FEDERAL UNIVERSITY OF CEARÁ

CENTER OF TECHNOLOGY

DEPARTMENT OF TELEINFORMATICS

UNDERGRADUATE COURSE IN COMPUTER ENGINEERING

CATHERINE BEZERRA MARKERT

MODELLING OF GREENHOUSE GAS EMISSIONS IN WASTEWATER TREATMENT PLANTS: A DATA-DRIVEN APPROACH

FORTALEZA

2025

CATHERINE BEZERRA MARKERT

MODELLING OF GREENHOUSE GAS EMISSIONS IN WASTEWATER TREATMENT PLANTS: A DATA-DRIVEN APPROACH

Undergraduate Thesis submitted to the Computer Engineering Course of the Center of Technology of the Federal University of Ceará, as a partial requirement for obtaining the Bachelor Degree in Computer Engineering.

Advisor: Prof. Dr. Michela Mulas

FORTALEZA
2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M297m Markert, Catherine Bezerra.
Modelling of greenhouse gas emissions in wastewater treatment plants : a data-driven approach / Catherine Bezerra Markert. – 2025.
149 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia de Computação, Fortaleza, 2025.
Orientação: Profa. Dra. Michela Mulas.

1. Greenhouse gas emissions. 2. Wastewater treatment plants. 3. Data-driven modelling. 4. Rolling window. I. Título.

CDD 621.39

CATHERINE BEZERRA MARKERT

MODELLING OF GREENHOUSE GAS EMISSIONS IN WASTEWATER TREATMENT PLANTS: A DATA-DRIVEN APPROACH

Undergraduate Thesis submitted to the Computer Engineering Course of the Center of Technology of the Federal University of Ceará, as a partial requirement for obtaining the Bachelor Degree in Computer Engineering.

Approved on: 20/01/2025

EXAMINATION BOARD

Prof. Dr. Michela Mulas (Advisor)
Universidade Federal do Ceará - UFC

Prof. Dr. Guilherme de Alencar Barreto
Universidade Federal do Ceará Ceará - UFC

Prof. Dr. George André Pereira Thé
Universidade Federal do Ceará Ceará - UFC

À minha mãe, Rejane Bezerra, por todo o apoio. Obrigada por sempre me passar confiança, segurança e por mostrar que nunca estive sozinha em minha caminhada.

ACKNOWLEDGEMENTS

À minha mãe, Rejane Bezerra, por todo amor e dedicação e por sempre acreditar em mim. Ao meu pai, Werner Markert, que sempre me fez sentir querida e amada.

Às amigas de ensino médio, Rafaela e Lia, por todos os momentos divertidos que vivemos ao longo desses quase 8 anos de amizade, por todo o apoio que me deram ao longo da vida e por sempre mostrarem curiosidade sobre a área da Computação (e ouvirem longos monólogos sobre).

Aos amigos da turma de 2020: Davi e Yago, que confiaram em mim para compor o trio da maratona de programação e de muitos outros trabalhos. Vocês me ensinaram muito sobre amizade e companheirismo e sou muito feliz de tê-los como amigos. Danilo, Matheus e Vitor, do P&P, meu primeiro grupo de trabalho na faculdade. Lembrarei com carinho os momentos maravilhosos que tivemos. A todos do grupo Ohana, pela parceria nesses cinco anos de graduação, da pandemia até a formatura. Sem vocês, minha caminhada certamente seria muito mais árdua.

Às amigas Bruna, Gabriela Ponciano e Yohanne, por terem me acolhido e por serem amigas tão especiais. À Rebeca, por ter me ensinado muito sobre lutar para alcançar objetivos e por confiar em mim. Às minhas "filhas" do ADOTE, Gabriela Bezerra, uma das pessoas mais incríveis que conheço, e Lorena, que me cativou muito com sua simpatia. À Byanca, Caroline, Cecília, Emanuele e Vitória, por serem mulheres fortes e me inspirarem a seguir na engenharia.

Aos amigos que tive a alegria de me aproximar nesses 5 anos de caminhada: Lucas Rocha, Pedro Falcão, Kauã, Samuel, Cícero, Orlando, Vinícius, por serem um ombro amigo quando precisei. Aos colegas de pesquisa da Room 2, pelo apoio em projetos e por permitirem termos bons momentos de descontração.

À minha orientadora, Michela Mulas, por confiar em mim em diversas monitorias e projetos, e por me orientar no desenvolvimento e na escrita desse TCC. À banca avaliadora, professores Guilherme Barreto e George Thé, por terem me ensinado tanto em sala de aula e fora dela. À professora Atlands Rocha, por ter me dado a minha primeira oportunidade de Iniciação Científica e me ensinar muito com sua dedicação.

Ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

“Ninguém se forma sozinho.”

(Autor Desconhecido)

ABSTRACT

With the impacts of global warming on our planet, strategies are developed to increase sustainability by reducing greenhouse gas (GHG) emissions. While these gases are essential to keep the Earth's warmth and support life, too much of them are emitted due to human activity, causing disruptive changes to the planet's climate. Although having a smaller contribution compared to other sectors such as energy, transportation or agriculture, sewage treatment has a notable impact on GHG emissions, generating a range of gases including methane (CH_4), carbon dioxide (CO_2), and nitrous oxide (N_2O). Data-driven approaches are employed to understand more of the GHG emissions' behaviour and to develop machine learning models that can predict these emissions and perform control actions in the plant to reduce them. This work explores data-driven techniques to analyse and model GHG emissions in a Wastewater Treatment Plant (WWTP) located in Helsinki, Finland. The research follows three key steps: data analysis, pre-processing, and modelling. First, a detailed analysis of the dataset enhances process understanding and identifies how variations in predictor variables influence GHG emissions. Next, pre-processing techniques are applied to improve the modelling process. Finally, machine learning models - both linear and non-linear - are developed to predict nitrous oxide (N_2O) and carbon dioxide (CO_2) emissions. Results demonstrate that employing rolling window approaches, even with simple models, generally improve model accuracy, offering promising perspectives for emission modelling in WWTPs.

Keywords: Greenhouse gases emissions; Wastewater Treatment Plants; Data-driven modelling; Rolling Window.

RESUMO

Com os impactos do aquecimento global em nosso planeta, estratégias são desenvolvidas para aumentar a sustentabilidade, como a redução das emissões de gases de efeito estufa (GEE). Enquanto tais gases são essenciais para manter a temperatura da Terra e auxiliar na manutenção da vida, eles são emitidos em excesso pela atividade humana, causando mudanças disruptivas ao clima do planeta. Embora possuam uma contribuição menor comparada a outros setores como energia, transporte e agricultura, o tratamento de esgoto possível um impacto notório nas emissões de GEEs, gerando uma quantidade de gases que incluem metano (CH_4), dióxido de carbono (CO_2) e óxido nitroso (N_2O). Abordagens orientadas a dados são empregadas para compreender melhor o comportamento das emissões de GEE e desenvolver modelos de aprendizado de máquina capazes de predizer essas emissões e de realizar ações de controle na planta para reduzi-las. Este trabalho explora abordagens baseadas em dados para analisar e modelar as emissões de GEE em uma Estação de Tratamento de Esgoto (ETE) localizada em Helsinque, Finlândia. A pesquisa aborda três etapas principais: análise de dados, pré-processamento e modelagem. Inicialmente, uma análise detalhada do conjunto de dados melhora a compreensão do processo e identifica como variações nas variáveis preditoras influenciam as emissões de GEE. Em seguida, técnicas de pré-processamento são aplicadas para aprimorar o processo de modelagem. Por fim, modelos de aprendizado de máquina - tanto lineares quanto não lineares - são desenvolvidos para prever as emissões de óxido nitroso (N_2O) e dióxido de carbono (CO_2). Os resultados demonstram que o uso de abordagens de Janela Móvel (Rolling Window, em inglês), geralmente melhoram a precisão dos modelos, oferecendo perspectivas promissoras para a modelagem de emissões em ETEs.

Palavras-chave: Emissões de gases de efeito estufa. Estações de Tratamento de esgoto. Modelagem orientada a dados. Janela Móvel.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

R^2	Determination Coefficient
AD	Activity Data
AdaBoost	Adaptative Boosting
ANN	Artificial Neural Network
ASM1	Activated Sludge Model no. 1
ASM2d	Activated Sludge Model no. 2d
ASM3	Activated Sludge Model no. 3
ASP	Activated Sludge Process
CO_2	Carbon Dioxide
DNN	Deep Neural Networks
EF	Emission Factor
EU	European Union
FD	Full dataset
GHG	Greenhouse Gas
GWP	Global Warming Potential
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LOWESS	Locally Weighted Scatterplot Smoothing
LSTM	Long Short-Term Memory
MARS	Multivariate Adaptive Regression Spline
mRMR	Minimum Redundancy - Maximum Relevance
MW	Moving Window
N_2O	Nitrous Oxide
OLS	Ordinary Least Squares Regression
PCA	Principal Component Analysis
PLS	Partial Least Squares
RD	Reduced dataet
RMSE	Root Mean Squared Error
RSS	Residuals Sum of Squares
RW	Rolling Window
SVD	Single Value Decomposition

SVM	Support Vector Machine
UN	United Nations
VIF	Variance Inflation Factor
WWTP	Wastewater Treatment Plant
XGBoost	Extreme Gradient Boosting

LIST OF SYMBOLS

$\bar{x}$	Mean
S^2	Variance
σ	Standard deviation
$\tilde{x}$	Median
ρ	Pearson Correlation Coefficient
ϕ	Loading parameter
β	Tuning parameter to the robust covariance matrix weight parameter
T^2	Hotelling's statistic
$\hat{\beta}$	Model parameters for the Ordinary Least Squares Regression
λ	Shrinking coefficient
δ	Huber function threshold
α_0	Extreme Gradient Boosting Initial prediction
$w_{q(x_i,t)}$	Extreme Gradient Boosting eaves scores' vector
$\mathbb{I}$	Indicator function
η	Learning rate

CONTENTS

1 INTRODUCTION

This chapter introduces the modelling of Greenhouse Gas (GHG) emissions outlines, the problem addressed in this work, and highlights the potential solutions, which will be detailed in the next chapters. The following sections will discuss the development of quantification and modelling strategies applicable to sewage treatment plants, focusing on strategies specific to data-driven techniques.

1.1 Greenhouse Gases Emissions Modelling in Wastewater Treatment Plants

Global warming has become a serious issue to our planet, and many efforts are being made to reduce its impacts on modern society. One of the main challenges nowadays is reducing GHG emissions into the atmosphere. These gases, such as nitrous oxide (N_2O) and carbon dioxide (CO_2), present high Global Warming Potential (GWP) and can be emitted in various contexts, with the first having a GWP approximately 300 times higher than the second (IPCC, 2021).

Countries have adopted goals for the reduction of GHG emissions. In Brazil, for example, Law 15.042/2024, sanctioned in December 2024, established the Brazilian Greenhouse Gas Emissions Trading System, a regulated carbon market aimed at limiting GHG emissions across different sectors (BRASIL, 2024). The European Union (EU) has implemented laws that generated more investments in minimizing these emissions. EU has made a commitment of reducing its net GHG emissions by at least 55% by 2030, under the European Climate Law (European Commission, 2021).

The development of carbon-neutral infrastructure services and resources relates to the 11th and 13th United Nations (UN) Sustainment Development Goals, Sustainable Cities and Communities and Climate Action, respectively (United Nations, 2015). A sector that impacts on GHG emissions is wastewater treatment, since municipal sewage produces approximately 1.43 billion metric tonnes yearly, and is expected to increase this production by 50% by 2050 (QADIR *et al.*, 2020). In addition, N_2O emissions may account for up to 80% of a Wastewater Treatment Plant (WWTP)'s total carbon footprint (KHALIL *et al.*, 2023).

Given that only 42% of sewage water is properly managed and/or reused (United Nations World Water Assessment Programme, 2017), significant efforts have been made to expand wastewater treatment capacity. For instance, technological advancements have been made to improve the population's access to sewage treatment, such as incorporating models to monitor, control, and predict the behaviour of the treated sewer water. The reduction of GHG emissions is important to guarantee the minimization of global warming impacts and reach the goals set by countries and organizations.

The GHG emissions identified in wastewater treatment facilities can either be direct or indirect, with the first ones being related to the biological process itself and the second ones to fossil fuel and electricity consumption (DAELMAN *et al.*, 2013). Moreover, special attention has been given to the N_2O ones because of their elevated GWP and are commonly released from many sewage facilities. This gas is produced as an intermediate byproduct during the nitrification reaction process carried out by nitrifying bacteria. Its fluxes are highly dynamic and the sensitivity in its modelling is the major source of the uncertainty of the emission estimates of GHGs (VASILAKI *et al.*, 2018; LU *et al.*, 2023). Different strategies have been explored to improve its quantification and modelling.

The Emission Factor (EF) quantification of a GHG is an empirical strategy that provides a relationship to compute emissions. It is based on the product of the Activity Data (AD), that is, the usage and input values related directly to emissions of carbon in a year from an individual emission origin, by the EF, the number of GHG emitted per unit of use of a given source, and the GWP of the respective GHG, as represented in Equation 1.1 (IPCC, 2019). The method has been used to create and update urban GHG emission inventories in Brazilian cities, for instance, helping in the understanding and comparison of these emission behaviours (ANDRADE *et al.*, 2021).

$$\text{Emissions} = \text{AD} \times \text{EF} \times \text{GWP} \tag{1.1}$$

This quantification presents a high associated uncertainty, which means that estimations in the quantification are not very precise or reliable, and variability and its results are commonly oversimplified and misestimated (SONG *et al.*, 2024). Therefore, it cannot reflect the real GHG's behaviour completely, with models that use operational data from WWTPs obtaining emission results closer to the experimental data than this approach (RAMÍREZ-MELGAREJO *et al.*, 2020). For more process understanding and accurate emission prediction, modelling approaches - such as mechanistic and data-driven - have been developed.

1.1.1 Mechanistic Modelling

The mechanistic modelling of GHGs can be performed using models that vary in both dynamics and complexity. They may represent steady-state systems, focusing on stable and time-invariant conditions, or dynamic systems, which evolve over time because of their internal states and/or external inputs, for instance. Regarding complexity, simpler models are easier to interpret and analyse, but may not account for all the variables and interactions involved in GHG production. In contrast, more complex models offer a deeper understanding of process modelling but are more difficult to develop due to the physical complexities of the processes and parameter optimization necessities.

Mechanistic models of GHG emissions in WWTPs have been implemented with success, for instance, in BANI SHAHABADI *et al.* (2010), where a comprehensive mathematical model was developed considering on-site and off-site activities, helping to acknowledge the contribution of individual processes to the GHG emissions. In Gulhan *et al.* (2023), where the Activated Sludge Model no. 1 (ASM1) was modified in a plant-wide approach for the WWTP of Corleone WWTP (Italy) to model N_2O emissions. Furthermore, Maktabifard *et al.* (2022) compared a variation of the Activated Sludge Model no. 2d (ASM2d) and the Activated Sludge Model no. 3 (ASM3) models, with the latter showing improved results and performing accurate predictions of the emissions.

These models are sensitive to significant changes in sewage treatment operations and require frequent updates to remain accurate (BAHRAMIAN *et al.*, 2023; HUANG *et al.*, 2024). Moreover, the development of such models requires a deep understanding of plant processes, as parameter calibration and mechanistic expressions are derived from this acquired knowledge. Consequently, the need for models capable of extracting insights from collected data to enhance process knowledge and adapting more robustly to changes in the WWTP's operation has powered the adoption of data-driven methods.

1.1.2 Data-driven Modelling

Data-driven models use machine learning resources to develop predictive models based on collected data from WWTPs. They present an optimized approach considering mechanistic models, which are more complex to develop. These models exploit information collected from sensors of real-time monitored variables, and, with multivariate statistical techniques, identify relationships between predictors and GHGs, such as N_2O (VASILAKI *et al.*, 2018). Also, employed pre-processing techniques can detect anomalous or faulty behaviours in processes (KAZEMI *et al.*, 2021) and features that are more important to the modelling approach (KHALIL *et al.*, 2024).

With data-driven modelling, the process understanding can be improved, as shown in Hwangbo *et al.* (2020), who propose a data-driven framework for process modelling based on deep-learning techniques which assisted in comprehending the plant-wide operation. They could understand the process and how biological N_2O was produced by joining a Monte-Carlo simulation with the easyGSA tool (AL *et al.*, 2019) using a deep-learning-based model determined by model discrimination.

To accurately predict N_2O emissions, the collection and pre-processing of data and the prediction of the emissions should be more robust, as data collected by real-time sensors can fail on storing samples or present discrepant values that may affect model performance. Therefore, the used data should be treated before developing the models. In addition, Haimi (2016), Kazemi *et al.* (2020), Kazemi *et al.* (2021) present that eventual performance failures could be identified by fault-detection techniques such as Support Vector Machine (SVM)s, PCA, Incremental PCA and Moving Window PCA. Furthermore, pre-processing techniques, such as Feature Selection, produce a less redundant and computationally complex dataset (KHALIL *et al.*, 2023), being important for the improvement of GHG modelling results.

When referring to data-driven models, not only the results, but also model complexity, computational cost, performance, and interpretability should be taken into account (MEHRANI *et al.*, 2022). Depending on these considerations, different machine-learning strategies can be employed, being linear like Robust (WANG *et al.*, 2023), or non-linear such as Multivariate Adaptive Regression Spline (MARS), SVM, and Extreme Gradient Boosting (XGBoost) (SZELĄG *et al.*, 2023). Regarding N_2O emissions, linear models tend to present an effective performance when sliding-layer models are used (WANG *et al.*, 2023), reaching a Determination Coefficient (R^2) - a coefficient that calculates how fit the developed model is to the data and ranges from 0 to 1 - of 0.7028 when the Bayesian Regression is used, but non-linear models tend to present the best results, such as the Light Gradient Boosting Machine, which had an R^2 of 0.73 in Wang *et al.* (2023), slightly outperforming the sliding-layer Bayesian Regression.

In addition, depending on the analysed features and performed pre-processing techniques, results can be improved. Khalil *et al.* (2023) achieved a R^2 of 0.88 with the non-linear

KNN model when the Minimum Redundancy - Maximum Relevance (mRMR) feature selection was employed and 0.95 when the ensemble method Adaptative Boosting (AdaBoost) was employed. The use of Deep Neural Networks (DNN) models is also employed in the modelling process, presenting reliable results in Mehrani *et al.* (2022), where an Artificial Neural Network (ANN) obtained and R^2 of 0.67 in test data in comparison to 0.06 in the SVM algorithm, and in Seshan *et al.* (2025), where Long Short-Term Memory (LSTM) units were used to forecast N_2O emissions and evaluated considering a sliding layer with a prediction horizon that ranged from 0.5 to 6 hours. Their R^2 values ranged from 0.98 when the prediction horizon was the smallest to 0.59 when the prediction horizon was the longest.

Different modelling strategies are tested to assess model performance and fit the process, with different models obtaining optimized results depending on the specific research. Strategies such as feature selection, sliding-layer and prediction horizons can improve outcomes. However, since non-linear models, especially deep-learning, are less interpretable and more costly computationally, there is a need to focus on data-driven modelling approaches that can reach acceptable results while maintaining interpretability and less computational cost, such as linear models.

1.2 Motivation

Given the contribution of the wastewater treatment sector to the emission of GHGs and the technological advancements aimed at improving the population's access to sewage treatment, aligning the optimization of these technologies with the reduction of GHGs is crucial for enhancing sustainability. One way of achieving this is by modelling the processes within WWTPs, including emissions of GHGs and the factors that produce these emissions. With the rise of different data-driven techniques, machine learning approaches are increasingly applied to understand and predict the behaviours of these gases. This is an important step in reducing GHG emissions and providing reasonable water quality after sewage treatment, since the knowledge gained by the modelling procedure can help optimize the treatment process.

Modelling of the GHG emissions using data-driven approaches can enhance our comprehension and forecasting these emissions. Techniques such as KNN, Decision Trees, Ensemble Learning models, and DNN have been employed to analyse performances, complexities, and interpretability of models for predicting N_2O emissions (KHALIL *et al.*, 2023). Data-driven modelling also allows experimentation on less computationally costly and complex models, such as the OLS Regression and its variations and the Partial Least Squares (PLS), which are easier to interpret, but not widely explored in literature.

By developing and refining these models, the control system of sewage treatment plants can be optimised to improve performance, reducing GHG emissions and ensuring that the treated water is delivered with minimal impact on global warming.

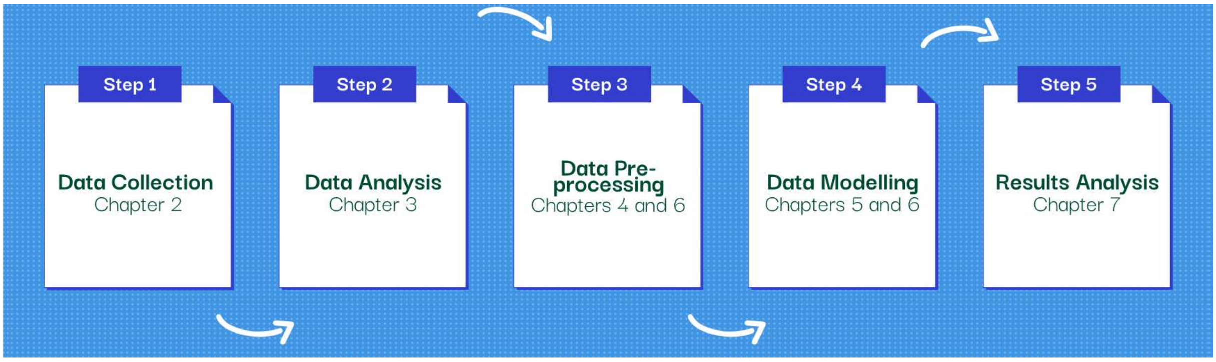
1.3 Objectives

This work's main objectives are:

1. Model the GHG emissions in a WWTP using a data-driven approach.
2. Understand the relations between the plant's features and the GHG emissions, especially the N_2O gas.
3. Understand how linear and non-linear machine learning models perform.
4. Understand if there are some characteristics of the N_2O gas, such as seasonality, that could affect the emissions' behaviour.

The work follows the framework done in Haimi (2016) and exploits a set of data provided by a WWTP located in the Viikinmäki neighbourhood, in the city of Helsinki (Finland). Since CO_2 emissions are also collected, the same algorithms chosen to model the N_2O emissions will also be tested to obtain insights and predict the CO_2 ones. Figure 1 presents the steps of the framework and in which chapters they will be discussed.

Figure 1 – Schematic of the developed framework



Source: The author.

1.4 Chapters Organization

Afther this introduction, the chapters in this document are mainly organized in two parts. The first one, represented by chapters 2 to 5, comprises the sewage plant's description and the theoretical background necessary for performing data-driven modelling strategies. The second part, chapters 6 to 8, presents the modelling, the obtained results and the conclusions. Moreover, Appendixes A, B and C contain the full set of plots.

Individually, the chapters follow this organization: Chapter 2 outlines the configuration of the WWTP and the data employed in this work. Chapter 3 introduces key data analysis concepts used in order to better understand the behaviour of the variables in the given dataset. Chapter 4 explains pre-processing concepts, including handling missing data, normalization, sample selection, and feature selection, that are the essential steps to refine the dataset and improve the accuracy of the modelling process. Moreover, Chapter 5 presents the theory behind both data-driven linear and non-linear models and how to evaluate their accuracy. Chapter 6 explains the developed models' configuration and parameters. Chapter 7 discusses the results of data-driven modelling. Finally, Chapter 8 brings a conclusion on the performed research.

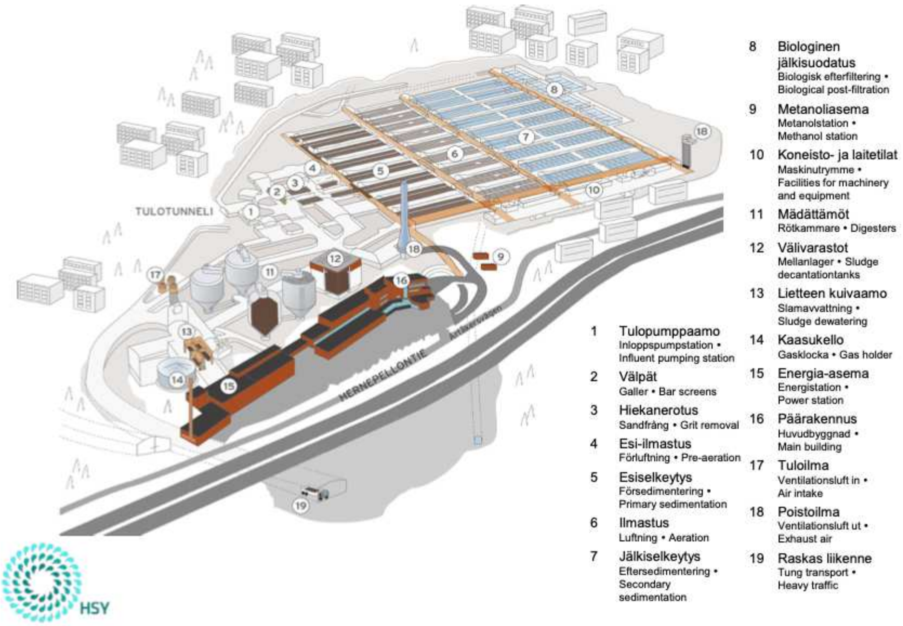
2 VIIKINMÄKI WASTEWATER TREATMENT PLANT

This chapter explains the configuration of the WWTP under study, describes the collected data and defines a reduced predictor set used in the development of the prediction models.

2.1 Plant description and data acquision

The data used in this work are provided by the Viikinmäki WWTP, located in Viikinmäki, a neighbourhood in Helsinki (Finland). The plant is one of the largest municipal wastewater treatment plants in the Nordic countries. The treatment process contains pumping of influent, bar screening, sand removal, primary sedimentation, biological reactor with denitrification - nitrification configuration, secondary sedimentation and denitrifying post-filtration (Figure 2).

Figure 2 – Schematic representation of the plant

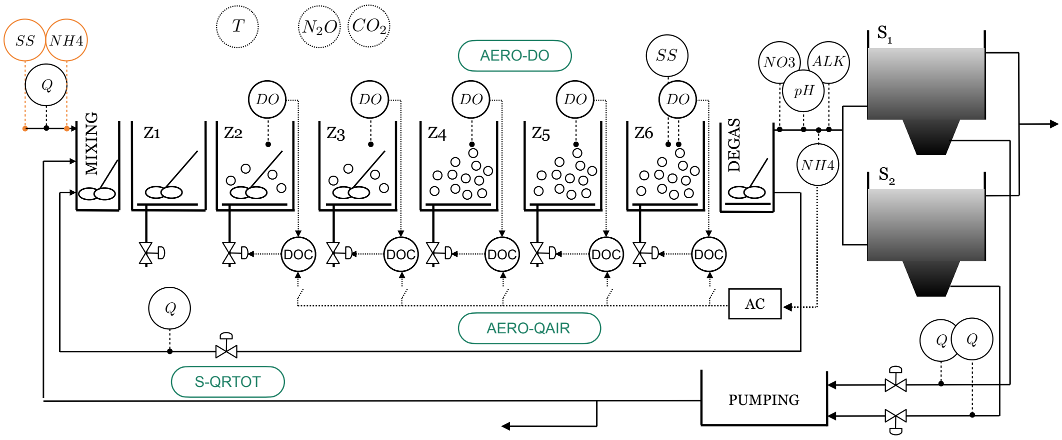

TULOTUNNELI
HERNEPELLONTIE
Artillerivägen
1 Tulopumppaamo Inloppspumpstation • Influent pumping station
2 Välpät Galler • Bar screens
3 Hiekanerotus Sandfrång • Grit removal
4 Esi-ilmastus Förluftning • Pre-aeration
5 Esiselkeytys Försedimentering • Primary sedimentation
6 Ilmastus Luftning • Aeration
7 Jälkiselkeytys Eftersedimentering • Secondary sedimentation
8 Biologinen jälkisuodatus Biologisk efterfiltering • Biological post-filtration
9 Metanoliasema Metanolstation • Methanol station
10 Koneisto- ja laitetilat Maskinutrymme • Facilities for machinery and equipment
11 Mädättämöt Rötkammare • Digesters
12 Välivarastot Mellanlager • Sludge decantationtanks
13 Lietteen kuivaamo Slamavvattning • Sludge dewatering
14 Kaasukello Gasklocka • Gas holder
15 Energia-asema Energistation • Power station
16 Päärakennus Huvudbyggnad • Main building
17 Tuloilma Ventilationsluft in • Air intake
18 Poistoilma Ventilationsluft ut • Exhaust air
19 Raskas liikenne Tung transport • Heavy traffic
HSY

Source: Helsinki Region Environmental Services Authority (HSY) (2021)

The Viikinmäki WWTP constantly monitors GHG emissions using a multi-component gas analyser, which measures the concentrations of pollutants in gas mixtures in the effluent air channel. The sensor is integrated into the plant's automation system and sends data to the treatment plant's internal database (KOSONEN *et al.*, 2016). Measurements include all unit operations, and the sensors store gaseous components in the effluent air, such as N_2O and CO_2.

The focus of this work is on the biological component of the treatment plant: the Activated Sludge Process ASP, represented in steps 6 and 7 in Figure 2. It involves aerating the wastewater to promote the growth of microorganisms that degrade organic matter and nutrients. However, this process also produces significant GHG emissions. Therefore, understanding and modelling emissions in ASP is crucial. In the plant, the ASP consists of nine treatment lines, each divided into six sequential zones, with the first being anoxic without aeration. The number of non-aerated zones, i.e., the anoxic volume, varies depending on the adjustable aeration mode. Each line begins with a mixing zone, where primary wastewater, secondary sedimentation return sludge, and internal recycling sludge are added. Figure 3 shows the scheme of a single line, presenting the measuring devices considered in this work.

Figure 3 – Schematic representation of an ASP line with sensors



Source: Viikinmäki WWTP.

Table 1 summarizes the data provided for each activated sludge line: the influent flow from primary sedimentation to the activated sludge lines is measured as I-Q; the sludge recirculation flow rates of the secondary sedimentation tanks are S1-QR and S2-QR and the internal recirculation (QA) are also provided. The concentration of dissolved oxygen (DO) is monitored in zones 2 to 6 (Z2/Z6-DO), while the total suspended solids are only evaluated in zone Z6 (Z6-SS). The effluent quality at the degassing zone is determined in terms of ammonia (D-NH$_4$), nitrate (D-NO$_3$), alkalinity (D-ALK), and pH (D-pH).

In each activated sludge line, the corresponding airflow rates can adjust the dissolved oxygen in all zones (Z2-DO to Z6-DO) by a feedback control loop. Aerobic zones 4 to 6 are always aerated, with a target DO of 3.5 mgL^{-1}. Zones 2 to 3 are aerated only when the ammonia content in this degassing zone exceeds a treatment limit (4 mgL^{-1}). Zone 1 is never aerated. When aerated, the dissolved oxygen targets in zones 2 to 3 are generally 3.5 mgL^{-1}.

The number of aerated zones is used to achieve removal efficiency by adjusting the anoxic volume. As a common practice for external recycling, the activated sludge flow rates returning from the clarifiers (S1/S2-QR) are proportional to the influent (I-Q). The internal flow depends on the recirculated sludge, the influent flow, and the number of aerated zones. Also, liquid temperature (T), suspended solids (SS), ammonia (NH_4) are provided and total amount of air added (AERO-QAIR) and average amount of dissolved oxygen (AERO-DO) are calculated.

Table 1 – Set of data considered for the study ($i = 1, \dots, 9$, unless stated otherwise)

Variable	**Description**	**Units**
Li-I-SS	Influent total suspended solid (only for i = 3, 9)	mg/L
Li-I-NH_4	Influent ammonia (only for i = 9)	mg/L
E-T	Liquid temperature	ºC
Li-Q-INT	Internal recycle flow-rate	m^3/s
Li-I-Q	Influent flow-rate	m^3/s
Li-S1/S2-QR	Return sludge flow-rate from S1 and S2	dm^3/s
Li-S-QRTOT	Total return sludge flow-rate	m^3/s
Li-Z2/Z6-DO	Dissolved oxygen (Z2 to Z6)	mg/L
Li-Z2/Z6-QAIR	Air flow-rate (Z2 to Z6)	Nm^3/s
Li-Z6-SS	Total suspended solids in Z6	g/L
Li-D-NH_4	Ammonia at the end of degas	mg/L
Li-D-NO_3	Nitrate at the end of degas	mg/L
Li-D-ALK	Alkalinity at the end of degas	mmol/L
Li-D-pH	pH at the end of degas	–
Li-AERO-DO	Average amount of dissolved oxygen (Z2 to Z6)	mg/L
Li-AERO-QAIR	Total amount of air added (Z2 to Z6)	Nm^3/s
N_2O	Nitrous oxide	mg/L
CO_2	Carbon dioxide	mg/L

Source: Viikinmäki WWTP.

The analysed data are collected from the Viikinmäki WWTP from January 11th, 2020, to March 11th, 2020. Each sample is an hourly average of the ASP measurement. The available dataset consists of 202 input predictors and 2 outputs (N_2O and CO_2). The dataset's predictors and output columns present 1441 samples. That is, the data can be represented in a predictor matrix $X \in \mathbb{R}^{n \times p}$, with n being equal to the number of samples (1441) and p the number of predictors (202), and in an output matrix $Y \in \mathbb{R}^{n \times m}$, where m is the number of target variables (2).

2.2 Reduced Predictor Set

Since the nine lines of the ASP exhibit similar behaviour for each predictor, a dataset based on their average values is also created, providing a reduced input matrix of 26 predictors and 2 outputs, presented in Table 2. That is, the predictor matrix $X \in \mathbb{R}^{n\times p}$ with $n = 1441$ and $p = 26$ and a target matrix $Y \in \mathbb{R}^{n\times m}$ with $m = 2$. Performing this reduction also allowed a more generalized modelling process, with the features viewed as a whole and slightly distancing the modelling strategies from a plant-driven development.

Table 2 – Reduced predictor set considered for the study ($i = 1,\ldots,9$, unless stated otherwise)

Variable	**Description**	**Units**
Li-I-SS	Influent total suspended solid (only for i = 3, 9)	mg/L
Li-I-NH4	Influent ammonia (only for i = 9)	mg/L
E-T	Liquid temperature	ºC
AVG-Q-INT	Average internal recycle flow-rate	m^3/s
AVG-I-Q	Average influent flow-rate	m^3/s
AVG-S1/S2-QR	Average return sludge flow-rate from S1 and S2	dm^3/s
AVG-S-QRTOT	Average total return sludge flow-rate	m^3/s
AVG-Z2/Z6-DO	Average dissolved oxygen (Z2 to Z6)	mg/L
AVG-Z2/Z6-QAIR	Average air flow-rate (Z1 to Z6)	Nm^3/s
AVG-Z6-SS	Average total suspended solids in Z6	g/L
AVG-D-NH4	Average ammonia at the end of degas	mg/L
AVG-D-NO3	Average nitrate at the end of degas	mg/L
AVG-D-ALK	Average alkalinity at the end of degas	mmol/L
AVG-D-pH	Average pH at the end of degas	–
AVG-AERO-DO	Average average amount of dissolved oxygen (Z2 to Z6)	mg/L
AVG-AERO-QAIR	Average total amount of air added (Z2 to Z6)	Nm^3/s
N_2O	Nitrous oxide	mg/L
CO_2	Carbon dioxide	mg/L

Source: Viikinmäki WWTP.

3 DATA ANALYSIS

This chapter discusses the theoretical background of Data Analysis. It begins by explaining the concept of exploratory analysis and which statistical methods can be used to present information from a dataset, showing how a univariate analysis can be performed. It also describes methods used in the bivariate analysis, the correlation and covariance analysis. Statistic results and visualizations from the Viikinmäki's full and reduced predictor sets will be shown.

3.1 Exploratory Analysis

Analysing the characteristics of a dataset is important to understand the distribution and the behaviour of its features. It can also help identify potential relationships between variables. Univariate analysis examines each predictor and output individually. This can be done by calculating statistical metrics or visualizing the feature through time series plots, box plots, or histograms. Aditionally, the correlation and covariance analysis explores the relationships between two variables, representing bivariate analysis. The following sections will investigate how the univariate and bivariate analysis can be represented.

3.1.1 Univariate Analysis

The univariate analysis examines a single variable to summarise its distribution and characteristics, using statistical metrics and visualizations of the variables individually.

3.1.1.1 Statistical metrics

Statistical metrics refer to the numerical values used to describe and summarise key aspects of the dataset. These metrics provide insight into the characteristics of the data. For the univariate analysis of the column $X \in \mathbb{R}^{n \times 1}$, we consider the following:

- Mean ($\bar{x}$): the mean value of x. It provides a measure of the central tendency.
  - $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- Standard deviation (σ): provides a measure of dispersion (spread). It is calculated as the square root of the variance, that measures the sample's variability from the mean.
  - The sample variance of a predictor x is calculated as $S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - So, $\sigma = \sqrt{S^2}$

- First quartile (Q_1): the value that divides the lowest 25% of the sorted values of the variable x.
- Median ($\tilde{x}$): is the middle value of the variable x when it is sorted in ascendant order.
- Third quartile (Q_3): the value that separates the lowest 75% of the highest 25%.

Regarding quartiles, they are defined as points where a sorted feature can be divided into 4 equally distributed parts. The first quartile ranges from the predictor's minimum value to the sample in position Q_1, the second from the sample directly after Q_1 to the median, the third from the median to the sample in position Q_3 and the fourth from the sample directly after Q_3 to the feature's maximum value.

For the sake of conciseness, only the statistical metrics for the reduced predictor set are presented in Table 3, where each features' minimum (Min) and maximum (Max) values are also shown. With this analysis, it is observed that the variables have different value ranges and that the predictors that present alike working configurations, such as AVG-Z5-QAIR and AVG-Z6-QAIR, AVG-Z4-DO, AVG-Z5-DO and AVG-Z6-DO, and AVG-S1-QR AND AVG-S2-QR, present similar statistics. Additionally, the outputs also present different ranges. Moreover, some features present low standard deviation values considering the mean, such as AVG-S-QRTOT, ACG-Z4-QAIR, AVG-Z5-QAIR, AVG-Z6-QAIR, AVG-Z4-DO, showing that these feature's samples are probably close to its mean.

3.1.2 Box plots

Box plots are a visual representation of a feature's distribution, focusing on the quartiles. They display the variable's quartiles. With them, it is possible to analyse if the values are more concentrated in a certain range by observing the distribution of the quartiles. They are constructed following these steps:

1. Two lines mark where the first and third quartile are located.
2. A thicker line (median) marks where the median is located,
3. The first and third quartiles are connected, forming a box-shaped form. Their distance is the Interquartile Range (IQR).
4. The whiskers represent the values between Q_1-1.5·IQR and Q_2+1.5·IQR and are the values that capture most of the data.
5. The outliers, if present, are marked with dots.

Considering the full dataset described in Chapter 2, the predictors whose values were

Table 3 – Descriptive statistics for the reduced predictor set

Variable	$\bar{x}$	σ	**Min**	Q_1	$\tilde{x}$	Q_3	**Max**
L3-I-SS	226.5092	88.7481	72.2005	111.7566	300.0000	300.0000	300.0000
L9-I-SS	124.5431	47.8566	51.3283	51.3283	111.7566	300.0000	299.2737
L9-I-NH_4	33.1866	6.7652	10.0377	34.2747	34.2747	49.3102	49.3102
E-T	11.9026	1.1833	7.7736	7.8800	12.0018	13.9355	13.9355
AVG-QINT	0.6539	0.0121	0.3395	0.6540	0.6548	0.6556	0.6570
AVG-I-Q	0.5269	0.1401	0.0	0.4451	0.4957	0.6020	0.9671
AVG-S1-QR	174.5949	37.2575	98.4166	151.0436	168.4957	202.0594	274.6792
AVG-S2-QR	181.6287	37.7279	103.925	158.1988	176.3501	210.0990	281.6904
AVG-S-QRTOT	1.0101	0.0769	0.5616	0.6015	0.9996	1.0671	1.2081
AVG-Z2-DO	0.3241	0.2520	0.0556	0.1346	0.3127	0.4114	2.4665
AVG-Z3-DO	1.6578	0.9530	0.1137	0.6015	2.1724	2.4730	2.7421
AVG-Z4-DO	2.4709	0.0864	2.1023	2.4406	2.4860	2.5173	2.8023
AVG-Z5-DO	2.4633	0.1024	1.9922	2.4550	2.4863	2.5079	2.8069
AVG-Z6-DO	2.4664	0.1290	1.9106	2.4531	2.4911	2.5165	3.3235
AVG-Z2-QAIR	0.0114	0.0408	0.0	0.0	0.0	0.0	0.3077
AVG-Z3-QAIR	0.2562	0.1757	0.0	0.0649	0.2901	0.4070	0.5457
AVG-Z4-QAIR	0.3135	0.0792	0.1018	0.2712	0.3213	0.3696	0.4896
AVG-Z5-QAIR	0.2518	0.0587	0.0768	0.2228	0.2552	0.2932	0.3694
AVG-Z6-QAIR	0.2133	0.0513	0.0663	0.1870	0.2167	0.2497	0.3188
AVG-Z6-SS	4.1953	0.4704	3.2540	3.8537	4.0491	4.5072	5.7203
AVG-D-NH_4	2.7641	1.6294	0.1287	0.9643	2.6994	3.8537	8.0754
AVG-D-NO_3	9.7459	2.4237	1.6408	7.8800	9.5723	11.6856	15.0123
AVG-D-ALK	1.2670	0.1807	0.4236	1.6375	1.2472	2.0786	1.9338
AVG-D-pH	6.2185	0.1563	2.7182	6.1520	6.2024	6.2789	6.5588
AVG-AERO-QAIR	1.0461	0.2735	0.2755	0.8340	1.0473	1.2564	1.6417
AVG-AERO-DO	1.8765	0.2691	1.2456	1.6375	2.0034	2.0786	2.5712
N_2O	18.4829	14.0693	1.8287	9.2003	15.5636	24.5654	93.7133
CO_2	4398.7323	1029.5664	1373.6484	3833.5342	4436.1631	5192.4258	6428.6421

Source: The author.

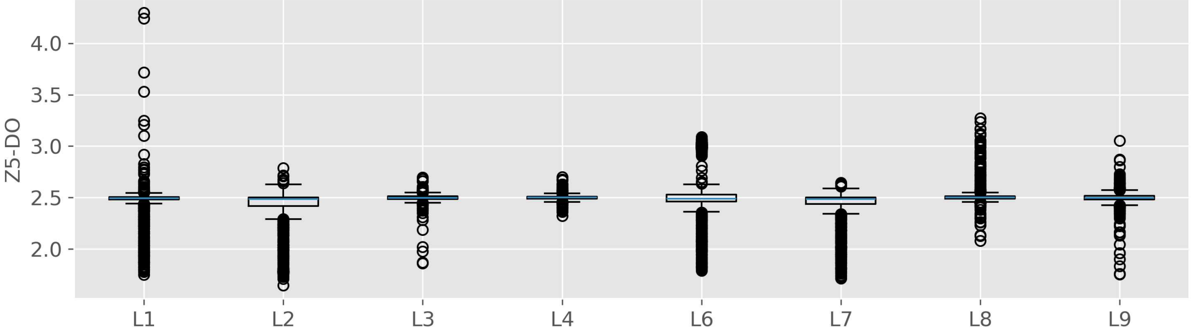
collected in different lines had a single plot representing all their lines. Figures 4 to 9 represent the box plots for four selected inputs and the two outputs, with the full set of plots available in Appendix A.

Figures 4 to 9 show how the variables quartiles' intervals are distributed, and the samples considered outliers. Notice that the line 9, which was implemented later in the plant, presents a different behaviour in every plot it appeared (Figures 4, 5, 6). In addition, each of the lines 1-8 in the total return sludge flow rate (Figure 5) and nitrate at the end of degas (Figure 6) have similar behaviour. This similarity may be explained because the same process happens in each line.

Moreover, in the dissolved oxygen feature (Figure 4), the outliers are very close to the IQR, while in the nitrate at the end of degas, they are located further away. It is also noticeable that most of the samples labelled as outliers are situated on the bottom of the plot, as shown in the liquid temperature (Figure 7) and the CO_2 plots (Figure 9). The N_2O plot (Figure

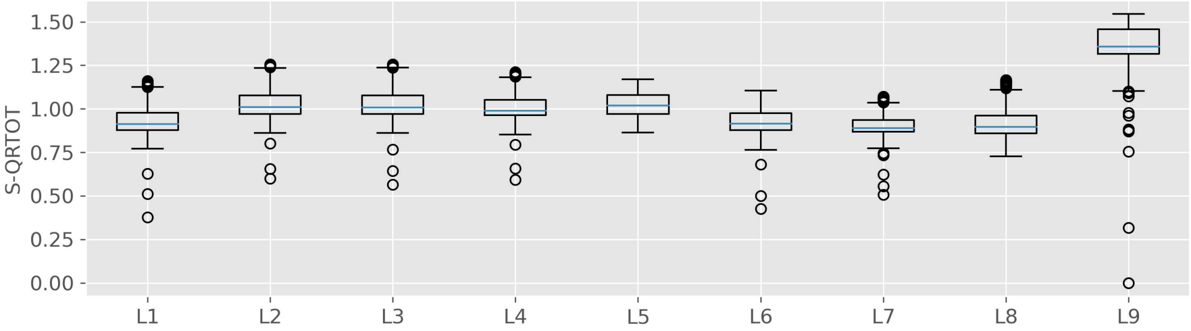
8), however, follows a different pattern, with the samples considered outliers shown in its upper part.

Figure 4 – Box plot of dissolved oxygen in zone 5



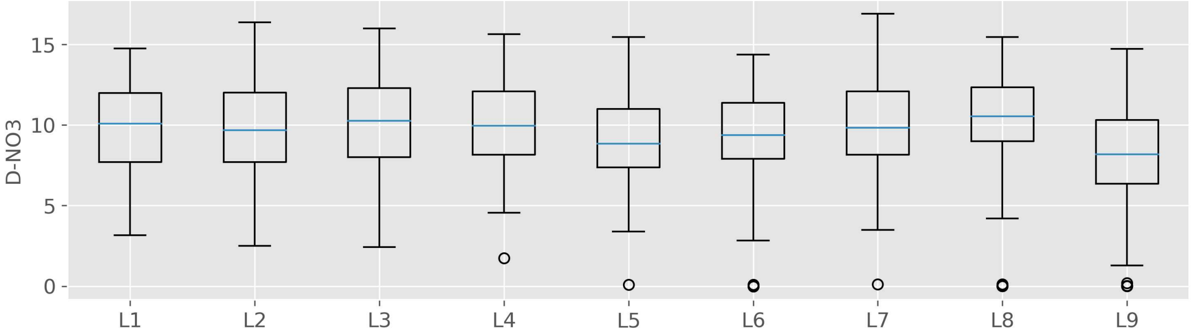
Source: The author.

Figure 5 – Box plot of total return sludge flow-rate



Source: The author.

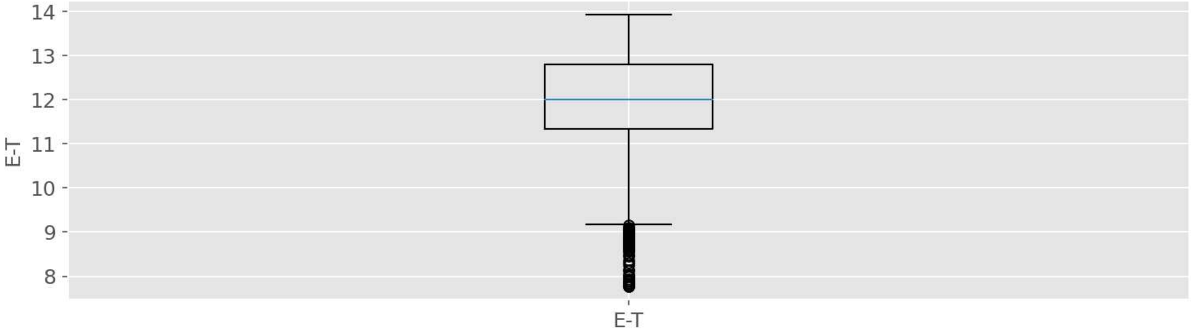
Figure 6 – Box plot of nitrate at the end of degas



Source: The author.

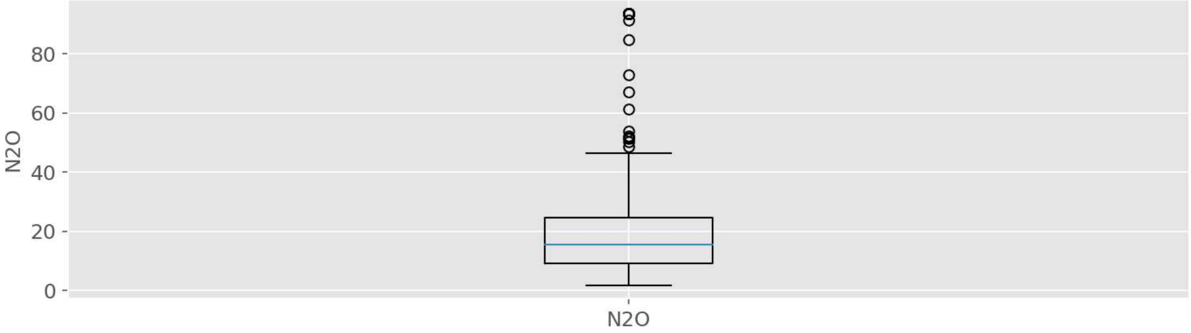
As for the reduced predictor set's box plots, observe that the dissolved oxygen in

Figure 7 – Box plot of liquid temperature



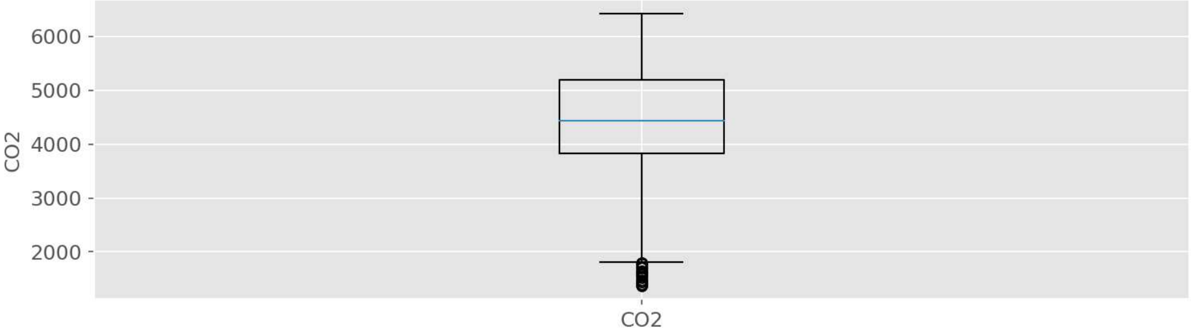
Source: The author.

Figure 8 – Box plot of nitrous oxide



Source: The author.

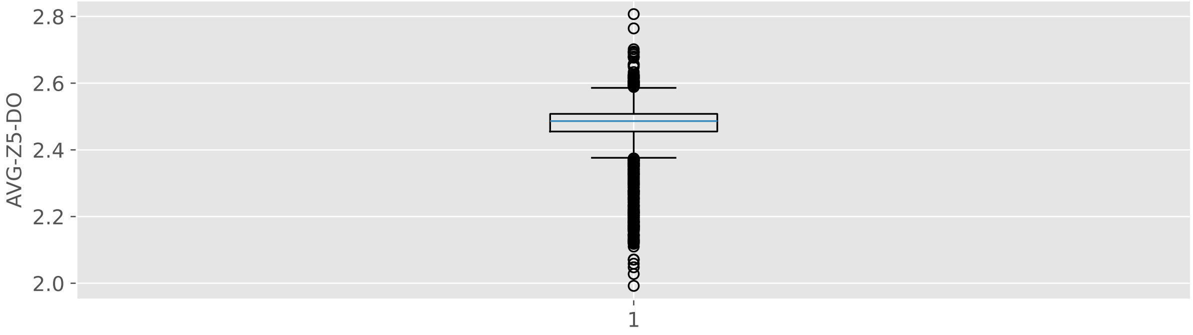
Figure 9 – Box plot of carbon dioxide



Source: The author.

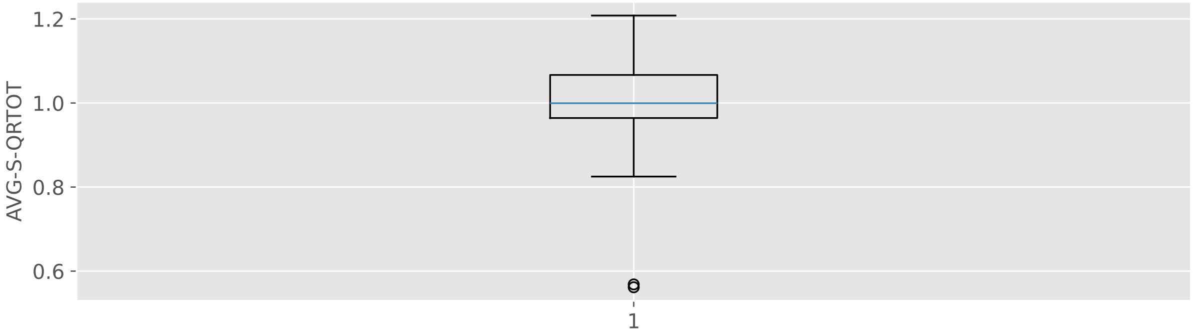
zone 5 (Figure 10) presents many outliers, while the average total return sludge flow-rate and the average nitrate at the end of degas (Figures 11 and 12, respectively) present fewer, further from its whiskers.

Figure 10 – Box plot of average dissolved oxygen in zone 5



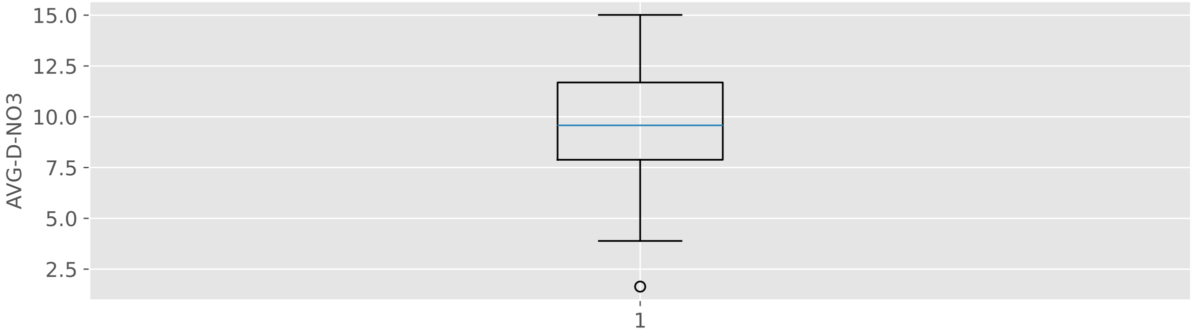
Source: The author.

Figure 11 – Box plot of average total return sludge flow-rate



Source: The author.

Figure 12 – Box plot of average nitrate at the end of degas



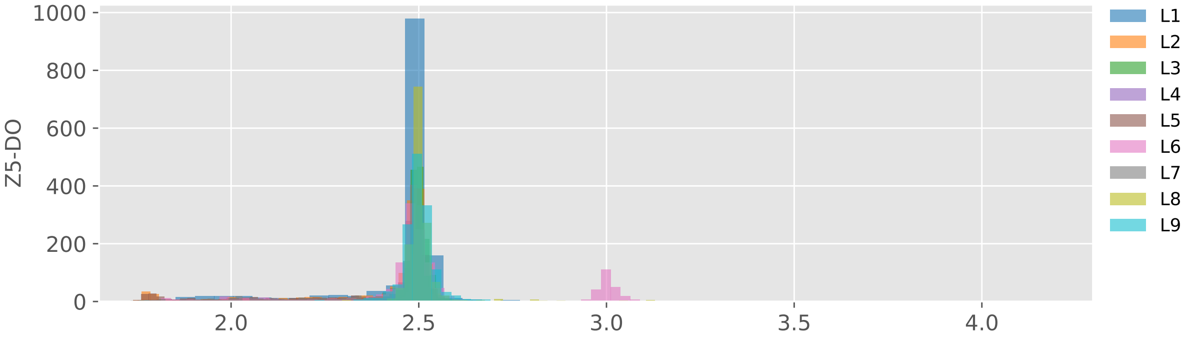
Source: The author.

3.1.3 Histograms

Histograms present how data are spread across different intervals on the x-axis, and y-axis representing the frequency or the count of data from that interval. They help visualize the shape of the data distributions, such as whether it is symmetric, uniform or has multiple modes.

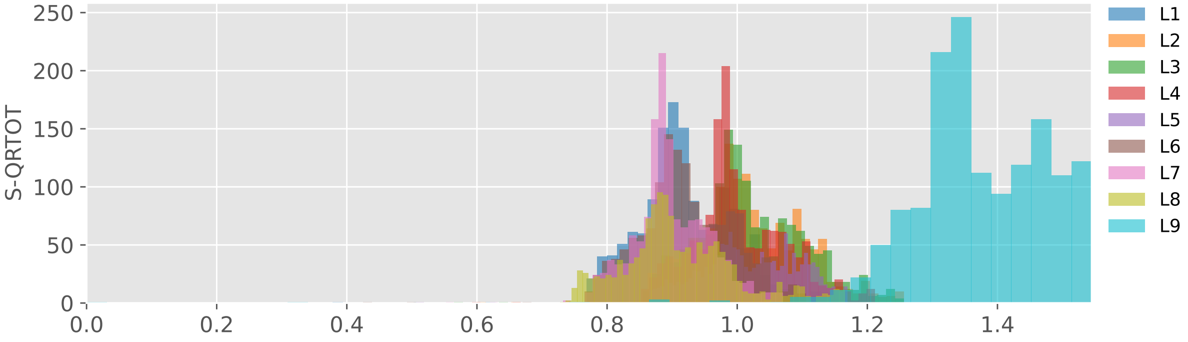
The histograms for the full dataset are presented in Figures 13 to 18 for the 9 lines of dissolved oxygen in zone 5, the return sludge flow-rate and nitrate at the end of degas, the liquid temperature and the two gaseous outputs, respectively. Notice that line 9 presents a different variation in the total return sludge flow rate feature (Figure 14). In addition, the dissolved oxygen in zone 5 (Figure 13) and the nitrate at the end of degas (Figure 15) histograms show that each line distribution variates, with some being skewed to the left or the right and some not presenting skewness. The histograms for the liquid temperature (Figure 16) and the CO_2 (Figure 18) present skewed to the left distributions, while the N_2O (Figure 17) shows a skewed to the right one. These distributions probably explain the way their respective box plots marked the samples considered outliers, since the skewness shown in the histograms displays that most of the samples are concentrated in this range, while samples in other ranges occur less and can be classified as outliers.

