Comparison of ML and ASM models for effluent nutrient estimation in the Hias Process ⋆

Tiina M. Komulainen * **Malik Baqeri** *
Katrine Marsteng Jansen ** **Arvind Keprate** *

* *GrønnMet - Green Energy Lab, Department of Mechanical, Electrical and Chemical Engineering, Oslo Metropolitan University, Norway (e-mail corresponding author: tiina.komulainen@oslomet.no).*
** *Hias IKS, Norway*

Abstract: The aim of this article is to develop and compare machine learning (ML) methods with activated sludge models (ASM) for estimation of effluent nutrients in the Hias Process. The Hias Process is a novel moving bed bioreactor with enhanced biological phosphorus removal and simultaneous nitrification and denitrification (MBBR-EBPR-SND). As the main energy cost of the nutrient removal process is aeration, it is necessary to design of energy-efficient control strategies that ensure compliance with legal requirements for nutrient removal in real-time while optimizing the aeration rates. The first step in control strategy design is development of models that represent the main process dynamics.
The case study data set of four months was collected from a 192 000 PE municipal MBBR process at Hias water resource recovery facility in Norway. The Hias Process consists of three anaerobic and seven aerobic zones, where biomass carriers flow continuously submersed in the used water and remove over 90 % of the phosphorus. The online measurements include used water flowrate, aeration rates, dissolved oxygen, suspended solids, and soluble nutrients PO_4, COD, NO_2 and NO_3. Reduced ASM model, support vector regression (SVR) and long short-term memory neural network (LSTM), with and without dynamic time-delay, were developed to predict the effluent PO_4 in the Hias process. The model prediction accuracies were compared using correlation coefficients and trend figures. The SVR model with fine gaussian kernel gave best results with strong R index of 0.9. The LSTM model reached a sufficient R index of 0.6 and the reduced ASM2d model a weak R index of 0.2. Including the dynamic time-delay improved the model accuracy. The machine learning models with dynamic time-delay will be developed further for energy-efficient control strategy development.

Keywords: water resource recovery facility; activated sludge model; support vector regression; long short-term memory network.

1. INTRODUCTION

As the water industry is responsible for approximately one percent of the total energy consumption in the European union, new legal requirements for energy neutrality are underway EuropeanCommission (2021). Therefore, development of energy-efficient control strategies is essential to minimize the energy consumption and to meet the strict nutrient recovery requirements at water resource recovery facilities (WRRF). The Hias Process is a novel, compact biological nutrient removal process that consists of a continuous-flow moving bed bioreactor with enhanced biological phosphorus removal and simultaneous nitrification and denitrification (MBBR-EBPR-SND) Rudi et al. (2019). As the main energy cost of the Hias Process is aeration, energy-efficient control strategies need to be developed to optimize the aeration rates and to ensure compliance with legal requirements for nutrient removal in real-time.

Development of energy-efficient control strategies requires models that capture the main dynamic behaviour of the nutrient removal phenomena. First principles models such as the ASM2d model Henze et al. (1999) and its simplified version the reduced ASM2d model Nair et al. (2019) can be used as development and testing environment for control strategies. However, it is time-consuming and sometimes unfeasible to develop such models due to scarce instrumentation. Hence, machine learning models have gained high research interest in the water industry, for example for applications such as virtual/soft sensors Paepae et al. (2021).

In our previous work, data-driven models were developed to estimate the effluent nutrients in the Hias Process Nermo (2023), Komulainen et al. (2023). Virtual sensors were developed for estimation of PO_4 and COD in the Hias Process using additional electrical conductivity (EC)

⋆ RFF Innlandet, Norway, is gratefully acknowledged for funding the PACBAL project (nr 337727).

measurements Komulainen et al. (2024). In this work we will use the virtual sensor estimating PO_4 at inlet, develop reduced ASM2d models, and refine two best performing machine learning models from Baqeri (2024), Support Vector Regression (SVR) and Long-Short Term Memory (LSTM). In this study we answer the following research questions: Can the reduced ASM2d model, SVR model and LSTM model follow the dynamic trends of the effluent PO_4 data? Which model gives the highest prediction accuracy?

2. MATERIALS AND METHODS

2.1 Software

Matlab software package version R2023a was used in the work. The simulation method was ode23s with automatic settings for the time step and error tolerance.

