water

MDPI

Article

Estimating Phosphorus and COD Concentrations Using a Hybrid Soft Sensor: A Case Study in a Norwegian Municipal Wastewater Treatment Plant

Abhilash Nair [1,*], Aleksander Hykkerud [1] and Harsha Ratnaweera [1,2]

[1] DOSCON AS, Østre Aker vei 19, 0581 Oslo, Norway; aleksander@doscon.no (A.H.); harsha.ratnaweera@nmbu.no (H.R.)
[2] Faculty of Science and Technology, Norwegian University of Life Sciences, 1432 Ås, Norway
* Correspondence: abhilash@doscon.no; Tel.: +47-48409183

Abstract: Online monitoring of wastewater quality parameters is vital for an efficient and stable operation of wastewater treatment plants (WWTP). Several WWTPs rely on daily/weekly analysis of water samples rather than online automated wet-analyzers due to their high capital and maintenance costs. Soft-sensors are emerging as a viable alternative for real-time monitoring of parameters that either lack a reliable measuring principle or are measured using expensive online sensors. This paper presents the development, implementation, and validation of a hybrid soft sensor used to estimate Total Phosphorus (TP) and Chemical Oxygen Demand (COD) in the influent and effluent streams of a full-scale WWTP. A systematic method for cleaning and processing sensor data, identifying statistically significant correlations, and developing a mathematical model, is discussed. A non-intrusive Industrial Internet of Things (IIoT) infrastructure for soft-sensor deployment and a web-based GUI for data visualization are also presented in this work. The values of TP and COD estimated by the soft sensor are validated by comparing the estimated values to the daily average of their corresponding lab measurements. The data validation results demonstrate the potential of soft sensors in providing real-time values of essential wastewater quality parameters with an acceptable degree of accuracy.

Keywords: hybrid soft-sensors; online monitoring; IIoT; digitalization

Citation: Nair, A.; Hykkerud, A.; Ratnaweera, H. Estimating Phosphorus and COD Concentrations Using a Hybrid Soft Sensor: A Case Study in a Norwegian Municipal Wastewater Treatment Plant. *Water* **2022**, *14*, 332. https://doi.org/10.3390/w14030332

Academic Editor: Dimitrios E. Alexakis

Received: 21 December 2021
Accepted: 22 January 2022
Published: 24 January 2022

1. Introduction

The chemical wastewater treatment process, which includes coagulation and flocculation followed by sedimentation, is one of the most commonly used wastewater treatment processes in Norway [1]. The operational efficiency of Wastewater Treatment Plants (WWTPs) based on chemical treatment is maintained by ensuring an optimal dosage of coagulants and flocculants. Several control strategies varying from simple flow proportional to sophisticated multi-parameter-based dosing control strategies [2] can be found in the literature. The growing number of users of the multi-parameter-based dosing control strategies provide impetus to monitor additional wastewater quality parameters in WWTPs.

WWTPs use several methods to monitor essential wastewater quality parameters. These methods vary from the offline analysis of water samples using standardized lab tests [3] to online sensors that can relay real-time data to the Supervisory Control And Data Acquisition (SCADA) system of the treatment plant. Standardized lab measurements provide information on the overall efficiencies of the treatment plant. However, the delay in obtaining the measurement results and the relatively large timescale (once every day or 2 days) associated with sample collection limits their utility in automation and process control algorithms or decision support systems. Adoption of new technologies such as ballasted flocculation and separation [4] or the potential use of nano-material [5] has further encouraged the use of automation, process control, and advanced monitoring

in WWTPs. A rapid growth in sensor technology in the past decade has resulted in several new online measuring techniques and a significant drop in the prices of online sensors. The current market provides sensors to monitor several parameters in wastewater that are otherwise measured offline using standardized lab tests. However, the high installation costs of these online sensors and the regular maintenance costs associated with the replacement of calibration solutions and reagents, cleaning of source lamps/lenses often make it economically infeasible, especially for small to medium-sized WWTPs [6]. A comprehensive review of online monitoring systems used in the coagulation/flocculation process presented in [7] shows that most wastewater quality parameters such as Total Phosphates (TP), Chemical Oxygen Demand (COD), and Total Nitrogen (TN) are measured offline using standardized lab analysis. However, parameters such as pH, Dissolved Oxygen (DO), conductivity, Oxidation-Reduction Potential (ORP), water level, and flowrate, that can be measured using reliable, inexpensive, and low maintenance sensors are often measured online in most treatment facilities including small to medium WWTPs [8].

Software sensors, also known as soft sensors or virtual software sensors are viable alternatives that are being actively explored [9]. Soft sensors are mathematical models implemented in software that can use real-time data from easy to measure physical sensors to estimate essential wastewater quality parameters that are either difficult to measure or are measured offline using lab analysis. A detailed review of different soft sensor algorithms used in WWTPs is presented in [10]. Several case studies presenting both simulator-based evaluations [11] and pilot-scale implementations [12] of soft sensors can be found in the literature. However, we found hardly any full-scale implementations of soft sensors especially in wastewater treatment processes using coagulation/flocculation and sedimentation.

The primary aim of this work is to design, test, and validate a hybrid soft sensor to estimate influent and effluent concentrations of TP and COD in a full-scale municipal wastewater treatment facility located in Norway. The real-time data from easy to measure online sensors installed in a full-scale WWTP are correlated to the data obtained from periodic lab-analysis of raw and treated water samples. A concise data-analysis tool for cleaning and transforming online data, comparison between various soft sensor models, state-of-the-art algorithm deployment strategy, and practical issues encountered during soft sensor deployment are investigated.

2. Materials and Methods

2.1. Søndre Follo Wastewater Treatment Plant (SFR)

SFR is a municipal wastewater treatment facility located in Vestby municipality, Norway. The treatment plant has a treatment capacity of 29,000 p.e and uses a conventional coagulation/flocculation process to remove colloids, particles, and soluble ortho-phosphates from municipal wastewater. The wastewater initially passes through a grit removal unit, removing larger particles, and enters a fat and sand removal unit. The coagulant is dosed at the entry point of the sand removal unit due to the high mixing rates achieved in this process. Flocculants are dosed to the wastewater at the entry point of the flocculation units, equipped with paddle mixers for slow mixing to ensure the formation of flocs. Wastewater with sufficiently developed flocs is distributed to six rectangular sedimentation tanks that provide adequate residence time for the flocs (sludge) to settle down in the bottom. The treated wastewater passes over weirs to the effluent channel. Sludge, which is removed from the bottom of the sedimentation unit is sent to anaerobic digestion after dewatering, thickening, and sludge stabilization.

The treatment plant is equipped with online pH, TSS, conductivity, ORP, level, and flow sensors that relay real-time information to a SCADA system provided by GUARD Automation (https://guard.no/) assessed on 15 December 2021. A multi-parameter-based dosing control system [13], provided by DOSCON, ensures an optimal dosing of coagulant and polymer in the treatment plant. The layout of the treatment process, the location of

online sensors, dosing points for coagulants and flocculants, and the flow distribution of wastewater along the WWTP are presented in Figure 1.