Figure 13 – Histogram of dissolved oxygen in zone 5



Source: The author.

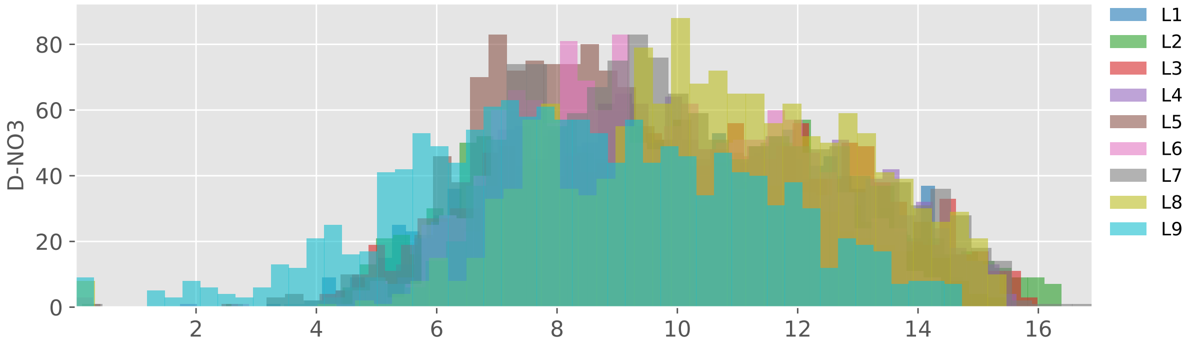
Figure 14 – Histogram of total return sludge flow-rate



Source: The author.

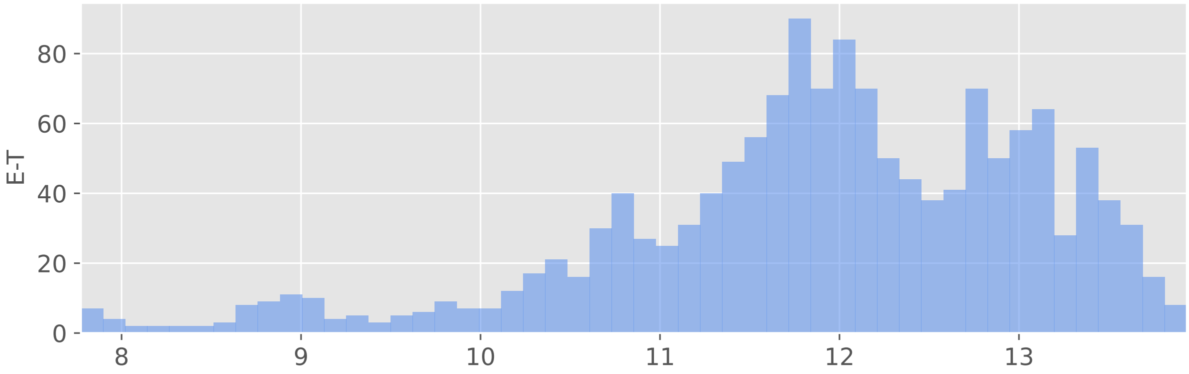
As for the reduced predictor set, the histograms are shown in Figures 19 to 21, presenting average dissolved oxygen in zone 5, average total return sludge flow-rate and average

Figure 15 – Histogram of nitrate at the end of degas



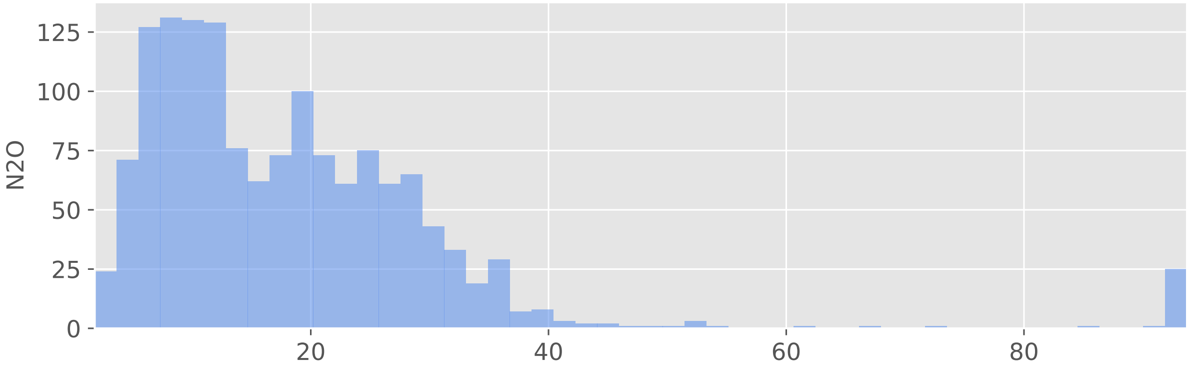
Source: The author.

Figure 16 – Histogram of liquid temperature



Source: The author.

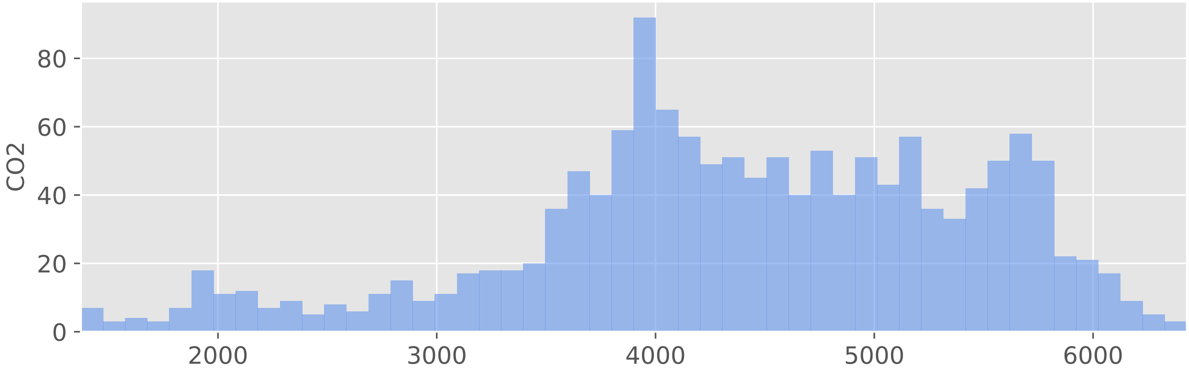
Figure 17 – Histogram of nitrous oxide



Source: The author.

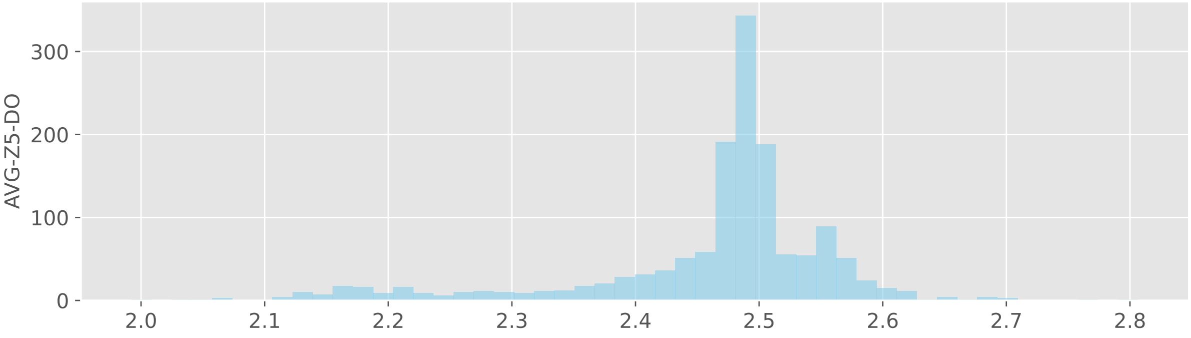
nitrate at the end of degas, respectively. It is observed that all the presented features have skewness to the left, which could explain why the samples classified by the box plots as outliers are located at the bottom.

Figure 18 – Histogram of carbon dioxide



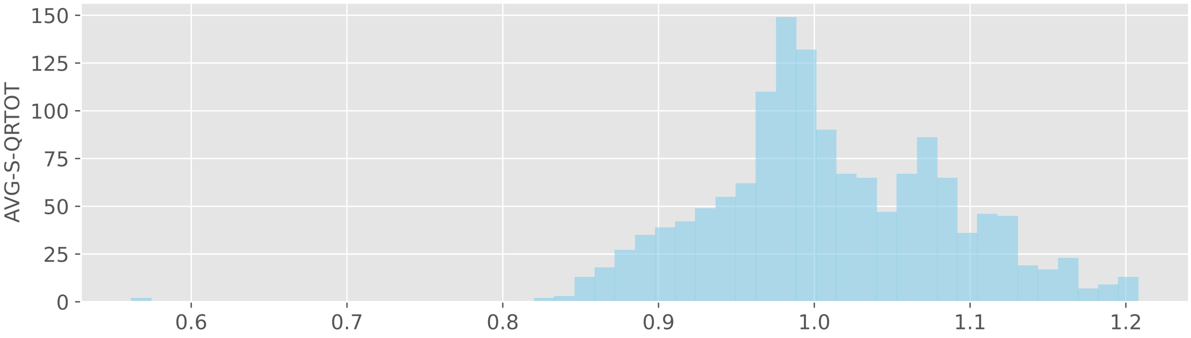
Source: The author.

Figure 19 – Histogram of average dissolved oxygen in zone 5



Source: The author.

Figure 20 – Histogram of average total return sludge flow-rate

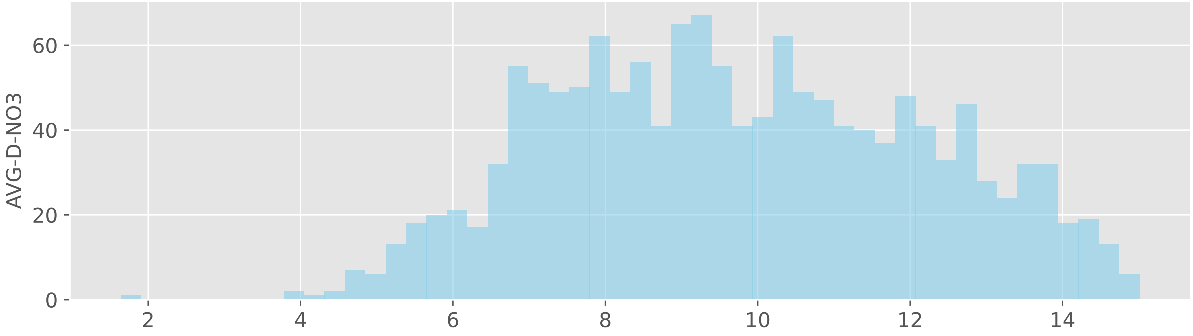


Source: The author.

3.1.4 Time series

Since the provided data consisted of time series - sequences of data collected at successive time intervals - they are plotted along the corresponding time axis. This approach allows for tracking changes over time, analysing trends, patterns, and possible seasonal effects.

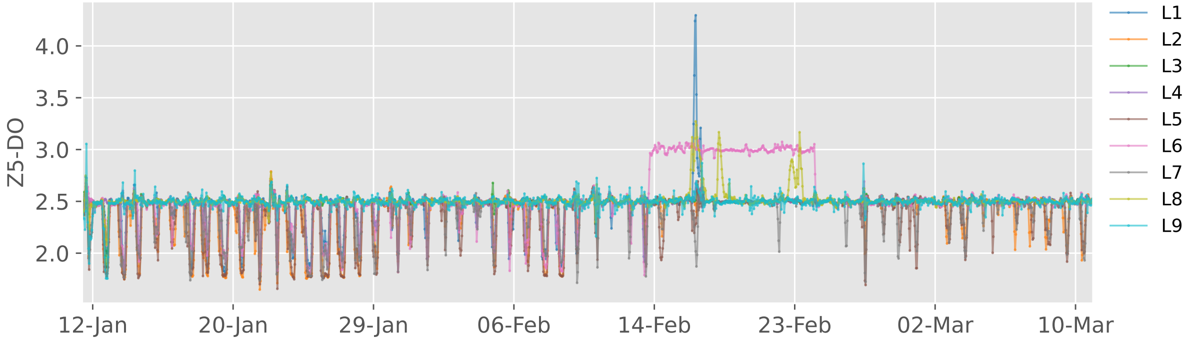
Figure 21 – Histogram of average nitrate at the end of degas



Source: The author.

Figure 22 to 27 presents the full dataset's time series, respectively, for the 9 lines of dissolved oxygen in zone 5, the return sludge flow-rate and nitrate at the end of degas, the liquid temperature and the two GHG outputs. It can be observed that Z5-DO (Figure 22), S-QRTOT (Figure 23) and D-NO3 (Figure 24) exhibit very similar behaviour, with the exception of the total return sludge flow-rate in line 9, which differs from the others. Additionally, the period between the 14th and the 23rd of February is characterised by anomalous peaks in the dissolved oxygen in zone 5 and irregular behaviours in the nitrate at the end of degas features. This may result from the reduction in the liquid temperature variable that happened during this period (Figure 25) probably caused by snow melt, that reduces the influent wastewater liquid temperature. This unusual behaviour is also noticeable in the GHG emissions (Figures 26 and 27), with more relevance on the reduction of CO_2 emissions.

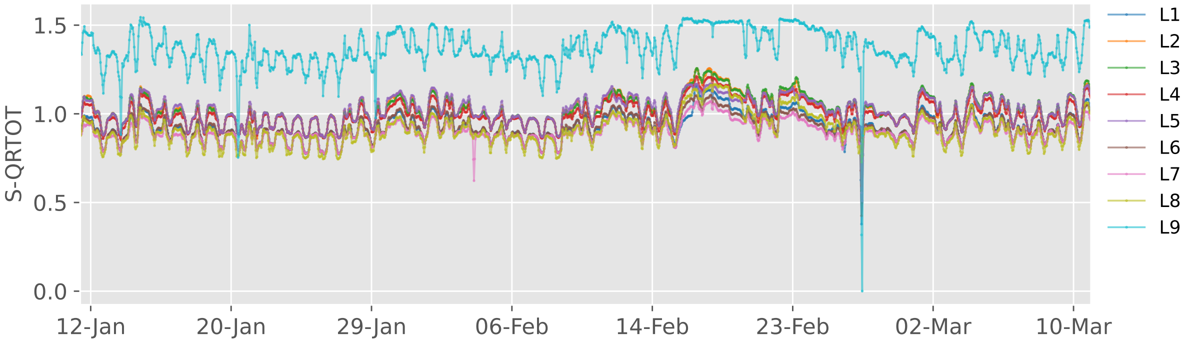
Figure 22 – Time series of dissolved oxygen in zone 5



Source: The author.

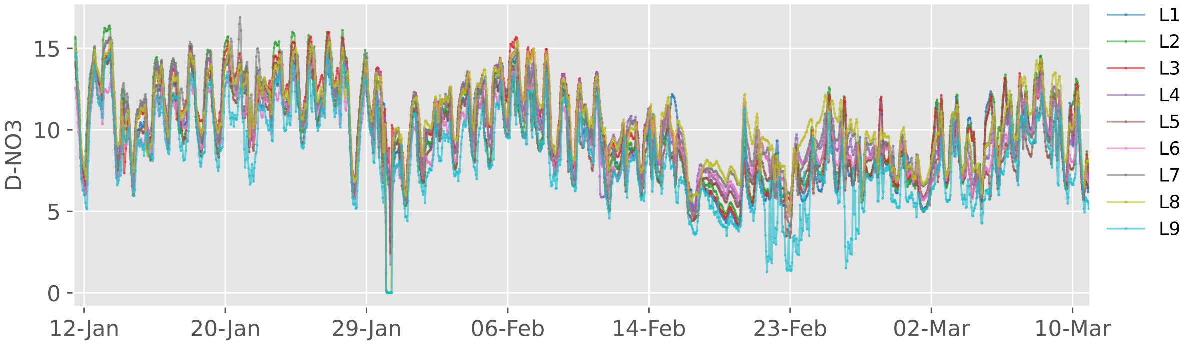
Figures 28 to 30 show the time series for, respectively, average dissolved oxygen in zone 5, average total return sludge flow-rate and average nitrate at the end of degas. Notice that

Figure 23 – Time series of total return sludge flow-rate



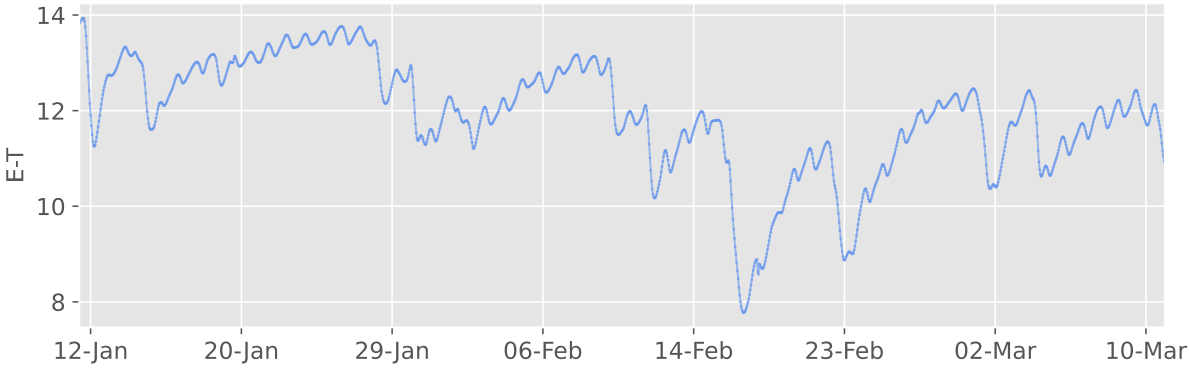
Source: The author.

Figure 24 – Time series of nitrate at the end of degas



Source: The author.

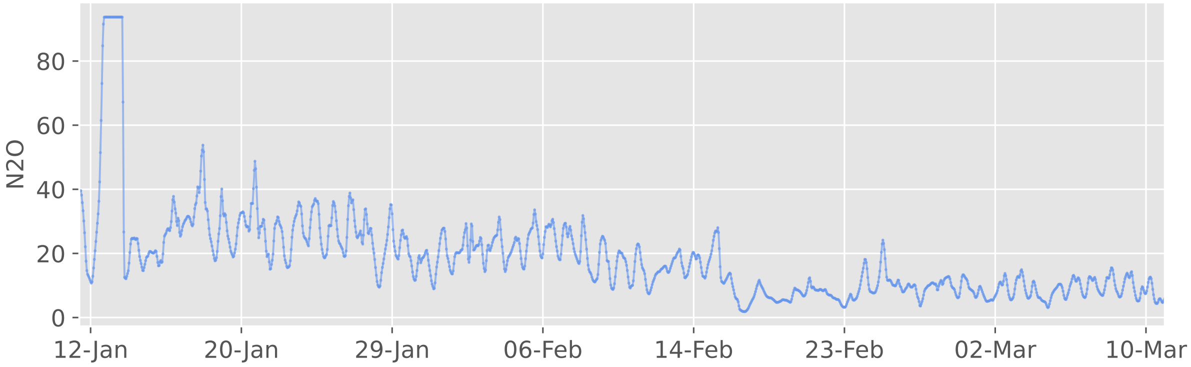
Figure 25 – Time series of liquid temperature



Source: The author.

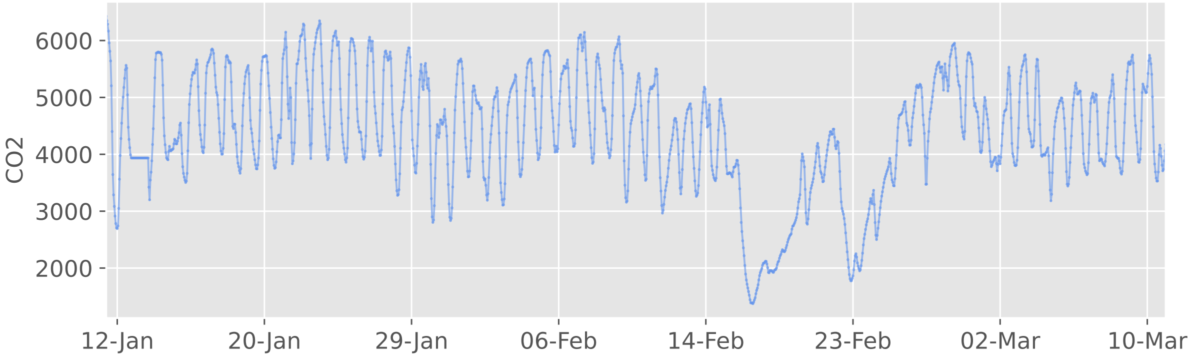
they show abnormalities during the second and third weeks of February, probably associated with the change in liquid temperature caused by the snow melt.

Figure 26 – Time series of nitrous oxide



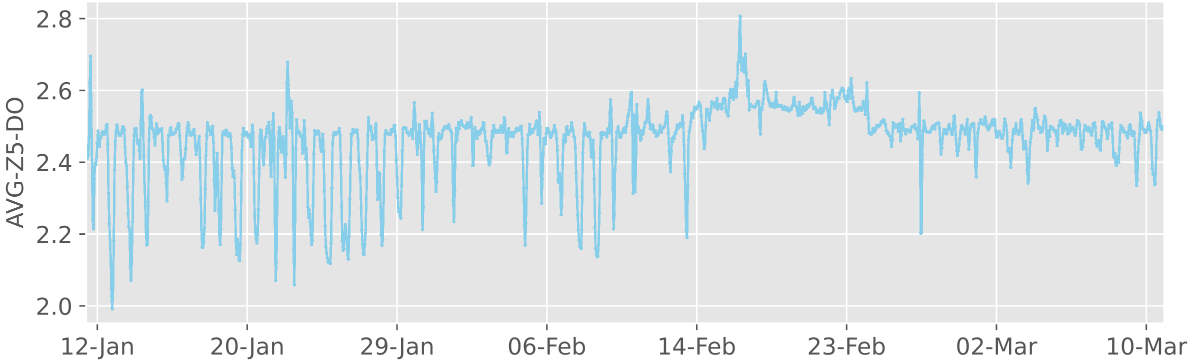
Source: The author.

Figure 27 – Time series of carbon dioxide



Source: The author.

Figure 28 – Time series of average dissolved oxygen in zone 5

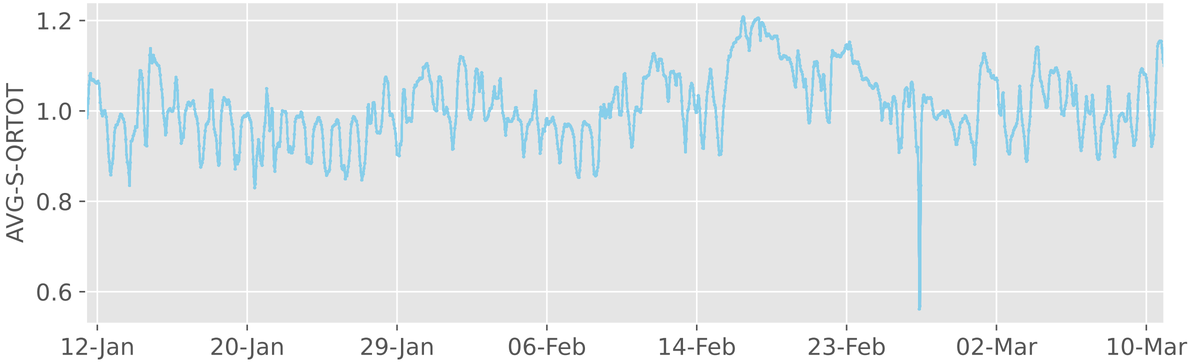


Source: The author.

3.1.5 Bivariate Analysis

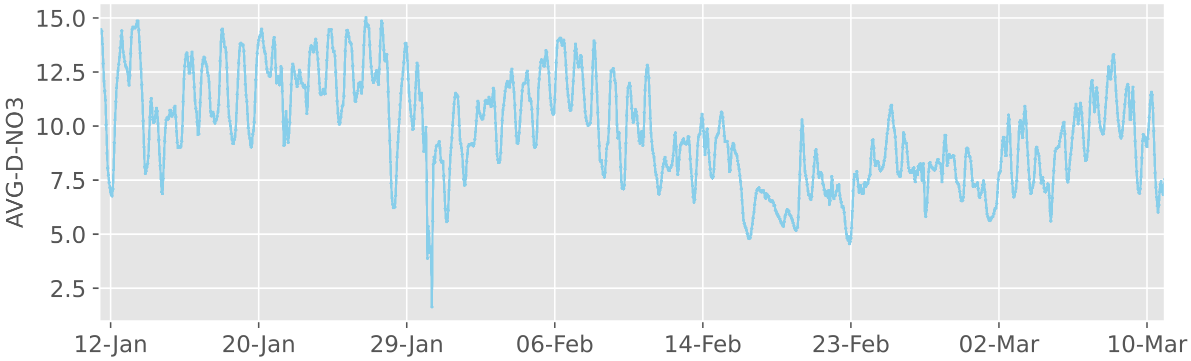
Bivariate analysis examines the relationship between two variables. As statistical measures, the covariance is used to evaluate how the variables variate together, and Pearson Correlation Coefficient to assess the strength and direction of the linear relationship.

Figure 29 – Time series of average total return sludge flow-rate



Source: The author.

Figure 30 – Time series of average nitrate at the end of degas



Source: The author.

3.1.5.1 Covariance Analysis

The covariance of two features x and y is calculated as shown in Equation (3.1). It represents the measurement of the joint variability of these two features, indicating the amount that changes in the variable x are associated with changes in the variable y.

$$cov(x,y) = \frac{\sum_i^p (x_i - \bar{x})(y_i - \hat{y})}{n-1} \tag{3.1}$$

For exemplification, a covariance matrix of two features will have the following format:

$$\begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix}$$

Considering some properties of $cov(x,x) = var(x)$ and $cov(x,y) = cov(y,x)$, we have that:

$$\begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix}$$

The covariance matrix of n features follows the same pattern. It is a step in the Correlation analysis and the PCA calculations, which will be explained further.

3.1.5.2 Correlation Analysis

The Pearson Correlation Coefficient (ρ) is used to quantify the strength and direction of the linear relationship between two variables. It is calculated as the ratio of the covariance between two variables and the product of their standard deviations (Equation (3.2)). It ranges between -1 and +1, and the closer the value is to +1, the stronger the positive correlation between the variables (as one increases, the other also increases). Conversely, when the value is closer to -1, the variables are negatively correlated (as one increases, the other decreases). When the coefficient is closer to zero, the variables exhibit weak or non-linear correlation, meaning that the changes in one variable have little or no effect on the other.

$$\rho = \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

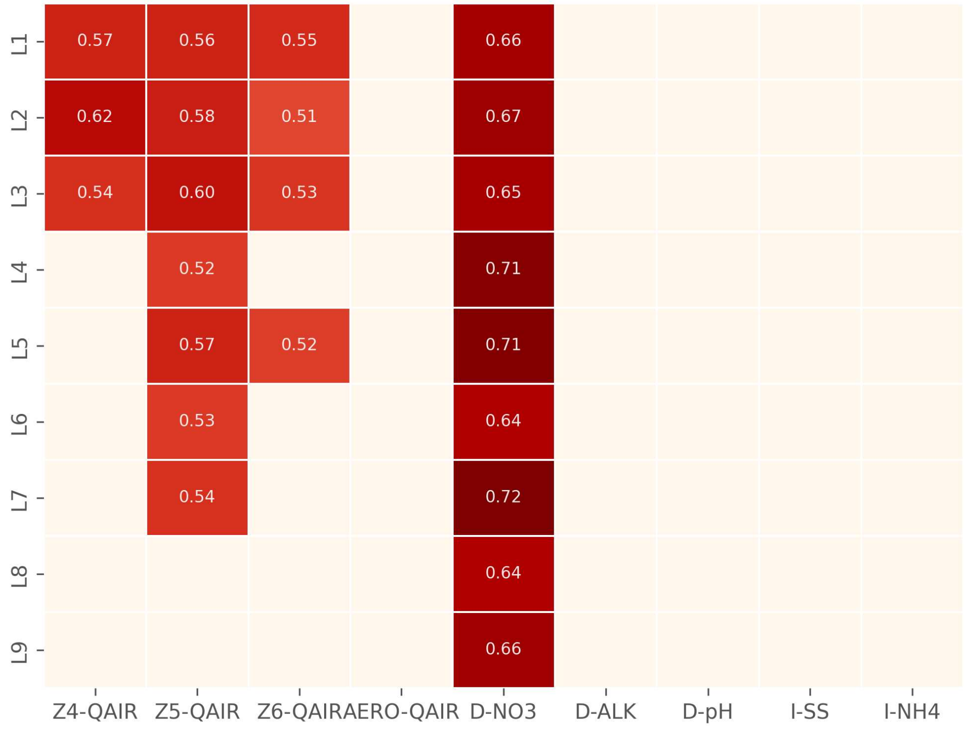
where ρ is the correlation coefficient, x_i is a sample of the x feature, $\bar{x}$ is the mean of x, y_i is a sample of the y feature, and $\bar{y}$ is the mean of y.

For the full dataset, because of its feature size, only the predictors with a correlation greater than 0.5 (either positive or negative) with N_2O and CO_2 emissions are considered. Figures 31 and 32 show heatmaps of the ρ values with N_2O and CO_2 gases, respectively. Values closer to +1 are deeper in dark red.

By observing the correlation plots, it is noticeable that the features don't show strong negative correlations with the output and that the CO_2 gas emissions present a stronger positive correlation with the dataset's features, especially the ones that measure air flow rate (Figure 32). Moreover, the features most strongly correlated to the N_2O emissions (Figure 31) are those associated with the nitrate at the end of degas, likely due to the transformation of nitrate compounds to nitrous ones, which contribute to GHG formation.

Further analysis shows that both GHG emissions are related to the liquid temperature, with a correlation of 0.50 with the N_2O gas and 0.66 with the CO_2 gas. This may explain why the outputs present behaviour changes when there is a temperature change.

Figure 31 – Features' correlation with N_2O gas

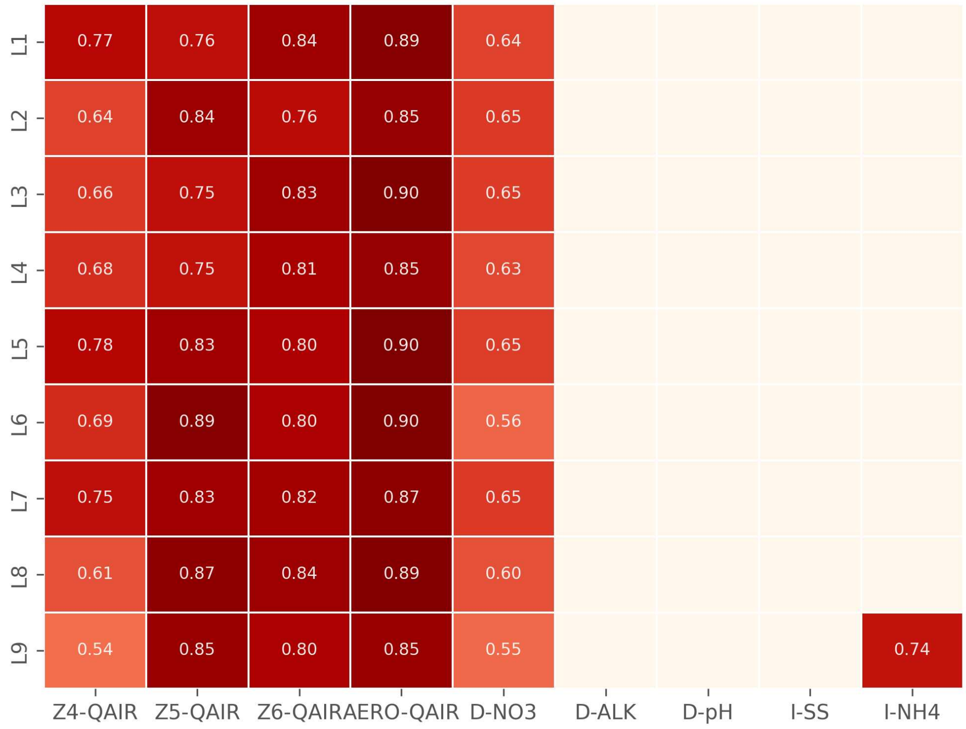


Source: The author.

For the reduced predictor set, it is possible to understand more about how the features correlated between them, as represented by the heatmap in Figure 33. It is noticeable that the features of the same kind (dissolved oxygen, air flow, return sludge flow rate) are positively correlated, while the liquid temperature is negatively correlated with the influent flow rate and the return sludge flow rate.

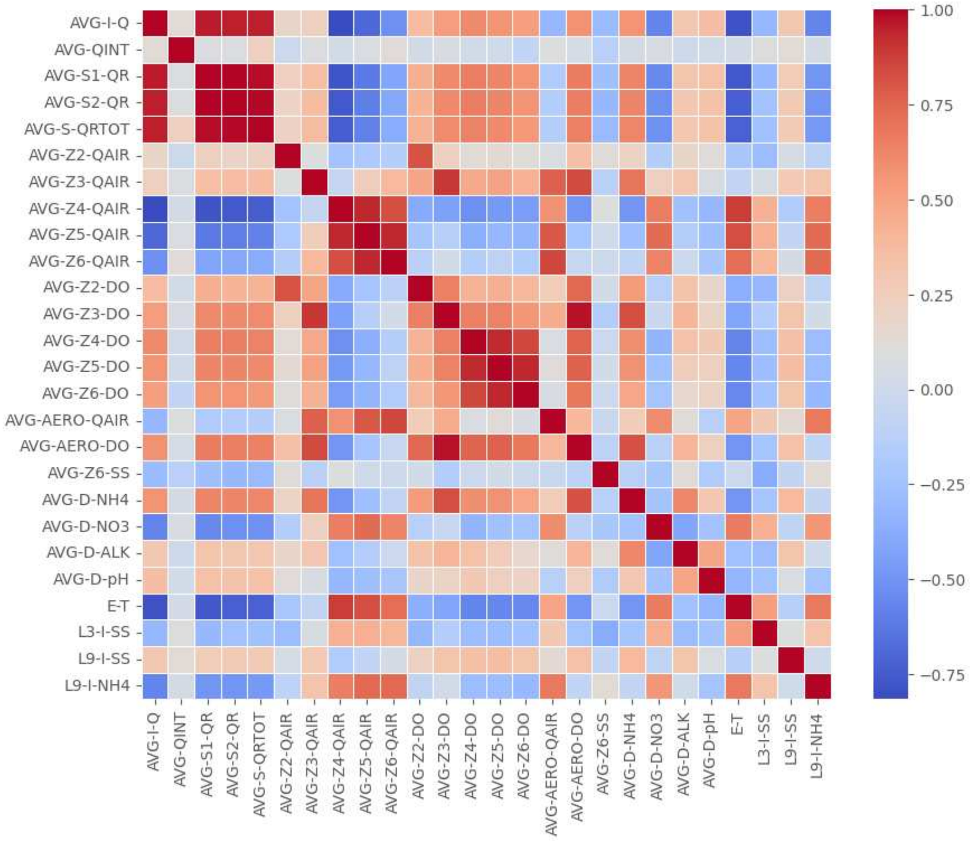
In addition, based on the correlation between the features and the outputs, four features show a strong with the N_2O emission, as detailed in Table 10, and nine with CO_2 emissions, shown in Table 11. It is observed that the liquid temperature is somewhat positively correlated to the two outputs, while the variable related to the nitrate at the end of degas shows the highest correlation with the N_2O emissions. In addition, variables related to the air flow emissions are more correlated with the CO_2 gas. These findings are consistent with the ones obtained with the full dataset, showing that the correlation between the variables and the outputs is alike.

Figure 32 – Features' correlation with CO_2 gas



Source: The author.

Figure 33 – reduced predictor set's correlation



Source: The author.

Table 4 – Features with the highest correlation with the N_2O gas

Feature Name	**Correlation Value**
AVG-Z4-QAIR	0.51
AVG-Z5-QAIR	0.55
AVG-D-NO_3	0.71
E-T	0.50

Source: The author.

Table 5 – Features with the highest correlation with the CO_2 gas

Feature Name	**Correlation Value**
AVG-I-Q	-0.54
AVG-Z3-QAIR	0.58
AVG-Z4-QAIR	0.71
AVG-Z5-QAIR	0.85
AVG-Z6-QAIR	0.85
AVG-AERO-QAIR	0.92
AVG-D-NO_3	0.66
E-T	0.66
L9-I-NH_4	0.74

Source: The author.

4 DATA PRE-PROCESSING

This chapter presents concepts of data pre-processing techniques. The first section explores the impact of missing values in the dataset and briefly outlines strategies for dealing with them. The second section discusses normalization, which helps improve model performance by scaling the dataset's values. The third section focuses on methods for detecting anomalous behaviour and performing sample selection. The final section defines algorithms that select relevant features for data-driven modelling.

4.1 Missing Values

Missing values refer to the absence of samples in one or more variables in the dataset. They can arise due to various reasons as sensor faults, for example, and must either be removed or imputed before training machine learning models.

If the dataset is large and the number of samples that present missing values is relatively small, removing these samples may not significantly affect the overall modelling process. However, if a substantial portion of the data is missing, imputation methods can be employed to estimate reasonable values for the missing sample considering, for instance, the mean or mode of the specific feature.

Considering the studied data, the full dataset presents 438 missing values, which will be removed. For the reduced predictor set, the calculation of the predictors' averages excludes missing values in each feature. As a result, the dataset contains 260 missing values, which were removed.

4.2 Normalization

The dataset considered in this study contains variables with different units and varying ranges that can impact model development. Features that have higher values may be wrongly perceived as more relevant during the modelling process. To address this issue, the dataset should be normalized, making all of its features fit in a specific value range.

A commonly used normalization method is the Min-Max scaling, which transforms all the feature's values into a range that typically varies between 0 and 1. This process is described by Equation (4.1), where each scaled sample (x_{scaled}) is derived from the original value (x) by considering its minimum (x_{min}) and maximum (x_{max}) values.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4.1}$$

4.3 Sample Selection

Sample selection techniques help to identify samples that can be considered outliers or present anomalous behaviour. To perform modelling strategies adequately, the presence of outliers must be reduced, since they affect model performance because of their different dynamics considering the other samples. Thus, the employment of sample selection techniques helps to generate a dataset that is considered to have more quality for the modelling process.

4.3.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) involves the diagonalisation, or eigen decomposition, of the covariance matrix of the standardized predictors. Each eigen vector represents the direction (principal component) in which the data varies the most, and each eigenvalue represents the amount of variance along the corresponding eigenvector (JAMES *et al.*, 2013). Hence, the technique can be used as a way of representing the dataset in smaller dimensions while preserving its variation as much as possible.

Each matrix's eigenvalue is called a component and is a standardized linear combination of the dataset's p features multiplied by elements named loadings, which indicate the contribution of p to the component, as presented in Equation (4.2), where Z_i is the calculated component, the vector $(\phi_{1i}, \phi_{2i}, ..., \phi_{pi})$ is the loadings vector and the vector $(X_1, X_2, ..., X_p)$ is the feature vector.

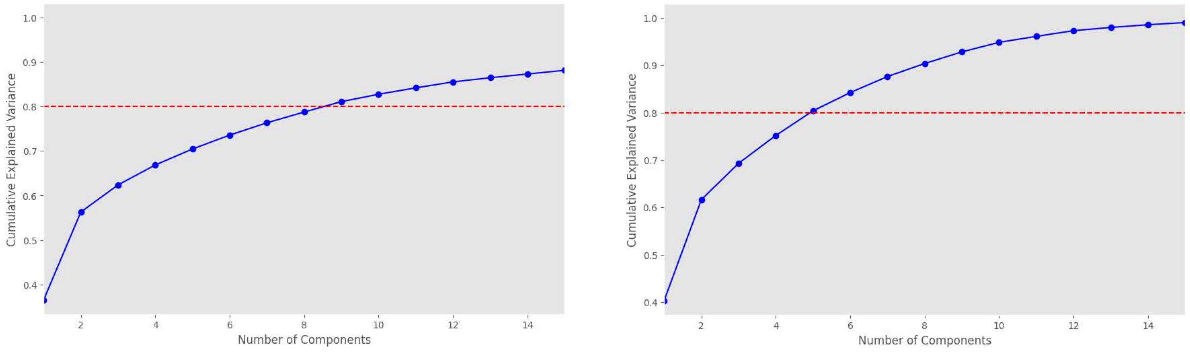
$$Z_i = \phi_{1i}X_1 + \phi_{2i}X_2 + ... + \phi_{pi}X_p \tag{4.2}$$

A dataset with p features will also have p principal components, with the percentage of explained variance being ordered in descendant order. Because of the algebraic calculations of the covariance matrix diagonalization, the first component will always have the most explained variance, the second will have the second most explained variance, and so on. In addition, a subsequent component is always orthogonal to its former.

Considering how to perform the PCA on a dataset, to select the number of components used for the analysis, a Single Value Decomposition (SVD) is performed on the normalized dataset's covariance matrix to diagonalize it and to understand how much the components explain the dataset's variance. The SVD is calculated as shown in Equation (4.3), with A being the covariance matrix, U and V the orthonormal matrices, and D a diagonal matrix with positive entries. The D matrix is the one considered for the explained variance analysis.

$$A = UDV^T \tag{4.3}$$

The cumulative sum of the explained variance percentage for each component can be represented visually for better understanding. According to the literature (ZHANG *et al.*, 2020), the selected components should account for a cumulative explained variance of at least 80% of the dataset's total variance. Figures 34a and 34b show that, considering the full and reduced predictor sets, respectively, nine and five components should be considered in the sample selection process. The plots only show the first 15 components for better visualization. Notice that the cumulative explained variance increases at a slower rate as more components are added.



(a) Single Value Decomposition of the full dataset

(b) Single Value Decomposition of the reduced predictor set

Figure 34 – Single Value Decomposition plots

Source: Plotted by author.

4.3.2 Robust PCA

The default PCA algorithm is highly affected by anomalous samples, and the estimation of the sample mean that is employed for the centring of the data is conditioned by relevant outliers, which differ greatly from the other samples (HAIMI, 2016), which may reduce the effectiveness of the default algorithm.

As a solution to this problem, Robust PCA methods apply robust statistic estimators for locating the data's centre and scaling the covariance matrix that is more resistant to outliers, as the robust covariance matrix, used in this work, which would be a weighted covariance matrix in which close observations would contribute more to the calculations (RUIZ-GAZEN, 1996). The weight parameter $w(i,j)$ is defined in Equation (4.4), where β is a tuning parameter that reduces the influence of further observations. The higher its value, the more the samples closest to the centre are considered:

$$w(i,j) = exp(-\frac{\beta}{2}(y(i)-y(j))^{cov_r(x,y)}\Sigma^{-1}(x(i)-x(j))) \tag{4.4}$$

Therefore, the robust covariance matrix calculation equation (RUIZ-GAZEN, 1996) will be presented as (4.5):

$$cov_r(x,y) = \frac{\sum_i^{p-1}\sum_{j=i+1}^{p} w(i,j)(x(i)-x(j))(y(i)-y(j))}{\sum_i^{p-1}\sum_{j=i+1}^{p} w(i,j)} \tag{4.5}$$

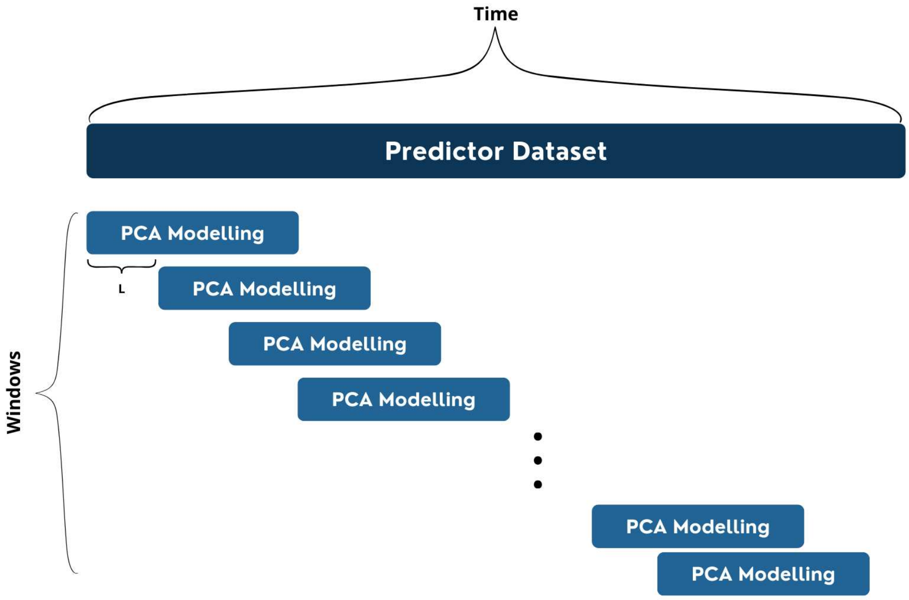
4.3.3 Moving Window PCA

To address the time-invariant characteristic of the PCA in time-varying processes, the use of Moving Window (MW) methods is suggested. The MW-PCA approach utilizes historical data from a period with a window length of L to build PCA models (KRUGER; XIE, 2012). Then, novel PCA models are constructed at the time intervals of a certain shift size Z, each window shifts along this size, and a new model is trained at each time step with only the newest data in the length Z (Figure 35).

4.3.4 Statistics for monitoring data variability

The Hotelling's (T^2) and Q-statistic statistics (JACKSON; MUDHOLKAR, 1979) and their confidence limits T^2_{lim} (ATKINSON *et al.*, 2004) and Q_{lim} (NOMIKOS; MACGREGOR, 1995) are used for monitoring the variability in data and identifying possible outliers. The Hotelling's T^2 measures the normalised Mahalanobis distance of the projected observation score at a time k, from the origin of the principal component subspace as in Equation (4.6), where $\wedge^{-1}$ represents the inverse of the covariance matrix with the inverse of the eigenvalues associated with the retained principal components.

Figure 35 – MW-PCA configuration



Source: The author.

$$T^2(k) = t(k) \wedge^{-1} t(k) \tag{4.6}$$

The Q-statistic, in Equation (4.7), measures the orthogonal distance of an observation $x_d(k)$ from its reconstruction $\hat{x}_d(k)$ on the principal component subspace.

$$Q(k) = \sum_{d=1}^{p} (x_d(k) - \hat{x}_d(k))^2 \tag{4.7}$$

The confidence limits are calculated for a certain confidence level z between 0 and 1, with lower range values providing stricter limits. The value of z is chosen arbitrarily.

Applying and analysing these statistics involves observing the defined limits, allowing the identification of which measurements could have provoked the anomaly detection.

These methods help to obtain some insights on the studied datasets:

4.3.5 Full dataset

After tests with window sizes ranging from one to four weeks, the most sensitive results[1] for the MW-PCA were obtained when the size was 14 days. In addition, the confidence limit is chosen as 97.5%, and only periods that were at least 1 hour long were considered.

[1] Results that contained the highest quantity of anomalous samples.

Nine principal components were selected to perform the variable selection techniques (Figure 34a). The results in Table 6 show that the MW technique is more robust in identifying the periods that may be considered abnormal, showing lesser periods, while the Conventional and Robust identify the same number of periods. Due to its lower complexity, the Conventional PCA was the chosen method for the pre-processing of the data models.

Table 6 – Anomalous periods identified by different PCA approaches

PCA Approach	T^2 Statistic	Q-Statistic
Conventional PCA	15	3
Robust PCA	15	3
MW-PCA	12	3

Source: The author.

Figures 36a, 36c, and 36e show where the anomalous periods were identified by the T^2 statistic, while Figures 36b, 36d and 36f represent the ones identified by the Q-statistic. The anomalous periods are highlighted in yellow. Most anomalous periods were between 1 and 2 hours long and were identified especially during the beginning of the dataset and during the 18th and the 23rd of February, which refer to variations in the liquid temperature probably caused by the snow melt phenomenon.

4.3.6 Reduced Predictor Set

The optimal window size of the MW-PCA is 28 days, with tested sizes ranging from one to four weeks, the confidence limit is 97.5%, and only periods that were at least 1 hour long were considered.

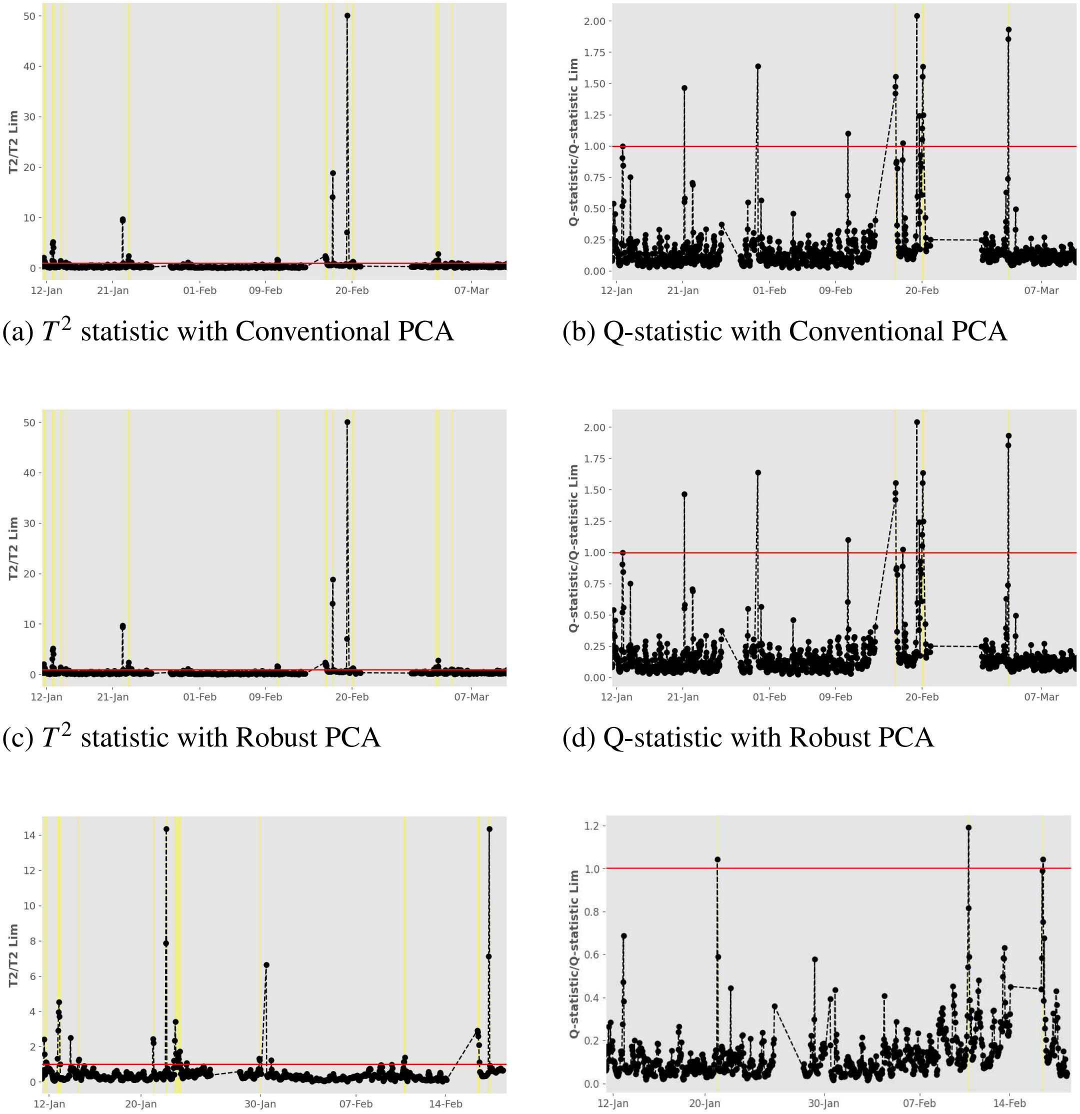
Five principal components were selected to perform the techniques, (Figure 34b). Also for this set of data, notice, in Table 7, that the MW-PCA has a more robust performance regarding the statistics and that the Conventional and Robust approaches yield the same results. As for the full dataset, the chosen method for the variable selection is also the Conventional PCA approach.

Table 7 – Anomalous periods identified by different PCA approaches

PCA Approach	T^2 Statistic	Q-Statistic
Conventional PCA	20	1
Robust PCA	20	1
MW-PCA	8	0

Source: The author.

Figure 36 – Sample selection statistics



(a) T^2 statistic with Conventional PCA

(b) Q-statistic with Conventional PCA

(c) T^2 statistic with Robust PCA

(d) Q-statistic with Robust PCA

(e) T^2 statistic with MW-PCA

(f) Q-statistic with MW-PCA

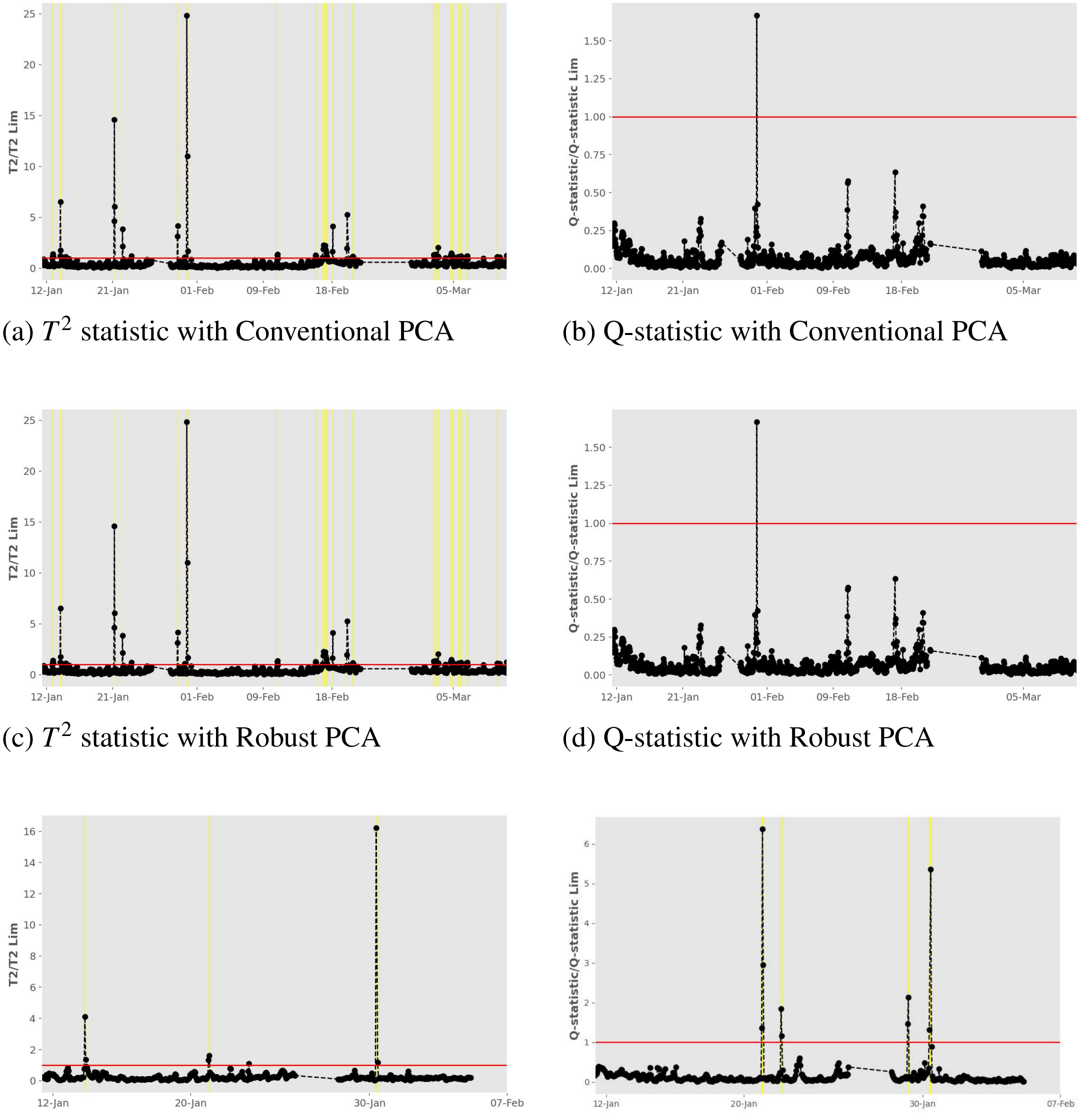
Source: The author.

Figures 37a, 37c, and 37e present the anomalous periods identified by the T^2 statistic, while Figures 37b, 37d and 37f show the ones identified by the Q-statistic. Notice that more anomalous periods are identified at the end of February and the beginning of March in the T^2 approach, while fewer periods are identified in the Q-statistic PCA.

4.4 Variable Selection

Machine learning techniques used in data-driven modelling procedures may become computationally expensive when a lot of dimensions (predictors) are used to train the models, as a large number of features have to be analysed. For this task, the Minimum Redundancy,

Figure 37 – Sample selection statistics



(a) T^2 statistic with Conventional PCA

(b) Q-statistic with Conventional PCA

(c) T^2 statistic with Robust PCA

(d) Q-statistic with Robust PCA

(e) T^2 statistic with MW-PCA

(f) Q-statistic with MW-PCA

Source: The author.

Maximum Relevance (mRMR) algorithm (DING; PENG, 2003) is used. It identifies the features that exhibit the highest relevance to the output variable while maintaining minimal redundancy among the previously chosen predictors.

The mRMR approach acknowledges the top m features that contain the highest relevance to the output while keeping minimal redundancy considering the previously chosen inputs, not regarding the machine model used (DING; PENG, 2003). The relevance, the mutual information, is calculated based on Equation (4.8) (PENG *et al.*, 2005), in which $p(x,y)$ is the joint probability function of a variable x and a variable y, $p(x)$ and $p(y)$ are the marginal probability distribution functions of x and y. In addition, the redundancy calculation follows Equation (4.9), which is the Variance Inflation Factor (VIF), where R^2 is the Coefficient of

Determination. This statistical metric analyses the intensity of multicollinearity, which is higher when at least two features present high correlation, by calculating the degree a predictor coefficient's variance is increased because of multicollinearity.

$$MI(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy \tag{4.8}$$

$$VIF = \frac{1}{1-R^2} \tag{4.9}$$

Then, it embodies a wrapper technique that applies the selection to obtain the Pareto Front solutions, which contain the ideal balance of model performance and quantity of input features.

The variable selection insights obtained in this research are presented in the two following subsections.

4.4.1 Full Dataset

To select the best number of features, the train set was tested with values chosen arbitrarily. These values represented the number of selected features, ranging from 11 to 200. The factor for best model selection considered was the smallest Root Mean Squared Error (RMSE) considering a linear regression with the number of selected features using the mRMR criterion.

