation is shown in the appendix in Figure D.11.

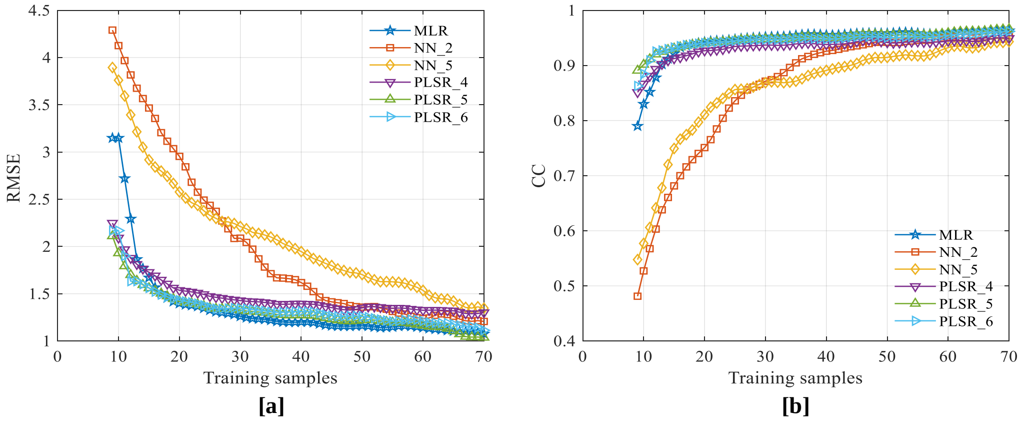


Figure 4.9: Setup B: average regression performance, all input variables: RMSE [a] and CC [b] performance for different regression algorithms with increasing training samples

The results for setup B are comparable to the results obtained for setup A: MLR and PLSR perform better than the trained NNs over the entire training sample range. According to Figure D.11 in the appendix, NNs with two and three hidden neurons resulted in comparable performances for setup B. MLR and PLSR show almost constant performances from a training sample size of 30 onwards, whereas, the NNs performance is increasing over the range of training samples, in particular for smaller training samples. For PLSR, the utilisation of five or six latent variables resulted in comparable performances. Between MLR and PLSR_5/ PLSR_6, a slightly higher CC and lower RMSE can be seen for PLSR when very small training samples are applied (<15 training samples). This corresponds to the shown performance variations for the three algorithms in the boxplots of Figure 4.10: in [a] and [b] of this figure, a lower variation and higher median performance can be observed for PLSR_5 compared to MLR when 10 training samples are applied. For 20 and more training samples, this difference disappears. The performance variation of NN_2 in [c] of Figure 4.10 is higher for all training samples compared to MLR and PLSR_5. Figure 4.10 [d] presents the residuals for PLSR_5 trained with 40 training samples. No pattern can be observed. The two lower plots in [d] show that the residuals are approximately normally distributed.

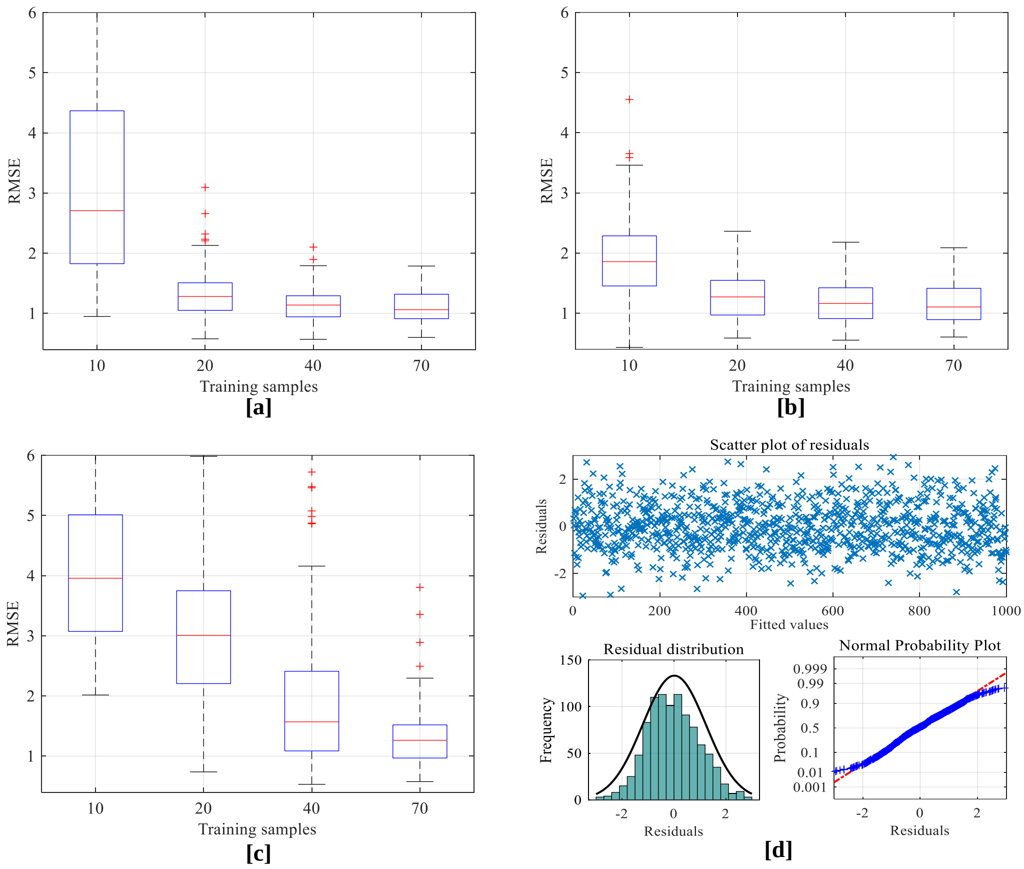


Figure 4.10: Setup B: RMSE variation and residual plots, all input variables: box plots for performance variations of algorithms MLR [a], PLSR_5 [b], and NN_2 [c], example of residual analysis for PLSR_5 algorithm trained with 40 samples [d]

Setup C

The developed PMD in setup C estimates the pH value of tank 1 after the mixing of several flows. According to Table 4.2, nine input variables were identified from the process knowledge as being relevant to the target variable pHT1. Similar to the previous setups, the average performance for MLR, PLSR, and NN over an increasing number of training samples is presented for setup C in Figure 4.11. The RMSE performance variation and residual plots are displayed in Figure 4.12 and the corresponding CC performance variation is shown in the appendix in Figure D.13.

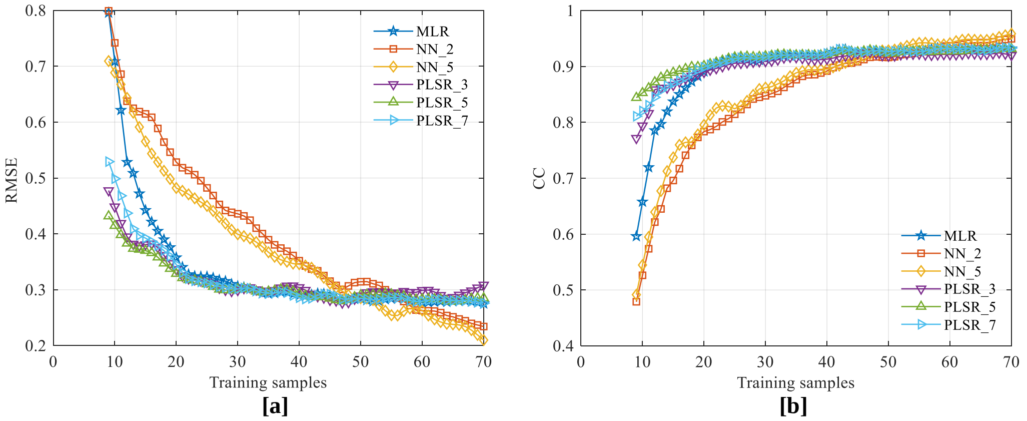


Figure 4.11: Setup C: average regression performance, all input variables: RMSE [a] and CC [b] performance for different regression algorithms with increasing training samples

For the shown average performance of setup C in Figure 4.11, MLR and PLSR outperform NN for small training samples. However, when the training sample size exceeds approximately 50 training samples, the shown NNs result in slightly higher performances. Although, the applied complexity measures of Table 4.3 resulted in comparable values for setups A, B, and C, this outperforming behaviour of NNs compared to MLR and PLSR for higher training samples was not seen in setups A and B. A potential explanation for these differences is that the applied complexity measures do not represent the entire regression complexity of the datasets. As a result, slightly different model performances are observed for datasets with similar regression complexity. For setup C, the best result for PLSR was achieved with five latent variables and for NNs with five to six hidden neurons (Figure D.13 in the appendix). When the training sample size is very small (smaller than approximately 20), PLSR_5 shows a higher average performance and lower performance variation compared to MLR as displayed in Figure 4.11 and Figure 4.12 [a] and [b]. According to setups A and B, the best performing NN (NN_5) within setup C showed a higher performance variation for training samples 10, 20, and 40, however, for 70 training samples a similar variation was achieved compared to MLR and PLSR_5 (Figure 4.12 and Figure D.13). Figure 4.12 [d] shows that no pattern and approximately normal distribution can be observed for the residuals that resulted from a trained PLSR_5 model with 40 training samples.

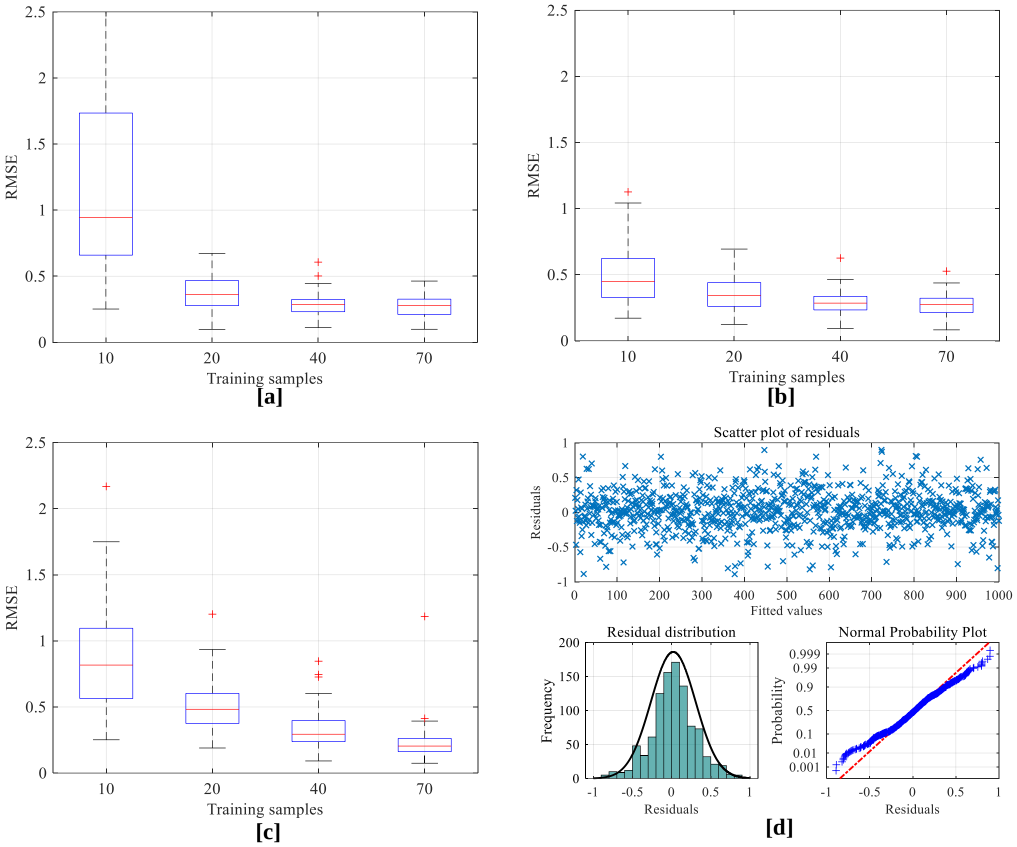


Figure 4.12: Setup C: RMSE variation and residual plots, all input variables: box plots for performance variations of algorithms MLR [a], PLSR_5 [b], and NN_5 [c], example of residual analysis for PLSR_5 algorithm trained with 40 samples [d]

Setup D

The developed PMD in setup D estimates the filling level of tank 1 which depends on the addition of mains water and the transition from tank 3 as well as the outflow to tank 2. According to Table 4.2, five input variables were identified from the process knowledge as being relevant to the target variable LI1. Similar to the previous setups, the average performance for MLR, PLSR, and NNs over an increasing number of training samples is presented for setup D in Figure 4.13. The RMSE performance variation and residual plots are displayed in Figure 4.14 and the corresponding CC performance variation is shown in the appendix in Figure D.15.

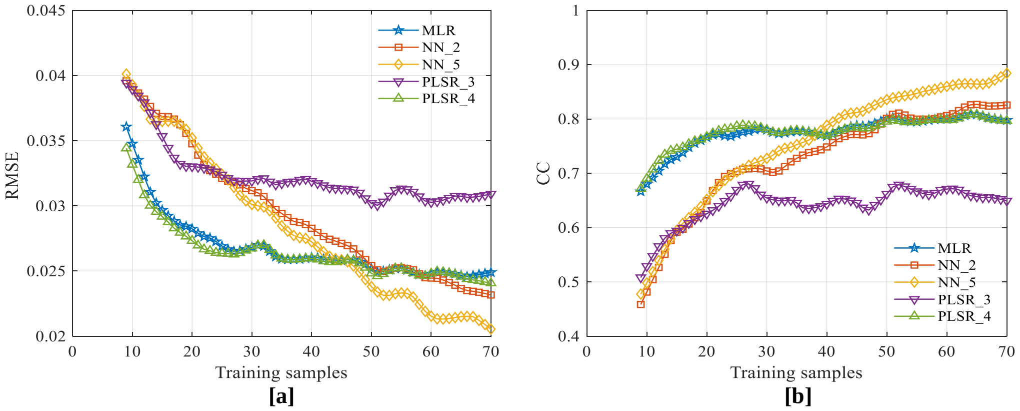


Figure 4.13: Setup D: average regression performance, all input variables: RMSE [a] and CC [b] performance for different regression algorithms with increasing training samples

The results for the trained PMDs for setup D indicate that NNs outperform MLR and PLSR for training samples larger than approximately 50. In particular, for five to seven hidden neurons (shown in Figure D.15 in the appendix), higher average performances and lower performance variations are obtained for NNs when the training sample size exceeds 50, as displayed in Figure 4.13 and Figure 4.14. From the setup D complexity measures in Table 4.3, higher NNs performances were expected compared to MLR and PLSR. Though, for training samples lower than 50, MLR and PLSR_4 result in higher average performances and lower performance variations. For PLSR, the addition of one further latent variable (PLSR_4 compared to PLSR_3) results in a clear performance increase. For MLR and PLSR_4, nearly similar performance variation can be seen in Figure 4.14 [a] and [b]. Similar to the results from the previous setups A-C, MLR and PLSR_4 show almost constant performances from 40 training samples onwards, whereas, the performance of NN_5 is continuously improving throughout the training sample range and the trend suggests that even better performances can be achieved with larger training samples. The presented residual analysis plots for a NN with five hidden neurons and 40 training samples in Figure 4.14 [d] does not show any residual pattern and confirms that the residuals are approximately normally distributed.

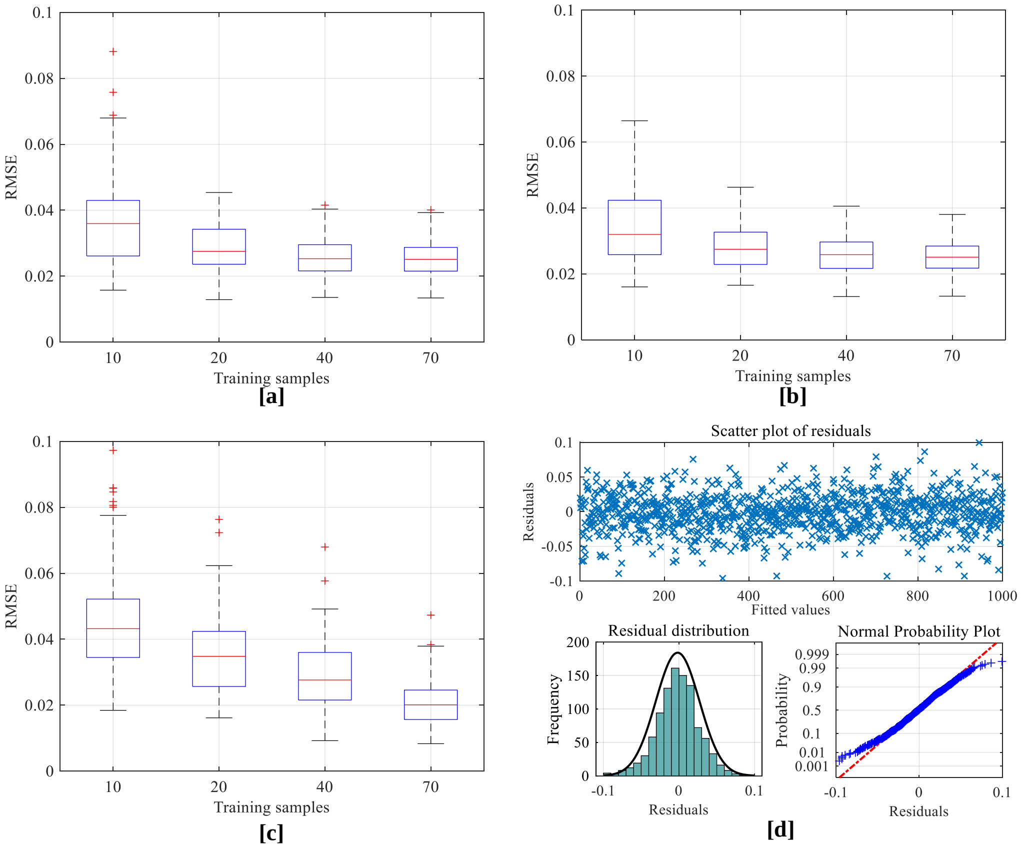


Figure 4.14: Setup D: RMSE variation and residual plots, all input variables: box plots for performance variations of algorithms MLR [a], PLSR_4 [b], and NN_5 [c], example of residual analysis for NN_5 algorithm trained with 40 samples [d]

Setup E

The developed PMD in setup E estimates the filling level of tank 1, which depends on the addition of mains water and the transition from tank 3 as well as the outflow to tank 2. According to Table 4.2, four input variables were identified from the process knowledge as being relevant to the target variable LI1. The average performance for MLR, PLSR, and NNs over an increasing number of training samples is presented for setup E in Figure 4.15. The RMSE performance variation and residual plots are displayed in Figure 4.16 and the corresponding CC performance variation is shown in the appendix in Figure D.17.

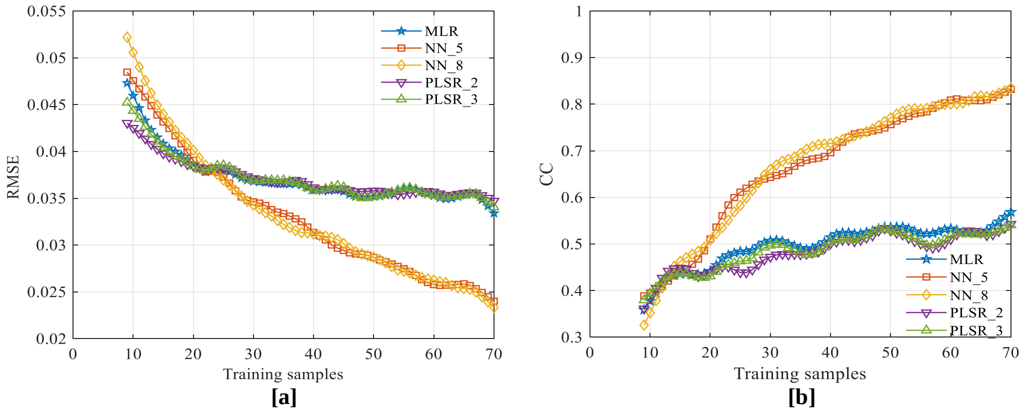


Figure 4.15: Setup E: average regression performance, all input variables: RMSE [a] and CC [b] performance for different regression algorithms with increasing training samples

From the complexity measurements displayed in Table 4.3, the dataset of setup E resulted in the highest complexity for most of the applied measures. This high dataset complexity and in particular the high values of the linearity measures L1 and L2 explain the distinct outperformance of the non-linear NNs compared to the linear algorithms MLR and PLSR for training samples larger than 20 in Figure 4.15 and Figure 4.16. According to Figure D.17 in the appendix, five to eight hidden neurons resulted in the best NN performance for setup E. MLR and the best performing PLSR (PLSR_3) failed to accurately estimate the target variable. The only exception from this can be seen for training samples lower than 20, where MLR and PLSR_3 resulted in slightly higher average performances and lower performance variations compared to NN_5. As a NN with five hidden neurons proved to be a suitable model for setup E, the residuals for this algorithm are displayed in Figure 4.16 [d] for 40 training samples. These residuals display a random pattern and are in line with a normal distribution.

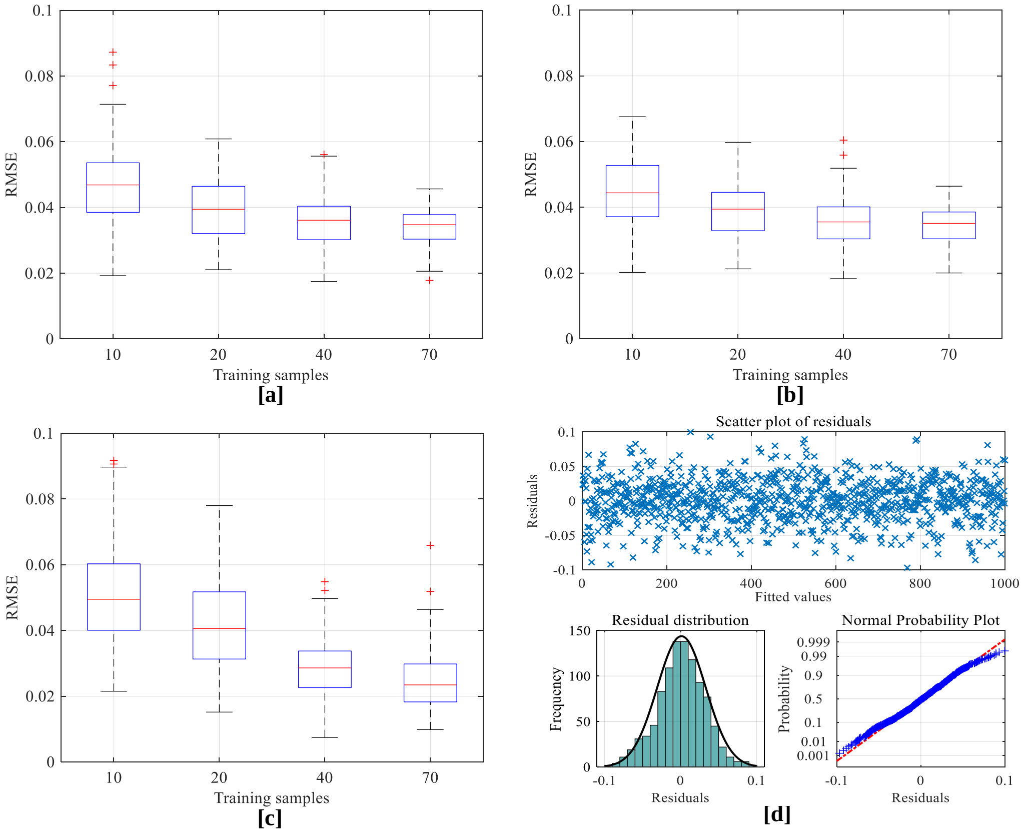


Figure 4.16: Setup E: RMSE variation and residual plots, all input variables: box plots for performance variations of algorithms MLR [a], PLSR_3 [b], and NN_5 [c], example of residual analysis for NN_5 algorithm trained with 40 samples [d]

Discussion of PMDs trained on original datasets with all input variables

The training and validation of the PMDs for setups A-E resulted in diverse outcomes. As already expected from the recommendations for regression algorithms in Table 2.4, MLR and PLSR are well suited to infer target variables when the complexity measures suggest a regression problem with high correlation and linearity, as seen for setups A, B, and C. Nonetheless, for setup C the NNs slightly outperformed MLR and PLSR for training samples between 50 and 70. One possible explanation is that the complexity measures do not represent the entire complexity of a dataset. Further investigations are needed to test this hypothesis and to analyse the reasons for algorithm performance variations when being applied to datasets with similar regression complexity. For very small training sample sizes, the results demonstrated that for all setups, independent of the dataset complexity, linear regression algorithms MLR and PLSR revealed higher performances compared to NNs. Hence, MLR and PLSR are less affected by very small training sample sizes than NNs. From the two linear algorithms, PLSR, with the optimal number of latent variables, showed higher average performances and smaller performance variations for very small training samples compared to MLR, in particular when many input variables were applied (setup C). This result corresponds to recommendations that were reported by Tange et al. [187] and in Table 2.4 which state that PLSR is less effected by small training samples in comparison to MLR and NN. Since PLSR is the least flexible model out of the three, it is less prone to overfitting and hence less sensitive to small training samples [187]. NNs were the applied algorithm with the highest flexibility and therefore resulted in the greatest overfitting for very small training samples. However, with increasing dataset complexity of setups D and E, NNs outperformed MLR and PLSR when enough training samples were available. The nearly constant performance of all trained MLR and PLSR algorithms for larger training samples suggests that the models have approached their minimal model error for all setups. In contrast to this, the trained NNs for all setups (in particular for setups C, D, and E) showed a steady increase in performance, which suggests that larger training samples might lead to further performance improvement.

All shown residual plots for the selected algorithms of the different setups resulted in a random pattern and the residuals were approximately normally distributed. These outcomes indicate that the examined algorithms were suitable for explaining the predictive information of the target variables.

Since various input variables were used in the setups, the next section analyses the possible model improvements that are achievable when the number of input variables for setups A-E is reduced.

4.3.3 PMDs trained on original datasets with reduced input variables

In this section, the influence of the number of input variables for the PMD models of setups A-E is analysed. The aim is to identify the most relevant input variables in relation to the target variables in order to remove the input variables with less influence to reduce the risk of overfitting. According to Andersen and Bro [158], the risk of overfitting increases for regression models with larger input variables to samples ratio. The authors state that input variable selection is an iterative process that identifies the best variable combination. As outlined in section 4.2.3, the input variable correlation to the target variable and the selectivity ratio were used to identify the relevance of the single input variables. The parameter for the input variable correlation to the target variable is identical to the complexity measure C1 in Table 4.3. High values of correlation and selectivity ratio indicate that the analysed input variable has high relevance for the estimation of the target variable. The information of both measures was used in the subsequent step, where subsets of input variables were trained with 10, 20, 40 and 70 training samples. Following the forward selection method, the most relevant input variable was selected first and further input variables were added. Different subset combinations have been trained and the performance of the PMD was analysed for the selected training samples. According to this procedure, the subset with the highest performance was identified.

Setup A

Within setup A, seven input variables were identified as being relevant to the target variable FT1 and have been applied in the PMD model within section 4.3.2. Table 4.4 presents the values of the correlation to the target variable as well as the selectivity ratio for each of the seven input variables.

Table 4.4: Setup A: Relevance of single input variables to the target variable

	FT2	FT4	TT1	$TT2_{\Delta Temp}$	$TT2_{\Delta time}$	TT4	TT5
Correlation to target variable	0.81	0.33	0.45	0.38	0.72	0.09	0.29
Selectivity ratio	2.85	0.21	0.22	0.16	1.49	0.04	0.07

The results show, that a wide range of relevance can be seen throughout the input variables. Both measures identify FT2 as the most relevant input variable to the target variable FT1. Therefore, this variable is selected and used in a simple linear regression (SLR) model. The performance of this model, as well as an extract of further subset performances of input variables, are presented in Table 4.5. In this table, the average performance and the performance variation (in form of the standard deviation) are shown for an MLR (SLR for one input variable) model and the training sample sizes 10, 20, and 70. The set of input variables with the highest accuracy is shown in bold, three input variables for setup A according to Table 4.5.

Table 4.5: Setup A: Regression performance of selected input variable subsets, subset with the highest accuracy highlighted in bold

Input variable(s)		Training samples	CC		RMSE	
			SLR/MLR		SLR/MLR	
Number	Name		Average	Std	Average	Std
1	FT2	10	0.89	0.13	0.53	0.16
		20	0.88	0.12	0.54	0.13
		70	0.89	0.11	0.51	0.13
3	**FT2; $TT2_{\Delta time}$; FT4**	10	0.90	0.12	0.51	0.20
		20	0.92	0.11	0.42	0.15
		70	0.94	0.09	0.36	0.11
7 (all)	FT2; $TT2_{\Delta time}$; FT4; TT1; TT5; TT4; $TT2_{\Delta Temp}$	10	0.85	0.24	0.79	0.83
		20	0.90	0.12	0.52	0.16
		70	0.94	0.09	0.37	0.11

In Figure 4.17, the average RMSE performance, as well as the RMSE performance variation of a MLR model trained with the most accurate subset of input variables for setup A, are shown. This model was trained with the three input variables that were highlighted in bold in Table 4.5. In addition, Figure 4.17 displays the RMSE performance of the previously trained PMD for setup A from section 4.3.2 which included all input variables. When all input variables were used, a PLSR_6 model showed the highest performance. Figure D.10 in the appendix presents the corresponding results for the CC. The comparison of the two models in Figure 4.17 and Figure D.10 demonstrates that higher performances are achievable with the reduced number of input variables for very small training samples. When 40 and more training samples are used, the PMDs trained on all input variables and the reduced subset result in similar outcomes.

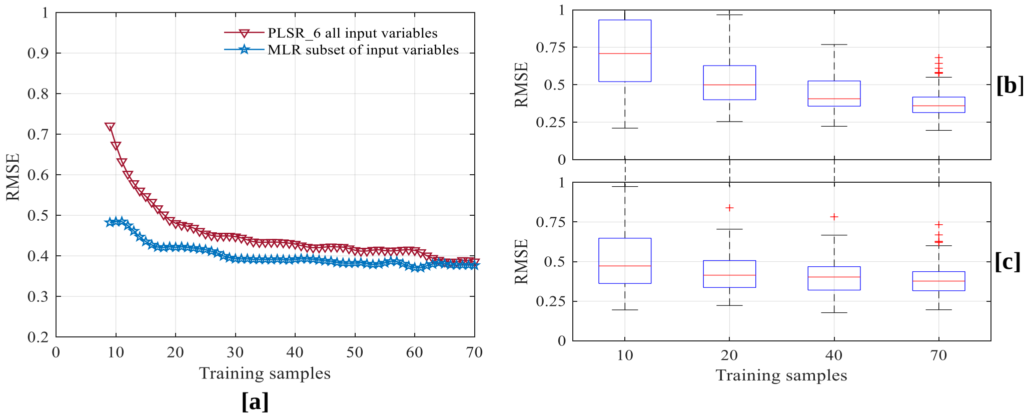


Figure 4.17: Setup A: RMSE all input variables and reduced input variable subset: average performance [a], performance variation for PLSR_6 trained on all input variables [b], performance variation for MLR trained on most accurate input variable subset [c]

Setup B

For setup B, eight relevant input variables were identified and applied in the first PMD in section 4.3.2. The relevance of the single input variables for setup B is shown in Table 4.6. High correlations to the target variable and selectivity ratios can be seen for FT4 and FT1. According to this information, different subsets of input variables were tested on their performance and a selection of these are displayed in Table 4.7. The best outcome was seen for a subset with six input variables, highlighted in bold in Table 4.7.