2.2 The Hias Process and instrumentation

The Hias Process is a biological nutrient removal process at a 192 000 PE municipal water resource recovery facility in Hamar, Norway. The Hias Process with instrumentation is illustrated in Fig. 1. The clarified used water (influent) and the recirculated biofilm carriers (from zone 10 via a conveyor belt) enter the anaerobic zone. The Hias Process consists of three anaerobic and seven aerobic zones, where biomass carriers flow continuously submersed in the used water and remove over 90 % of the phosphorus. The three anaerobic basins are mixed to ensure sufficient distribution of biofilm carriers in the water. Aeration in the following seven basins ensures sufficient dissolved oxygen concentrations for aerobic nutrient removal. Used water and submerged biofilm carriers float through the process with gravity.

The Hias Process instrumentation includes continuous online measurements of flowrate, temperature, aeration, dissolved oxygen, suspended solids, and nutrient compositions of PO_4, COD, NO_2 and NO_3. Additional online measurements of electrical conductivity at inlet and in zone 3 were installed during the PACBAL research project 2022-2023. Soluble COD, NO_2 and NO_3 are measured continuously at inlet and zone 7. Suspended solids SS are measured at zone 10 and in the effluent after the disc filter. Effluent PO_4 is measured using an online-analyzer with 10 minutes sampling time. The Hias Process online measurements utilized in this study are listed in Table 1. Hias laboratory assesses nutrient composition of PO_4 and soluble COD at inlet, zones 3,4,7,10, and outlet, and NH_3 at inlet and outlet from daily grab samples five days a week. Hence, there are 5 samples from laboratory and 1008 samples of online data for each variable per week.

2.3 Data collection and pre-processing

The Industrial IoT platform KYB, developed by Digitread Connect, was used for uploading and standardizing operational data from SCADA system of municipal Hias water resource recovery facility at Hamar, Norway. The online data set was collected in .csv format and the laboratory data set in .xlsx format.

Table 1. Online measurements.

Symbol	Description	Unit
F	Water flowrate inlet	m^3/h
T	Temperature inlet	^oC
$CODIN$	COD inlet	g/m^3
$NOIN$	NO_2 and NO_3 inlet	g/m^3
$ECIN$	El.conductivity inlet	mS/cm
FO_i	Aeration rate zones 4,5,6,7,8,9,10	m^3/h
DO_i	Dissolved oxygen zones 4,5,6,8,9	m^3/h
PO	PO_4 effluent	g/m^3

The inter-quartiles method was chosen for outlier removal. The outliers are identified as measurements more than 1.5 inter-quartile range above the upper quartile (75 percent) or below the lower quartile (25 percent). Missing and removed values were replaced with previous feasible values. Prior to ML modeling, the data is normalized to zero mean and standard deviation of one.

2.4 Modeling methods

The aim of this study is to develop and compare modeling methods that enable energy-efficient control strategy design for nutrient removal in the Hias Process. International Water Association has led work in developing Activated Sludge Models (ASM) that represent biological nutrient removal Henze et al. (1999). The ASM models were reduced and developed for a sequential MBBR pilot plant by Nair et al. (2019). In this study, these models were to be adapted for the continuous large-scale operation of the municipal Hias WRRF.

In the literature, Long-Short Term Memory (LSTM) neural network and Support Vector Regression (SVR) have gained lots of attention, and hence these two data-driven modeling methods are applied in this study. LSTM is a recurrent neural network suitable for modeling dependencies in and forecasting of sequential or time-series data. LSTM architecture includes memory cells and gates that regulate the flow of sequential data. These gates can learn which data in a sequence is important to keep or discard, enabling the network to maintain a longer context of information as described in Hochreiter and Schmidhuber (1997).

SVR is frequently used to predict relationship between continuous input and output variables. SVR minimizes error between the model prediction and the data by fitting a hyperplane in a high-dimensional space of the input variables and the output variables . The kernel trick converts the dataset to higher dimensions by combining the features using for example linear, quadratic, cubic, or Gaussian functions as described in Cortes and Vapnik (1995).

2.5 Model comparison

The Hias Process effluent PO_4 measurements were used as the output variable for the models. Both R^2 index and correlation coefficient R were used to compare the modeling accuracy of the different methods. The model prediction of the real data points is weak for index between 0-0.3, moderate for index between 0.3-0.5, sufficient for index between 0.5-0.7, and strong for index between 0.7-1.