The mRMR selection yielded a result of 18 features for the N_2O gas (represented in Table 8) and 40 for the CO_2 gas (shown in Table 9). Analysing the results, it is noticeable that these emissions are affected by different features. While the N_2O gas has more selected features which refer to the measured nitrate at the end of degas, the CO_2 gas relates to features more related to air flow rate. Considering what was observed in the data analysis procedure, most of the selected predictors present strong relations with the outputs.

Not all chosen predictors highly correlate with the outputs (Figures 31 and 32). Therefore, it is possible to conclude that, regarding the redundancy and relevance factors considered in the algorithm, features highly correlated to the outputs will not necessarily contribute to the best performance of the algorithms, but may be part of a subset of inputs that may yield an optimized outcome. Also, features showing lower correlation may be considered by the algorithm.

Table 8 – Features selected by the mRMR algorithm for the N_2O gas

Analysed Line	Selected Features
L1	D-NO3, Z5-QAIR, Z6-QAIR
L2	D-NO3, Z4-QAIR
L3	Z6-DO, D-NO3, Z5-QAIR, Z5-DO
L4	D-NO3
L5	Z4-DO, D-NO3
L6	D-NO3
L7	D-NO3, Z5-QAIR
L8	D-NO3, Z6-SS
L9	D-NO3

Source: The author.

Table 9 – Features selected by the mRMR algorithm for the CO_2 gas

Analysed Line	Selected Features
L1	AERO-QAIR, Z6-QAIR, Z4-QAIR, Z5-QAIR, Z3-QAIR, D-NO3
L2	AERO-QAIR, Z5-QAIR, Z6-QAIR, D-NO3
L3	AERO-QAIR, Z6-QAIR, Z5-QAIR, D-NO3
L4	AERO-QAIR, Z6-QAIR
L5	AERO-QAIR, Z5-QAIR, Z4-QAIR, Z6-QAIR, Z5-QAIR, D-NO3, Z4-QAIR
L6	Z5-QAIR, AERO-QAIR, Z6-QAIR, Z4-QAIR, Z3-QAIR
L7	Z6-SS, AERO-QAIR, Z5-QAIR, Z6-QAIR, Z4-QAIR
L8	AERO-QAIR, Z5-QAIR, Z6-QAIR
L9	AERO-QAIR, Z5-QAIR, Z6-QAIR, I-NH4

Source: The author.

4.4.2 Reduced Predictor Set

Considering the reduced predictor set, the mRMR Feature Selection algorithm evaluated 10 to 24 as the best number of features, with the smallest RMSE also being the best model's selection factor. Among 26 features, 15 were selected for the N_2O gas (Table 10) and 10 for the CO_2 gas (Table 11).

Evaluating the correlation values in Tables 10 and 11, notice that the features with the highest correlation, but not only them, contribute to creating regression models that yield optimal results.

Table 10 – Features selected by the mRMR algorithm for the N_2O gas

Selected Features
AVG-D-NO3
AVG-Z5-QAIR
AVG-Z2-DO
AVG-Z6-QAIR
AVG-Z4-QAIR
AVG-I-Q
L9-I-NH4
E-T
AVG-S1-QR
AVG-AERO-QAIR
AVG-S-QRTOT
AVG-D-ALK
AVG-S2-QR
AVG-Z4-DO
AVG-Z6-SS

Source: The author.

Table 11 – Features selected by the mRMR algorithm for the CO_2 gas

Selected Features
AVG-AERO-QAIR
AVG-Z6-SS
AVG-Z5-QAIR
AVG-Z6-QAIR
L9-I-NH4
AVG-Z4-QAIR
AVG-Z3-QAIR
AVG-D-NO3
E-T
AVG-I-Q

Source: The author.

5 DATA-DRIVEN MODELLING

This chapter discusses the theoretical background of data-driven modelling, presenting both linear and non-linear modelling strategies for GHG emissions. It also explains the metrics used to evaluate model performance. All the models are supervised techniques, which perform modelling strategies considering the dataset's features and outputs.

Regression data-driven modelling strategies estimate the relations between the inputs X and output Y by a certain function $f(X)$, estimated by the chosen regression model (Equation 5.1, where $\hat{Y}$ is the predicted output). Regarding GHGs, the goal is to obtain the emission value considering collected features. The dataset used for modelling is divided into the train set - used for training the model and obtaining its parameters - and the test set, used to obtain the model performance by fitting unknown samples to the model and comparing the resulting output with the real output value. Testing on unseen data is important to understand how the model can generalize its predictions apart from the characteristics that are learned during the training process.

$$\hat{Y} = f(X) \tag{5.1}$$

Moreover, the chosen machine learning models can either be linear or non-linear. Linear models predict the outputs using a linear function, making them less flexible but often more interpretable and computationally efficient. Their simplicity generally reduces the risk of overfitting, yet depending on the size and complexity of the dataset. Model flexibility is closely linked to its variance, which measures how sensitive the model is to variations in the training data. Therefore, higher flexibility can lead to higher variance, when the model can capture noise in the data, increasing the risk of overfitting. In contrast, bias refers to the error between the predicted output and the real output. A model with high bias may fail to capture key patterns in the data, resulting in underfitting.

5.0.1 Bias-variance Trade-off

When selecting the model to perform the predictions, the bias-variance trade-off should be accounted for. Less flexible models, such as linear, tend to present lower variance because they are less sensitive to fluctuations in the training data. However, their simplicity justifies their difficulty in capturing complex, non-linear patterns in the data, resulting in higher bias. On the other hand, models with higher variance, such as non-linear, generally exhibit lower bias, allowing them to capture more complex patterns in the data. However, this flexibility makes them more prone to overfitting, which is when the model is too adapted to the training dataset that it is not able to generalize to other samples' behaviours.

Therefore, when selecting a model approach, the linearity of the features with the output and the bias-variance trade-off are to be acknowledged. The following sections present the theoretical background on linear and non-linear regression models that will be used in this work. Most of the linear regression models were chosen to test a more interpretable approach in GHG modelling, while the non-linear models were chosen based on the literature.

5.1 Linear Models

5.1.1 Ordinary Least Squares Regression (OLS)

The Ordinary Least Squares (OLS) is a method used in linear regression to estimate the unknown parameters of the linear model minimizing the sum of squares of the differences between the observed and the estimated output generated by the linear function of the predictors.

When considering real-life applications, the datasets present more than one predictor, making X a vector with p parameters containing the dataset's predictors $[X_1, X_2, ..., X_p]^T$. Thus, each feature will have a separate slope coefficient in the linear equation, forming a coefficient of vector $\beta^T = [\beta_1, \beta_2, ..., \beta_p]^T$. For instance, a feature vector X will have its linear equation following Equation (5.3):

$$\hat{Y} = \hat{\beta}_0 + X^T \beta \tag{5.2}$$

This equation can also be represented as in 5.3:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 ... + \hat{\beta}_n X_n \tag{5.3}$$

These coefficients are unknown at first and can be obtained by using the predictor vector and arranging values for the coefficients that make the equation line as close as possible to the samples' location in the vectorial space. The Least Squares criterion (Equation (5.4), where

β is the coefficient vector, X is the predictor vector and Y is the outcome) The coefficients are obtained by the predictor and output matrices multiplications, with the resulting matrix of $X^T X$ being invertible (presenting full rank). If this is not the case, the sole use of the Least Squares criterion is not valid, but alternatives like penalization models can be implemented.

$$\beta = (X^T X)^{-1} X^T Y \tag{5.4}$$

In addition, the Least Squares derives the constants by choosing the ones that minimize the model's residuals (Equation (5.5)), which are the difference between the real output y_i and the predicted output $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i \tag{5.5}$$

The Residuals Sum of Squares (RSS) represents the regression's loss function. It is the sum of residuals of a dataset with n samples and is defined in Equation (5.6):

$$RSS = e_1^2 + e_2^2 + ... + e_n^2 \tag{5.6}$$

Considering the coefficients, the RSS equation can also be shown as:

$$RSS = \sum_{i=1}^{p} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} ... - \hat{\beta}_n X_{ip})^2 \tag{5.7}$$

The OLS algorithm is highly explainable and cheap in computational costs, but has some disadvantages: the non-linearity of the input-output relationship may lower its accuracy, and the presence of outliers and collinearity between variables also contribute to its poorer performance.

The three following subsections discuss other linear approaches that perform changes in the OLS equation in order to improve model predictions while maintaining simplicity and explainability.

5.1.2 Penalized Models

The penalized models, also known as shrinkage methods, seek to reduce the impact of factors, such as the presence of outliers, that reduce the OLS's accuracy (JAMES *et al.*, 2013).

These factors can cause an increase in the model's variance and may result in overfitting. In the penalized models, the linear function's coefficients can be constrained and shrunk, improving model performance. Although the model bias increases because of the bias-variance tradeoff, this growth is considered acceptable.

5.1.2.1 Ridge Regression

The Ridge Regression, in (5.8), is known as L2-Regularization. Its residuals sum square is added with a parameter λ that multiplies the square of each coefficient (the second order (L2) penalty). The tuning parameter λ is determined arbitrarily and the higher its value, the merrier shrinkage the model will have in its coefficients and the less flexible will the regression model be. So, as λ approaches infinity, the coefficients converge to zero.

$$SSE_{L2} = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{5.8}$$

5.1.2.2 Lasso Regression

The Lasso Regression, in Equation (5.9), is known as L1-Regularization. It has the residuals sum square added with a parameter λ that multiplies the module of each coefficient (the first order (L1) penalty). For this reason, this method is also considered a feature selection method, since some coefficients may be set to zero according to the chosen λ.

$$SSE_{L1} = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5.9}$$

5.1.3 Huber Regression

The Huber Regression provides a more robust approach than the OLS when dealing with data containing outliers, as it limits their impact on the model. Instead of using the Ordinary Least Squares criterion to calculate the model's coefficients, it uses the Huber loss function $L_\delta(e)$ (HUBER, 1964), in Equation (5.10). The penalization is equal to the Least Squares' half when the error e is inferior or equal to some threshold δ, assuming a quadratic behaviour. This threshold controls the number of samples considered outliers, becoming more robust the lower its value. However, when the error is superior, it assumes a linear one.

$$L_\delta(e) = \begin{cases} \dfrac{e^2}{2}, & \text{if } |e| \leq \delta, \\ \delta \cdot \left(|e| - \dfrac{\delta}{2} \right), & \text{otherwise.} \end{cases} \tag{5.10}$$

5.1.4 Bayesian Regression

In Bayesian linear regression, additional information is incorporated into the model through a prior probability distribution. By combining the likelihood function of the observed data with the prior, Bayes' theorem is used to update the belief, yielding a posterior distribution for model parameters β and allowing the prediction of Y. In its simplest and most common version, the output Y is derived from a normal probability distribution (Equation (5.11)) in which the mean is the product of the transpose of the coefficient matrix β^T and the predictor matrix X, and the variance is the product of the standard deviation σ and the identity matrix I (BISHOP; TIPPING, 2003).

$$Y \sim N(\beta^T X, \sigma^2 I) \tag{5.11}$$

5.1.5 Partial Least Squares

PLS is a dimension reduction regression model that uses a supervised approach. It maximizes the covariance between the predictors and the output by performing trials of finding component directions that relate both the outputs and the features (JAMES *et al.*, 2013).

The resulting set of features $Z_1, ..., Z_m$ is the linear combination of the dataset's features followed by the fit of a linear model with the Least Squares method. Each new feature is, like in the PCA approach, composed of each feature p of the predictor vector X multiplied by a loading parameter ϕ_i, as represented in the PCA equation (4.2). The more related to the output the variable is, the higher the value of ϕ.

5.2 Non-Linear Models

5.2.1 K-Nearest Neighbours

The KNN algorithm performs predictions by considering the k closest samples to the one that is being forecast. Since it solely calculates the distances between samples, without constructing an equation and obtaining its parameters, it is considered a non-parametric algorithm.

The distance between samples can be calculated using various metrics, with the Euclidean and the Mahalanobis distances being the most commonly used and the chosen metric depending on the chosen problem. The weighting of the distance may also vary depending on the selected metric, affecting the importance of each neighbour in the prediction.

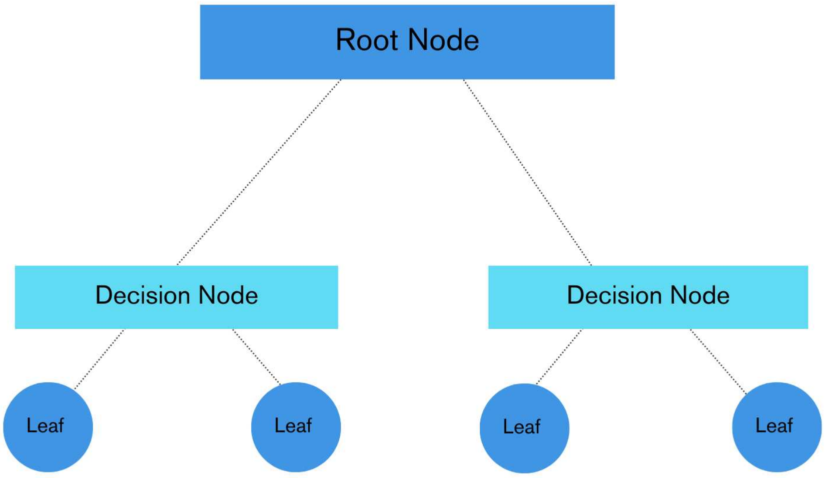
In addition, because of the algorithm's dependence on distance metrics, all the dataset's predictors must be centred and scaled before using the algorithm, since the feature's different measurements and the outliers may affect model performance.

5.2.2 Random Forest

The Random Forest model is an ensemble learning method that employs multiple Decision Trees to create models with improved predictive performance and generalization. A Decision Tree is a hierarchical structure that recursively partitions the feature space into J non-overlapping regions using splitting rules determined by the dataset's characteristics. At each tree node, the model iteratively refines splits based on a chosen criterion until a stopping criterion, such as the minimum number of samples per leaf (s) or the maximum tree depth (D), is met. Each sample is assigned to a single leaf node, which determines the output prediction as the mean of the target values within the considered region (JAMES *et al.*, 2013). Figure 38 presents a Decision Tree with a depth $D = 3$, resulting in four non-overlapping regions. It shows how the predictor space id divided at each tree level, with splits determined by the training data's characteristics.

Although Decision Trees can handle non-linear relationships between inputs and outputs well and are interpretable, their predictive accuracy performance is often lower because of high variance and overfitting tendency. Therefore, ensemble-learning methods such as the Random Forest are employed. They combine multiple Decision Trees in search of creating more robust and generalized model.

Figure 38 – Decision Tree structure



Source: The author.

In addition, the Random Forest model utilizes the Bagging (Bootstrap Aggregating) technique to introduce more randomness in the modelling process. In it, several Decision Trees are independently trained on random subsets of the available training data with replacement (bootstrap samples), and the predicted outcomes are obtained by the average of all the modelled trees' predictions. This method also improves Bagging by randomizing the features selected to model each Decision Tree. At each split, only a random predictor subset is considered, creating less correlated and more generalized trees, improving model performance and reducing overfitting. This model yields good results in datasets with high dimensionality or complex predictor interactions.

5.2.3 Extreme Gradient Boosting

Another ensemble approach is Boosting, which grows trees sequentially. In this method, each new Decision Tree is trained to improve the weaknesses of previously developed trees. The XGBoost method (CHEN; GUESTRIN, 2016) presents an advanced boosting algorithm that optimizes both the output prediction accuracy and computational efficiency. It also contains regularization techniques that help to reduce overfitting, making it a more generalized model.

As explained, theXGBoost builds trees sequentially, where each successive tree is trained to predict the residual errors[1] from the previous iteration. The model is represented by Equation (5.12):

[1] (Differences between the observed and predicted values)

$$y = \alpha_0 + \sum_{t=1}^{T} \sum_{i=1}^{n} w_{q(x_i,t)} \mathbb{I}(x_i \in \text{leaf}_t) \tag{5.12}$$

where α_0 is the initial prediction, $t = 1,2,3,4,...T$ is an indicator of the boosting step in the learning procedure, with T being the total trees, $w_{q(x_i,t)}$ is the leaves scores' vector associated with the leaf to which a sample x_i is assigned at a boosting step t, $q(x_i)$ is the function which assigns each data point to the corresponding leaf in the t-th tree, and $\mathbb{I}(x_i \in \text{leaf}_t)$ is indicator function equal to 1 if the sample x_i belongs to the leaf in the t-th tree, and 0 otherwise.

Each tree in the ensemble focuses on minimizing the residual errors from the previous tree. The optimization process follows a loss objective function, which is minimized in each iteration. To prevent overfitting, a regularization term is added to the loss function, allowing improved control of the model's complexity and more generalization.

In addition, some hyperparameters influence the XGBoost model's performance, such as the learning rate η, which controls the contribution of each tree's predictions to the resulting output. A smaller learning rate yields more conservative updates, while a larger rate may result in faster learning, but the model has a higher risk of overfitting. Also, the subsample ratio, s_r, is the proportion of the training data used to build each tree. By subsampling the data, XGBoost brings randomness into the modelling process, which helps to reduce overfitting and allows more generalization.

5.3 Hyperparameter optimization

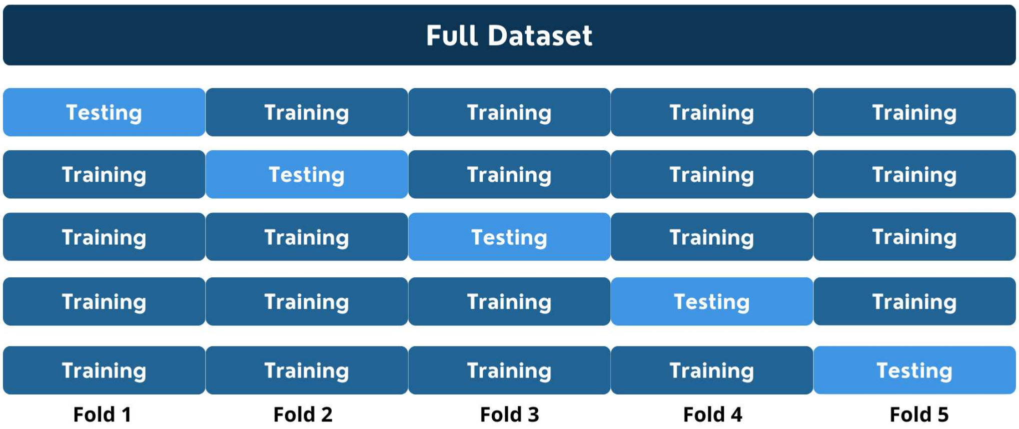
Many of the presented data-driven models have hyperparameters that can be defined by researchers. Different datasets will contain different values of them, and a way of optimizing these hyperparameters is by the Grid Search optimization algorithm. This approach tests all the provided hyperparameter value combinations in search of the best model possible and selects the values that obtain the best results. To obtain more generalized results, it generally employs cross-validation techniques.

5.3.1 Cross-validation

When first considering the dataset division into train and test sets, it is possible to notice that some samples are destined for the training set of the models and the remaining ones for the test set. Thus, the models are trained with a percentage of the samples and tested with another percentage. Therefore, to promote more mixed train and test sets, cross-validation is presented as a possibility of using the whole pre-processed dataset to improve machine learning models' results (HASTIE *et al.*, 2009).

The K-Fold cross-validation splits the data into K roughly equal-sized parts. For a certain k part, the model is fitted to the other K - 1 parts of the data, and its prediction error is calculated when predicting the k part of the data. This process is done iteratively until all the K parts are used as a test set, and the evaluation metrics are calculated by the mean of each train and test performance (HASTIE *et al.*, 2009). Figure 39 shows how a cross-validation of $K = 5$ performs in a certain dataset.

Figure 39 – Cross-validation of $K = 5$



Full Dataset

Testing	Training	Training	Training	Training
Training	Testing	Training	Training	Training
Training	Training	Testing	Training	Training
Training	Training	Training	Testing	Training
Training	Training	Training	Training	Testing
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Source: The author.

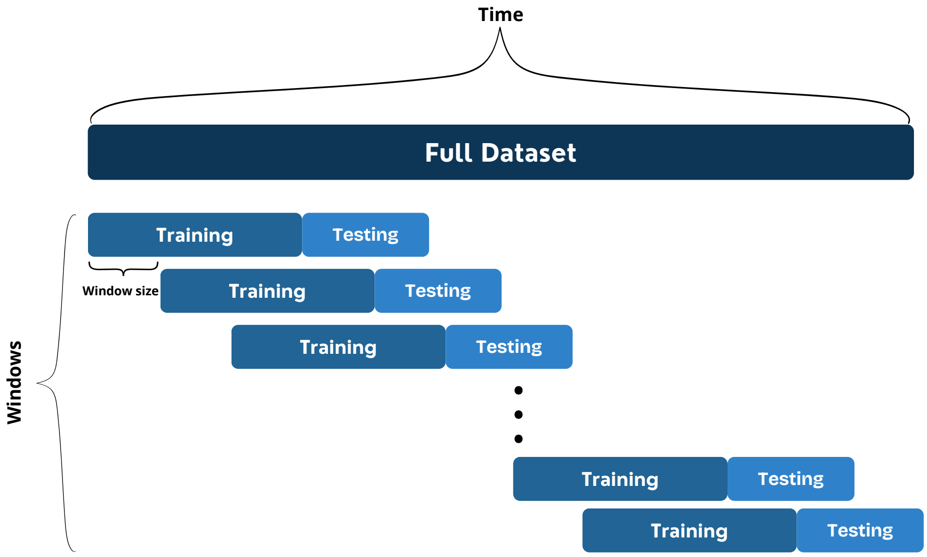
5.4 Rolling Window Models

With the RW approach, the models are fit in different periods, considering a defined window size. This improves the results of predictions for dynamic and time-variant variables since the windows can capture the system's dynamic behaviour better and improve prediction results.

For example, a RW model analyses a time series performance in a shorter period, while also having the simplicity of the OLS model. Cai and Juhl (2023) describe the rolling regression estimator as using a certain amount of observations, and each of the periods is indexed as r. The RW computes the regression in r periods or windows and then slides the window across the dataset. The size r is arbitrary. Figure 40 shows how the RW process occurs.

This technique can be implemented in many machine learning models, such as in the OLS, the Penalized, the Huber, the Bayesian, the PLS, the KNN, the Random Forest and the

Figure 40 – RW Regression process



Source: The author.

XGBoost Regressions.

5.5 Evaluation Metrics

To understand if the model can perform its predictions well, it must be evaluated. Some relevant methods to assess regression methods are RMSE and R^2. The first metric analyses how close the predictions were to the real values, and the second one checks if the tested model was able to fit the actual data.

The equation for the RMSE (5.13) shows that it is calculated as the root of the squared sum of the model's residuals divided by the number of predictions. This metric's output has the same unit as the model's output and it adds more weight to higher differences between predicted and real values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{5.13}$$

The equation for the R^2 (5.14) presents that the Determination Coefficient is 1 subtracted by dividing the sum of residuals by the sum of the squared subtraction of the real values and their mean value. This metric shows the proportion of outputs that can be predicted or explained by the features.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{5.14}$$

6 METHODOLOGY

This chapter outlines the methodology used in the experiments to support the discussion. It explains which parameters were chosen in each model. The results obtained regarding the pre-processing and the modelling are presented and discussed in the following chapter.

The pre-processing strategies and data-driven models were developed using Python programming language version 3.11.11 in the Google Colab environment. The `pandas` library (version 2.2.2) was used for loading and handling the collected data, while the `Matplotlib` (version 3.10.0) and `Seaborn` (version 0.13.2) libraries were employed to create visualizations for data analysis. In the pre-processing steps, the `scikit-learn` library (version 1.6.0) was used for normalization and the implementation of conventional PCA. For the Robust PCA technique, the `PyOD` library (version 2.0.3) was utilized, and for the mRMR algorithm, the `mrmr` library (version 0.2.8) was used.

The selected data-driven models used the scikit-learn library, except for XGBoost, which utilized the xgboost library version 2.1.3. The OLS, Ridge, Lasso, Huber, and Bayesian models were implemented with the `linear_model` module, PLS with the `cross_decomposition` module, Random Forest with the `ensemble` module, and KNN with the `neighbors` module. The Grid Search algorithm was implemented using the `model_selection` module. The R^2 value was calculated using the `r2_score` function from the `metrics` module, while the RMSE was computed using the square root of the Mean Squared Error (MSE) from the `metrics` module, with the help of the numpy library (version 1.26.4).

Missing samples were removed, the dataset was normalized and split into 25% for testing and 75% for training the machine learning models. This selection was made because the dataset has 12 weeks, so the first 9 were used to train the models and the last 3 to test them. The variables were not split randomly because of the dataset's time series characteristic, with shuffling samples in the train and test sets possibly causing model bias.

6.0.1 Model hyperparameters

The machine learning models were selected considering the models found in literature and their variations, ranging from linear to non-linear approaches. The Grid Search method was used for defining the best model hyperparameters, with a cross-validation of 5 folds, in both datasets. The tested hyperparameter values were chosen arbitrarily and simulation results are analysed by considering each model's RMSE and R^2 values.

The optimal encountered hyperparameters are given in Table 12. The window size unit is in hours. Also, the RW models used the same parameters as the default model's ones obtained by the Grid Search in search of understanding how the RW approach improves model results.

Table 12 – Regression model hyperparameters, considering the Full dataset (FD) and the Reduced dataet (RD)

Model	**Hyperparameter**	**Description**	$\mathbf{N_2O}_{FD}$	$\mathbf{CO_2}_{FD}$	$\mathbf{N_2O}_{RD}$	$\mathbf{CO_2}_{RD}$
RW-OLS	r	Window Size	3	6	5	4
Ridge	λ	Tuning Parameter	30	1.22	14.90	2.65
RW-Ridge	r	Window Size	168	168	336	336
Lasso	λ	Tuning Parameter	0.20	1.63	0.41	4.49
RW-Lasso	r	Window Size	168	24	336	24
Huber	δ	Threshold	1.35	1.35	1.35	1.35
RW-Huber	r	Window Size	336	336	336	336
RW-Bayesian	r	Window Size	501	501	501	501
PLS	m	Selected Components	4	2	3	5
RW-PLS	r	Window Size	168	168	24	168
KNN	k	Neighbours	49	27	43	24
RW-KNN	r	Window Size	336	336	336	336
Random Forest	D	Maximum Tree Depth	5	30	5	30
Random Forest	l_s	Minimum Leaves per Sample	2	1	2	1
Random Forest	s	Minimum Number of Samples per Split	5	2	5	5
Random Forest	n	Number of Trees	10	15	50	15
RW-Random Forest	r	Window Size	2	2	2	2
XGBoost	η	Learning Rate	0.1	0.1	0.1	0.1
XGBoost	D	Maximum Tree Depth	5	5	7	5
XGBoost	n	Number of Trees	100	80	100	80
XGBoost	s_r	Subsample Ratio	0.8	0.5	0.7	0.5
RW-XGBoost	r	Window Size	336	336	24	168

Source: The author.

7 RESULTS

This chapter discusses the results obtained by the employed modelling strategies in the previous chapter. The first section brings the results of the variable and feature selection approaches that yielded the best model metrics. The last section shows the data-driven outcomes and discusses how the RW approach improved some of their results.

7.1 GHG Modelling

The N_2O modelling results are presented in Table 13, with the top three best results in bold. Notice that the models performed better when the RW technique was employed, which can be explained by its ability to better capture the process dynamic behaviour. In the full dataset, the RW version of the OLS, Random Forest and XGBoost Regressions show the best results, demonstrating that the RW method allowed a simple linear machine learning model to effectively capture the dynamics of N_2O emissions. Additionally, it improved the performance of two more complex non-linear models. As for the reduced predictor set, although the metrics had a bit more bias, the results are comparable to the full dataset's.

The CO_2's modelling results are presented in Table 14, with the top three best results in bold. Notice that the models generally performed better in the CO_2 predictions compared to the N_2O ones, which may be due to the more linear relationship between the predictors and the CO_2 output. Notice that OLS, Lasso, and Bayesian Regressions showed the best OLS outcomes in the full dataset, with the OLS being the best model strategy, and that the RW technique improved almost all the models' results and obtained the best overall metrics. For the reduced predictor set, the RW OLS, XGBoost and Bayesian Regression presented the best overall results, which had slightly lower metrics than the ones from the full dataset. The RW technique also helped enhance the results of almost all the developed modelling techniques.

Considering that the datasets present missing values, the RW performance may be affected. Because the samples that present missing values are dropped, these dates are not considered in the modelling process, and therefore, samples from further dates are considered earlier in the window size. This can result in lower metric results and shows that, although this technique improves results, it also presents limitations. It is also important to note that all models will inevitably present bias, and that modelling strategies seek to balance models that present accuracy but without overfitting.

Table 13 – N_2O predictions, considering the FD and the RD datasets

Regression Model	RMSE$_{FD}$	R^2_{FD}	RMSE$_{RD}$	R^2_{RD}
OLS	7.7035	-6.2392	9.1470	-8.9743
RW-OLS	**1.8277**	**0.9853**	**2.9666**	**0.9577**
Ridge	6.4201	-4.0280	8.1210	-6.8622
RW-Ridge	4.2270	0.7707	4.1322	0.6761
Lasso	10.5697	-12.6281	9.8541	-10.5759
RW-Lasso	4.8230	0.7014	4.0599	0.6873
Huber Regression	4.0397	-0.9907	6.2214	-3.6142
RW-Huber Regression	1.7050	0.9081	2.2227	0.8330
Bayesian Regression	6.9885	-4.9578	7.9701	-6.5726
RW-Bayesian Regression	1.8666	0.8703	3.0065	0.7178
PLS	4.5768	-1.5553	4.9600	-1.9328
RW-PLS	3.9570	0.7985	5.7512	0.8429
KNN	6.8163	-4.6677	9.3620	-9.4485
RW-KNN	2.1463	0.8473	3.5267	0.6463
Random Forest	6.7519	-4.5611	8.5468	-7.7083
RW-Random Forest	**1.2097**	**0.9746**	**1.2517**	**0.9545**
XGBoost	5.6355	-2.8742	8.0431	-6.7120
RW-XGBoost	**2.1372**	**0.9121**	**4.5680**	**0.9009**

Source: The author.

Table 14 – CO_2 predictions, considering the FD and the RD datasets

Regression Model	RMSE$_{FD}$	R^2_{FD}	RMSE$_{RD}$	R^2_{RD}
OLS	213.8926	0.8942	305.6813	0.8095
RW-OLS	**116.9214**	**0.9856**	**125.3046**	**0.9817**
Ridge	222.4725	0.8856	183.2723	0.9315
RW-Ridge	207.4735	0.9570	221.6447	0.9431
Lasso	228.3290	0.8794	162.1333	0.9464
RW-Lasso	**154.9090**	**0.9746**	214.8448	0.9457
Huber Regression	313.5518	0.7727	196.0698	0.9216
RW-Huber Regression	168.6678	0.9509	**145.5647**	**0.9581**
Bayesian Regression	207.0421	0.9009	183.9768	0.9310
RW-Bayesian Regression	**164.0403**	**0.9603**	153.5918	0.9582
PLS	163.9791	0.9378	170.7134	0.9406
RW-PLS	234.0277	0.9430	184.0005	0.9623
KNN	192.4386	0.9144	371.5462	0.7186
RW-KNN	216.1806	0.9087	200.6064	0.9088
Random Forest	211.5556	0.8965	303.0723	0.8127
RW-Random Forest	166.4405	0.9328	167.8497	0.9361
XGBoost	193.7307	0.9132	249.2224	0.8734
RW-XGBoost	213.8536	0.9519	**191.6883**	**0.9590**

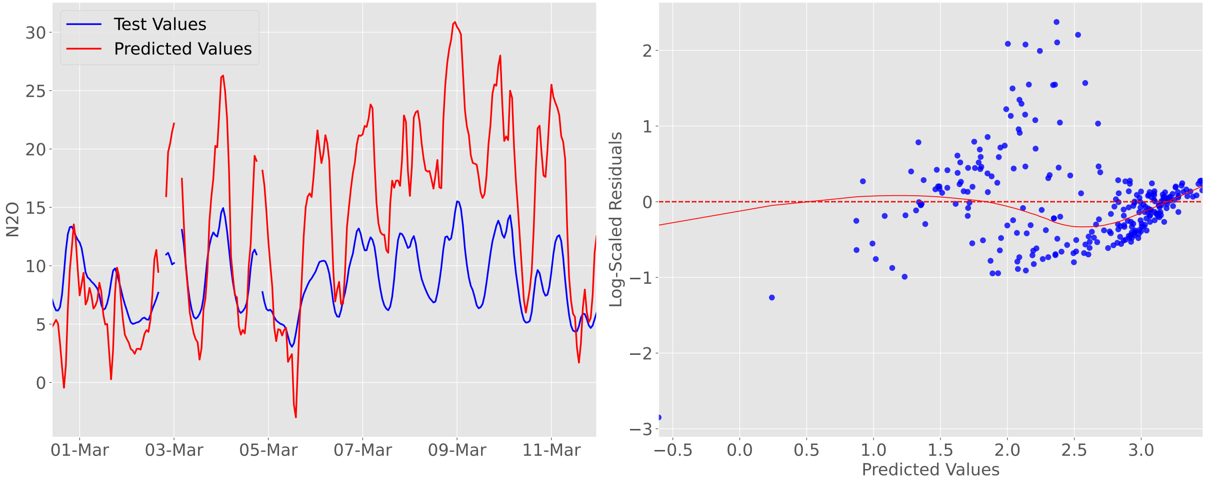
Source: The author.

Figures 41 to 44 present some of the N_2O residual plots for the full dataset, with the closer the values are to zero, the better fit to the data is the model, and real and predicted values in the temporal series plot, showing how the regression models fit compared to real data. More plots are shown in Appendix C. In the residual plots, the red dashed line represented the zero axis and the red straight line is the Locally Weighted Scatterplot Smoothing (LOWESS), which is a non-parametric regression method used to capture the trend of the plotted data. The values are in logarithmic scale for more interpretability.

The predicted values for the OLS and KNN (Figures 41 and 43, respectively), present offsets considering the test dataset values. Specifically considering the KNN model, it is observed that the predictions follow a similar pattern to the tested values but present an offset, which may have been caused, for example, by characteristics of the training dataset during the model fitting process.

Also, notice that the RW (Figures 42 and 44) approach can create a model that adapts smoothly to the dataset, especially considering the OLS model, while the models without it don't present feasible fits in the test set, having residuals that are further from the origin. The gaps in the plots refer to the indexes in which output samples were excluded from the testing dataset, either because they stored missing values or because their respective input sample stored a missing value in at least one of its features.

Figure 41 – N_2O OLS Regression in Full Dataset

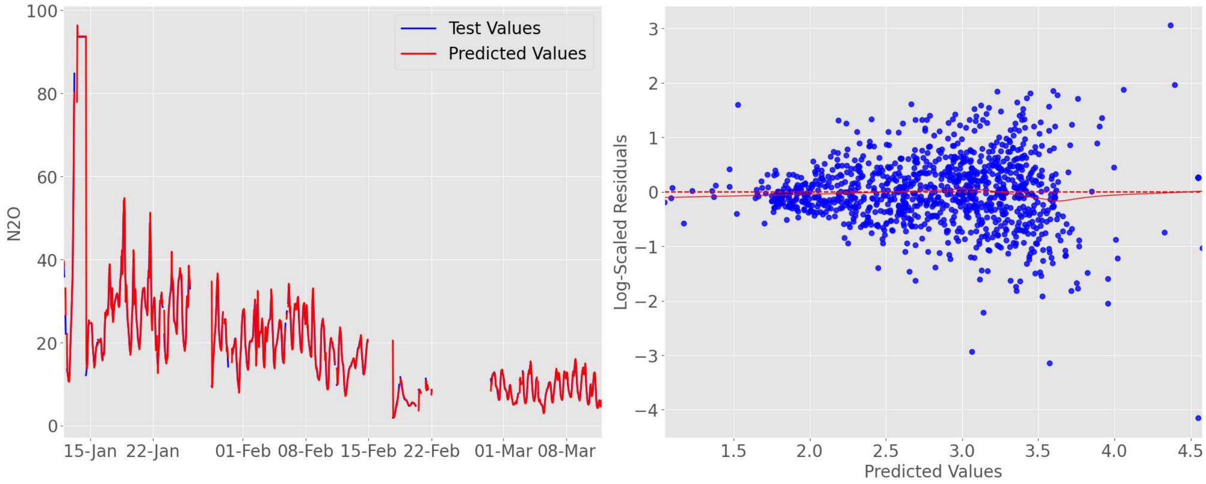


Source: The author.

Moreover, Figures 45 to 48 show some of the CO_2 residual plots and real and predicted values in the temporal series for the full dataset, showing how the regression models fit compared to real data. More plots are also presented in Appendix C Although the models developed without the RW approach already present good models, the incorporation of the method improves the modelling strategy, especially the OLS, which is the best model, yielding an adaptative and dynamic model.

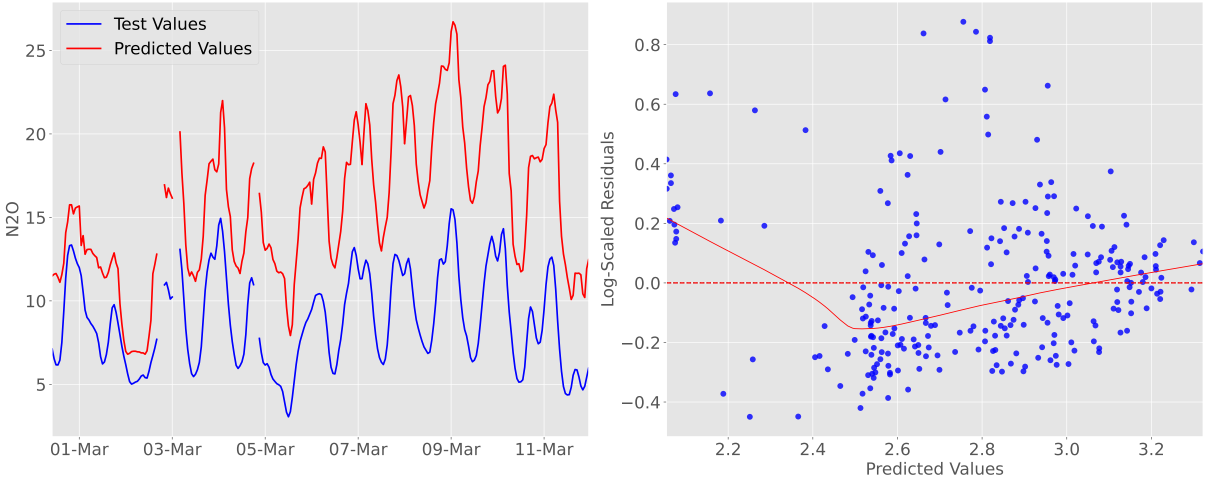
Therefore, the results show that the use of linear modelling strategies is achievable in both outputs by using the RW technique, which captures the dataset's dynamics and develops

Figure 42 – N_2O RW OLS Regression in Full Dataset



Source: The author.

Figure 43 – N_2O KNN Regression in Full Dataset



Source: The author.

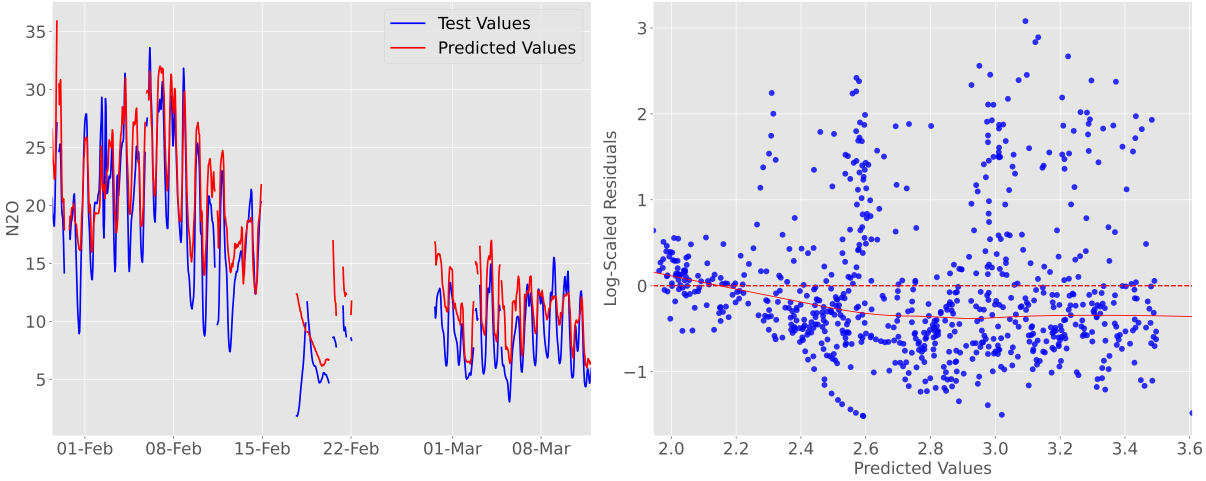
a model that is not costly computationally, and explainable.

The plots in Figures 49 to 52 follow the same structure as the ones presented before, but refer to the reduced predictor set models. Again, notice that the RW approach, especially in the OLS approach (Figure 50) produces good results for models, capturing the gas' dynamics.

In addition, Figures 53 to 56 present the results for some of the CO_2 gas regression models. Observe how the models without the RW approach already have a good fit with the test data but have their performance improved when the technique is incorporated.

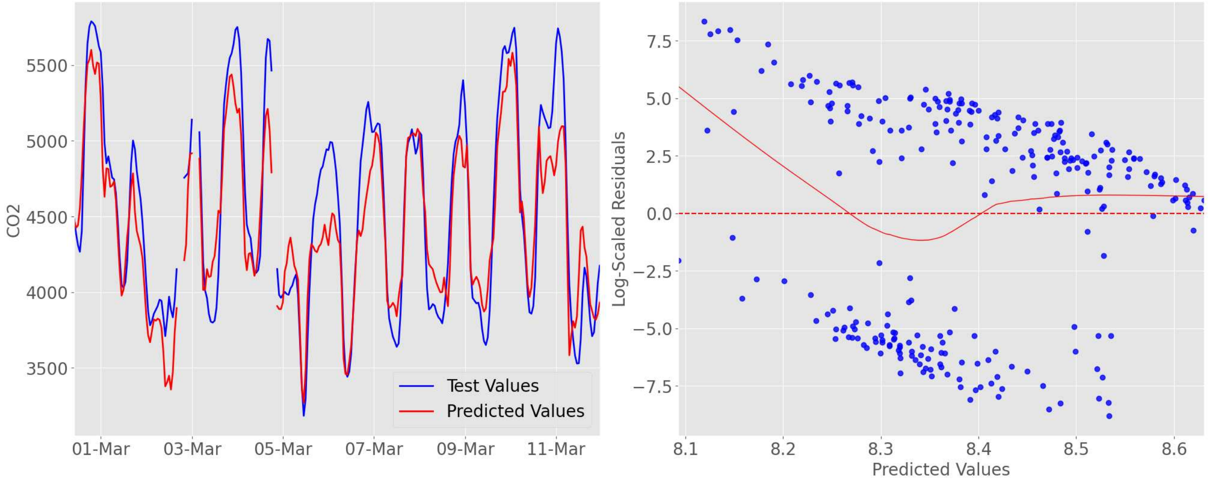
These results show that a linear approach with the RW approach, even regarding its limitations, is also employable in the reduced predictor set. This configuration not only produces more generalized modelling strategies, which could be expanded to other plants with similar

Figure 44 – N_2O RW KNN Regression in Full Dataset



Source: The author.

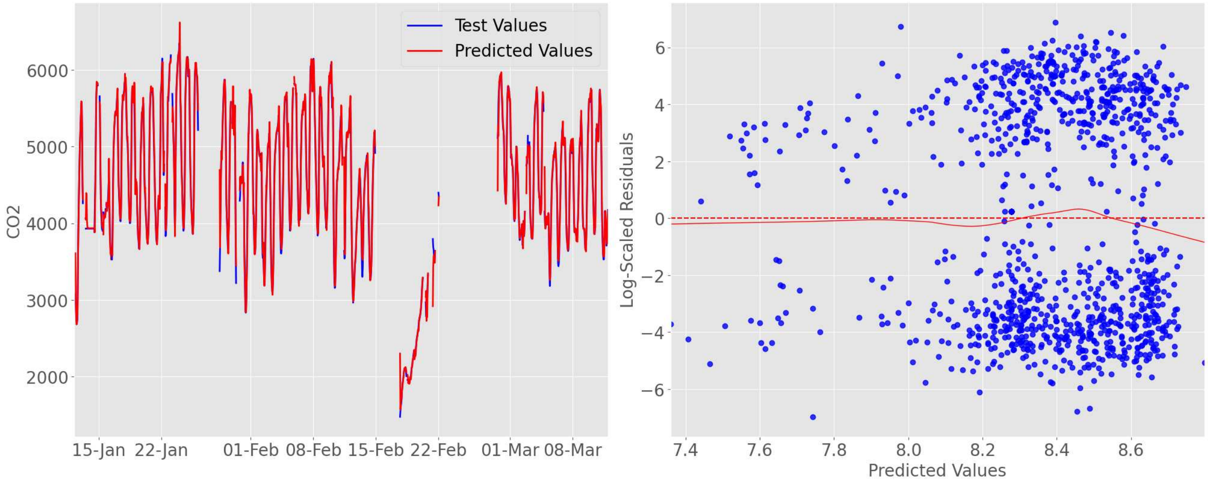
Figure 45 – CO_2 OLS Regression in Full Dataset



Source: The author.

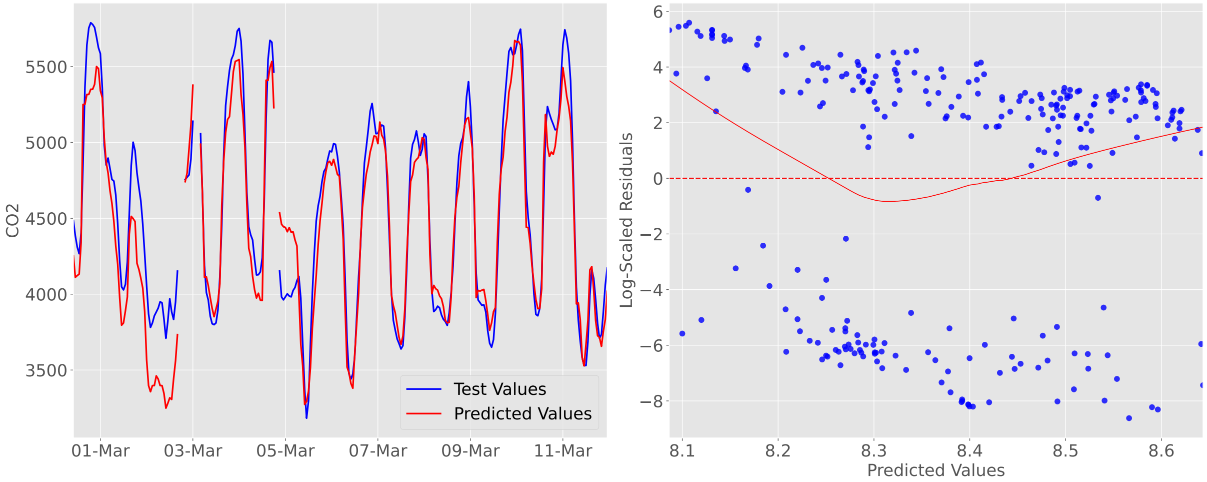
characteristics to the Viikinmäki WWTP, but is also less costly computationally because of its reduced number of predictors.

Figure 46 – CO_2 RW OLS Regression in Full Dataset



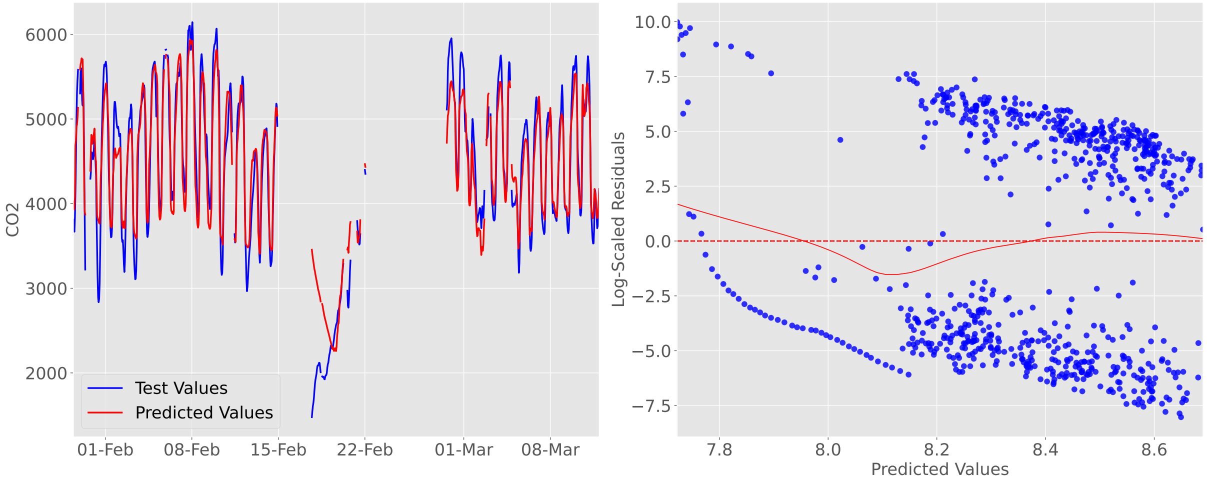
Source: The author.

Figure 47 – CO_2 KNN Regression in Full Dataset



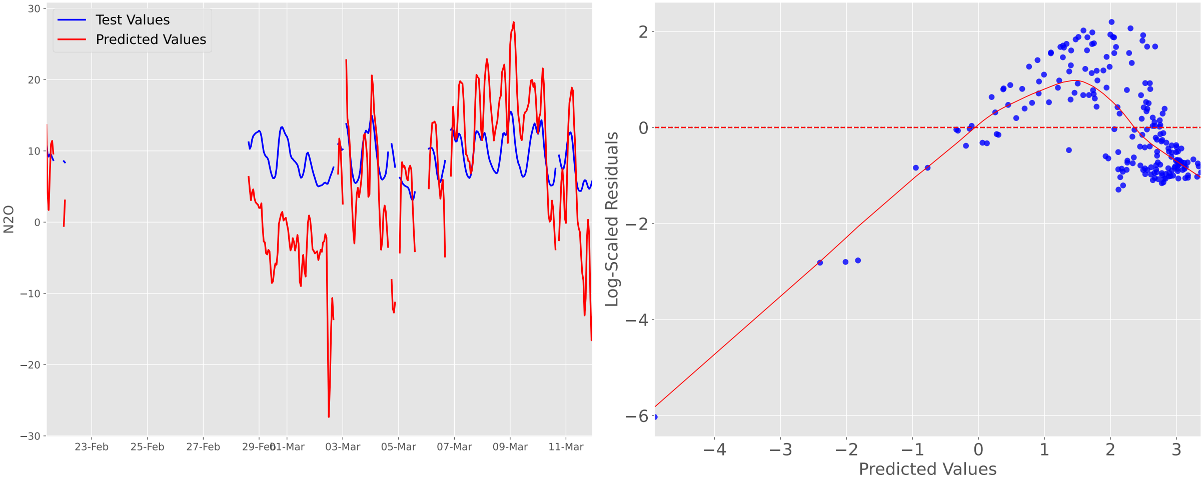
Source: The author.

Figure 48 – CO_2 RW KNN Regression in Full Dataset



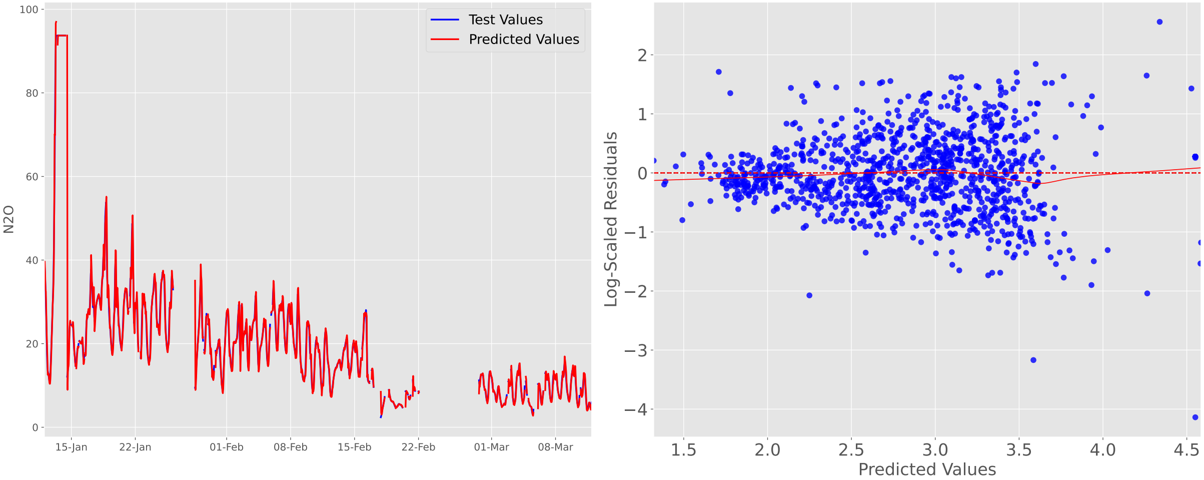
Source: The author.

Figure 49 – N_2O OLS Regression in Reduced Predictor Set



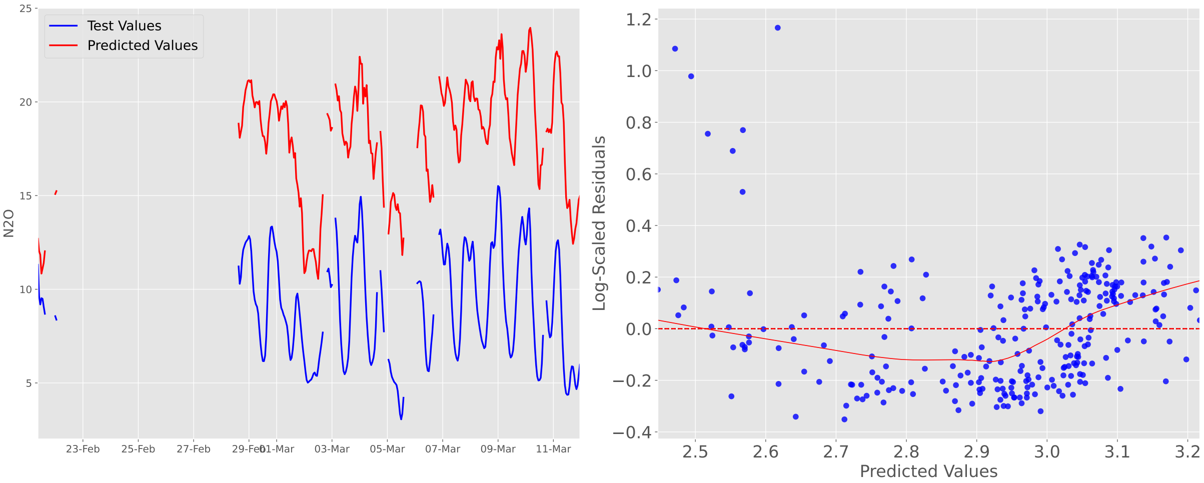
Source: The author.

Figure 50 – N_2O RW OLS Regression in Reduced Predictor Set



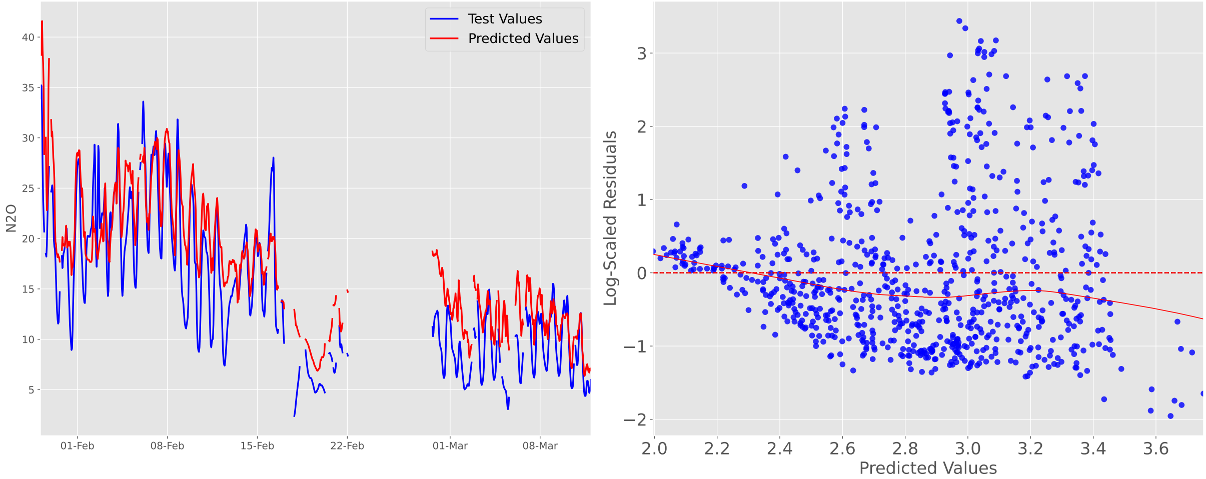
Source: The author.

Figure 51 – N_2O KNN Regression in Reduced Predictor Set



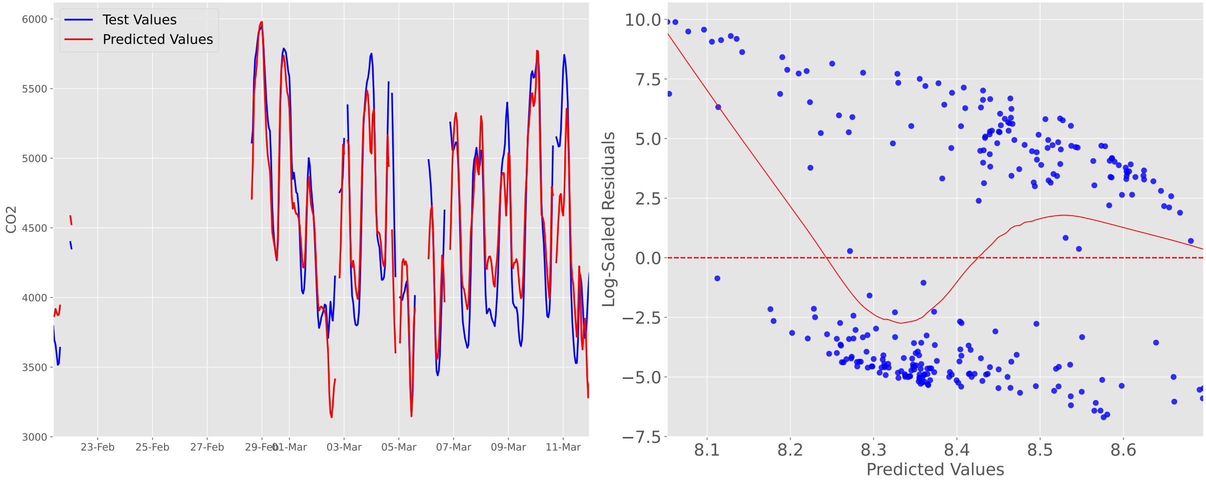
Source: The author.