Table 4.6: Setup B: Relevance of single input variables to target variable

	FT1	FT3	FT4	TT1	TT2*	TT3	TT4	TT5
Correlation to target variable	0.59	0.14	0.82	0.10	0.12	0.15	0.31	0.13
Selectivity ratio	0.039	$5*10^{-5}$	0.150	0.001	$2*10^{-4}$	$4*10^{-4}$	0.008	$3*10^{-4}$

*: TT2 at t_start

Table 4.7: Setup B: Regression performance of selected input variable subsets, subset with the highest accuracy highlighted in bold

Input variable(s)		Training samples	CC		RMSE	
			SLR/MLR		SLR/MLR	
Number	Name		Average	Std	Average	Std
1	FT4	10	0.85	0.12	2.37	0.69
		20	0.86	0.13	2.29	0.58
		70	0.84	0.12	2.21	0.41
6	**FT4; FT1; TT4; TT3; TT2*, FT3**	10	0.94	0.09	1.69	0.79
		20	0.96	0.04	1.17	0.39
		70	0.97	0.03	1.03	0.26
8 (all)	FT4; FT1; TT4; TT3; TT2*, FT3; TT5; TT1	10	0.88	0.22	2.33	5.48
		20	0.95	0.05	1.22	0.44
		70	0.97	0.03	1.11	0.25

*: TT2 at t_start

Further investigations into the subset of six input variables for setup B revealed that for these input variables, a PLSR model with five latent variables was the most effective algorithm. For all input variables, the same model arrangement (PLSR_5) showed the highest accuracy as analysed in section 4.3.2. In Figure 4.18, the RMSE model performance for these two PLSR_5 models is compared, one trained on the most accurate subset of input variables and the other on all input variables. The corresponding results for the CC are displayed in Figure D.12 in the appendix.

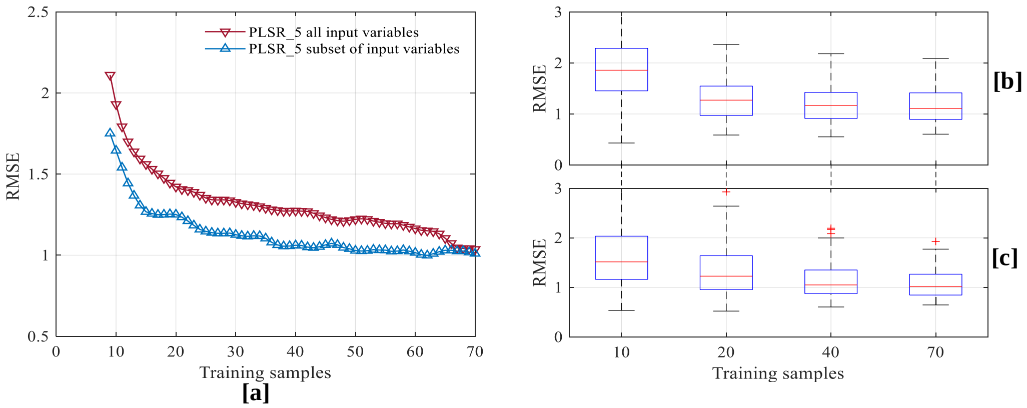


Figure 4.18: Setup B: RMSE all input variables and reduced input variable subset: average performance [a], performance variation for PLSR_5 trained on all input variables [b], performance variation for PLSR_5 trained on most accurate input variable subset [c]

According to Figure 4.18 [a], lower average performances are seen for the model with the subset of six input variables. In Figure 4.18 [b] and [c], slightly lower performance variations are seen for the subset of six input variables.

Setup C

For Setup C, the highest number of input variables has been selected based on the process knowledge. In total, nine input variables were chosen as being relevant to the target variable pHT1. To identify the input variables with the highest relevance, Table 4.8 presents the results for the relevance measures. The high values for the variable pHT2 (pH at t_start) are a notable finding. A correlation value of 0.95 and a selectivity ratio of 13.1 express a very high relevance towards the target variable. This result corresponds to the outcome of analysed performances of different input variable subsets, shown in in Table 4.9 for setup C. An SLR model with pHT2 (pH at t_start) as input variable resulted in the highest average performance and lowest performance variation throughout the analysed number of training samples.

Table 4.8: Setup C: Relevance of single input variables to the target variable

	FT1	FT3	FT4	TT1	TT3	TT4	TT5	pHT2*	pHT1*
Correlation to target variable	0.04	0.09	0.02	0.09	0.08	0.21	0.08	0.95	0.83
Selectivity ratio	0.001	0.016	0.013	0.002	0.020	0.023	0.006	13.1	2.65

*: pH value at t_start

Table 4.9: Setup C: Regression performance of selected input variables subsets, subset with the highest accuracy highlighted in bold

Input variable(s)		Training samples	CC		RMSE			
			SLR/MLR		SLR/MLR		NN (5. h. neurons)	
Number	Name		Average	Std	Average	Std	Average	Std
1	**pHT2***	10	0.95	0.04	0.26	0.09	-	-
		20	0.95	0.04	0.25	0.08	-	-
		70	0.95	0.04	0.25	0.08	-	-
3	pHT2*; TT4; pHT1*	10	0.93	0.09	0.31	0.12	0.57	0.27
		20	0.94	0.05	0.29	0.10	0.44	0.19
		70	0.95	0.04	0.23	0.08	0.26	0.10
9 (all)	pHT2*; TT4; pHT1*; TT1; FT3; TT3; TT5; FT1; FT4	10	0.63	0.22	0.90	2.40	0.72	0.37
		20	0.89	0.10	0.34	0.11	0.49	0.26
		70	0.94	0.05	0.31	0.08	0.22	0.94

*: pH value at t_start

The study of performance differences between the SLR model with pHT2 (pH at t_start) as an input variable and the best performing model with all input variables from section 4.3.2 (PLSR_5) is presented for the RMSE in Figure 4.19 and for the CC in Figure D.14. As expected, the SLR shows only very small variations of the average performance with increasing training sample number and correspondingly small boxes and whiskers for the performance variations in Figure 4.19 [c]. The PLSR_5 model trained with all input variables resulted in higher average performances throughout the entire sample range and higher performance variations for small training samples.

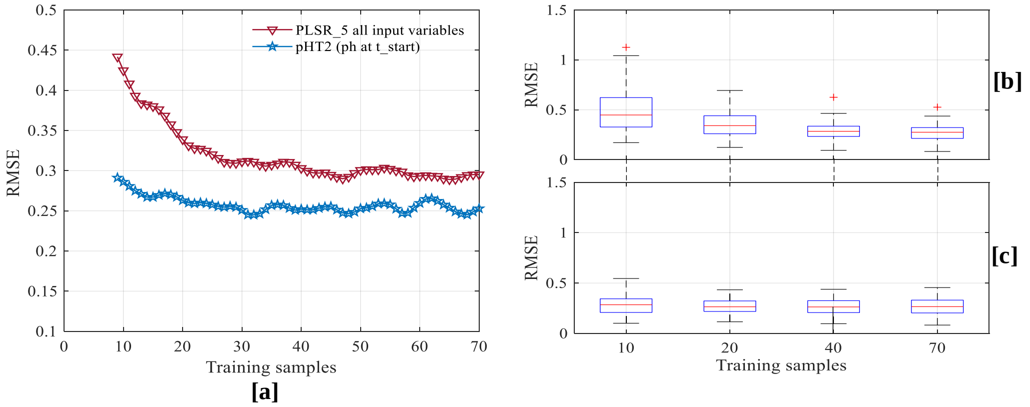


Figure 4.19: Setup C: RMSE all input variables and reduced input variable subset: average performance [a], performance variation for PLSR_5 trained on all input variables [b], performance variation for SLR with pHT2 (pH at t_start) [c]

Setup D

Within the original dataset of setup D, five input variables were identified as being relevant to the target variable LI1. As presented in Table 4.10, the maximum correlation between the single input variables and the target variable is 0.61. Compared to the previous setups A, B, and C, this maximum correlation value is lower which partly explains why the dataset of setup D shows a higher complexity and better performance of NNs with higher training samples. For setup D, an SLR model with the most relevant variable LI2 (at t_start) did not provide an accurate estimation, as displayed in Table 4.11. The analysis of different subsets of input variables revealed that the highest accuracy is seen for a subset of four input variables, highlighted in bold in Table 4.11. Nonetheless, the performance is approximately similar to the PMD that was trained with all input variables.

Table 4.10: Setup D: Relevance of single input variables to target variable

	FT4_1	FT4_2	LI1*	LI2*	t_change_FT4
Correlation to target variable	0.24	0.53	0.15	0.61	0.10
Selectivity ratio	33	278	15	366	10

*:at t_start

Table 4.11: Setup D: Regression performance of selected input variables subsets; subset with the highest accuracy highlighted in bold

Input variable		Training samples	CC		RMSE			
			SLR/MLR		SLR/MLR		NN (5. h. neurons)	
Number	Name		Average	Std	Average	Std	Average	Std
1	LI2*	10	0.63	0.18	0.034	0.007	-	-
		20	0.64	0.17	0.032	0.007	-	-
		70	0.64	0.18	0.031	0.005	-	-
4	**LI2*; FT4_2; FT4_1; LI1***	10	0.74	0.20	0.033	0.011	0.042	0.015
		20	0.75	0.15	0.029	0.009	0.034	0.013
		70	0.81	0.12	0.024	0.005	0.024	0.008
5 (all)	**LI2*; FT4_2; FT4_1; LI1*; t_change_FT4**	10	0.72	0.21	0.033	0.013	0.039	0.015
		20	0.75	0.17	0.029	0.007	0.033	0.011
		70	0.83	0.12	0.023	0.004	0.021	0.006

*:at t_start

The comparable performances of the most accurate subset of input variables and all input variables for setup D can be seen in the plots of Figure 4.20 for the RMSE and in Figure D.16 in the appendix for the CC.

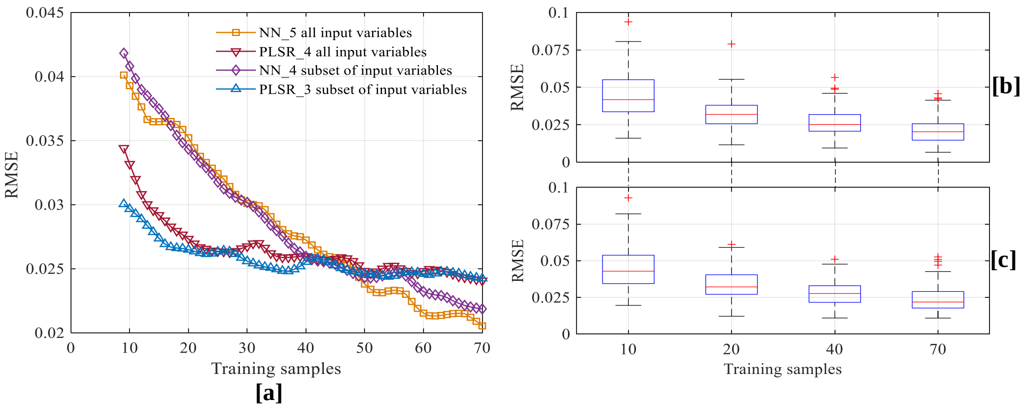


Figure 4.20: Setup D: RMSE all input variables and reduced input variable subset: average performance [a], performance variation for NN_5 trained on all input variables [b], performance variation for NN_4 trained on most accurate input variable subset [c]

For the subset of four input variables, PLSR_3 and NN_4 showed the most accurate performance for smaller and larger training samples, whereas, PLSR_4 and NN_5 were the highest performing models when all input variables were used (as seen in section 4.3.2). Figure 4.20 [a] shows that the average performances of the PLSR and NN models between the two sets of input variables are approximately similar, except for higher number of training samples where the NN_5 model (with all input variables) slightly outperforms the NN_4 model (with the subset of input variables). The performance variation plots for the NNs in Figure 4.20 [b] and [c] confirm this outcome.

Setup E

For setup E, four relevant input variables were identified and used in the first PMD in section 4.3.2. Table 4.12 displays the relevance of the single input variables. Both relevance measures identified LI1 (at t_start) as the most relevant input variable. However, compared to the previous setups, the lowest maximum correlation to the target variable is seen for the input variables of setup E.

Table 4.12: Setup E: Relevance of single input variables to the target variable

	FT3	LI1*	LI2*	LI3*
Correlation to target variable	0.16	0.52	0.17	0.06
Selectivity ratio	21	220	16	2

*:at t_start

Table 4.13: Setup E: Regression performance of selected input variables subsets, subset with the highest accuracy highlighted in bold

Input variable(s)		Training samples	CC SLR (for 1 input variable)/ NN (5 h. neurons)		RMSE SLR (for 1 input variable)/ NN (5 h. neurons)	
Number	Name		Average	Std	Average	Std
1	LI1*	10	0.51	0.21	0.039	0.008
		20	0.53	0.22	0.037	0.008
		70	0.54	0.24	0.036	0.006
3	**LI1*; LI2*; FT3**	10	0.46	0.23	0.043	0.019
		20	0.70	0.23	0.032	0.013
		70	0.86	0.16	0.023	0.008
4 (all)	LI1*; LI2*; FT3; LI3*;	10	0.37	0.24	0.050	0.020
		20	0.59	0.23	0.040	0.018
		70	0.84	0.16	0.024	0.008

A selection of different input variable subsets including their performance is presented in Table 4.13. Three input variables, highlighted in bold, resulted in the most accurate performance of all the analysed subsets.

The improved estimation performance of a PMD trained on the most accurate subset of input variables compared to the performance of the PMD trained with all input variables (as seen in section 4.3.2) can be seen for the RMSE in Figure 4.21 and for the CC in Figure D.18. For both sets of input variables, a NN with five hidden neurons showed the lowest estimation error. The comparison of the two PMDs in the three graphs of Figure 4.21 indicates that higher performances are achievable with the reduced number of input variables throughout the training sample range.

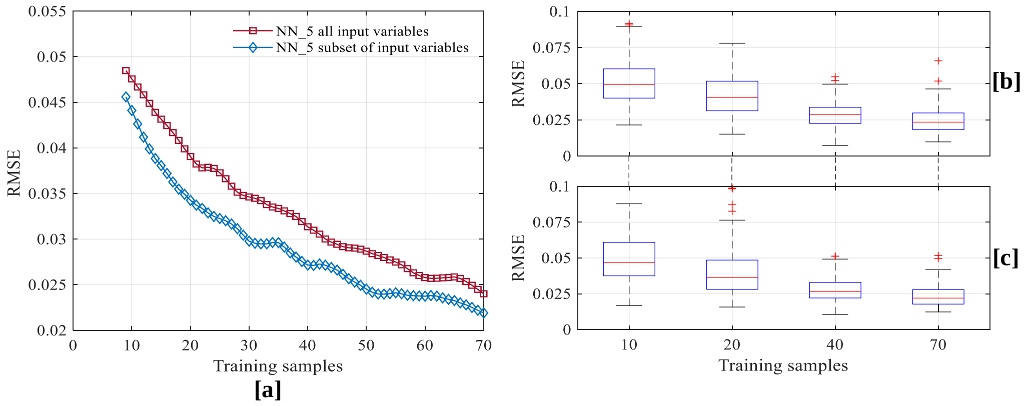


Figure 4.21: Setup E: RMSE all input variables and reduced input variable subset: average performance [a], performance variation for NN_5 trained on all input variables [b], performance variation for NN_5 trained on most accurate input variable subset [c]

Discussion of PMDs trained on original datasets with reduced input variables

The study of the performance influence for input variable reduction revealed that improvements were achievable in four out of the five setups. The results demonstrate that, especially for very small training samples, fewer input variables result in less overfitting during model training and therefore show a higher accuracy during validation. The input variable reduction in setup C to only one input variable demonstrates this behaviour clearly. Furthermore, this behaviour agrees to findings from the literature, i.e. by Andersen and Bro [158]. As Souza et al. [141] reported, input variable reduction becomes indispensable for PMDs that are trained with small samples and many input variables. The reduction of input variables must be controlled through a suitable strategy in order to identify a well-performing subset of input variables. The applied forward selection procedure enabled to find well-performing subsets of input variables with small effort. The used parameters for analysing the relevance of the single input variables (correlation to target variable and selectivity ratio) identified the input variable with the highest relevance for each setup. In addition, these parameters identified further input variables with high relevance to the target variable that were suitable for the combination within well-performing input variable subsets. However, setups A and B showed that the input variables in the most accurate subsets did not always follow the sequence of the identified input variable relevance. An explanation for this deviation is that the parameters of input variable relevance do not consider the combined effects of several input variables.

4.3.4 PMD training set manipulation through bootstrap

As the first manipulation method, the small training datasets were replicated by bootstrap. The number of BR was varied, up to 20 replications have been applied for the different setups. Similarly to the results of the previous sections, the performance of the algorithms was studied in terms of average (median) performance from 100 repetitions and performance variation after the training samples were modified by BR. The size of the training samples was increased from 9 to 70. With bootstrap, the training samples were replicated by drawing random samples with replacement from the original set of training samples. For validation, 10 random samples were applied that were not used in the training sample set nor replicated by bootstrap.

For each setup, the influence of BR has been studied with the best performing algorithm and the number of input variables that resulted in the lowest estimation error following the results from sections 4.3.2 and 4.3.3.

Setups A, B, and C

For setup A, B, and C, the following linear regression algorithms with a reduced set of input variables resulted in the highest performance according to sections 4.3.2 and 4.3.3: an MLR model with three input variables for setup A; a PLSR_5 model with six input variables for setup B; and for setup C, a SLR model with pHT2 (pH at t_start) as an input variable. Based on these model arrangements, up to 10 BR were applied to the training samples. Figure 4.22 presents the outcome of the training set modification with BR. In Figure 4.22, [a], [b], and [c] display the RMSE average performance and performance variation (for 10 and 40 training samples) for setup A, [d], [e], and [f] for setup B, and [g], [h], and [i] for setup C. In these graphs, boots 1 represents the performances of the algorithms without any modification. Figure D.19 in the appendix presents the corresponding results for the CC.

As the graphs display, similar results are obtained for the three linear algorithms of setups A, B, and C: No improvement can be seen for the average performances nor for the performance variations throughout the range of training samples. When the number of BR is increased, the algorithm performance is comparable to the non-modified training samples.

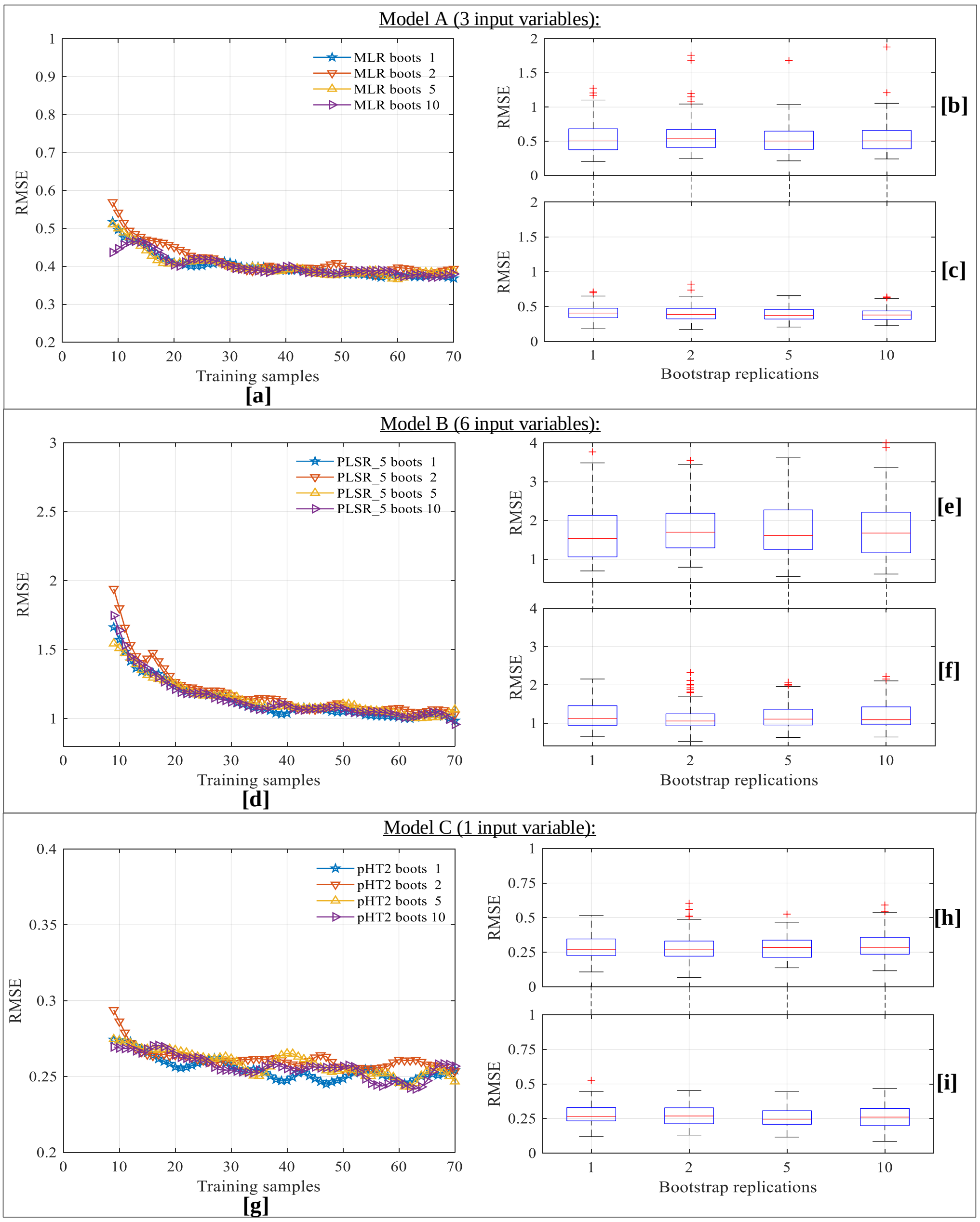


Figure 4.22: Setups A, B, C: BR influence on MLR, PLSR, and SLR; RMSE validation: average performances [a], [d], [g], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup B PLSR_5 performance variation for training samples 10 [e] and 40 [f], setup C SLR performance variation for training samples 10 [h] and 40 [i]

Setups D and E

According to sections 4.3.2 and 4.3.3, NNs showed the lowest estimation errors when higher numbers of training samples were used for the more complex datasets of setups D and E. For setup D, the optimum performance was obtained with an NN_5 model trained with all five input variables, whereas, an NN_5 model with the reduced set of three input variables showed the highest performance for setup E. The influence of up to 20 BR of the training samples was analysed for the two setups. In Figure 4.23, [a] shows the RMSE average performance for the bootstrap training sample manipulation of setup D and [b], [c] display the performance variations (for 10 and 70 training samples) for setup D. In the same figure, the equivalent plot arrangements are shown in [d], [e], and [f] for setup E. In the displayed graphs, boots 1 represents the performance of the algorithm without any modification. Figure D.20 in the appendix presents the corresponding results for the CC.

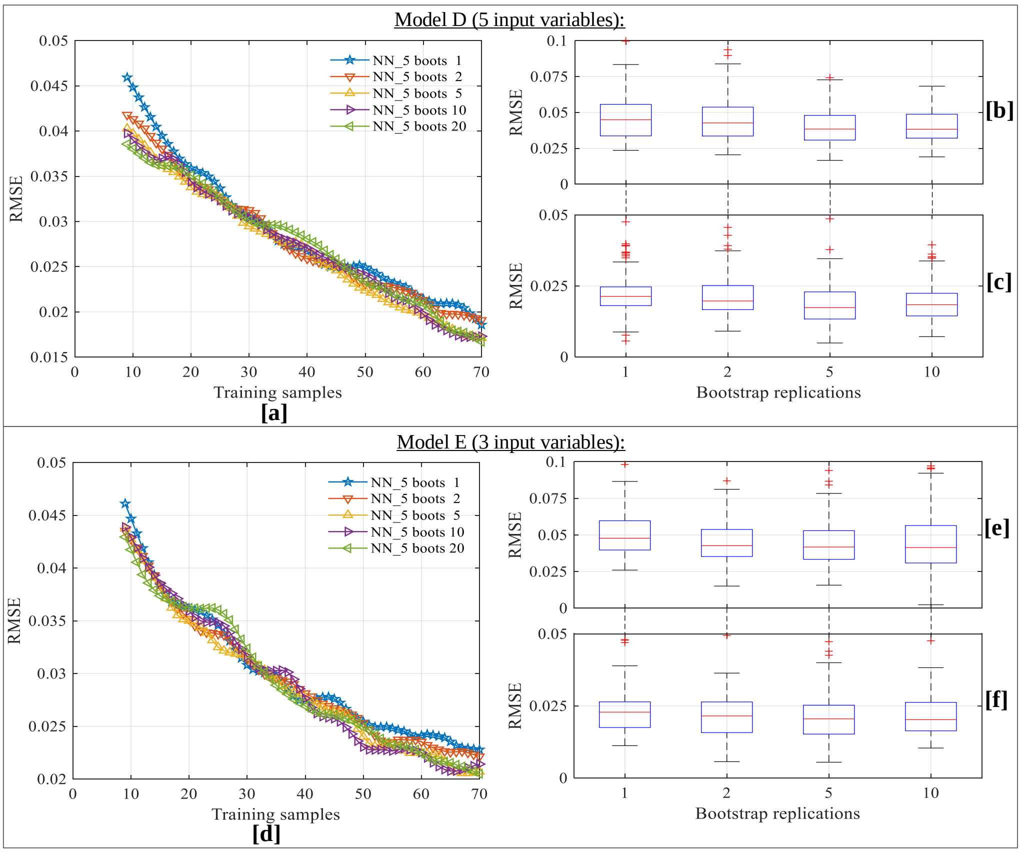


Figure 4.23: Setups D, E: BR influence on NN; RMSE validation: average performances [a], [d], setup D NN_5 performance variation for training samples 10 [b] and 70 [c], setup E NN_5 performance variation for training samples 10 [e] and 70 [f]

For both setups, a lower average performance can be seen for BR in the case of very small (~<15) and large (~>50) training samples. For training samples of the middle range (20-50), no clear improvement is recognisable. The boxplots representing the performance variations of setups D and E indicate that estimation improvements are possible by bootstrap modifications for the trained NN_5 models. When focusing on the impact of the number of BR, five replications indicate the highest estimation accuracy. BR higher than five, do not indicate a further performance improvement.

Discussion of PMD training set manipulation through BR

The applied bootstrap modification technique demonstrated diverse results for the regression algorithms. For MLR / SLR and PLSR, bootstrap does not improve the performance. For the more complex NN algorithm, a performance improvement could be seen for very small and large training samples with BR. These findings correspond to results presented by Andrijić et al. [202] where the authors showed that BR did not improve linear PMD models, whereas NNs resulted in increased accuracy when the training samples where bootstrapped. Similar outcomes for NNs in combination with bootstrap were reported by Napoli and Xibilia [170], Soares [164], and Tsai and Li [195] as presented in the literature section 2.3.3. The increased performance of NNs with BR is attributable to the higher number of training samples. Since NNs show the highest complexity and flexibility out of the three tested algorithms, the increased number of training samples lowers the risk of overfitting and therefore increases the PMD estimation performance in the small data scenarios [170].

An additional study for the linear algorithms MLR and PLSR revealed that the dataset complexity does not influence the outcome of training sample replications with bootstrap: for the more complex datasets D and E, no performance improvement was detectable for MLR and PLSR with bootstrap replicated training samples. To demonstrate this result, Figure 4.24 depicts the bootstrapped performance of MLR for setup E.

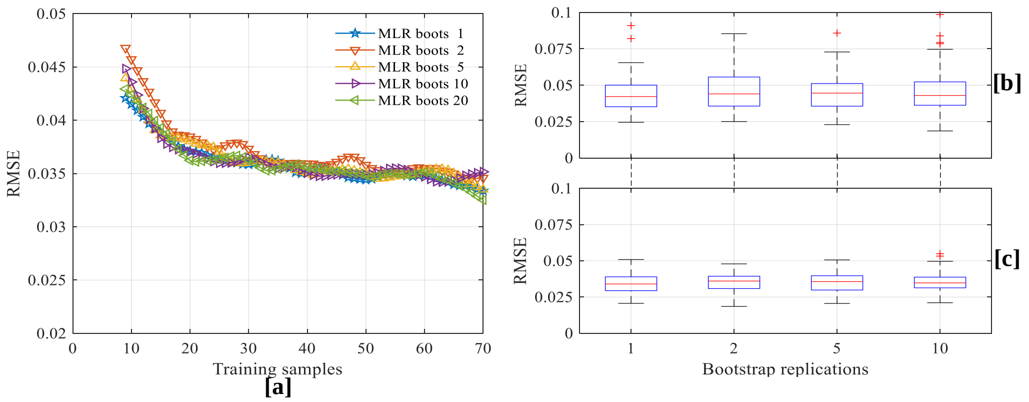


Figure 4.24: Setup E: BR influence on MLR for reduced input variables (3 variables): RMSE average performance [a], MLR performance variation for training samples 10 [b] and 70 [c]

4.3.5 PMD training set manipulation through artificial noise injection

As the second manipulation technique, the training datasets were changed by injecting artificial noise with zero mean to the training samples. Various noise levels were added with fixed Gaussian variances, σ^2, from 0.01 to 0.1, the noise was added independently of the signal amplitude. Additionally, the location of the noise injection was alternated. The noise addition was executed within the MATLAB code by generating normally distributed random numbers that showed a certain variance σ^2 and a zero mean. These values were then added to the training samples, according to the selected noise adding location. Similarly to the above performance studies, the PMD estimation accuracy was analysed in terms of average (median) performance from 100 repetitions and performance variation after the training samples were modified by ANI. The training samples were varied between 9 and 70. For validation, 10 random samples were applied that were not used in the training sample set nor affected by the ANI.

Setups A, B, and C

For setups A, B, and C, the noise was added to the linear regression algorithms with a reduced set of input variables that resulted in the highest performance according to sections 4.3.2 and 4.3.3. These were the same base models that have been applied in the previous manipulation through bootstrap. The results of the noise modifications are displayed in Figure 4.25. In this figure, [a], [b], and [c] display the RMSE average performance and performance variation (for 10 and 40 training samples) for setup A, [d], [e], and [f] for setup B, and [g], [h], and [i] for setup C. In these plots, variance 0 (σ^2= 0) represents the performances of the algorithms without any modification. Figure D.22 in the appendix presents the corresponding results for the CC.

For all three setups, the location of the artificial noise addition was varied and the influence of three different locations was studied: Noise added to the input variables, noise added to the target variable, and noise added to the input and target variables. For the three setups, noise addition to the input variables resulted in the lowest estimation error. Thus, Figure 4.25 depicts the outcomes for noise addition to the input variables. To compare the influence of the noise addition to the two other locations, Figure D.21 in the appendix presents the results exemplarily for setup A.