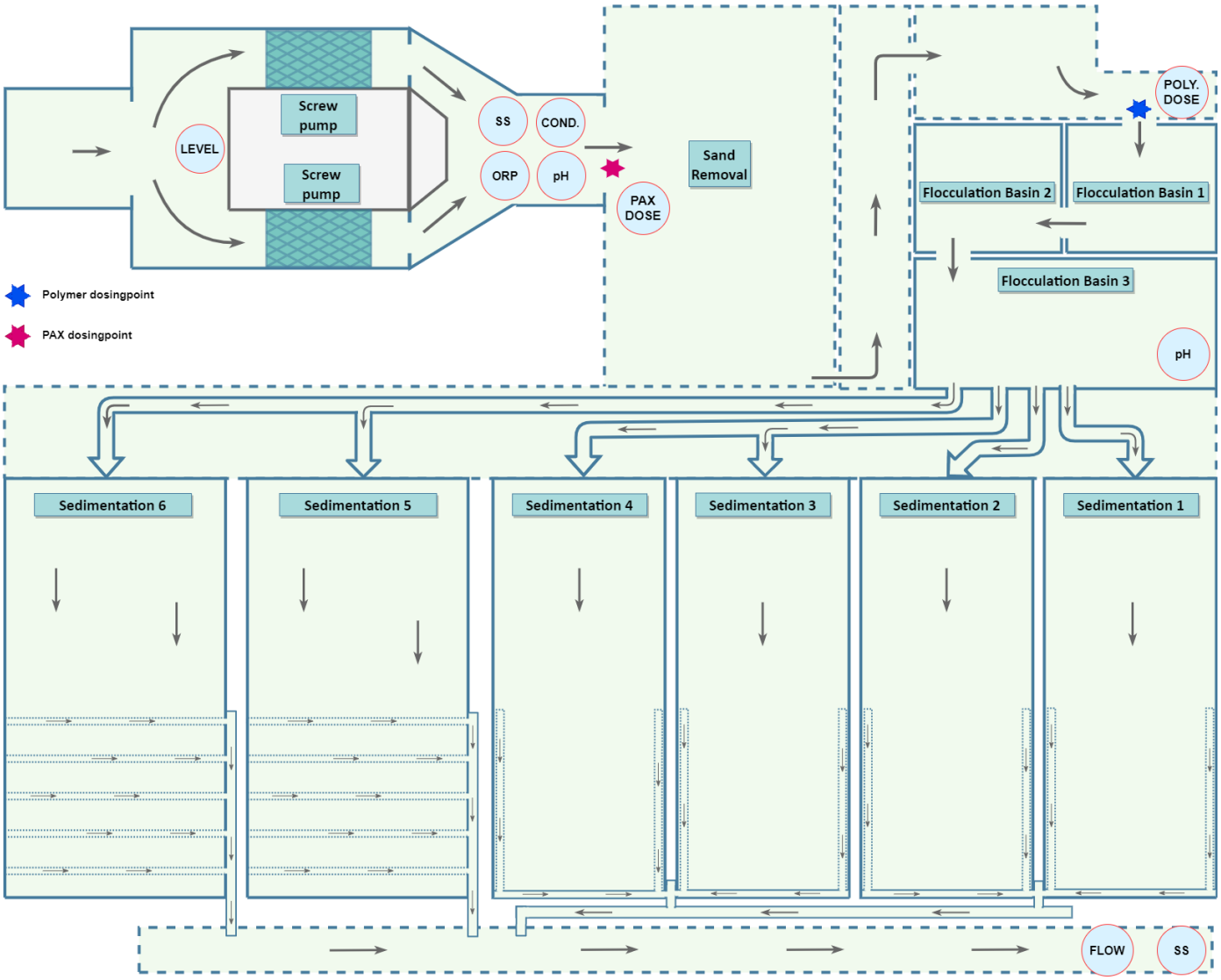


Figure 1. Plant layout and location of online sensors in SFR wastewater treatment plant.

In addition to the online sensors located in several sections of the WWTP, auto-samplers are installed at raw and treated wastewater sampling points to collect composite samples. The auto-samplers are programmed to collect 50 mL samples when the cumulative wastewater flow reaches 200 m^3. The composite samples are then analyzed for TP and COD using standardized methods as described in [3]. TP is analyzed as a daily average (from 8:00 a.m. of every working day to 7:00 a.m. of the next working day) and COD is analyzed once every week as weekly averages (on Fridays). The mean and standard deviation of the influent and effluent TPI and COD along with the removal rates are presented in Table 1. The distribution of data obtained from online sensors (pH, TSS, Flowrate, conductivity, ORP, PAX and Polymer dose flow-meters) and the values measured using standardized lab measurements (TP and COD) for the year 2020–2021 are presented in Figure 2a,b. The *x*-axis shows the range of the data used to calibrate the soft-sensor, and the *y*-axis shows the percentage of data corresponding to the value presented in the *x*-axis.

Table 1. Average values of influent, effluent, removal rates of TP and COD in SFR WWTP.

Parameter	Influent	Effluent	Removal (%)
TP	7.1 ± 2.4	0.49 ± 0.19	93 ± 3.2
COD	399 ± 184	80 ± 25	81 ± 7.9

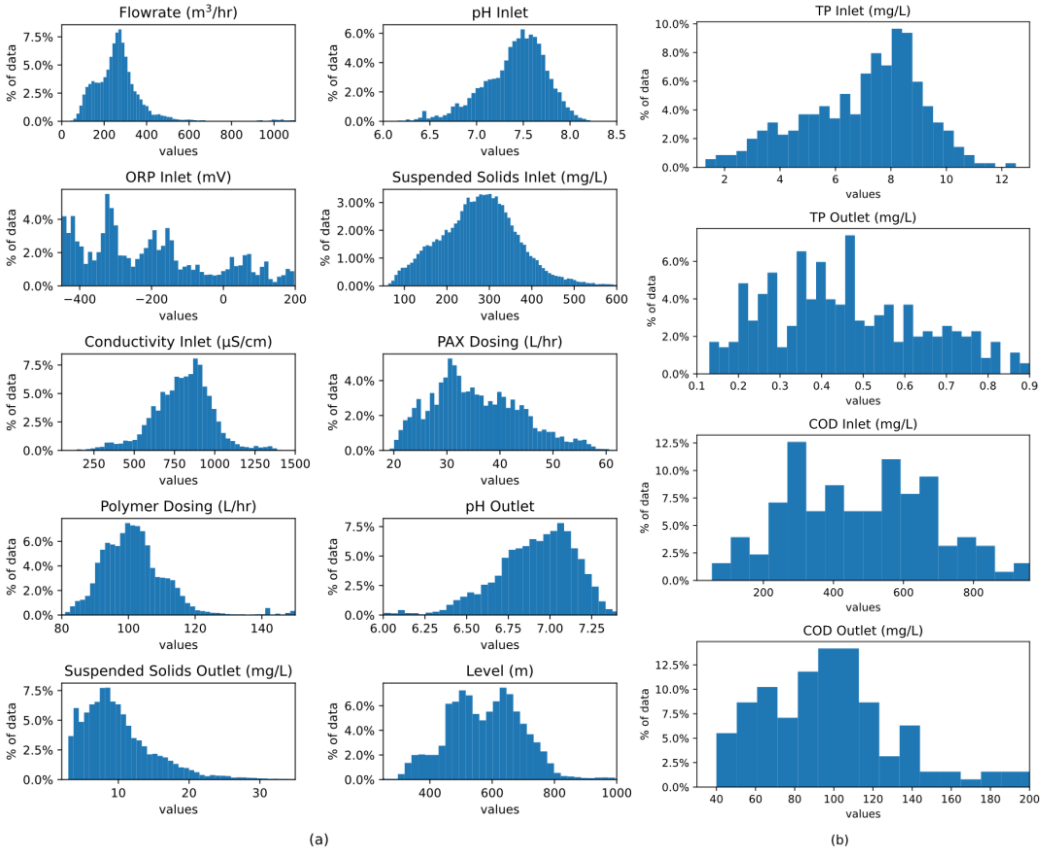


Figure 2. Distribution of data for the year 2020–2021 (**a**) from online sensors, (**b**) from lab measurements.

2.2. Mathematical Modelling

Mathematical models based on a mechanistic understanding of the coagulation-flocculation process can be found in the literature [14]. The increase in data availability caused by the recent adoption of online monitoring in most WWTP resulted in the increased use of data-driven models to describe the coagulation-flocculation process. The basic structure of these models varies from easily interpretable statistical models such as PCA/PCR to complex models such as Artificial Neural Network (ANN), Support Vector Machines (SVM) or Ensemble Tree (ET). Multiple Linear Regression (MLR) is a commonly used data-driven modelling technique used to predict output variables using multiple independent predictors. Instances of MLR implementation to predict effluent wastewater quality are found in the literature [15]. The mathematical representation of the MLR model is presented in Equations (1)–(3).

$$\hat{Y} = f(x_1, x_2, .. x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \quad (1)$$

$$\hat{Y} = \beta_0 + \sum_{i=0}^{n} \beta_i x_1 + \sum_{i=0}^{n} \beta_{ii} {x_i}^2 + \sum_{i=0}^{n} \sum_{j=i}^{n} \beta_{ij} x_i x_j + \epsilon \quad (2)$$

$$\min_{\beta} \| Y - \hat{Y} \|_2^2 \quad (3)$$

where $\hat{Y}$ is the predicted responses corresponding to the measured values Y, x_i are the predictors, β are the model coefficients, and ϵ is model error. Non-linear dependencies between the response and the predictors are introduced in the model by including the square and interaction terms of x_i as described in Equation (2). Several algorithms ranging from simple ordinary least square algorithm presented in Equation (3) to more complex stochastic algorithms [16] can be used to minimize the error between Y and $\hat{Y}$ and obtain the optimal set of the regression coefficients (β).

2.3. Workflow of Soft Sensor Development

The steps involved in the processing of online sensor data and development of mathematical models for estimating TP and COD in the influent and effluent are presented in Figure 3.