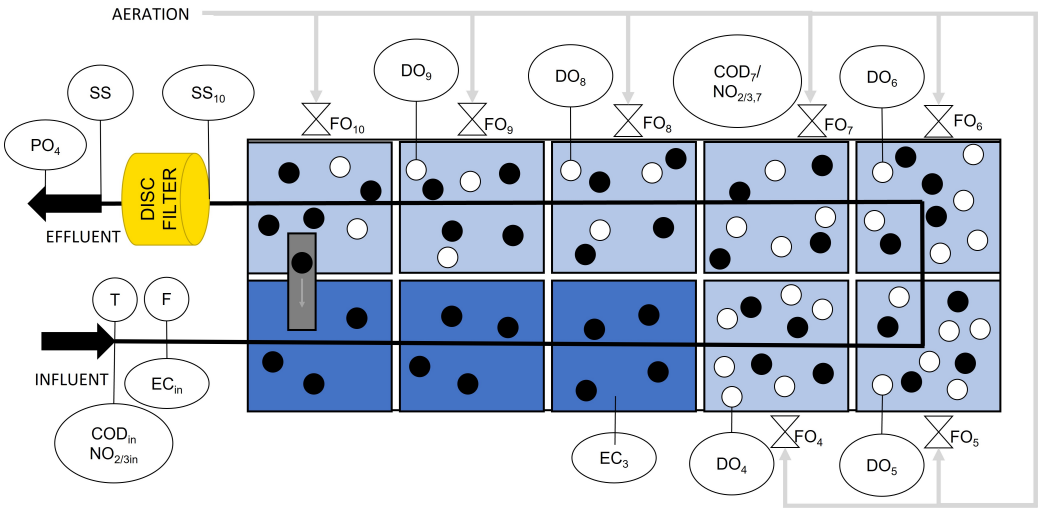

Fig. 1. The Hias Process with instrumentation.

3. RESULTS

3.1 Data selection and pre-processing

The online data and laboratory data were collected for a period of 21.3.-31.7.2023. During period of 1.4.-15.5.2023 many online measurements were missing, hence, the online data set is for period 15.5.2023-31.7.2023. Period of 21.3.-31.7.2023 is used for the laboratory data set.

The outliers in the data set were using inter-quartiles method. The missing values were filled in with previous feasible values. Prior to the machine learning model development, the online data was normalized.

3.2 Dynamic time delay

The time delay through the ten zones of the Hias Process has a mean of 6 hours (36 samples), a standard deviation of 1 hour (6 samples), a minimum of 4.8 hours (29 samples) and a maximum of 15 hours (90 samples). Hence, in this study the effect of dynamic time delay on model accuracy was analysed. For the reduced ASM2d models, the varying time delay t_d is calculated for a lumped volume V combining three real process zones ($3 \cdot 215m^3$) and continuous measurement of used water flowrate $F(t)$ in $[m^3/h]$ according to Equation 1:

$$t_d(t) = V/F(t) = 645m^3/F(t) \quad (1)$$

As the sharp variations and small oscillations in the time delay can cause numerical challenges in simulation, the time-delay was smoothed with a moving-mean approach. Time windows of 6 samples (1h), 12 samples (2h) and 18 samples (3h) were plotted against the calculated time-delay. Varying time-delay with moving mean of 12 samples (2h) was chosen for this study as it removes the fast oscillations that are present in moving mean of 6 samples, but follows the main trends more closely than moving mean of 18 samples.

3.3 Virtual measurements of inlet PO_4 and NH_3

As the online data set does not include online measurements of inlet PO_4 and NH_3, these were estimated from the laboratory data and online measurements of electrical conductivity and COD. In Komulainen et al. (2024) virtual sensors were developed for estimation of PO_4 at the Hias Process inlet using additional electrical conductivity (EC) measurement. Based on 32 unique laboratory data points, a linear regression was fitted between $ECIN$, $POIN$ and $CODIN$, given in Equation 2. The parameter values were $c_1 = 0.3488$ and $c_2 = 0.6138$ with strong modeling accuracy of $R^2 = 0.86$.