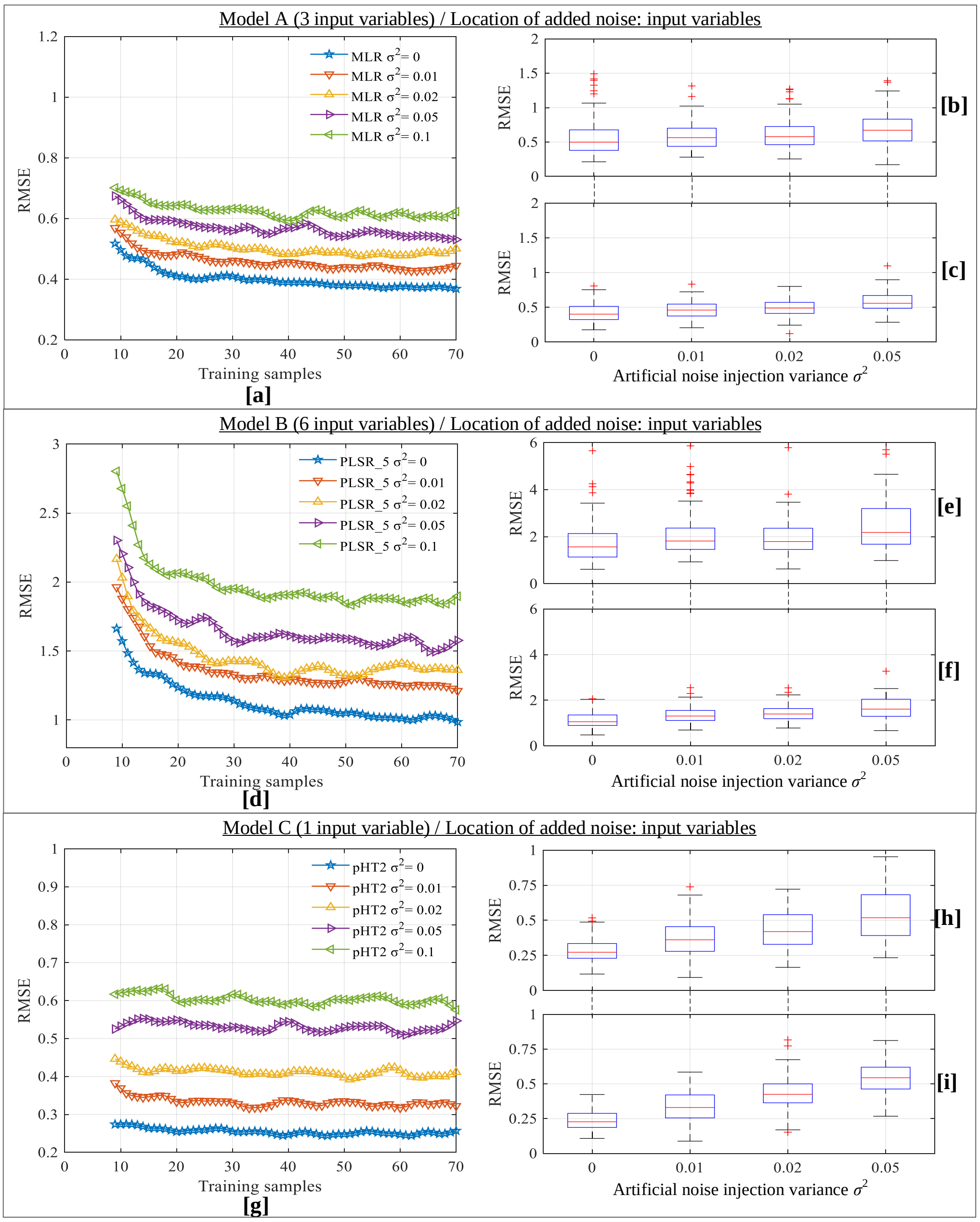


Figure 4.25: Setups A, B, C: ANI influence on MLR, PLSR, and SLR; RMSE validation: average performances [a], [d], [g], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup B PLSR_5 performance variation for training samples 10 [e] and 40 [f], setup C SLR performance variation for training samples 10 [h] and 40 [i]

As depicted in Figure 4.25, the addition of artificial noise with different variances does not improve the average performance nor the performance variation of the three models compared to the training samples that were not manipulated by noise addition (σ^2= 0). For the entire training sample range, the accuracy difference between noise-manipulated and non-manipulated linear models decreases distinctly with growing noise variance. When the artificial noise was added to the target or input and target variables, e.g. displayed in Figure D.21 in the appendix for setup A, the results showed worse performances than compared to the outcomes presented in Figure 4.25.

Setups D and E

Artificial noise was added to the training datasets of the NNs with the highest accuracy from setup D and E. Following from the results in sections 4.3.2 and 4.3.3, the highest NN performance for setup D was obtained with an NN_5 model trained with all five input variables and an NN_5 model with the reduced set of three input variables for setup E. Similar to the artificial noise addition for setups A, B, and C, the analysis of different noise adding locations for setups D and E revealed that the highest accuracy resulted when noise was added to the input variables. A comparison for the influence of noise adding location is presented in the appendix for setup D in Figure D.23. The outcomes for the noise addition to the input variables for setups D and E are displayed in Figure 4.26. In this figure, [a], [b], and [c] display the RMSE average performance and performance variation (for 10 and 40 training samples) for setup D and [d], [e], and [f] for setup E. Variance 0 (σ^2= 0) represents the performance of the models without any modification. Figure D.24 in the appendix displays the corresponding results for the CC.

The results for setup D and E demonstrate that the added noise leads to lower average performances and performance variations than compared to the training samples that were not manipulated by noise addition (σ^2=0). The estimation difference between noise manipulated and non-manipulated NNs decreases distinctly with growing noise variance and with increasing training samples.

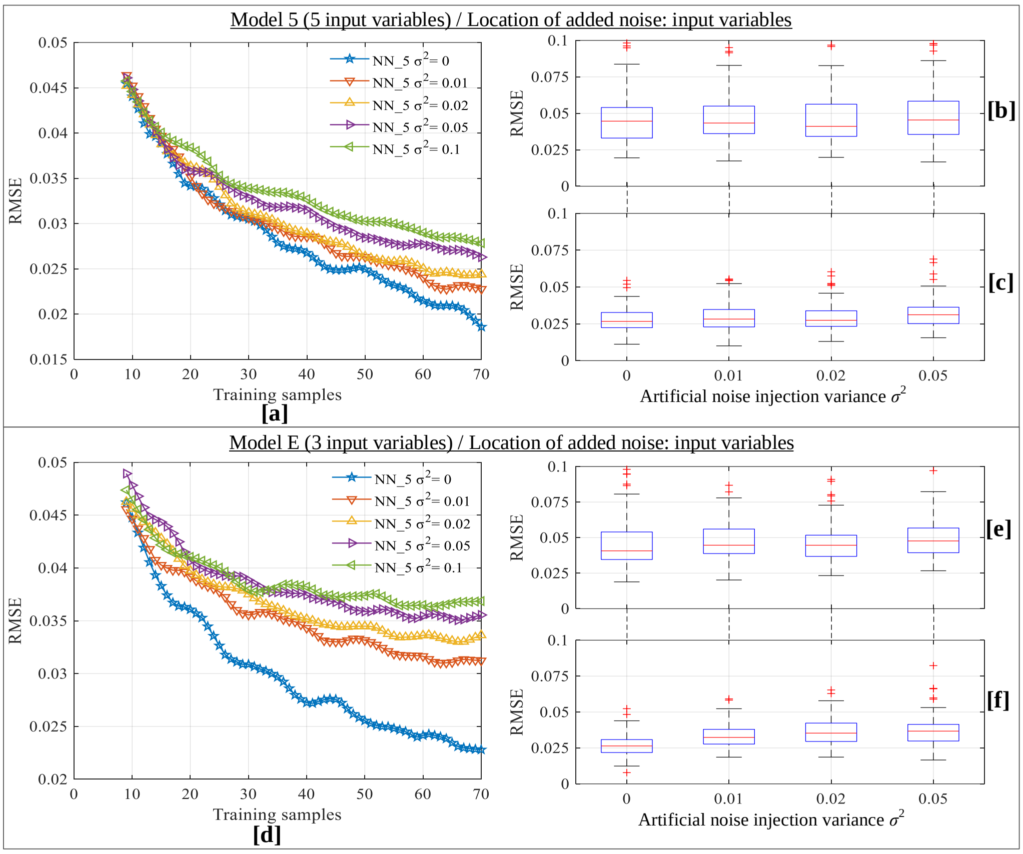


Figure 4.26: Setups D, E: ANI influence on NN; RMSE validation: average performances [a], [d], setup D NN_5 performance variation for training samples 10 [b] and 40 [c], setup E NN_5 performance variation for training samples 10 [e] and 40 [f]

Discussion of PMD training set manipulation through ANI

The results for the ANI of the training samples demonstrate that the addition of artificial noise did not improve the estimation accuracy for the linear algorithms MLR / SLR and PLSR nor the non-linear NNs. The dataset complexity did not influence this result. For all setups, the noise added to the input variables showed higher estimation accuracies than the addition to the target or input and target variables. The decreased performance through noise addition does not correspond to various literature recommendations, where noise addition increased model performances, i.e. reported by Soares [164], Fortuna [171], and Caponetto [206]. For the optimum location of noise addition, the literature review in section 2.3.3 revealed that diverse locations are suggested in different publications [164], [170], [171], [202].

For the setups D and E, the accuracy difference between noise manipulated and non-manipulated NNs increased with higher training sample numbers. This behaviour can be explained by the high complexity / flexibility of the NNs and the risk of overfitting:

When the training samples are very small, the additional noise can reduce the overfitting during the training phase. As the training sample size increases, the NN model can generalise better from training data to unseen data and the additional noise prevents higher performances. This observation leads to the hypothesis that if the noise levels in the original datasets of setups D and E would be higher than the ones present, the artificial noise manipulation for NNs might lead to higher estimation accuracies compared to NNs trained with non-manipulated training samples. To test this hypothesis, an additional study focused on the performance changes for ANI when higher noise levels would be present in the original datasets of setups A and E. To increase the noise of the original datasets, artificial noise was added to input and target variables. Similarly to the noise addition for the training sets, this noise showed a fixed Gaussian variance, σ^2, of 0.015 and was added independently of the signal amplitude. After this step, the PMD models were trained with linear (MLR for setup A) and non-linear (NN for setup E) algorithms following the same artificial noise manipulation of the training set as described above. Figure 4.27 depicts the outcomes.

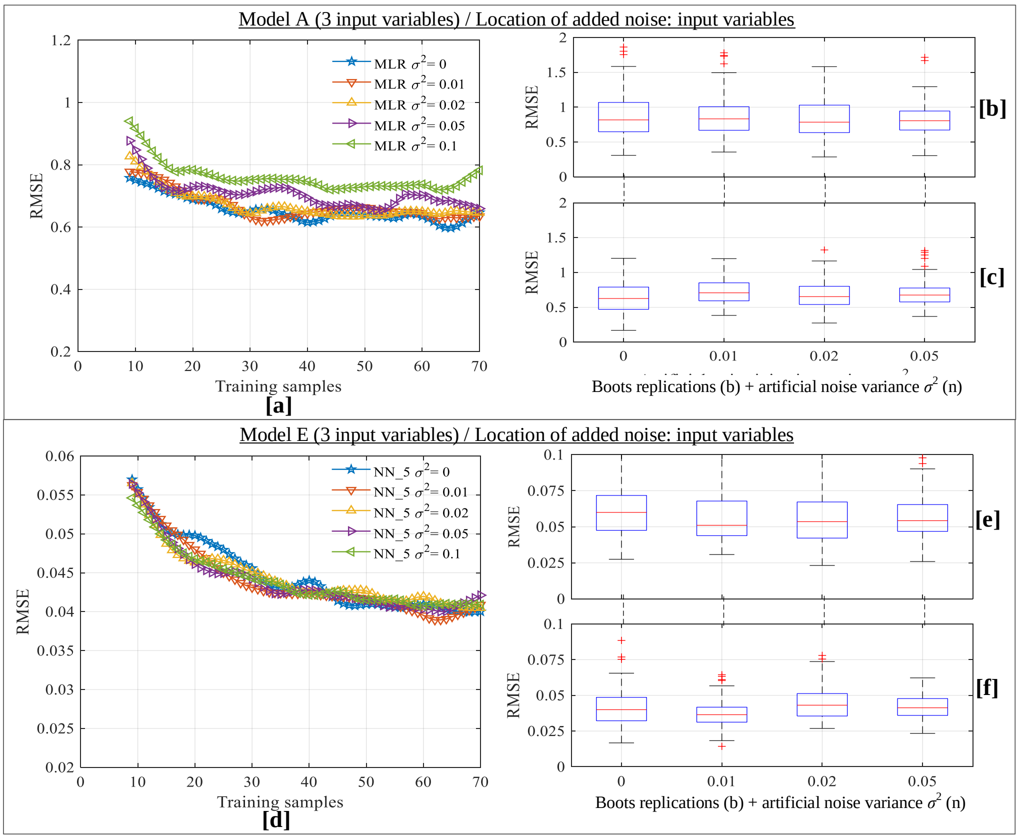


Figure 4.27: Setup A, E: ANI influence when higher noise levels present in original datasets: RMSE average performances [a], [c], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup E NN_5 performance variation for training samples 10 [e] and 40 [f]

For the MLR algorithm of setup A, the artificial noise manipulation does not lead to increased accuracy for the analysed case of higher noise levels in the original dataset. Nonetheless, the average performance difference between no training sample manipulation and the manipulation with increasing noise variances is smaller than compared to the noise manipulations in the original dataset (Figure 4.25 [a]). In the lower part of Figure 4.27, the results for the NN of setup E are depicted. Here, the noise manipulation of the training samples shows similar, in some parts better, outcomes compared to the non-manipulated samples ($\sigma^2 = 0$). In comparison with the original dataset of Figure 4.26 [d], an improved performance for noise manipulated training samples of the NN can be seen when the noise level is increased in the original dataset.

Taken together, the study of the higher noise levels in the original datasets of setups A and E suggests that the impact of ANI is dependent on the noise level in the original dataset. This would explain why the described noise addition in the literature, i.e. in [164], [171], [206], has led to positive impacts while the noise addition in this research did not lead to a PMD performance improvement. However, a more in-depth analysis of the relation between dataset noise levels and the influence of ANI is needed to prove this hypothesis.

Another aspect that can be observed from the results of the ANI is the misleading accuracy given by the CC for setup C. In Figure D.22 [g] in the appendix, the average CC performances for different ANIs to the SLR training set show similar results, independent of the noise variance. However, the RMSE performances for the same graph layout in Figure 4.25 [g] demonstrate that the average performance is continuously decreasing with increasing noise variance. For this case, the correlation between the estimated variable by the SLR and the target variable (expressed by the CC) does not change when noise is artificially added to the input variable. Though, the estimation accuracy is changing as the RMSE confirms. This case shows that the CC can be a misleading parameter of accuracy for potential PMD users, as already reported Maciel et al. [172].

4.3.6 PMD training set manipulation through bootstrap and artificial noise injection

As the third and last training set manipulation technique, bootstrap and ANI were applied in combination. Based on the outcomes of the two previous sections, the small datasets were first replicated by bootstrap up to 20 times before zero-mean noise was added to the input variables with variances, σ^2, between 0.01 and 0.1. The noise was added independently from the signal amplitude. The training samples were varied between 9 and 70. The PMD estimation accuracy was studied in terms of average (median) performance from 100 repetitions and performance variation. For validation, 10 random samples were applied that were not used in the training sample set nor affected by the BR or ANI.

Setups A, B, and C

For setups A, B, and C, the bootstrap and noise manipulation was applied to the linear regression algorithms with a reduced set of input variables that resulted in the highest performance according to sections 4.3.2 and 4.3.3. These were the same base models that were applied in the previous single manipulation through BR and ANI. The results for the combined training set manipulation through BR and ANI are presented in Figure 4.28. In this figure, [a], [b], and [c] display the RMSE average performance and performance variation (for 10 and 40 training samples) for setup A, [d], [e], and [f] for setup B, and [g], [h], and [i] for setup C. In these graphs, boots 1 together with variance 0 (σ^2=0) represent the performances of the algorithms without any modification. Figure D.25 in the appendix presents the corresponding results for the CC.

As the graphs of Figure 4.28 reveal, no performance improvement is observable for the three setups through the combined manipulation with BR and noise injection: For all three linear algorithms (MLR, PLSR, and SLR) the average performance, as well as the performance variation of the non-modified training samples, display the highest accuracies. The applied noise variance has a higher influence than the number of boots replications: The largest applied noise variance of σ^2=0.1 leads to poorer performances than the smaller variance of σ^2=0.02. The magnitude of the decrease in accuracy with increasing noise variance is similar to the training set modification through noise injection on its own, as displayed in the previous section. For MLR and PLSR, slightly better performance is visible for 10 BR compared to two BR when the same noise variance is applied.

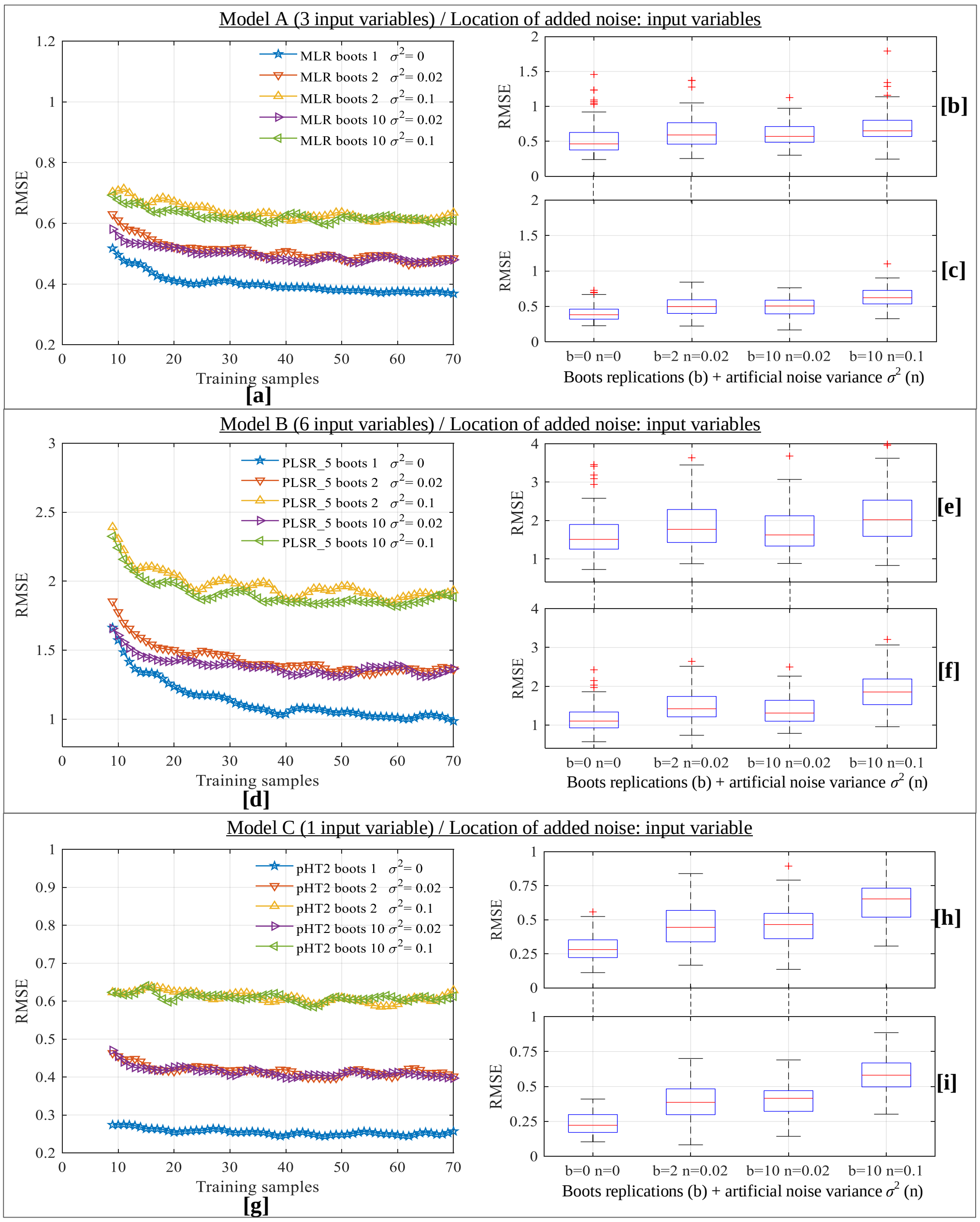


Figure 4.28: Setups A, B, C: BR and ANI influence on MLR, PLSR, and SLR; RMSE validation: average performances [a], [d], [g], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup B PLSR_5 performance variation for training samples 10 [e] and 40 [f], setup C SLR performance variation for training samples 10 [h] and 40 [i]

Setups D and E

For setups D and E, the influence of BR together with ANI was analysed for the NNs that resulted in the highest accuracy from sections 4.3.2 and 4.3.3: for setup D an NN_5 model trained with all five input variables and for setup E an NN_5 model with the reduced set of three input variables. These were the same base models that have been applied in the previous single manipulation through BR and ANI. In Figure 4.29, [a], [b], and [c] display the RMSE average performance and performance variation (for 10 and 40 training samples) for setup D and [d], [e], and [f] for setup E. In these graphs, boots 1 together with variance 0 (σ^2=0) represents the performance of the algorithm without any modification. Figure D.26 in the appendix presents the corresponding results for the CC.

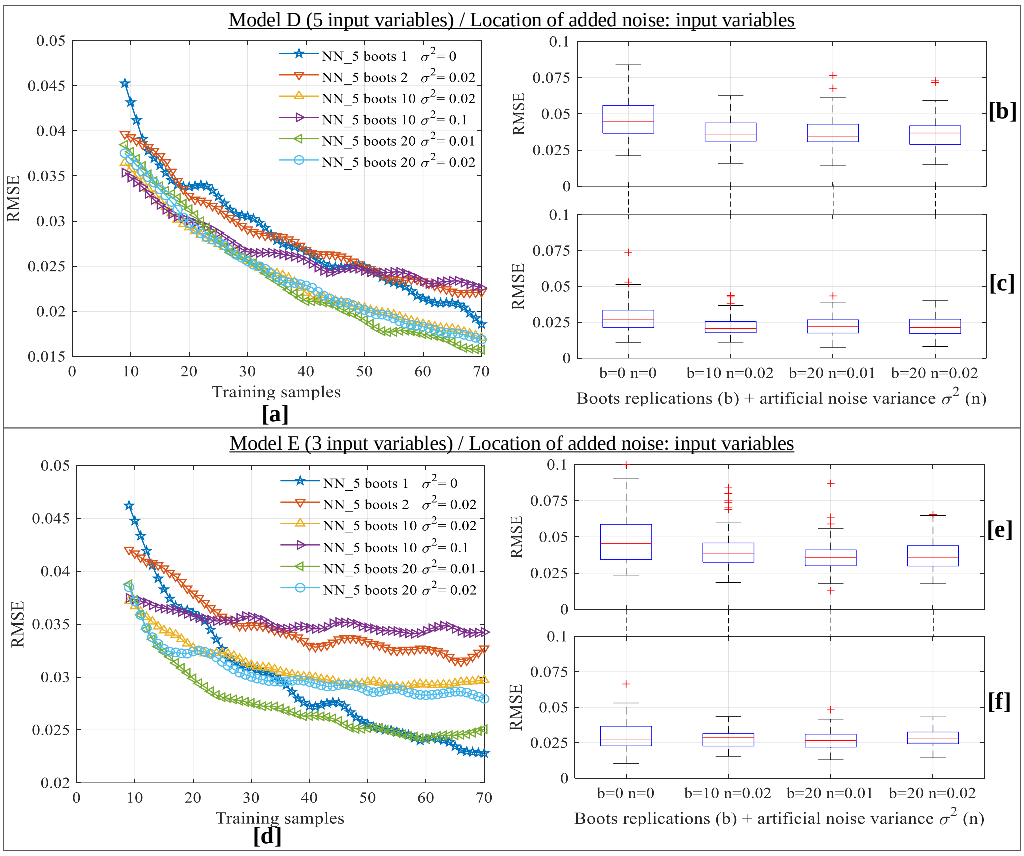


Figure 4.29: Setups D, E: BR and ANI influence on NN; RMSE validation: average performances [a], [d], setup D NN_5 performance variation for training samples 10 [b] and 40 [c], setup E NN_5 performance variation for training samples 10 [e] and 40 [f]

For both setups, improved performance can be seen for the manipulation of BR together with ANI. The results for setup D reveal that for very small training samples, noise injections with a high variance together with high numbers of BR result in better performances. When the training sample size exceeds 20, smaller noise variances together with higher numbers of BR demonstrated the best performance. Throughout the entire range of training samples, the average performance of the NNs that were modified by high numbers of BR and small noise variances resulted in better performances than the NN that was trained without any modification. The outcomes for setup E differ slightly to the results from setup D: Only NNs modified by small noise variances in combination with larger numbers of BR provide higher accuracies than the NN that was trained without any modification.

As seen in [b], [c], [e], [f] of Figure 4.29, the NN training set modification through higher numbers of BR in combination with lower noise variances decreases the NN performance variations for both setups compared to the non-modified training samples.

Discussion of PMD training set manipulation through BR and ANI

The combined training set modification through BR and ANI did not improve the performance of the linear algorithms of setups A, B, and C. In contrast to the findings for the linear algorithms, the estimation accuracy for the NNs of setups D and E increased through the BR and ANI modification. Small noise variances combined with higher numbers of BR led to the smallest estimation errors throughout the entire training sample range. For the analysed NNs of setups D and E, the modification with BR and ANI resulted in higher accuracies than compared to the modification by BR on its own (shown in section 4.3.4). This outcome suggests that for training sample manipulations of NNs, BR in combination with ANI are the first choice when dealing with small datasets.

Similarly to the discussion in the previous section, the impact of BR in combination with ANI was additionally analysed for the assumption that higher noise levels would be present in the original datasets of setups A and E. As before, artificial noise was added to the input and target variables by a fixed Gaussian variance, σ^2, of 0.015. Subsequently, the PMD models were trained following the procedure outlined above by the manipulation of the training samples with BR and ANI. Figure 4.30 presents the outcome for MLR (setup A) and NNs (setup E).

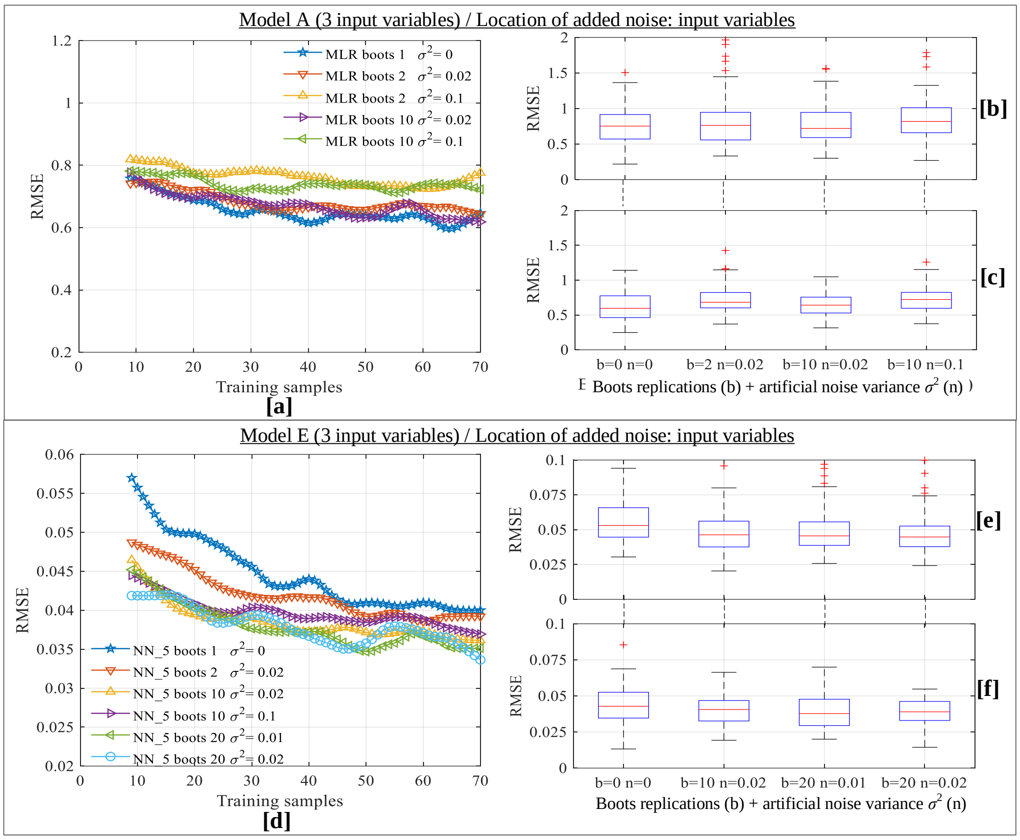


Figure 4.30: Setup A, E: BR and ANI influence when higher noise levels present in original dataset: RMSE average performances [a], [d], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup E NN_5 performance variation for training samples 10 [e] and 40 [f]

Figure 4.30 demonstrates that for the scenario of increased noise levels in the original dataset, no improved performance is observable for MLR of setup A when the training samples are manipulated by BR and ANI. For setup E, the performance enhances throughout the entire range of training samples when bootstrap and artificial noise manipulation are applied for increased noise levels in the original dataset. This suggests that independent of the noise levels in the original datasets, the performance of NNs can be increased by modifying training samples with BR and ANI.

Furthermore, the CC results for the average performances of BR and ANI manipulations for setup C, displayed in Figure D.25, show a misleading information of consistent accuracies regardless of the BR and noise variances. As mentioned in the previous section 4.3.5, this information does not correspond with the RMSE result and can lead to incorrect confidence by potential PMD developers.

4.4 Summary

As the use of PMDs is currently limited in industrial processes [140], the research in this chapter provides a comprehensive PMD development methodology for small dataset scenarios. Based on five datasets with varying regression complexity, the estimation accuracy of three regression algorithms was studied, including the impact of training sample manipulation with bootstrap replications (BR) and artificial noise injection (ANI).

Inspired by the layout of the analysed steam system in the previous chapter, an experimental water loop rig was built to generate datasets for PMD development. The integrated measuring devices and flexible design of the experimental rig enabled the collection of process-related data. To further increase the flexibility and data sourcing, a simulation model was developed which represented the flow behaviour of the experimental rig.

Based on the acquired data from the experimental rig, five small datasets were selected with diverse input and target variable relationships and varying regression complexities. In an initial step, the datasets were categorised by complexity measures that were introduced by Lorena et al. [175] and focused on linearity, correlation to the target variable, smoothness of the data distribution, and geometry / topology. Subsequently, the classified datasets were applied for PMD training and validation. According to the most popular PMD algorithms [149], the performance of two linear (MLR and PLSR) and one non-linear (NNs) algorithm were studied for each of the five small datasets. For an increasing number of training samples from 9 to 70, the average performance from 100 training and validation repetitions and the variation of the performance throughout these repetitions, demonstrated how the algorithms cope with varying training sample sizes and dataset complexities. Furthermore, the training samples of the five datasets were manipulated by BR and ANI to evaluate the influence on the PMD estimation error for the three algorithms.

To conclude the findings of the PMD study with the five datasets (also referred as setups A-E according to the experimental rig layouts), Table 4.14 displays the main outcomes in terms of algorithm and training sample manipulation suitability.