Figure 52 – N_2O RW KNN Regression in Reduced Predictor Set



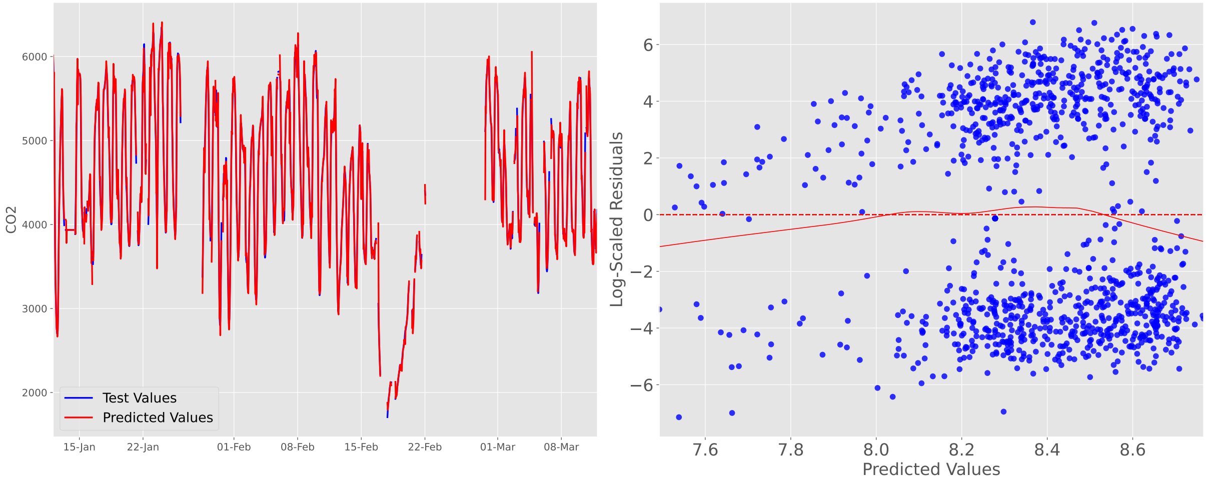
Source: The author.

Figure 53 – CO_2 OLS Regression in Reduced Predictor Set



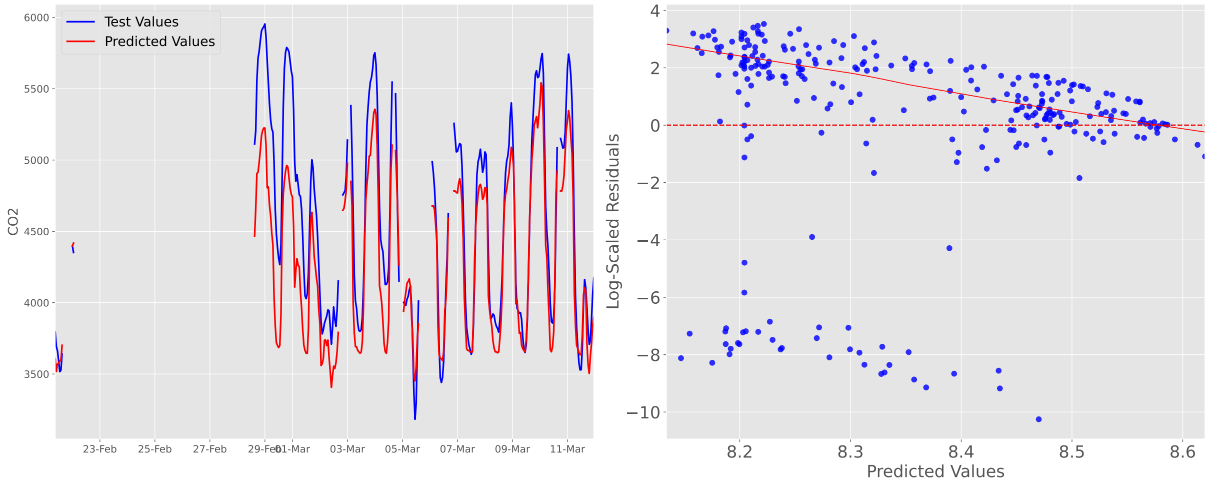
Source: The author.

Figure 54 – CO_2 RW OLS Regression in Reduced Predictor Set



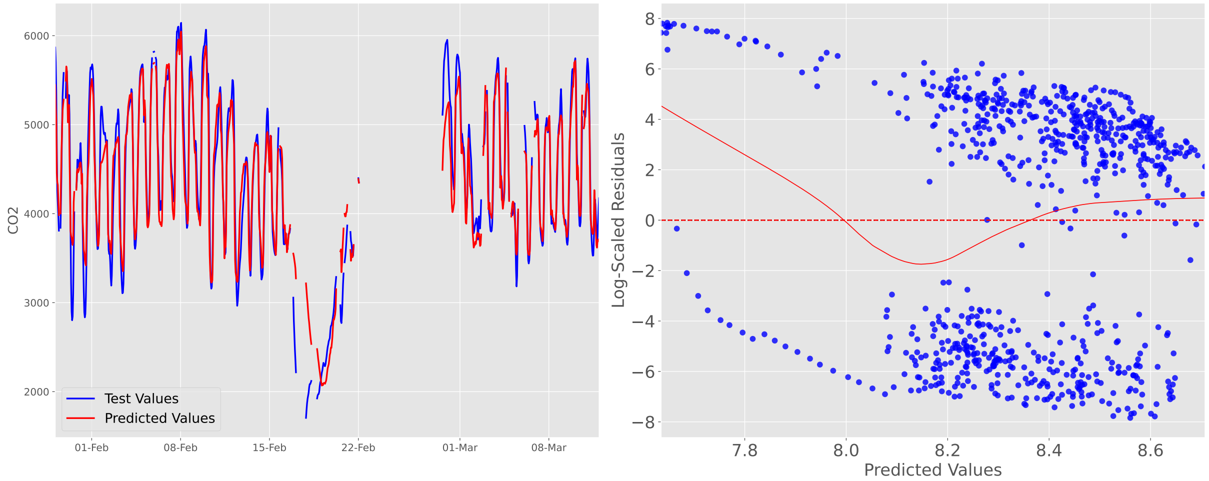
Source: The author.

Figure 55 – CO_2 KNN Regression in Reduced Predictor Set



Source: The author.

Figure 56 – CO_2 RW KNN Regression in Reduced Predictor Set



Source: The author.

8 CONCLUSION

This work presented data-driven modelling techniques for GHG prediction in the wastewater treatment sector. In this document, the concepts and methods explained were applied to data from the Viikinmäki WWTP. Moreover, modelling approaches were presented and discussed.

The experiments demonstrated that data analysis methods enhance the understanding of the relations between the WWTP's features and the N_2O and CO_2 emissions. It was possible to define the predictors that were more closely related to the outputs and understand more how changes in the variables' behaviour affected the emissions. Additionally, pre-processing strategies improved model performance, with variable and feature selection enabling the detection of anomalous periods that could affect model performance and the creation of less redundant, more efficient models.

The outcomes also presented that it is possible to perform the modelling of GHGs using linear and non-linear machine learning models. Although the CO_2 gas modelling yields better results at first because of its more linear relationship with the features, the strategies applied to predict N_2O gas emission also had acceptable results when performing the RW technique, which allowed the machine learning models to capture more of the gas' dynamic behaviour. In addition, linear models, which have lower computational costs and are more explainable, performed well with both gases using the RW, showing that it is possible to obtain a good model for the emissions using a simple approach. Feasible results were also observed in the reduced predictor set configuration, which presented a more generalized option, distancing itself from plant-driven modelling. In this dataset, the RW technique also improved results compared to its absence, showing that, despite dataset limitations, this strategy can enhance model performance.

This work, however, could not discuss the factors that impacted the variations in the N_2O gas' behaviour. This occurred because of the dataset's size, which had few samples collected during winter in Finland, with the presented results not being able to attest to the seasonality of the gas' emissions, for example. Future work will focus on collecting additional data to explore this issue further and to evaluate how the developed models perform across different seasons, as varying conditions may impact the behaviour of the features.

REFERENCES

AL, R.; BEHERA, C. R.; ZUBOV, A.; GERNAEY, K. V.; SIN, G. Meta-modeling based efficient global sensitivity analysis for wastewater treatment plants – an application to the bsm2 model. **Computers & Chemical Engineering**, v. 127, p. 233–246, 2019.

ANDRADE, J. C. S.; CONCEIÇÃO, G. C. A.; MARINHO, M. M. de O. Comparing urban greenhouse gas emission inventories in Brazilian cities. **Latin American J. of Management for Sustainable Development**, 2021.

ATKINSON, A. C.; RIANI, M.; CERIOLI, A. **Multivariate Data and the Forward Search**. New York, NY: Springer New York, 2004. 31–88 p.

BAHRAMIAN, M.; DERELI, R. K.; ZHAO, W.; GIBERTI, M.; CASEY, E. Data to intelligence: The role of data-driven models in wastewater treatment. **Expert Systems with Applications**, v. 217, p. 119453, 2023.

BANI SHAHABADI, M.; YERUSHALMI, L.; HAGHIGHAT, F. Estimation of greenhouse gas generation in wastewater treatment plants – Model development and application. **Chemosphere**, v. 78, n. 9, p. 1085–1092, 2010.

BISHOP, C.; TIPPING, M. Bayesian Regression and Classification. In: ____. **Advances in Learning Theory: Methods, Models and Applications**. [S.l.]: IOS Press, 2003. (NATO Science Series, III: Computer and Systems Sciences), p. 267–285. ISBN 978-1-58603-341-5.

BRASIL. **Lei nº 15.042, de 2024**. 2024. *Disponibiliza medidas relativas a (detalhar tema, se necessário). Diário Oficial da União, Brasília, DF, 2024.* Disponível em: <https://legislacao.presidencia.gov.br/atos/?tipo=LEI&numero=15042&ano=2024&ato=997QTQE1UNZpWTb27>. Acesso em: 12 jan. 2025.

CAI, Z.; JUHL, T. The distribution of rolling regression estimators. **Journal of Econometrics**, v. 235, n. 2, p. 1447–1463, 2023.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2016. (KDD '16), p. 785–794.

DAELMAN, M. R. J.; VOORTHUIZEN, E. M. van; DONGEN, L. G. J. M. van; VOLCKE, E. I. P.; LOOSDRECHT, M. C. M. van. Methane and nitrous oxide emissions from municipal wastewater treatment – results from a long-term study. **Water Science and Technology**, v. 67, n. 10, p. 2350–2355, 05 2013.

DING, C.; PENG, H. Minimum redundancy feature selection from microarray gene expression data. In: **Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003**. [S.l.: s.n.], 2003. p. 523–528.

European Commission. 'fit for 55': Delivering the eu's 2030 climate target on the way to climate neutrality. **Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions**, European Commission Brussels, Belgium, 2021.

GULHAN, H.; COSENZA, A.; MANNINA, G. Modelling greenhouse gas emissions from biological wastewater treatment by GPS-X: The full-scale case study of Corleone (Italy). **Science of The Total Environment**, v. 905, p. 167327, 2023.

HAIMI, H. **Data-derived soft sensors in biological wastewater treatment: With application of multivariate statistical methods**. Phd Thesis (PhD Thesis) — Aalto University, 02 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. [S.l.]: Springer, 2009.

Helsinki Region Environmental Services Authority (HSY). **Viikinmäki Wastewater Treatment Plant**. 2021. Accessed: 2025-01-05.

HUANG, L.; LI, H.; LI, Y. Greenhouse gas accounting methodologies for wastewater treatment plants: A review. **Journal of Cleaner Production**, v. 448, p. 141424, 2024.

HUBER, P. J. Robust Estimation of a Location Parameter. **Annals of Mathematical Statistics**, v. 35, p. 492–518, 1964.

HWANGBO, S.; AL, R.; SIN, G. An integrated framework for plant data-driven process modeling using deep-learning with Monte-Carlo simulations. **Computers and Chemical Engineering**, v. 143, p. 107071, 2020. ISSN 0098-1354.

IPCC. **2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories: Overview**. 2019.

IPCC. Index. In: ____. **Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change**. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2021. p. In press.

JACKSON, J. E.; MUDHOLKAR, G. S. Control Procedures for Residuals Associated With Principal Component Analysis. **Technometrics**, ASA Website, v. 21, n. 3, p. 341–349, 1979.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: with Applications in R**. [S.l.]: Springer, 2013.

KAZEMI, P.; BENGOA, C.; STEYER, J.-P.; GIRALT, J. Data-driven techniques for fault detection in anaerobic digestion process. **Process Safety and Environmental Protection**, v. 146, p. 905–915, 2021.

KAZEMI, P.; GIRALT, J.; BENGOA, C.; MASOUMIAN, A.; STEYER, J.-P. Fault detection and diagnosis in water resource recovery facilities using incremental PCA. **Water Science and Technology**, v. 82, n. 12, p. 2711–2724, 08 2020.

KHALIL, M.; ALSAYED, A.; LIU, Y.; VANROLLEGHEM, P. A. Machine learning for modeling N2O emissions from wastewater treatment plants: Aligning model performance, complexity, and interpretability. **Water Research**, v. 245, p. 120667, 2023.

KHALIL, M.; ALSAYED, A.; LIU, Y.; VANROLLEGHEM, P. A. An integrated feature selection and hyperparameter optimization algorithm for balanced machine learning models predicting N2O emissions from wastewater treatment plants. **Journal of Water Process Engineering**, v. 63, p. 105512, 2024.

KOSONEN, H.; HEINONEN, M.; MIKOLA, A.; HAIMI, H.; MULAS, M.; CORONA, F.; VAHALA, R. Nitrous Oxide Production at a Fully Covered Wastewater Treatment Plant: Results of a Long-Term Online Monitoring Campaign. **Environmental Science & Technology**, v. 50, n. 11, p. 5547–5554, 2016.

KRUGER, U.; XIE, L. **Advances in Statistical Monitoring of Complex Multivariate Processes: With Applications in Industrial Process Control. Statistics in Practice**. [S.l.]: John Wiley and Sons, 2012.

LU, H.; WANG, H.; WU, Q.; LUO, H.; ZHAO, Q.; LIU, B.; SI, Q.; ZHENG, S.; GUO, W.; REN, N. Automatic control and optimal operation for greenhouse gas mitigation in sustainable wastewater treatment plants: A review. **Science of The Total Environment**, v. 855, p. 158849, 2023.

MAKTABIFARD, M.; BLOMBERG, K.; ZABOROWSKA, E.; MIKOLA, A.; MąKINIA, J. Model-based identification of the dominant N2O emission pathway in a full-scale activated sludge system. **Journal of Cleaner Production**, v. 336, p. 130347, 2022.

MEHRANI, M.-J.; BAGHERZADEH, F.; ZHENG, M.; KOWAL, P.; SOBOTKA, D.; MAKINIA, J. Application of a hybrid mechanistic/machine learning model for prediction of nitrous oxide (N2O) production in a nitrifying sequencing batch reactor. **Process Safety and Environmental Protection**, v. 162, p. 1015–1024, 2022.

NOMIKOS, P.; MACGREGOR, J. F. Multivariate SPC Charts for Monitoring Batch Processes. **Technometrics**, ASA Website, v. 37, n. 1, p. 41–59, 1995.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 8, p. 1226–1238, 2005.

QADIR, M.; DRECHSEL, P.; CISNEROS, B. J.; KIM, Y.; PRAMANIK, A.; MEHTA, P.; OLANIYAN, O. Global and regional potential of wastewater as a water, nutrient and energy source. **Natural Resources Forum**, v. 44, n. 1, p. 40–51, 2020.

RAMÍREZ-MELGAREJO, M.; REYES-FIGUEROA, A.; GASSÓ-DOMINGO, S.; GÜERECA, L. P. Analysis of empirical methods for the quantification of N2O emissions in wastewater treatment plants: Comparison of emission results obtained from the IPCC Tier 1 methodology and the methodologies that integrate operational data. **Science of The Total Environment**, v. 747, p. 141288, 2020.

RUIZ-GAZEN, A. A very simple robust estimator of a dispersion matrix. **Computational Statistics and Data Analysis**, v. 21, n. 2, p. 149–162, 1996.

SESHAN, S.; POINAPEN, J.; ZANDVOORT, M. H.; van Lier, J. B.; KAPELAN, Z. Forecasting nitrous oxide emissions from a full-scale wastewater treatment plant using LSTM-based deep learning models. **Water Research**, v. 268, p. 122754, 2025.

SONG, C.; ZHU, J.-J.; WILLIS, J. L.; MOORE, D. P.; ZONDLO, M. A.; REN, Z. J. Oversimplification and misestimation of nitrous oxide emissions from wastewater treatment plants. **Nature Sustainability**, v. 7, n. 10, p. 1348–1358, 2024.

SZELĄG, B.; ZABOROWSKA, E.; MąKINIA, J. An algorithm for selecting a machine learning method for predicting nitrous oxide emissions in municipal wastewater treatment plants. **Journal of Water Process Engineering**, v. 54, p. 103939, 2023.

United Nations. **Sustainable Development Goals**. 2015. Accessed: 2025-01-07.

United Nations World Water Assessment Programme. **Wastewater: The Untapped Resource**. Paris, France: The United Nations World Water Development Report, 2017. Accessed: 2025-01-07. Available at: <https://www.unesco.org/reports/wwdr/2017/wastewater-untapped-resource>.

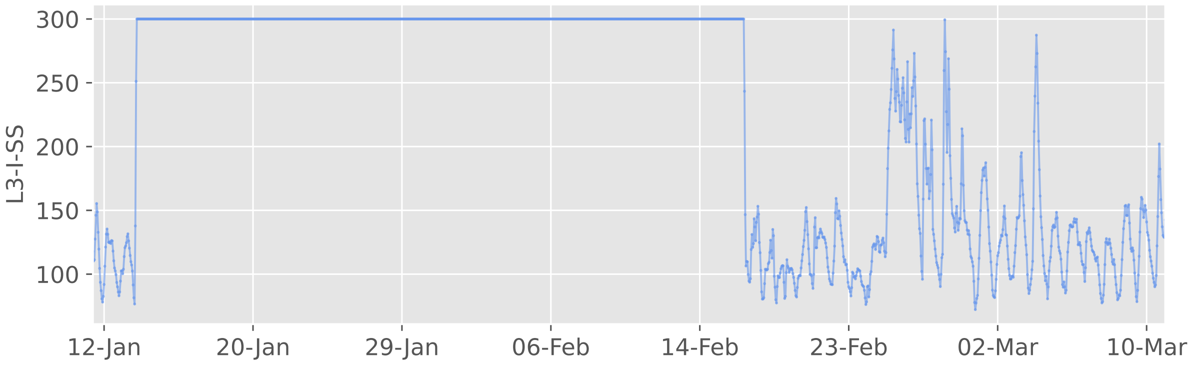
VASILAKI, V.; VOLCKE, E.; NANDI, A.; van Loosdrecht, M.; KATSOU, E. Relating N2O emissions during biological nitrogen removal with operating conditions using multivariate statistical techniques. **Water Research**, v. 140, p. 387–402, 2018.

WANG, Y.-Q.; WANG, H.-C.; SONG, Y.-P.; ZHOU, S.-Q.; LI, Q.-N.; LIANG, B.; LIU, W.-Z.; ZHAO, Y.-W.; WANG, A.-J. Machine learning framework for intelligent aeration control in wastewater treatment plants: Automatic feature engineering based on variation sliding layer. **Water Research**, v. 246, p. 120676, 2023.

ZHANG, X.; HE, L.; ZHANG, J.; WHITING, M. D.; KARKEE, M.; ZHANG, Q. Determination of key canopy parameters for mass mechanical apple harvesting using supervised machine learning and principal component analysis (pca). **Biosystems Engineering**, v. 193, p. 247–263, 2020.

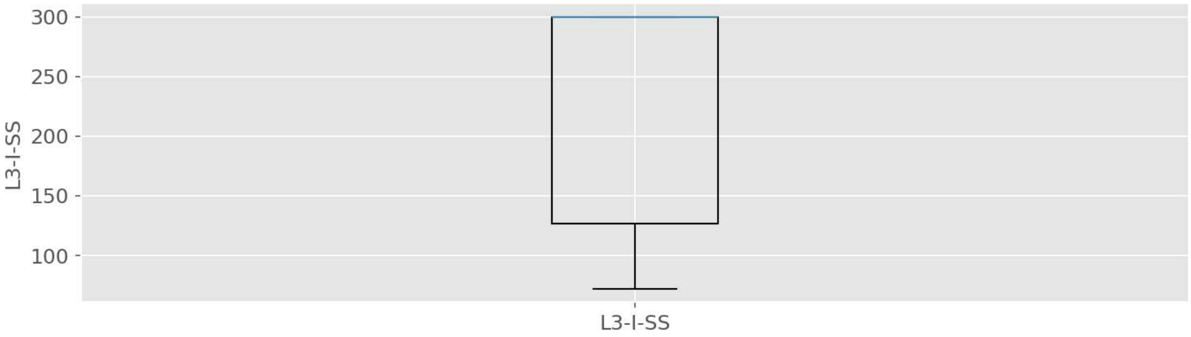
APPENDIX A – DATASET'S FEATURES EXPLORATORY ANALYSIS

Time-series of the influent total suspended solids in line 3



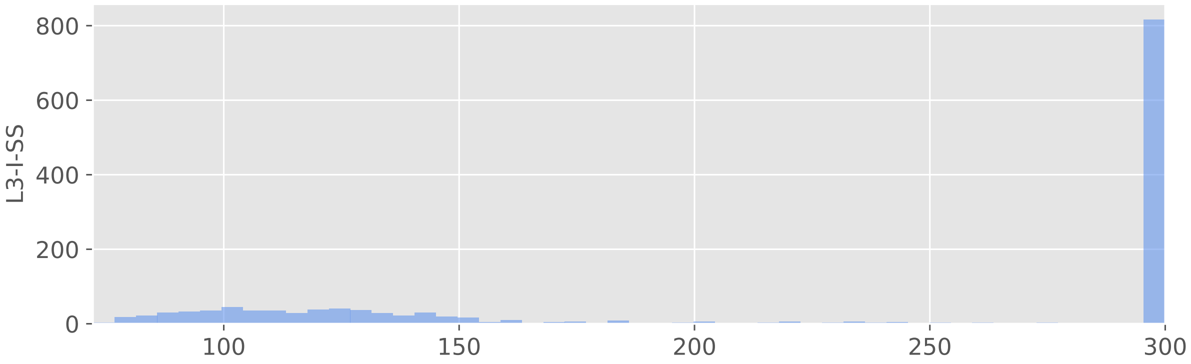
Source: The author.

Box plot of the influent total suspended solids in line 3



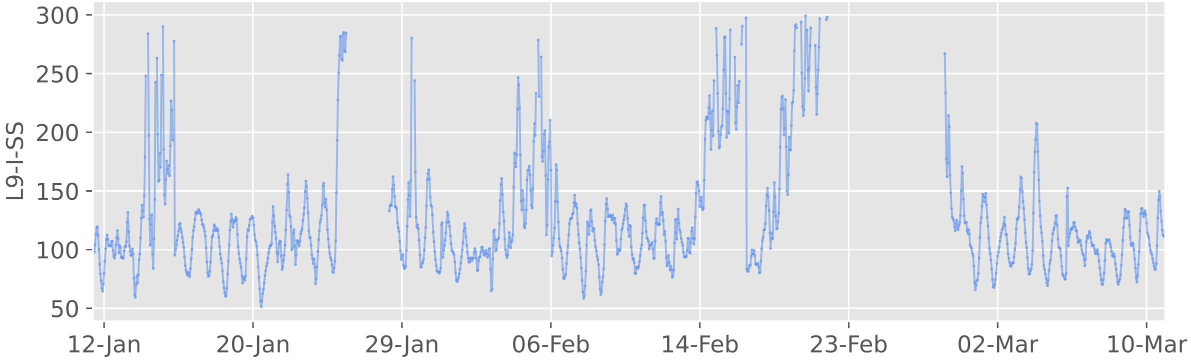
Source: The author.

Histogram of the influent total suspended solids in line 3



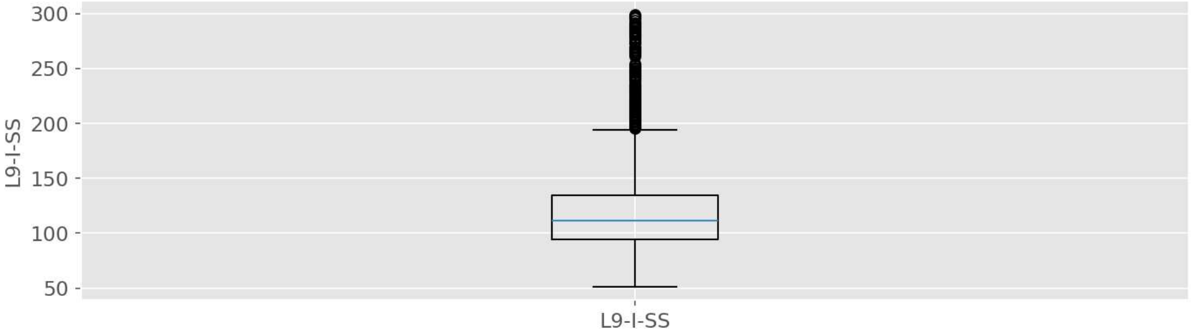
Source: The author.

Time-series of the influent total suspended solids in line 9



Source: The author.

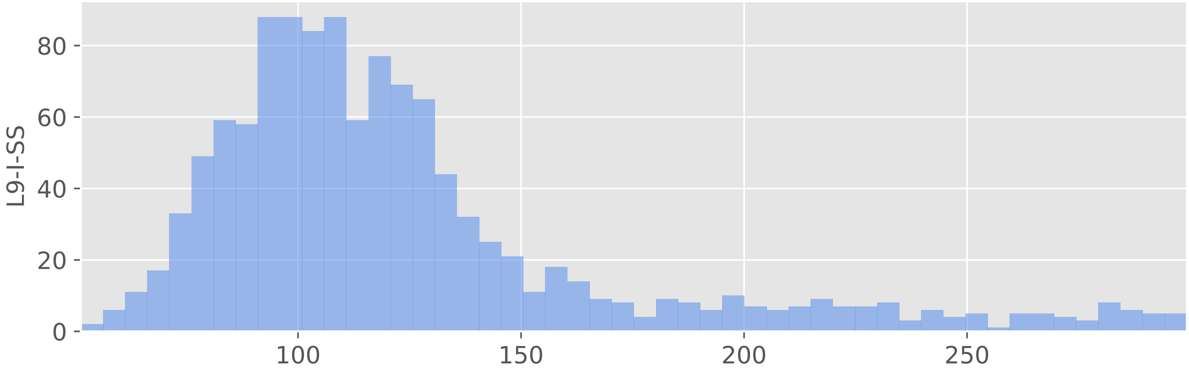
Box plot of the influent total suspended solids in line 9



Source: The author.

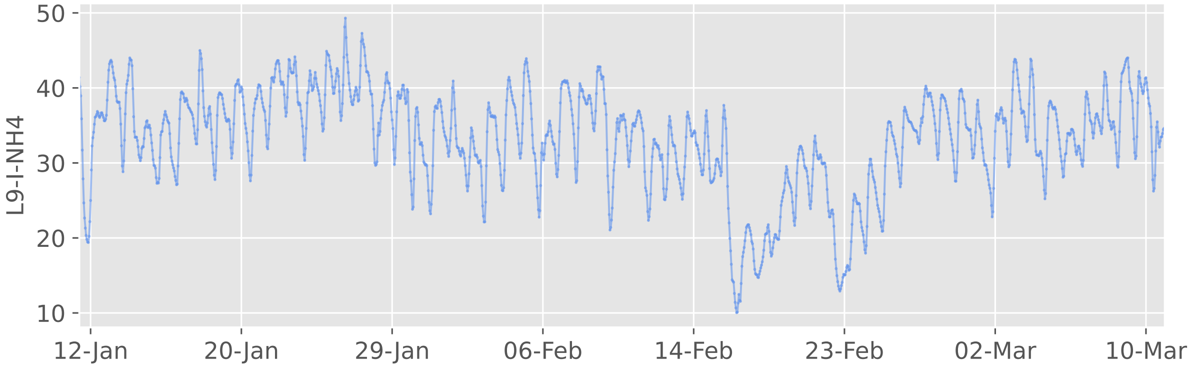
Box plot of the influent total suspended solids in line 9

Histogram of the influent total suspended solids in line 9



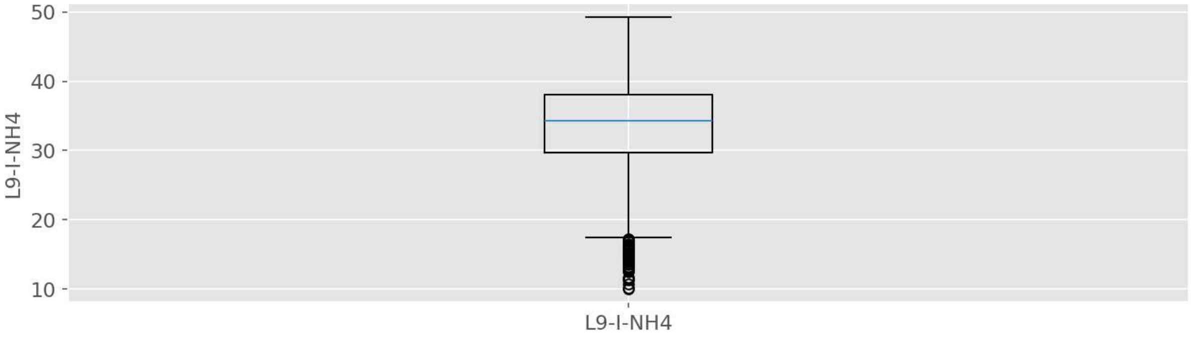
Source: The author.

Time-series of the Influent Ammonia in Line 9



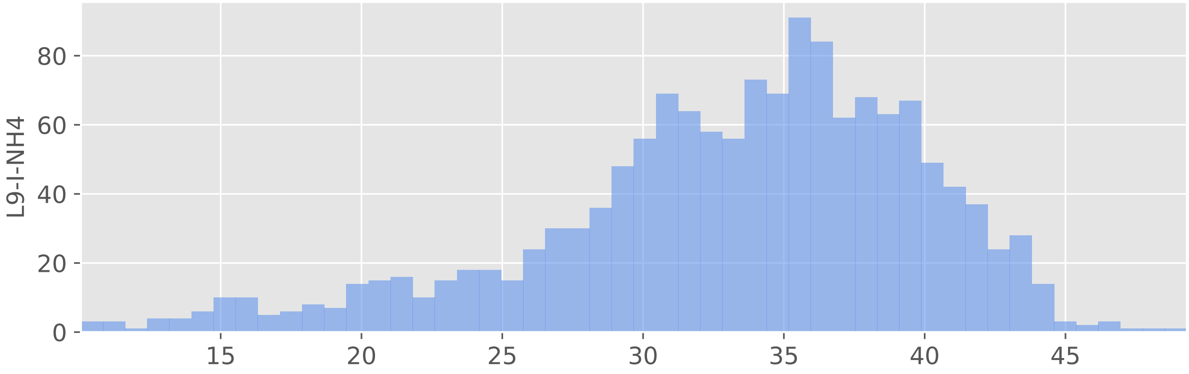
Source: The author.

Box plot of the Influent Ammonia in Line 9



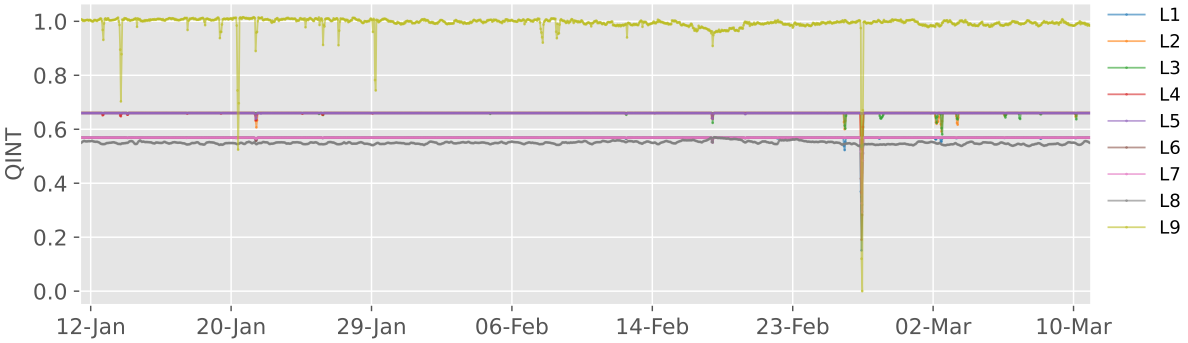
Source: The author.

Histogram of the influent ammonia in Line 9



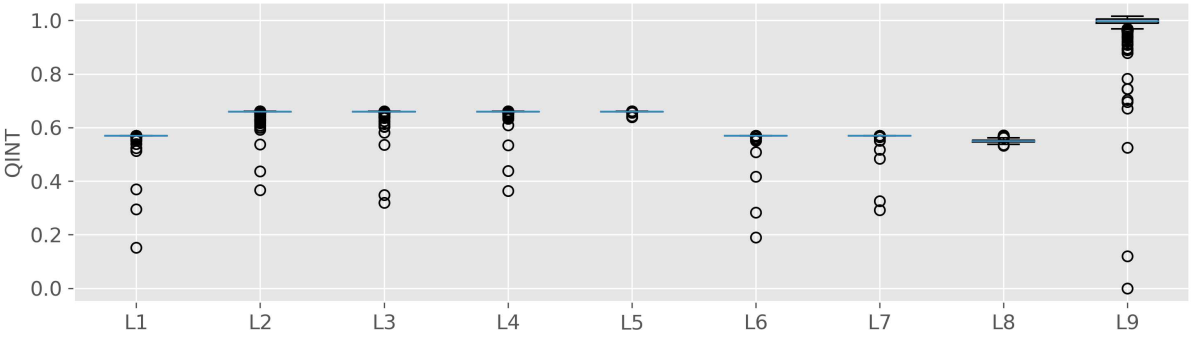
Source: The author.

Time-series of the internal recycle flow rate



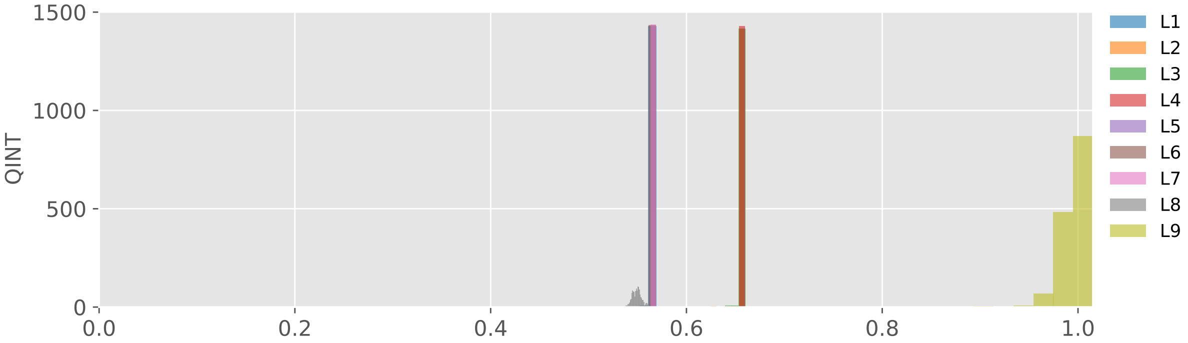
Source: The author.

Box plot of the internal recycle flow rate



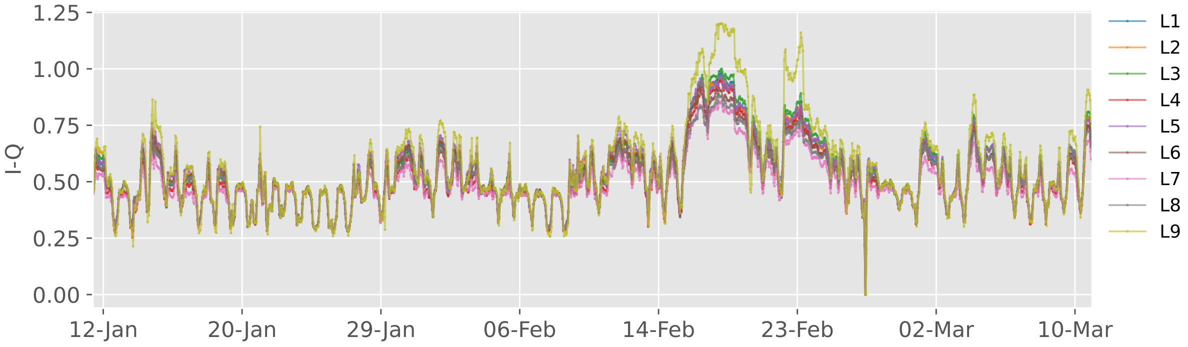
Source: The author.

Histogram of the internal recycle flow rate



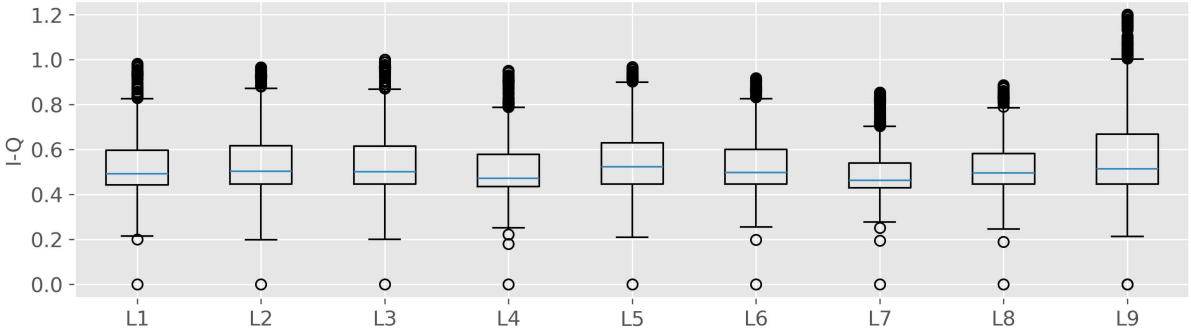
Source: The author.

Time-series of the influent flow rate



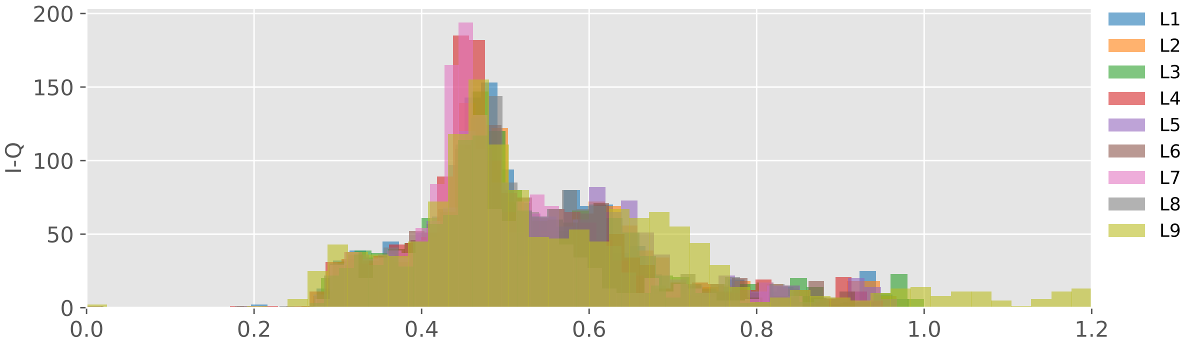
Source: The author.

Box plot of the influent flow rate



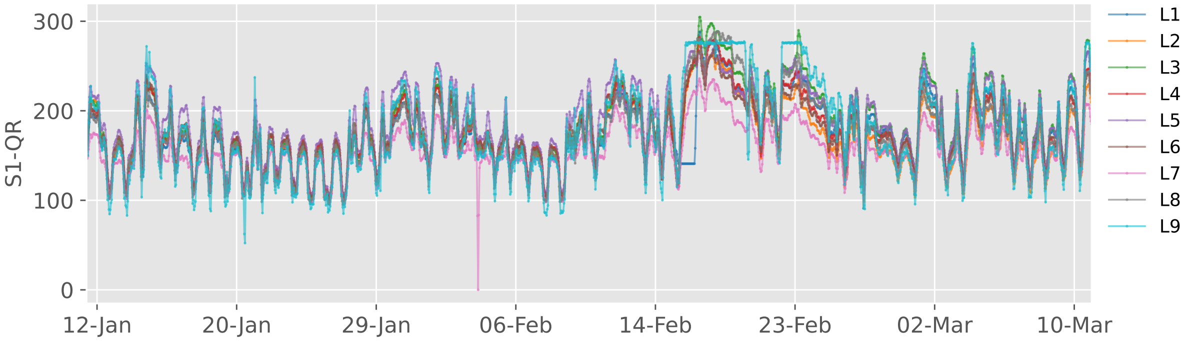
Source: The author.

Histogram of the influent flow rate



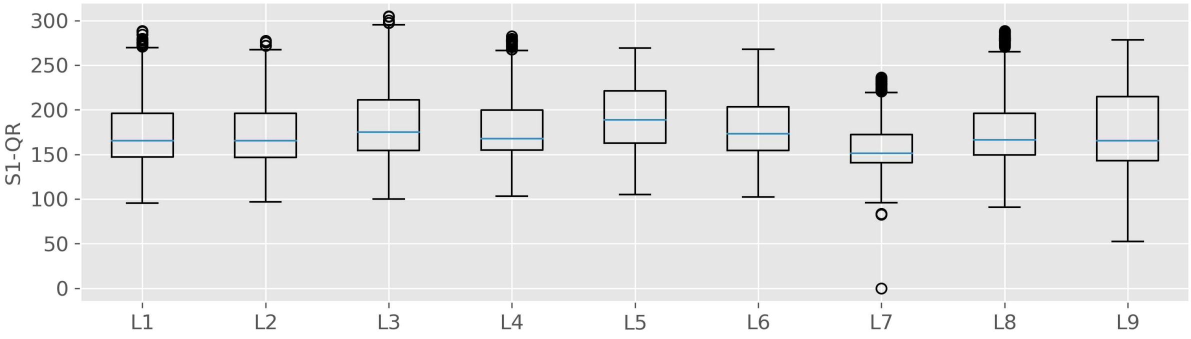
Source: The author.

Time-series of the return sludge flow rate from S1



Source: The author.

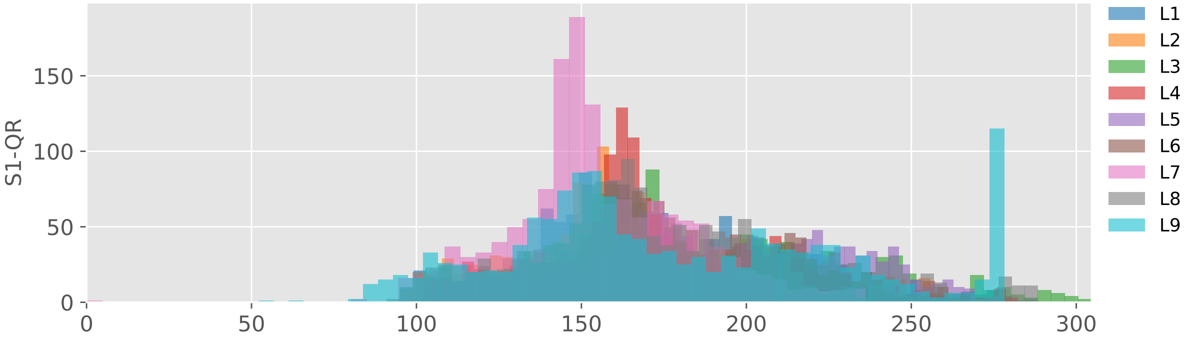
Box plot of the return sludge flow rate from S1



Source: The author.

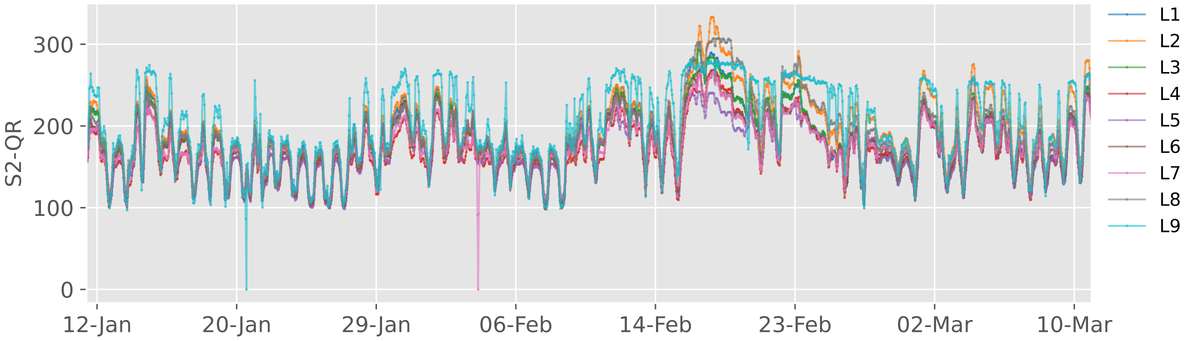
Box plot of the return sludge flow rate from S1

Histogram of the return sludge flow rate from S1



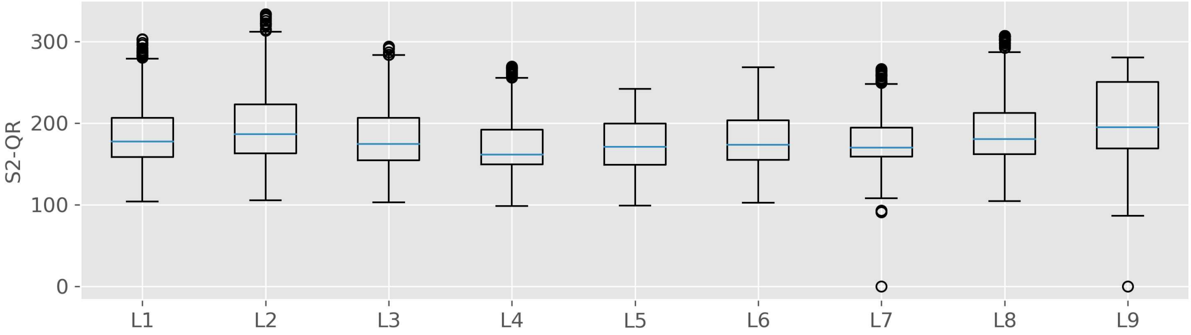
Source: The author.

Time-series of the return sludge flow rate from S2



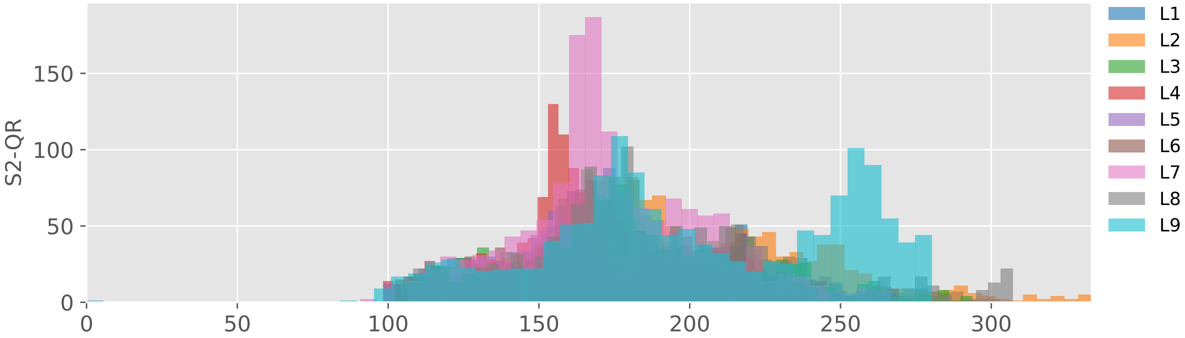
Source: The author.

Box plot of the return sludge flow rate from S2



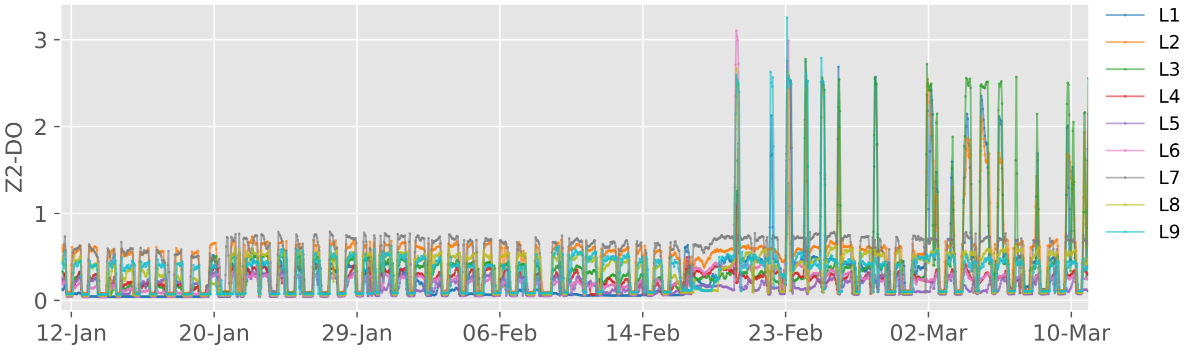
Source: The author.

Histogram of the return sludge flow rate from S2



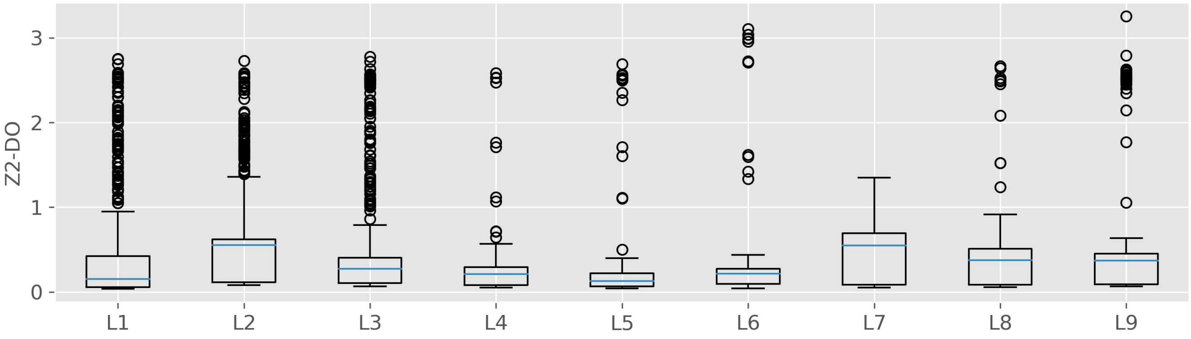
Source: The author.

Time-series of the dissolved oxygen in zone 2



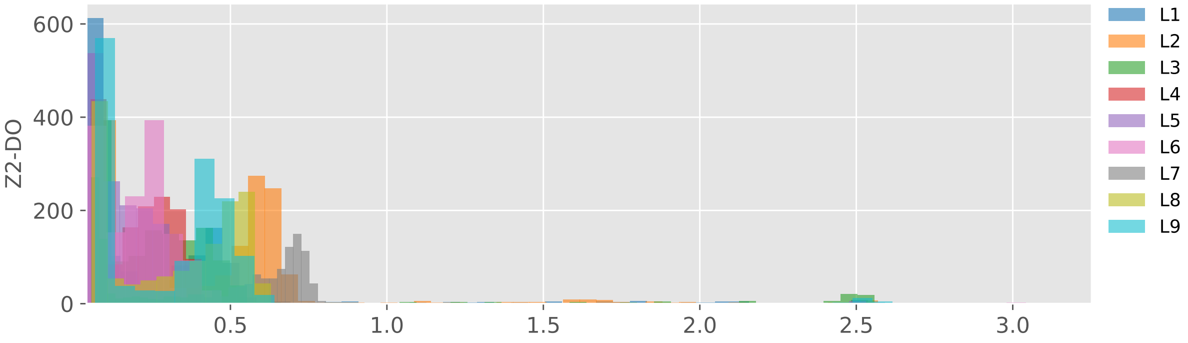
Source: The author.

Box plot of the dissolved oxygen in zone 2



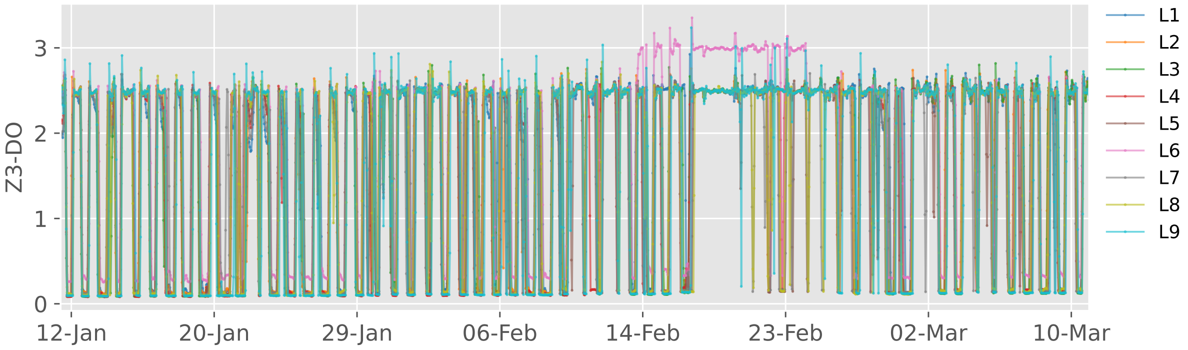
Source: The author.

Histogram of the dissolved oxygen in zone 2



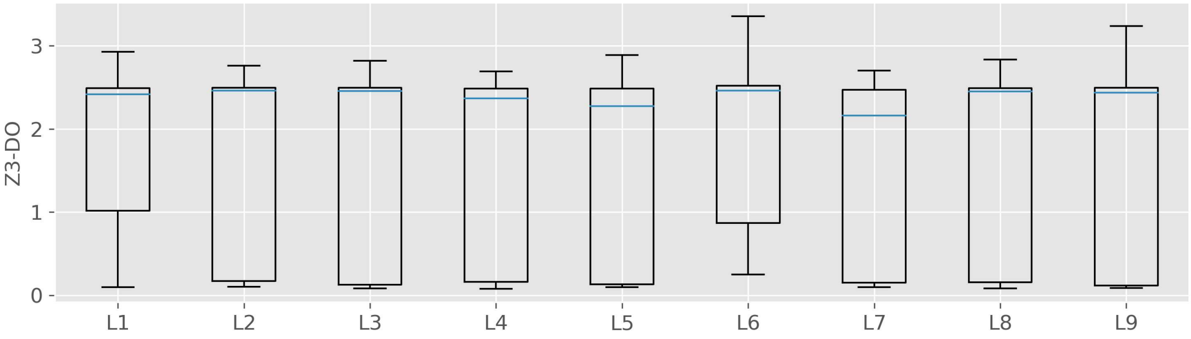
Source: The author.

Time-series of the dissolved oxygen in zone 3



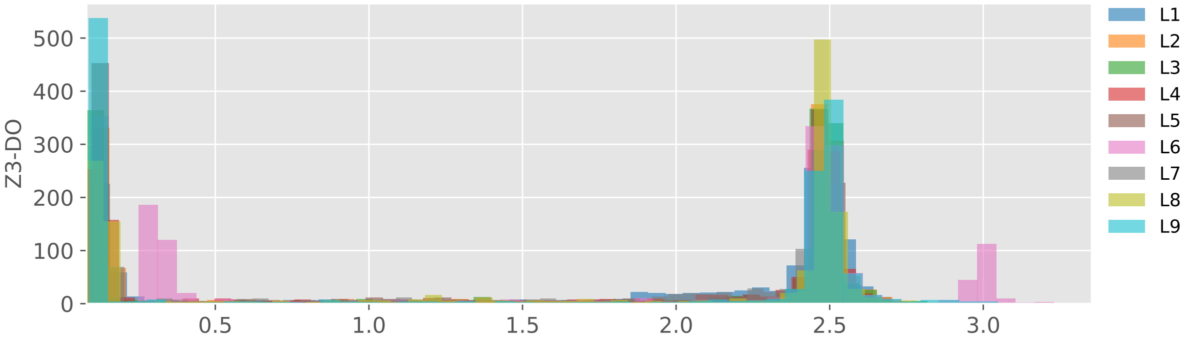
Source: The author.

Box plot of the dissolved oxygen in zone 3



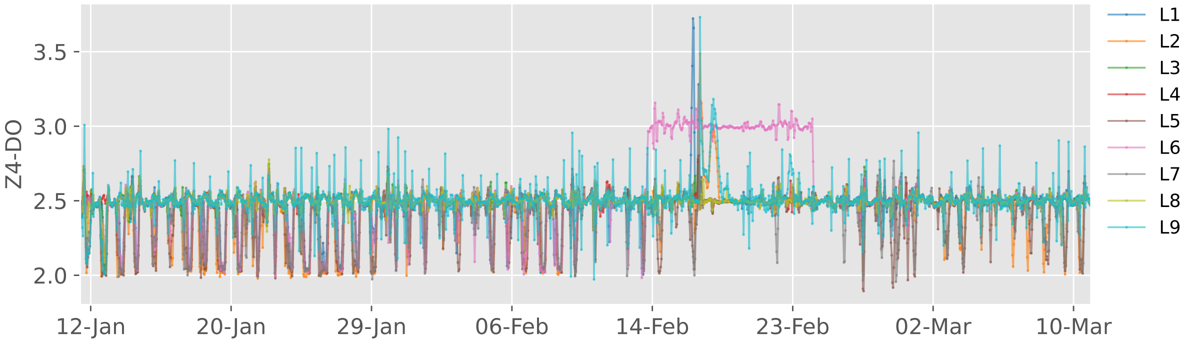
Source: The author.