$$ECIN = c_1 POIN + c_2 CODIN \quad (2)$$

In this work we developed a simple estimation of NH at the inlet. Based on 28 unique laboratory data points, a linear regression was fitted between $NHIN$ and $CODIN$, given in Equation 3. The parameter values were $c_3 = 0.1211$ with strong modeling accuracy of $R^2 = 0.97$.

$$NHIN = c_3 CODIN \quad (3)$$

3.4 Reduced ASM2d model development

The activated sludge models (ASM) describe the dynamic changes in the nutrient concentrations and dissolved oxygen in the process zones. The simplified ASM2d models developed for a pilot scale batch-MBBR process by Nair et al. (2019) were further modified to fit the available measurements in the continuous large-scale municipal WRRF process that is the subject of this study. Significant simplifications are necessary to match the model variables with the available online measurements in the Hias process. The simplified models included nutrient uptake and release by the microbes in the biomass carriers. The following assumptions were applied:

- Phosphate PO_4, soluble organic substrate COD, ammonia NH_3, sum of nitrate NO_2 and nitrite NO_3, and dissolved oxygen DO_2 are the components in the simplified models.
- Ready biodegradable substrate (S_F) and volatile fatty acids (S_A) presented in Nair et al. (2019) are lumped together as soluble substrate (S_S). In this study, interpreted as the measured variable, soluble COD.

- Particulate biodegradable components (X_S) are omitted due to missing online measurement of influent total suspended solids.
- The biomass variables of stored poly-phosphate (X_{PP}) and stored organic compounds COD (X_{PHA}) are omitted to simplify the equations.
- Temperature effect is neglected, as it varies very little between days. It increases slowly from 8.8 ^oC to 13.6 ^oC during the four months period.
- Plug flow is assumed for water movement between zones. The time delay is dynamically time-delayed t_d.
- To match the model variables and real measurements, the process volume is divided to three parts: one lumped anaerobic volume, where V1 includes zones 1-3, two lumped aerobic volumes where V2 includes zones 4-6 and V3 includes zones 7-9 as illustrated in Figure 2. Zone 10, representing 10% of total process volume is omitted to simplify the calculations.

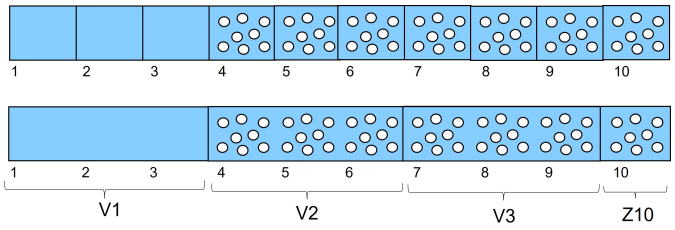

Fig. 2. Above: The Hias Process with 10 zones. Below: Modeling approach with 3 lumped zones.

Anaerobic volume V1 In anaerobic zones i, the biomass consumes soluble COD and the component balance follows Equation 4.

$$\frac{dCOD_i(t)}{dt} = \frac{F(t)}{V}(COD_{i-1}(t-t_d) - COD_i(t)) - r_3\frac{COD_i(t)}{K_A + COD_i(t)} \tag{4}$$

Simultaneously, the biomass releases PO to water with stoichiometric relation Y_{PO} to COD uptake following Equation 5.

$$\frac{dPO_i(t)}{dt} = \frac{F(t)}{V}(PO_{i-1}(t-t_d) - PO_i(t)) + Y_{PO}r_3\frac{COD_i(t)}{K_A + COD_i(t)} \tag{5}$$

It is assumed that concentrations of ammonia $dNH(t)/dt = 0$, nitrate/nitrite $dNO(t)/dt = 0$ and dissolved oxygen $dDO(t)/dt = 0$ remain unchanged through the anaerobic zone.

Aerobic volumes V2 and V3 In aerobic zones j, the biomass takes up soluble COD and consume oxygen DO, the component balance follows Equation 6.

$$\frac{dCOD_j(t)}{dt} = \frac{F(t)}{V}(COD_{j-1}(t-3t_d) - COD_j(t)) - r_1\frac{COD_j(t)}{K_S + COD_j(t)}\frac{DO_j(t)}{K_O + DO_j(t)} \tag{6}$$

In aerobic zones, the biomass takes up phosphate PO and consume oxygen DO, the component balance follows Equation 7.