Table 4.14: Summary of PMD performance analysis for setups A-E

		PMD-setup (datasets)				
		A	**B**	**C**	**D**	**E**
Complexity measures	Correlation, linearity, smoothness:	high	high	high (max correlation C1 very high)	medium	low
		Performance study:				
Original t.s.; all input variables	Input variables:	7	8	9	5	4
	MLR:	●	●	◐ (medium + large t.s.)	◐ (small + medium t.s.)	◐ (very small t.s.)
	PLSR:	●	●	●	◐ (small + medium t.s.)	◐ (very small t.s.)
	NN:	○	○	◐ (large t.s.)	◐ (medium + large t.s.)	●
Original t.s.; reduced input variables	Performance improvement:	●	●	●	○	●
	Input variables for most accurate subset:	3	6	1	-	3
	Most accurate algorithm:	MLR	PLSR_5	SLR	NN_5 (all input variables)	NN_5
		t.s. manipulation: applied to best algorithm + best subset of input variables (for setup D: all input variables):				
t.s. manipulation with BR	Performance improvement:	○	○	○	◐ (very small + large t.s.)	◐ (very small + large t.s.)
t.s. manipulation with ANI	Performance improvement:	○	○	○	○	○
t.s. manipulation with BR+ ANI	Performance improvement:	○	○	○	●	●
	Highest accuracy:				many BR + small σ^2	many BR + small σ^2

t.s.: training samples; ●: suitable / improvement; ◐: suitable under certain conditions; ○: not suitable / no improvement

From the research presented in this chapter, the following conclusions can be drawn for the development of PMDs with small datasets, additionally depicted in Figure 4.31:

- The applied complexity measures enabled the classification of the five datasets according to their regression difficulty. Almost all of the 10 measures for the regression complexity areas including correlation, linearity, smoothness, and geometry / topology indicated a clear separation between datasets requiring a simple or more complex regression approach. Nonetheless, for datasets of similar regression complexity, varying performances were observed for the same algorithm. Therefore, further work is needed to analyse the cause of algorithm performance variation and to identify if all aspects of regression complexity are captured by the applied measures.
- Regardless of whether a linear or non-linear algorithm is used, input variable reduction can significantly improve the PMD performance for the analysed scenario of small training datasets. The applied strategy of input variable reduction, consisting of single input variable relevance (correlation to target variable and selectivity ratio) and forward selection method, identified a well-performing subset with little effort. In four of the five setups, algorithms with a smaller number of input variables showed improved performance, especially when very small training samples were used.
- The two linear algorithms MLR and PLSR led to small estimation errors when the complexity measures of a dataset showed high correlation and linearity. The application of pre-modelling modification techniques BR and / or ANI did not show any performance improvement. PLSR should be applied for very small training samples.
- For datasets with high regression complexities, NNs with a small number of hidden neurons outperformed the linear algorithms, except for the scenario of very small training data in which PLSR should be preferred. This result is consistent with several publications that challenge the dogmatic perception of NN requiring a large number of training samples to perform well [187], [195].

  The manipulation of small training datasets with BR improved the NN performances. When BR were applied in combination with ANI, the highest performance enhancement was detectable for NNs. For this combined manipulation, high BR should be applied in combination with small noise variances added to the input variables.

- For the analysed multiple training and validation repetitions per training sample number, the average performance increased and the performance variation decreased for all three algorithms with an increasing number of training samples. Independent of the dataset complexity, NNs showed the highest performance variation for very small training samples, followed by MLR and PLSR.
- For PMD validation, the RMSE should be used against the CC. In two cases, the correlation between the estimated and measured target variable led to misleading validation results.

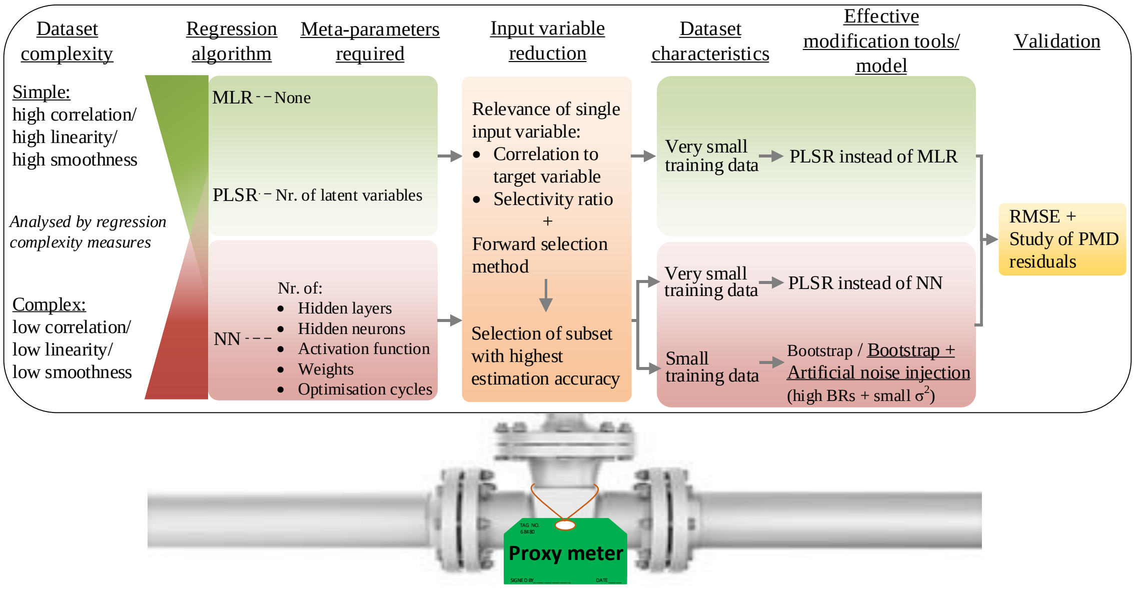


Figure 4.31: Summarising PMD development methodology for small datasets

Chapter 5

Summary and Conclusions

5.1 Summary

For the latest advancement of manufacturing systems, named as ‘Industry 4.0’ or Smart Manufacturing, large amounts of precise process data are required. However, the acquisition of sufficient data represents a challenge for many manufacturing plants, as reported by Kusiak [25]. Developed metering frameworks from academia are often not applicable in real-world manufacturing processes due to monitoring gaps [94]. Therefore, this research focused on the development of effective data gathering tools for complex industrial systems in order to provide a comprehensive process understanding. By considering industrial needs, the design and development of functional resource measurement methodologies and the study of predictive models with a reduced-order characteristic have been addressed in this work to accelerate a holistic data acquisition in industrial environments.

Following from the investigations of two purified water (PW) and one steam system in an industrial case facility, a comprehensive metering strategy for detailed information gathering of technical building services (TBS) has been developed. Although TBS require large shares of resources, holistic analysis methods are not comprehensively addressed in the literature [62]. The novel four phase metering methodology in this research fills this gap by identifying available data sources, abstracting the central process steps, and mapping the resource flows within TBS. In the last phase, the absence of functional online meters is surrogated by a basic proxy metering strategy that enables the approximation of missing parameters by combining related data sources in a regression model.

The gathered process information from the holistic meter approach allowed the development of an automated calculation methodology for the total cost of TBS. By focusing on the interactions of identified resources and further added values, this automated cost index tracks the real value that is embedded in the analysed system.

Based on the potential of basic proxy metering devices (PMDs) for estimating missing process data and the industrial need for simpler implementation strategies [140], the development of comprehensive PMD modelling methodologies has been successfully addressed in the second part of this research. For the common case of small datasets, the regression performance of PMD algorithms for diverse dataset regression complexities has been analysed. As well as the complexity classification of five small datasets from an experimental rig, the estimation accuracies of the most common linear and non-linear PMD algorithms have been studied, including the impact of training sample manipulations with bootstrap and artificial noise injection.

5.2 Conclusions

5.2.1 Resource measurement strategy for technical building services

An industrial case study was undertaken with the aim of developing a detailed metering methodology for TBS. A novel four-phase metering framework was developed following the investigations of two PW and one steam system within a case facility. For each analysed system, the framework identified the required resources, visualised the system internal flow patterns, and revealed further added values that are required for the TBS operation.

Following the highlighted importance of applicable metering strategies for industrial systems from several researchers [26], [27], [33], [45], the developed metering methodology for TBS within this research represents a novel data acquisition approach with adaptability to existing industrial conditions and constraints that has not been previously reported to this extent in the literature.

Based on the information obtained from the application of the TBS metering framework for the two PW systems of the case facility, key performance indicators showed that 48% and 68% of the incoming mains water were converted into PW. The data gathered through the metering framework identified that the significant efficiency difference of the analysed PW systems was mainly due to divergent performances of the reverse osmosis and polishing-ultrafiltration units. For the analysed steam system of the case facility, a fuel to steam efficiency of 72% on average was calculated for the three implemented boilers. Steam losses throughout the distribution lines reduced the amount of received steam at the consumer side by an average of 8%. In addition, the gathered information enabled the identification of possible system optimisations with an easy to adopt characteristic for all three analysed TBS. For example, in both PW systems, significant amounts of resources could be saved if the PW intakes were turned off during temporary production stops.

Following from these results it can be concluded that the strategic measuring methodology developed in this work enabled a holistic resource flow transparency of the investigated TBS, which led to an increased understanding of resource interactions and the disclosure of system optimisation possibilities.

To enable automated data acquisition from offline or broken measuring devices, the fourth phase of the TBS metering framework incorporated a basic PMD developing strategy. Following these modelling steps, three PMDs were developed for the investigated TBS of the case facility. Two of the PMDs estimated the target variable with sufficient accuracy, while the third PMD resulted in poor performances due to unsuitable sampling times and large time offsets between the related measuring devices. In all three models, small datasets resulting from manual recordings were a limitation. These results demonstrate that first-order PMDs can be a useful choice for estimating missing data of industrial systems. However, the application of the three PMDs has also demonstrated that appropriate conditions of the related measuring devices are needed for a successful PMD implementation.

The integration of the gained data in an automated calculation of the total TBS cost revealed the actual value that is embedded in the system by showing the contribution of each cost factor. For the analysed PW system of the case facility, the investigation of the total cost over one year showed that the costs of the purified water were on average 3.4 times higher than the company’s mains water supply cost, with mains water and further added values being the central cost factors. For the steam system, natural gas was by far the main cost factor, followed by electrical energy. Up to now, similar cost models that can be found in the literature [100], [103], [105] rely on manual data acquisition. The developed methodology in this research incorporates for the first time an automated cost calculation based on online meters / PMDs. The case application demonstrated that this automation enables the tracking of the effectiveness of system modifications, as seen in the investigated PW system, where a membrane change in one of the units reduced the PW cost by an average of 9%.

Following from the results of the metering framework applications, it can be concluded that even with basic tools in the form of a strategic measurement methodology; the functionality, flow patterns, and cost factors of complex TBS can be broken down and a holistic data acquisition is possible. For data acquisition extension, even simple structured PMDs infer data gaps of missing parameters with acceptable accuracy and provide a valuable alternative to the installation of an additional physical measuring device when the conditions of the related measuring devices are appropriate.

5.2.2 Proxy measurements in small dataset scenarios

The potential of basic PMDs from the applications in the case facility has motivated the development of comprehensive modelling methodologies to accelerate the use in industry. One key challenge for the currently limited application of PMDs in industrial systems is the academic focus of ever advancing mathematical PMD solutions which require time and cost-intensive development [140]. Furthermore, many PMD modelling datasets from industrial environments are of small size which represents an additional barrier [171]. To overcome these limitations, this work focused on testing and validating different PMD model algorithms and data manipulation techniques for varying regression complexities in order to provide targeted and straightforward PMD development steps for industrial application.

As a foundation for the PMD testing, small datasets with simple regression complexity were successfully acquired from an experimental water loop rig. The data from the implemented measuring devices of the experimental rig provided an effective resource for building a variety of PMD regression models with diverse input and target variable relationships. Further datasets with increased regression complexity were generated by a simulation model that represented the flow behaviour of the experimental rig.

With the recently presented set of regression complexity measures by Lorena et al. [175], five small datasets from the experimental rig were successfully characterised in terms of their regression complexity. Based on this classification, the performance of various PMD configurations has been tested and the following main conclusions were drawn for the development of PMDs on small datasets:

- Within this work, a first-time application of complexity measures for PMD dataset characterisation has proven that insights into the regression complexity of a dataset are possible when the applied measures focus on correlation, linearity, smoothness, and geometry / topology.
- This work presented a comprehensive sensitivity analysis focusing on the impact of input variable reduction for regression models and concludes that the

identification of the most significant variables is essential to increase the PMD performance.

- Following from the novel application of the regression complexity measures, multiple linear regression and Partial Least Squares regression resulted in low estimation errors for regression problems with high correlation and linearity.
- In the age of increased usage of artificial intelligence approaches, this work concludes that in contrary to other publications, Neural Networks can be a suitable choice for small datasets, however, they should only be applied if the analysed regression complexity is high.
- If Neural Networks are applied, this work demonstrated through a comprehensive study of training data manipulation methodologies that training sample replications with bootstrap improve the model performance. Highest Neural Network performance enhancement was detected for bootstrap in combination with small additions of artificial noise to the input variables.

  For linear models, no performance improvement was observed when the training data was manipulated with bootstrap and / or artificial noise injection.
- In conclusion for all analysed models, in particular for Neural Networks; higher numbers of training samples increased the model accuracies and decreased the performance variations.

Following from the investigation outcomes, it can be concluded that unique merit / novelty in this work can be seen in a first-time summary of development methodologies for basic PMDs in dependence of the existing regression complexity. This work has demonstrated that basic models trained on small datasets can enable approximations with high accuracies when the appropriate model and training data manipulation strategy is used. This initiates the question whether the additional development effort for a more complex model approach with a high-end solution is worth the potential model accuracy increase. This decision, in addition to the general complexity demand of an industrial monitoring system, is in the hands of the engineers. This work has proven that basic metering strategies and comprehensive PMD modelling methodologies can enable detailed insights and approximations of complex industrial systems.

5.3 Recommendations for future work

There are several areas where this work could be expanded and developed further. For example:

1. Development of further resource measurement strategies

   There is an opportunity for developing further data acquisition frameworks for manufacturing systems that focus on industrial implementation. Similarly to the presented methodology in this research and corresponding to the requirements outlined in the literature, these frameworks need to be developed in close collaboration between academia and industry to ensure applicability. In addition, the frameworks should aim for a holistic approach that considers multiple resources within a system in order to support the study and characterisation of resource interdependencies. The focus of these frameworks can range from specific manufacturing process steps to larger sub-systems of a manufacturing facility, similar to TBS.

2. PMD performance study with further regression datasets, algorithms, and industrial systems

   While this work shows a novel comparison of PMD performances as a function of regression complexity, limitations exist due to the small number of studied cases and applied algorithms as well as the experimental rig data for PMD modelling. To further simplify the application of PMDs in industry, additional studies of PMD behaviours should be conducted for a variety of regression complexities to expand the presented developing methodologies. This expansion could comprise the analysis of PMD performances of further regression algorithms, i.e. support vector regression as an additional algorithm for datasets with higher regression complexities, and the inclusion of further datasets with a variety of regression complexities for PMD behaviour analysis. As the PMD development methodology is based on data from an experimental rig, the methodology needs to be validated on larger scale industrial setups to prove its applicability.

3. Measurement noise and uncertainty impact on the PMD performance

   When using a metering device, every process measurement is affected by noise. For the input and target variables of a PMD, the noise levels of each online measuring device can have a high impact on the PMD performance. In particular, complex regression models with high flexibility, such as NNs, can be strongly influenced by high levels of measurement noise. This can lead to increasing PMD uncertainty. An in-depth study between measurement noise and PMD performance is needed to further analyse this dependency. In this context, the relationship between dataset noise levels and the artificial noise injection as a training manipulation strategy should be studied in further detail, in addition to the location of artificial noise addition.

4. Further analysis of the dataset regression complexity

   The applied complexity measures from Lorena et al. [175] enabled the classification of datasets into simpler and more complex regression problems. Nonetheless, for datasets with similar regression complexity, varying performances were observed for the same algorithm. Therefore, further investigation is needed to analyse what causes algorithm performance variations when applied to datasets with similar regression complexity and how well the characteristics of regression complexity is represented in the used measures.

5. Preventing information overload of monitoring systems with PMDs

   With the rapid increase of monitoring data from industrial processes, facilities can be confronted with a data overload, affecting the optimal process control. To increase the transparency of a monitoring system, the concept of proxy metering could be adopted. Instead of estimating an unknown process parameter or function as a backup for an online meter, the proxy meter concept could be developed further with the target of reducing / condensing the information output from a complex monitoring system for operators. For example, a PMD could be developed to estimate the state of a process section based on a variety of online meters or the information of these devices could only show up if certain events occur. With this methodology, engineers would gain an effortless tool to summarise the relevant information of their monitoring systems, similar to the approaches in the field of condition monitoring.

References

[1] J. R. Duflou, J. W. Sutherland, D. Dornfeld, C. Herrmann, J. Jeswiet, S. Kara, M. Hauschild, and K. Kellens, “Towards energy and resource efficient manufacturing: A processes and systems approach,” *CIRP Ann. - Manuf. Technol.*, vol. 61, no. 2, pp. 587–609, Jan. 2012, doi: 10.1016/j.cirp.2012.05.002.

[2] R. Stark, G. Seliger, and J. Bonvoisin, *Sustainable Manufacturing*. Springer International Publishing, 2017.

[3] International Energy Agency (IEA), “Tracking Industry,” 2019, Accessed: Jan. 09, 2020. [Online]. Available: https://www.iea.org/reports/tracking-industry-2019.

[4] International Energy Agency (IEA), “Global Energy Review 2020 - The impacts of the Covid-19 crisis on global energy demand and CO2 emissions,” 2020, Accessed: Jun. 08, 2020. [Online]. Available: https://www.iea.org/reports/global-energy-review-2020.

[5] United Nations World Water Assessment Programme (WWAP), “The United Nations World Water Development Report 2014: Water and Energy.” UNESCO, Paris/ France, 2014, Accessed: May 11, 2017. [Online]. Available: http://unesdoc.unesco.org/images/0022/002257/225741E.pdf.

[6] M. Schipper, *Energy-Related Carbon Dioxide Emissions in U.S. Manufacturing*. U.S. Energy Information Administration. DOE/EIA-0573(2005), 2006.

[7] A. D. Jayal, F. Badurdeen, O. W. Dillon, and I. S. Jawahir, “Sustainable manufacturing: Modeling and optimization challenges at the product, process and system levels,” *CIRP J. Manuf. Sci. Technol.*, vol. 2, no. 3, pp. 144–152, 2010, doi: 10.1016/j.cirpj.2010.03.006.

[8] C. Herrmann, C. Schmidt, D. Kurle, S. Blume, and S. Thiede, “Sustainability in manufacturing and factories of the future,” *Int. J. Precis. Eng. Manuf. - Green Technol.*, vol. 1, no. 4, pp. 283–292, 2014, doi: 10.1007/s40684-014-0034-z.

[9] C. Herrmann, S. Thiede, S. Kara, and J. Hesselbach, “Energy oriented simulation of manufacturing systems - Concept and application,” *CIRP Ann. - Manuf. Technol.*, vol. 60, no. 1, pp. 45–48, 2011, doi: 10.1016/j.cirp.2011.03.127.

[10] International Energy Agency (IEA), “Towards a More Energy Efficient Future: Applying Indicators to Enhance Energy Policy,” Paris, France, 2009.

[11] S. H. Bonilla, C. M. V. B. Almeida, B. F. Giannetti, and D. Huisingh, “The roles of cleaner production in the sustainable development of modern societies,” *J. Clean. Prod.*, vol. 18, no. 1, pp. 1–5, 2010, doi: 10.1016/j.jclepro.2009.09.001.

[12] K. K. B. Hon, “Performance and evaluation of manufacturing systems,” *CIRP Annals - Manufacturing Technology*, vol. 54, no. 2. pp. 139–154, 2005, doi: 10.1016/s0007-8506(07)60023-7.

[13] B. S. Linke, D. R. Garcia, A. Kamath, and I. C. Garretson, “Data-driven sustainability in manufacturing: Selected examples,” in *Procedia Manufacturing*, 2019, vol. 33, pp. 602–609, doi: 10.1016/j.promfg.2019.04.075.

[14] J. G. Webster and H. Eren, *Measurement, Instrumentation, and Sensors Handbook: Spatial, Mechanical, Thermal, and Radiation Measurement*, 2nd ed. CRC Press, 2014.

[15] R. Teti, K. Jemielniak, G. O’Donnell, and D. Dornfeld, “Advanced monitoring of machining operations,” *CIRP Ann. - Manuf. Technol.*, vol. 59, no. 2, pp. 717–739, Jan. 2010, doi: 10.1016/j.cirp.2010.05.010.

[16] S. Karnouskos and A. W. Colombo, “Architecting the next generation of service-based SCADA/DCS system of systems,” in *IECON Proceedings (Industrial Electronics Conference)*, 2011, pp. 359–364, doi: 10.1109/IECON.2011.6119279.

[17] P. Stavropoulos, D. Chantzis, C. Doukas, A. Papacharalampopoulos, and G. Chryssolouris, “Monitoring and control of manufacturing processes: A review,” in *Procedia CIRP*, 2013, vol. 8, pp. 421–425, doi: 10.1016/j.procir.2013.06.127.

[18] L. Wang and R. X. Gao, *Condition Monitoring and Control for Intelligent Manufacturing*. Springer London, 2006.

[19] E. E. Doebelin, *Measurement Systems - Application and Design*, 4th ed. McGraw-Hill International Editions, 1990.

[20] F. Tao, Q. Qi, A. Liu, and A. Kusiak, “Data-driven smart manufacturing,” *J. Manuf. Syst.*, vol. 48, pp. 157–169, Jul. 2018, doi: 10.1016/j.jmsy.2018.01.006.

[21] A. M. Gontarz, D. Hampl, L. Weiss, and K. Wegener, “Resource consumption monitoring in manufacturing environments,” *Procedia CIRP*, vol. 26, pp. 264–269, Jan. 2015, doi: 10.1016/j.procir.2014.07.098.

[22] P. O’Donovan, K. Leahy, K. Bruton, and D. T. J. O’Sullivan, “An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities,” *J. Big Data*, vol. 2, no. 1, p. 25, Dec. 2015, doi: 10.1186/s40537-015-0034-z.

[23] K.-D. Thoben, S. Wiesner, and T. Wuest, “‘Industrie 4.0’ and Smart Manufacturing – A Review of Research Issues and Application Examples,” *Int. J. Autom. Technol.*, vol. 11, no. 1, pp. 4–16, Jan. 2017, doi: 10.20965/ijat.2017.p0004.

[24] Smart Manufacturing Leadership Coalition (SMLC), “Implementing 21st Century Smart Manufacturing - Workshop Summary Report,” Washington, DC, USA, 2011.

[25] A. Kusiak, “Smart manufacturing must embrace big data,” *Nature*, vol. 544, no. 7648, pp. 23–25, Apr. 2017, doi: 10.1038/544023a.

[26] K. Bunse, M. Vodicka, P. Schönsleben, M. Brülhart, and F. O. Ernst, “Integrating energy efficiency performance in production management - Gap analysis between industrial needs and scientific literature,” *J. Clean. Prod.*, vol. 19, no. 6–7, pp. 667–679, 2011, doi: 10.1016/j.jclepro.2010.11.011.

[27] N. Bhanot, P. V. Rao, and S. G. Deshmukh, “An integrated approach for analysing the enablers and barriers of sustainable manufacturing,” *J. Clean. Prod.*, vol. 142, pp. 4412–4439, 2017, doi: 10.1016/j.jclepro.2016.11.123.

[28] Bundesministerium für Bildung und Forschung, “Zukunftsbild „Industrie 4.0".” 2013, Accessed: Jan. 15, 2018. [Online]. Available: https://www.bmbf.de/pub/Zukunftsbild_Industrie_4.0.pdf.

[29] H. Kagermann, W. Lukas, and W. Wahlster, “Industrie 4.0 - Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution,” *VDI Nachrichten*, no. 13, 2011.

[30] G. Culot, G. Nassimbeni, G. Orzes, and M. Sartor, “Behind the definition of Industry 4.0: Analysis and open questions,” *Int. J. Prod. Econ.*, vol. 226, p. 107617, Aug. 2020, doi: 10.1016/j.ijpe.2020.107617.

[31] F. Zezulka, P. Marcon, I. Vesely, and O. Sajdl, “Industry 4.0 – An Introduction in the phenomenon,” *IFAC-PapersOnLine*, vol. 49, no. 25, pp. 8–12, 2016, doi: 10.1016/j.ifacol.2016.12.002.

[32] P. K. Paritala, S. Manchikatla, and P. K. D. V. Yarlagadda, “Digital Manufacturing- Applications Past, Current, and Future Trends,” *Procedia Eng.*, vol. 174, pp. 982–991, Jan. 2017, doi: 10.1016/j.proeng.2017.01.250.

[33] A. Kusiak, “Smart manufacturing,” *Int. J. Prod. Res.*, vol. 56, no. 1–2, pp. 508–517, Jul. 2017, doi: 10.1080/00207543.2017.1351644.

[34] P. Zheng, H. Wang, Z. Sang, R. Y. Zhong, Y. Liu, C. Liu, K. Mubarok, S. Yu, and X. Xu, “Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives,” *Front. Mech. Eng.*, vol. 13, no. 2, pp. 137–150, Jun. 2018, doi: 10.1007/s11465-018-0499-5.

[35] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster, “Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Securing the future of German manufacturing industry.” Forschungsunion Germany, 2013.

[36] A. Albers, B. Gladysz, T. Pinner, V. Butenko, and T. Stürmlinger, “Procedure for Defining the System of Objectives in the Initial Phase of an Industry 4.0 Project Focusing on Intelligent Quality Control Systems,” *Procedia CIRP*, vol. 52, pp. 262–267, 2016, doi: 10.1016/j.procir.2016.07.067.

[37] A. Schütze, N. Helwig, and T. Schneider, “Sensors 4.0 - Smart sensors and measurement technology enable Industry 4.0,” *J. Sensors Sens. Syst.*, vol. 7, no. 1, pp. 359–371, May 2018, doi: 10.5194/jsss-7-359-2018.

[38] C. Fuchs, “Industry 4.0: The digital German ideology,” *TripleC*, vol. 16, no. 1, pp. 280–289, Feb. 2018, doi: 10.31269/vol16iss1pp280-289.

[39] D. Spath, O. Ganschar, S. Gerlach, M. Hämmerle, T. Krause, and S. Schlund, “Produktionsarbeit der Zukunft – Industrie 4.0.” Frauenhofer Verlag, Stuttgart, 2013, Accessed: Jan. 15, 2018. [Online]. Available: https://www.iao.fraunhofer.de/lang-de/images/iao-news/produktionsarbeit-der-zukunft.pdf.

[40] A. G. Frank, L. S. Dalenogare, and N. F. Ayala, “Industry 4.0 technologies: Implementation patterns in manufacturing companies,” *Int. J. Prod. Econ.*, vol. 210, pp. 15–26, Apr. 2019, doi: 10.1016/j.ijpe.2019.01.004.

[41] W. Li and S. Kara, “Methodology for Monitoring Manufacturing Environment by Using Wireless Sensor Networks (WSN) and the Internet of Things (IoT),” *Procedia CIRP*, vol. 61, pp. 323–328, Jan. 2017, doi: 10.1016/j.procir.2016.11.182.

[42] N. Bhanot, P. V. Rao, and S. G. Deshmukh, “Enablers and barriers of sustainable manufacturing: Results from a survey of researchers and industry professionals,” in *Procedia CIRP*, Jan. 2015, vol. 29, pp. 562–567, doi: 10.1016/j.procir.2015.01.036.

[43] A. Kusiak, “A four-part plan for smart manufacturing,” *ISE Mag.*, vol. 49, no. 7, pp. 43–47, 2017.

[44] Y. Lu, “Industry 4.0: A survey on technologies, applications and open research issues,” *Journal of Industrial Information Integration*, vol. 6. pp. 1–10, 2017, doi: 10.1016/j.jii.2017.04.005.

[45] A. Trianni, E. Cagno, and S. Farné, “Barriers, drivers and decision-making process for industrial energy efficiency: A broad study among manufacturing small and medium-sized enterprises,” *Appl. Energy*, vol. 162, pp. 1537–1551, Jan. 2016, doi: 10.1016/j.apenergy.2015.02.078.

[46] V. Balaji, P. Venkumar, M. S. Sabitha, S. Vijayalakshmi, and R. M. Rathikaa Sre, “Smart Manufacturing Through Sensor Based Efficiency Monitoring System (SBEMS),” in *Proceedings of the Eighth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016)*, Springer, 2017, pp. 34–43.

[47] eSight Energy Ltd, “eSight Energy [Computer software].” Cambridge, UK, 2017.

[48] I. L. Yen, S. Zhang, F. Bastani, and Y. Zhang, “A Framework for IoT-Based Monitoring and Diagnosis of Manufacturing Systems,” in *Proceedings - 11th IEEE International Symposium on Service-Oriented System Engineering, SOSE 2017*, Apr. 2017, pp. 1–8, doi: 10.1109/SOSE.2017.26.

[49] S. Mittal, M. A. Khan, D. Romero, and T. Wuest, “A critical review of smart manufacturing & Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs),” *Journal of Manufacturing Systems*, vol. 49. Elsevier, pp. 194–214, Oct. 01, 2018, doi: 10.1016/j.jmsy.2018.10.005.

[50] Siemens AG, “SIMATIC PCS neo,” 2019, Accessed: Jan. 28, 2020. [Online]. Available: https://new.siemens.com/global/en/products/automation/distributed-control-system/simatic-pcs-neo.html.

[51] E. O’Driscoll and G. E. O’Donnell, “Industrial power and energy metering a state-of-the-art review,” *J. Clean. Prod.*, vol. 41, pp. 53–64, 2013, doi: 10.1016/j.jclepro.2012.09.046.

[52] H. A. Abbas, “Future SCADA challenges and the promising solution: The agent-based SCADA,” *Int. J. Crit. Infrastructures*, vol. 10, no. 3–4, pp. 307–333, Jan. 2014, doi: 10.1504/IJCIS.2014.066354.

[53] F. Horden, “SCADA and DCS Battle for Globalization,” *Pipeline Gas J.*, vol. 241, 2014.

[54] Honeywell International INC, “Experion Elevate,” 2017, Accessed: Jan. 28, 2020. [Online]. Available: https://www.honeywellprocess.com/en-US/explore/products/control-monitoring-and-safety-systems/integrated-control-and-safety-systems/experion-pks/Pages/experion-elevate.aspx.

[55] ABB, “ABB brings the digital wellhead to the masses with cloud-based SCADA analytics,” 2019, Accessed: Jan. 28, 2020. [Online]. Available: https://new.abb.com/news/detail/17977/abb-brings-the-digital-wellhead-to-the-masses-with-cloud-based-scada-analytics.

[56] Yokogawa Europe, “SCADA Software (FAST/TOOLS),” 2019, Accessed: Jan. 28, 2020. [Online]. Available: https://www.yokogawa.com/eu/solutions/products-platforms/control-system/supervisory-control-and-data-acquisition-scada/fast-tools/.

[57] G. Olsson, *Water and Energy: Threats and Opportunities*, 1st ed. IWA Publishing, 2012.

[58] European Commission- Eurostat, “Statistics in focus: Water use in industry.” ISSN:2314-9647, 2014.

[59] Organisation for Economic Co-operation and Development (OECD), “OECD Environmental Outlook Baseline.” Paris, France, 2007.

[60] D. J. Barrington, A. Prior, and G. Ho, “The role of water auditing in achieving water conservation in the process industry,” *J. Clean. Prod.*, vol. 52, pp. 356–361, 2013, doi: 10.1016/j.jclepro.2013.03.032.

[61] C. Schulze, S. Thiede, B. Thiede, D. Kurle, S. Blume, and C. Herrmann, “Cooling tower management in manufacturing companies: A cyber-physical system approach,” *J. Clean. Prod.*, vol. 211, pp. 428–441, Feb. 2019, doi: 10.1016/j.jclepro.2018.11.184.

[62] S. Thiede, M. Schönemann, D. Kurle, and C. Herrmann, “Multi-level simulation in manufacturing companies: The water-energy nexus case,” *J. Clean. Prod.*, vol. 139, pp. 1118–1127, Dec. 2016, doi: 10.1016/j.jclepro.2016.08.144.