Histogram of the dissolved oxygen in zone 3



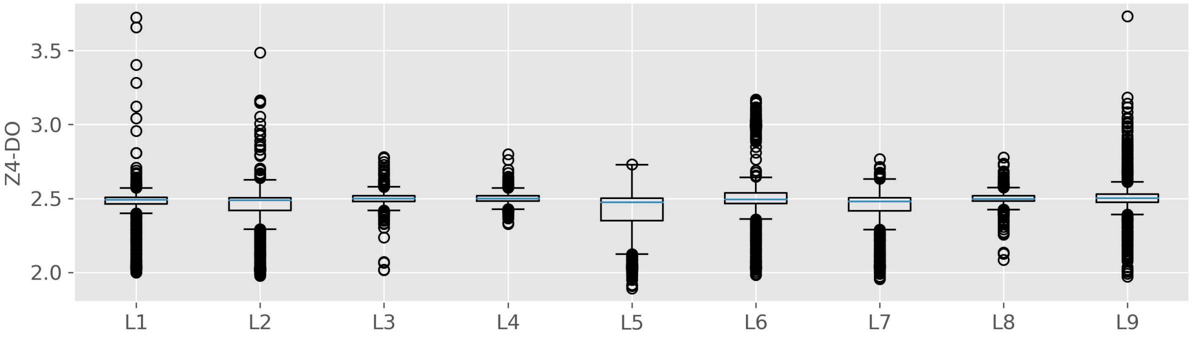
Source: The author.

Time-series of the dissolved oxygen in zone 4



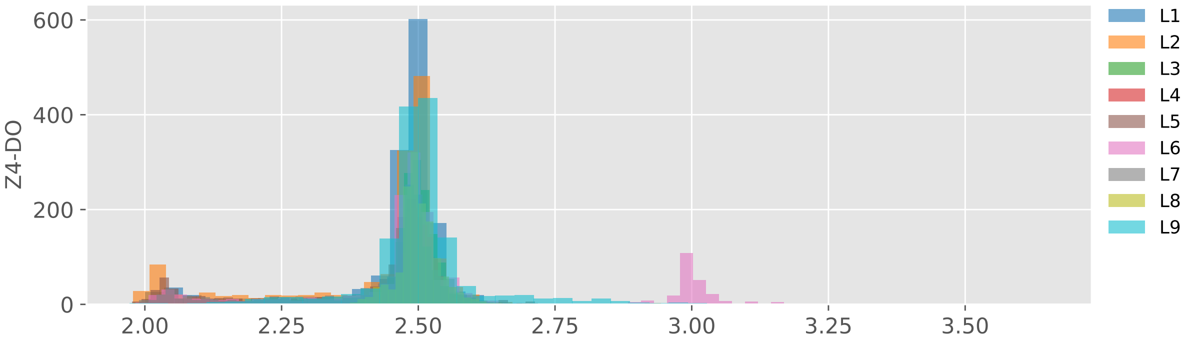
Source: The author.

Box plot of the dissolved oxygen in zone 4



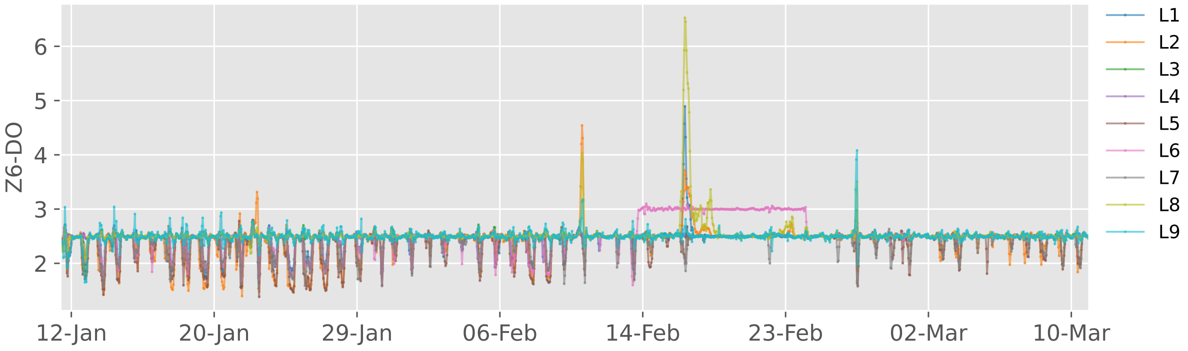
Source: The author.

Histogram of the dissolved oxygen in zone 4



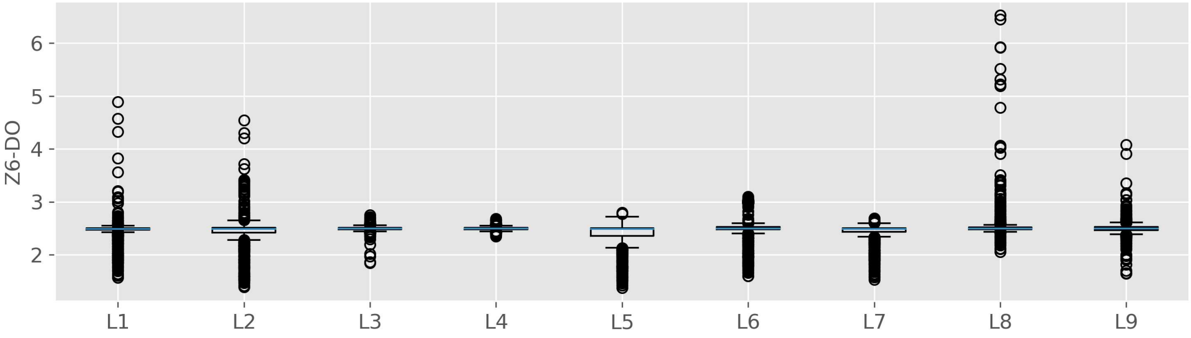
Source: The author.

Time-series of the dissolved oxygen in zone 6



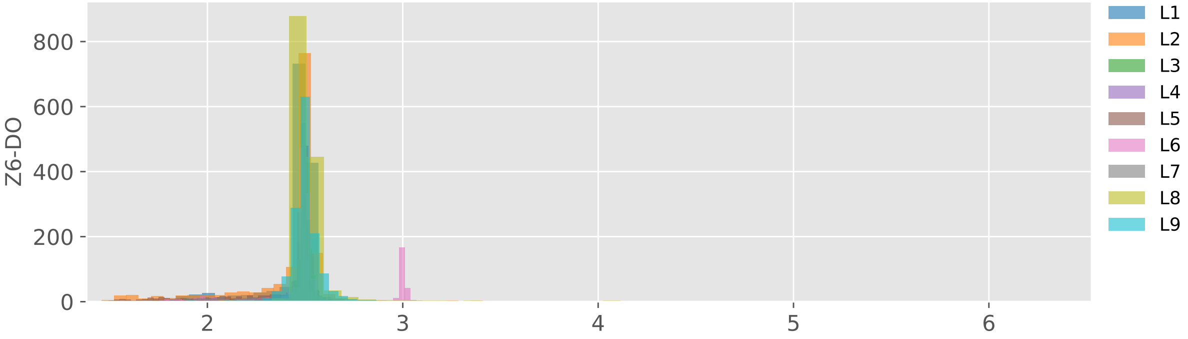
Source: The author.

Box plot of the dissolved oxygen in zone 6



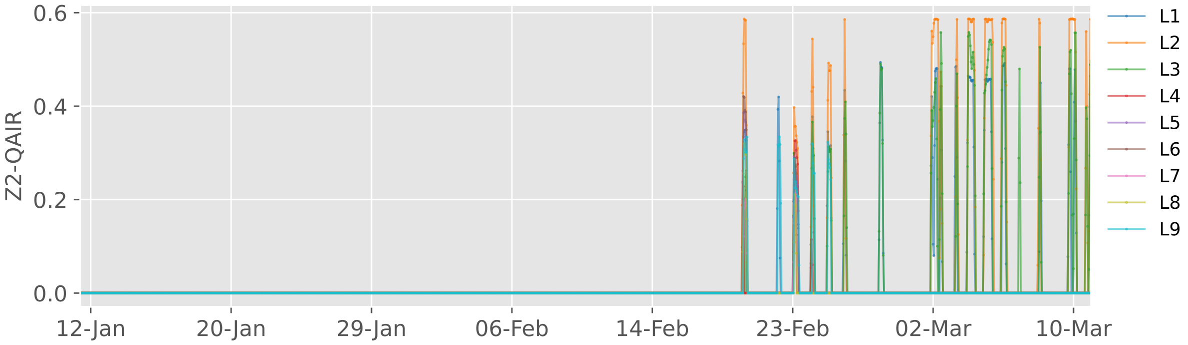
Source: The author.

Histogram of the dissolved oxygen in zone 6



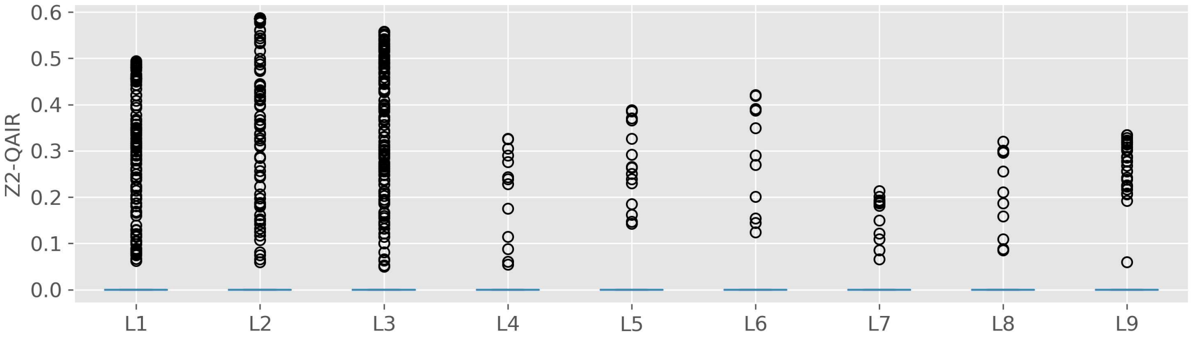
Source: The author.

Time-series of the air flow rate in zone 2



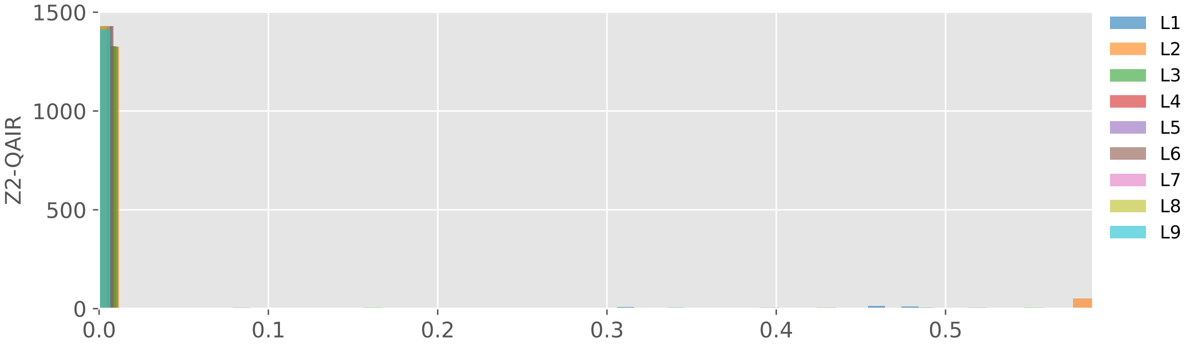
Source: The author.

Box plot of the air flow rate in zone 2



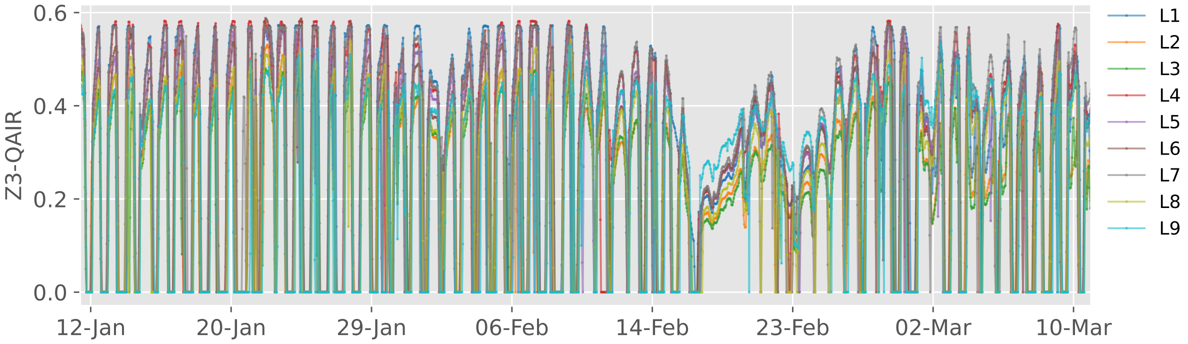
Source: The author.

Histogram of the air flow rate in zone 2



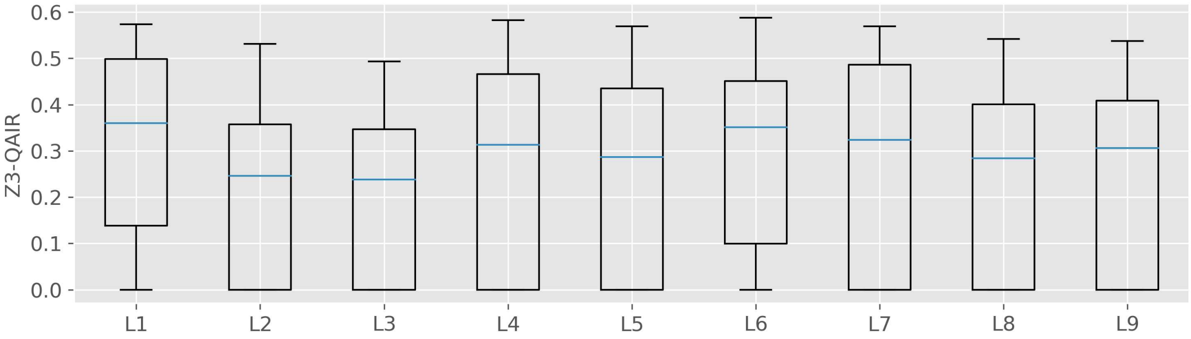
Source: The author.

Time-series of the air flow rate in zone 3



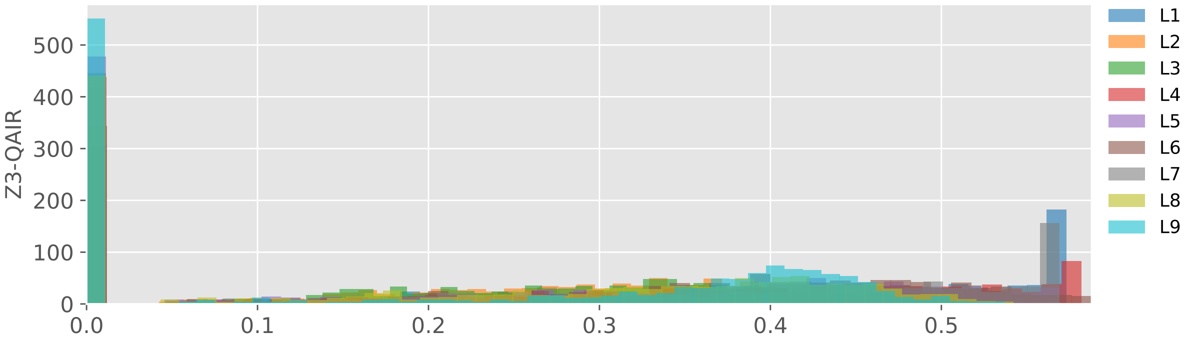
Source: The author.

Box plot of the air flow rate in zone 3



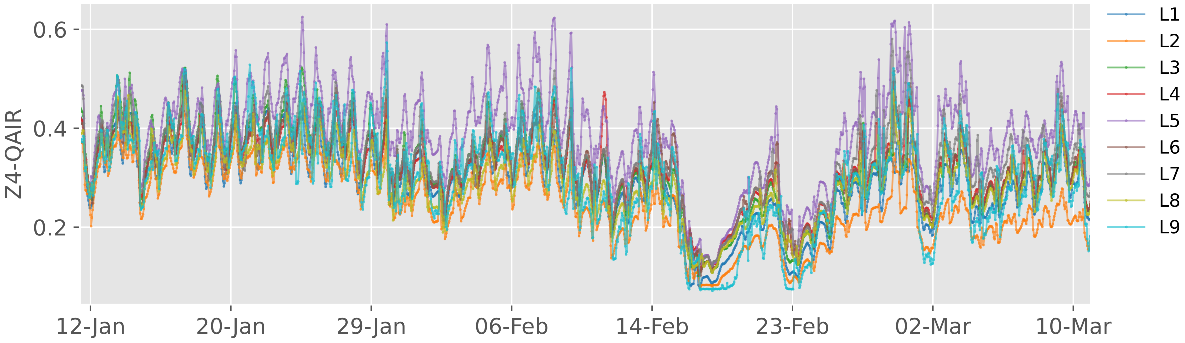
Source: The author.

Histogram of the air flow rate in zone 3



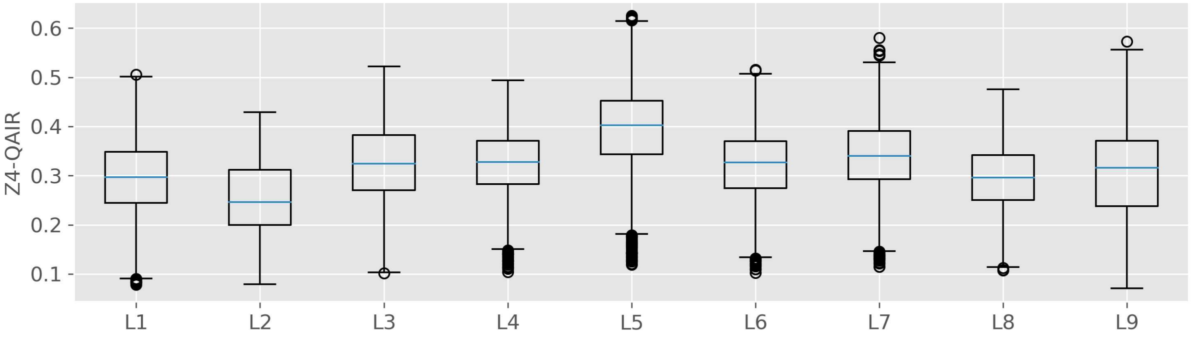
Source: The author.

Time-series of the air flow rate in zone 4



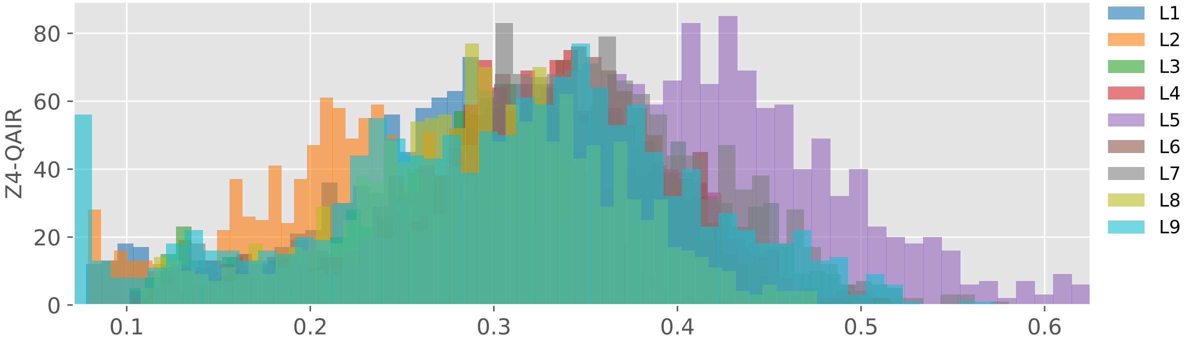
Source: The author.

Box plot of the air flow rate in zone 4



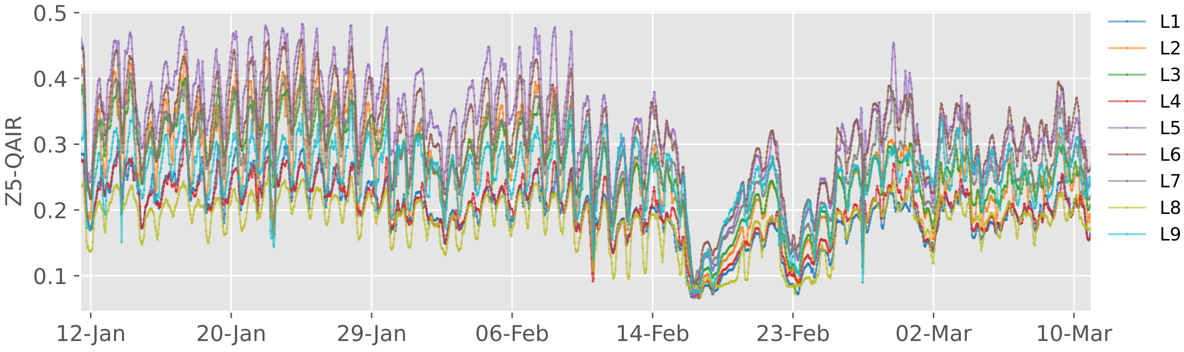
Source: The author.

Histogram of the air flow rate in zone 4



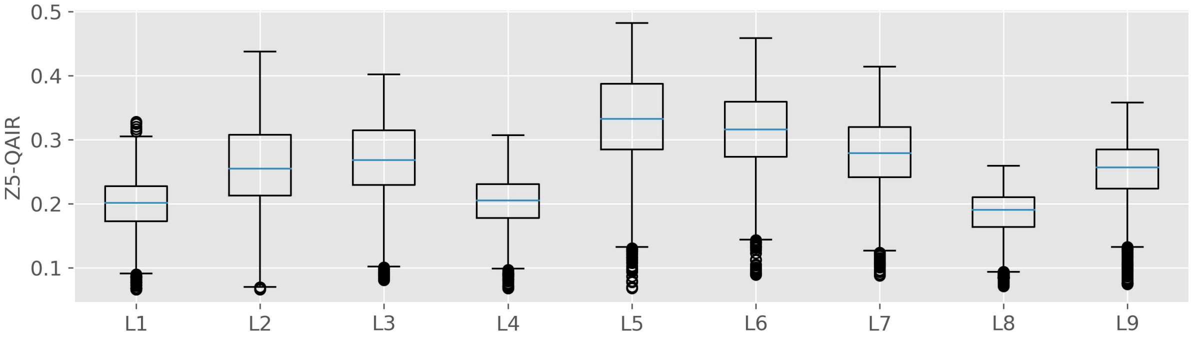
Source: The author.

Time-series of the air flow rate in zone 5



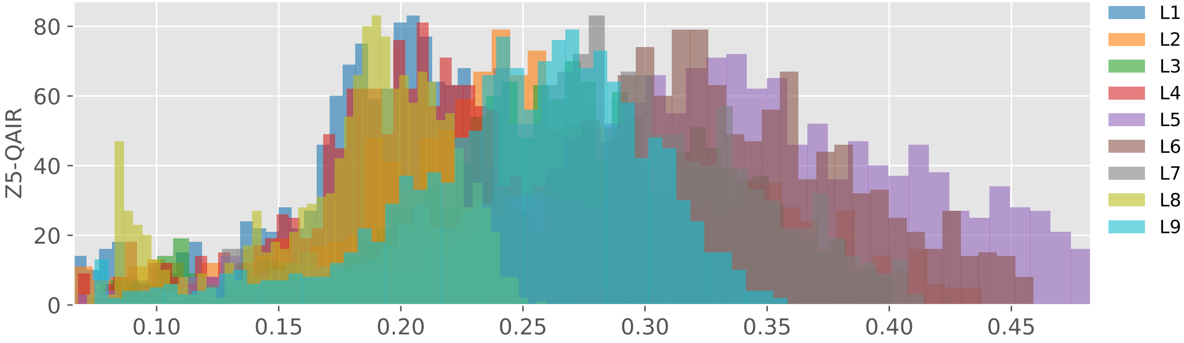
Source: The author.

Box plot of the air flow rate in zone 5



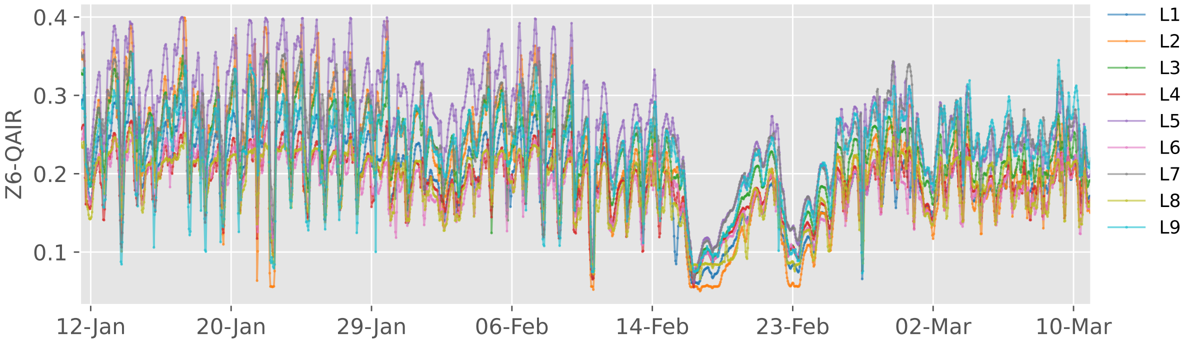
Source: The author.

Histogram of the air flow rate in zone 5



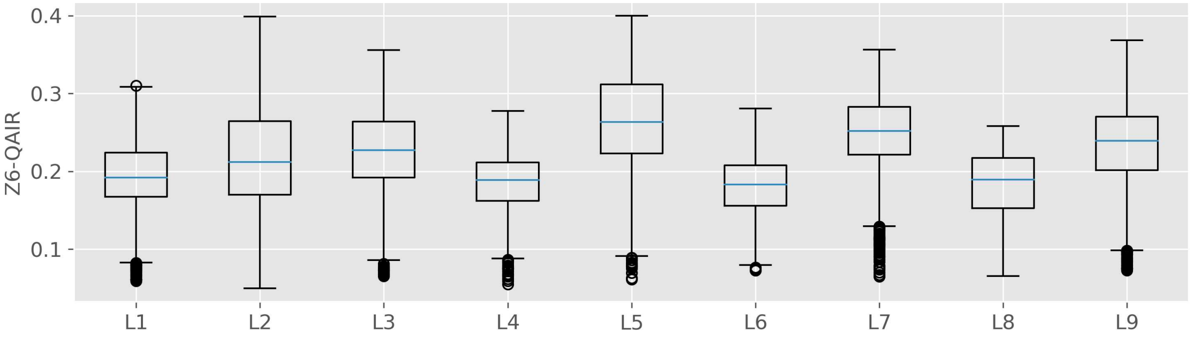
Source: The author.

Time-series of the air flow rate in zone 6



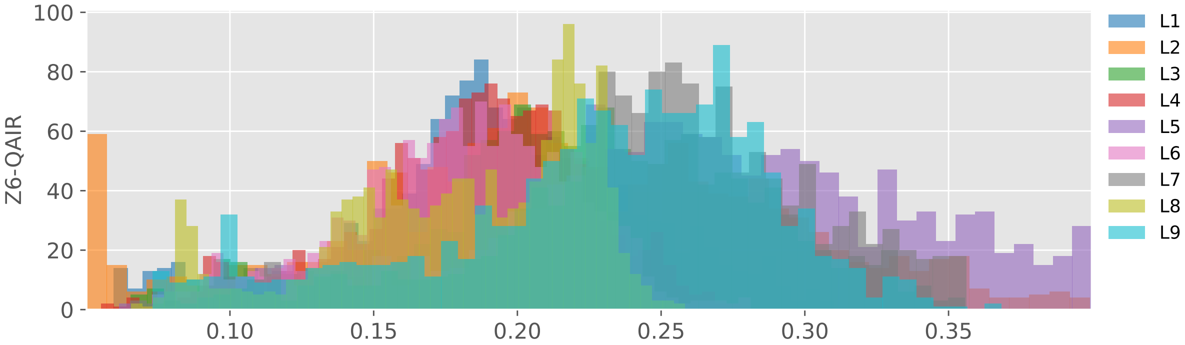
Source: The author.

Box plot of the air flow rate in zone 6



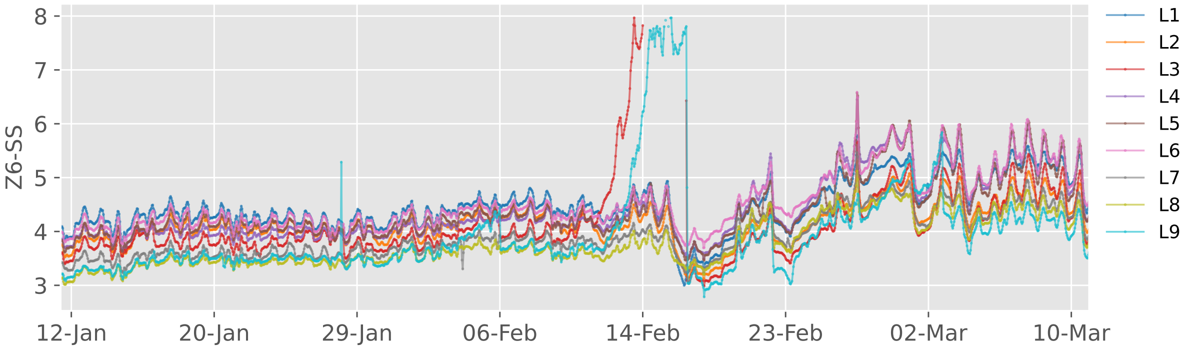
Source: The author.

Histogram of the air flow rate in zone 6



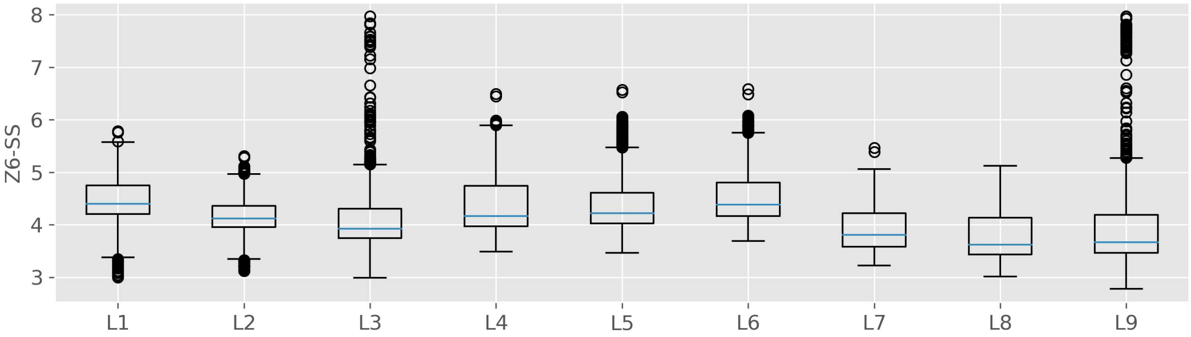
Source: The author.

Time-series of the total suspended solids in zone 6



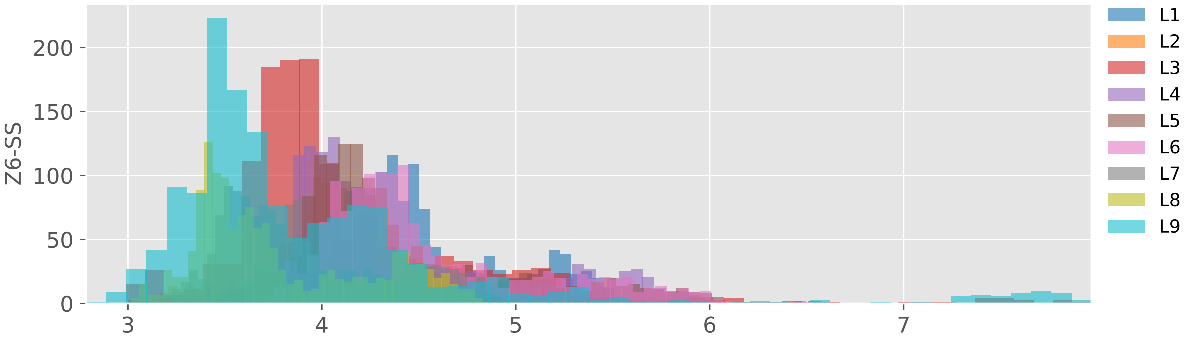
Source: The author.

Box plot of the total suspended solids in zone 6



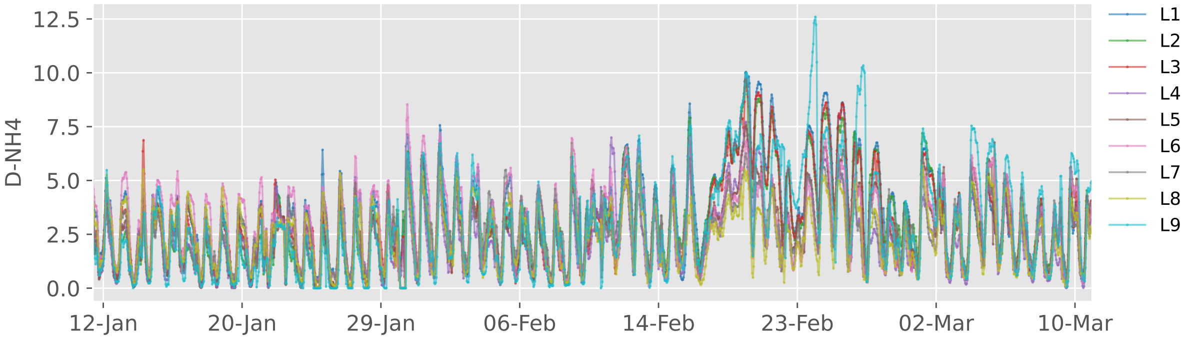
Source: The author.

Histogram of the total suspended solids in zone 6



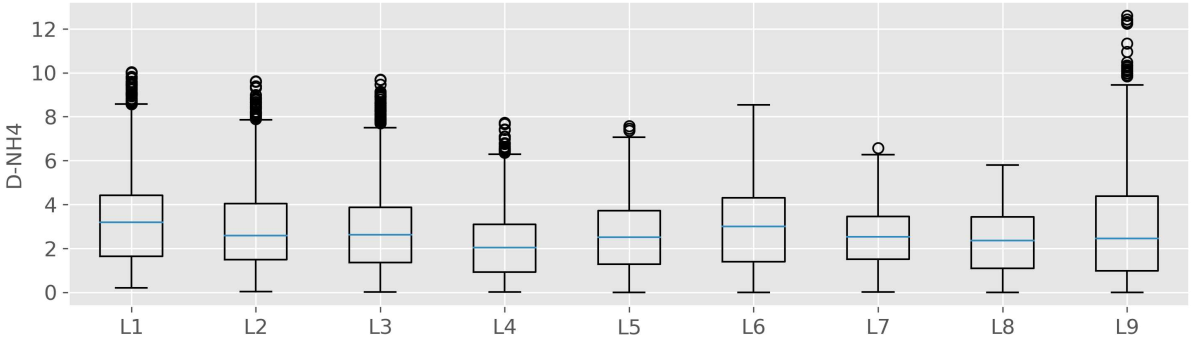
Source: The author.

Time-series of the ammonia at the end of degas



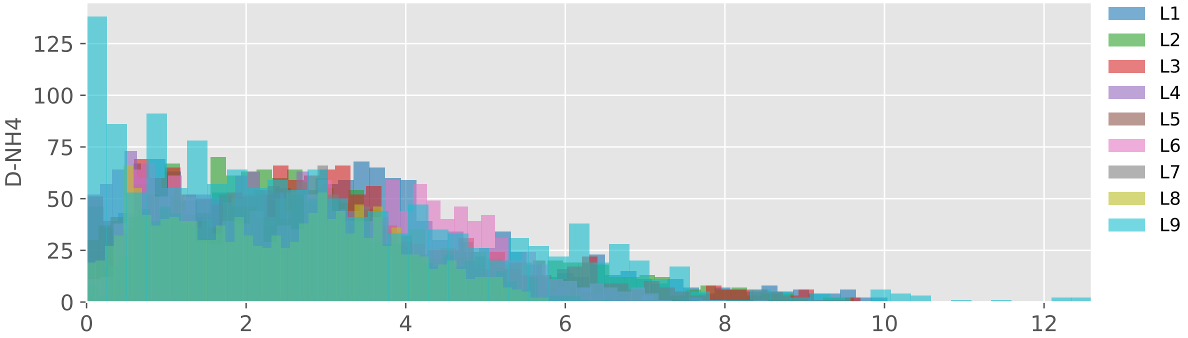
Source: The author.

Box plot of the ammonia at the end of degas



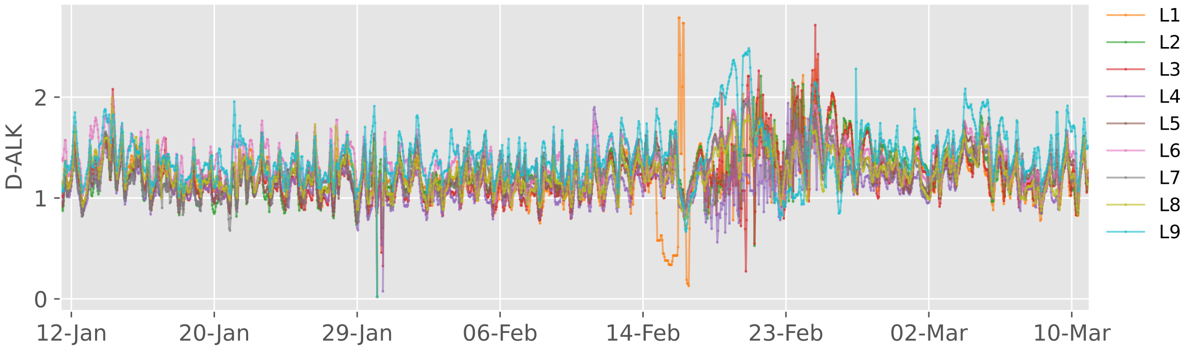
Source: The author.

Histogram of the ammonia at the end of degas



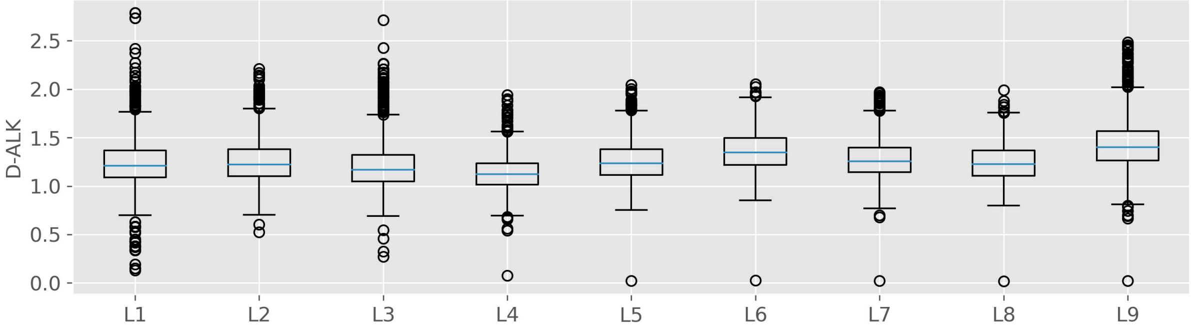
Source: The author.

Time-series of the alkalinity at the end of degas



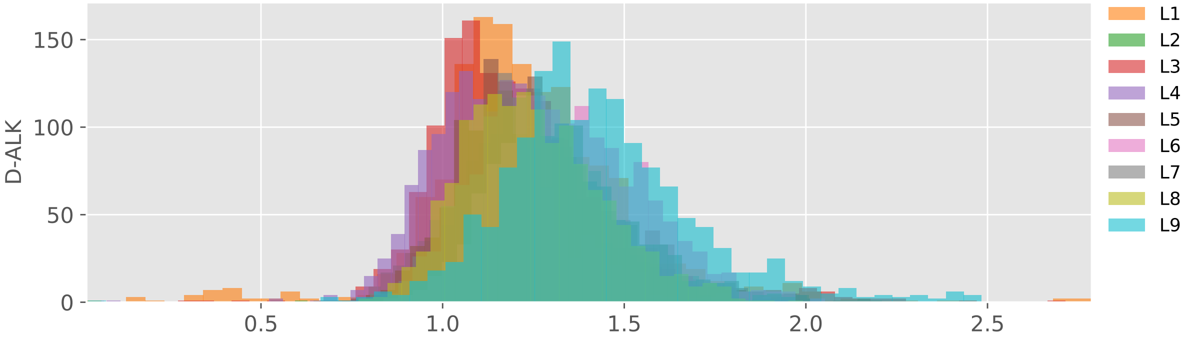
Source: The author.

Box plot of the alkalinity at the end of degas



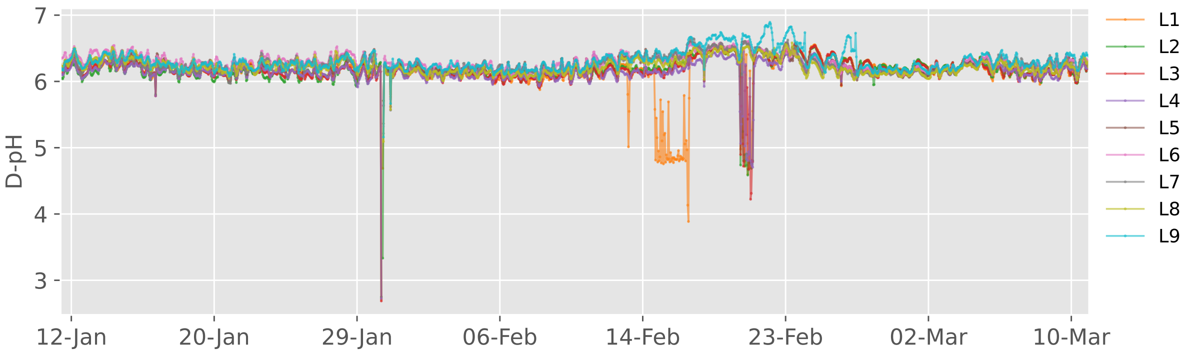
Source: The author.

Histogram of the alkalinity at the end of degas



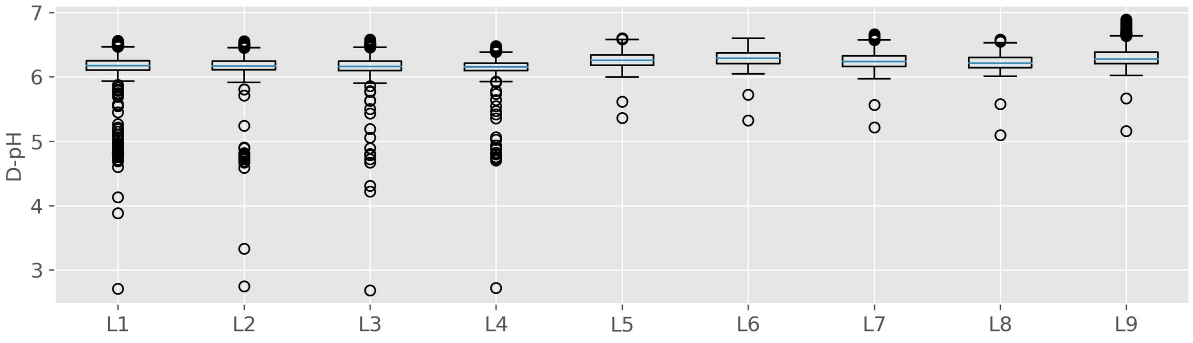
Source: The author.

Time-series of the pH at the end of degas



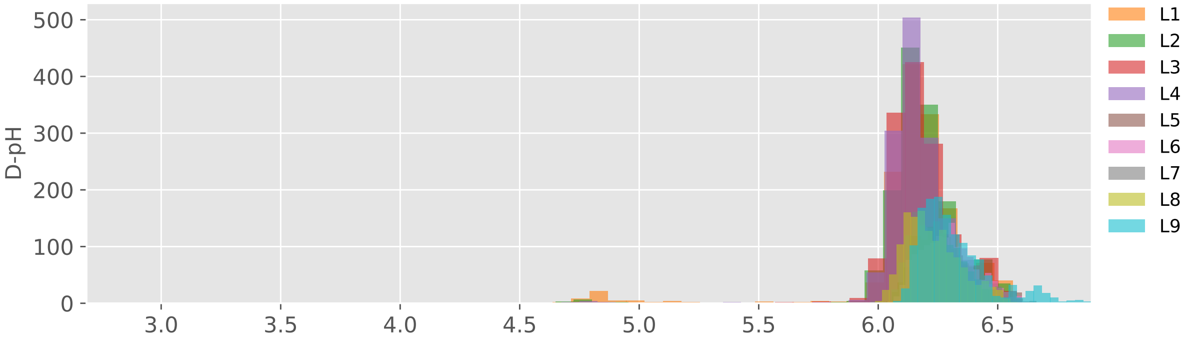
Source: The author.

Box plot of the pH at the end of degas



Source: The author.

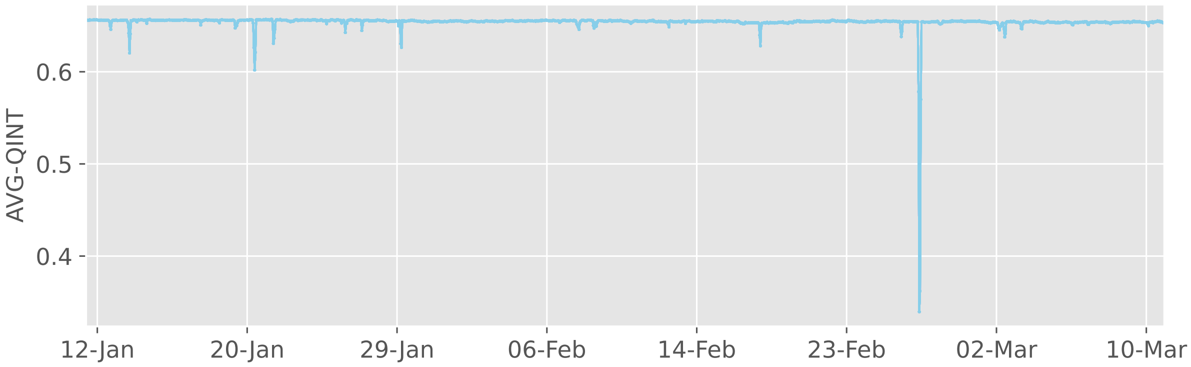
Histogram of the pH at the end of degas



Source: The author.

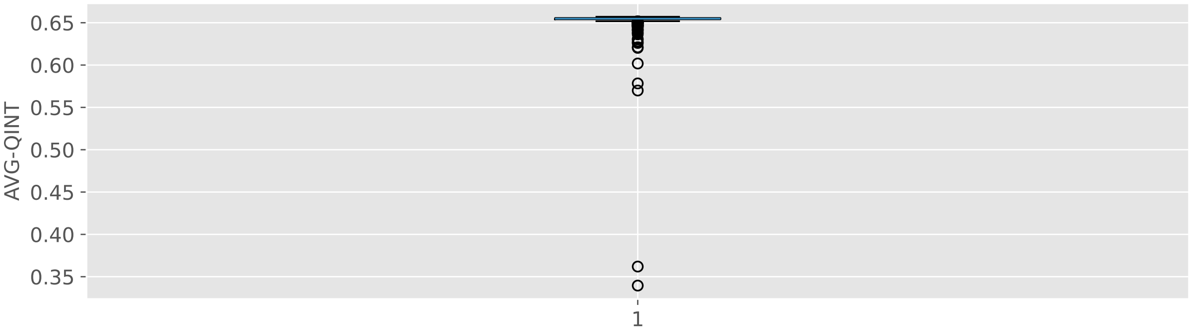
APPENDIX B – REDUCED FEATURE SET’S FEATURES EXPLORATORY ANALYSIS

Time-series of the average internal recycle flow rate



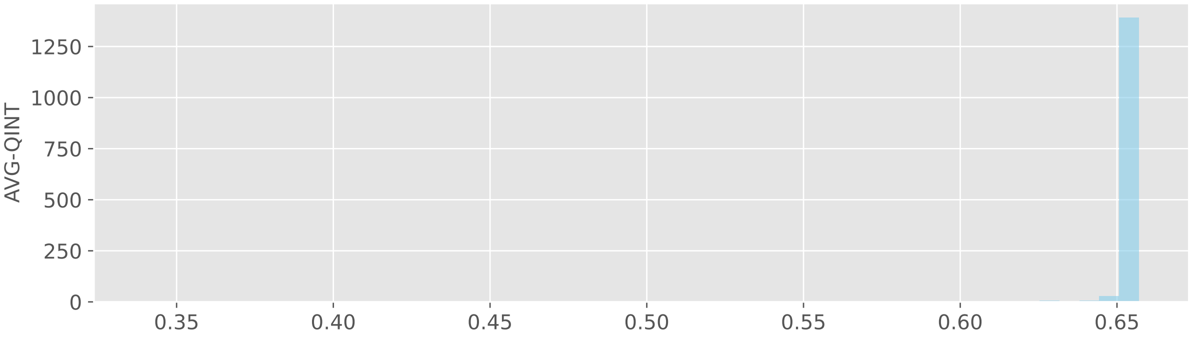
Source: The author.

Box plot of the average internal recycle flow rate



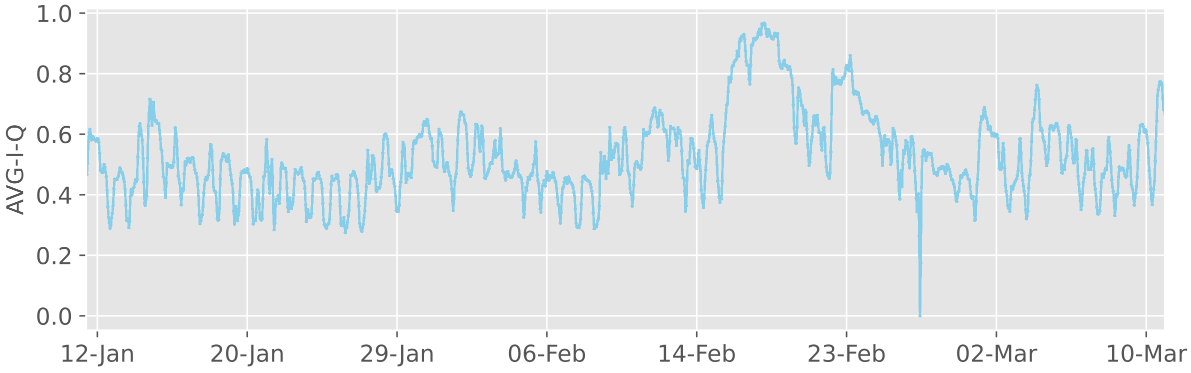
Source: The author.

Histogram of the average internal recycle flow rate



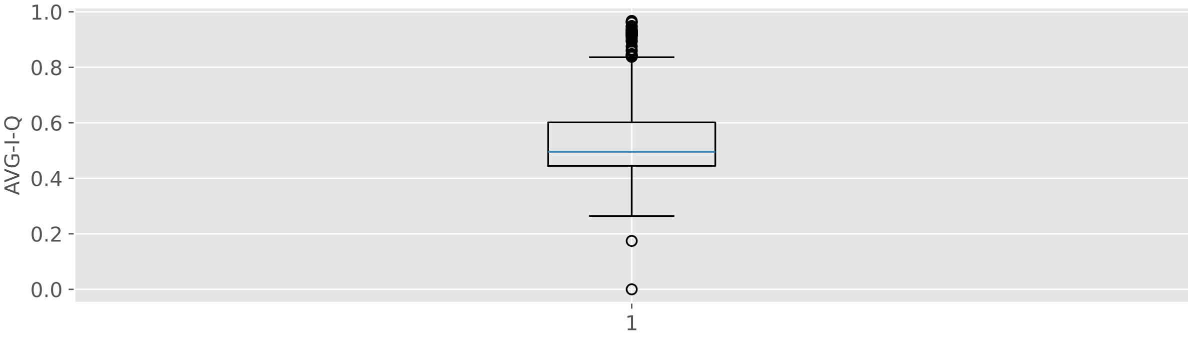
Source: The author.

Time-series of the average influent flow rate



Source: The author.

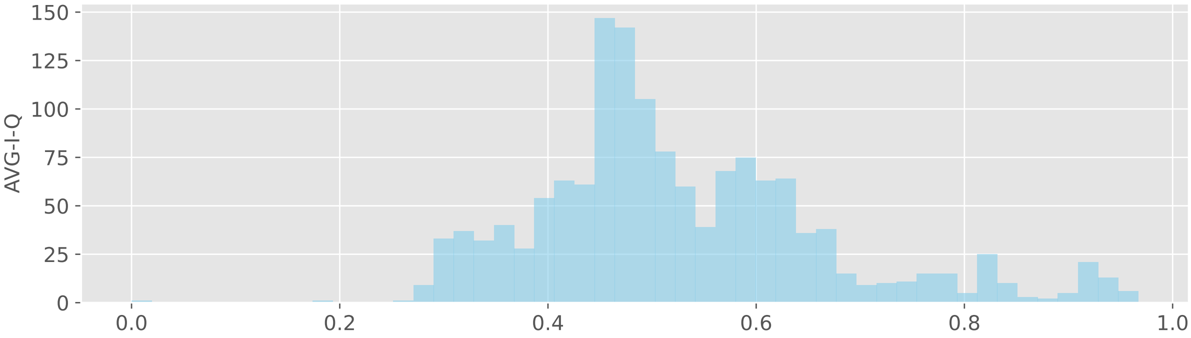
Box plot of the average influent flow rate



Source: The author.

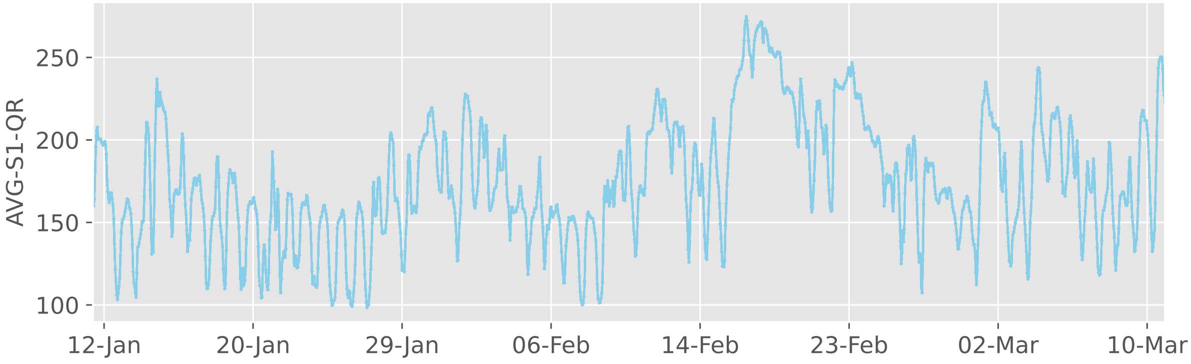
Box plot of the average influent flow rate

Histogram of the average influent flow rate



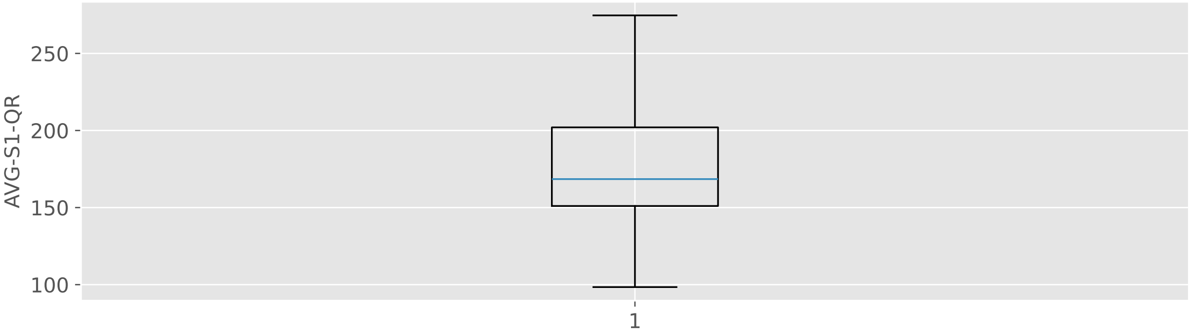
Source: The author.

Time-series of the average return sludge flow rate from S1



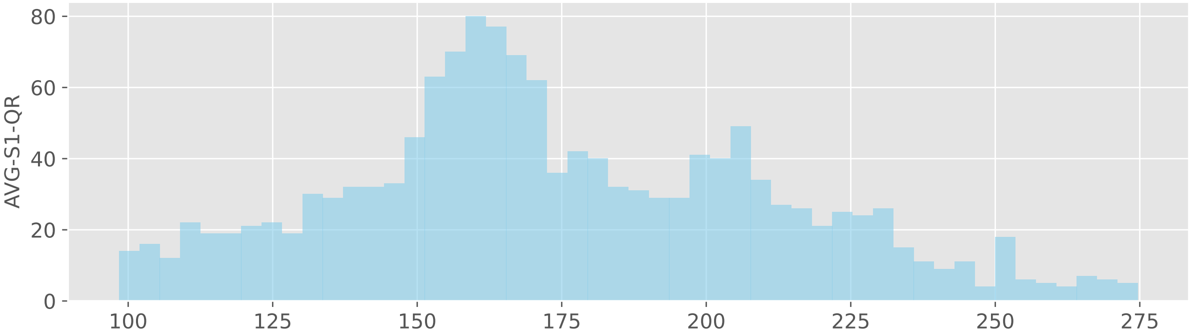
Source: The author.

Box plot of the average return sludge flow rate from S1



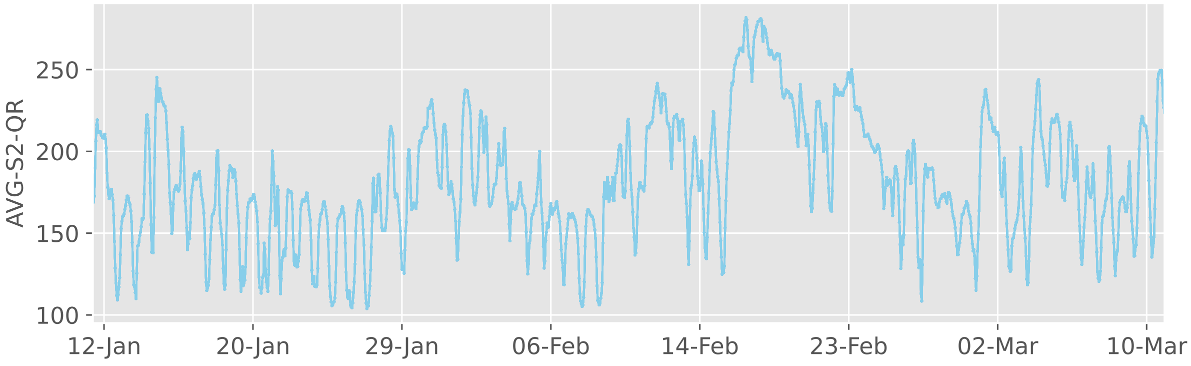
Source: The author.