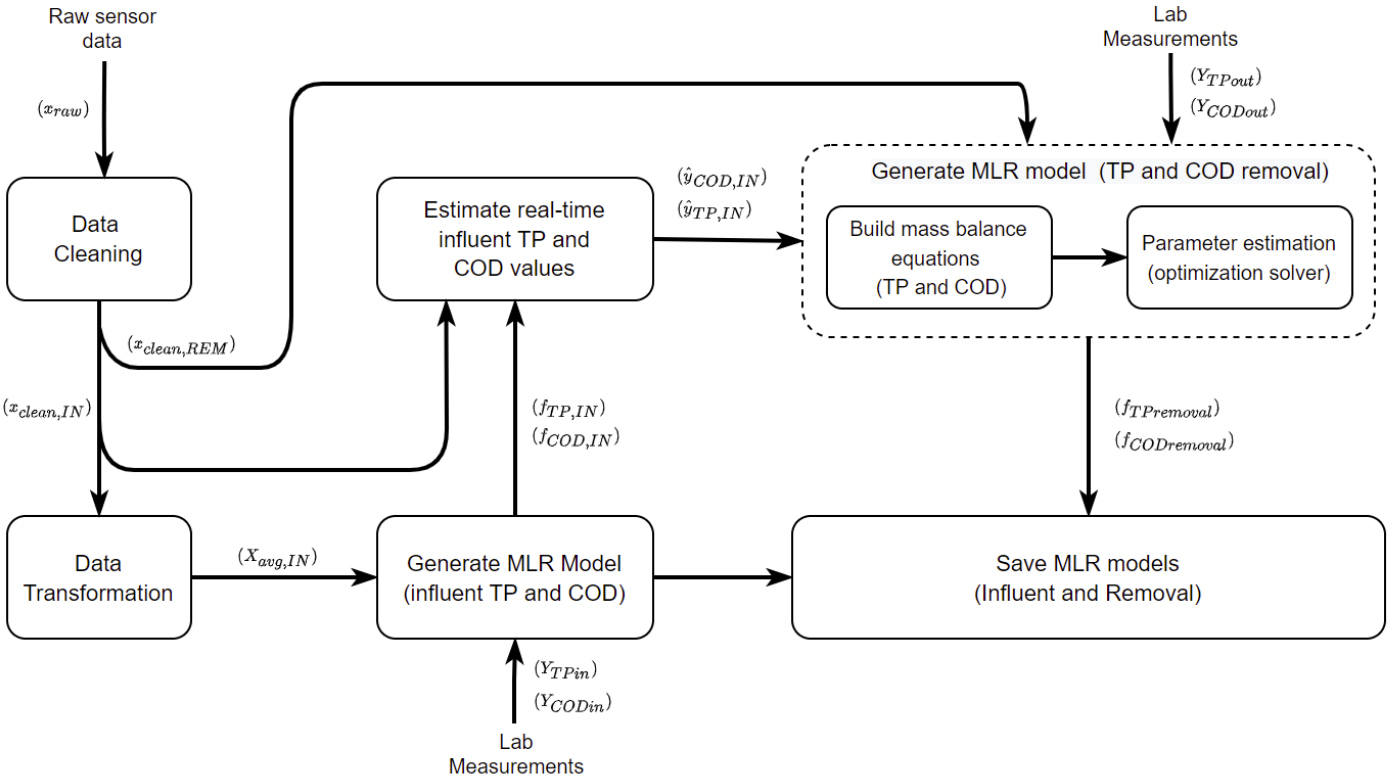


Figure 3. Systematic workflow and stages involved in the development of a soft sensor for estimating TP and COD in influent and effluent.

2.3.1. Step 1—Data Cleaning and Transformation

Data rich does not imply information rich [17]. In this context using noisy data ridden with outliers would result in inaccurate mathematical models, incapable of generating meaningful predictions. Therefore, data cleaning and outlier removal is the first and foremost stage before using online sensor data for model generation. Several outlier detection methods tailored to detect outliers in wastewater treatment processes can be found in the literature [18,19]. The moving window approach described in [19] was used to clean the data before using it for model calibration.

2.3.2. Step 2—Generate Influent TP and COD Model

The sensor layout of WWTP (Figure 1) shows that the flowrate (x_{Q}), level in the influent channel (x_{LVL}), suspended solids (x_{SSI}), conductivity (x_{CNI}), ORP (x_{ORP}), and pH (x_{PHI}) of the raw wastewater are monitored online. From a mechanistic point of view, the particulate fraction of TP and COD varies proportionally to the TSS in the wastewater. Inverse dependencies between flowrate/level and TP/COD can be observed by

comparing the average raw wastewater flowrate with the lab-measured values of TP and COD. The inverse correlation can be explained by the fact that a sudden increase in raw wastewater flowrate is often caused by rainfall or snowmelt events that can dilute wastewater and reduce TP/COD concentrations. Several prior works found in the literature indicate a correlation between variations in conductivity [20], pH, ORP [21], and the soluble fractions of TP/COD. A linear combination of predictors influencing the soluble and particulate components of TP and COD would be an ideal choice of predictors for the model. Therefore, the influent raw wastewater quality parameters $x_{clean,\,\mathrm{IN}} = \begin{bmatrix} x_{\mathrm{LVL}} & x_{\mathrm{Q}} & x_{\mathrm{CIN}} & x_{\mathrm{SSI}} & x_{\mathrm{PHI}} & x_{\mathrm{ORP}} \end{bmatrix}$ would be the best choice of independent predictors that can be used to correlate influent TP and COD values. An MLR model was developed for prediction COD and TP. The MLR model coefficients were generated using several model calibration algorithms as mentioned in [16]. The algorithm that showed the best fit between the lab-measured data was selected, and the functions were saved to be used in the next steps.

2.3.3. Step 3—Build Mass Balance Equations for TP and COD

A dynamic model was developed by conducting a material balance of TP and COD in the coagulation/sedimentation process. A generic form of the material balance equation used in modelling the effluent parameters is presented in Equation (4).

$$\frac{d\hat{y}_{\mathrm{OUT}}}{dt} = \frac{x_{\mathrm{Q}}}{V}(\hat{y}_{\mathrm{IN}} - \hat{y}_{\mathrm{OUT}}) - r\,\hat{y}^{k}_{\mathrm{OUT}} \tag{4}$$

In Equation (4), $\hat{y}_{\mathrm{OUT}}$ is the effluent COD/TP values and $\hat{y}_{\mathrm{IN}}$ is the influent TP/COD values. An n^{th} order kinetics is used to model the TP/COD removal rates in the process. The rate constant r and the reaction order n can be determined by fitting the effluent wastewater quality data to the dynamic model. The holdup volume V for WWTP is calculated by conducting tracer tests and at different flowrates.

2.3.4. Step 4—Generate TP and COD Removal Models

Dosage of coagulant (x_{PAX}) and dosage of polymer (x_{POL}) are important control variables, used to adjust the removal efficiencies of solids and phosphates in WWTPs [13]. A positive correlation between the dosage of coagulant/flocculant and the removal percentages, substantiated with systematically designed jar tests, can be found in the literature [22]. The results of these jar tests are often used as a basis to determine the optimal dosage of coagulants and polymer. Most industrial coagulants have an optimal pH range beyond which a substantial reduction in removal percentages can be observed. Treatment plant operators also tune dosing control algorithms to ensure that the operating pH range is not crossed. Therefore, $x_{clean,\,\mathrm{REM}} = \begin{bmatrix} x_{\mathrm{PAX}} & x_{\mathrm{POL}} & x_{\mathrm{PHO}} \end{bmatrix}$ is chosen as predictor for the mathematical functions defining the removal rates r. The batch model calibration approach, involving minimization of a quadratic error function between estimated and lab-measured value, is a commonly used technique for estimating parameters of dynamic models built using mass balance and reaction kinetics [23,24]. The parameter estimation method explained in [15] was used as a basis to construct the dynamic optimization problem.

2.3.5. Step 5—Save MLR Models/Coefficients

The coefficients of the MLR model developed for both the influent and for removal rates are saved in a structured form which can be later used for real time deployment.

2.4. Software Packages

The open-source programming language Python (www.python.org) accessed on 18 December 2021 was used to process the raw sensor data, generate mathematical models, and deploy algorithms for real-time estimation. The open-source BSD-licensed library, pandas, version 1.3 (https://pandas.pydata.org/) accessed on 18 December 2021, was used to clean and condition raw data and generate concurrent datasets of the same timestamp.

Scikit-learn, a free Python library (https://scikit-learn.org/) accessed on 18 December 2021 provides several algorithms to train MLR models and obtain the regression coefficients β described in Equation (1). The dynamic optimization problem (mentioned in step 4) was implemented in Python and solved using the optimization algorithms provided as a part of Python's open-source SciPy library (https://scipy.org/) accessed on 18 December 2021. The '*scipy.optimize*' package provides a multitude of algorithms that can be used for solving unconstrained optimization problems. In this work, four different optimization solving algorithms, a. Nelder–Mead (NM) [25], b. Trust-region Newton conjugate gradient (TR) [26], c. Sequential Least Squares Quadratic Programming (SLSQP) [27], and d. Broyden–Fletcher–Goldfarb–Shanno (BFGS) [28] were used separately, and the results were assessed based on the regression fit and the time required by the optimization solver to converge to an optimal solution. A generic version of the script that can be used to calibrate the model can be downloaded from the github repository (https://github.com/abhilash2134/MLR-Model.git) accessed on 21 January 2022.

2.5. Soft Sensor Deployment

A multitude of options for the deployment of soft sensor scripts and control algorithms are available in the market today. The deployment methods vary from direct implementation of scripts in a PLC/microcontroller/PC (edge computing) [29], or remote nonintrusive implementation presented in [30] using either cloud services or own infrastructure. In this paper, a more secure and robust version of the non-intrusive soft sensor deployment strategy discussed in [30] was used as a basis to build the IIoT infrastructure of DOSMON. An overview of the IIoT architecture of DOSMON used to acquire telemetry data from field devices and deploy the soft sensor algorithms in real-time is presented in Figure 4.