[63] Z. A. Manan and S. R. W. Alwi, “Water pinch analysis evolution towards a holistic approach for water minimization,” *Asia-Pacific J. Chem. Eng.*, vol. 2, no. 6, pp. 544–553, Nov. 2007, doi: 10.1002/apj.99.

[64] J. Hesselbach, C. Herrmann, R. Detzer, L. Martin, S. Thiede, and B. Lüdemann, “Energy Efficiency through Optimised Coordination of Production and Technical Building Services,” in *LCE 2008: 15th CIRP International Conference on Life Cycle Engineering: Conference Proceedings*, 2008, pp. 624–269.

[65] S. Thiede, G. Posselt, and C. Herrmann, “SME appropriate concept for continuously improving the energy and resource efficiency in manufacturing companies,” *CIRP J. Manuf. Sci. Technol.*, vol. 6, no. 3, pp. 204–211, Jan. 2013, doi: 10.1016/j.cirpj.2013.02.006.

[66] S. Mousavi, S. Kara, and B. Kornfeld, “Assessing the Impact of Embodied Water in Manufacturing Systems,” *Procedia CIRP*, vol. 29, pp. 80–85, 2015, doi: 10.1016/j.procir.2015.01.001.

[67] V. T. Penna, S. Martins, and P. Mazzola, “Identification of bacteria in drinking and purified water during the monitoring of a typical water purification system,” *BMC Public Health*, vol. 2, no. 1, pp. 1–11, Dec. 2002, doi: 10.1186/1471-2458-2-13.

[68] Council of Europe, *European Pharmacopoeia*, 6th ed. Strasbourg, France, 2016.

[69] B. Chris and M. Kimber, *Basic Water Treatment*, 5th ed. ICE Publishing, 2013.

[70] M. Sörensen, P. Satish, and J. Weckenmann, “State of the Art Sanitisation of Purified Water (PFW),” *Pharm. Ind.*, vol. 78, no. 7, pp. 1046–1055, 2016.

[71] R. Singh, *Membrane Technology and Engineering for Water Purification - Application, Systems Design and Operation*, 2nd ed. Butterworth-Heinemann, 2015.

[72] S. Brueske, R. Sabouni, C. Zach, and H. Andres, “U.S. Manufacturing Energy Use and Greenhouse Gas Emissions Analysis,” 2012. doi: 10.2172/1055125.

[73] E. Nieuwlaar, A. L. Roes, and M. K. Patel, “Final Energy Requirements of Steam for Use in Environmental Life Cycle Assessment,” *J. Ind. Ecol.*, vol. 20, no. 4, pp. 828–836, Aug. 2016, doi: 10.1111/jiec.12300.

[74] D. Einstein, E. Worrell, and M. Khrushch, “Steam systems in industry: Energy use and energy efficiency improvement potentials,” in *Proceedings of the 2001 ACEEE Summer Study on Energy Efficiency in Industry*, 2001.

[75] European Commission Directorate - General for Energy, “Mapping and analyses of the current and future (2020 - 2030) heating/cooling fuel deployment (fossil/renewables) - Work package 2: Assessment of the technologies for the year 2012,” 2016, Accessed: Mar. 12, 2019. [Online]. Available: https://ec.europa.eu/energy/sites/ener/files/documents/Report WP2.pdf.

[76] P. Regulagadda, I. Dincer, and G. F. Naterer, “Exergy analysis of a thermal power plant with measured boiler and turbine losses,” *Appl. Therm. Eng.*, vol. 30, no. 8–9, pp. 970–976, Jun. 2010, doi: 10.1016/j.applthermaleng.2010.01.008.

[77] M. C. Barma, R. Saidur, S. M. A. Rahman, A. Allouhi, B. A. Akash, and S. M. Sait, “A review on boilers energy use, energy savings, and emissions reductions,” *Renew. Sustain. Energy Rev.*, vol. 79, pp. 970–983, Nov. 2017, doi: 10.1016/j.rser.2017.05.187.

[78] M. Bojić and S. Dragićević, “Optimization of steam boiler design,” *Proc. Inst. Mech. Eng. Part A J. Power Energy*, vol. 220, no. 6, pp. 629–634, Jan. 2006, doi: 10.1243/09576509JPE203.

[79] J. Bujak, “Minimizing energy losses in steam systems for potato starch production,” *J. Clean. Prod.*, vol. 17, no. 16, pp. 1453–1464, Nov. 2009, doi: 10.1016/j.jclepro.2009.06.013.

[80] A. Hasanbeigi, G. Harrell, B. Schreck, and P. Monga, “Moving beyond equipment and to systems optimization: Techno-economic analysis of energy efficiency potentials in industrial steam systems in China,” *J. Clean. Prod.*, vol. 120, pp. 53–63, May 2016, doi: 10.1016/j.jclepro.2016.02.023.

[81] M. Siddhartha Bhatt, “Energy audit case studies I - Steam systems,” *Appl. Therm. Eng.*, vol. 20, no. 3, pp. 285–296, Feb. 2000, doi: 10.1016/s1359-4311(99)00020-4.

[82] P. Therkelsen and A. McKane, “Implementation and rejection of industrial steam system energy efficiency measures,” *Energy Policy*, vol. 57, pp. 318–328, Jun. 2013, doi: 10.1016/j.enpol.2013.02.003.

[83] E. Masanet and M. E. Walker, “Energy-water efficiency and U.S. industrial steam,” *AIChE J.*, vol. 59, no. 7, pp. 2268–2274, Jul. 2013, doi: 10.1002/aic.14148.

[84] M. E. Walker, Z. Lv, and E. Masanet, “Industrial steam systems and the energy-water nexus,” *Environ. Sci. Technol.*, vol. 47, no. 22, pp. 13060–13067, Nov. 2013, doi: 10.1021/es403715z.

[85] U.S. Department of Energy, “The Water-Energy Nexus: Challenges and Opportunities-Overview and Summary,” 2014. doi: 10.1007/s13398-014-0173-7.2.

[86] A. M. Hamiche, A. B. Stambouli, and S. Flazi, “A review of the water-energy nexus,” *Renewable and Sustainable Energy Reviews*, vol. 65. Pergamon, pp. 319–331, Nov. 01, 2016, doi: 10.1016/j.rser.2016.07.020.

[87] M. Sachidananda, “A framework for modelling and reduction of water usage in the manufacturing industry,” PhD thesis, Loughborough University, UK, 2013.

[88] P. H. Gleick, *Water in Crisis: A Guide to the World’s Fresh Water Resources*, 1st ed. Oxford University Press, 1993.

[89] S. Thiede, D. Kurle, and C. Herrmann, “The water–energy nexus in manufacturing systems: Framework and systematic improvement approach,” *CIRP Ann. - Manuf. Technol.*, vol. 66, no. 1, pp. 49–52, 2017, doi: 10.1016/j.cirp.2017.04.108.

[90] V. Novotny, “Water and energy link in the cities of the future - Achieving net zero carbon and pollution emissions footprint,” *Water Sci. Technol.*, vol. 63, no. 1, pp. 184–190, Jan. 2011, doi: 10.2166/wst.2011.031.

[91] I. Schlei-Peters, D. Kurle, M. G. Wichmann, S. Thiede, C. Herrmann, and T. S. Spengler, “Assessing combined water-energy-efficiency measures in the automotive industry,” *Procedia CIRP*, vol. 29, pp. 50–55, 2015, doi: 10.1016/j.procir.2015.02.013.

[92] S. Mousavi, S. Kara, and B. Kornfeld, “A hierarchical framework for concurrent assessment of energy and water efficiency in manufacturing systems,” *J. Clean. Prod.*, vol. 133, pp. 88–98, 2016, doi: 10.1016/j.jclepro.2016.05.074.

[93] S. Mousavi, S. Thiede, W. Li, S. Kara, and C. Herrmann, “An integrated approach for improving energy efficiency of manufacturing process chains,” *Int. J. Sustain. Eng.*, vol. 9, no. 1, pp. 11–24, Jan. 2016, doi: 10.1080/19397038.2014.1001470.

[94] M. Sachidananda and S. Rahimifard, “Reduction of Water Consumption within Manufacturing Applications,” in *Leveraging Technology for a Sustainable World*, Springer Berlin Heidelberg, 2012, pp. 455–460.

[95] J. Williams, S. Bouzarovski, and E. Swyngedouw, “Politicising the nexus: Nexus technologies, urban circulation, and the coproduction of water-energy,” *Nexus Netw. Think Piece Ser.*, vol. 1, pp. 1–27, 2014.

[96] N. Castree *et al.*, “Changing the intellectual climate,” *Nat. Clim. Chang.*, vol. 4, no. 9, pp. 763–768, Sep. 2014, doi: 10.1038/nclimate2339.

[97] R. Cairns and A. Krzywoszynska, “Anatomy of a buzzword: The emergence of ‘the water-energy-food nexus’ in UK natural resource debates,” *Environ. Sci. Policy,* vol. 64, pp. 164–170, Oct. 2016, doi: 10.1016/j.envsci.2016.07.007.

[98] S. Parry and J. Murphy, “Problematizing interactions between social science and public policy,” *Crit. Policy Stud.*, vol. 9, no. 1, pp. 97–107, Jan. 2015, doi: 10.1080/19460171.2014.964277.

[99] D. Kurle, C. Herrmann, and S. Thiede, “Unlocking water efficiency improvements in manufacturing — From approach to tool support,” *CIRP J. Manuf. Sci. Technol.*, vol. 19, pp. 7–18, 2017, doi: 10.1016/j.cirpj.2017.02.004.

[100] B. P. Walsh, K. Bruton, and D. T. J. O’Sullivan, “The true value of water: A case-study in manufacturing process water-management,” *J. Clean. Prod.*, vol. 141, pp. 551–567, 2016, doi: 10.1016/j.jclepro.2016.09.106.

[101] B. P. Walsh, S. N. Murray, and D. T. J. O’Sullivan, “The water energy nexus, an ISO50001 water case study and the need for a water value system,” *Water Resour. Ind.*, vol. 10, pp. 15–28, 2015, doi: 10.1016/j.wri.2015.02.001.

[102] B. P. Walsh, D. O. Cusack, and D. T. J. O’Sullivan, “An industrial water management value system framework development,” *Sustain. Prod. Consum.*, vol. 5, pp. 82–93, 2016, doi: 10.1016/j.spc.2015.11.004.

[103] C. Leusder, “Assessing the impacts of water on industry-the case of Asia Pacific Breweries (APB),” *Water Pract. Technol.*, vol. 11, no. 2, pp. 245–259, 2016, doi: 10.2166/wpt.2016.014.

[104] M. Yang and R. K. Dixon, “Investing in efficient industrial boiler systems in China and Vietnam,” *Energy Policy*, vol. 40, no. 1, pp. 432–437, Jan. 2012, doi: 10.1016/j.enpol.2011.10.030.

[105] US Department of Energy - Energy Efficiency and and R. Energy, “Industrial Technologies Program: How to calculate the true cost of steam.” Industrial Technologies Program, Washington, DC, USA, 2003.

[106] C. Dascalu, C. Caraiani, C. Iuliana Lungu, F. Colceag, and G. Raluca Guse, “The externalities in social environmental accounting,” *Int. J. Account. Inf. Manag.*, vol. 18, no. 1, pp. 19–30, Mar. 2010, doi: 10.1108/18347641011023252.

[107] A. P. Hubbard, R.G. O’Brien, *Essentials of Economics*, 1st ed. Prentice Hall, Inc., 2006.

[108] K. G. Shapiro, “Incorporating costs in LCA,” *Int. J. Life Cycle Assess.*, vol. 6, no. 2, pp. 121–123, 2001, doi: 10.1007/BF02977850.

[109] P. Limprayoon and F. Y. Phillips, “Corporate vs. social attitudes toward environmental externalities,” *Int. J. Glob. Environ. Issues*, vol. 11, no. 2, pp. 109–138, 2011, doi: 10.1504/ijgenvi.2011.043506.

[110] B. Linke, Y. C. Huang, and D. Dornfeld, “Establishing greener products and manufacturing processes,” *International Journal of Precision Engineering and Manufacturing*, vol. 13, no. 7. pp. 1029–1036, 2012, doi: 10.1007/s12541-012-0134-z.

[111] A. Rentizelas and D. Georgakellos, “Incorporating life cycle external cost in optimization of the electricity generation mix,” *Energy Policy*, vol. 65, pp. 134–149, Feb. 2014, doi: 10.1016/j.enpol.2013.10.023.

[112] P. Rogers, R. D. De Silva, and R. Bhatia, “Water is an economic good: How to use prices to promote equity efficiency and sustainability,” *Water Policy*, 2002, doi: 10.1016/S1366-7017(02)00004-1.

[113] P. Rogers, A. Huber, and R. Bhatia, “Water as a Social and Economic Good : How to Put the Principle into Practice,” *Glob. Water Partnersh. Tech. Advis. Comm. TAC*, 1998.

[114] K. Holmgren and S. Amiri, “Internalising external costs of electricity and heat production in a municipal energy system,” *Energy Policy*, vol. 35, no. 10, pp. 5242–5253, Oct. 2007, doi: 10.1016/j.enpol.2007.04.026.

[115] F. Zhao, J. Ogaldez, and J. W. Sutherland, “Quantifying the water inventory of machining processes,” *CIRP Ann. - Manuf. Technol.*, vol. 61, no. 1, pp. 67–70, 2012, doi: 10.1016/j.cirp.2012.03.027.

[116] International Organization for Standardization (ISO), “ISO 14046:2014, Environmental management -- Water footprint -- Principles, requirements and guidelines.” International Organization for Standardization, Geneva, Switzerland, 2014.

[117] J. L. Chen, Y.-B. Chen, and H.-C. Huang, “Quantifying the Life Cycle Water Consumption of a Machine Tool,” *Procedia CIRP*, vol. 29, pp. 498–501, 2015, doi: 10.1016/j.procir.2015.02.197.

[118] T. T. Pham, T. D. Mai, T. D. Pham, M. T. Hoang, M. K. Nguyen, and T. T. Pham, “Industrial water mass balance as a tool for water management in industrial parks,” *Water Resour. Ind.*, vol. 13, pp. 14–21, 2016, doi: 10.1016/j.wri.2016.04.001.

[119] M. Patterson, “Water Usage Effectiveness (WUE^{TM}): A Green Grid Data Center Sustainability Metric,” 2011, Accessed: Dec. 10, 2017. [Online]. Available: http://tmp2014.airatwork.com/wp-content/uploads/The-Green-Grid-White-Paper-35-WUE-Usage-Guidelines.pdf.

[120] U.S. Department of Energy, “Benchmark the Fuel Cost of Steam Generation.” Washington, DC, USA, 2012, doi: DOE/GO-102012-3391.

[121] J. Sturman, G. Ho, and K. Mathew, *Water Auditing and Water Conservation*. IWA Publishing, 2004.

[122] N. A. Shuaib, P. T. Mativenga, J. Kazie, and S. Job, “Resource efficiency and composite waste in UK supply chain,” in *Procedia CIRP*, 2015, vol. 29, pp. 662–667, doi: 10.1016/j.procir.2015.02.042.

[123] M. Altendorf, P. Berrie, H. Bjönnes, F. B. Jensen, and F. Bonschab, *Flow handbook: a practical guide: measurement technologies - applications - solutions*, 3rd ed. Endress + Hauser Flowtec, 2006.

[124] UN-Water, “Managing water under Uncertainty and Risk, The United Nations World Water Development Report 4,” 2012, Accessed: Feb. 10, 2018. [Online]. Available: http://www.unesco.org/new/en/natural-sciences/environment/water/wwap/wwdr/wwdr4-2012/.

[125] J. Redmond, E. Walker, and C. Wang, “Issues for small businesses with waste management,” *J. Environ. Manage.*, vol. 88, no. 2, pp. 275–285, Jul. 2008, doi: 10.1016/j.jenvman.2007.02.006.

[126] L. Liu, S. M. Kuo, and M. Zhou, “Virtual sensing techniques and their applications,” *2009 Int. Conf. Networking, Sens. Control*, pp. 31–36, Mar. 2009, doi: 10.1109/ICNSC.2009.4919241.

[127] O. Seifert, “New Developments in Flow Sensors for Industrial Furnaces,” *Energy Procedia*, vol. 120, pp. 469–476, Aug. 2017, doi: 10.1016/j.egypro.2017.07.226.

[128] A. Salonen and M. Deleryd, “Cost of poor maintenance: A concept for maintenance performance improvement,” *J. Qual. Maint. Eng.*, vol. 17, no. 1, pp. 63–73, Mar. 2011, doi: 10.1108/13552511111116259.

[129] E. O’Driscoll, D. O. Cusack, and G. E. O’Donnell, “Implementation of energy metering systems in complex manufacturing facilities-A case study in a biomedical facility,” *Procedia CIRP*, vol. 1, no. 1, pp. 524–529, 2012, doi: 10.1016/j.procir.2012.04.093.

[130] S. Humphrey, H. Papadopoulos, B. Linke, S. Maiyya, A. Vijayaraghavan, and R. Schmitt, “Power measurement for sustainable high-performance manufacturing processes,” in *Procedia CIRP*, 2014, vol. 14, pp. 466–471, doi: 10.1016/j.procir.2014.03.041.

[131] N. Weinert, S. Chiotellis, and G. Seliger, “Methodology for planning and operating energy-efficient production systems,” *CIRP Ann. - Manuf. Technol.*, vol. 60, no. 1, pp. 41–44, 2011, doi: 10.1016/j.cirp.2011.03.015.

[132] V. A. Balogun and P. T. Mativenga, “Modelling of direct energy requirements in mechanical machining processes,” *J. Clean. Prod.*, vol. 41, pp. 179–186, 2013, doi: 10.1016/j.jclepro.2012.10.015.

[133] K. Kellens, W. Dewulf, M. Overcash, M. Z. Hauschild, and J. R. Duflou, “Methodology for systematic analysis and improvement of manufacturing unit process life-cycle inventory (UPLCI) —CO 2 PE! initiative (cooperative effort on process emissions in manufacturing). Part 1: Methodology description,” *Int. J. Life Cycle Assess.*, vol. 17, no. 1, pp. 69–78, Jan. 2012, doi: 10.1007/s11367-011-0340-4.

[134] K. Kellens, W. Dewulf, M. Overcash, M. Z. Hauschild, and J. R. Duflou, “Methodology for systematic analysis and improvement of manufacturing unit process life cycle inventory (UPLCI) CO2PE! initiative (cooperative effort on process emissions in manufacturing). Part 2: Case studies,” *Int. J. Life Cycle Assess.*, vol. 17, no. 2, pp. 242–251, Feb. 2012, doi: 10.1007/s11367-011-0352-0.

[135] M. F. Rajemi, P. T. Mativenga, and A. Aramcharoen, “Sustainable machining: Selection of optimum turning conditions based on minimum energy considerations,” *J. Clean. Prod.*, vol. 18, no. 10–11, pp. 1059–1065, 2010, doi: 10.1016/j.jclepro.2010.01.025.

[136] Sustainable Energy Authority of Ireland, “Energy Audit Handbook,” Dublin, Ireland, 2016. Accessed: Dec. 14, 2017. [Online]. Available: https://www.seai.ie/resources/publications/SEAI-Energy-Audit-Handbook.pdf.

[137] P. Rao, M. R. Muller, and G. Gunn, “Conducting a metering assessment to identify submetering needs at a manufacturing facility,” *CIRP J. Manuf. Sci. Technol.*, vol. 18, pp. 107–114, Aug. 2017, doi: 10.1016/j.cirpj.2016.10.005.

[138] L. Fortuna, S. Graziani, A. Rizzo, and M. G. Xibilia, *Soft Sensors for Monitoring and Control of Industrial Processes*. Springer London, 2007.

[139] S. Boslaugh and P. A. Watters, “Proxy Metering,” in *Statistics in a Nutshell - A Desktop Quick Reference*, O’Reilly Media, 2008, p. 480.

[140] R. Luttmann, D. G. Bracewell, G. Cornelissen, K. V. Gernaey, J. Glassey, V. C. Hass, C. Kaiser, C. Preusse, G. Striedner, and C. F. Mandenius, “Soft sensors in bioprocessing: A status report and recommendations,” *Biotechnol. J.*, vol. 7, no. 8, pp. 1040–1048, Aug. 2012, doi: 10.1002/biot.201100506.

[141] F. A. A. Souza, R. Araújo, and J. Mendes, “Review of soft sensor methods for regression applications,” *Chemom. Intell. Lab. Syst.*, vol. 152, pp. 69–79, Mar. 2016, doi: 10.1016/j.chemolab.2015.12.011.

[142] M. Shakil, M. Elshafei, M. A. Habib, and F. A. Maleki, “Soft sensor for NOx and O2 using dynamic neural networks,” *Comput. Electr. Eng.*, vol. 35, no. 4, pp. 578–586, Jul. 2009, doi: 10.1016/j.compeleceng.2008.08.007.

[143] T. Aziz Al-Qutami, R. Ibrahim, I. Ismail, and M. A. Ishak, “Development of soft sensor to estimate multiphase flow rates using neural networks and early stopping,” *Int. J. SMART Sens. Intell. Syst.*, vol. 10, no. 1, 2017.

[144] K. Zheng, J. Watt, C. Wang, and Y. K. Cho, “Development and implementation of a virtual outside air wet-bulb temperature sensor for improving water-cooled chiller plant energy efficiency,” *Sustain. Cities Soc.*, vol. 23, pp. 11–15, May 2016, doi: 10.1016/j.scs.2016.02.012.

[145] P. Kadlec, B. Gabrys, and S. Strandt, “Data-driven Soft Sensors in the process industry,” *Comput. Chem. Eng.*, vol. 33, no. 4, pp. 795–814, Apr. 2009, doi: 10.1016/j.compchemeng.2008.12.012.

[146] R. Doraiswami, C. Diduch, and M. Stevenson, “Soft Sensor,” in *Identification of Physical Systems: Applications to Condition Monitoring, Fault Diagnosis, Soft Sensor and Controller Design*, Chichester, UK: John Wiley & Sons, Ltd, 2014, p. 536.

[147] I. Boiarkina, N. Depree, W. Yu, D. I. Wilson, and B. R. Young, “Rapid particle size measurements used as a proxy to control instant whole milk powder dispersibility,” *Dairy Sci. Technol.*, vol. 96, no. 6, pp. 777–786, Feb. 2017, doi: 10.1007/s13594-016-0302-5.

[148] A. K. Pani and H. K. Mohanta, “A survey of data treatment techniques for soft sensor design,” *Chem. Prod. Process Model.*, vol. 6, no. 1, Jan. 2011, doi: 10.2202/1934-2659.1536.

[149] P. Kadlec, R. Grbic, and B. Gabrys, “Review of adaptation mechanisms for data-driven soft sensors,” *Comput. Chem. Eng.*, vol. 35, no. 1, pp. 1–24, Jan. 2011, doi: 10.1016/j.compchemeng.2010.07.034.

[150] B. Lin, B. Recke, J. K. H. Knudsen, and S. B. Jørgensen, “A systematic approach for soft sensor development,” *Comput. Chem. Eng.*, vol. 31, no. 5–6, pp. 419–425, May 2007, doi: 10.1016/j.compchemeng.2006.05.030.

[151] S. K. Lahiri, “Soft Sensors,” in *Multivariable Predictive Control: Applications in Industry*, John Wiley & Sons, Ltd, 2017, pp. 145–165.

[152] D. Khachane, A. Suryawanshi, and D. Pathak, “Data selection and filtration technique for tuning virtual sensor model of NOx estimation in MATLAB,” in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Aug. 2016, pp. 1–4, doi: 10.1109/ICCUBEA.2016.7860148.

[153] H. Kaneko, “Automatic outlier sample detection based on regression analysis and repeated ensemble learning,” *Chemom. Intell. Lab. Syst.*, vol. 177, pp. 74–82, Jun. 2018, doi: 10.1016/j.chemolab.2018.04.015.

[154] L. Davies and U. Gather, “The identification of multiple outliers,” *J. Am. Stat. Assoc.*, vol. 797, no. 5, pp. 1–5, Sep. 1993, doi: 10.2307/2290763.

[155] R. Pearson, “Scrub Data with Scale-invariant Nonlinear Digital Filters,” *EDN*, vol. 47, no. 2, pp. 71–75, 2002.

[156] M. Hubert and S. Van Der Veeken, “Outlier detection for skewed data,” *J. Chemom.*, vol. 22, no. 3–4, pp. 235–246, Mar. 2008, doi: 10.1002/cem.1123.

[157] P. Saha, N. Roy, D. Mukherjee, and A. K. Sarkar, “Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data,” in *Procedia Computer Science*, Jan. 2016, vol. 83, pp. 107–114, doi: 10.1016/j.procs.2016.04.105.

[158] C. M. Andersen and R. Bro, “Variable selection in regression-a tutorial,” *J. Chemom.*, vol. 24, no. 11–12, pp. 728–737, Nov. 2010, doi: 10.1002/cem.1360.

[159] E. O’Driscoll, K. Kelly, D. O. Cusack, and G. E. O’Donnell, “Characterising the Energy Consumption of Machine Tool Actuator Components Using Pattern Recognition,” *Procedia CIRP*, vol. 12, pp. 127–132, 2013, doi: 10.1016/j.procir.2013.09.023.

[160] P. Kadlec and B. Gabrys, “Soft sensors: Where are we and what are the current and future challenges?,” in *IFAC Proceedings Volumes*, Jan. 2009, vol. 42, no. 19, pp. 572–577, doi: 10.3182/20090921-3-TR-3005.00098.

[161] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997, doi: 10.1109/4235.585893.

[162] K. Jørgensen, V. Segtnan, K. Thyholt, and T. Næs, “A comparison of methods for analysing regression models with both spectral and designed variables,” *J. Chemom.*, vol. 18, no. 10, pp. 451–464, Oct. 2004, doi: 10.1002/cem.890.

[163] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer New York, 2009.

[164] S. Soares, R. Araújo, P. Sousa, and F. Souza, “Design and application of soft sensor using ensemble methods,” in *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*, Sep. 2011, pp. 1–8, doi: 10.1109/ETFA.2011.6059061.

[165] I. Witten, E. Frank, and M. Hall, *Data Mining - Practical Machine Learning Tools and Techniques*, 3rd ed. Elsevier Inc., 2011.

[166] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer New York, 2013.

[167] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005, doi: 10.3354/cr030079.

[168] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge University Press, 2007.

[169] A. Di Bella, L. Fortuna, S. Graziani, G. Napoli, and M. G. Xibilia, “Development of a soft sensor for a thermal cracking unit using a small experimental data set,” in *2007 IEEE International Symposium on Intelligent Signal Processing, WISP*, 2007, pp. 1–6, doi: 10.1109/WISP.2007.4447584.

[170] G. Napoli and M. G. Xibilia, “Soft Sensor design for a Topping process in the case of small datasets,” *Comput. Chem. Eng.*, vol. 35, no. 11, pp. 2447–2456, Nov. 2011, doi: 10.1016/j.compchemeng.2010.12.009.

[171] L. Fortuna, S. Graziani, and M. G. Xibilia, “Comparison of soft-sensor design methods for industrial plants using small data sets,” *IEEE Trans. Instrum. Meas.*, vol. 58, no. 8, pp. 2444–2451, Aug. 2009, doi: 10.1109/TIM.2009.2016386.

[172] A. I. Maciel, I. G. Costa, and A. C. Lorena, “Measuring the complexity of regression problems,” in *Proceedings of the International Joint Conference on Neural Networks*, Jul. 2016, pp. 1450–1457, doi: 10.1109/IJCNN.2016.7727369.

[173] P. Brazidal, C. Giraud-Carrier, C. Soares, and R. Vilalta, *Meta-learning: Applications to data mining*. Springer Berlin Heidelberg, 2009.

[174] E. Leyva, A. González, and R. Pérez, “A set of complexity measures designed for applying meta-learning to instance selection,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 354–367, Feb. 2015, doi: 10.1109/TKDE.2014.2327034.

[175] A. C. Lorena, A. I. Maciel, P. B. C. de Miranda, I. G. Costa, and R. B. C. Prudêncio, “Data complexity meta-features for regression problems,” *Mach. Learn.*, vol. 107, no. 1, pp. 209–246, Jan. 2018, doi: 10.1007/s10994-017-5681-1.

[176] J. Abonyi, B. Farsang, and T. Kulcsar, “Data-driven development and maintenance of soft-sensors,” in *SAMI 2014 - IEEE 12th International Symposium on Applied Machine Intelligence and Informatics, Proceedings*, Jan. 2014, pp. 239–244, doi: 10.1109/SAMI.2014.6822414.

[177] L. I. Kuncheva and I. Žliobait, “On the window size for classification in changing environments,” *Intell. Data Anal.*, vol. 13, no. 6, pp. 861–872, Nov. 2009, doi: 10.3233/IDA-2009-0397.

[178] H. Kaneko and K. Funatsu, “Maintenance-free soft sensor models with time difference of process variables,” *Chemom. Intell. Lab. Syst.*, vol. 107, no. 2, pp. 312–317, Jul. 2011, doi: 10.1016/j.chemolab.2011.04.016.

[179] M. Kano and Y. Nakagawa, “Data-based process monitoring, process control, and quality improvement: Recent developments and applications in steel industry,” *Comput. Chem. Eng.*, vol. 32, no. 1–2, pp. 12–24, Jan. 2008, doi: 10.1016/j.compchemeng.2007.07.005.

[180] R. D. Tobias, “An Introduction to Partial Least Squares Regression,” in *Proceedings of the Twentieth Annual SAS Users Group International Conference*, 1995, pp. 1250–1257.

[181] D. Dong and T. J. Mcavoy, “Nonlinear principal component analysis - Based on principal curves and neural networks,” *Comput. Chem. Eng.*, vol. 20, no. 1, pp. 65–78, Jan. 1996, doi: 10.1016/0098-1354(95)00003-K.

[182] S. Wold, M. Sjöström, and L. Eriksson, “PLS-regression: A basic tool of chemometrics,” *Chemom. Intell. Lab. Syst.*, vol. 58, no. 2, pp. 109–130, Oct. 2001, doi: 10.1016/S0169-7439(01)00155-1.

[183] R. Bro and A. K. Smilde, “Principal component analysis,” *Anal. Methods*, vol. 6, no. 9, pp. 2812–2831, Apr. 2014, doi: 10.1039/C3AY41907J.

[184] D. Dornfeld and D.-E. Lee, *Precision Manufacturing*, 1st ed. Springer US, 2008.

[185] R. Rojas, *Neural Networks - A Systematic Introduction*, 1st ed. Springer Berlin Heidelberg, 1996.

[186] H. Yu and B. M. Wilamowski, “Levenberg–Marquardt Training,” in *Industrial Electronics Handbook Intelligent Systems*, CRC Press, 2011, pp. 12–15.

[187] R. I. Tange, M. A. Rasmussen, E. Taira, and R. Bro, “Benchmarking support vector regression against partial least squares regression and artificial neural network: Effect of sample size on model performance,” *J. Near Infrared Spectrosc.*, vol. 25, no. 6, pp. 381–390, Dec. 2017, doi: 10.1177/0967033517734945.

[188] H. M. Hashemian, “State-of-the-art predictive maintenance techniques,” *IEEE Trans. Instrum. Meas.*, vol. 60, no. 1, pp. 226–236, Jan. 2011, doi: 10.1109/TIM.2010.2047662.

[189] Z. Liu, N. Meyendorf, and N. Mrad, “The role of data fusion in predictive maintenance using digital twin,” in *AIP Conference Proceedings*, Apr. 2018, vol. 1949, no. 1, p. 020023, doi: 10.1063/1.5031520.

[190] Z. Wang, J. Gong, H. Wu, and Q. Li, “The development and application of virtual flow metering system in offshore gas field,” in *Proceedings of the Biennial International Pipeline Conference, IPC*, Sep. 2014, vol. 1, doi: 10.1115/IPC201433399.