$$\frac{dPO_j(t)}{dt} = \frac{F(t)}{V}(PO_{j-1}(t-3t_d) - PO_j(t)) - r_4\frac{PO_j(t)}{K_{PS} + PO_j(t)}\frac{DO_j(t)}{K_O + DO_j(t)} \tag{7}$$

In aerobic zones, the biomass convert ammonia NH to nitrite and nitrate NO, and consume oxygen DO. The component balance follows Equation 8.

$$\frac{dNH_j(t)}{dt} = \frac{F(t)}{V}(NH_{j-1}(t-3t_d) - NH_j(t)) - r_5\frac{NH_j(t)}{K_{NH} + NH_j(t)}\frac{DO_j(t)}{K_{OAOB} + DO_j(t)} \tag{8}$$

Simultaneously in deeper layers of biofilm, biomass converts nitrite and nitrate NO into nitrogen gas. The component balance follows Equation 9.

$$\frac{dNO_j(t)}{dt} = \frac{F(t)}{V}(NO_{j-1}(t-3t_d) - NO_j(t)) + r_5\frac{NH_j(t)}{K_{NH} + NH_j(t)}\frac{DO_j(t)}{K_{OAOB} + DO_j(t)} - r_6\frac{NO_j(t)}{K_{NO} + NO_j(t)}\frac{K_O(t)}{K_O + DO_j(t)} \tag{9}$$

In aerobic zones, the dissolved oxygen DO component balance consists of mass transfer in and out of the zone, mass transfer from aeration FO, the biomass consuming oxygen for nutrient uptake. The oxygen component balance follows Equation 10.

$$\begin{aligned}\frac{dDO_j(t)}{dt} = {} & \frac{F(t)}{V}(DO_{j-1}(t-3t_d) - DO_j(t)) \\ & + K_L\frac{FO(t)}{V}(DO^*_{max}(t) - DO_j(t)) \\ & - r_1\frac{COD_j(t)}{K_S + COD_j(t)}\frac{DO_j(t)}{K_O + DO_j(t)} \\ & - Y_{PA}r_4\frac{PO_j(t)}{K_{PS} + PO_j(t)}\frac{DO_j(t)}{K_O + DO_j(t)} \\ & - Y_{NH}r_5\frac{NH_j(t)}{K_{NH} + NH_j(t)}\frac{DO_j(t)}{K_{AOB} + DO_j(t)}\end{aligned} \tag{10}$$

The saturation coefficients, stochiometric constants and rate constants from Nair et al. (2019) were used in this work, presented in Table 2. The initial conditions for simulation models are calculated for aggregated laboratory and online data set, presented in Table 3. The models were implemented in Matlab and Simulink. The parameters from Nair et al. (2019) did not give a reasonable fit to Hias Process data, hence, the reaction rates r and stoichiometric constant Y_{PO} were optimized further using the Hias data set. The parameters were fitted to the initial steady state data 3 by setting the ordinary differential

equations 4 - 10 to zero. The tuned reaction rates r and constant Y_{PO} together with correlation coefficient R for model fitness are presented in Table 4.

Table 2. Reduced ASM2d model parameter values for saturation coefficients K, rate constants r, and stochiometric constants Y.

Symbol	Description	Value	Unit
K_A	COD anaerobic	2.20	$gCOD/m^3$
K_S	COD aerobic	0.11	$gCOD/m^3$
K_{PS}	PO aerobic	0.2	gP/m^3
K_O	DO aerobic	2.96	gO_2/m^3
K_{OAOB}	DO aerobic nitrifiers	1.57	gO_2/m^3
K_{NH}	NH aerobic	1	gN/m^3
K_{NO}	NO aerobic	1.02	gN/m^3
K_L	Aeration	38	gN/m^3
r_3	COD anaerobic	112	$gCOD/(m^3h)$
r_1	COD aerobic	20.8	$gCOD/(m^3h)$
r_4	PO aerobic	171	$gP/(m^3h)$
r_5	NH aerobic	17.9	$gN/(m^3h)$
r_6	NOX aerobic	12.9	$gN/(m^3h)$
Y_{PO}	P/COD anaerobic P release	0.577	$gP/gCOD$
Y_{PA}	DO/P aerobic P storage	1.496	gO_2/gP
Y_{NH}	DO/NH aerobic nitrification	4.32	gO_2/gNH

Table 3. Initial conditions for nutrients, flowrates and temperature dated 12.05.2023 at 12:00. Values marked with * are assumed values.