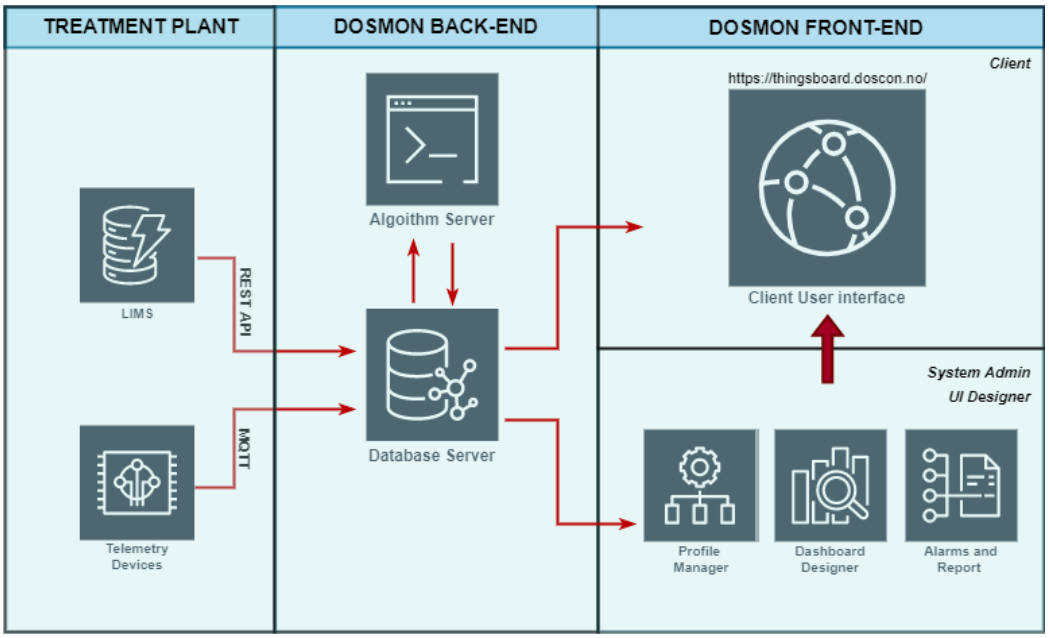


Figure 4. Soft sensor deployment strategy in DOSMON.

The backend of DOSMON core mainly consists of two servers. The primary server is used for the acquisition and storage of data from the IoT devices and Lab Information Management System (LIMS) software (database server), while the secondary server is used for running the soft sensor algorithms (algorithm server). The communication and data exchange between algorithm and data server is established using DOSMON's REST API. The algorithm server pulls the raw sensor telemetry from the data server and pushes algorithm results back to the data server. The software architecture presented in Figure 4 allows us to have reliable and robust IoT communication and dedicated computing power

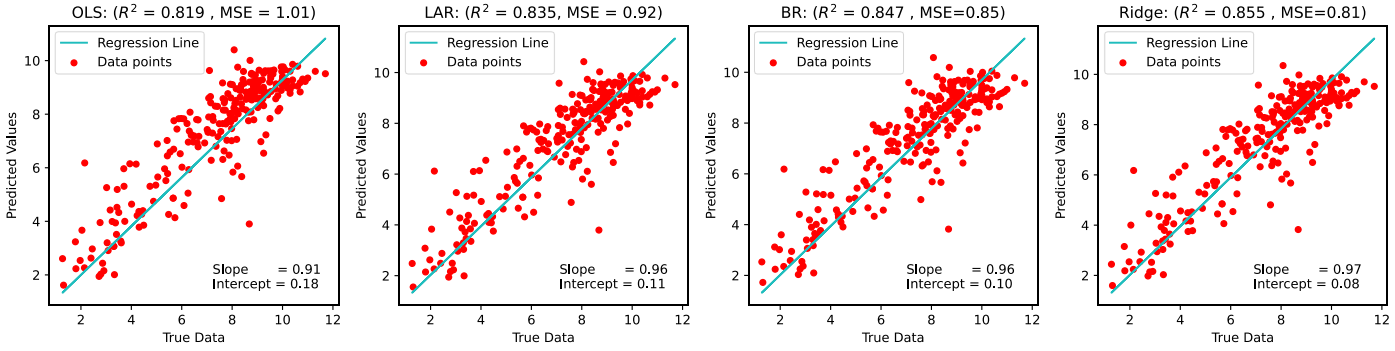
for executing soft sensor algorithms. The IIoT architecture also allows us to manage changes in the algorithms and tune model parameters a minimal downtime at the plan.

The MLR models developed using the methods described in Section 2.3 are stored using the 'pickle' functionality of Python. The pickled models simplify the model parameter storage as we do not have to create our own data storage formats. These models can be created in any Python environment, both offline and online, and can be loaded in the deployment stage.

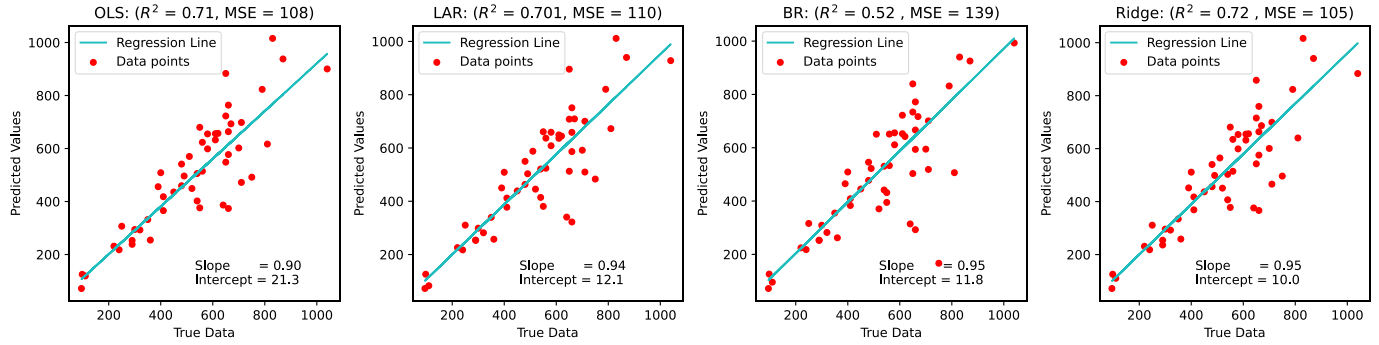
3. Results and Discussion

3.1. Model Calibration Results

Real-time data from the online sensors were obtained from the treatment plant's SCADA system and lab-measured values of TP and COD were obtained from LIMS software. Online and lab-measured data for a period of 12 months (from 10 April 2020 to 10 April 2021) are used to calibrate the model and obtain the MLR model coefficients. The results of model calibration showing a comparison between the model predicted and lab-measured values are presented in Figure 5. The plots also show the mean-square error (MSE) and the Pearson correlation coefficient (R^2) for the models calibrated using the four different calibration algorithms mentioned in Section 2.4. The slope and intercept of the regression line are also presented in the Figure 2 plots. The scatter plot of a perfect prediction model would be a 45-degree line with a slope = 1 and intercept = 0 and an R^2 value of 1. A quantitative assessment of different prediction models is conducted by comparing the values of R^2 and regression line equations.



(**a**)



(**b**)

Figure 5. *Cont.*

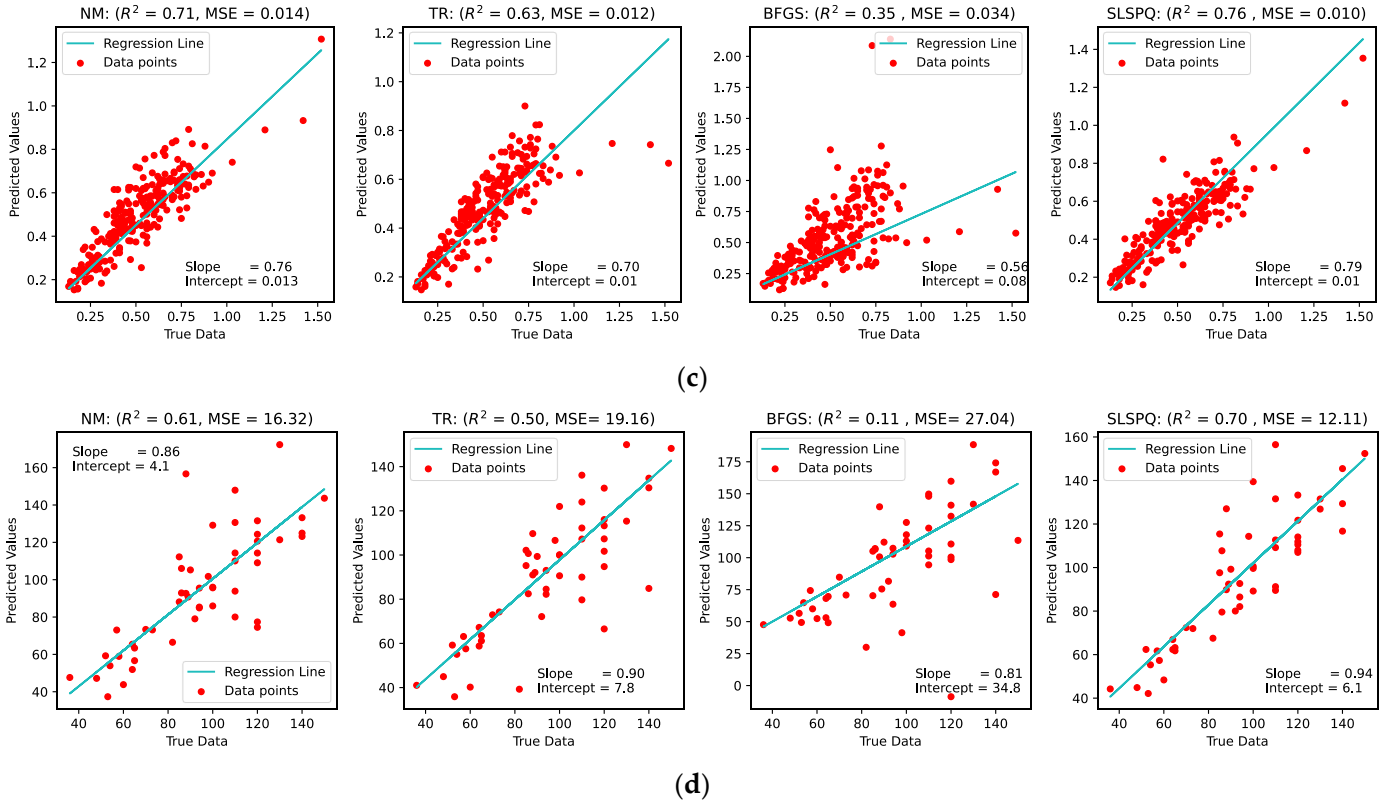


Figure 5. Lab-measured versus predicted values using different model calibration algorithms for (**a**) influent TP, (**b**) influent COD, (**c**) effluent TP, (**d**) effluent COD.