[191] G. Äijälä and D. Lumley, “Integrated soft sensor model for flow control,” *Water Sci. Technol.*, vol. 53, no. 4–5, pp. 473–482, 2006, doi: 10.2166/wst.2006.152.

[192] P. Juntunen, M. Liukkonen, M. J. Lehtola, and Y. Hiltunen, “Dynamic soft sensors for detecting factors affecting turbidity in drinking water,” *J. Hydroinformatics*, vol. 15, no. 2, pp. 416–426, 2013, doi: 10.2166/hydro.2012.052.

[193] L. Xie, Y. Zhao, D. Aziz, X. Jin, L. Geng, E. Goberdhansingh, F. Qi, and B. Huang, “Soft sensors for online steam quality measurements of OTSGs,” *J. Process Control*, vol. 23, no. 7, pp. 990–1000, Aug. 2013, doi: 10.1016/j.jprocont.2013.05.006.

[194] X. Li, S. Deng, S. Wang, Z. Lv, and L. Wu, “Review of Small Data Learning Methods,” in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, Jul. 2018, vol. 01, pp. 106–109, doi: 10.1109/COMPSAC.2018.10212.

[195] T. I. Tsai and D. C. Li, “Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems,” *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1293–1300, Oct. 2008, doi: 10.1016/j.eswa.2007.08.043.

[196] J. J. Faraway and N. H. Augustin, “When small data beats big data,” *Stat. Probab. Lett.*, vol. 136, pp. 142–145, May 2018, doi: 10.1016/j.spl.2018.02.031.

[197] R. Kitchin and T. P. Lauriault, “Small data in the era of big data,” *GeoJournal*, vol. 80, no. 4, pp. 463–475, Aug. 2015, doi: 10.1007/s10708-014-9601-7.

[198] T. F. Edgar and E. N. Pistikopoulos, “Smart manufacturing and energy systems,” *Comput. Chem. Eng.*, vol. 114, pp. 130–144, 2018, doi: 10.1016/J.COMPCHEMENG.2017.10.027.

[199] A. Di Bella, S. Graziani, G. Napoli, and M. G. Xibilia, “Stacking approaches for the design of Soft Sensors using small data set,” in *2008 Mediterranean Conference on Control and Automation - Conference Proceedings, MED’08*, Jun. 2008, pp. 1810–1815, doi: 10.1109/MED.2008.4602160.

[200] M. A. Babyak, “What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models,” *Psychosom. Med.*, vol. 66, no. 3, pp. 411–421, May 2004, doi: 10.1097/01.psy.0000127692.23278.a9.

[201] R. Adawiah Mat Noor and Z. Ahmad, “Neural network based soft sensor for prediction of biopolycaprolactone molecular weight using bootstrap neural network technique,” in *Conference on Data Mining and Optimization*, Jun. 2011, pp. 70–73, doi: 10.1109/DMO.2011.5976507.

[202] Ž. U. Andrijić, M. Cvetnić, N. Bolf, Ž. U. Andrijić, M. Cvetnić, and N. Bolf, “SOFT SENSOR MODELS FOR A FRACTIONATION REFORMATE PLANT USING SMALL AND BOOTSTRAPPED DATA SETS,” *Brazilian J. Chem. Eng.*, vol. 35, no. 2, pp. 745–756, Jun. 2018, doi: 10.1590/0104-6632.20180352s20150727.

[203] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *Proc. 14th Int. Jt. Conf. Artif. Intell. - Vol. 2*, vol. 2, pp. 1137–1143, 1995, doi: 10.1067/mod.2000.109031.

[204] C. Kamath and Y. J. Fan, “Regression with small data sets: a case study using code surrogates in additive manufacturing,” *Knowl. Inf. Syst.*, vol. 57, no. 2, pp. 475–493, Nov. 2018, doi: 10.1007/s10115-018-1174-1.

[205] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Stat. Surv.*, vol. 4, no. 0, pp. 40–79, 2009, doi: 10.1214/09-SS054.

[206] R. Caponetto, G. Dongola, A. Gallo, and M. G. Xibilia, “FPGA based soft sensor for the estimation of the kerosene freezing point,” in *Proceedings - 2009 IEEE International Symposium on Industrial Embedded Systems, SIES 2009*, Jul. 2009, pp. 228–236, doi: 10.1109/SIES.2009.5196219.

[207] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[208] J. Barron, C. Ashton, and L. Geary, “The Effects of temperature on pH measurement,” *TSP*, vol. 1, no. 2, pp. 1–7, 2005.

[209] N. Japkowicz and M. Shah, *Evaluating learning algorithms: A classification perspective*. Cambridge University Press, 2011.

[210] H. A. Martens and P. Dardenne, “Validation and verification of regression in small data sets,” *Chemom. Intell. Lab. Syst.*, vol. 44, no. 1–2, pp. 99–121, 1998, doi: 10.1016/S0169-7439(98)00167-1.

[211] P. Kadlec, “On robust and adaptive soft sensors,” PhD thesis, Bournemouth University, UK, 2009.

[212] T. G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms,” *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[213] R. R. Bouckaert and E. Frank, “Evaluating the replicability of significance tests for comparing learning algorithms,” in *Advances in Knowledge Discovery and Data Mining. PAKDD 2004. Lecture Notes in Computer Science*, vol. 3056, Springer Berlin Heidelberg, 2004.

[214] C. Nadeau and Y. Bengio, “Inference for the generalization error,” *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003, doi: 10.1023/A:1024068626366.

[215] W. Bohl and W. Elmendorf, *Technische Strömungslehre*, 13th ed. Vogel Buchverlag, 2005.

List of Publications

Conference proceedings

- International Manufacturing Conference (2017):

  D. Seiler, D. Donovan, and G.E. O'Donnell

  "A Perspective on Energy and Resource Measurement Strategies in Industrial Water Systems: A Basis for Analysis, Modelling and Optimisation"

- CIRP: Conference on Life Cycle Engineering (2018):

  D. Seiler, D. Donovan, and G.E. O'Donnell

  "Facing the information leaks in industrial water systems: A concept for proxy measurements to support water metering audits"

  DOI: *https://doi.org/10.1016/j.procir.2017.11.036*

- CIRP: Conference on Life Cycle Engineering (2019):

  D. Seiler, D. Donovan, and G.E. O'Donnell

  "Resource transparency of industrial systems: A model to demonstrate the total costs of water purification"

  DOI: *https://doi.org/10.1016/j.procir.2019.01.013*

Journal publications

- Journal of Cleaner Production (2019):

  D. Seiler, D. Donovan, and G.E. O'Donnell

  "Industrial steam systems: Strategic framework with proxy measurements to evaluate the total cost of steam"

  DOI: *https://doi.org/10.1016/j.jclepro.2019.05.285*

- Journal of Cleaner Production (2020):

  D. Seiler and G.E. O'Donnell

  "Filling the gap: Comparison of proxy measurement techniques for small data scenarios"

  Manuscript accepted, in publication

Appendices

Appendix A: Dataset complexity measures

This section outlines the applied dataset complexity measures for regression models that were introduced by Lorena et al. [175]. The measures are presented in Table 2.3. Datasets normalisation between [0,1] has been performed before the measures were applied (except for measure *T1*).

Correlation measures:

C1: Maximum inputs correlation to target variable: This measure identifies the maximum correlation value between each input variable and the target variable:

$$\mathrm{C1} = \max\left|\rho(x^j, y)\right|_{j=1,\dots,d} \tag{A.1}$$

The Spearman correlation ρ is used as it is non-parametric and does not assume a particular distribution between the input variables and the target variable from the dataset. The absolute value of the correlation is used as both extreme values [-1,1] indicate a strong correlation.

Higher values of C1 indicate simpler regression problems meaning there is at least one input variable strongly correlated to the target variable.

C2: Average inputs correlation to target variable: This parameter measures the average correlation of all input variables to the target variable:

$$\mathrm{C2} = \sum_{j=1}^{d} \frac{\left|\rho(x^j, y)\right|}{d} \tag{A.2}$$

Similar to C1, this measure uses the Spearman correlation ρ and the absolute value. Higher values of C2 indicate a simpler regression problem.

C3: Input efficiency: This measure examines if a SLR with a certain input variable can explain the relationship to the target variable well. First, the input variable with the highest correlation to the target variable is selected. Second, a linear fit between the selected input variable (highest correlation) and the target variable is performed. Third, all samples of the input variable that show a residual error of less than 0.1 ($|\epsilon_i| \leq 0.1$) are removed from the dataset. Afterwards, the most correlated input variable to the remaining samples is identified and the previous steps are repeated. This process is iterated until all input variables are analysed or the input dataset X becomes empty. C3 expresses the ratio of samples for which the residuals error of the linear fit is larger than the threshold of 0.1.

$$\mathrm{C3} = \frac{\#\,\{x_i | |\epsilon_i| > 0.1\}_{T_l}}{n} \tag{A.3}$$

where the numerator gives the number of samples that remain in the dataset at the end. T_l is the dataset from which this number is calculated, l indicates the number of iterations that are executed by the algorithm. n is the number of samples in the original dataset. Higher values for C3 indicate more complex problems.

Linearity measures:

L1: Mean absolute error: This parameter uses an MLR between input and target variables. The mean absolute error of this linear fit is expressed by L1:

$$L1 = \sum_{i=1}^{n} \frac{|\epsilon_i|}{n} \tag{A.4}$$

Lower values of L1 indicate simpler (linear) problems.

L2: Residual variance: Similar to L1, an MLR between input and target variables is first performed for L2. Afterwards, the residual values are squared and the average is calculated according to:

$$L2 = \sum_{i=1}^{n} \frac{\epsilon_i^2}{n} \tag{A.5}$$

Lower values of L2 indicate simpler (linear) problems.

Smoothness measures:

S1: Inputs distribution: For this measure, the data samples are first arranged according to the ascending order of the target values y_i. Subsequently, the Euclidean distance is computed between neighbouring samples. Finally, the Euclidean distances of all pairs are summarised and divided by the total number of samples:

$$S1 = \frac{1}{n}\sum_{i=2}^{n} \|x_i - x_{i-1}\|_2 \tag{A.6}$$

Lower values of S1 indicate simpler problems (similar points in the input space share similar target values).

S2: Error of nearest neighbour regressor: This measure is applied to identify how close the samples lie together. First, a distance matrix for the sample inputs in X is calculated. Subsequently, the target value of each input sample x_i is predicted by the nearest neighbour input sample $NNR(x_i)$. The Mean Squared Error (MSE) for a 1-nearest neighbour regressor (NNR) expresses the accuracy. Finally, all MSEs are summarised and divided by the number of samples:

$$S2 = \frac{1}{n}\sum_{i=1}^{n} (NNR(x_i) - y_i)^2 \tag{A.7}$$

where $NNR(x_i)$ represents the 1-nearest neighbour prediction for x_i.
Lower values of S2 indicate simpler problems.

Geometry, topology, and density measures:

L3: Non-linearity of linear regressor: This measure first selects pairs of samples with similar target outputs and then creates a new test point by randomly interpolating within the pair. Simultaneously, a linear regression model is trained on the original dataset and applied to the new, interpolated points. By using the MSE, L3 measures how sensitive the regression model is to the interpolated points:

$$L3 = \frac{1}{l}\sum_{i=1}^{l}(f(x_i') - y_i')^2 \tag{A.8}$$

where l is the number of interpolated input samples x_i' and target samples y_i'.
Lower values of L3 indicate simpler problems, indicating that the training samples are distributed smoothly.

S3: Non-linearity of nearest neighbour regressor: This parameter applies the same calculation pattern as in L3, however, this time a Nearest Neighbour regressor (NNR) is trained on the original dataset instead of a linear regression model:

$$S3 = \frac{1}{l}\sum_{i=1}^{l}(NNR(x_i') - y_i')^2 \tag{A.9}$$

where l is the number of interpolated input samples x_i' and target samples y_i'.
Lower values of S3 indicate simpler problems, indicating that the training samples are distributed smoothly.

T1: Average number of samples per dimension: This measure expresses the average number of samples per dimension and gives an indicative on data sparsity:

$$T1 = \frac{n}{d} \tag{A.10}$$

Lower values of T1 indicate more complex problems.

Appendix B: Calibration of experimental rig measuring devices

Before the experimental rig was set into operation, the temperature, flow rate, and pH sensors were calibrated to ensure high accuracy of the measurements. The calibration and validation procedure for these measuring devices is explained below.

Thermocouples

To measure the water and air temperature in and around the experimental rig, K-type thermocouples have been applied. These devices were connected to a thermocouple module (NI-9213) within the compact DAQ which included cold-junction compensation and enabled the temperature acquisition through the LabVIEW environment. To assure high accuracy, the thermocouples were calibrated for a fixed temperature range within a calibration bath. A master temperature device (Stanford Research Systems calibrated thermistor probe) has been used in the calibration bath as a reference sensor. The thermocouples were calibrated between 18°C and 78°C by using six different temperatures within this range. As soon as constant temperatures were reached for each setpoint, 10 values of the reference temperature from the master probe and the according temperatures from the thermocouples were recorded. Subsequently, the average deviations between the master probe and the thermocouples were determined and each thermocouple has been corrected according to the respective deviation by using linear regression.

Finally, the calibrated thermocouples were validated to determine the measurement uncertainty. Six validation temperatures within the range of 18°C to 78°C were approached in the calibration bath and the values of the thermocouples and the master probe were recorded and compared. The absolute deviations for each validation temperature setpoint are displayed in Table B.1.

Table B.1: Validation deviations (absolute) for the thermocouples of the experimental rig

Measuring device	Validation temperatures					
	21°C	30°C	35°C	40°C	50°C	60°C
TT1 [°C]	0.01	0.00	0.05	0.06	0.11	0.13
TT2 [°C]	0.08	0.09	0.08	0.14	0.17	0.19
TT3 [°C]	0.05	0.02	0.06	0.05	0.08	0.15
TT4 [°C]	0.09	0.10	0.16	0.15	0.18	0.19
TT5 [°C]	0.09	0.06	0.07	0.03	0.01	0.00

Flow meters:

Three flow sensors (RS Pro Beverage Flow Meter) were implemented within the experimental rig to measure the water flow rates at various locations. For signal acquisition, a voltage input module (NI-9205) has been used in the compact DAQ which enabled to correlate the signal frequency to the corresponding flow rate within the LabView environment. The single flow meters were calibrated by manually recording the past volume within a certain time interval when constant flow conditions were present. Subsequently, the signal frequencies were related to the manual flow rate recordings by linear relationships. The devices were calibrated by five setpoints within a flow range of 0.5 l/min to 2.0 l/min. The validation of the flow devices followed the same methodology as the calibration, manual recordings of the passed water volume were compared to the measured volumes by the flow meters. The following table presents the determined validation parameters.

Table B.2: Validation deviations (absolute) for the flow meters of the experimental rig

Measuring device	Validation flow rates					
	0.5 l/min	0.7 l/min	0.9 l/min	1.0 l/min	1.4 l/min	2.0 l/min
FT1 [l/min]	0.00	0.01	0.01	0.01	0.02	0.03
FT2 [l/min]	0.01	0.01	0.02	0.02	0.03	0.04
FT3 [l/min]	0.00	0.00	0.01	0.02	0.02	0.03
FT4 [l/min]	0.01	0.02	0.01	0.02	0.03	0.04

pH meters:

Two pH meters (DF Robot SEN0161) were implemented in the experimental rig to measure the pH value of the water in tank 1 and 2. The output voltage of these devices was changing according to the measured pH value and was recorded by a voltage input module (NI-9205) within the compactDAQ. To calibrate both meters, the pH value of water within a beaker was alternated by the addition of sodium carbonate (to increase the pH value) or sodium bisulfate (to decrease the pH value). Similar to the calibration of the thermocouples, a master pH meter (Mettler Toledo FiveEasy Plus) was used as a reference pH meter within the beaker. The pH meters were calibrated between pH 2.5 and pH 10.5 by applying five different pH setpoints within this range. As soon as constant pH values were reached at each setpoint, 10 values of the master pH meter and the indicated voltages from the pH meters were recorded. Subsequently, the average voltages for each setpoint were related to the average master pH value by a linear relationship. To validate the pH meters, five validation setpoints in the range of pH 2.5 to pH 10.5 were approached in the beaker and the values of the pH meters and the master probe were compared. The absolute deviations for each setpoint are shown in the following table.

Table B.3: Validation deviations (absolute) for pH meters of the experimental rig

Measuring device	Validation pH values				
	2.5	2.7	7.3	8.0	10.5
pHT1 [pH]	0.02	0.01	0.03	0.03	0.07
pHT2 [pH]	0.06	0.08	0.04	0.11	0.09

Appendix C: Experimental rig simulation model for setup D and E

C.1 Model derivation

The model for the experimental rig setups D and E has been simulated based on equations which relate the tank filling levels to the tank exit flow rates. An example tank with the associate parameters is shown in Figure C.1.

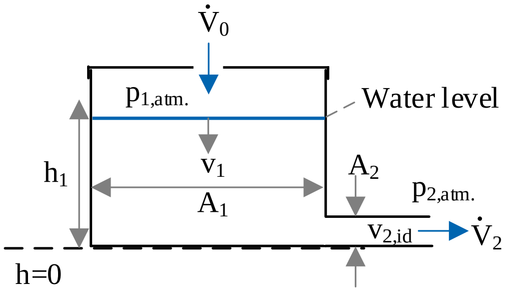


Figure C.1: Example tank of the experimental rig including model parameters

To derivate a model relating the tank filling level with the exit flow rate, a volume balance is defined for the example tank of Figure C.1 as a first step:

$$h_1 * A_1 = h_{1\_start} * A_1 + \dot{V}_0 * t - \dot{V}_2 * t \tag{C.1}$$

To determine the volume flow rates at the tank exit, the outlet speed can be calculated according to the Bernoulli equation:

$$p_{1,atm} + \rho * g * h_1 + \frac{\rho}{2} * v_1^2 = p_{2,atm} + \rho * g * h_2 + \frac{\rho}{2} * v_{2,id}^2 \tag{C.2}$$

Since the tank is an open system, the pressure term in equation (C.2) can be neglected. Additionally, the hydrostatic pressure at the tank exit is very small in this setup and was therefore neglected as well. Thus equation (C.2) can be reduced to:

$$\rho * g * h_1 + \frac{\rho}{2} * v_1^2 = \frac{\rho}{2} * v_{2,id}^2 \tag{C.3}$$

With the equation of continuity (equation (C.4)) and the inclusion of the coefficient of discharge (equation (C.5)), the speed at the tank exit is calculated according to equation (C.6) and the outlet volume flow rate is calculated with equation (C.7). The coefficient of discharge μ has been assumed according to the tank nozzle geometry [215].

$$v_1 * A_1 = v_{2,id} * A_2 \tag{C.4}$$

$$v_{2,real} = v_{2,id} * \mu \tag{C.5}$$

$$v_{2,real} = \sqrt{\frac{2 * g * h_1}{\left(\frac{1}{\mu}\right)^2 - \left(\frac{A_2}{A_1 * \mu}\right)^2}} \tag{C.6}$$

$$\dot{V}_2 = v_{2,real} * A_2 \tag{C.7}$$

Following from this, the change of the filling level in the example tank of Figure C.1 can be calculated with equation (C.1):

$$h_1 = h_{1\_start} + \frac{\dot{V}_0 * t}{A_1} - \frac{\dot{V}_2 * t}{A_1} \tag{C.8}$$

Correspondingly, the filling levels of the experimental rig tanks 1, 2, and 3 were calculated according to the following differential equation within the developed MATLAB Simulink model:

$$\frac{dh_1}{dt} = \frac{\dot{V}_0}{A_1} - \frac{\dot{V}_2}{A_1} \tag{C.9}$$

The block diagram of the Simulink model for setup D is displayed in Figure C.2. At the start of the simulation, an initial filling level, randomly selected, was provided for the integrators of tank 1 and 2. The difference of the tank filling levels was given with the unit m/s. This unit has been applied as the standard unit throughout the Simulink block diagram. Therefore, the volume flow rates were divided by the corresponding area.

Within setups D and E, the outflow of tank 1 was flowing into tank 2 and the outflow of tank 2 was flowing to the drain. Tank 1 was filled by mains water (dependent on the 2-position controller) and by the outflow of tank 3 (FT4). Within setup D, two constant flow rates were assumed for FT4, changing once at a random time during the simulation period. Within setup E, tank 3 was added to the Simulink block diagram and the flow rate FT4 was continuously decreasing from a randomly chosen filling height at the simulation start.

The implemented 2-position controller identified through a loop function if the filling level in tank 2 was above or below the set target filling level of 0.2 m and opened or closed the mains water addition to tank 1 through valve V1 correspondingly.

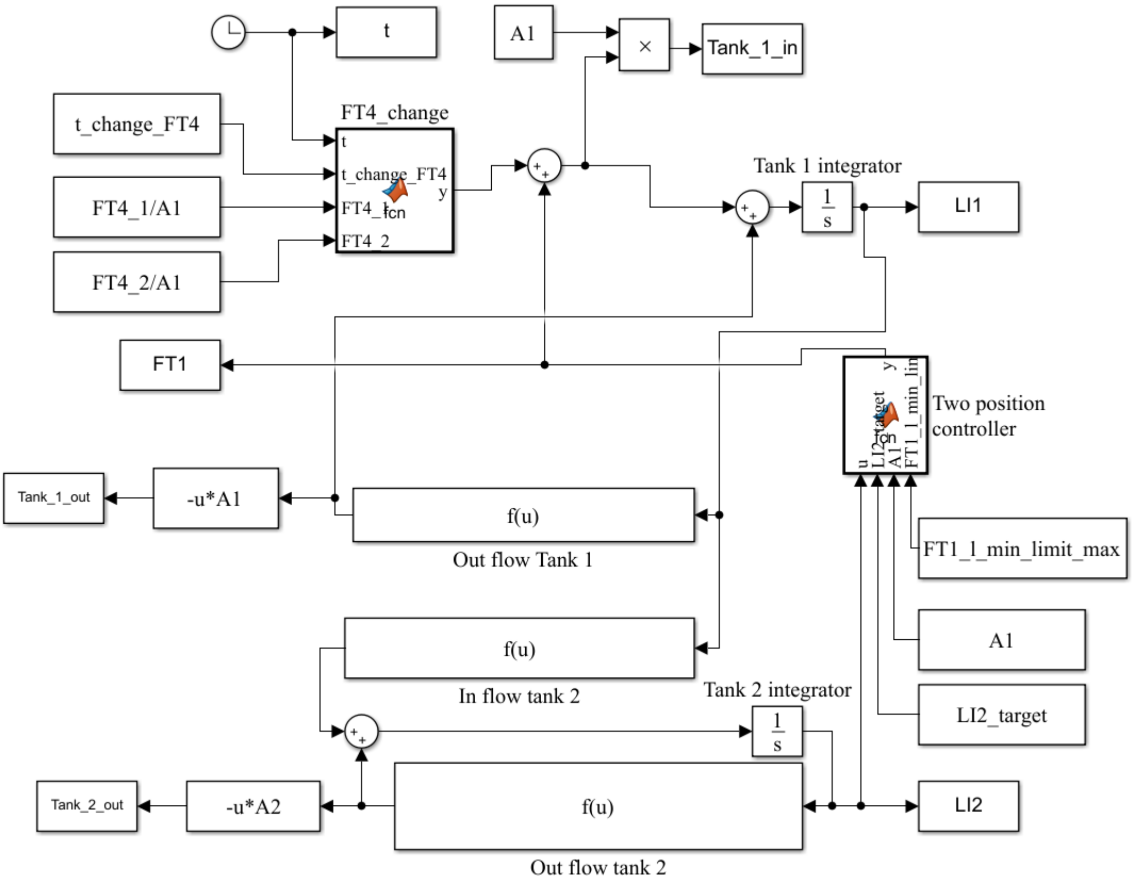


Figure C.2: Block diagram of the Simulink model for setup D

C.2 Model validation

To validate the simulation model of setup D and E, the models' differential equation (C.9) has been integrated to compare the simulated filling levels to the simulation time. To enable this comparison, a steady inflow ($\dot{V}_0$) has been assumed for the example tank shown in Figure C.1. Initially, equations (C.6), and (C.7) were substituted into equation (C.9) and two new variables (P and Q) were introduced to simplify the integration:

$$\frac{dh_1}{dt} = \frac{\dot{V}_0}{A_1} - \frac{A_2}{A_1}\sqrt{\frac{2g}{\left(\frac{1}{\mu}\right)^2 - \left(\frac{A_2}{A_1 * \mu}\right)^2}}\sqrt{h_1} = P - Q\sqrt{h_1} \tag{C.10}$$

with:

$$P = \frac{\dot{V}_0}{A_1} \tag{C.11}$$

$$Q = \frac{A_2}{A_1}\sqrt{\frac{2g}{\left(\frac{1}{\mu}\right)^2 - \left(\frac{A_2}{A_1 * \mu}\right)^2}} \tag{C.12}$$

Rearranging equation (C.10):

$$\frac{dh_1}{P - Q\sqrt{h_1}} = dt \tag{C.13}$$

with $x = \sqrt{h_1}$ and $dx = \frac{dh_1}{2\sqrt{h_1}}$, equation (C.13) is integrated:

$$\int_0^t P\,dt = \int_{x(0)}^{x(t)} \frac{2 * x}{1 - \frac{Q}{P}\,x}\,dx \tag{C.14}$$

$$t = -\frac{2}{Q}\left[x(t) - x(0) + \frac{P}{Q} * \log\left(\frac{P - Q * x(t)}{P - Q * x(0)}\right)\right] \tag{C.15}$$

Replacing $x(t) = \sqrt{h_1}(t),\ x(0) = \sqrt{h_1}(0)$:

$$t = -\frac{2}{Q}\left[\sqrt{h_1}(t) - \sqrt{h_1}(0) + \frac{P}{Q} * \log\left(\frac{P - Q * \sqrt{h_1}(t)}{P - Q * \sqrt{h_1}(0)}\right)\right] \tag{C.16}$$

After resubstituting P according to equation (C.11) and Q according to equation (C.12), equation (C.16) has been used to validate the simulation model. For this validation process, the following parameters were used:

$h_{1\_start} = 0.2\,m;\ A_1 = 0.0674\,m^2;\ A_2 = 0.785\,cm^2;\quad \dot{V}_0 = 2.0\,\frac{l}{min}$

$\mu = 0.6$ (assumed according to the nozzle geometry [215])

For these validation parameters, Figure C.3 [a] presents the Simulink simulation results of the example tanks' filling level h_1. The graph shows that the filling level is decreasing for a duration of approximately 10 minutes and staying constant for the further simulation time. Subsequently, the resulting values of the simulated filling level h_1 have been applied in equation (C.16) in order to calculate the corresponding time, the results are shown in Figure C.3 [b]. Similar to the Figure C.3 [a], the graph of Figure C.3[b] shows that the filling level h_1 decreases until approximately 10 minutes and shows a constant level afterwards. As the graphs of Figure C.3 demonstrate, the simulated filling level h_1 coincides with the calculated time of equation (C.16). As an example, the filling level of 0.05 m corresponds to the simulation time of five minutes. This comparison indicates that the simulation model is capable of correctly calculating the tanks' exit flow.

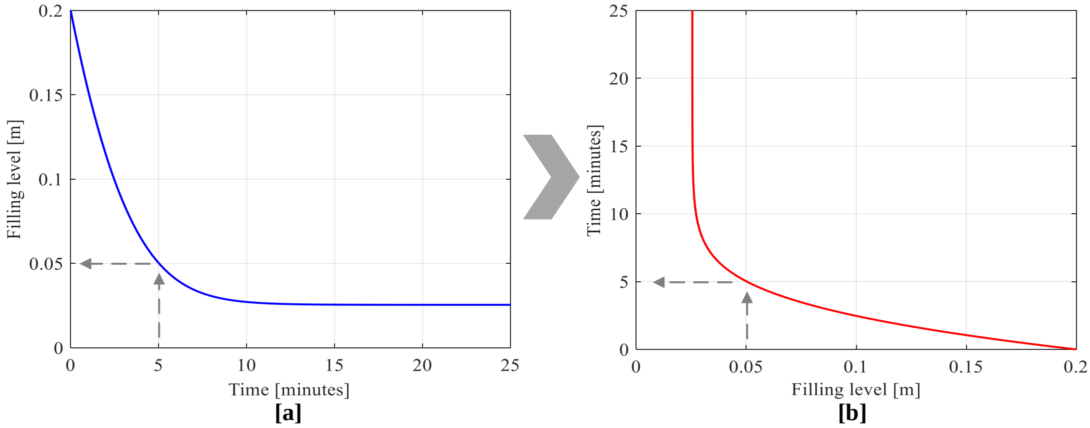


Figure C.3: Validation of simulation model: Simulation of filling level h_1 based on validation parameters [a], integration of simulated filling level h_1 according to equation (C.16) [b]

Appendix D: Supplementary figures

This section provides additional figures for the presented research of chapter 3 and 4.