Histogram of the average return sludge flow rate from S1



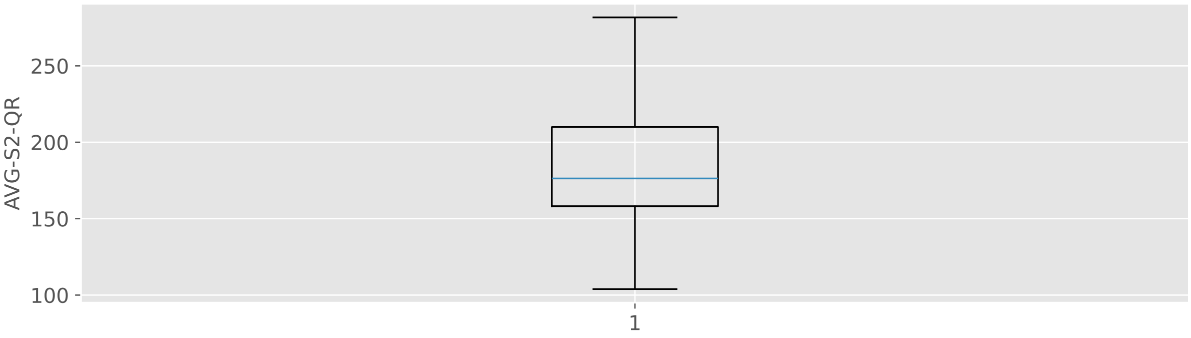
Source: The author.

Time-series of the average return sludge flow rate from S2



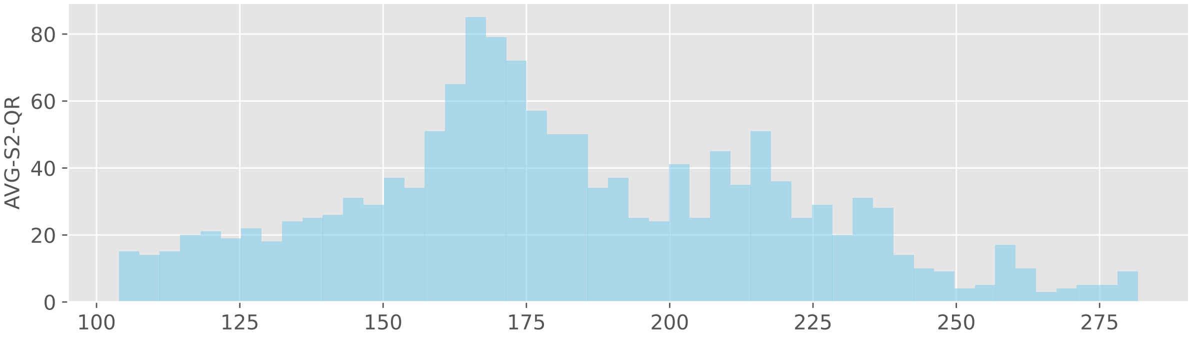
Source: The author.

Box plot of the average return sludge flow rate from S2



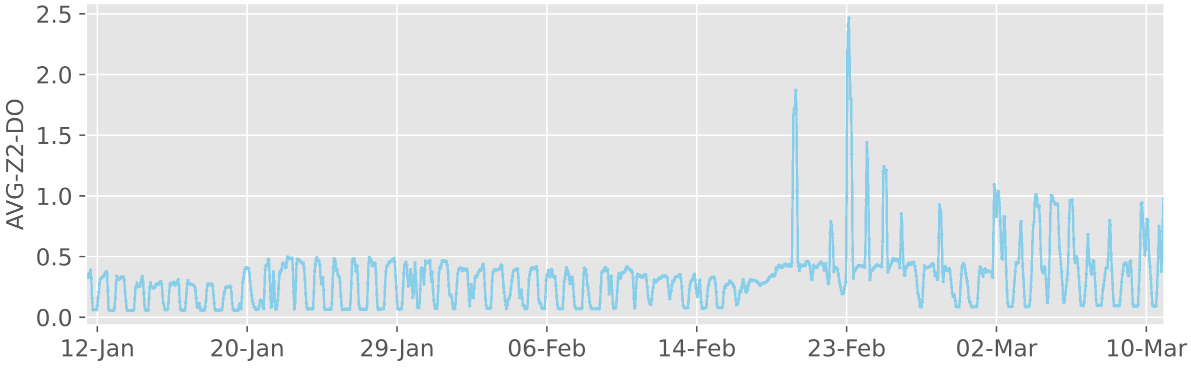
Source: The author.

Histogram of the average return sludge flow rate from S2



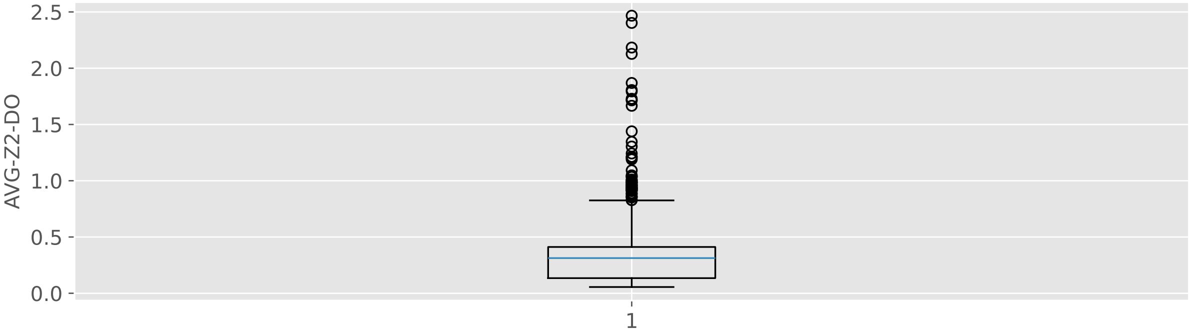
Source: The author.

Time-series of the average dissolved oxygen in Zone 2



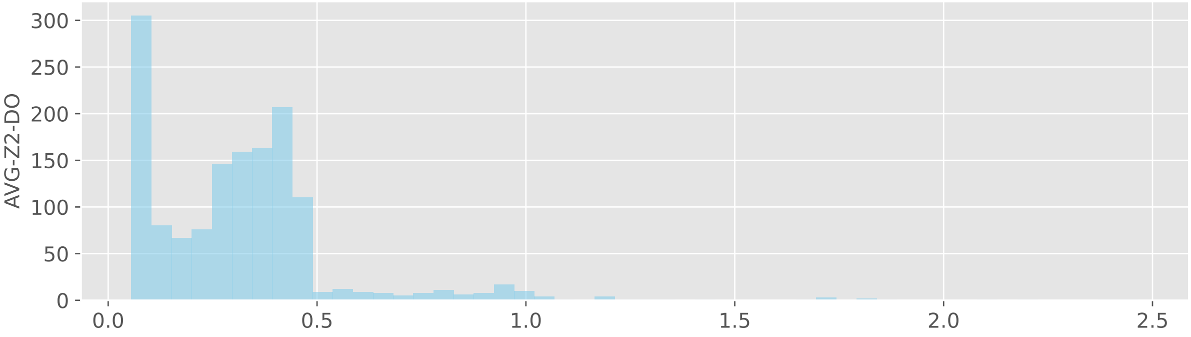
Source: The author.

Box plot of the average dissolved oxygen in Zone 2



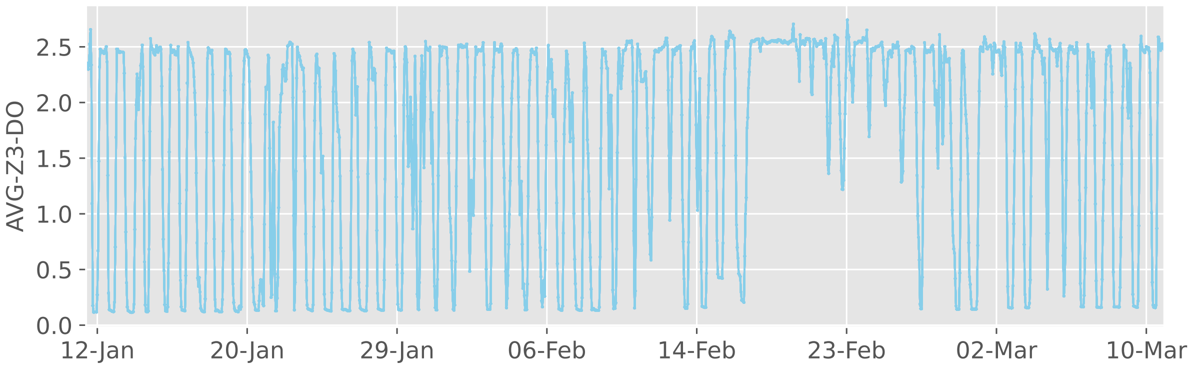
Source: The author.

Histogram of the average dissolved oxygen in Zone 2



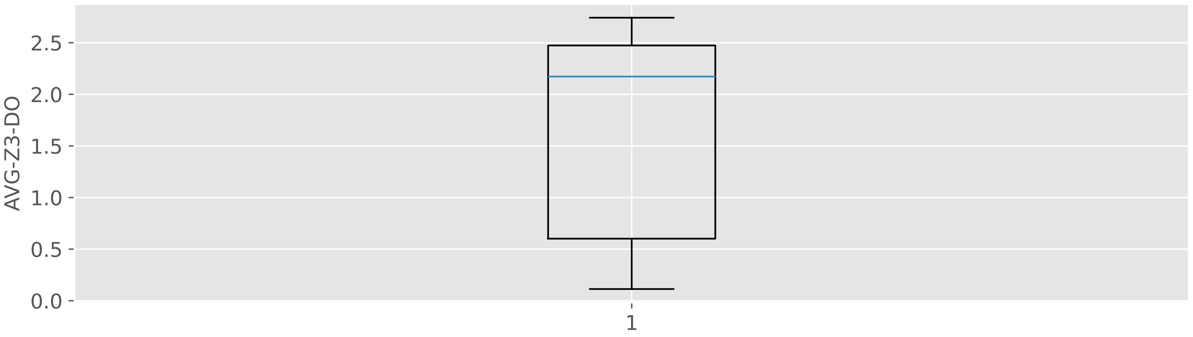
Source: The author.

Time-series of the average dissolved oxygen in Zone 3



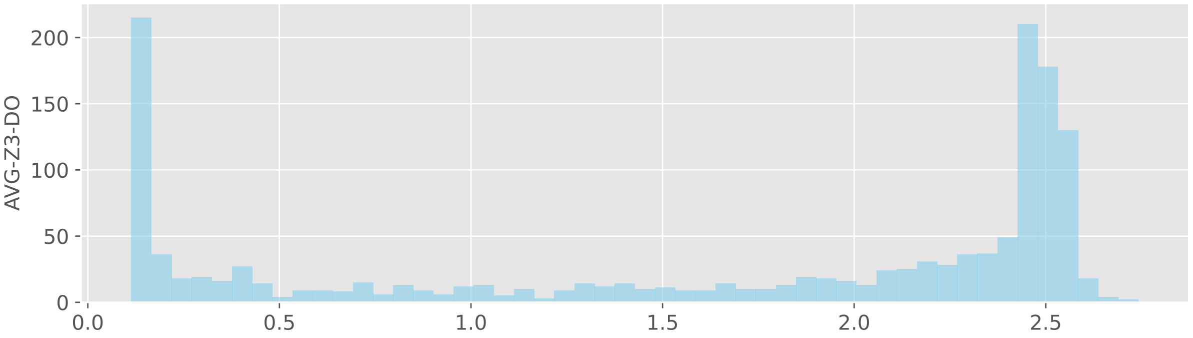
Source: The author.

Box plot of the average dissolved oxygen in Zone 3



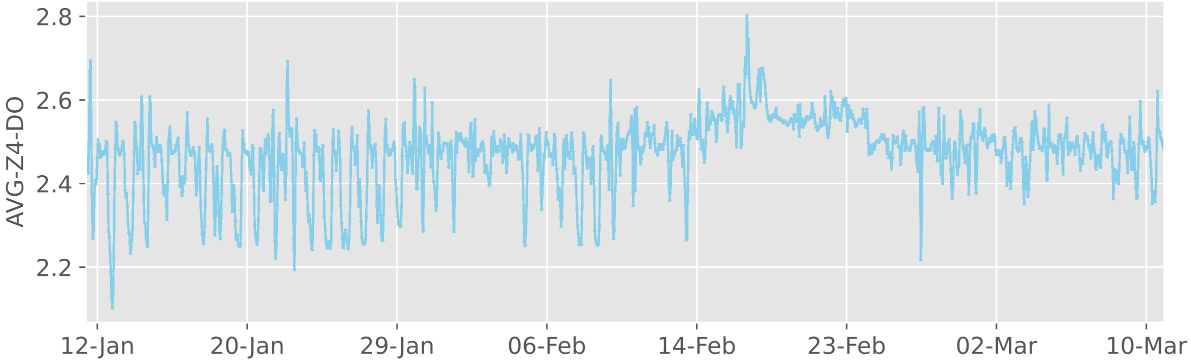
Source: The author.

Histogram of the average dissolved oxygen in Zone 3



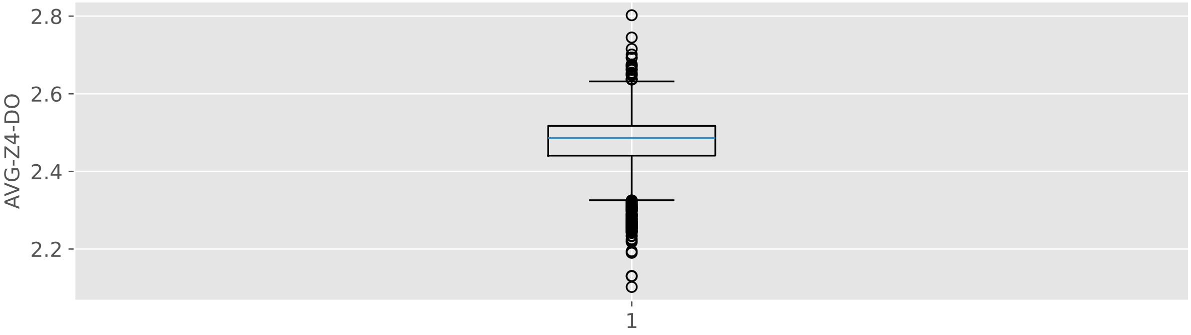
Source: The author.

Time-series of the average dissolved oxygen in Zone 4



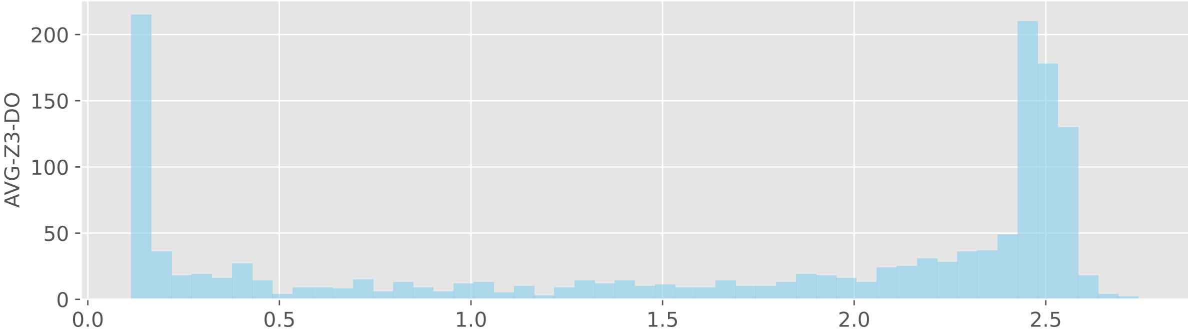
Source: The author.

Box plot of the average dissolved oxygen in Zone 4



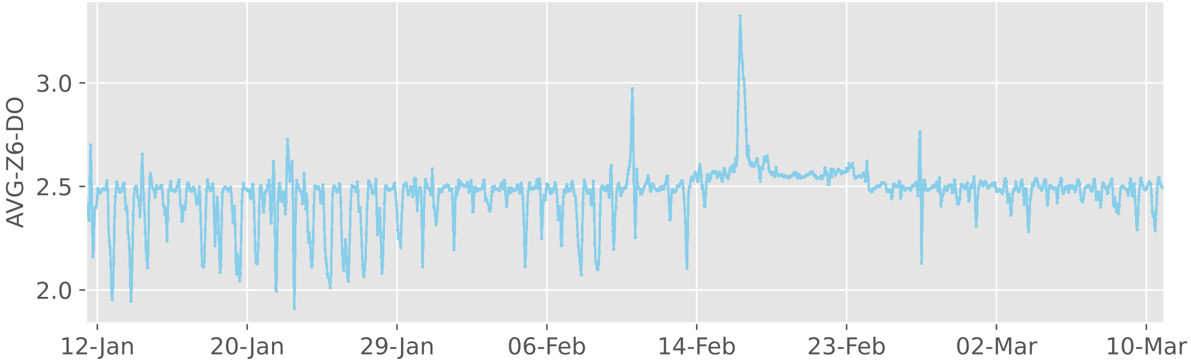
Source: The author.

Histogram of the average dissolved oxygen in Zone 4



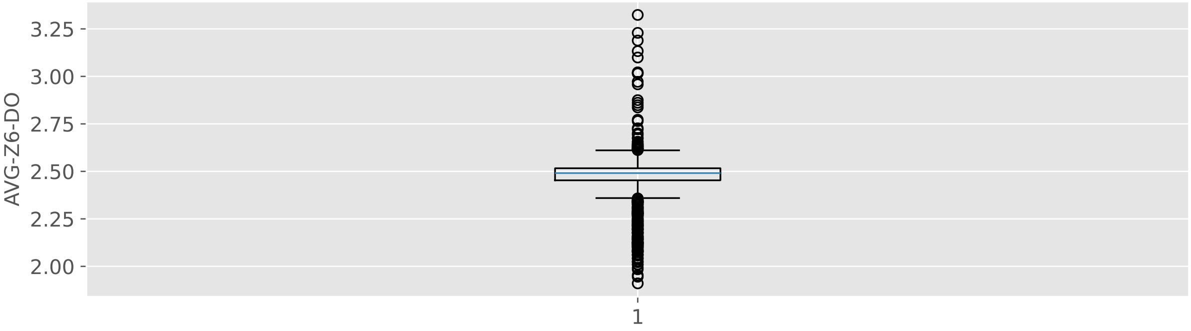
Source: The author.

Time-series of the average dissolved oxygen in Zone 6



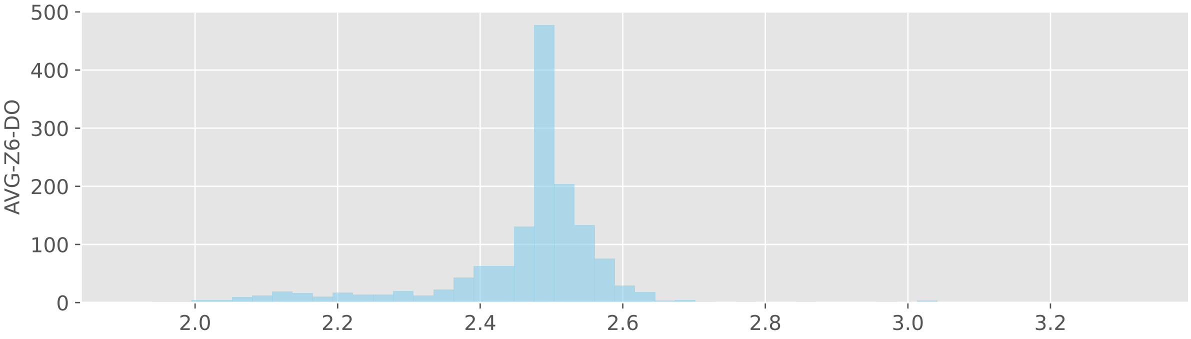
Source: The author.

Box plot of the average dissolved oxygen in Zone 6



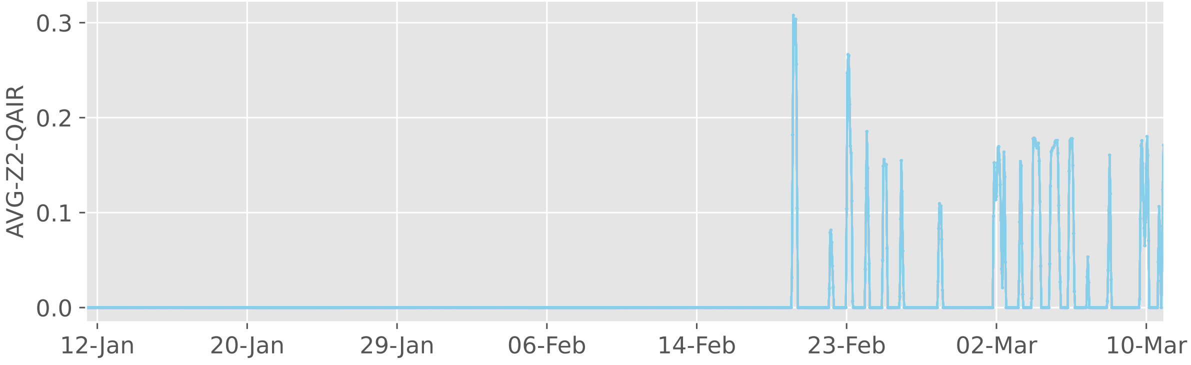
Source: The author.

Histogram of the average dissolved oxygen in Zone 6



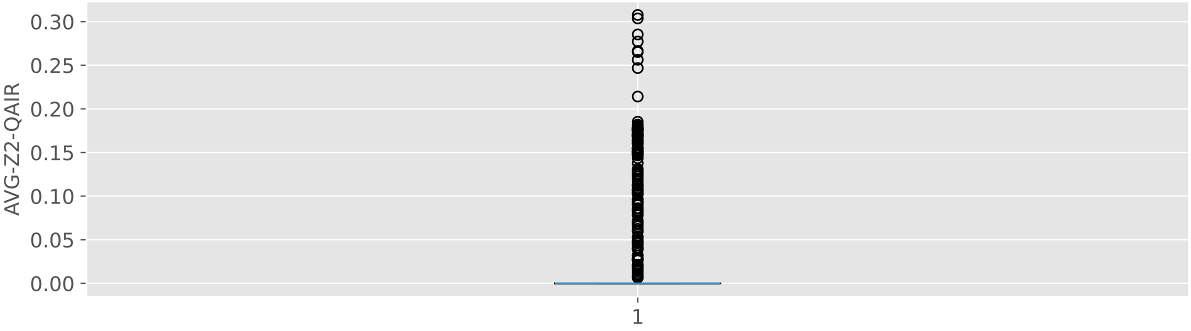
Source: The author.

Time-series of the average air flow rate in Zone 2



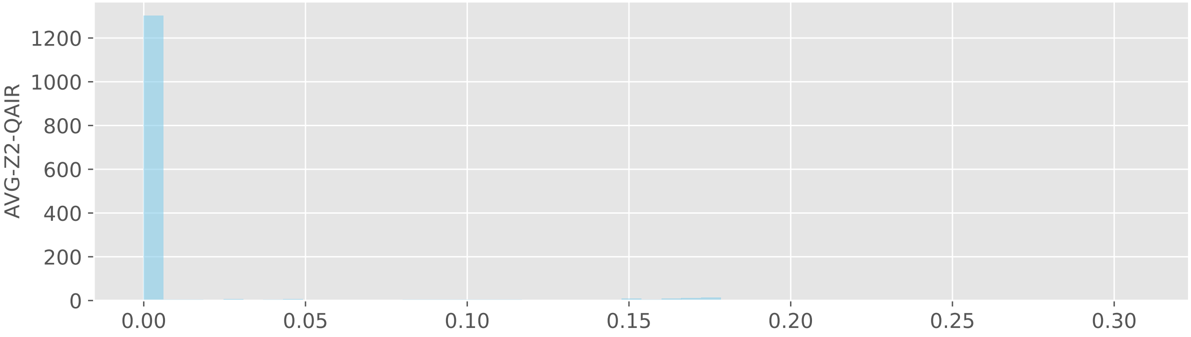
Source: The author.

Box plot of the average air flow rate in Zone 2



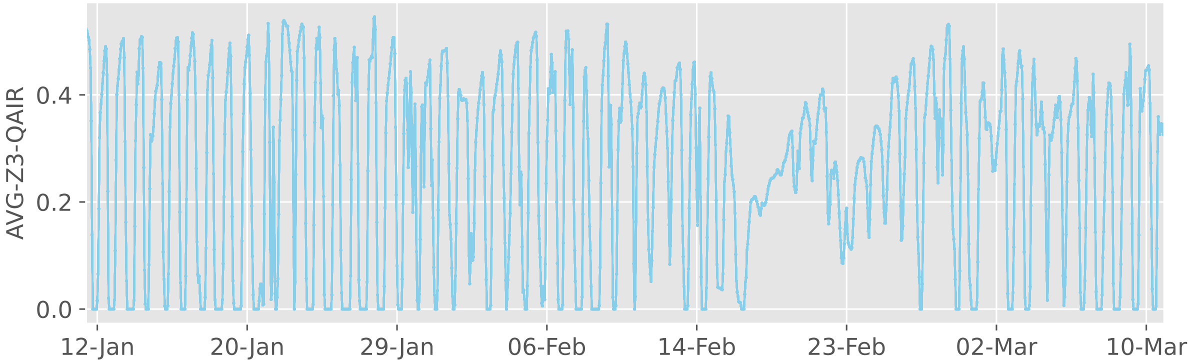
Source: The author.

Histogram of the average air flow rate in Zone 2



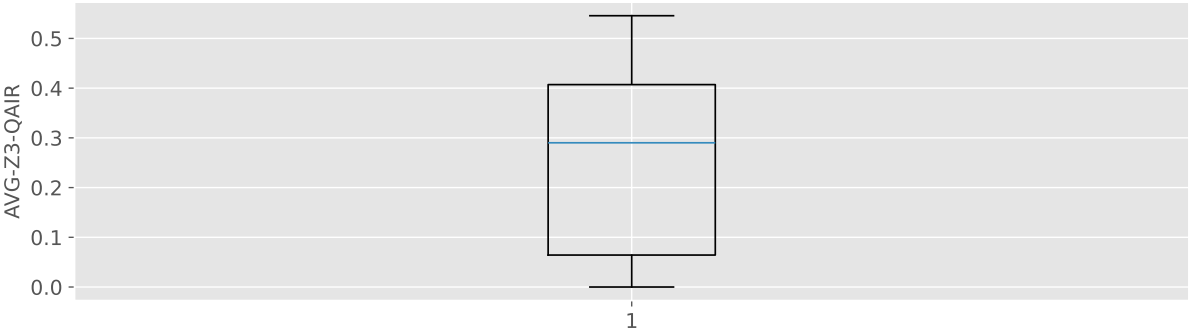
Source: The author.

Time-series of the average air flow rate in Zone 3



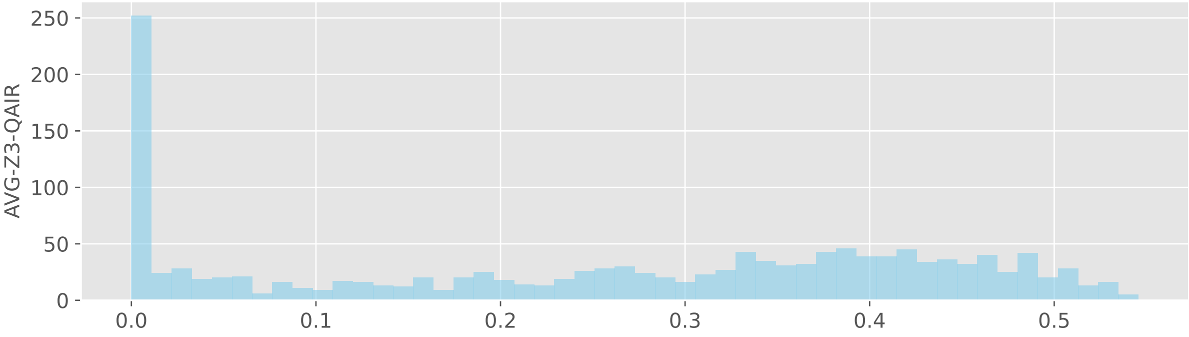
Source: The author.

Box plot of the average air flow rate in Zone 3



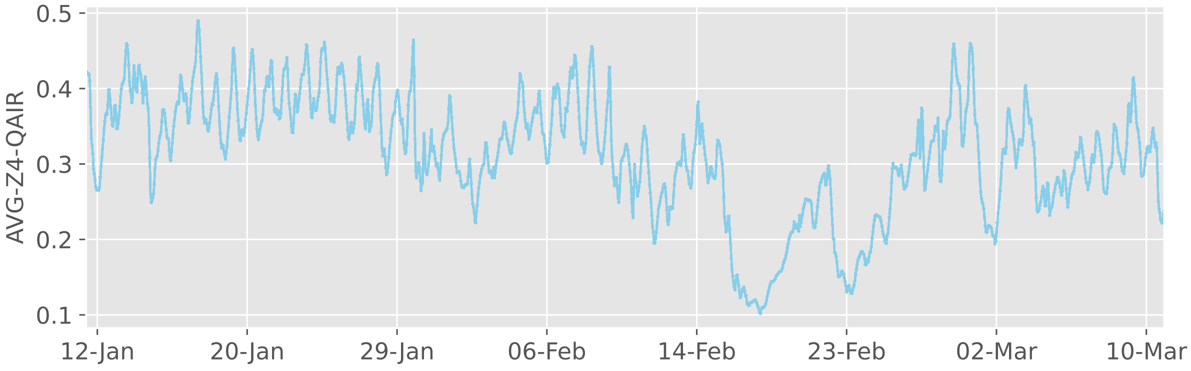
Source: The author.

Histogram of the average air flow rate in Zone 3



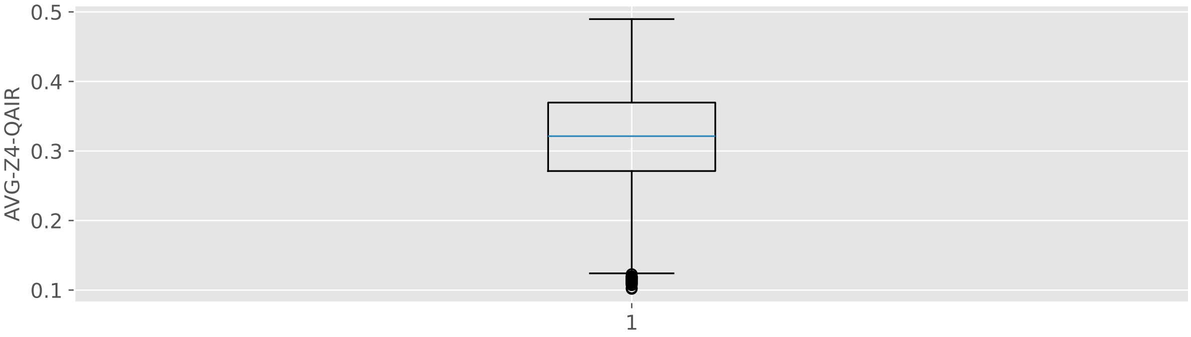
Source: The author.

Time-series of the average air flow rate in Zone 4



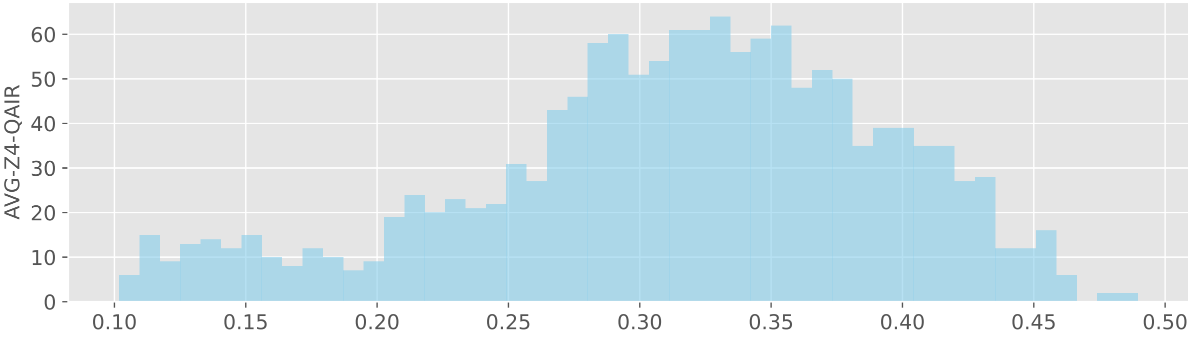
Source: The author.

Box plot of the average air flow rate in Zone 4



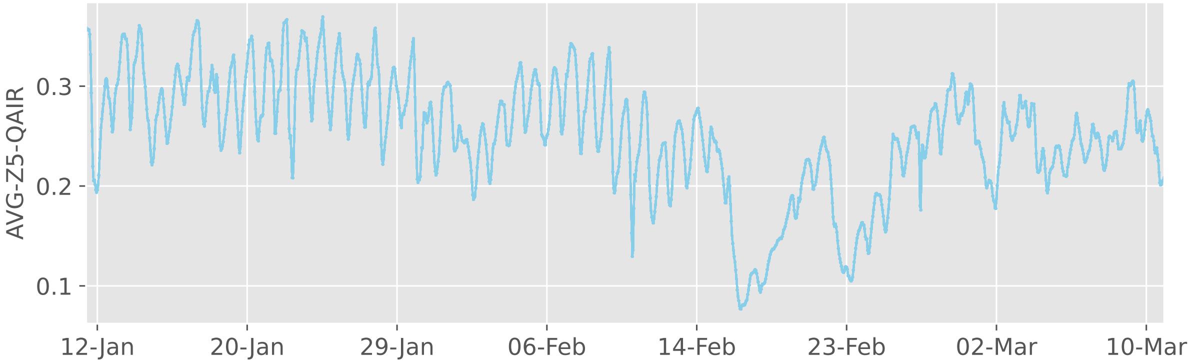
Source: The author.

Histogram of the average air flow rate in Zone 4



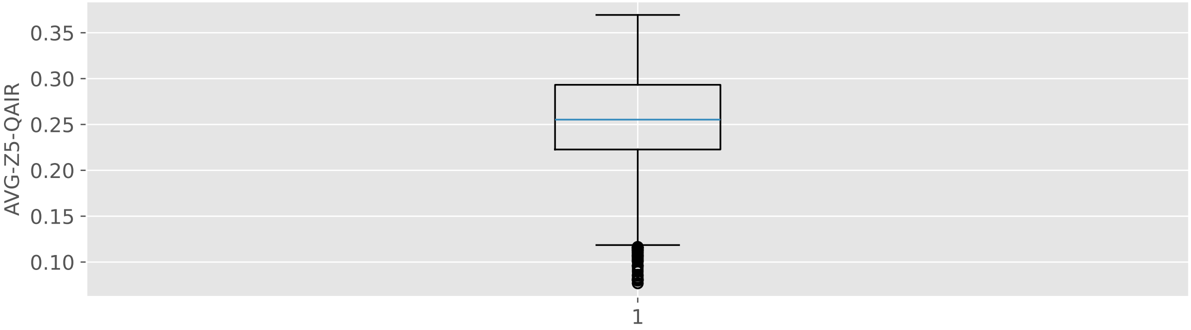
Source: The author.

Time-series of the average air flow rate in Zone 5



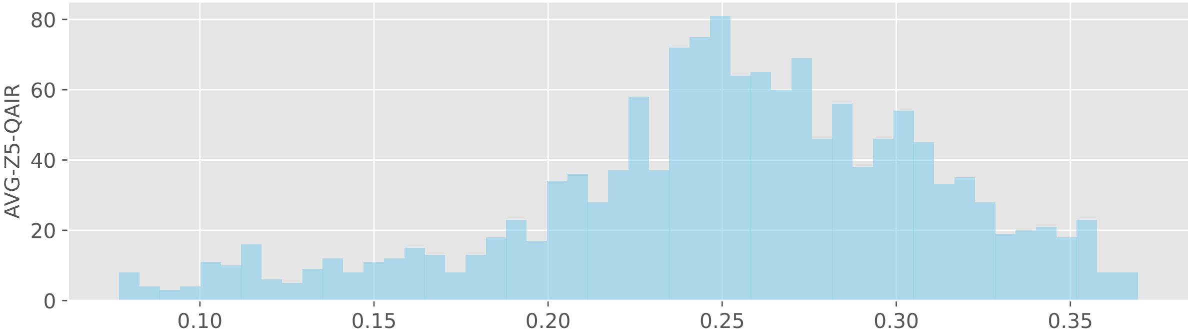
Source: The author.

Box plot of the average air flow rate in Zone 5



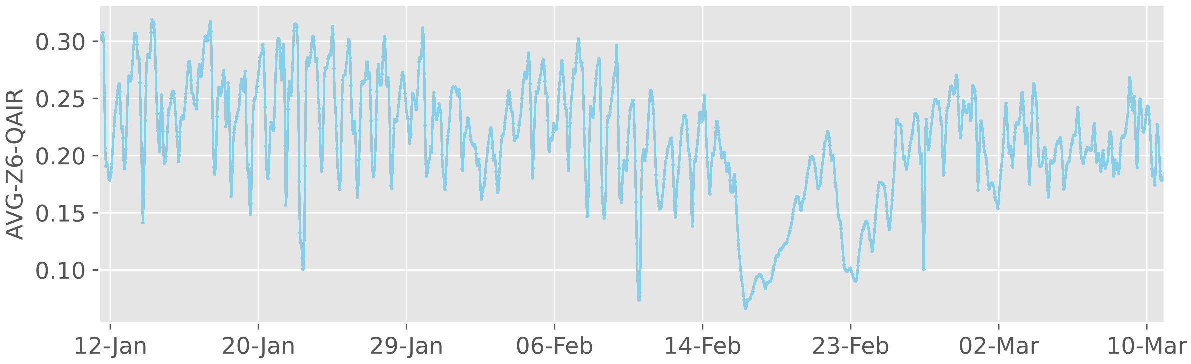
Source: The author.

Histogram of the average air flow rate in Zone 5



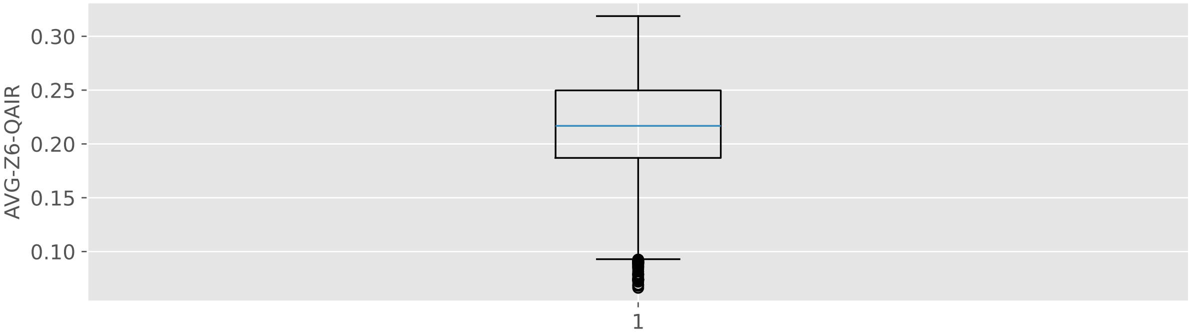
Source: The author.

Time-series of the average air flow rate in Zone 6



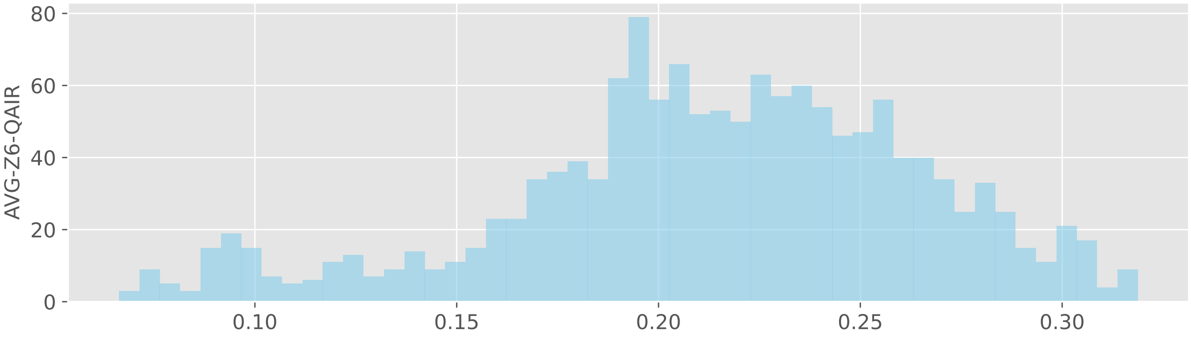
Source: The author.

Box plot of the average air flow rate in Zone 6



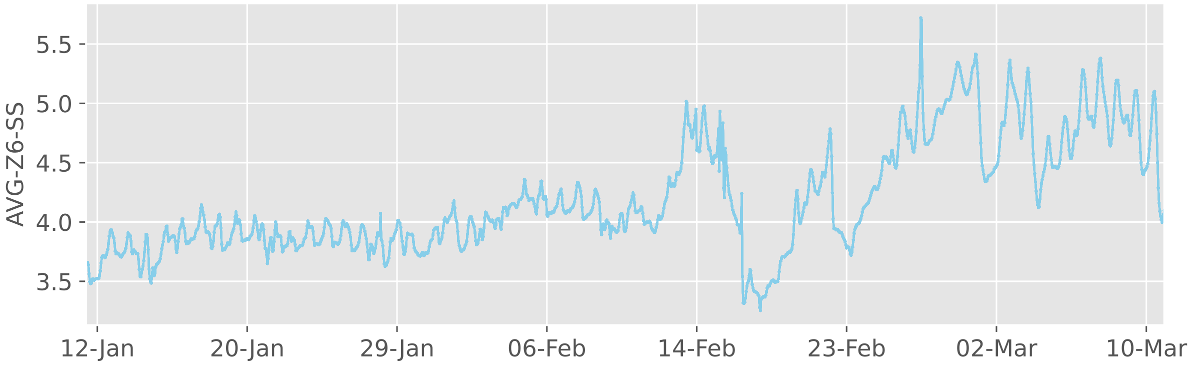
Source: The author.

Histogram of the average air flow rate in Zone 6



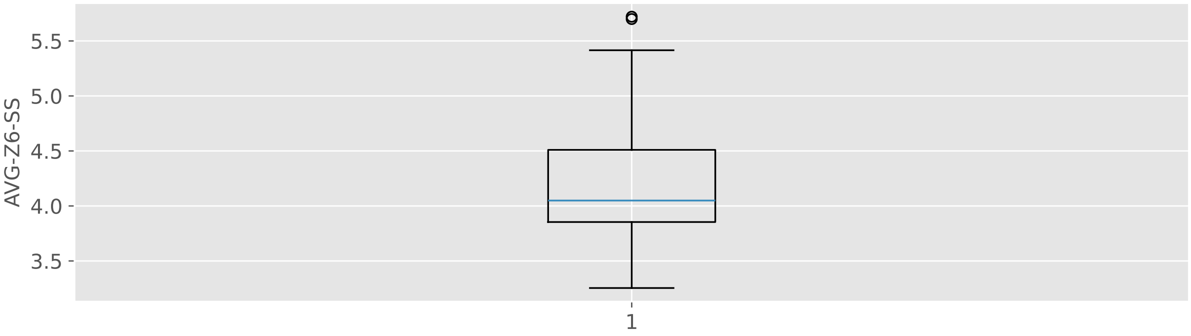
Source: The author.

Time-series of the average total suspended solids in Zone 6



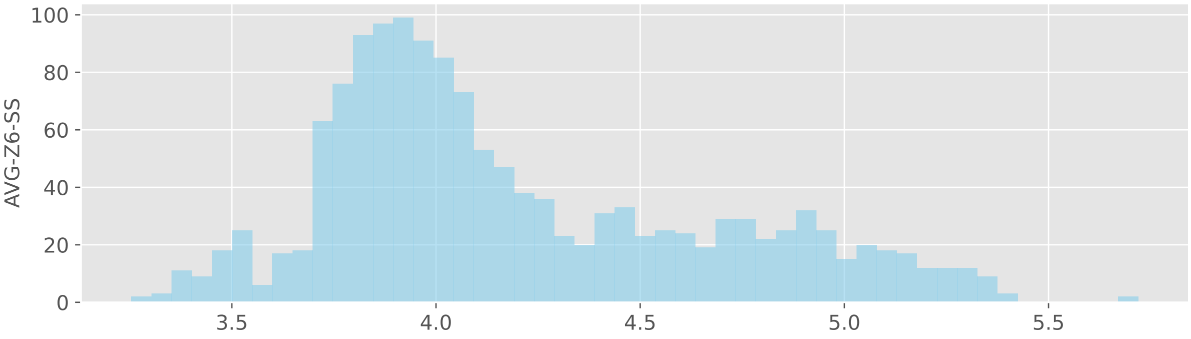
Source: The author.

Box plot of the average total suspended solids in Zone 6



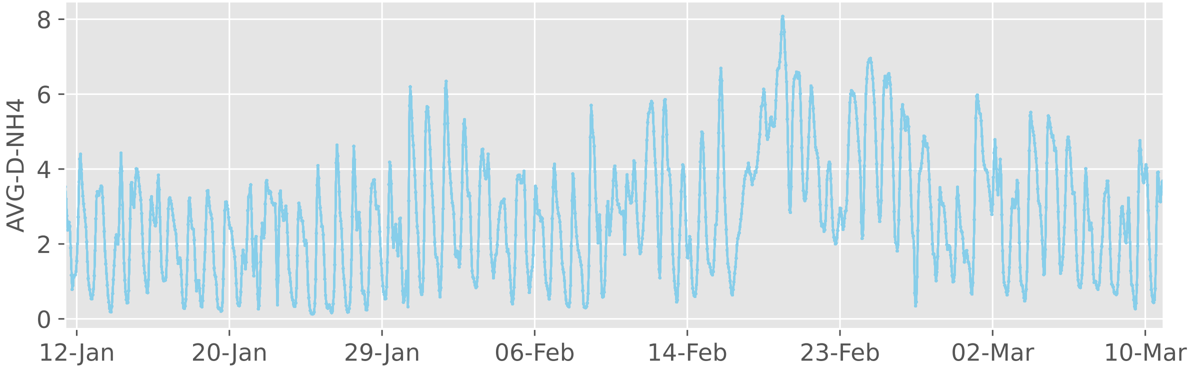
Source: The author.

Histogram of the average total suspended solids in Zone 6



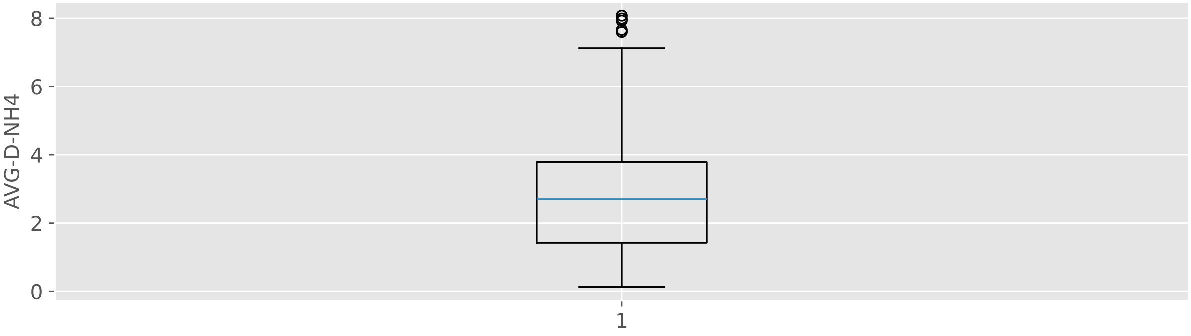
Source: The author.

Time-series of the average ammonia at the end of degas



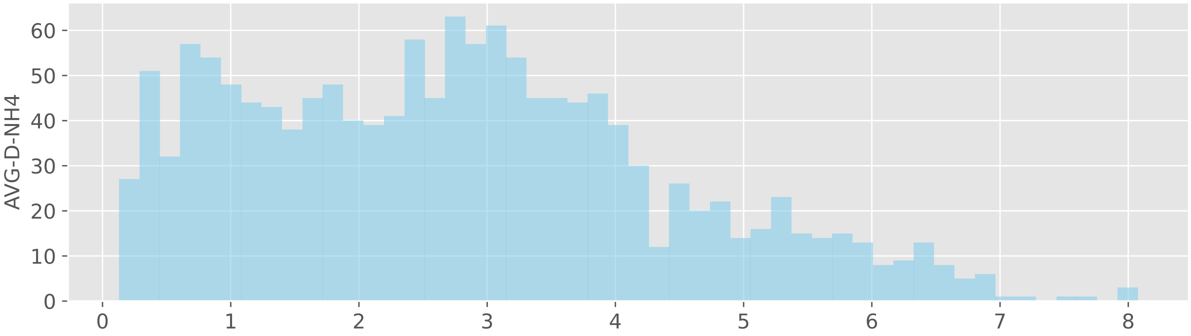
Source: The author.

Box plot of the average ammonia at the end of degas



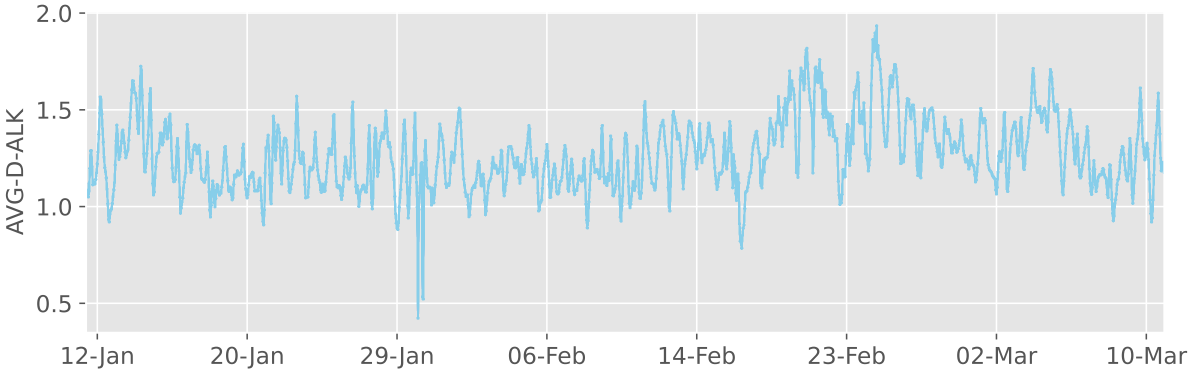
Source: The author.

Histogram of the average ammonia at the end of degas



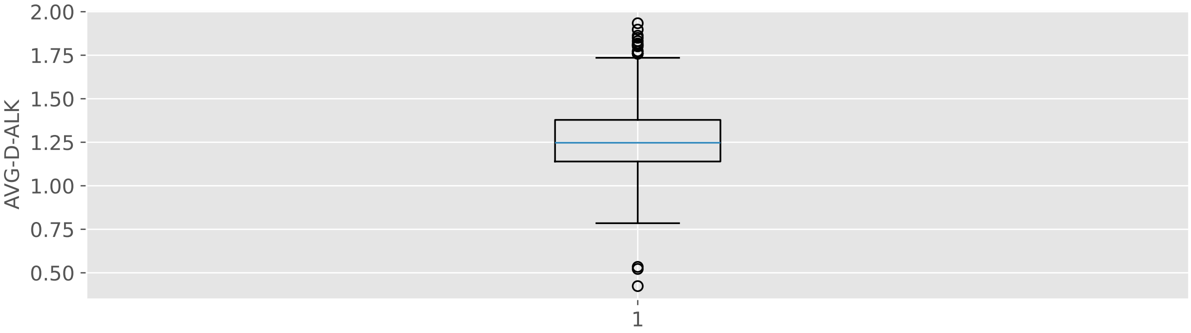
Source: The author.

Time-series of the average alkalinity at the end of degas



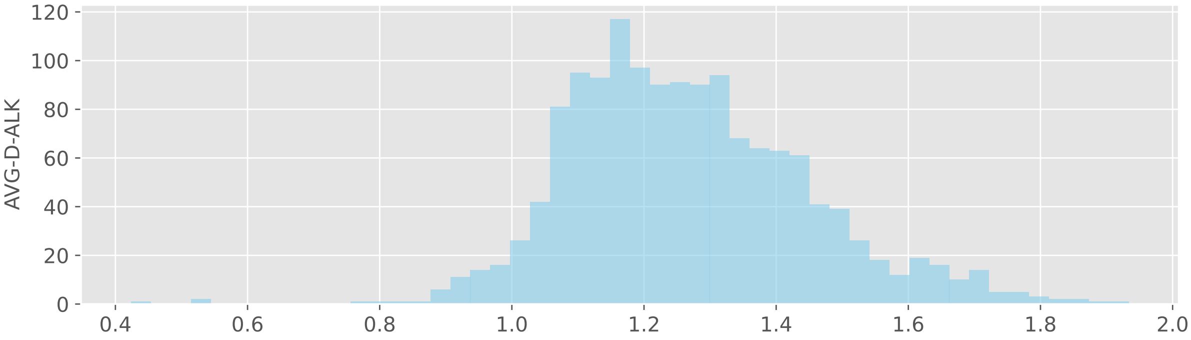
Source: The author.

Box plot of the average alkalinity at the end of degas



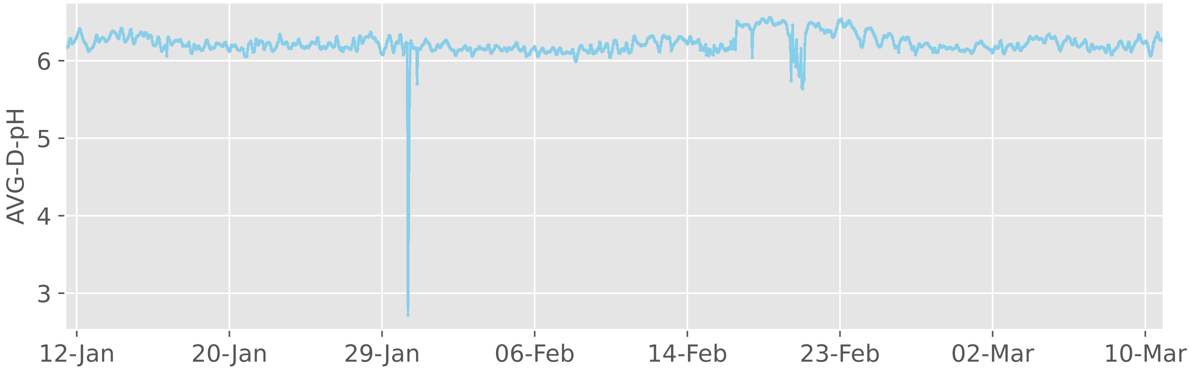
Source: The author.

Histogram of the average alkalinity at the end of degas



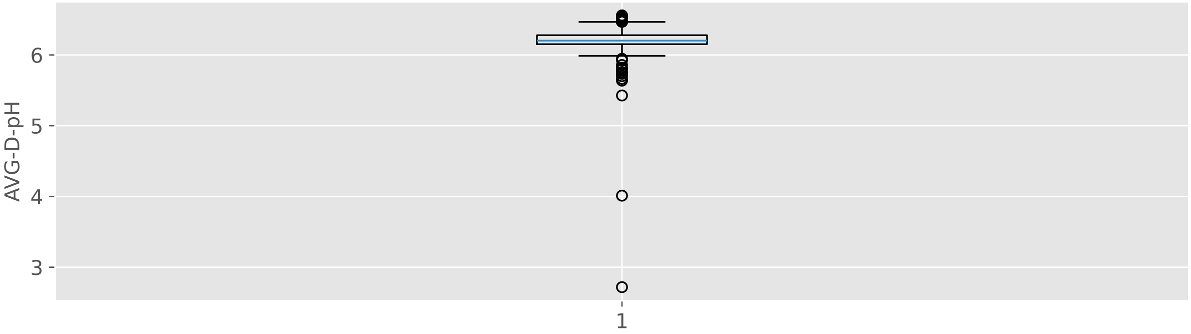
Source: The author.

Time-series of the average pH at the end of degas



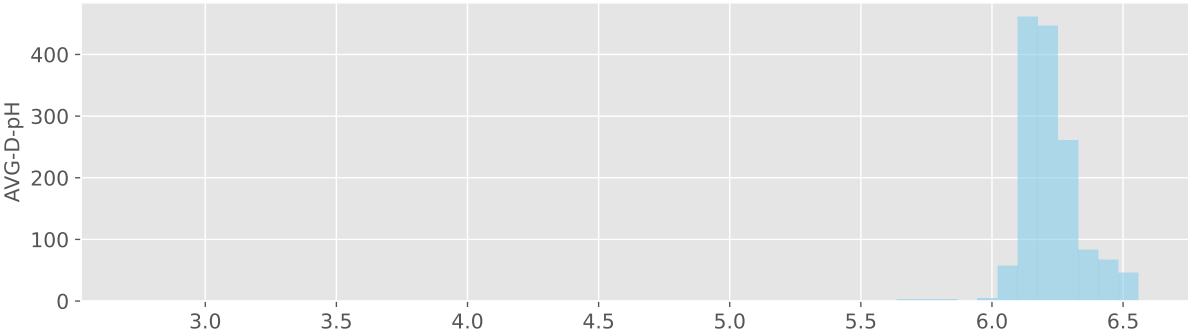
Source: The author.

Box plot of the average pH at the end of degas



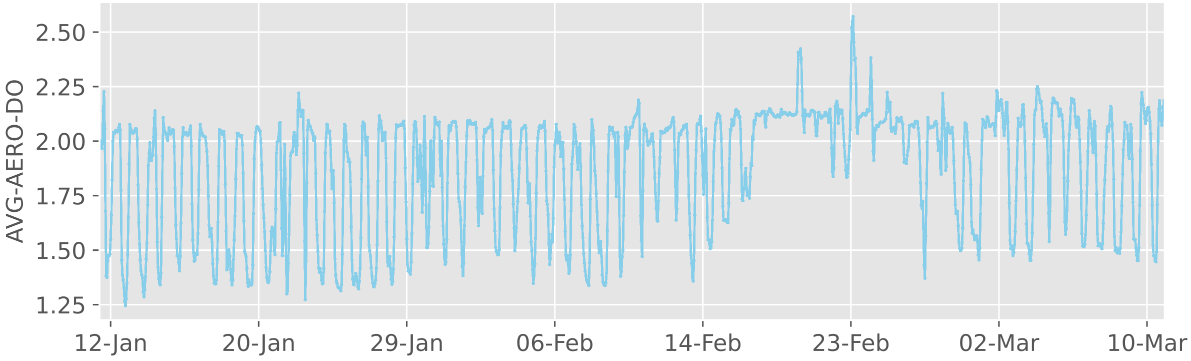
Source: The author.