The plots presented in Figure 5a as well as a comparison between correlation coefficients (R^2 and MSE) show minimal difference between the results obtained with all four algorithms. However, the model coefficient obtained by the Ridge algorithm shows relatively better results with (R^2 =0.86 and MSE = 0.81) as compared to an R^2 =0.82 for OLS algorithm, R^2 =0.83 for LAR, and R^2 =0.84 for BR algorithm. Similar results are observed in Figure 5b which shows a comparison between predicted and lab measured values of different algorithms used to obtain MLR model coefficients for influent COD. The Ridge algorithm showed an R^2 of 0.72 compared to a value of 0.70 for LAR, 0.52 for BR, and 0.71 for OLS. The regression line for the Ridge algorithm (for influent TP) showed a slope of 0.97 and an intercept of 0.08 which ensures a random distribution of error along the ideal prediction line. The slope and intercept value close to the ideal prediction line of y = x also shows that MLR model coefficients obtained using Ridge algorithm has a comparatively better model fit. Therefore, the model coefficients obtained by the Ridge algorithm is deployed in DOSMON's algorithm server for real-time estimation of influent TP and COD.

The plots presented in 5b as well as the metrics presented in Table 2 show that model calibration using SLSPQ and NM algorithms provides better results compared to BFGS and TR algorithms. A faster convergence to an optimal solution is observed in the NM algorithm compared to the other three algorithms. However, the minima obtained by the SLSPQ algorithm shows a better fit (R^2 = 0.76) compared to the other three algorithms, with R^2 values of 0.35 for TR, 0.71 for NM, and 0.63 for BFGS algorithm. Along with an R^2 value close to 1, the model obtained from the SLSPQ algorithm also presents a slope and intercept value of 0.79 and 0.01, respectively, which is closer to the ideal prediction line (with a slope = 1 and intercept = 0) compared to the NM, TR and BFGS algorithms. Slower convergence of the minimization algorithm implies a higher computational requirement to solve the optimization problem. The solver time presented in Table 2 also shows a significant increase in solver time of 18.4 h for the SLSPQ algorithm compared to 9.1 h for

the NM algorithm. Therefore, the use of the NM algorithm is preferred in situations with limits in computational resources or circumstances where frequent recursive calibration of MLR models is required [20]. As computation power and recalibration of the MLR models were not an issue in this study, the results which provided the lowest values of MSE were deployed in real-time for estimating effluent TP and COD values.

Table 2. Comparison of TP and COD effluent model calibration using various solver algorithms.

Parameter	Algorithm	RMSE	R^2	Solver Time (Hours)
TP	NM	0.118	0.71	9.1
	BFGS	0.109	0.63	8.3
	TR	0.184	0.35	10.1
	SLSPQ	0.101	0.76	18.4
COD	NM	4.03	0.61	3.14
	BFGS	4.37	0.50	2.58
	TR	5.20	0.11	3.87
	SLSPQ	3.48	0.70	6.44

3.2. GUI for Visualizing Soft Sensor Data in DOSMON

The 'Dashboard Designer' provided as a part of the DOSMON core (shown in Figure 4) can be used to build dashboards for visualizing data on dial-gauge widgets, trend curves, or layout maps. A user-friendly dashboard was created to visualize the values estimated by the soft sensor deployed in the algorithm server. DOSMON also provides the possibility of downloading the raw data from the web interface that can be used for further data analysis and validations. The interactive interface allows the selection of a desired timescale to visualize both real-time and historical data. A snapshot of the dashboard's front-end page showing both the estimated and the lab-measured data is presented in Figure 6.

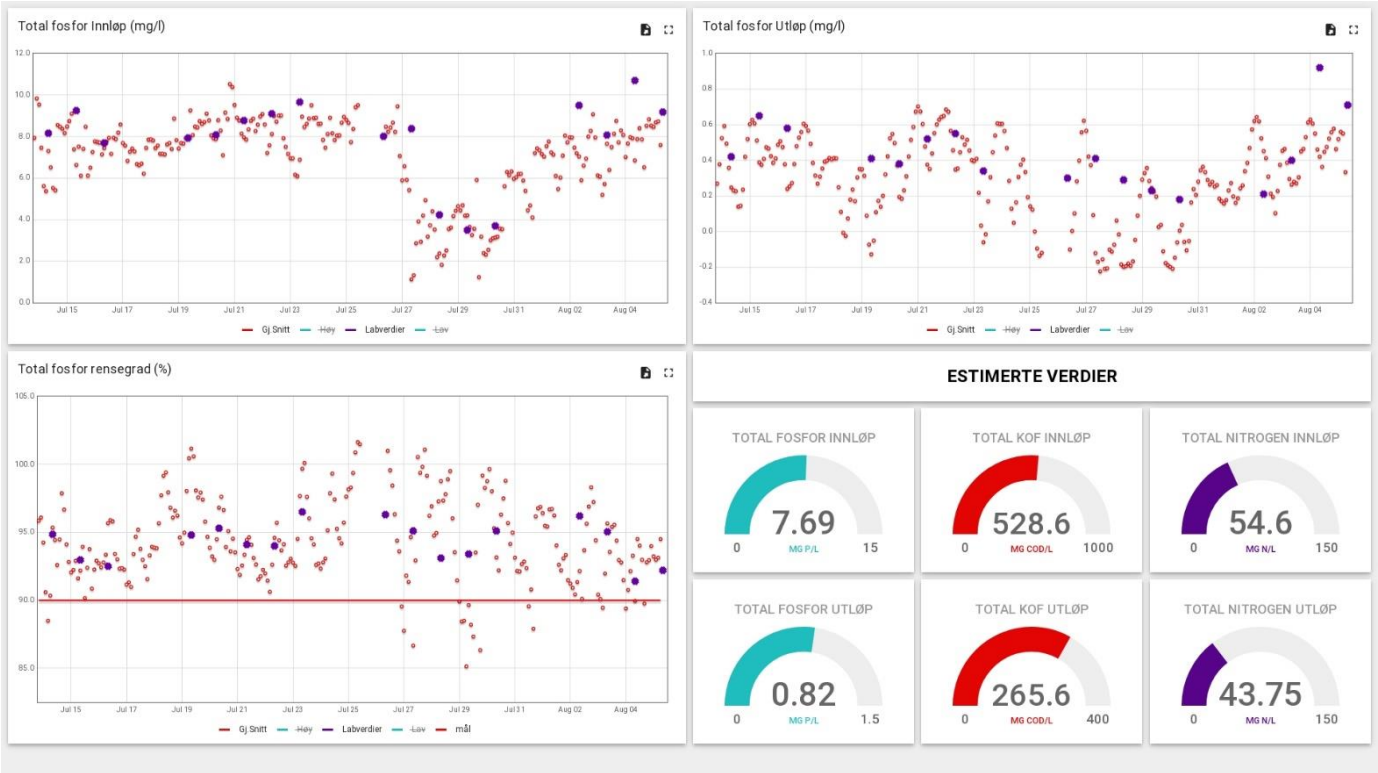


Figure 6. GUI dashboard presenting estimated values of TP influent and effluent along with lab-measured values for a period.

The dashboard designed for visualizing soft sensor data shows real-time estimated values of TP and COD, and NH (currently in the development stage and is not validated) as dial-gauge widgets in the bottom right corner. The dashboard also displays three time-series plots showing the mean real-time estimate (of TP influent, TP effluent and TP removal percentage) as discrete red circles along with the upper and lower limits of prediction error as continuous blue lines. The default timescale of the dashboard is three days, which can be changed manually by the end-user. The lab-measured value of TP influent and TP effluent obtained daily from the treatment plants LIMS software is also provided as discrete purple points in the time-series plots. A guest user account, to access and visualize the live dashboard shown in Figure 6, can be provided on request.

3.3. Soft Sensor Validation Results

The soft sensor estimations of the influent and effluent TP/COD were monitored for four months (from 15 April 2021 to 18 August 2021). A comparison between the values estimated by the soft sensor along with the prediction error (calculated according to the method described in [31]) and the lab-measured values are presented in Figure 7.