D.1 Additional figures for the case facility's IC1 water system

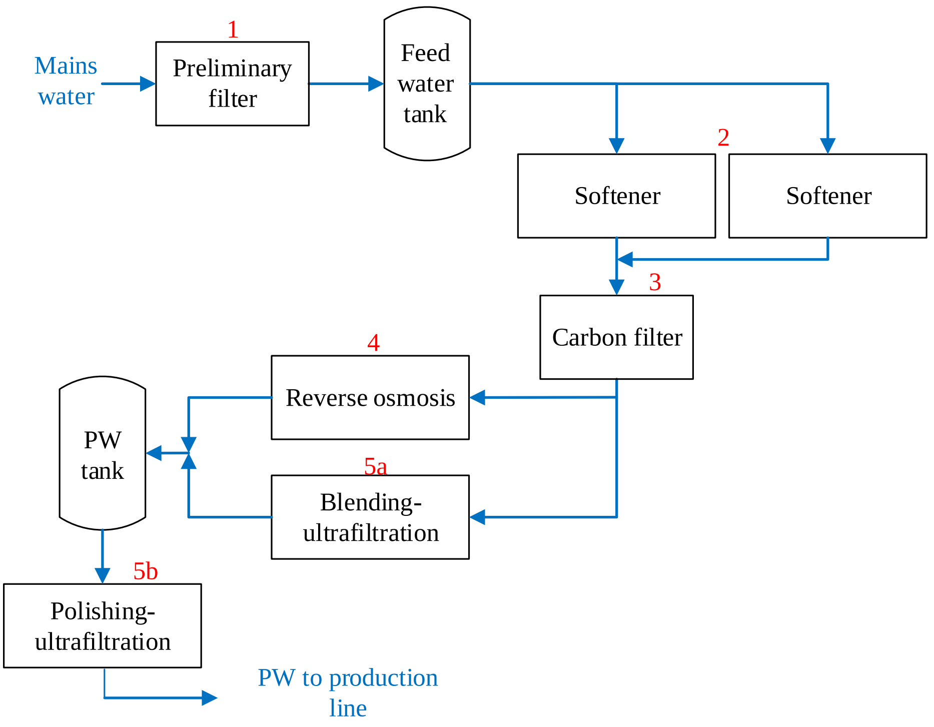


Figure D.1: WPDs within the IC1 water system of the case facility

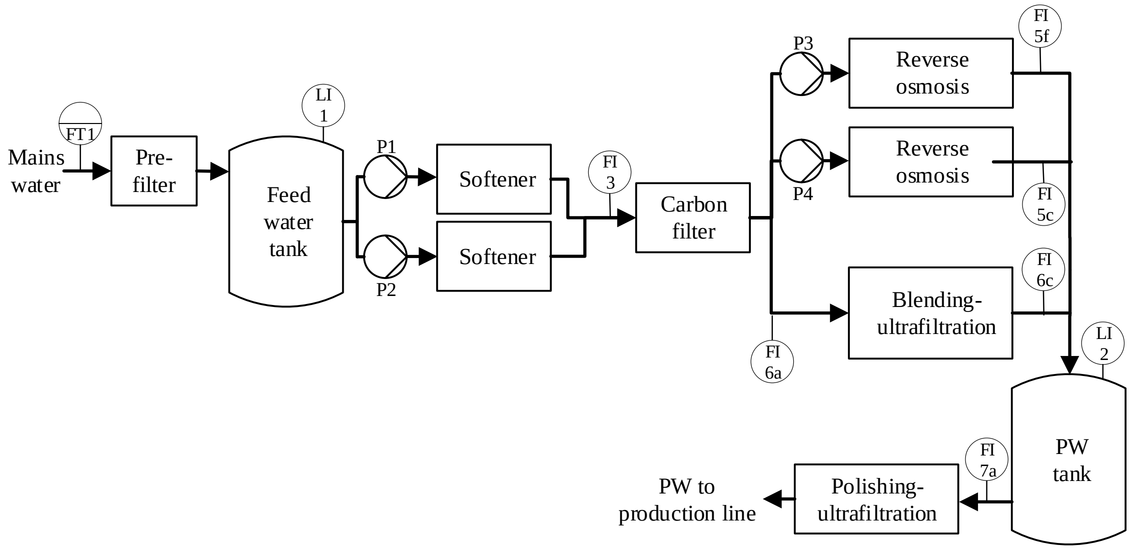


Figure D.2: BFD of the IC1 water purification process including flow and level indicators

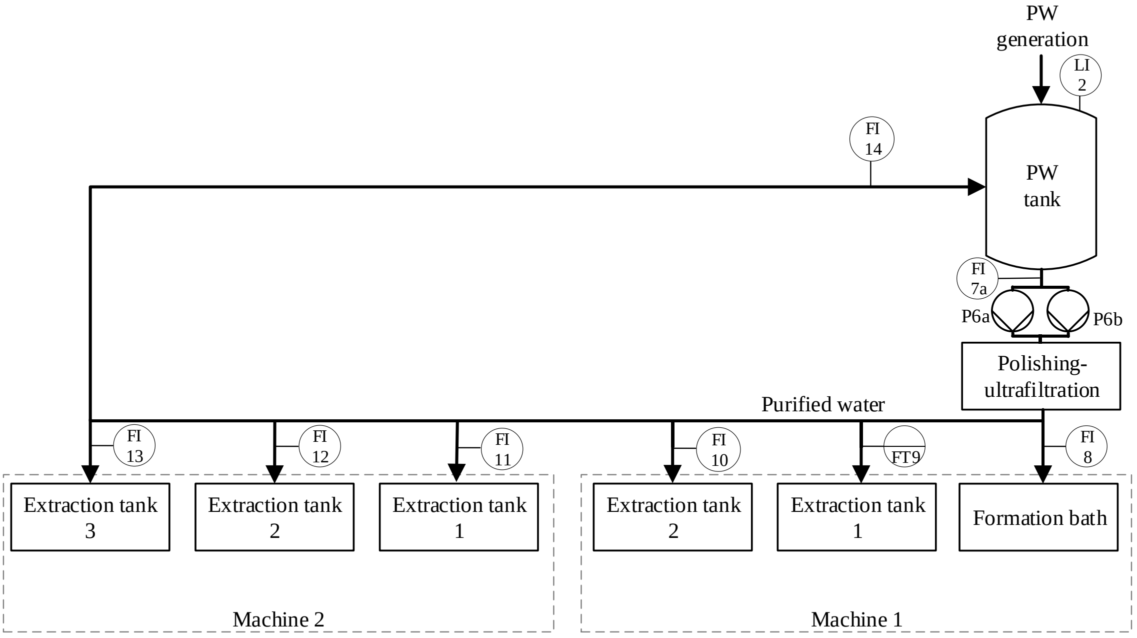


Figure D.3: BFD of the IC1 PW distribution including flow and level indicators

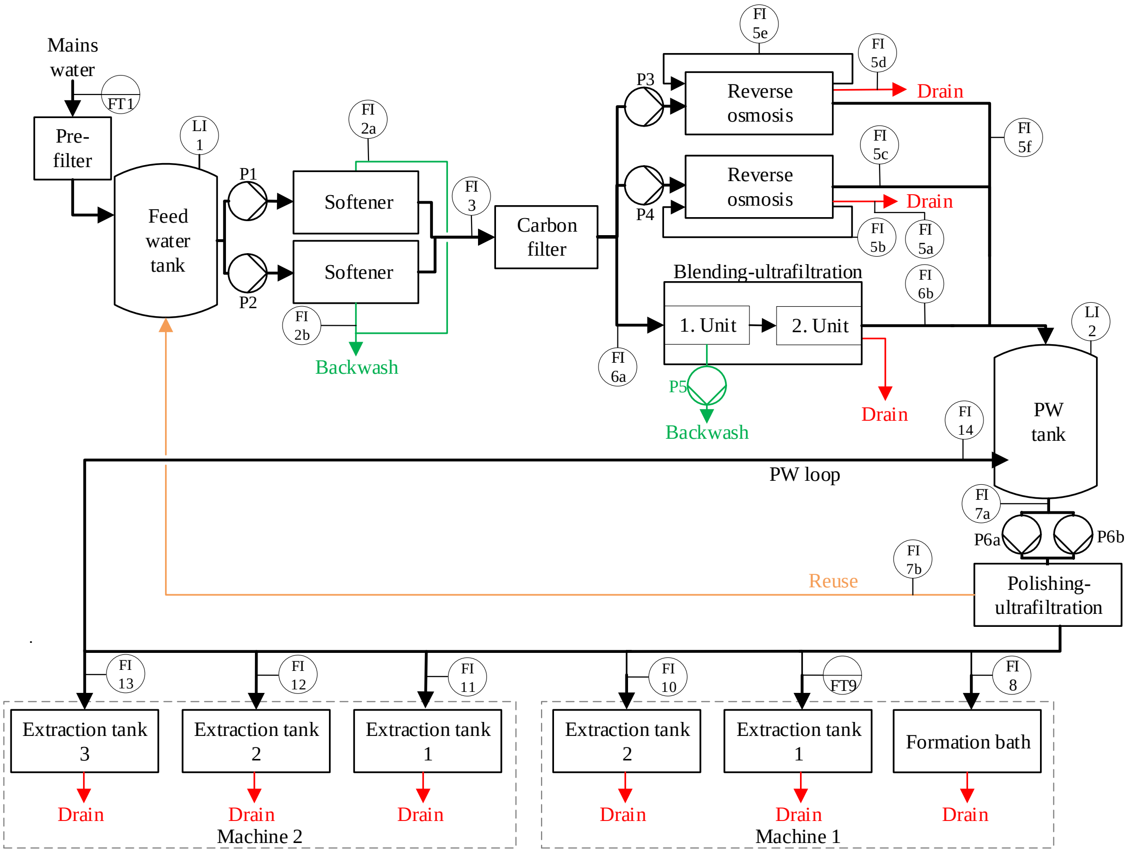


Figure D.4: BFD of the IC1 water system including sub-flows and measuring instruments

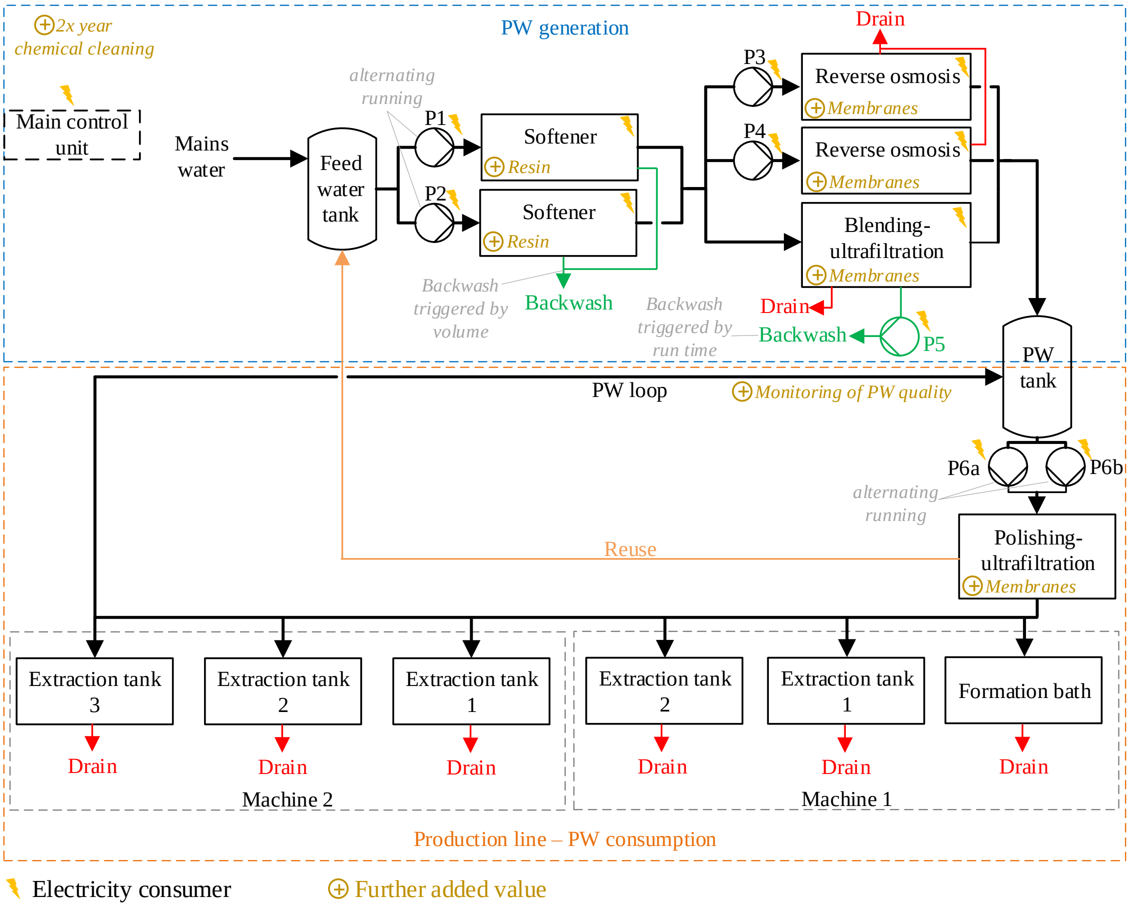


Figure D.5: Abstracted BFD for the IC1 water system, including electrical energy consumers and further added values

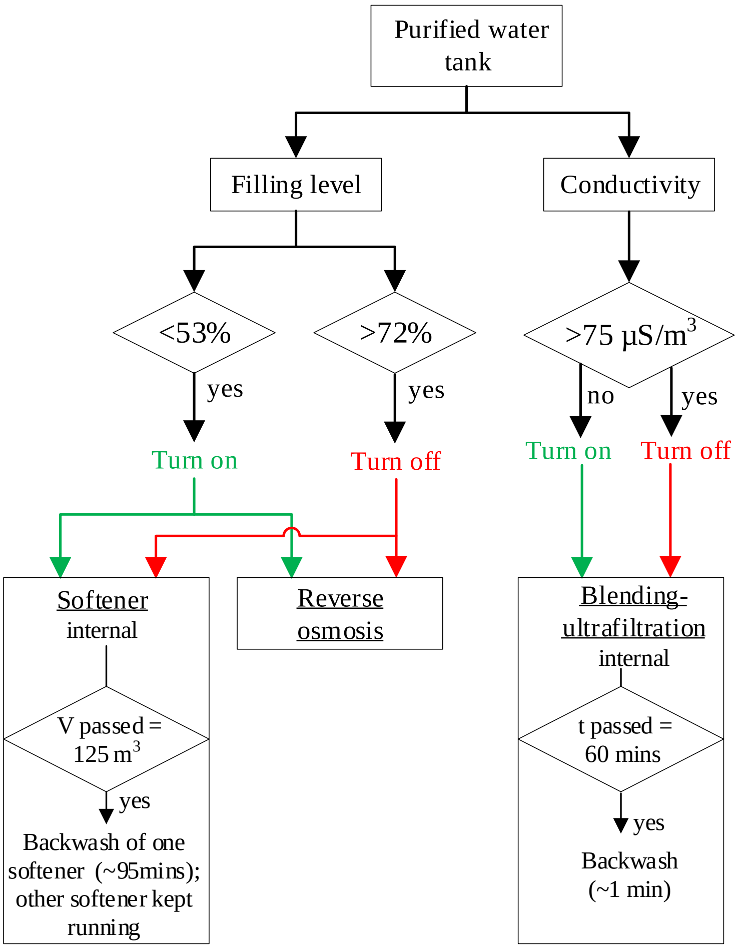


Figure D.6: Process control layout for the IC1 water system

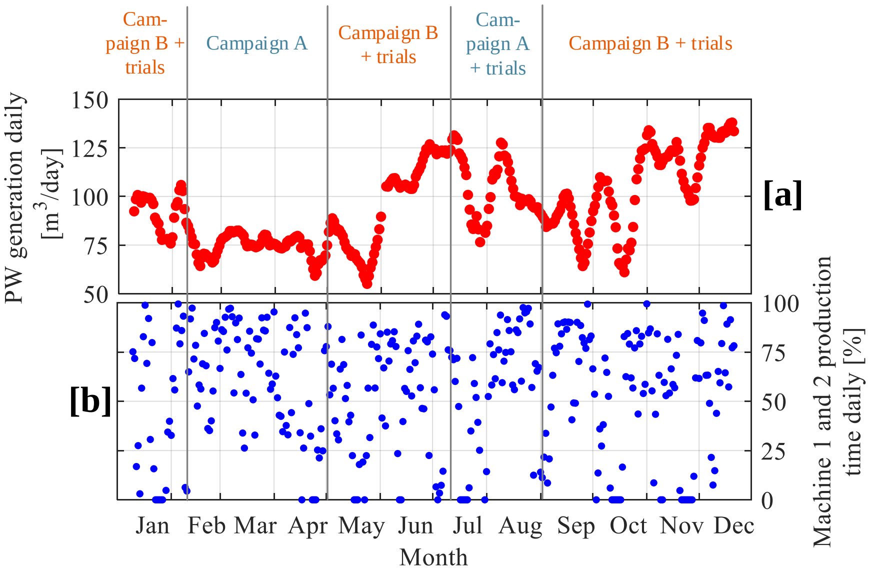


Figure D.7: IC2 PW generation vs. machine production activity including production campaigns: daily generated PW (estimated by PMD from section 3.3.1) [a], production activity from machine 1 and 2 [b]

D.2 Additional figure for the simplified proxy metering approach

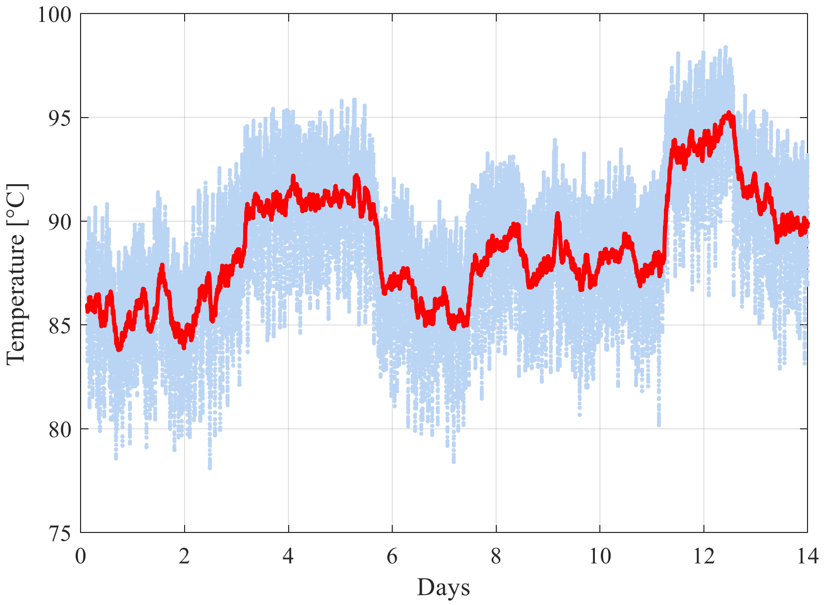


Figure D.8: Condensate tank temperature of the case facility’s steam system: light blue: condensate tank temperature with a sample interval of 30 seconds, red: moving average filter for condensate tank temperature

D.3 Additional figures for small dataset PMD performances

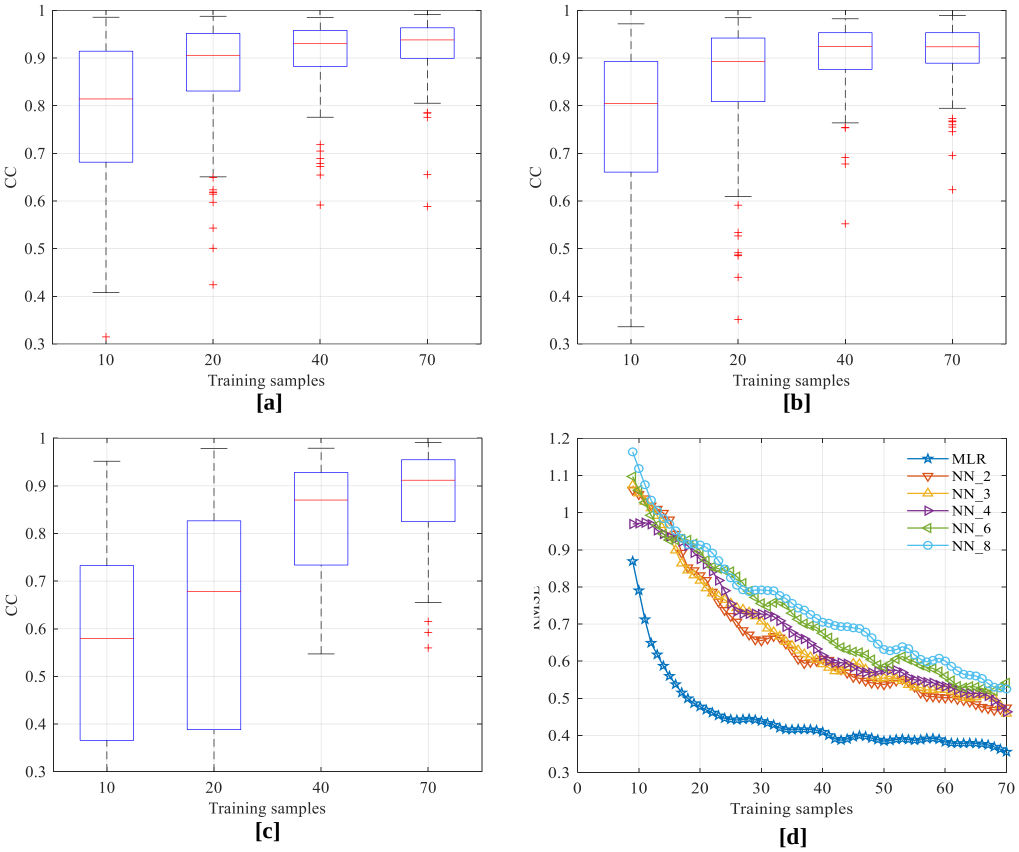


Figure D.9: Setup A: CC variation and NN hidden neurons, all input variables: box plots for the performance variation of algorithms MLR [a], PLSR_6 [b], and NN_2 [c], NN average performance for variety of hidden neurons [d]

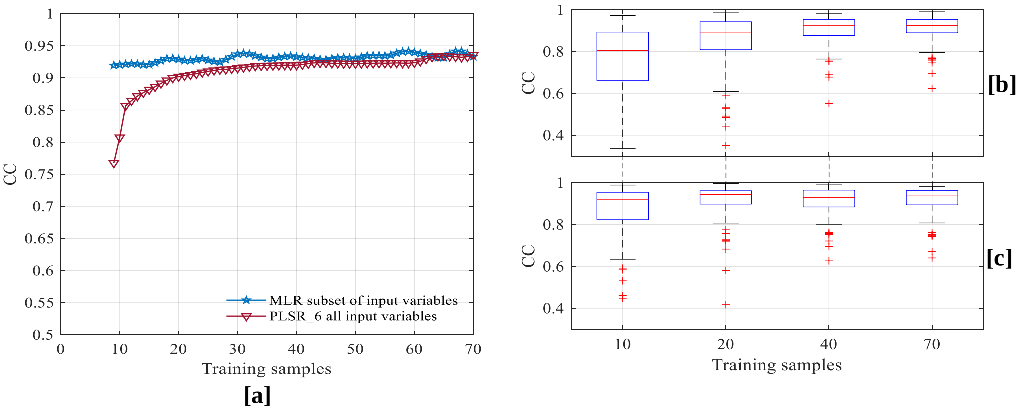


Figure D.10: Setup A: CC all input variables and reduced input variable subset: average performance [a], performance variation for PLSR_6 trained on all input variables [b], performance variation for MLR trained on most accurate input variable subset [c]

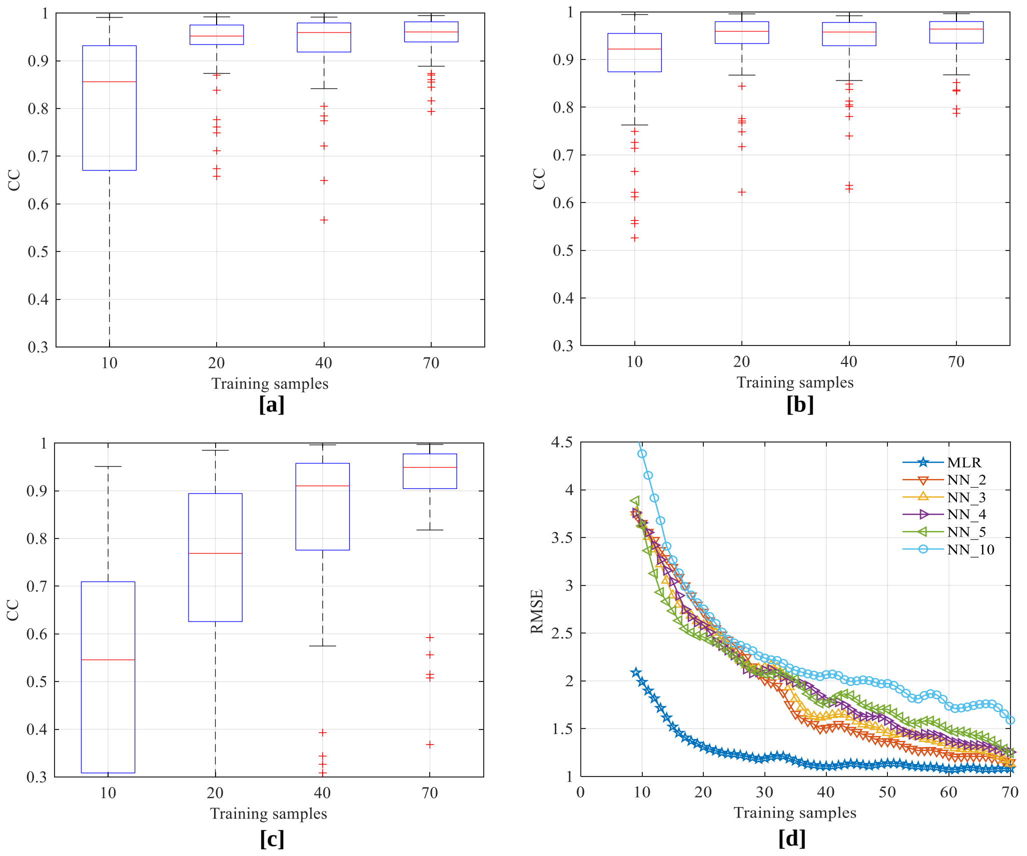


Figure D.11: Setup B: CC variation and NN hidden neurons, all input variables: box plots for the performance variations of MLR [a], PLSR_5 [b], and NN_2 [c]; NN average performance for a variety of hidden neurons [d]

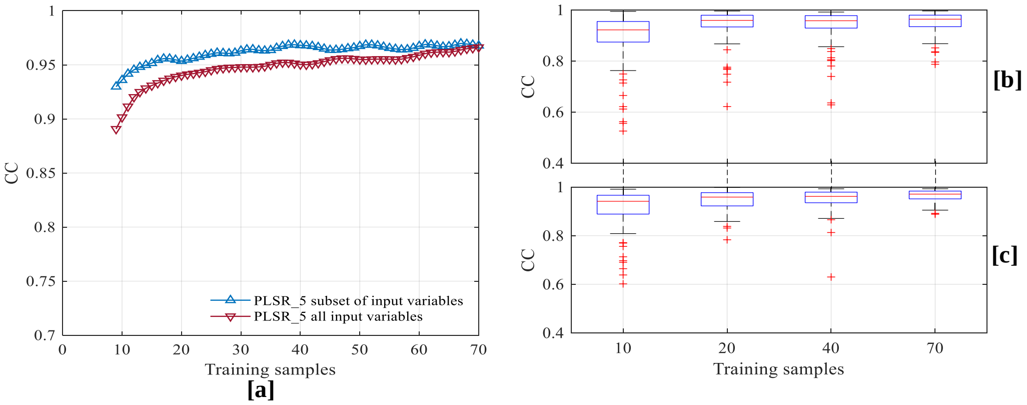


Figure D.12: Setup B: CC all input variables and reduced input variable subset: average performance [a], performance variation for PLSR_5 trained on all input variables [b], performance variation for PLSR_5 trained on most accurate input variable subset [c]

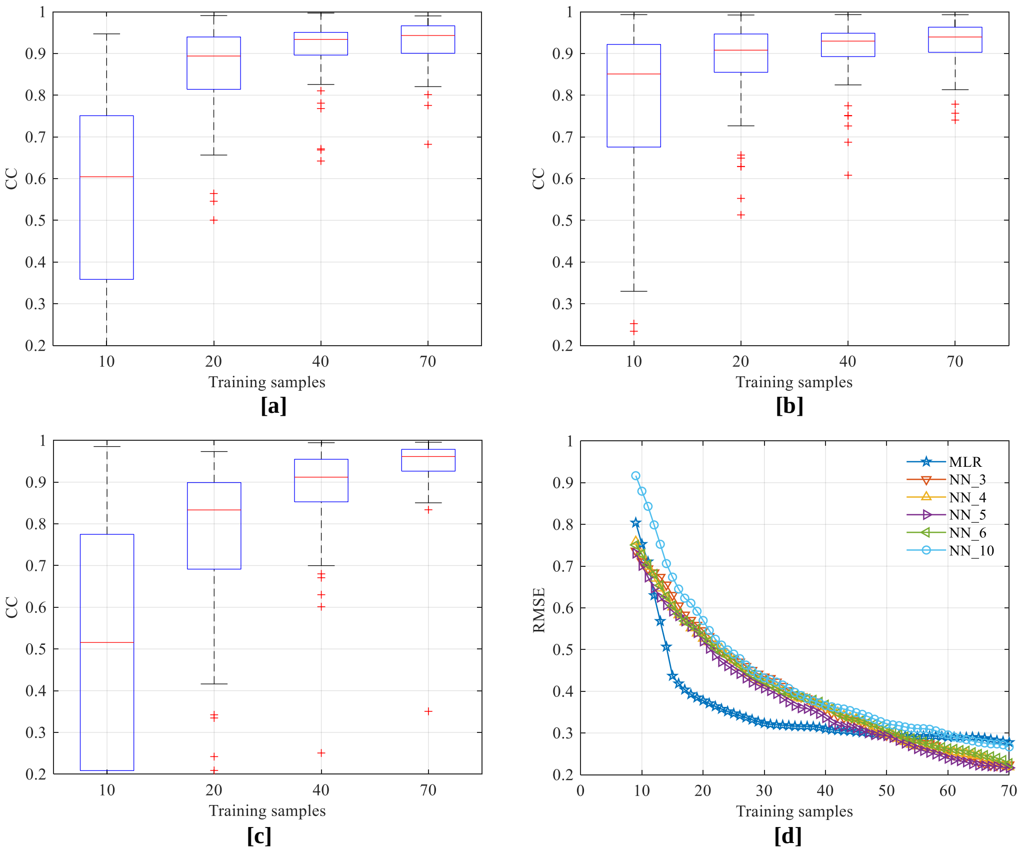


Figure D.13: Setup C: CC variation and NN hidden neurons, all input variables: box plots for the performance variation of algorithms MLR [a], PLSR_5 [b], and NN_5 [c], NN average performance for variety of hidden neurons [d]

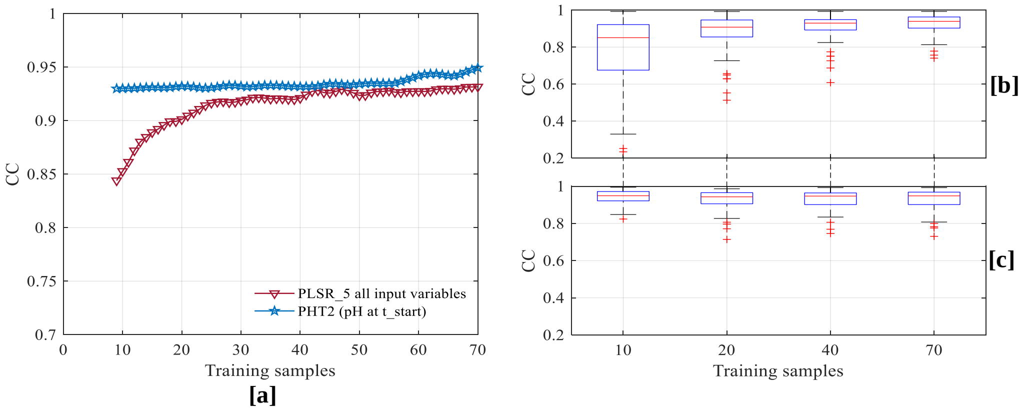


Figure D.14: Setup C: CC all input variables and reduced input variable subset: average performance [a], performance variation for PLSR_5 trained on all input variables [b], performance variation for SLR with pHT2 (pH at t_start) [c]

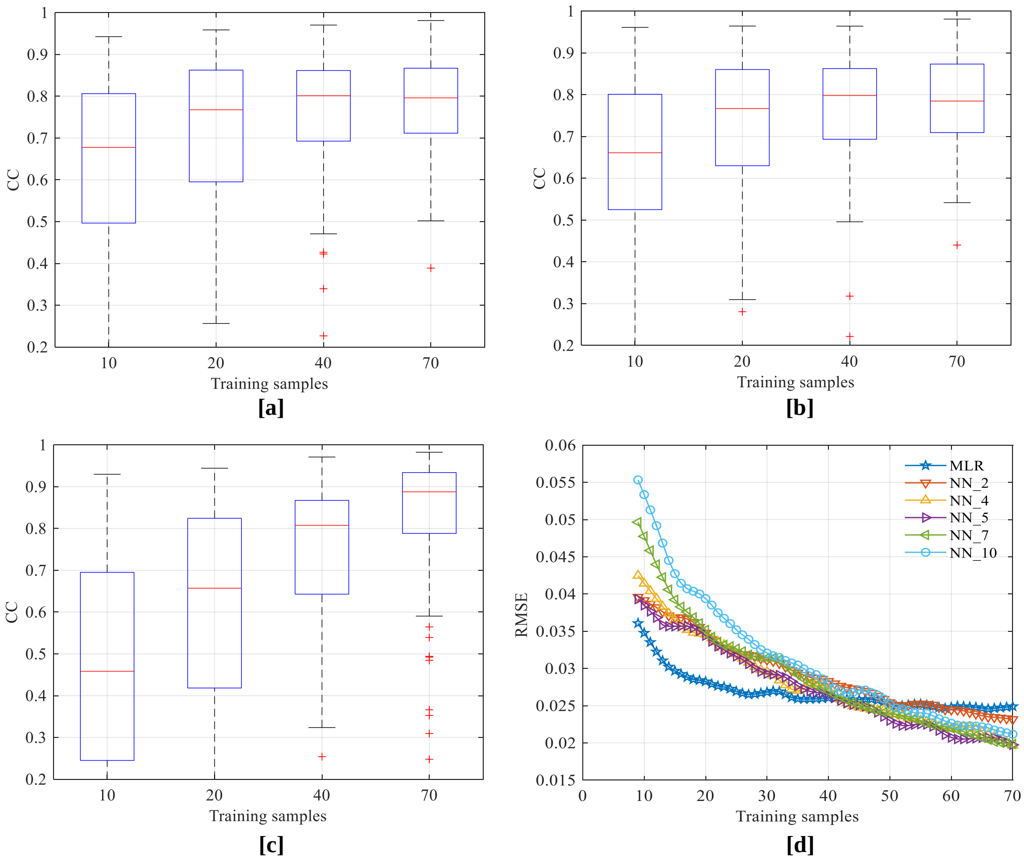


Figure D.15: Setup D: CC variation and NN hidden neurons, all input variables: box plots for the performance variation of algorithms MLR [a], PLSR_4 [b], and NN_5 [c], NN average performance for variety of hidden neurons [d]