Histogram of the average pH at the end of degas



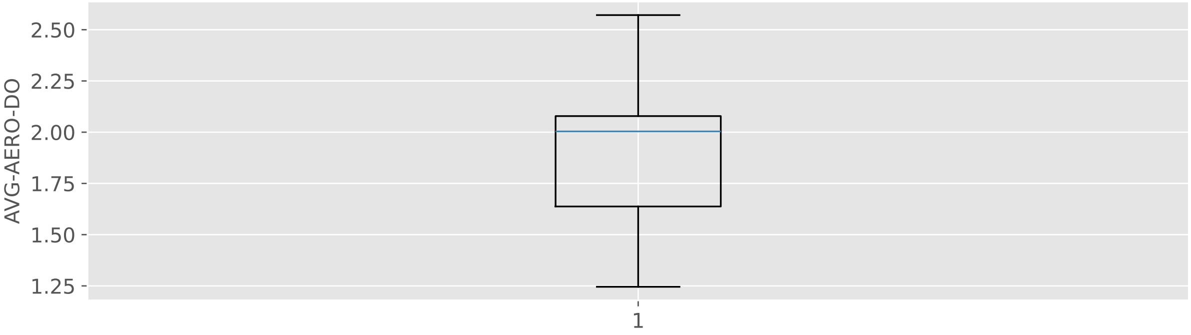
Source: The author.

Time-series of the average amount of dissolved oxygen (Zone 2 to Zone 6).



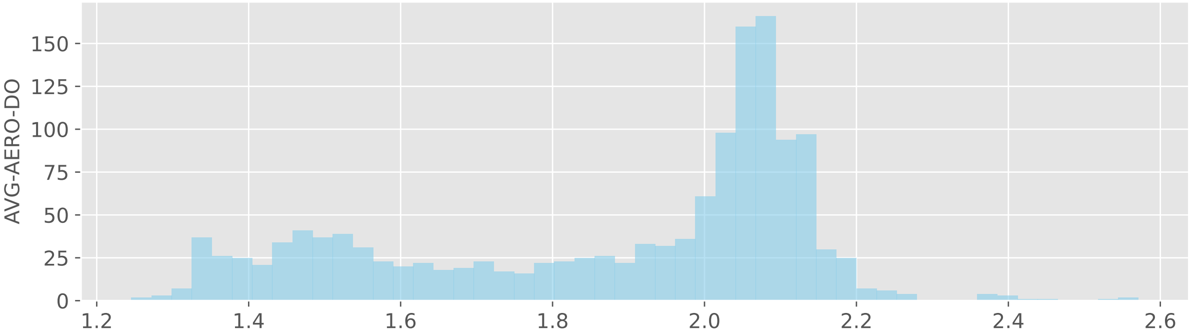
Source: The author.

Box plot of the average amount of dissolved oxygen (Zone 2 to Zone 6).



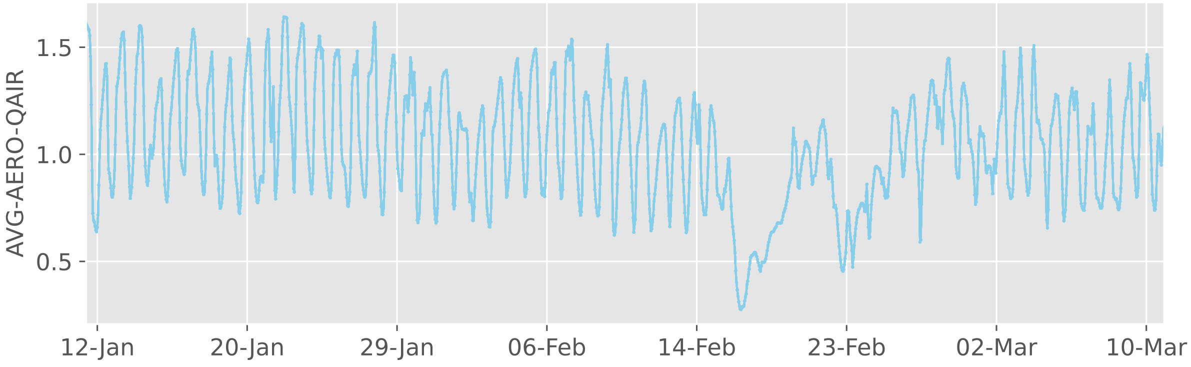
Source: The author.

Histogram of the average amount of dissolved oxygen (Zone 2 to Zone 6).



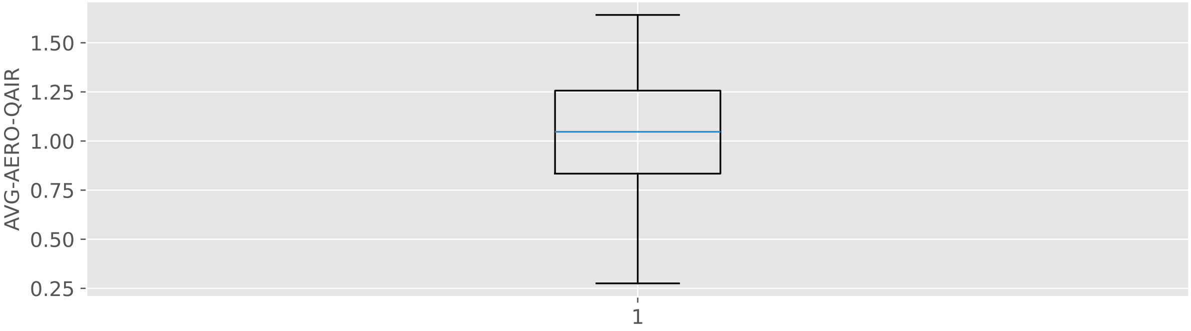
Source: The author.

Time-series of the average amount of air added (Zone 2 to Zone 6).



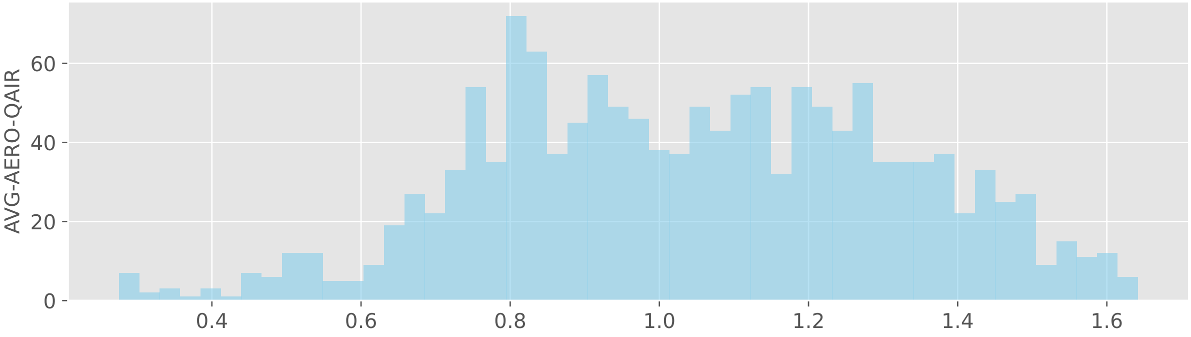
Source: The author.

Box plot of the average amount of air added (Zone 2 to Zone 6).



Source: The author.

Histogram of the average amount of air added (Zone 2 to Zone 6).

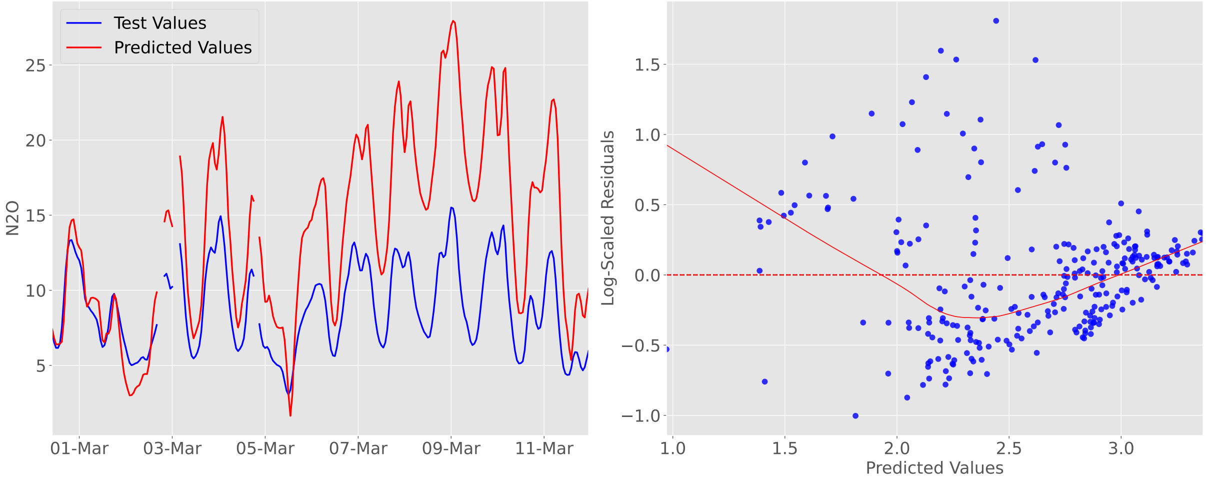


Source: The author.

APPENDIX C – OTHER REGRESSIONS

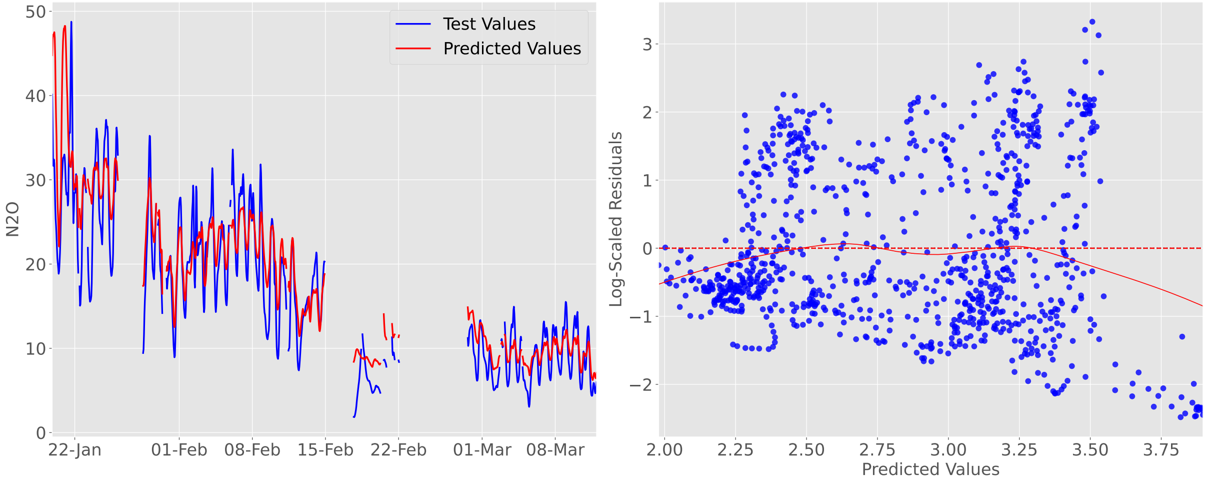
C.1 Full Dataset N_2O Regressions

Ridge Regression in Full Dataset



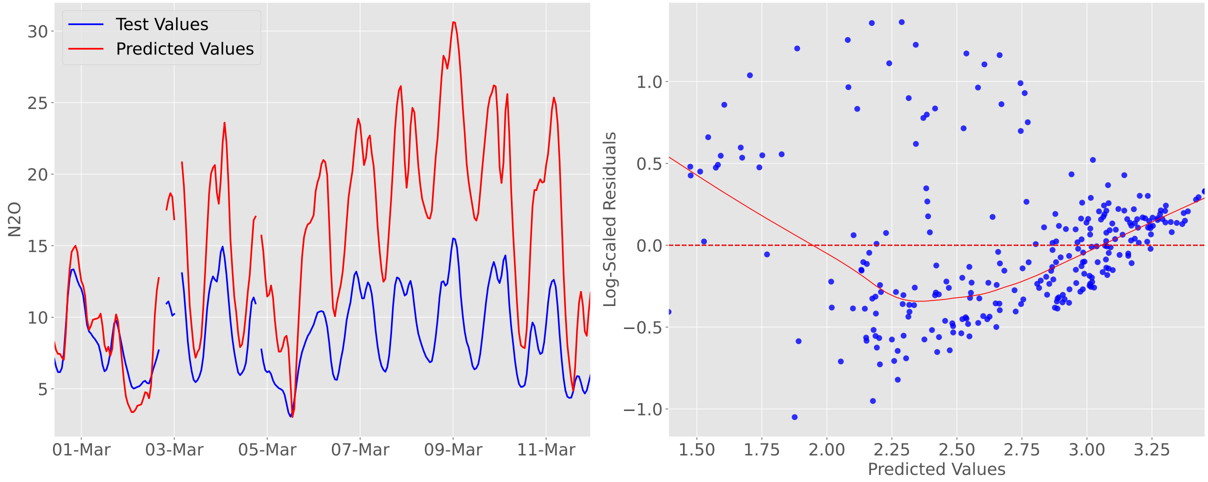
Source: The author.

RW Ridge Regression in Full Dataset



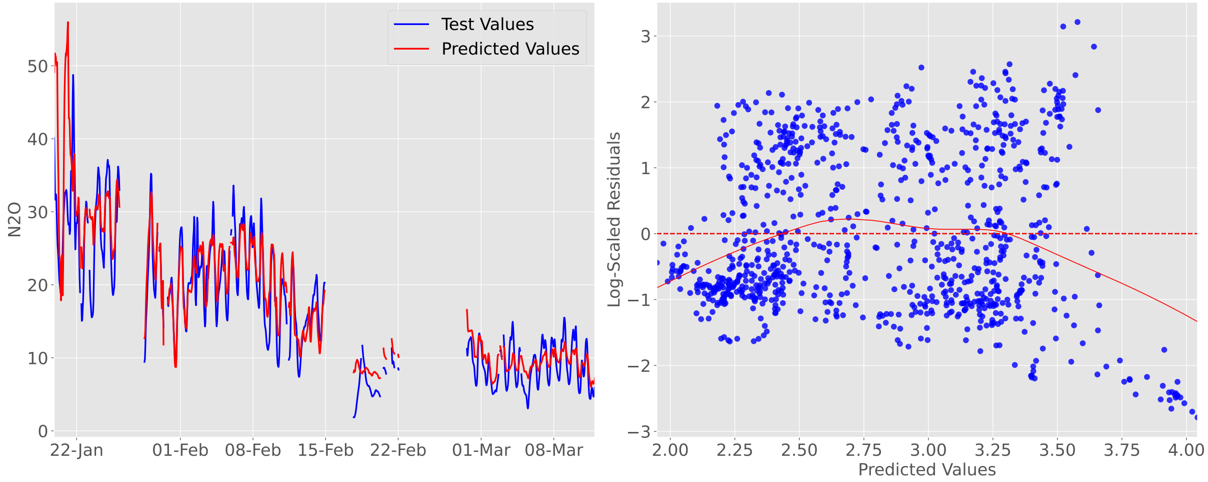
Source: The author.

Lasso Regression in Full Dataset



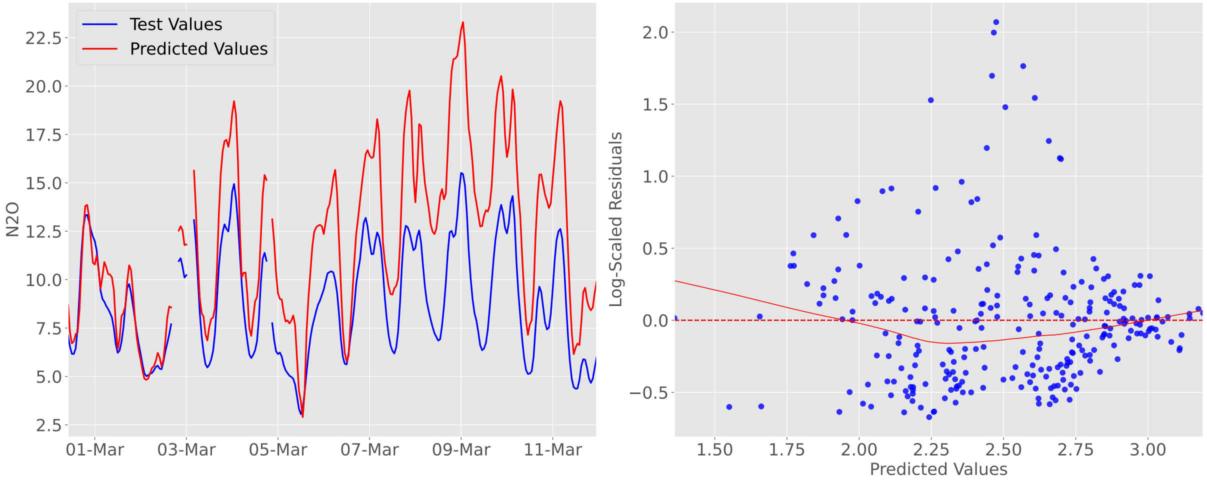
Source: The author.

RW Lasso Regression in Full Dataset



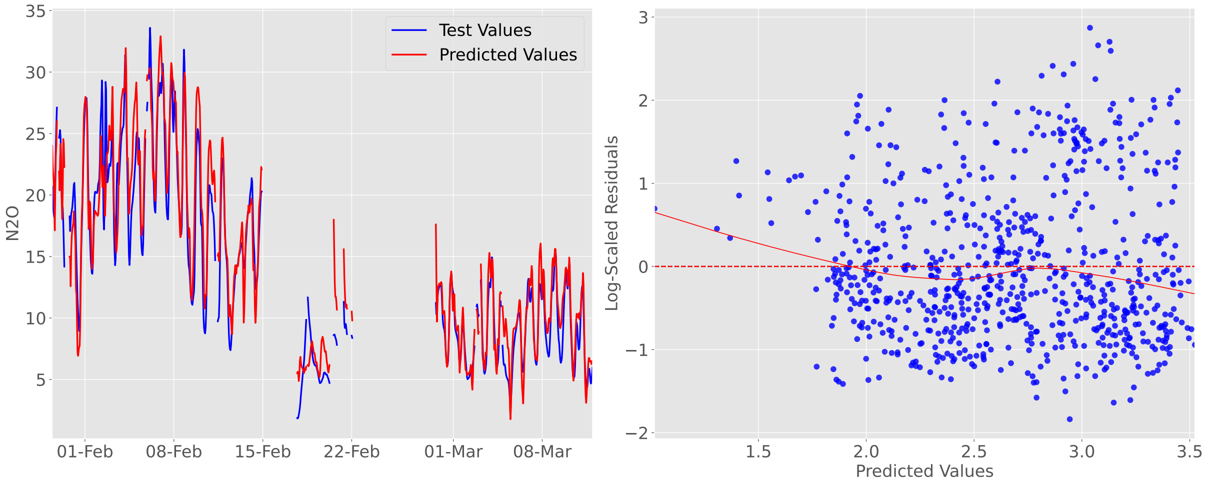
Source: The author.

Huber Regression in Full Dataset



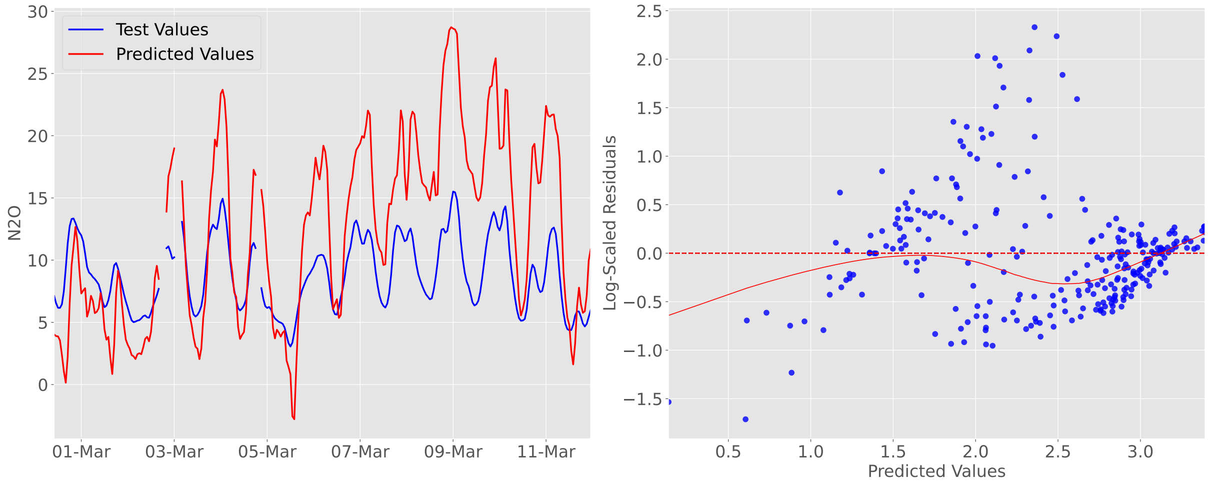
Source: The author.

RW Huber Regression in Full Dataset



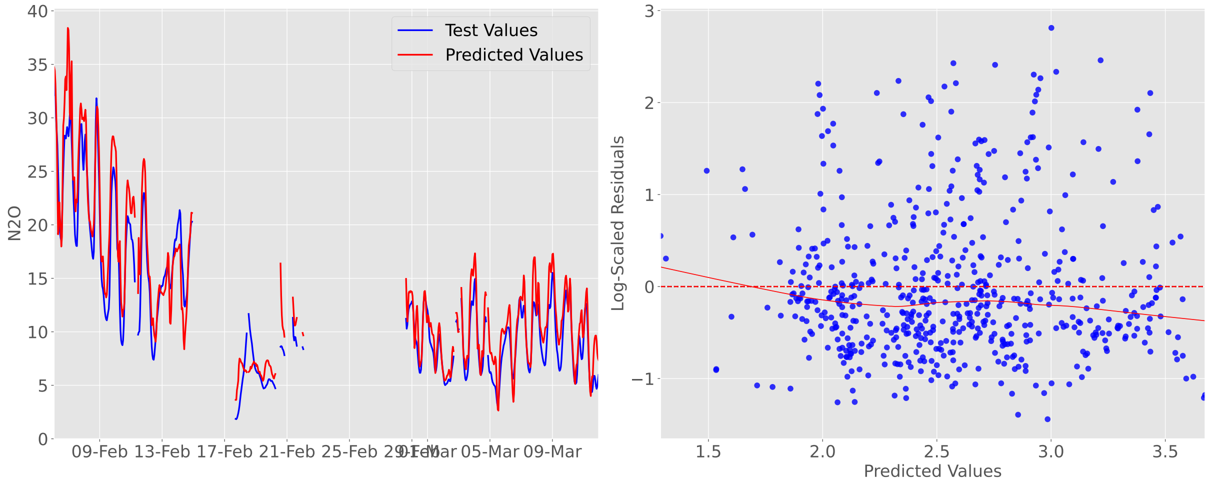
Source: The author.

Bayesian Regression in Full Dataset



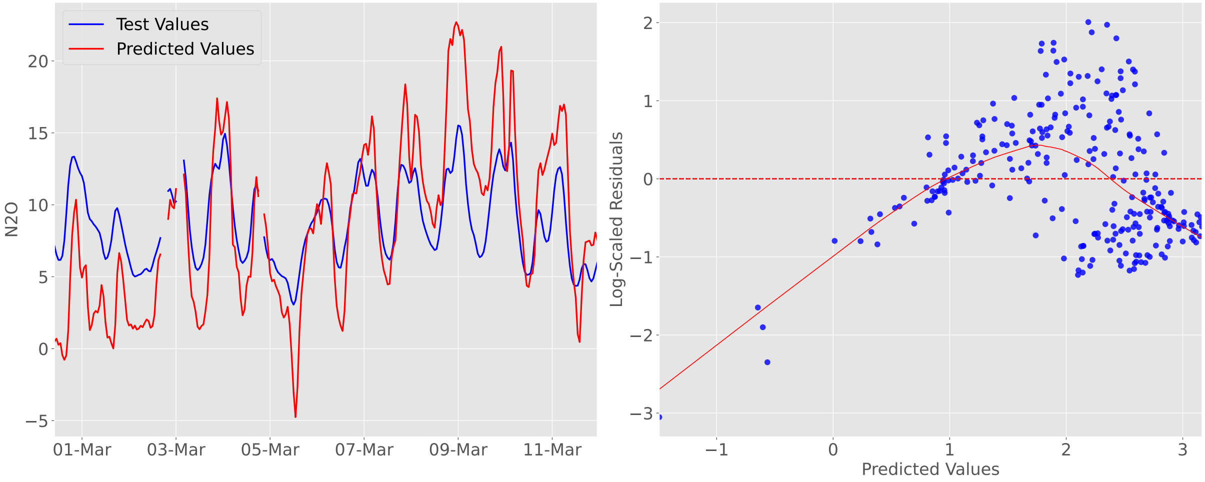
Source: The author.

RW Bayesian Regression in Full Dataset



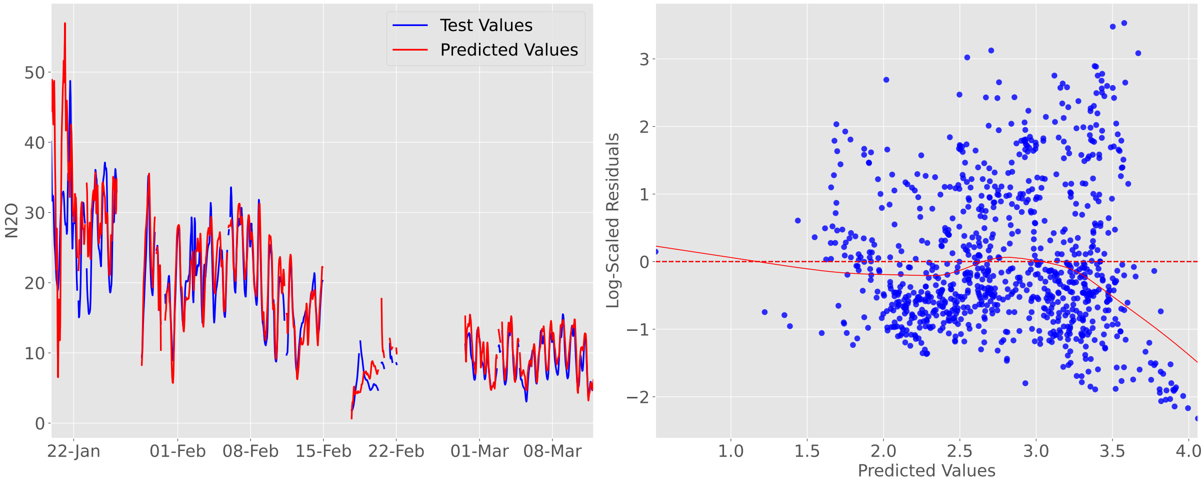
Source: The author.

PLS Regression in Full Dataset



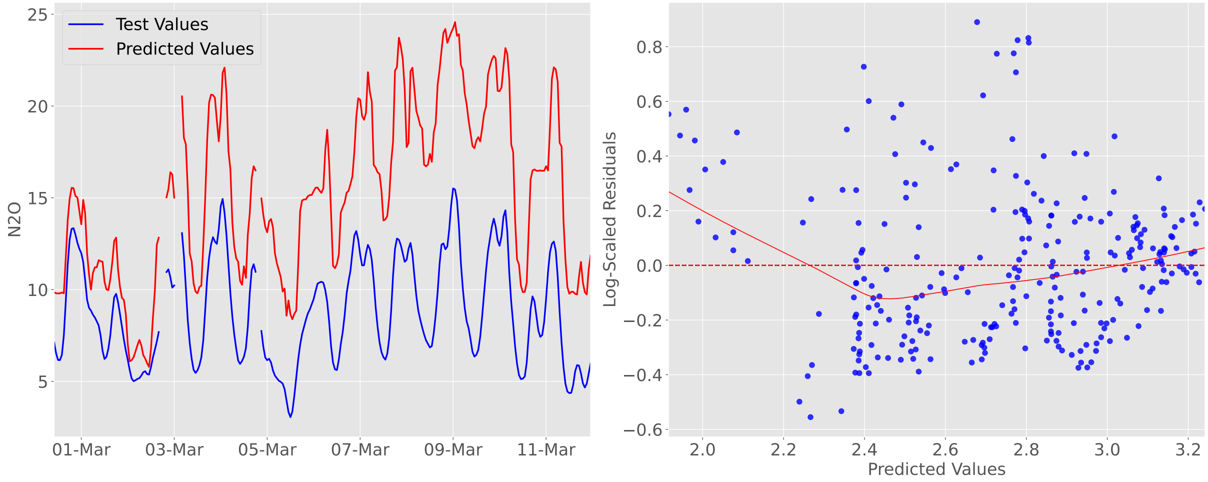
Source: The author.

RW PLS Regression in Full Dataset



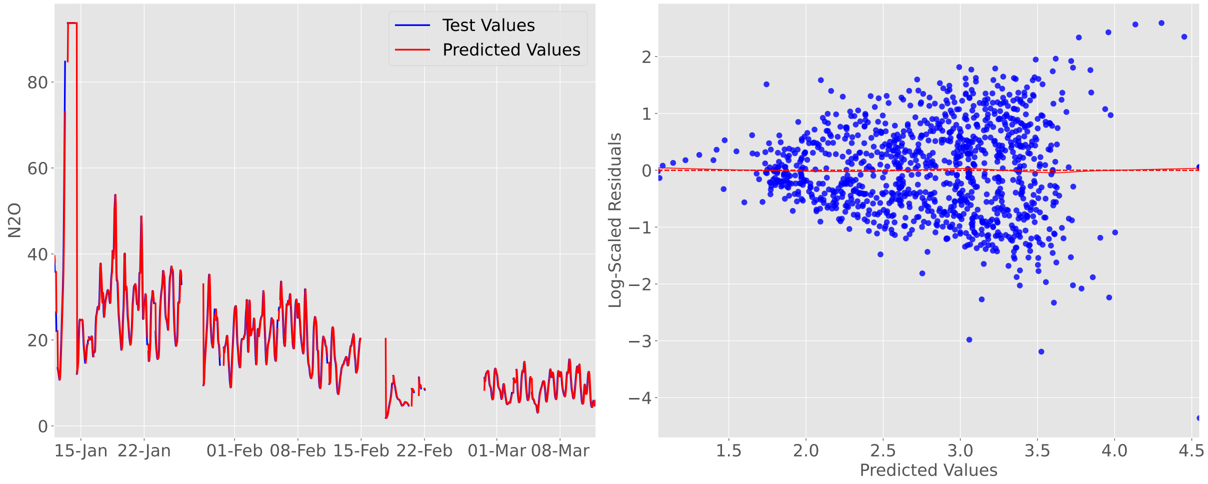
Source: The author.

Random Forest Regression in Full Dataset



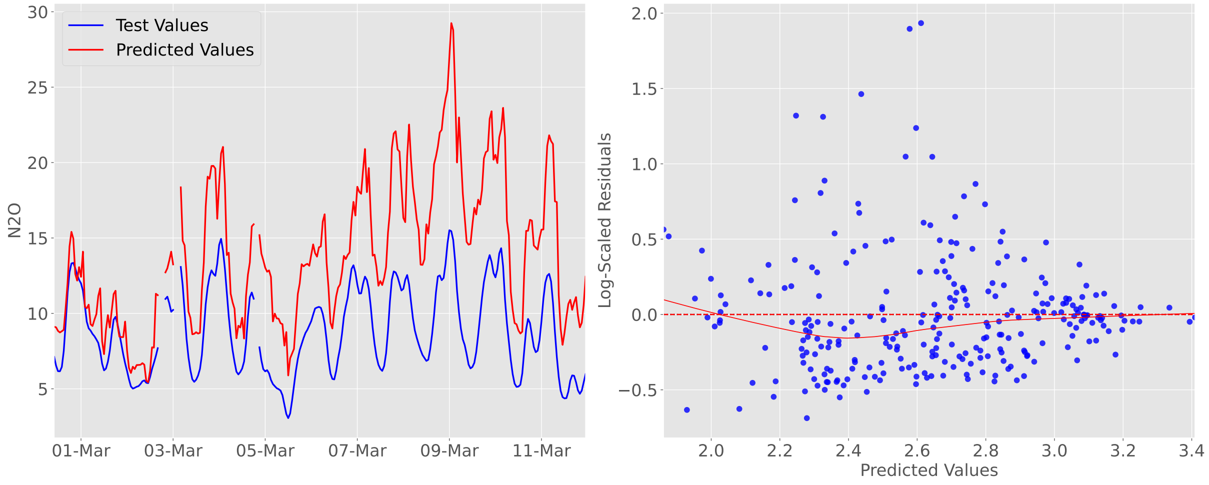
Source: The author.

RW Random Forest Regression in Full Dataset



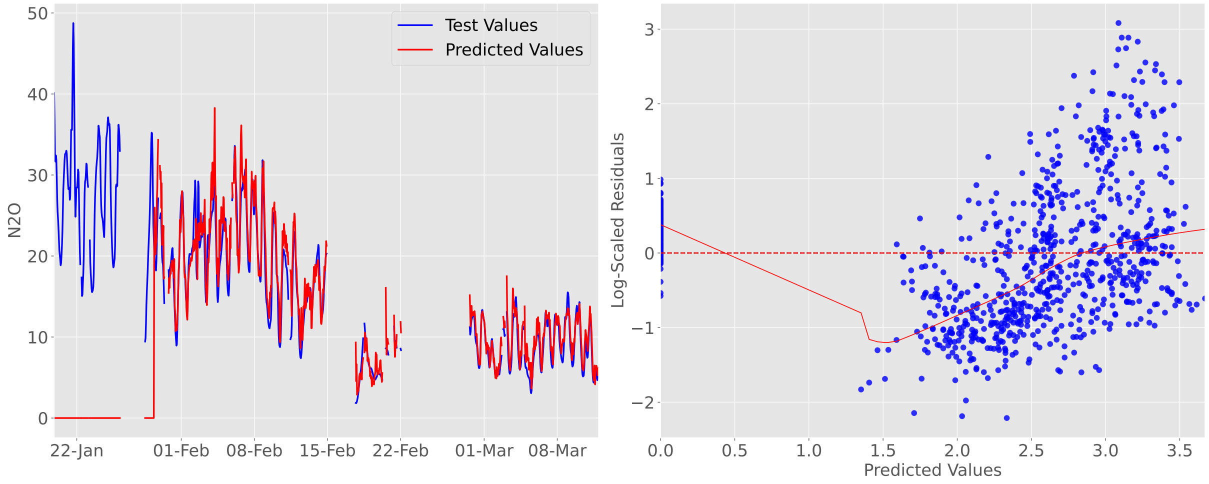
Source: The author.

XGBoost Regression in Full Dataset



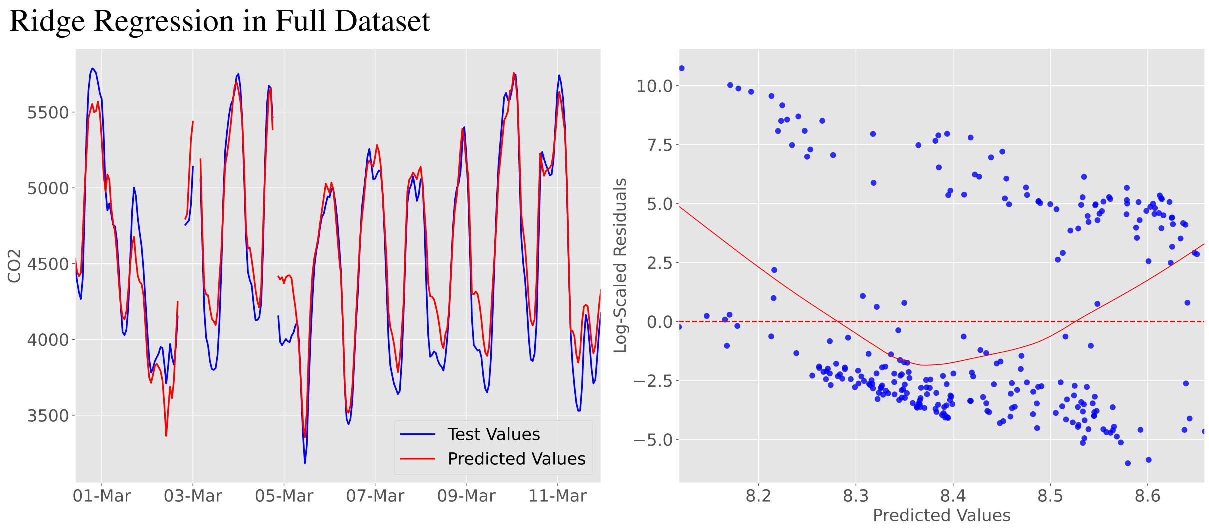
Source: The author.

RW XGBoost Regression in Full Dataset

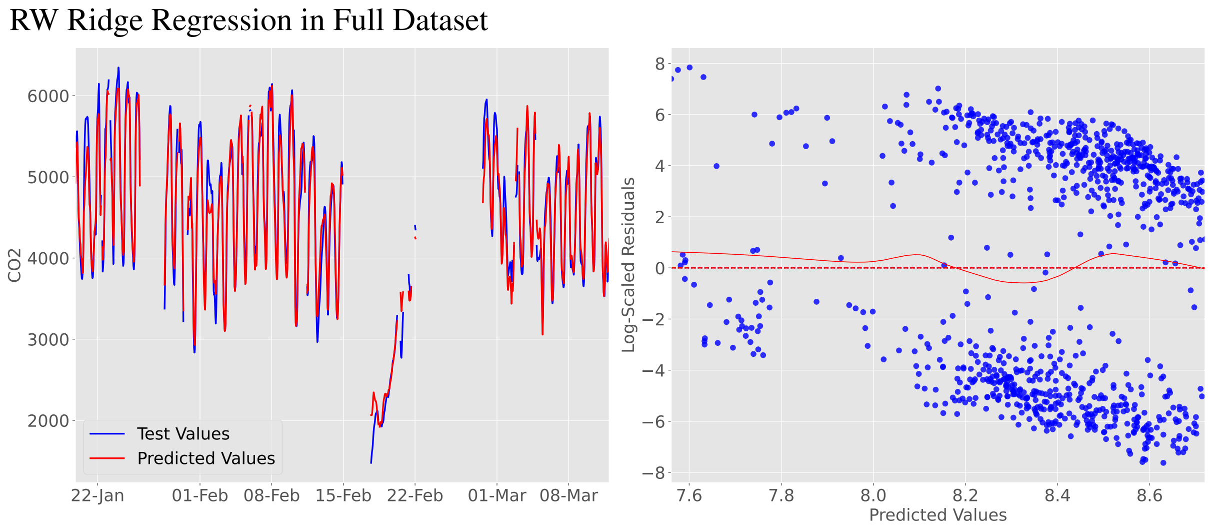


Source: The author.

C.2 Full Dataset CO_2 Regressions

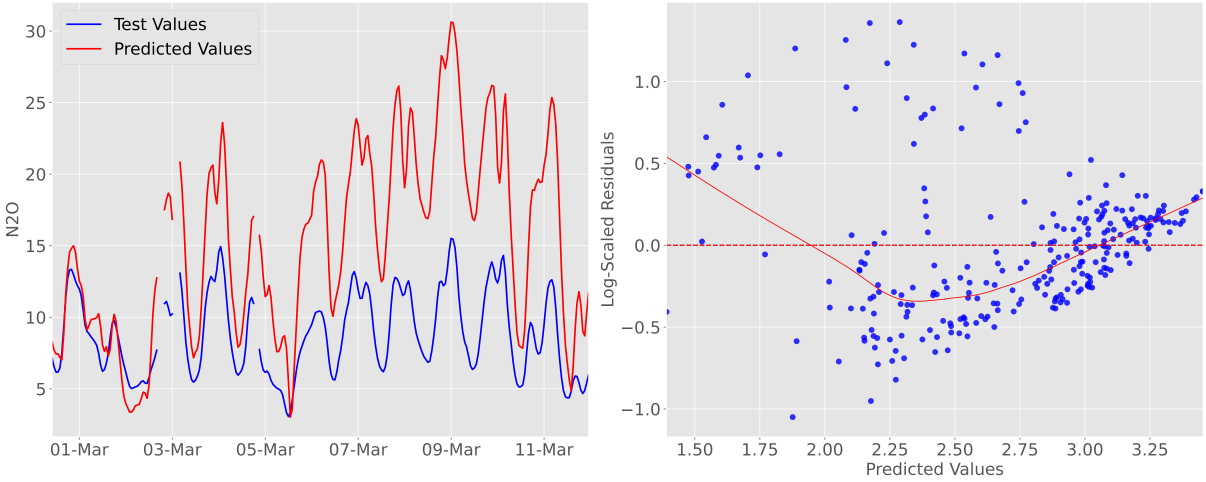

Ridge Regression in Full Dataset

Source: The author.


RW Ridge Regression in Full Dataset

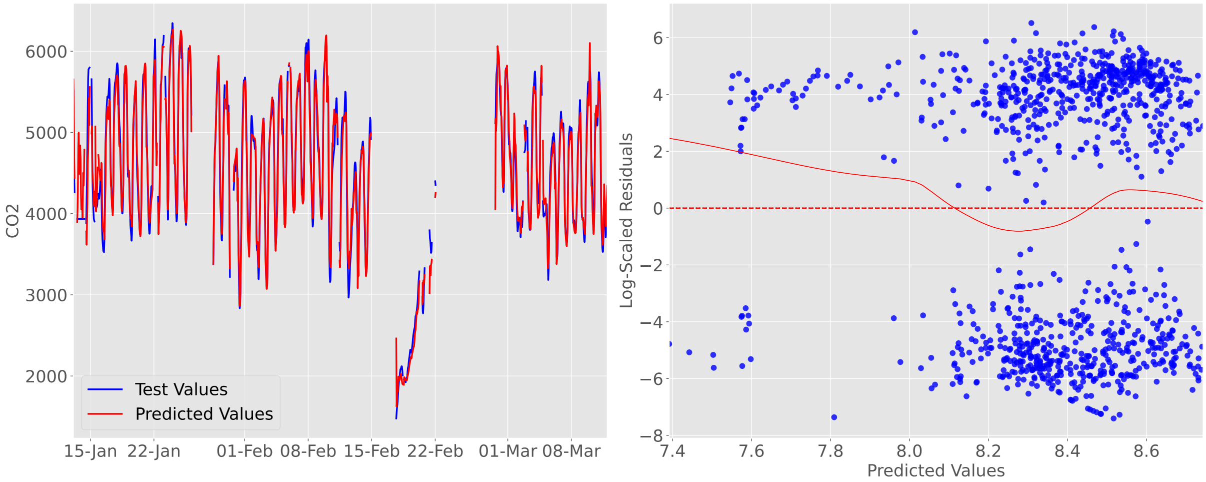
Source: The author.

Lasso Regression in Full Dataset



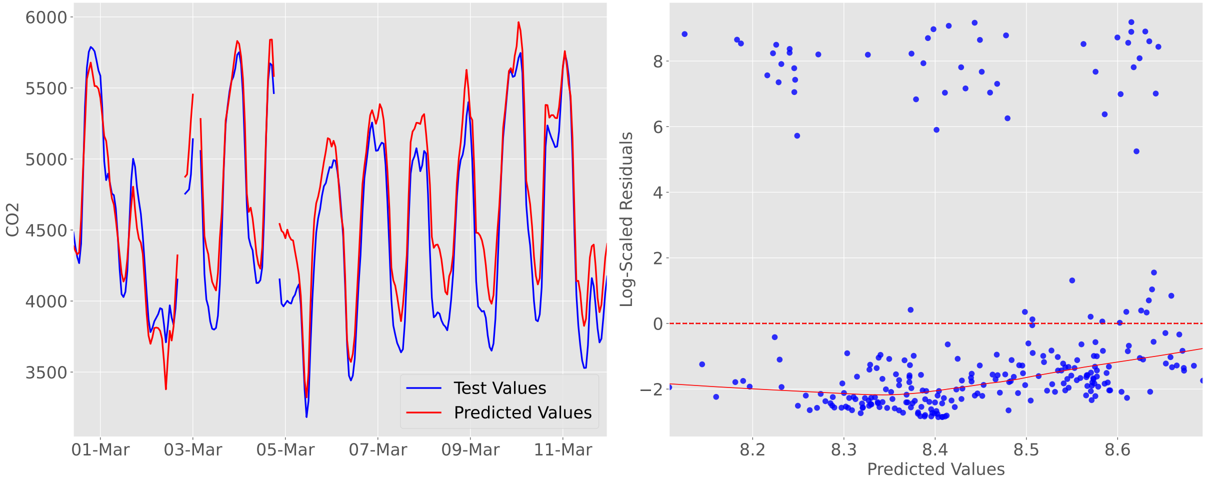
Source: The author.

RW Lasso Regression in Full Dataset



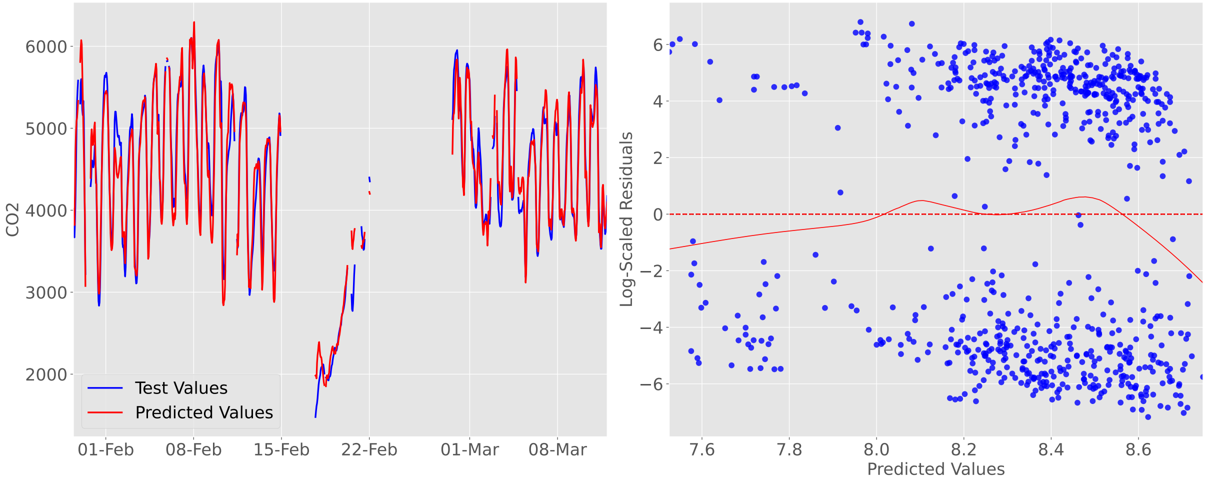
Source: The author.

Huber Regression in Full Dataset



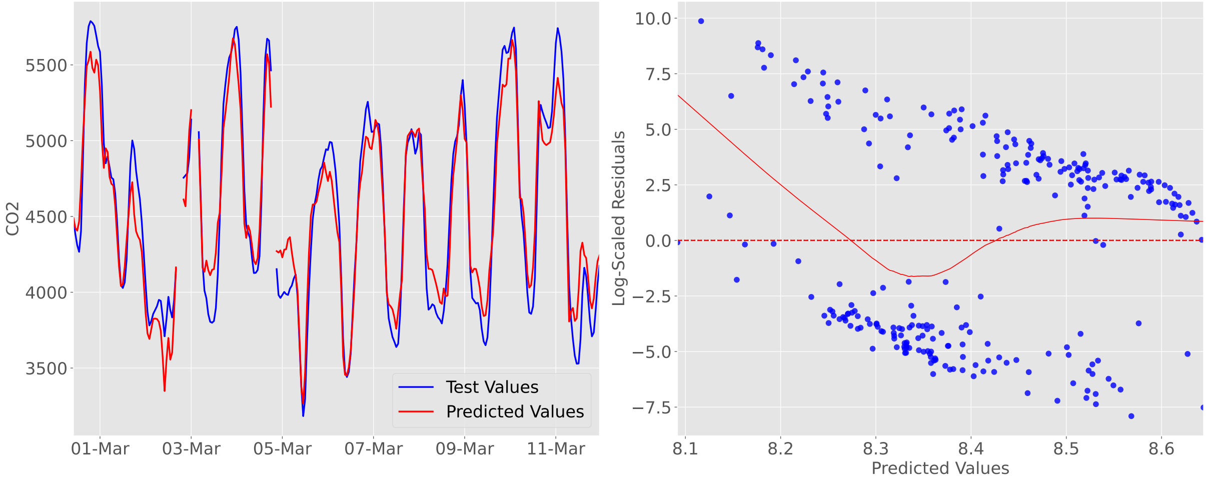
Source: The author.

RW Huber Regression in Full Dataset



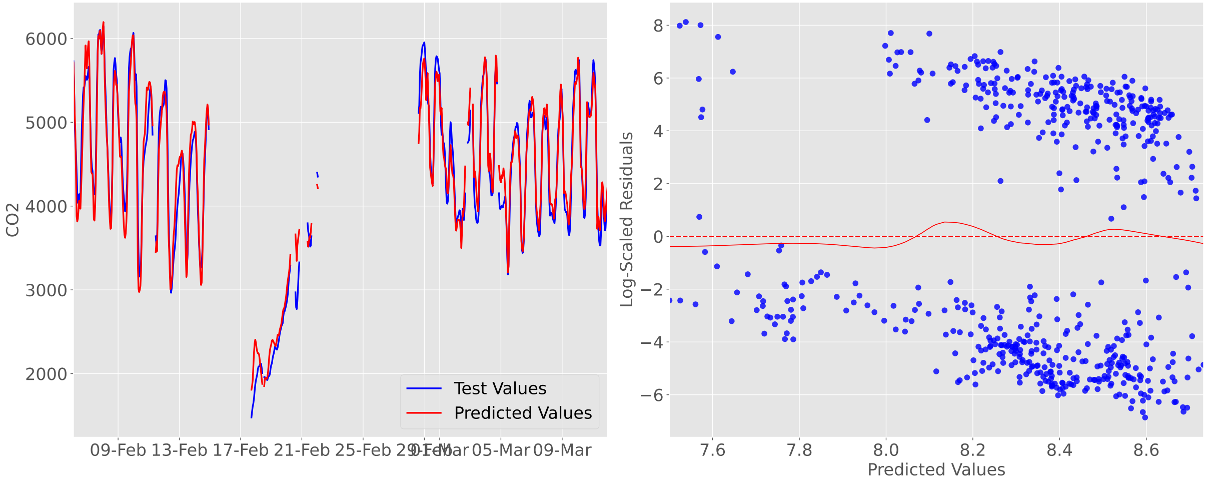
Source: The author.

Bayesian Regression in Full Dataset



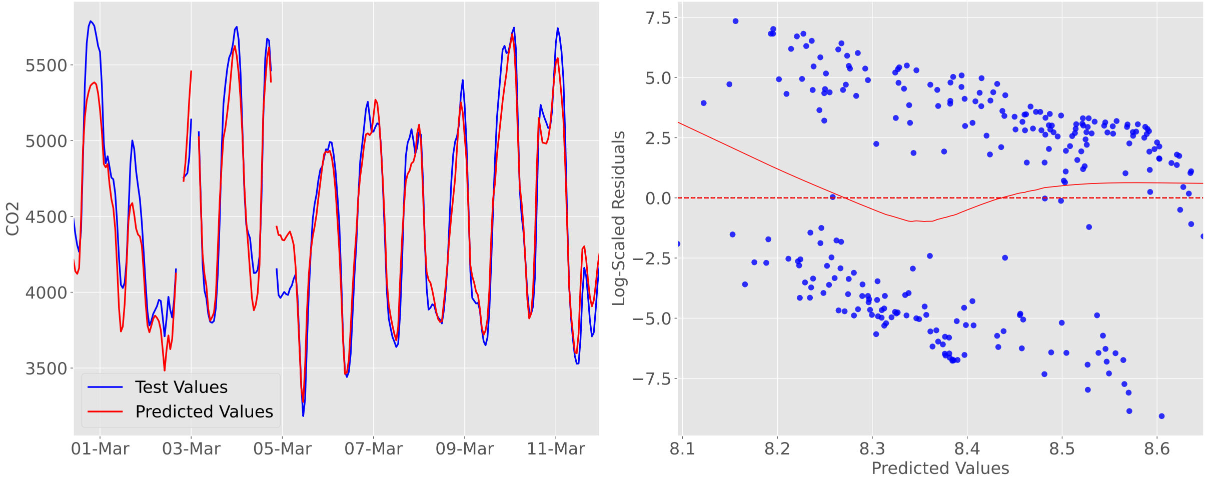
Source: The author.

RW Bayesian Regression in Full Dataset



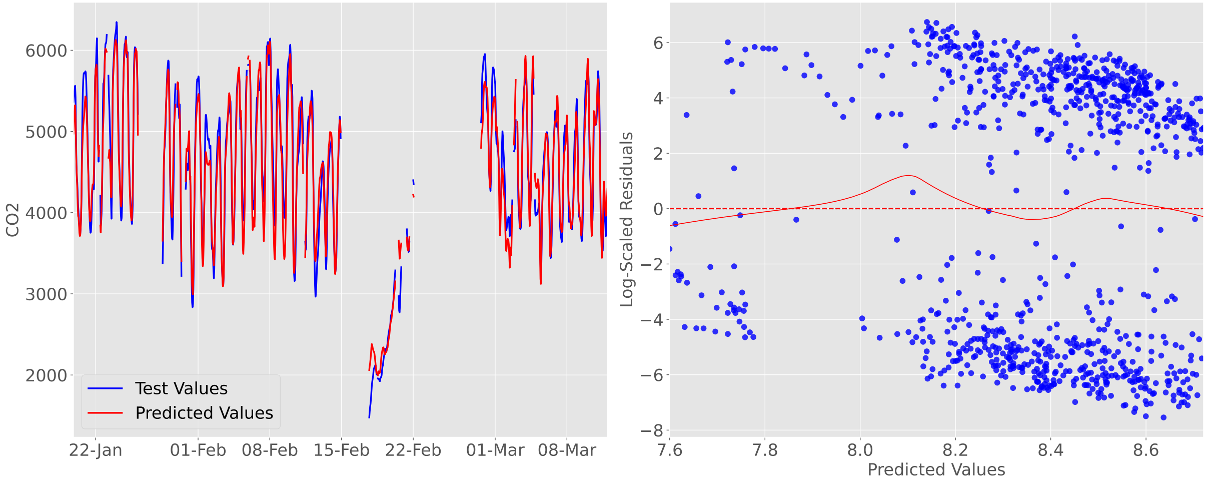
Source: The author.

PLS Regression in Full Dataset



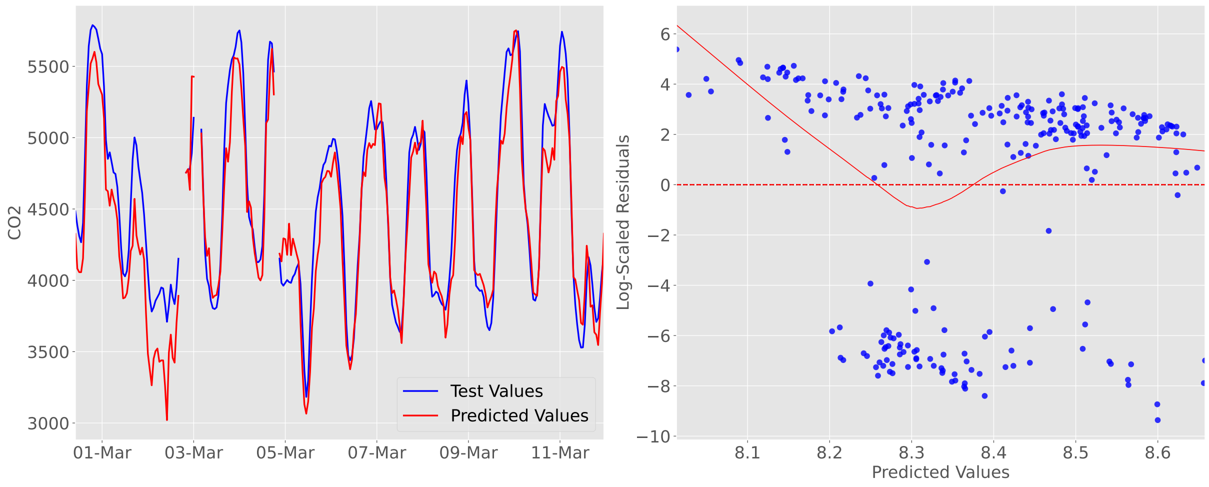
Source: The author.

RW PLS Regression in Full Dataset



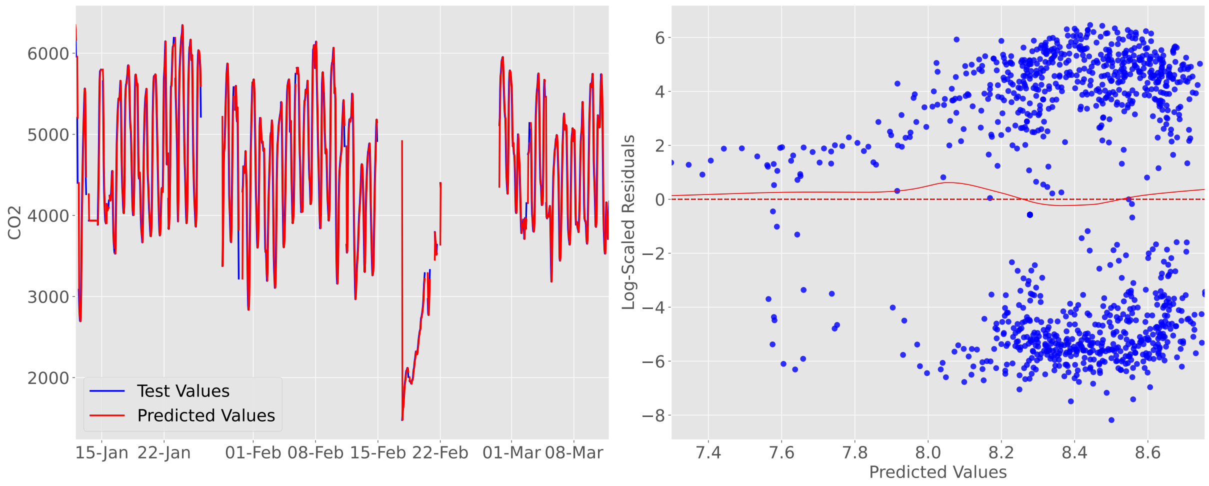
Source: The author.