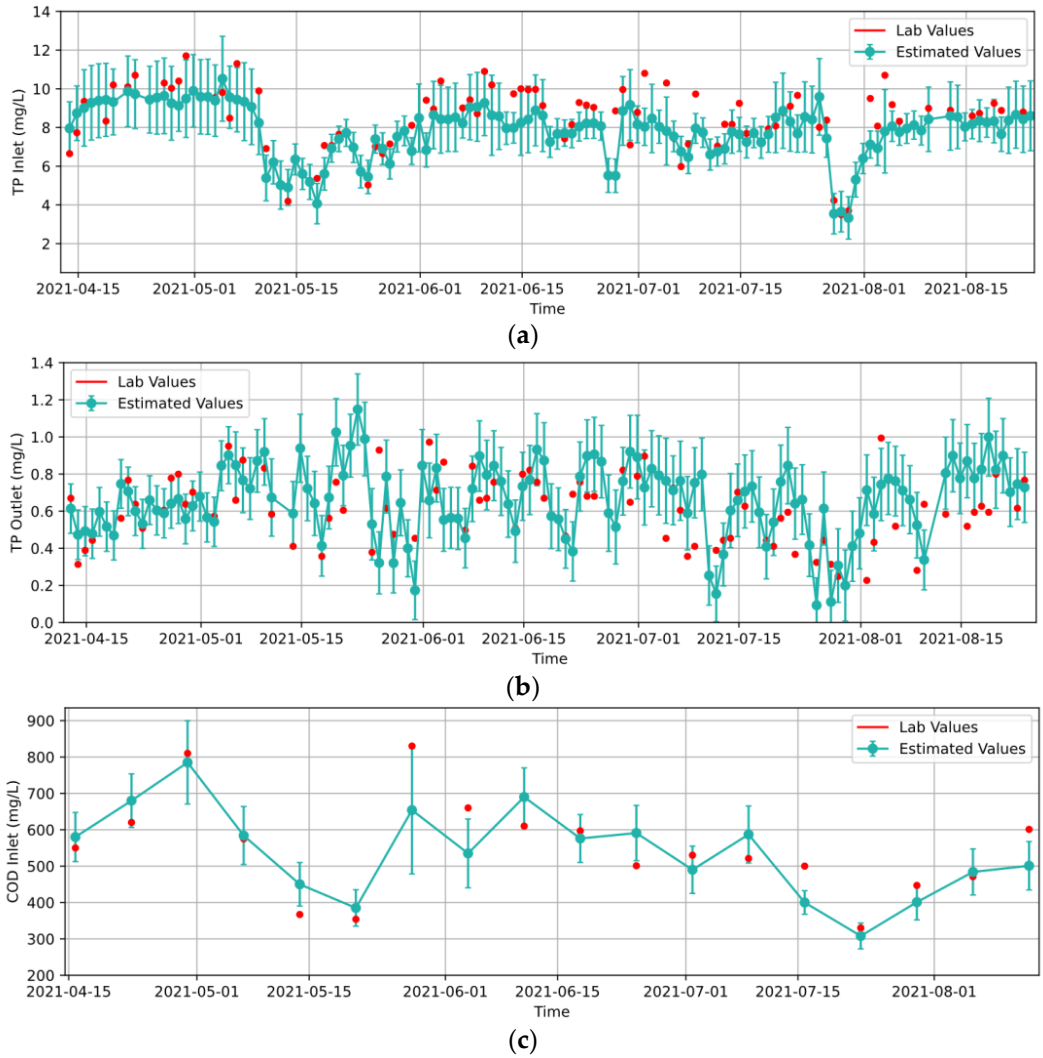
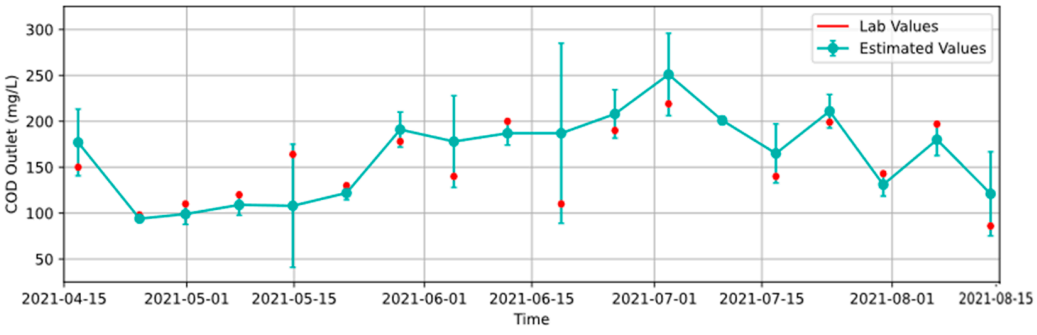


Figure 7. *Cont.*



(d)

Figure 7. Lab-measured versus estimated values during the validation period for (**a**) influent TP, (**b**) effluent TP, (**c**) influent COD, (**d**) effluent COD.

The discrete red points in Figure 7 represent values measured using standardized lab tests, and the continuous blue lines show daily averages of the values estimated by the soft sensor. The discrete blue points show the mean estimated values, and the error bar shows the prediction error of the soft sensor. The plots in Figure 7 show that the soft sensor estimates follow similar trends compared to the lab-measured values. It can be observed that about 90% of the actual lab-measured values lie within the prediction limits of the soft sensor estimates. A few distinct inaccuracies in estimation values are observed in the influent TP estimation when the lab-measured values are above 12 mg/L. A possible reason for the reduction in prediction accuracy could be the insufficiency of data points in the range TP > 12 mg/L used to calibrate the influent MLR model.

The accuracy of the soft sensor estimations can be further substantiated with the plots presented in Figure 8, where a comparison between estimated and lab-measured values along with the perfect prediction line and R^2 values are presented. An R^2 value of 0.81 for TP inlet, 0.73 for TP effluent, 0.76 for COD inlet, and 0.69 for COD effluent shows that a reasonable estimation of wastewater quality parameters in both influent and effluent streams can be achieved using the hybrid soft sensor. It can be observed that TP estimations have a lower prediction error and a better correlation coefficient ($R^2 = 0.81$) compared to COD ($R^2 = 0.72$). This is most likely due to the difference in the number of data points used in calibrating the MLR models for TP and COD. The daily measurements of TP provided a higher number of data points (283 data points) as compared to the weekly average values (51 data points) available for calibrating the MLR model for COD.

Evaluating a soft-sensor algorithm in a full-scale WWTP is especially challenging compared to simulator-based evaluation or pilot-scale testing. Full-scale treatment facilities are subjected to a wide range of unpredicted diurnal and season fluctuations. A better control (or even prior knowledge) of influent disturbances is available while evaluating soft sensor algorithms in a simulator or a pilot-scale unit. The uncertainties in influent disturbances, inherent to a full-scale treatment plant, provide an excellent platform to assess the robustness of the soft sensor algorithm. A comparison between the lab-measured and estimated values presented in Figures 6–8 shows the ability of the soft sensor to consistently deliver reliable estimates of both influent and effluent TP/COD values even in the presence of the disturbances of a full-scale treatment facility.