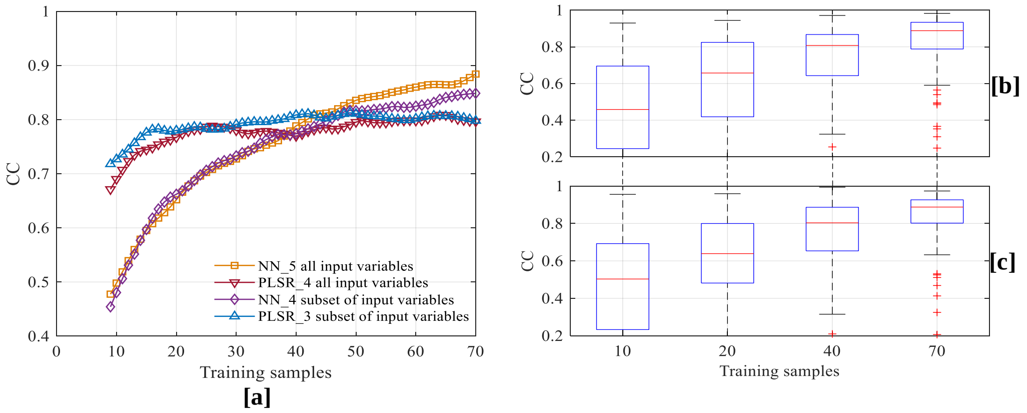


Figure D.16: Setup D: CC all input variables and reduced input variable subset: average performance [a], performance variation for NN_5 trained on all input variables [b], performance variation for NN_4 trained on most accurate input variable subset [c]

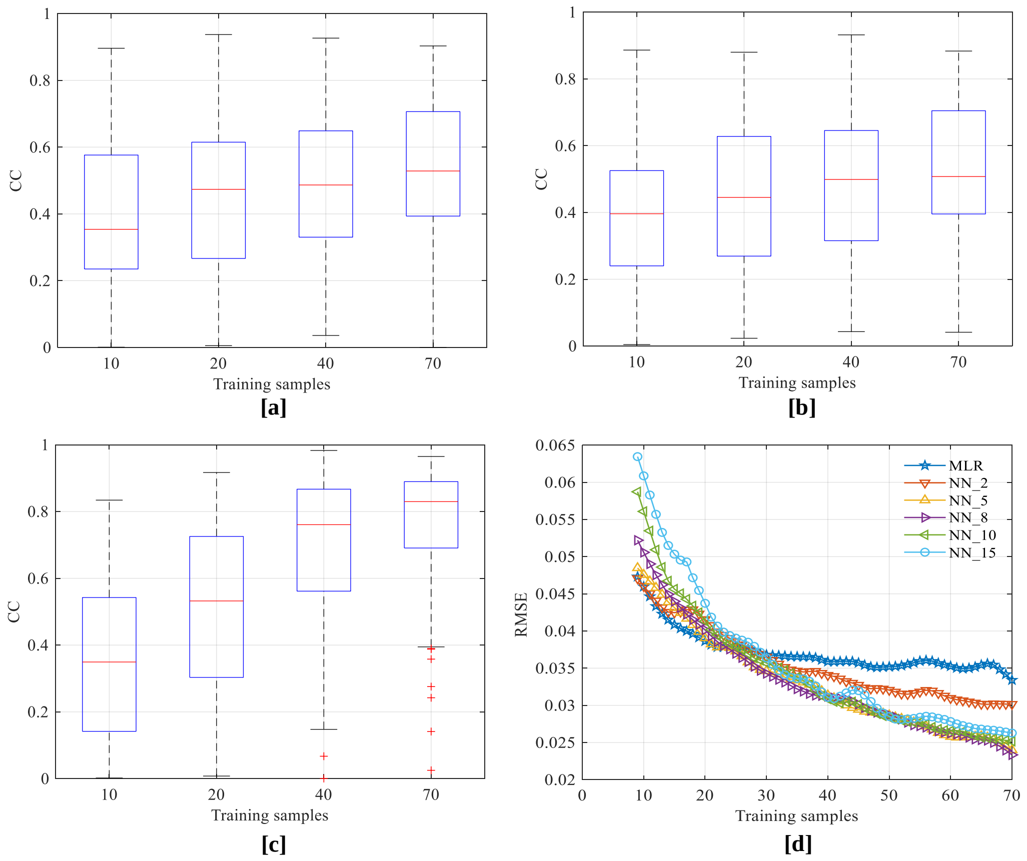


Figure D.17: Setup E: CC variation and NN hidden neurons, all input variables: box plots for the performance variation of MLR [a], PLSR_3 [b], and NN_5 [c]; NN average performance for variety of hidden neurons [d]

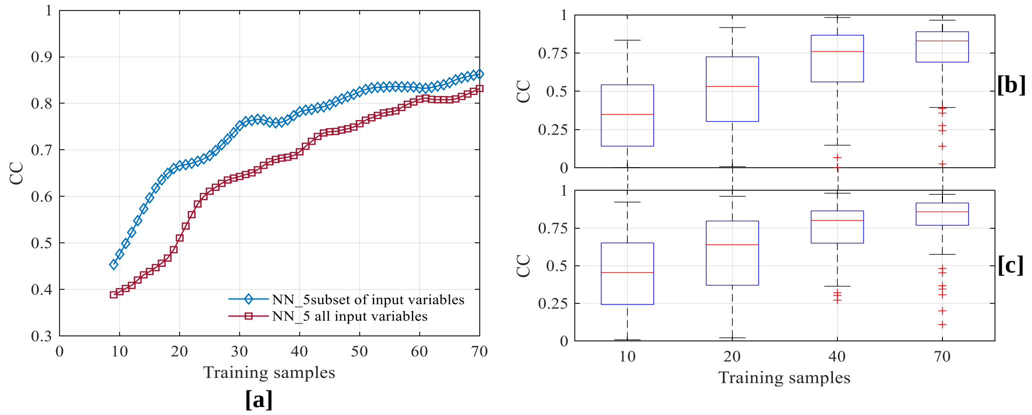


Figure D.18: Setup E: CC all input variables and reduced input variable subset: average performance [a], performance variation for NN_5 trained on all input variables [b], performance variation for NN_5 trained on most accurate input variable subset [c]

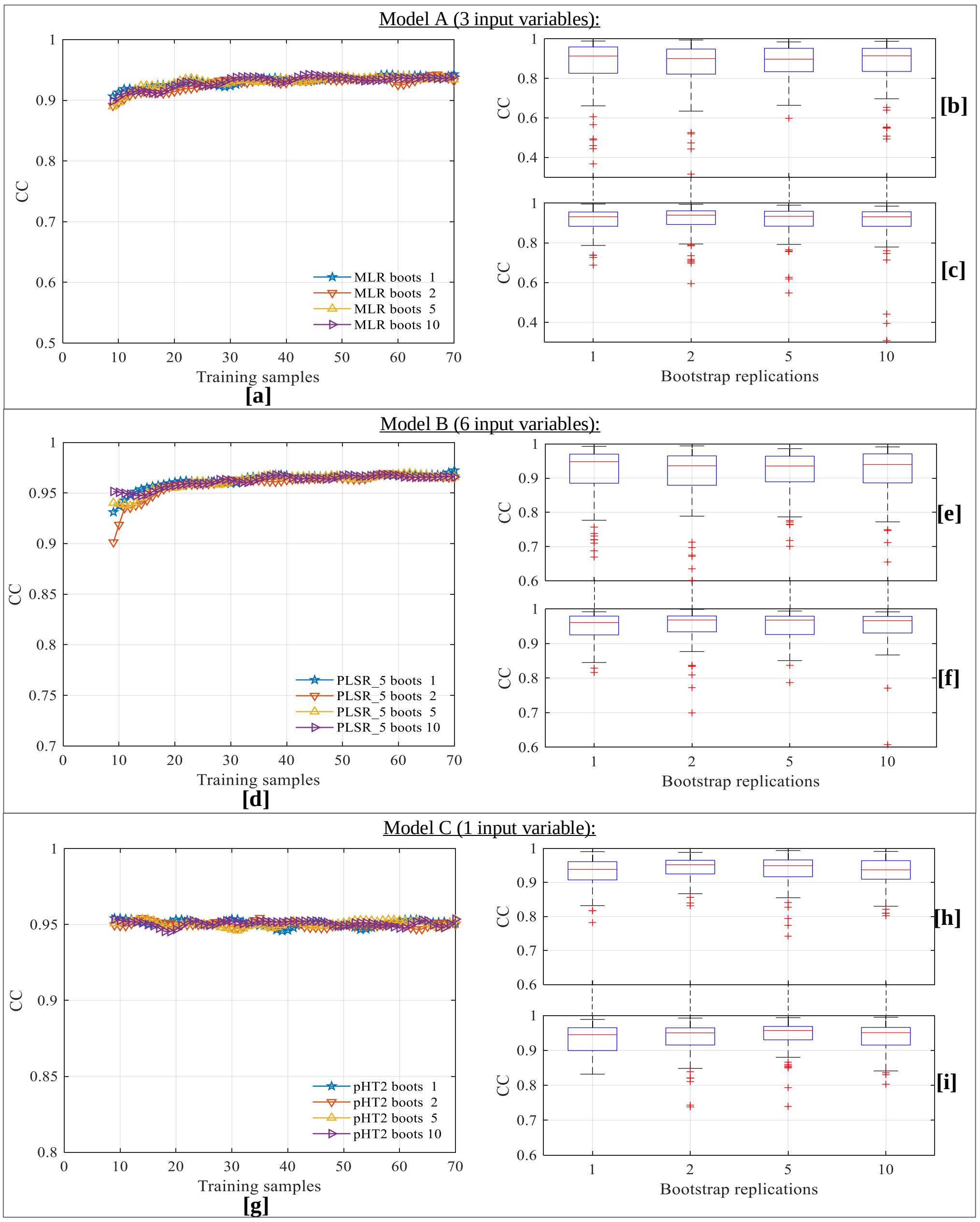


Figure D.19: Setups A, B, C: BR influence on MLR, PLSR, and SLR; CC validation: average performances [a], [d], [g], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup B PLSR_5 performance variation for training samples 10 [e] and 40 [f], setup C SLR performance variation for training samples 10 [h] and 40 [i]

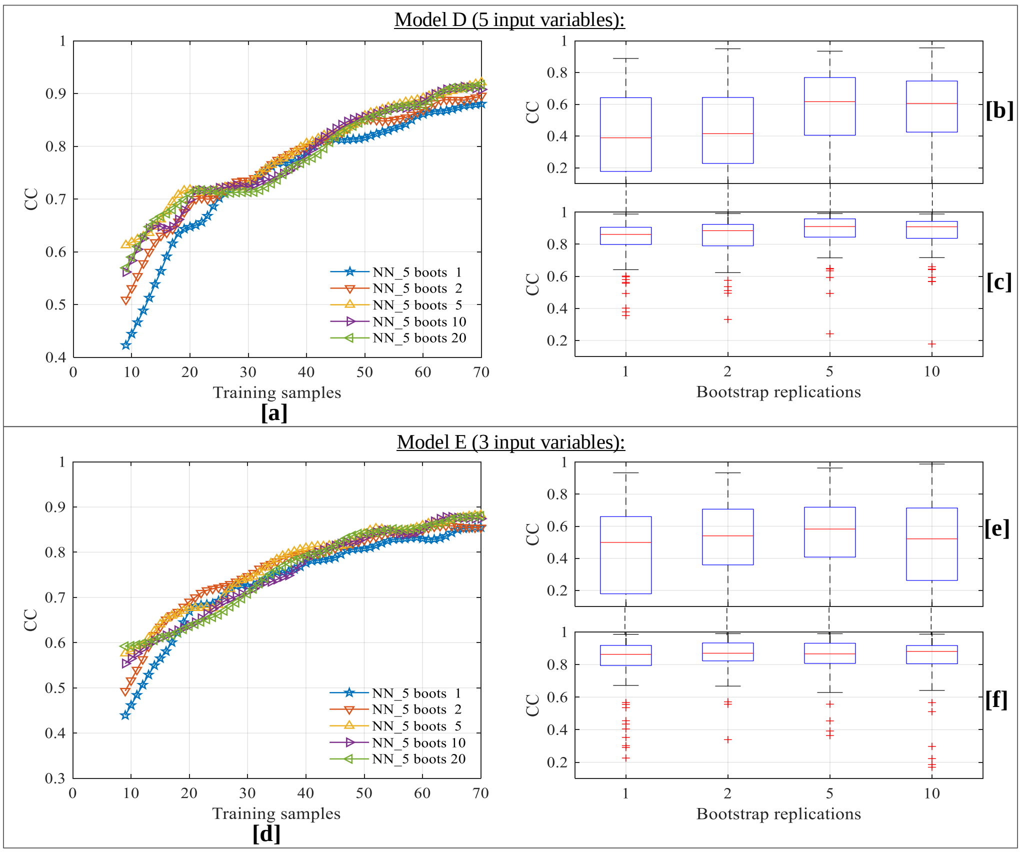


Figure D.20: Setups D, E: BR influence on NN; CC validation: average performances [a], [d], setup D NN_5 performance variation for training samples 10 [b] and 70 [c], setup E NN_5 performance variation for training samples 10 [e] and 70 [f]

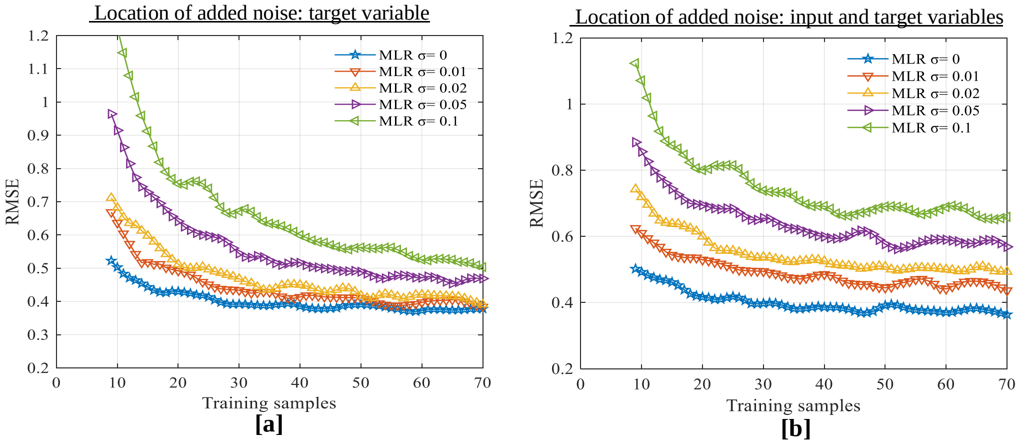
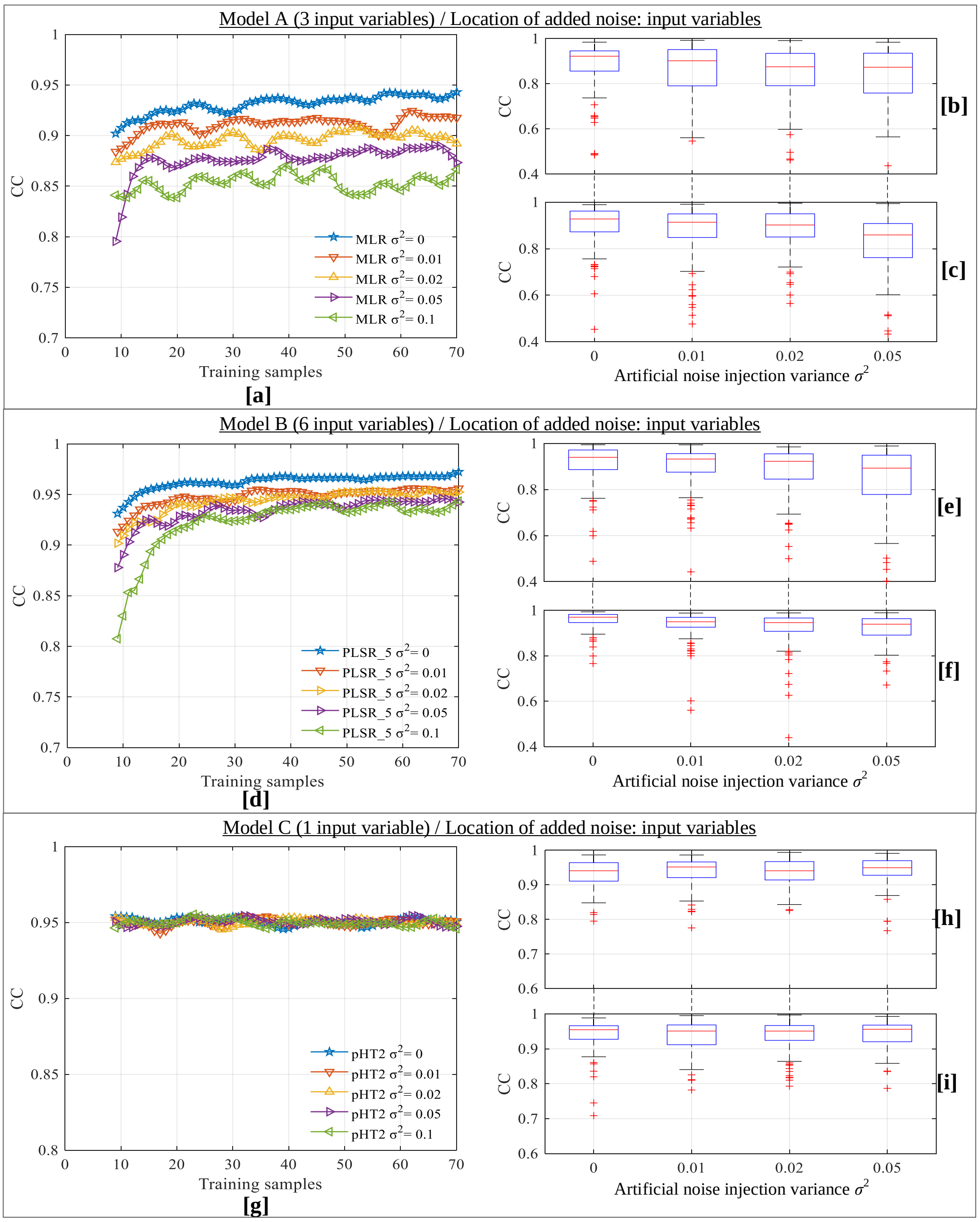


Figure D.21: Setup A: Influence of ANI location: RMSE average performances for artificial noise injected to target variable [a] or input and target variables [b]



Model A (3 input variables) / Location of added noise: input variables

Model B (6 input variables) / Location of added noise: input variables

Model C (1 input variable) / Location of added noise: input variables

Figure D.22: Setups A, B, C: ANI influence on MLR, PLSR, and SLR; CC validation: CC average performances [a], [d], [g], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup B PLSR_5 performance variation for training samples 10 [e] and 40 [f], setup C SLR performance variation for training samples 10 [h] and 40 [i]

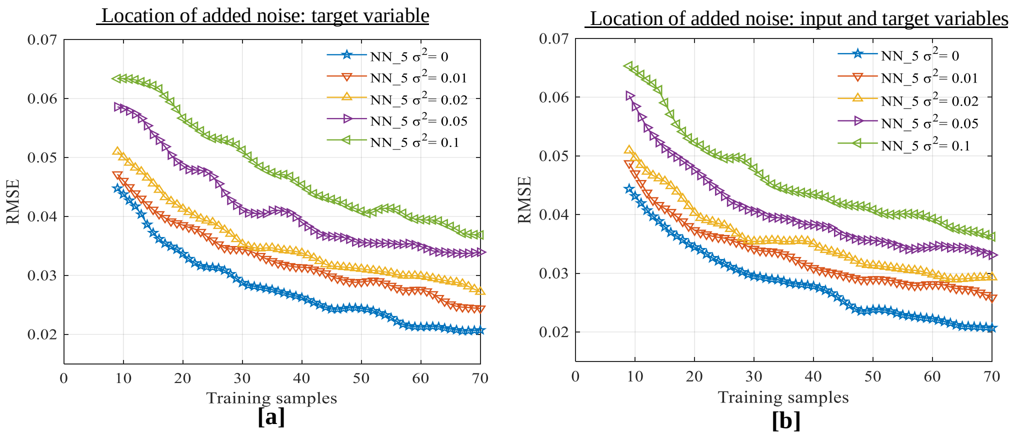


Figure D.23: Setup D: Influence of ANI location: RMSE average performances for artificial noise injected to target variable [a] or input and target variables [b]

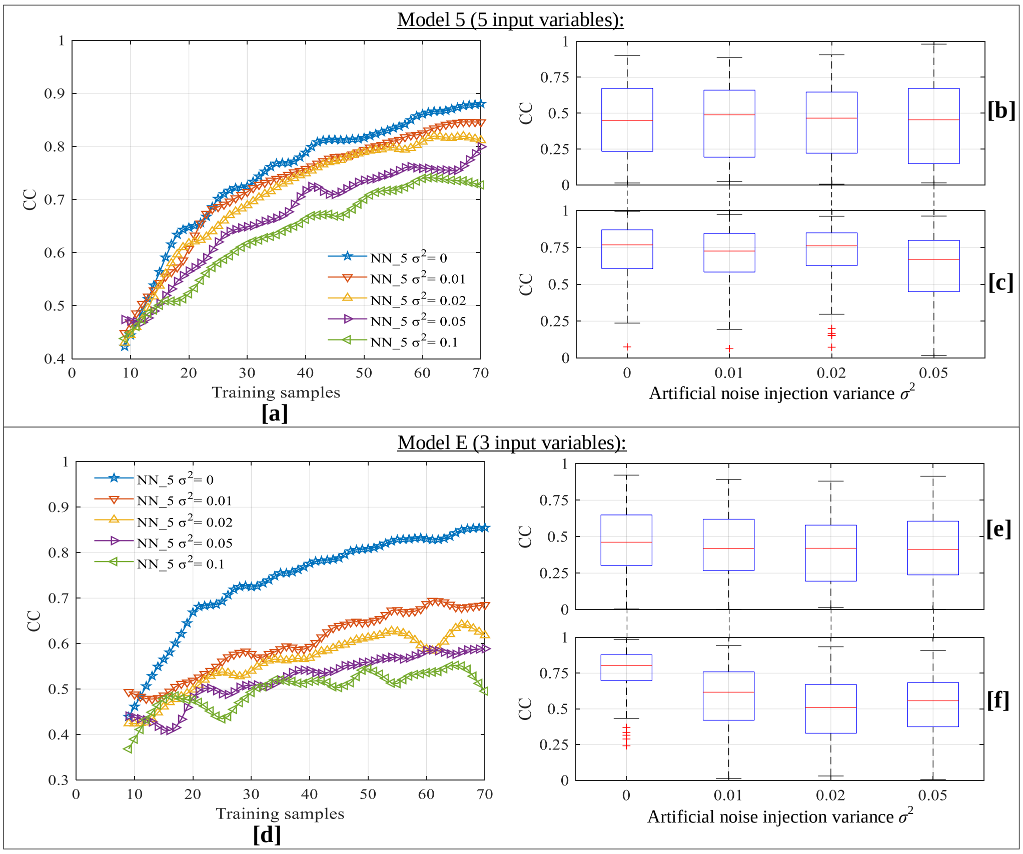


Figure D.24: Setups D, E: ANI influence on NN; CC validation: average performances [a], [d], setup D NN_5 performance variation for training samples 10 [b] and 40 [c], setup E NN_5 performance variation for training samples 10 [e] and 40 [f]

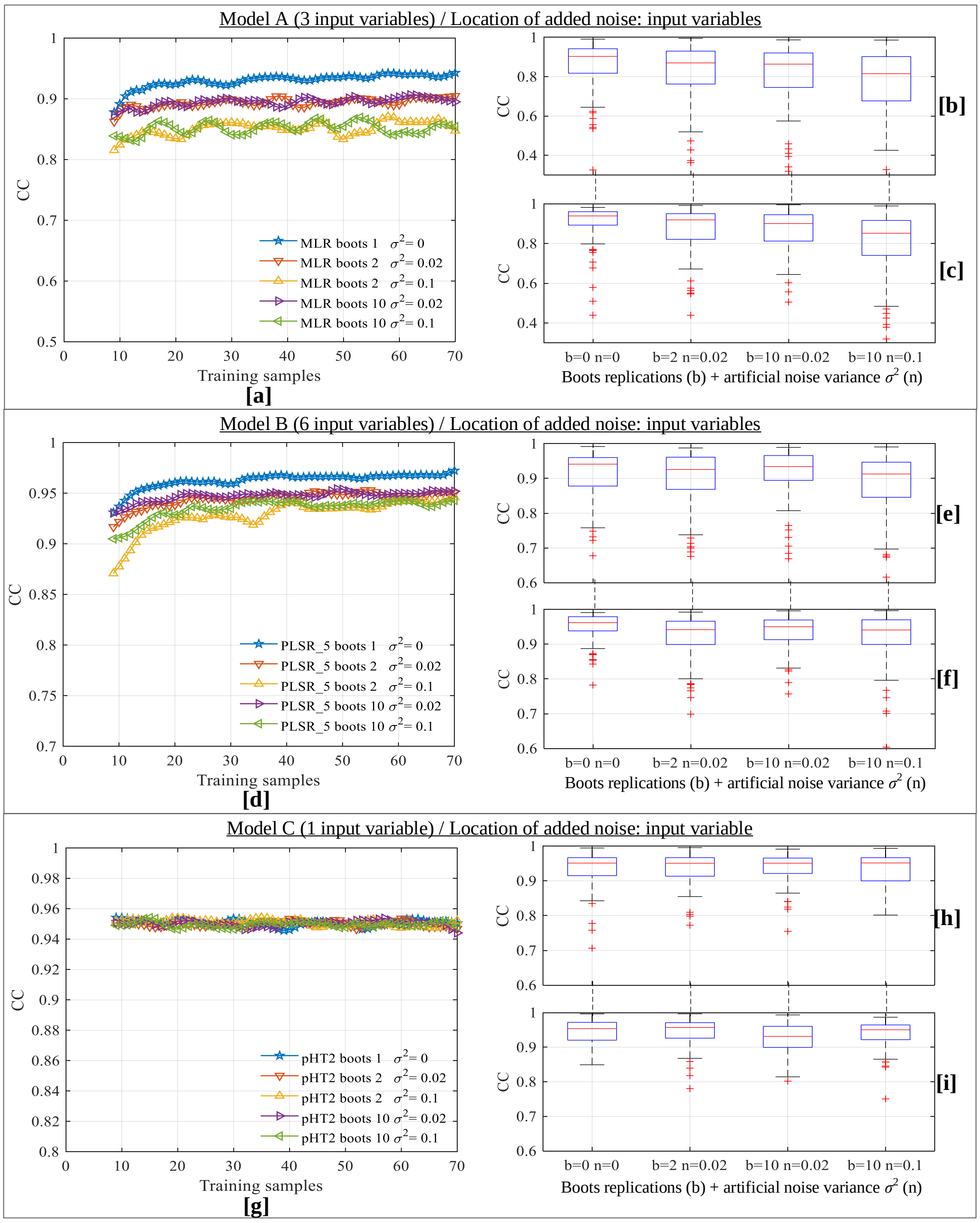


Figure D.25: Setups A, B, C: BR and ANI influence on MLR, PLSR, and SLR; CC validation: average performances [a], [d], [g], setup A MLR performance variation for training samples 10 [b] and 40 [c], setup B PLSR_5 performance variation for training samples 10 [e] and 40 [f], setup C SLR performance variation for training samples 10 [h] and 40 [i]

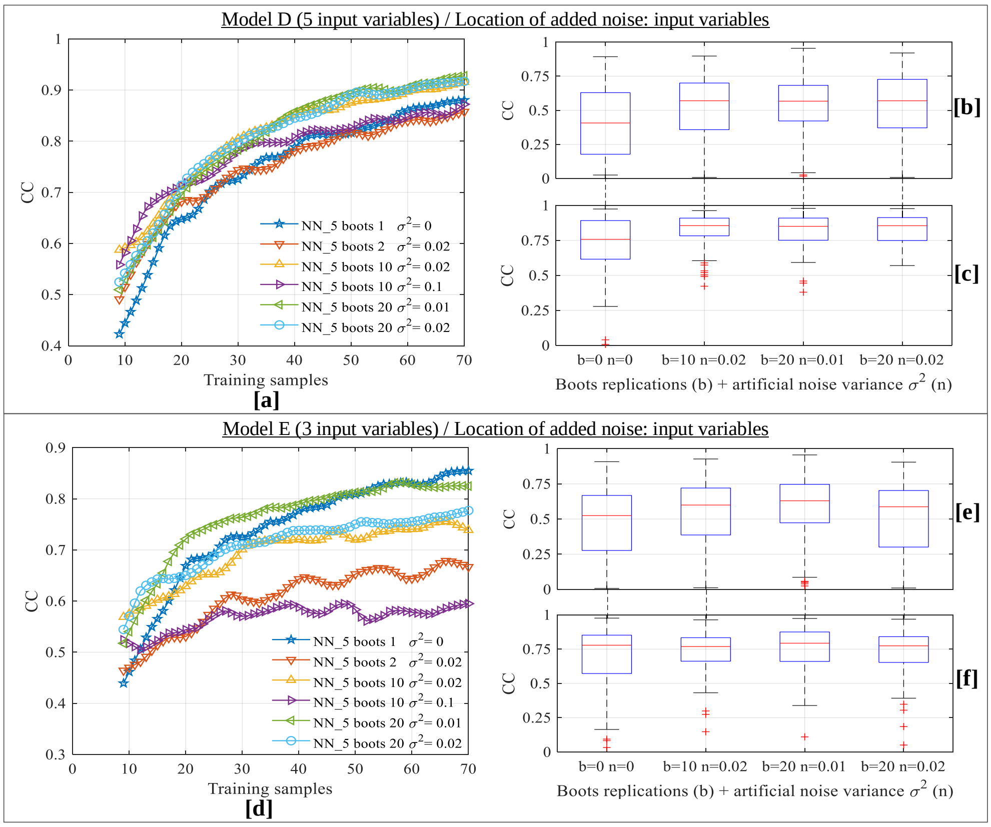


Figure D.26: Setups D, E: BR and ANI influence on NN; CC validation: average performances [a], [d], setup D NN_5 performance variation for training samples 10 [b] and 40 [c], setup E NN_5 performance variation for training samples 10 [e] and 40 [f]