Random Forest Regression in Full Dataset



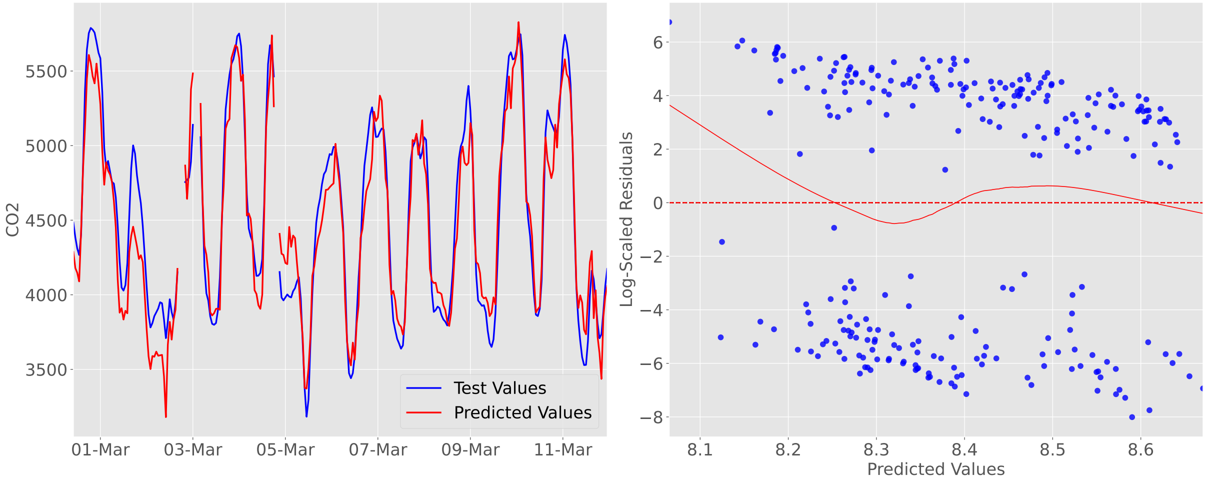
Source: The author.

RW Random Forest Regression in Full Dataset



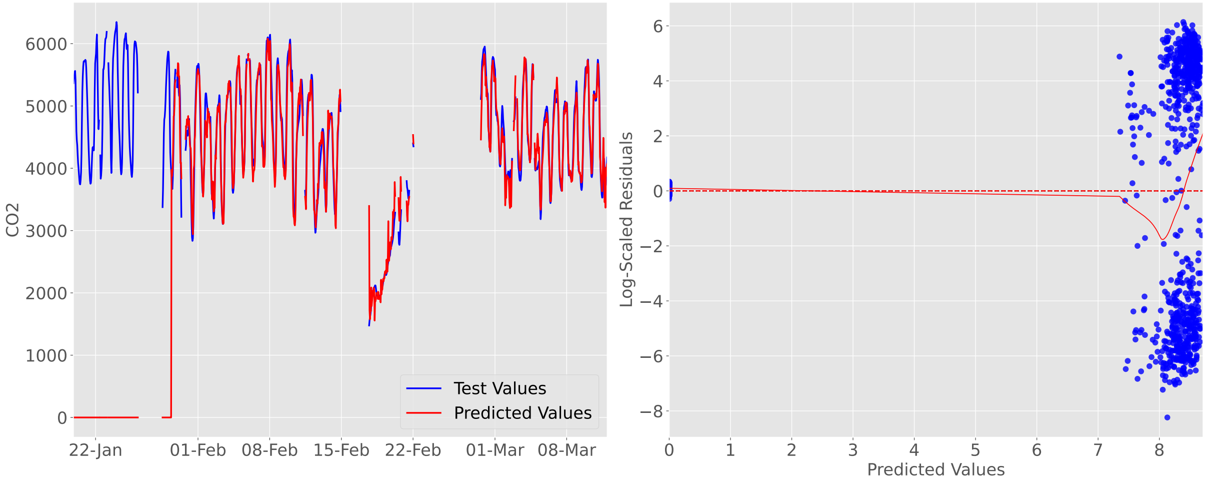
Source: The author.

XGBoost Regression in Full Dataset



Source: The author.

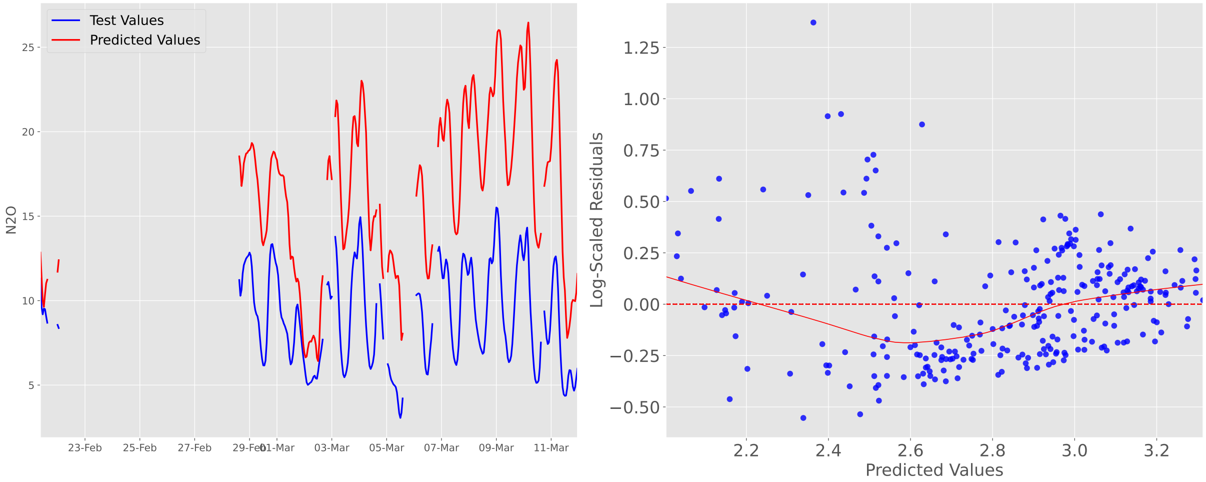
RW XGBoost Regression in Full Dataset



Source: The author.

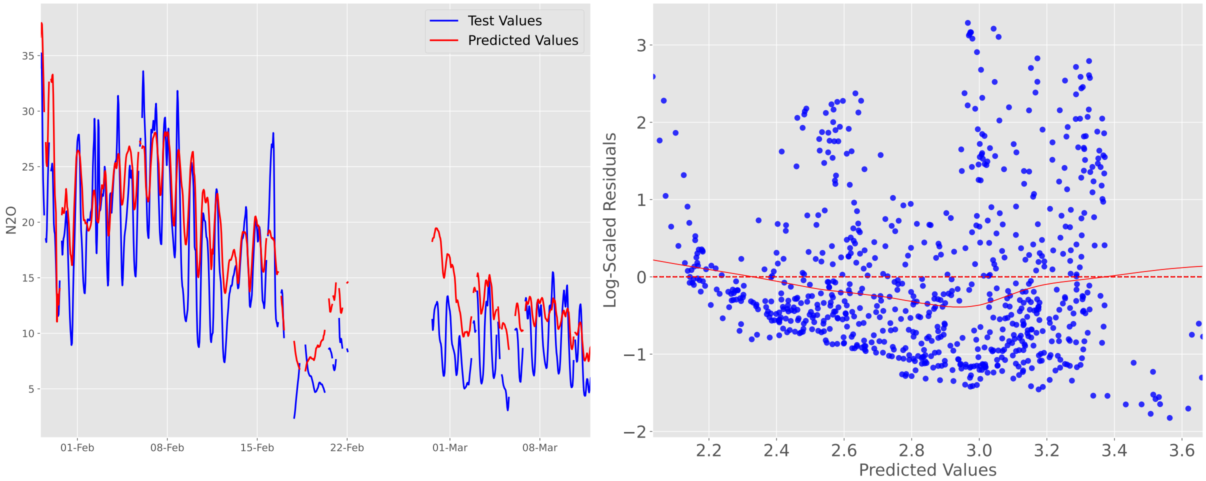
C.3 Reduced Feature Set N_2O Regressions

Ridge Regression in Reduced Feature Set



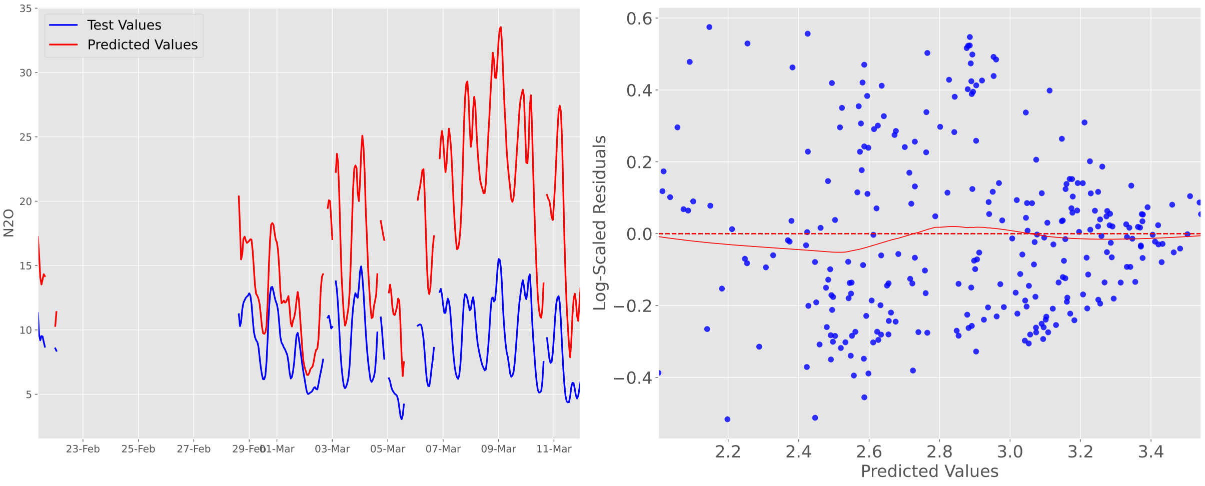
Source: The author.

RW Ridge Regression in Reduced Feature Set



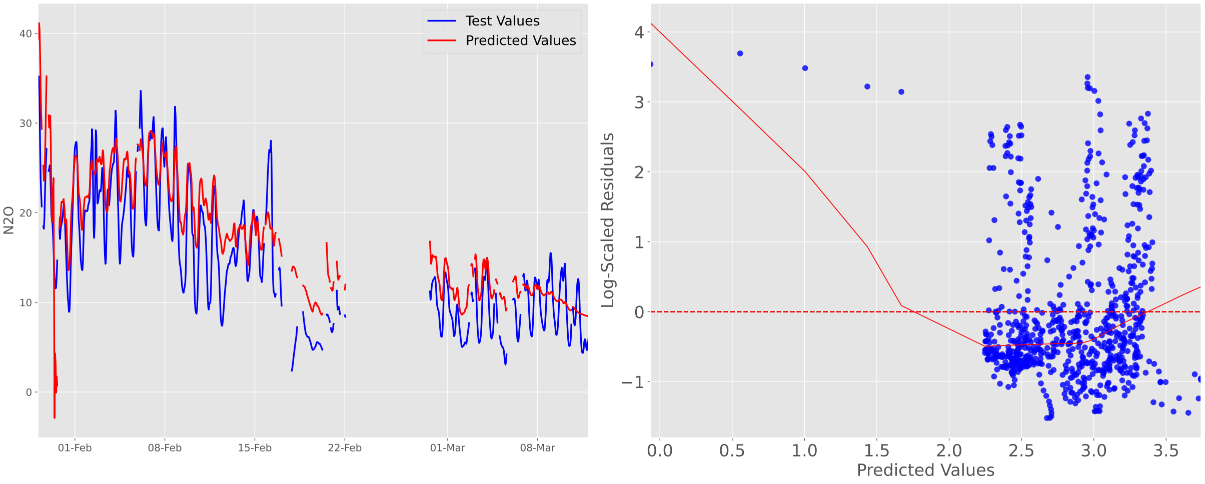
Source: The author.

Lasso Regression in Reduced Feature Set



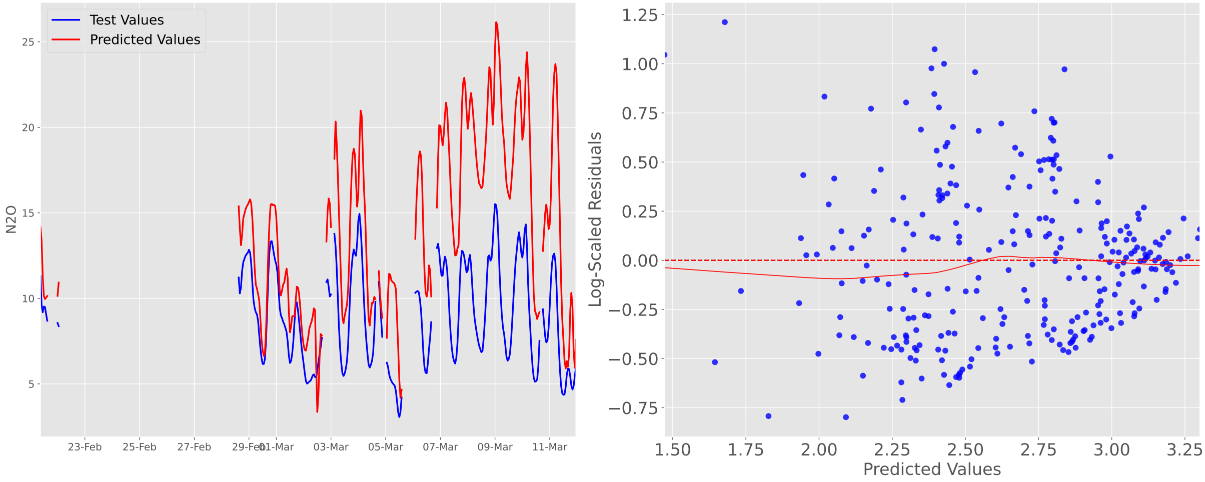
Source: The author.

RW Lasso Regression in Reduced Feature Set



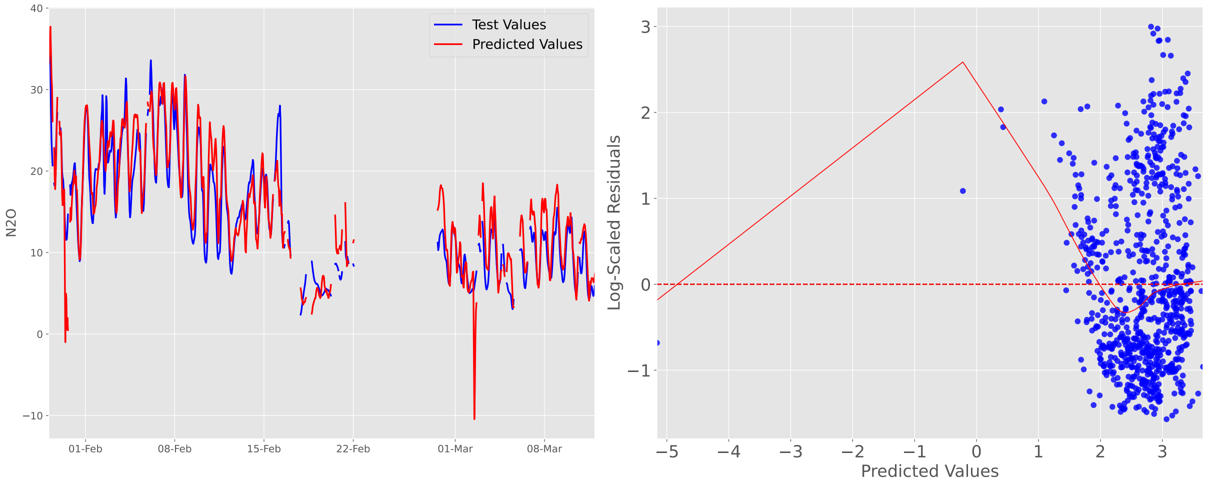
Source: The author.

Huber Regression in Reduced Feature Set



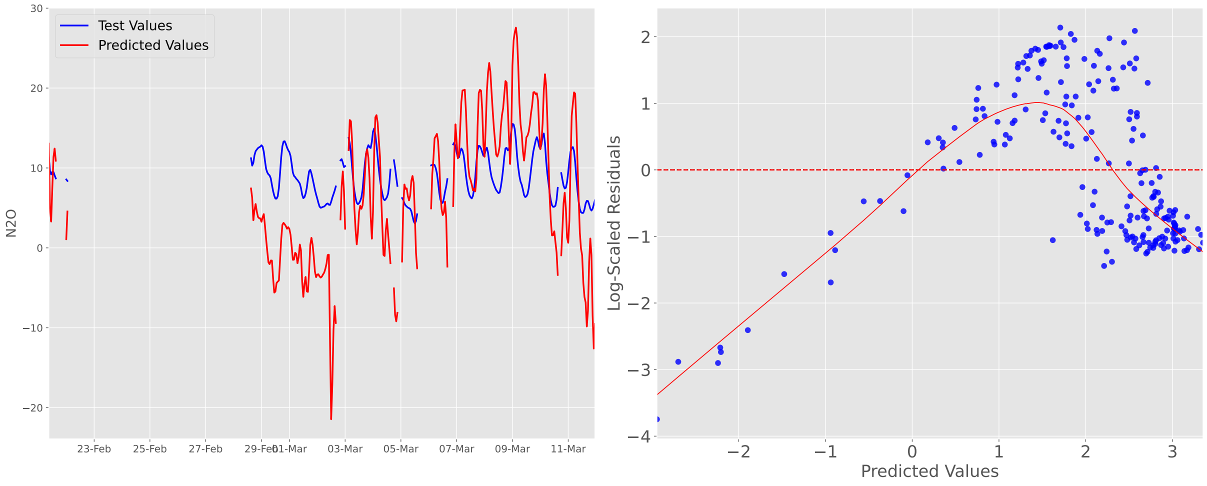
Source: The author.

RW Huber Regression in Reduced Feature Set



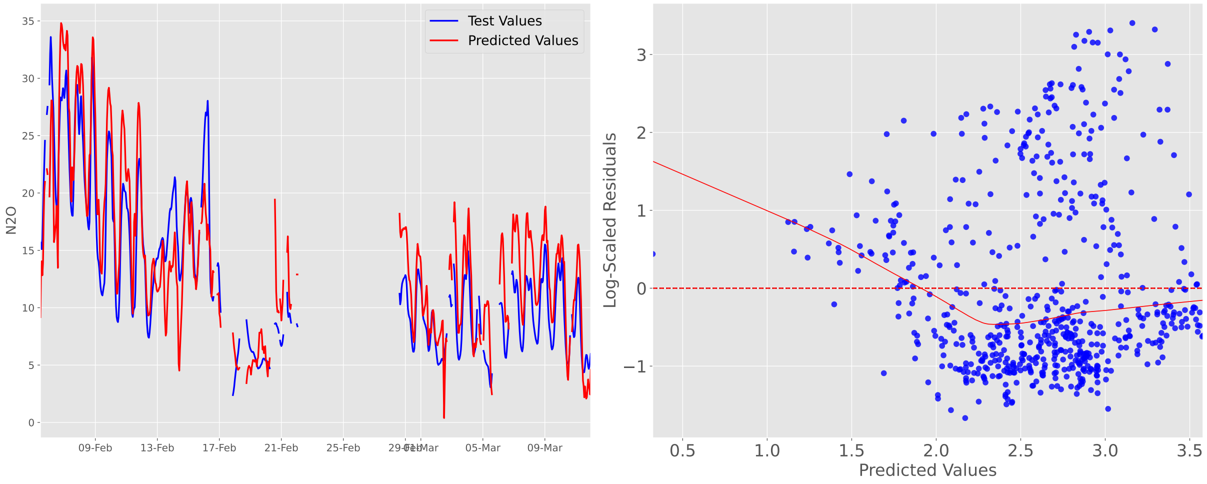
Source: The author.

Bayesian Regression in Reduced Feature Set



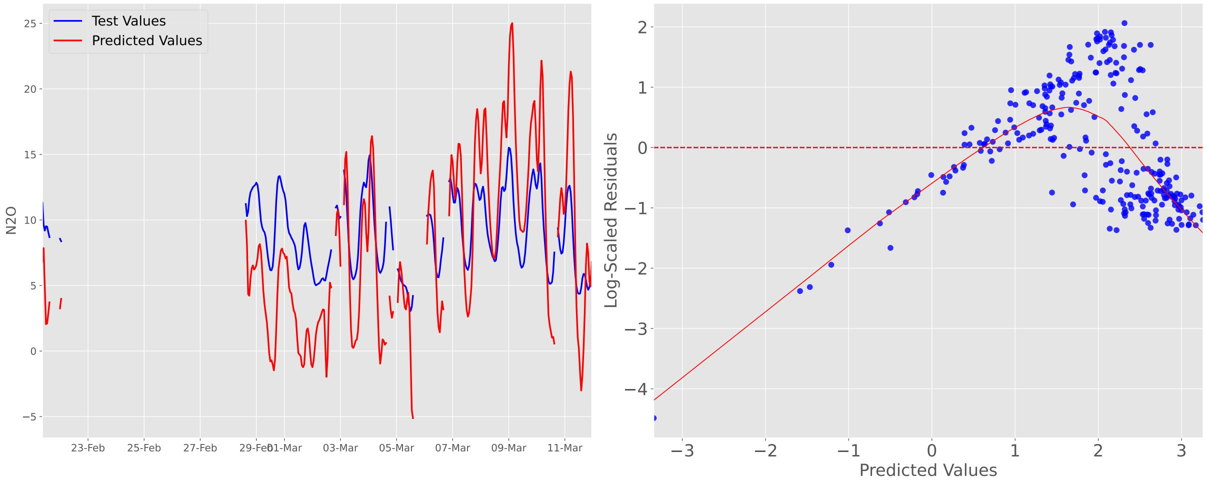
Source: The author.

RW Bayesian Regression in Reduced Feature Set



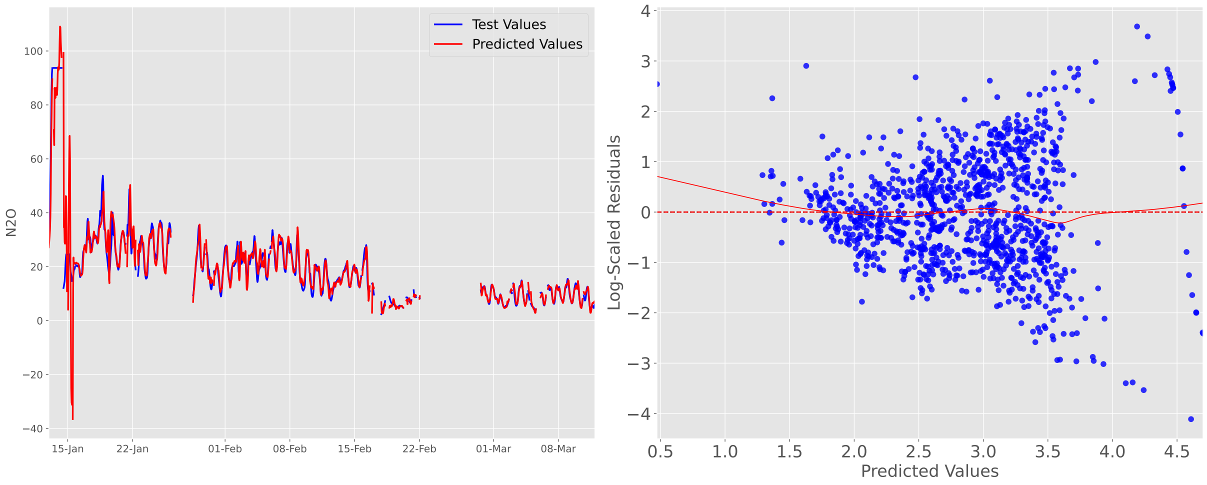
Source: The author.

PLS Regression in Reduced Feature Set



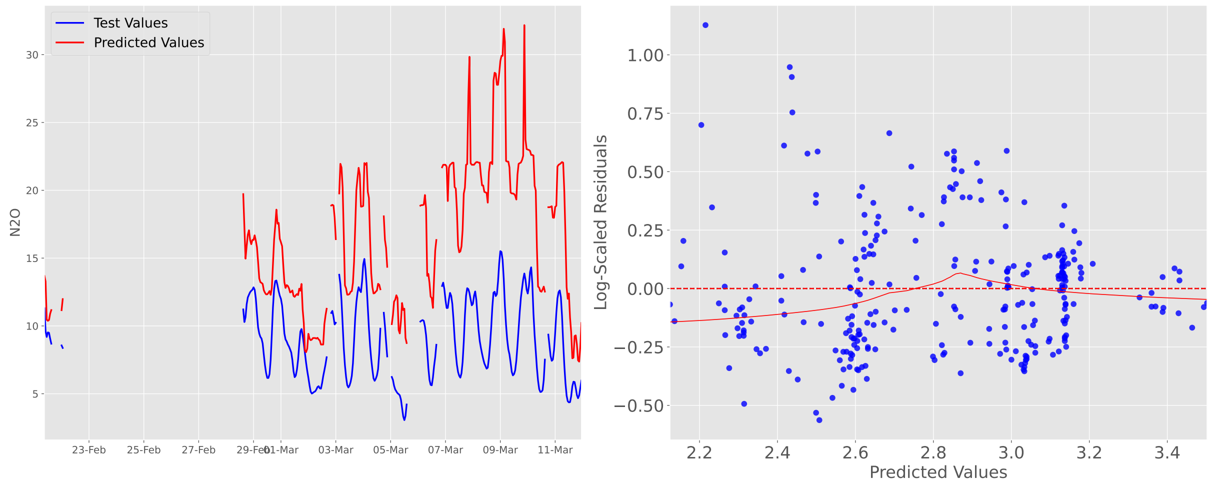
Source: The author.

RW PLS Regression in Reduced Feature Set



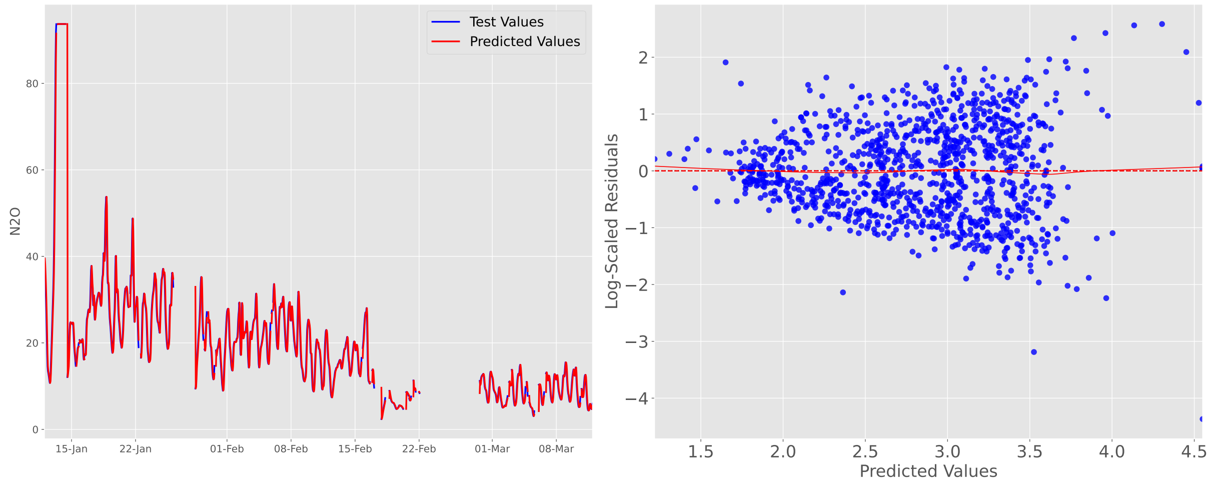
Source: The author.

Random Forest Regression in Reduced Feature Set



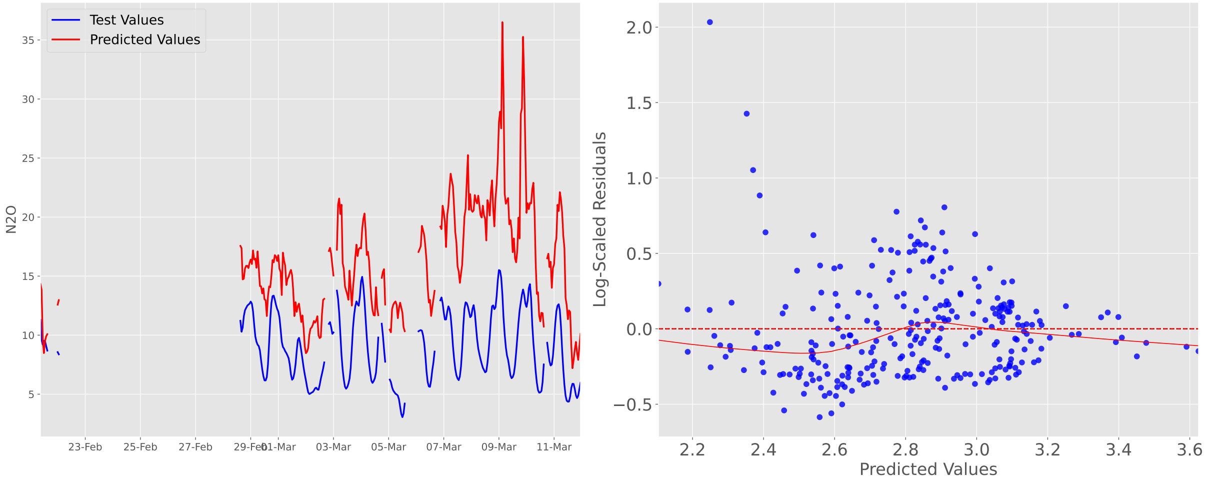
Source: The author.

RW Random Forest Regression in Reduced Feature Set



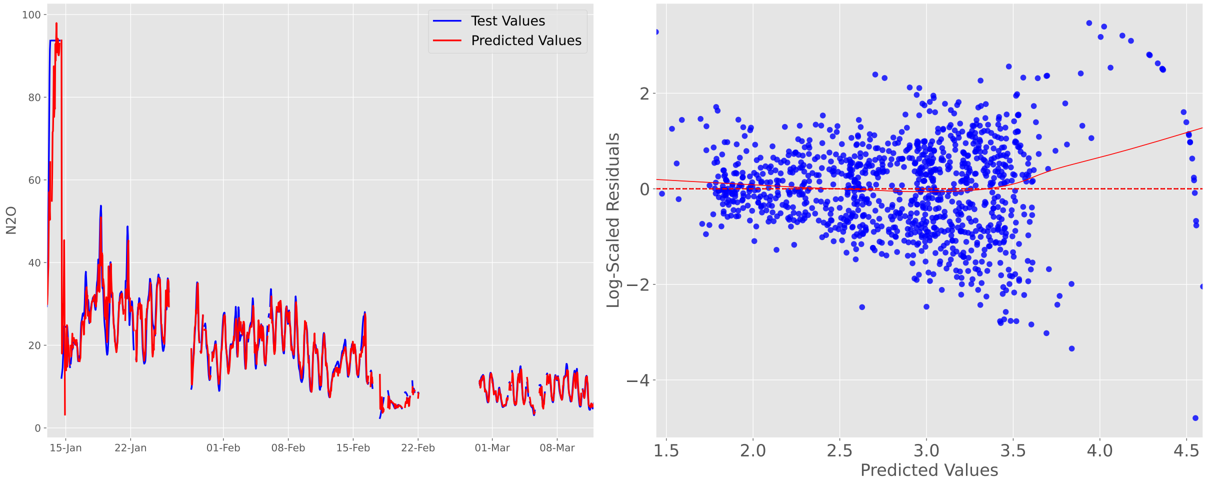
Source: The author.

XGBoost Regression in Reduced Feature Set



Source: The author.

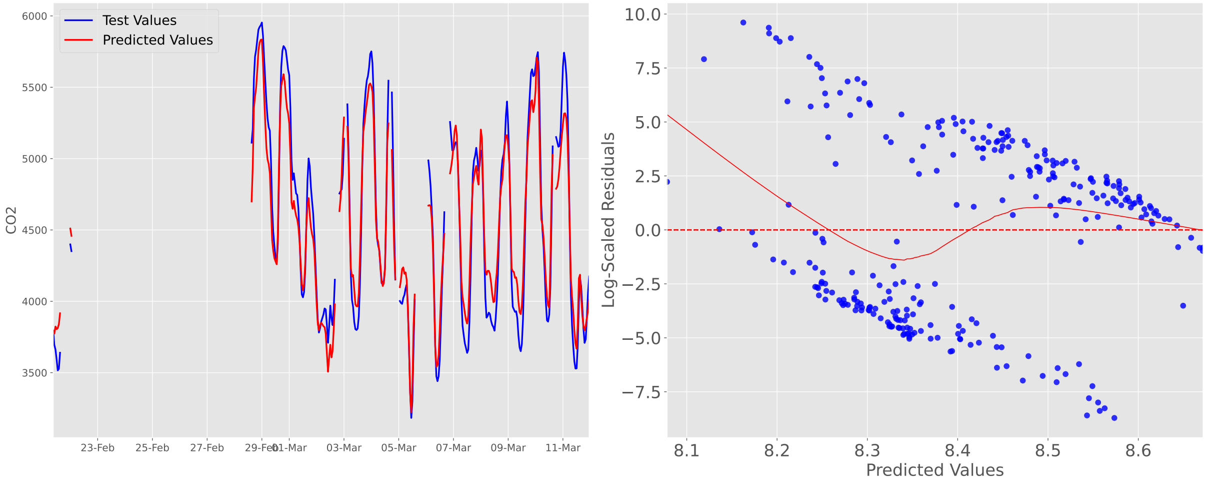
RW XGBoost Regression in Reduced Feature Set



Source: The author.

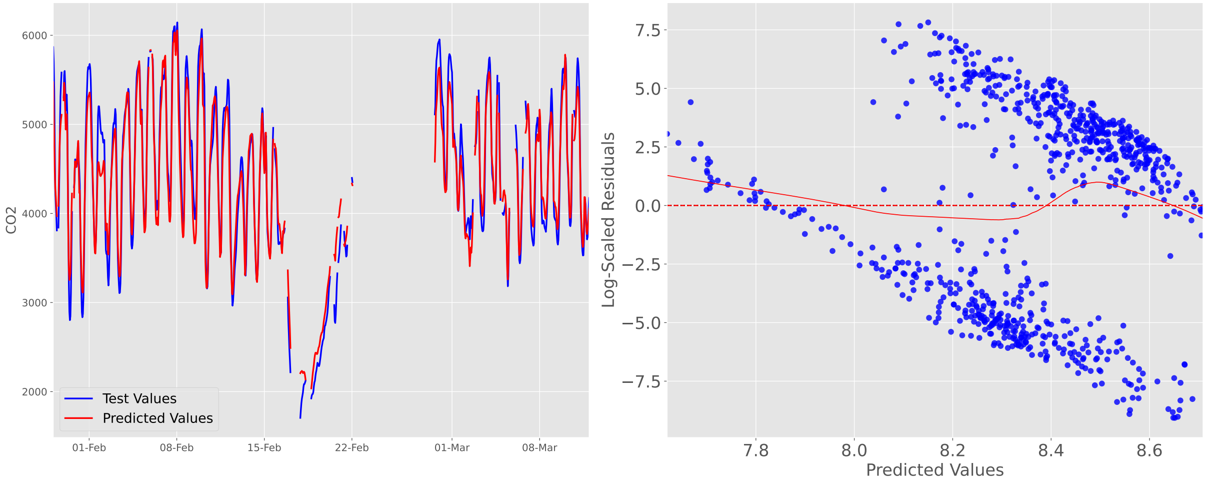
C.4 Reduced Feature Set CO_2 Regressions

Ridge Regression in Reduced Feature Set



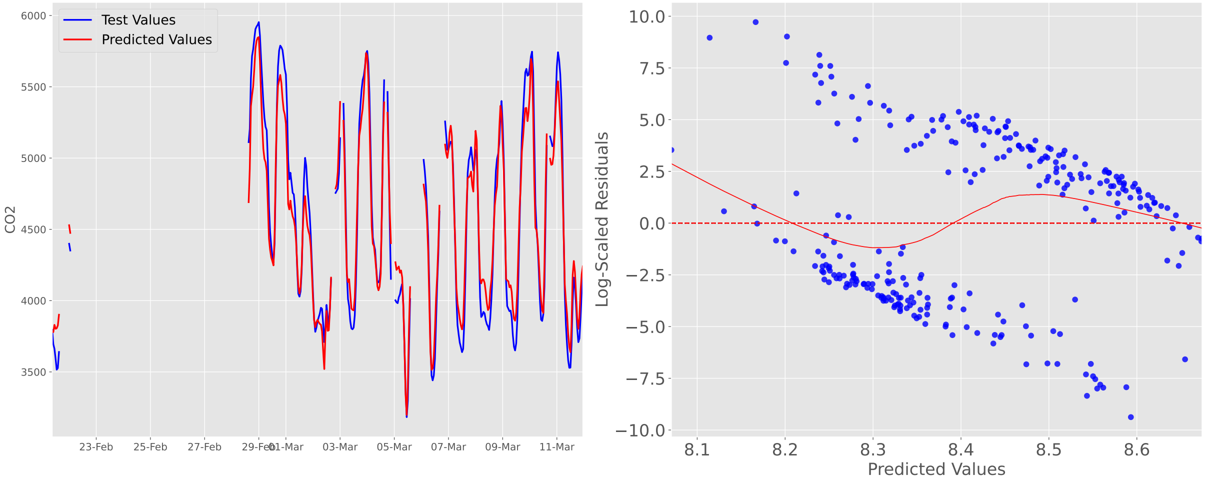
Source: The author.

RW Ridge Regression in Reduced Feature Set



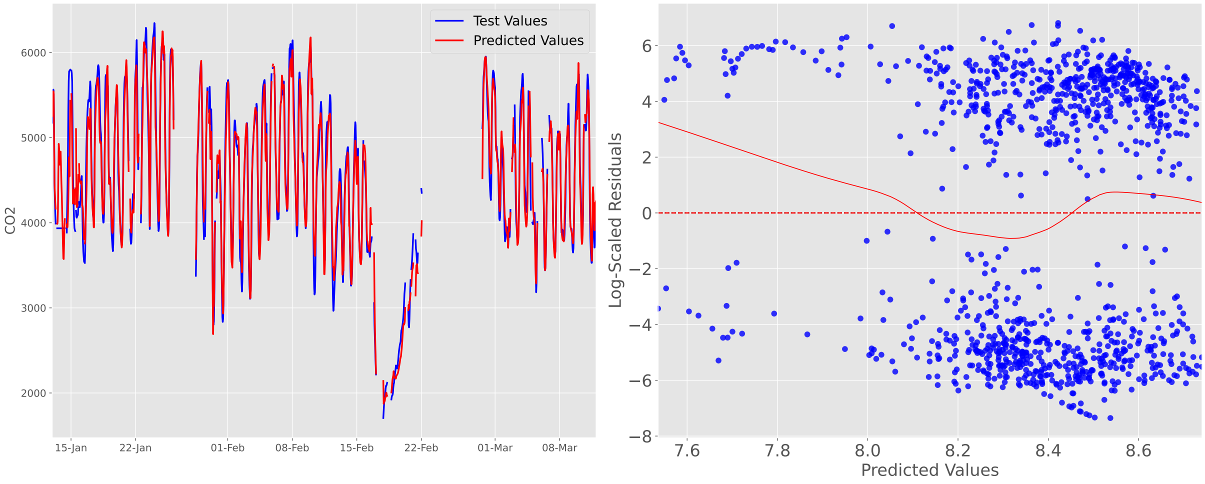
Source: The author.

Lasso Regression in Reduced Feature Set



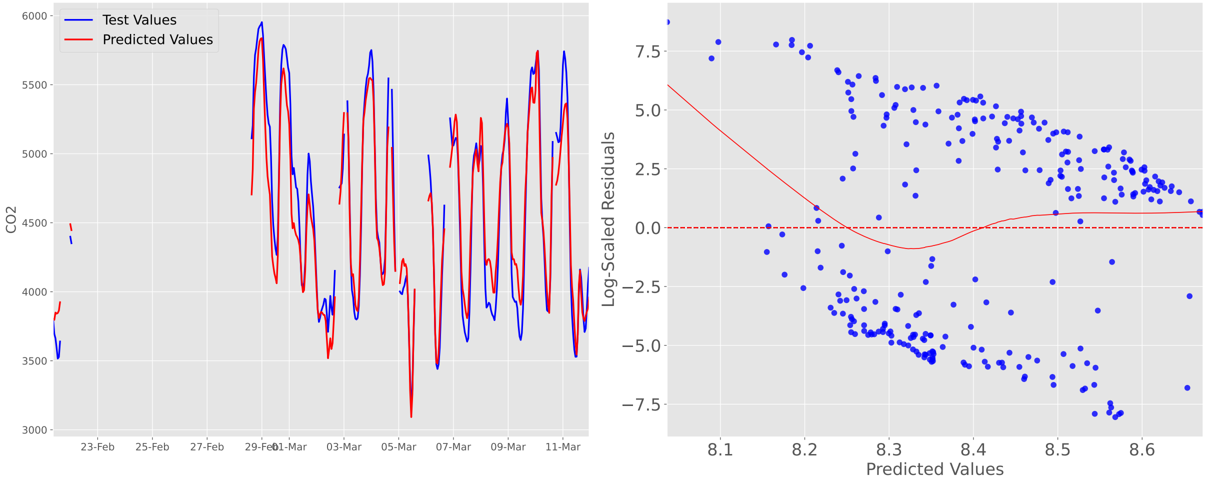
Source: The author.

RW Lasso Regression in Reduced Feature Set



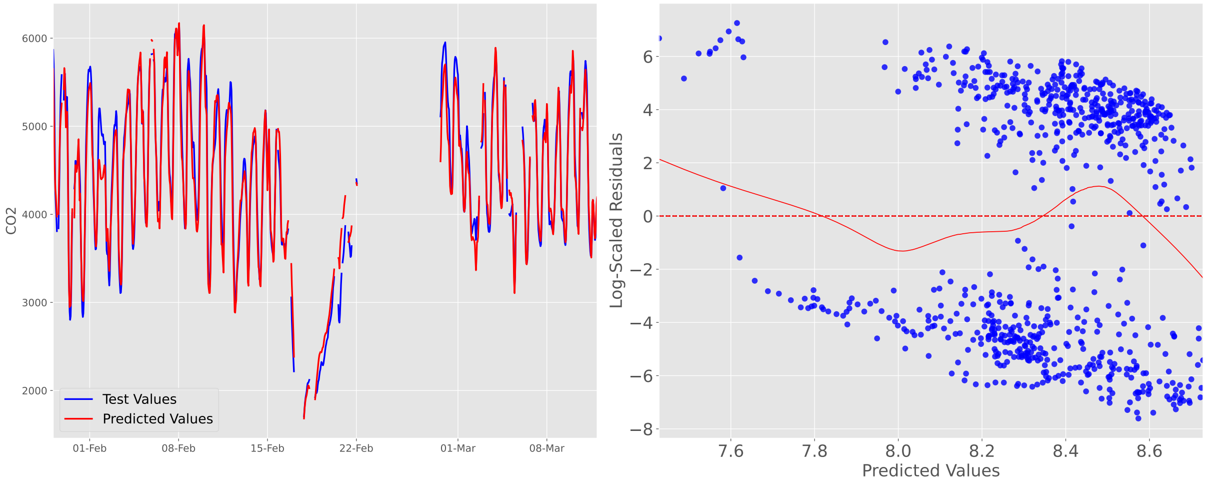
Source: The author.

Huber Regression in Reduced Feature Set



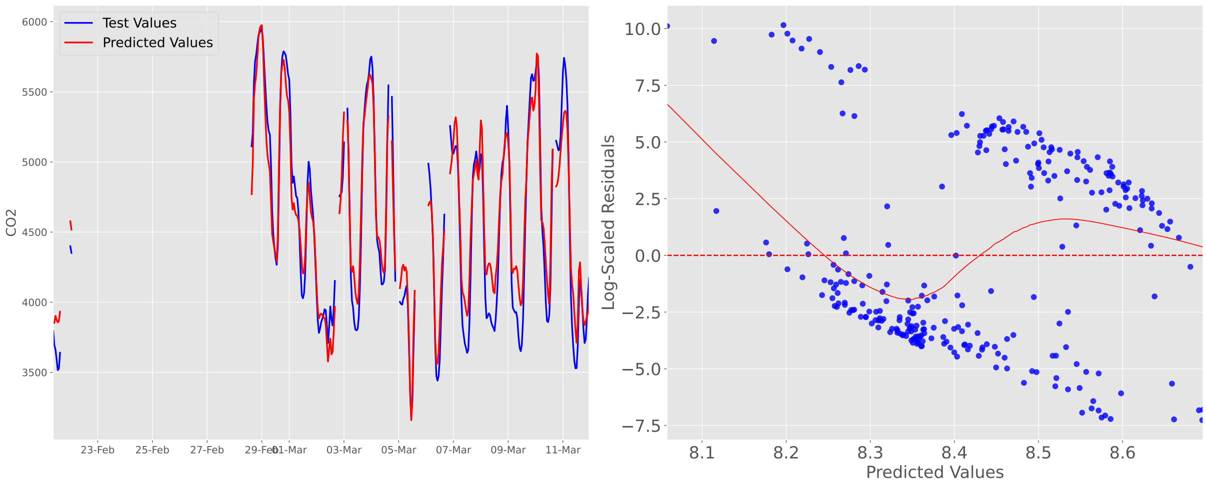
Source: The author.

RW Huber Regression in Reduced Feature Set



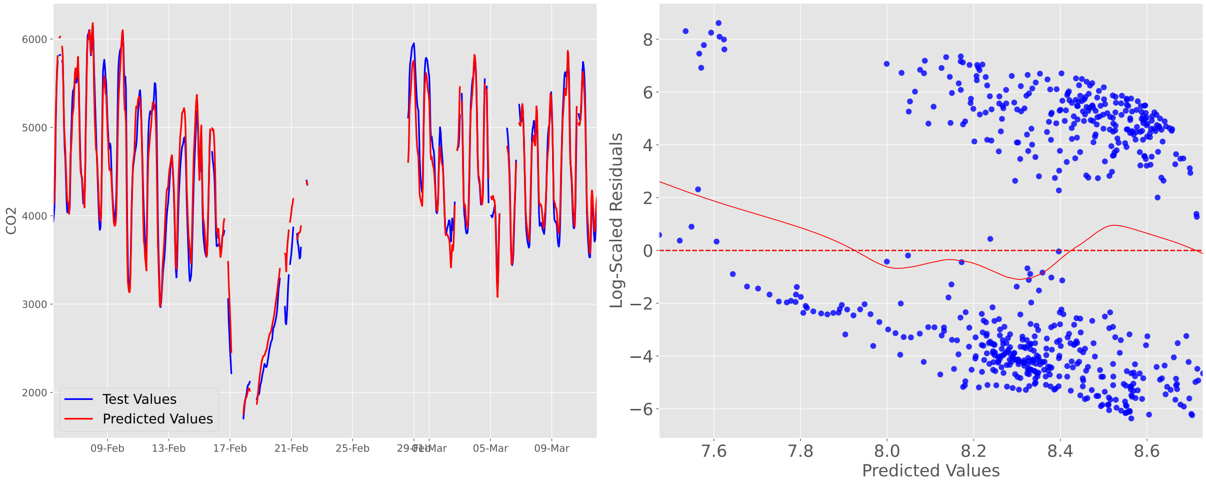
Source: The author.

Bayesian Regression in Reduced Feature Set



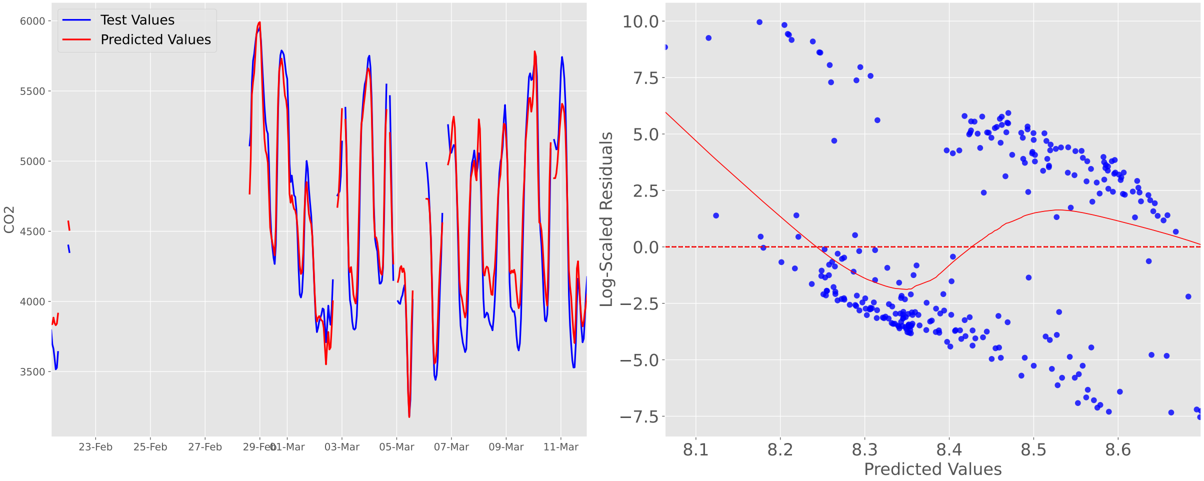
Source: The author.

RW Bayesian Regression in Reduced Feature Set



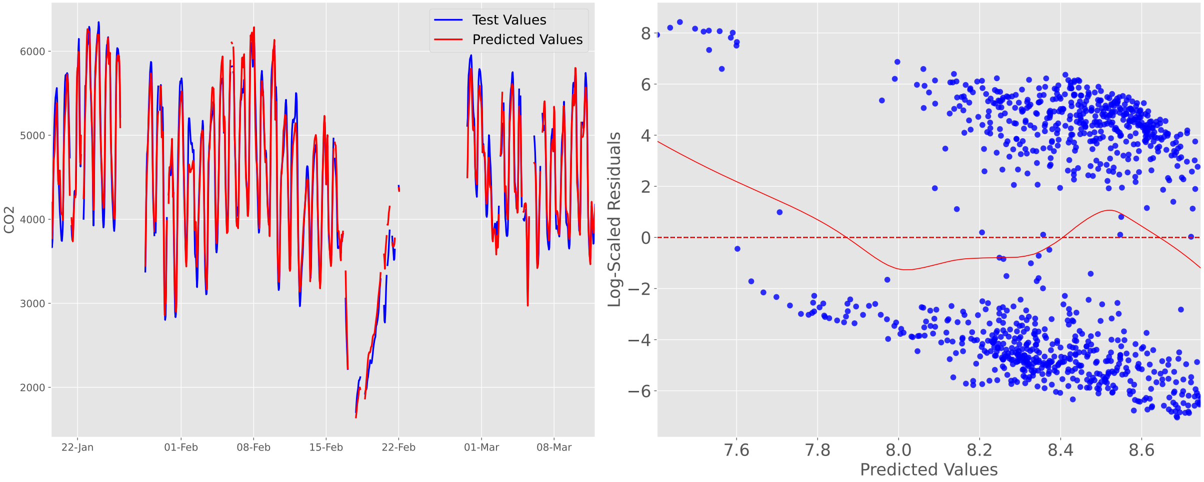
Source: The author.

PLS Regression in Reduced Feature Set



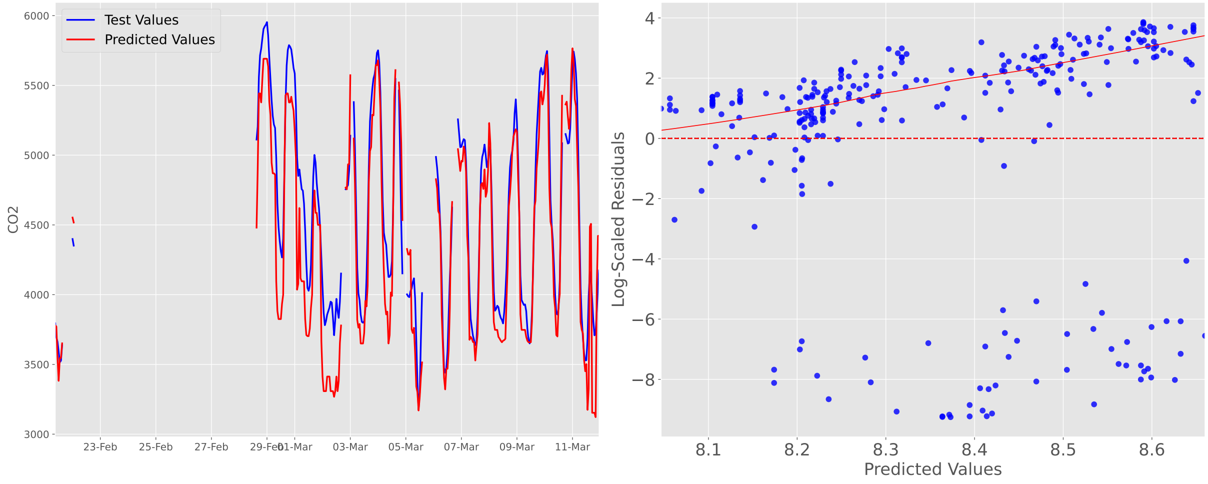
Source: The author.

RW PLS Regression in Reduced Feature Set



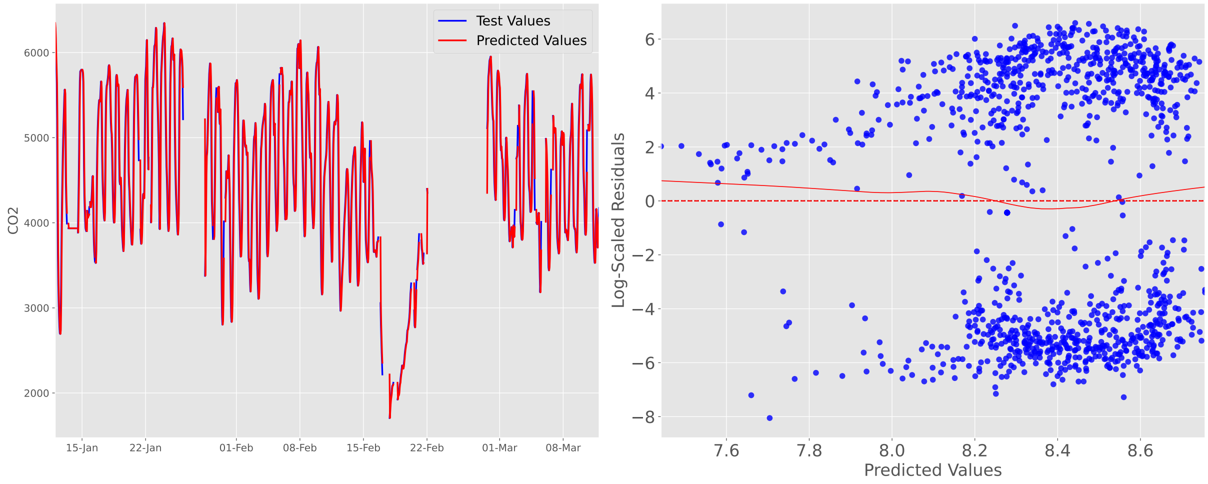
Source: The author.

Random Forest Regression in Reduced Feature Set



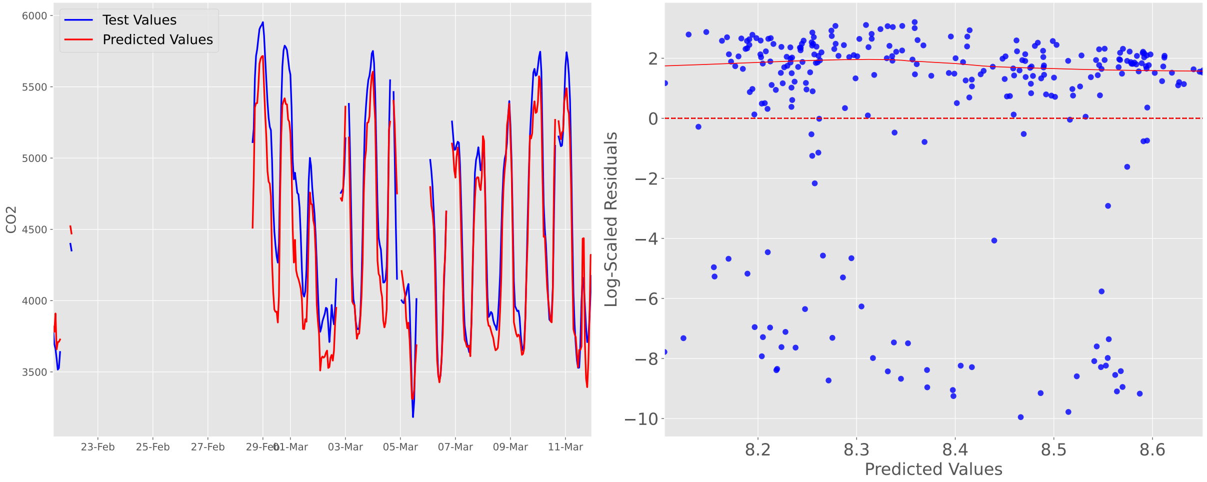
Source: The author.

RW Random Forest Regression in Reduced Feature Set



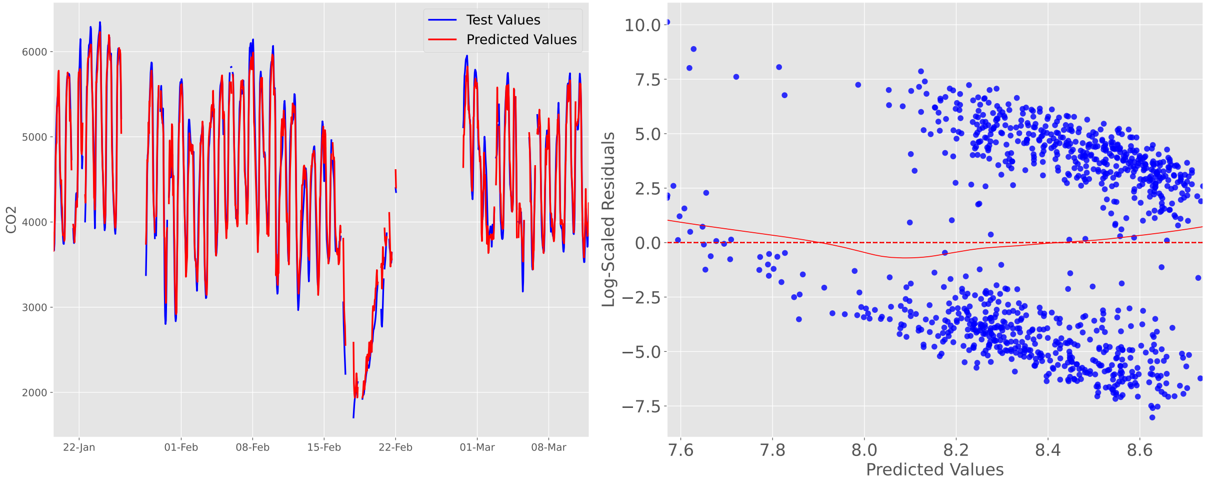
Source: The author.

XGBoost Regression in Reduced Feature Set



Source: The author.

XGBoost Regression in Reduced Feature Set



Source: The author.