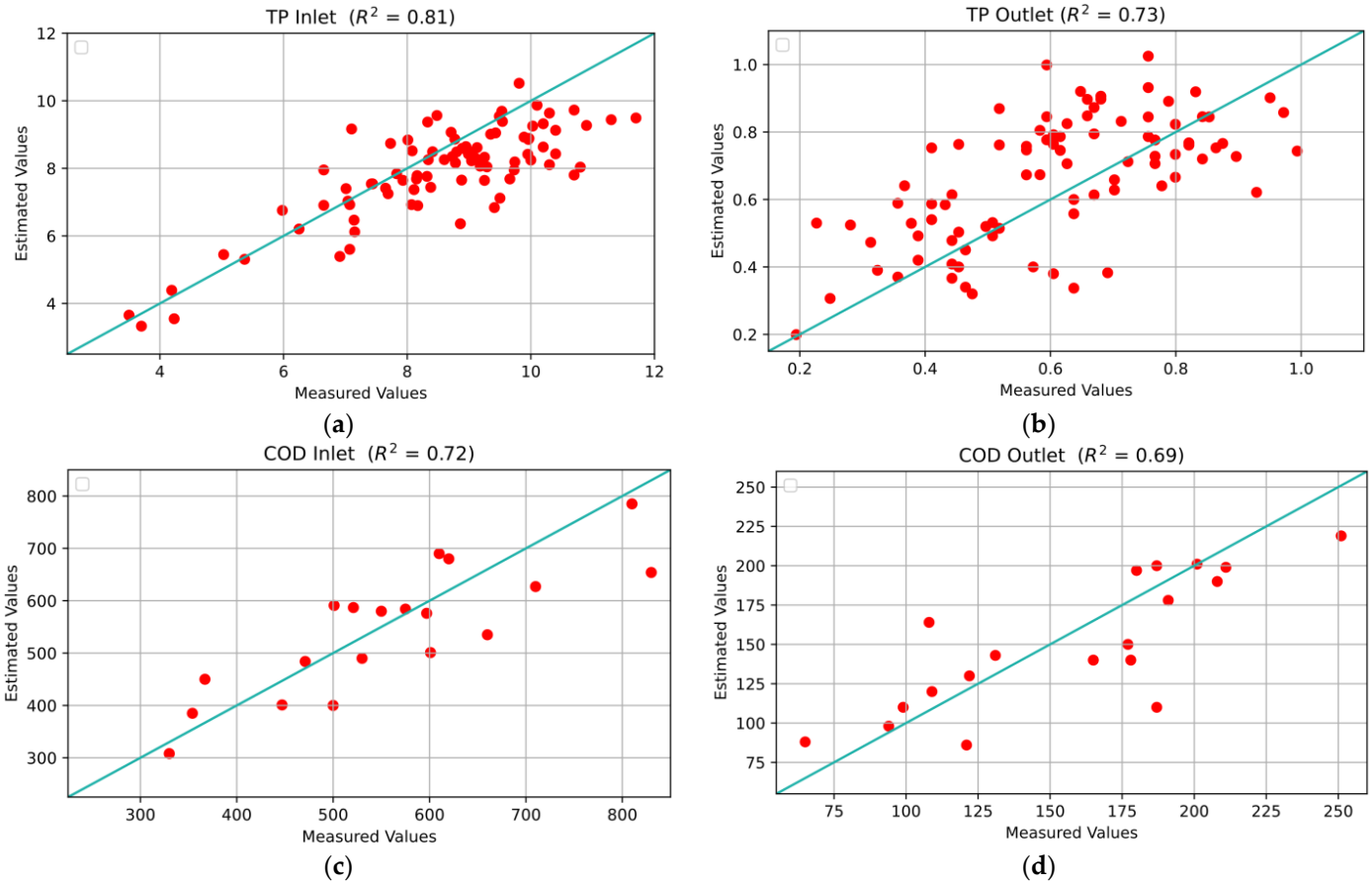


Figure 8. Regression plots for (**a**) influent TP, (**b**) effluent TP (**c**) influent COD, (**d**) effluent COD.

3.4. Benefits of Using a Dynamic Model for Estimating Effluent Wastewater Quality Parameters

The soft sensor development workflow, presented in Section 2.3, shows a difference between the approaches used to develop influent and effluent estimation models. While a conventional MLR model correlating the TP/COD values to their corresponding online sensor data was used for influent wastewater quality parameters, a dynamic model was used for the effluent wastewater quality parameters. A few case studies presenting the estimation of effluent wastewater quality parameters using data-driven soft sensors are found in the literature [32–34]. These case studies show that it is possible to develop a statistically significant correlation between effluent TP/COD and their corresponding online sensor data. Correlation coefficients of R^2 = 0.72 and MSE = 0.013 for effluent TP and R^2 = 0.62 and MSE = 15.5 for effluent COD were obtained when the daily/weekly averaged values of flowrate, effluent suspended solids effluent, PAX/Polymer dosage, and pH (after dosing) were correlated to the lab-measured effluent TP and COD values. However, several issues were encountered while deploying the effluent MLR models in the algorithm servers for real-time estimations. The estimation error was particularly visible for effluent TP, where the soft sensors estimated values that were mechanistically impossible to attain. The estimated values frequently showed as negative effluent TP values or effluent TP being higher than influent TP. The frequent extremities of the estimated TP and COD values can be dealt with by implementing an averaging filter, by imposing min-max limits, or by integrating conditional loops limiting the TPO to always be below TPI estimates. However, these constraints failed to achieve a general smoothness in the TP and COD estimations in the effluent. A comparison between the effluent TP estimated by the conventional MLR model versus the dynamic model is presented in Figure 9.

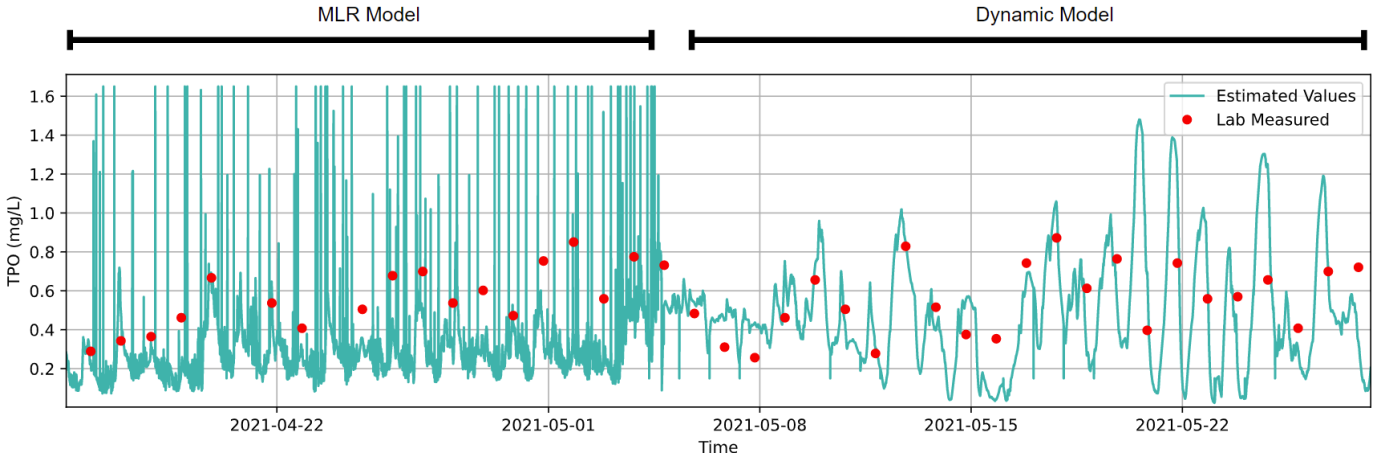


Figure 9. Estimated effluent TP values in MLR versus dynamic models.

The dynamic model used to estimate the effluent wastewater quality parameters are based on the mass balance of TP/COD in the coagulation-flocculation process. An n-order kinetics is assigned to the removal term to ensure that the TP/COD removal decreases significantly when the estimated values reach a value close to 0. This mathematical constraint prevents the estimated values of effluent TP and COD from dropping to negative values. It should also be noted that the estimated values of influent TP and COD are included in the dynamic model, which ensures that the effluent TP and COD values are lower than their influent values. The concept of integrating mechanistic and data-driven techniques in the soft sensor algorithm has resulted in a significant improvement in the estimated effluent values. The improvement in effluent TP estimation can be observed in the plots presented in Figure 9, where during the first 15 days (15 April 2021–5 May 2021), the soft sensor used the MLR model to estimate TP values, after which (from 5 May 2021 onwards) the estimator algorithm switched to the dynamic model. The rapid variations in the estimated TP values (sudden jumps between the upper limits of 1.6 mg/L and the lower limit of 0.1 mg/L) frequently occurring in the first 15 days are non-existent after the soft-sensor algorithm switches to the dynamic model.

3.5. Limitations of Hybrid Estimator and Possible Improvements

The hybrid soft sensor developed in this work uses the daily average value of TP and weekly averages of COD to calibrate the regression models. Although the soft-sensor algorithm has demonstrated a reasonably accurate estimation of the average TP and COD values during the validation stage, the estimation accuracies of their diurnal variations were not validated. The variations observed in estimated TP and COD values are due to the diurnal variations in the predictors (flowrate, suspended solids, conductivity, and pH) included in the estimation model. However, the possibility of a mismatch between the actual TP/COD values and values estimated by the soft sensor cannot be completely ruled out.

A possible solution to improve the soft sensor is to increase the time-resolution of the lab-measured data used in calibrating the MLR models. A rigorous sampling campaign should be conducted where composite samples are collected once or twice every hour to measure TP and COD concentrations. This hourly (or half-hourly) data can provide a better insight into the diurnal variations in wastewater quality parameters at the influent and the effluent. The MLR models calibrated using these high time-resolution lab measured data can better capture diurnal variations and subsequently improve the estimation results.

4. Conclusions

An alternative cost-effective method for monitoring essential wastewater quality parameters such as TP and COD in a full-scale wastewater treatment plant was tested and

validated in this work. The secure, non-intrusive, cost-effective system enables the codes written in scientific programming languages such as Python to be deployed for real-time estimation. The results presented in this work demonstrates that MLR models can be used to develop statistically significant correlations between online sensor data and wastewater quality parameters such as TP and COD. Among various different algorithms used to calibrate the MLR models, the Ridge algorithm showed the best model fit for the influent TP/COD, while the SLSQP algorithm provided the best fit for the removal model used to estimate effluent TP/COD values. Hybrid models combining mechanistic elements with data-driven techniques have shown better prediction accuracy compared to purely black-box models. The systematic approach presented in the work can be further expanded to estimate additional wastewater quality parameters (nitrogen, VFA, etc.) provided adequate lab-measured values are available. The IIoT architecture described in this work presents a seamless infrastructure to deploy, modify, and tune soft sensors in a full-scale treatment plant without causing any operational downtime.

Author Contributions: Conceptualization, A.N. and H.R.; methodology, A.N.; software, A.H. and A.N.; validation, A.N.; formal analysis, A.N. and H.R.; investigation, A.N. and A.H.; resources, H.R.; data curation, A.N.; writing—original draft preparation, A.N.; writing—review and editing, A.H. and H.R.; visualization, A.N.; supervision, H.R.; project administration, H.R.; funding acquisition, H.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Water JPI Water Harmony project and DOSCON AS.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data reported in this paper are accessible from the NSD—Norwegian center for research data.

Acknowledgments: Authors wish to acknowledge the encouraging discussions and logistical assistance from Geir Simensen and Rune Larsen at the Sondre Follo WWTP, Vestby, Norway.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ødegaard, H. Norwegian Experiences with Chemical Treatment of Raw Wastewater. *Water Sci. Technol.* **1992**, *25*, 255–264. [CrossRef]
2. Liu, W.; Ratnaweera, H. Improvement of Multi-Parameter-Based Feed-Forward Coagulant Dosing Control Systems with Feed-Back Functionalities. *Water Sci. Technol.* **2016**, *74*, 491–499. [CrossRef] [PubMed]
3. American Public Health Association. *Standard Methods for the Examination of Water and Wastewater*, 21st ed.; American Public Health Association: Washington, DC, USA, 2005; ISBN 0-87553-047-8.
4. Gasperi, J.; Laborie, B.; Rocher, V. Treatment of Combined Sewer Overflows by Ballasted Flocculation: Removal Study of a Large Broad Spectrum of Pollutants. *Chem. Eng. J.* **2012**, *211–212*, 293–301. [CrossRef]
5. Yaqoob, A.A.; Parveen, T.; Umar, K.; Mohamad Ibrahim, M.N. Role of Nanomaterials in the Treatment of Wastewater: A Review. *Water* **2020**, *12*, 495. [CrossRef]
6. Olsson, G. ICA and Me—A Subjective Review. *Water Res.* **2012**, *46*, 1585–1624. [CrossRef] [PubMed]
7. Ratnaweera, H.; Fettig, J. State of the Art of Online Monitoring and Control of the Coagulation Process. *Water* **2015**, *7*, 6574–6597. [CrossRef]
8. Xiao, H.; Bai, B.; Li, X.; Liu, J.; Liu, Y.; Huang, D. Interval Multiple-Output Soft Sensors Development with Capacity Control for Wastewater Treatment Applications: A Comparative Study. *Chemom. Intell. Lab. Syst.* **2019**, *184*, 82–93. [CrossRef]
9. Haimi, H.; Corona, F.; Mulas, M.; Sundell, L.; Heinonen, M.; Vahala, R. Shall We Use Hardware Sensor Measurements or Soft-Sensor Estimates? Case Study in a Full-Scale WWTP. *Environ. Model. Softw.* **2015**, *72*, 215–229. [CrossRef]
10. Haimi, H.; Mulas, M.; Corona, F.; Vahala, R. Data-Derived Soft-Sensors for Biological Wastewater Treatment Plants: An Overview. *Environ. Model. Softw.* **2013**, *47*, 88–107. [CrossRef]
11. Busch, J.; Kühl, P.; Schlöder, J.P.; Bock, H.G.; Marquardt, W. State Estimation for Large-Scale Wastewater Treatment Plants. *IFAC Proc. Vol.* **2009**, *42*, 596–601. [CrossRef]
12. Nair, A.M.; Fanta, A.; Haugen, F.A.; Ratnaweera, H. Implementing an Extended Kalman Filter for Estimating Nutrient Composition in a Sequential Batch MBBR Pilot Plant. *Water Sci. Technol.* **2019**, *80*, 317–328. [CrossRef] [PubMed]

13. Manamperuma, L.; Wei, L.; Ratnaweera, H. Multi-Parameter Based Coagulant Dosing Control. *Water Sci. Technol.* **2017**, *75*, 2157–2162. [CrossRef] [PubMed]
14. Moruzzi, R.B.; de Oliveira, S.C. Mathematical Modeling and Analysis of the Flocculation Process in Chambers in Series. *Bioprocess Biosyst. Eng.* **2013**, *36*, 357–363. [CrossRef]
15. Zare Abyaneh, H. Evaluation of Multivariate Linear Regression and Artificial Neural Networks in Prediction of Water Quality Parameters. *J. Environ. Health Sci. Eng.* **2014**, *12*, 40. [CrossRef] [PubMed]
16. Jobson, J.D. Multiple Linear Regression. In *Applied Multivariate Data Analysis: Regression and Experimental Design*; Jobson, J.D., Ed.; Springer: New York, NY, USA, 1991; pp. 219–398, ISBN 978-1-4612-0955-3.
17. Olsson, G.; Nielsen, M.; Yuan, Z.; Lynggaard-Jensen, A.; Steyer, J.-P. *Instrumentation, Control and Automation in Wastewater Systems*; IWA Publishing: London, UK, 2005; ISBN 978-1-78040-268-0.
18. Baggiani, F.; Marsili-Libelli, S. Real-Time Fault Detection and Isolation in Biological Wastewater Treatment Plants. *Water Sci. Technol.* **2009**, *60*, 2949–2961. [CrossRef] [PubMed]
19. Haimi, H.; Mulas, M.; Corona, F.; Marsili-Libelli, S.; Lindell, P.; Heinonen, M.; Vahala, R. Adaptive Data-Derived Anomaly Detection in the Activated Sludge Process of a Large-Scale Wastewater Treatment Plant. *Eng. Appl. Artif. Intell.* **2016**, *52*, 65–80. [CrossRef]
20. Nair, A.M.; Gonzalez-Silva, B.M.; Haugen, F.A.; Ratnaweera, H.; Østerhus, S.W. Real-Time Monitoring of Enhanced Biological Phosphorus Removal in a Multistage EBPR-MBBR Using a Soft-Sensor for Phosphates. *J. Water Process Eng.* **2020**, *37*, 101494. [CrossRef]
21. Hong, S.H.; Lee, M.W.; Lee, D.S.; Park, J.M. Monitoring of Sequencing Batch Reactor for Nitrogen and Phosphorus Removal Using Neural Networks. *Biochem. Eng. J.* **2007**, *35*, 365–370. [CrossRef]
22. Manamperuma, L.D.; Ratnaweera, H.C.; Martsul, A. Mechanisms during Suspended Solids and Phosphate Concentration Variations in Wastewater Coagulation Process. *Environ. Technol.* **2016**, *37*, 2405–2413. [CrossRef]
23. Marsili Libelli, S.; Ratini, P.; Spagni, A.; Bortone, G. Implementation, Study and Calibration of a Modified ASM2d for the Simulation of SBR Processes. *Water Sci. Technol.* **2001**, *43*, 69–76. [CrossRef]
24. Nair, A.; Cristea, V.-M.; Agachi, P.Ş.; Brehar, M. Model Calibration and Feed-Forward Control of the Wastewater Treatment Plant—Case Study for CLUJ-Napoca WWTP: Case Study for CLUJ-Napoca WWTP. *Water Environ. J.* **2018**, *32*, 164–172. [CrossRef]
25. Gao, F.; Han, L. Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters. *Comput. Optim. Appl.* **2012**, *51*, 259–277. [CrossRef]
26. Gould, N.I.M.; Lucidi, S.; Roma, M.; Toint, P.L. Solving the Trust-Region Subproblem Using the Lanczos Method. *SIAM J. Optim.* **1999**, *9*, 504–525. [CrossRef]
27. Gill, P.E.; Murray, W.; Saunders, M.A. SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization. *SIAM Rev.* **2005**, *47*, 99–131. [CrossRef]
28. Mokhtari, A..; Ribeiro, A. RES: Regularized Stochastic BFGS Algorithm. *IEEE Trans. Signal Process.* **2014**, *62*, 6089–6104. [CrossRef]
29. Schütz, D.; Wannagat, A.; Legat, C.; Vogel-Heuser, B. Development of PLC-Based Software for Increasing the Dependability of Production Automation Systems. *IEEE Trans. Ind. Inform.* **2013**, *9*, 2397–2406. [CrossRef]
30. Nair, A.M.; Hykkerud, A.; Ratnaweera, H. A Cost-Effective IoT Strategy for Remote Deployment of Soft Sensors—A Case Study on Implementing a Soft Sensor in a Multistage MBBR Plant. *Water Sci. Technol.* **2020**, *81*, 1733–1739. [CrossRef]
31. Duník, J.; Šimandl, M. Estimation of State and Measurement Noise Covariance Matrices by Multi-Step Prediction. *IFAC Proc. Vol.* **2008**, *41*, 3689–3694. [CrossRef]
32. Luccarini, L.; Porrà, E.; Spagni, A.; Ratini, P.; Grilli, S.; Longhi, S.; Bortone, G. Soft Sensors for Control of Nitrogen and Phosphorus Removal from Wastewaters by Neural Networks. *Water Sci. Technol.* **2002**, *45*, 101–107. [CrossRef]
33. Zhu, S.; Han, H.; Guo, M.; Qiao, J. A Data-Derived Soft-Sensor Method for Monitoring Effluent Total Phosphorus. *Chin. J. Chem. Eng.* **2017**, *25*, 1791–1797. [CrossRef]
34. Liu, W.; Ratnaweera, H.; Kvaal, K. Model-Based Measurement Error Detection of a Coagulant Dosage Control System. *Int. J. Environ. Sci. Technol.* **2019**, *16*, 3135–3144. [CrossRef]