Nutrient	IN	V1	V2	V3	unit
PO	5.56	36.4	3.83	0.14	gP/m^3
COD	450	202	72	67.5	$gCOD/m^3$
NH	53	53*	49*	45	gN/m^3
NOX	3.63	3.63*	0.698	0.5*	gN/m^3
DO	0*	0*	8.75	5.92	gO_2/m^3
F	417.6				m^3/h
T	8.3				^oC
$FOV2$			9571		m^3/h
$FOV3$				2897	m^3/h

Table 4. Tuning of of the reduced ASM2d model parameters and resulting correlation coefficient R.

Param.	Nair	Tuned
K_L	38	5.80 &0.18
r_3	112	162
r_1	20.8	113 & 4.34
r_4	171	20.8& 6.09
r_5	17.9	3.11&0.31
r_6	12.9	72.7&48.4
Y_{PO}	0.577	0.125
R	-0.0079	0.0735

3.5 Machine learning model results

The input variables were first dynamically time-delayed using Simulink block "variable time delay". The variables at inlet, including flow (F), temperature (T), sum of nitrates and nitrites ($NOXIN$), soluble organic matter ($CODIN$) and estimated phosphorus ($POIN$) in were delayed using dynamic time delay $Td3$. The total aeration rate in the lumped zone V2 ($FOV2$) was delayed using dynamic time delay $Td2$, and the total aeration rate in lumped zone V3 ($FOV3$) was delayed using dynamic time delay $Td1$. The dynamically time-delayed input variables were collected from Simulink. Both input variables and the output variables were normalized and the data set was divided into two, 50% training and 50% testing.

Support Vector Regression Different regression models were fitted to the dynamically time-delayed training data in Matlab Regression Learner toolbox. The models were compared using the dynamically time-delayed test data set. The results are presented in Table 5 with R^2 model accuracy index obtained from the toolbox. The best results were achieved with support vector regression and gaussian process regression with R^2 values up to 0.9 indicating excellent model fitness. In comparison, linear regression and support vector regression with linear kernel function results in R^2 values around 0.4 indicating insufficient model fitness. The models without time delay were also developed, but these resulted in poorer model accuracy as given in Table 5. On average the model accuracy reduction was 5% for the best performing models, SVR with fine gaussian kernel function and gaussian process. The support vector regression model with fine gaussian kernel function was chosen further for comparison with other models.

Table 5. Regression model type, R^2 model accuracy for train and test data sets with and without (w/o) dynamic time-delay.

Time delay Model	with train	with test	w/o train	w/o test
SVR fine gaussian	0.87	0.86	0.80	0.83
SVR medium gaussian	0.62	0.59	0.58	0.56
SVR cubic	0.54	0.50	0.52	0.48
SVR quadratic	0.50	0.47	0.46	0.43
SVR linear	0.39	0.35	0.40	0.38
Linear regression	0.42	0.38	0.42	0.39
Gaussian process quadratic	0.92	0.92	0.86	0.87

Long-short term memory network The LSTM neural network was implemented in Matlab. The network architecture included: a sequence input layer with 7 features (inputs), 3 LSTM layers with different number of nodes, fully connected layer with 1 response (output) and a regression layer. The initial plant options of the model were chosen as follows: optimizer-Adam, MaxEpochs-40, Mini Batch Size of 1008, Sequence Length of 144, Gradient Threshold of 1, Initial Learn Rate of 0.001, Learn Rate Schedule piecewise, Learn Rate Drop Factor of 0.001, Learn Rate Drop Period of 10. The LSTM model architecture was tested with 3 to 5 LSTM layers with 7-14-28-14-7 nodes. Increasing the depth of the network from 3 LSTM layers to 4 and to 5 LSTM layers decreased the model fit. Increasing the depth of an LSTM network, such as here, from 3 to 5 layers, can lead to issues like overfitting, where the model learns noise and details specific to the training data but fails to generalize. Thus explaining, why the output prediction of

the 5 layer LSTM gave a limited response in the trend figure.

Further, the architecture with 3 LSTM layers was used for testing of the number of nodes. Increasing the number of nodes from 7-14-7 to 14-28-14 and futher to 28-56-28 improved the model fitness both numerically and visually in the trend figure. However, increasing the size to 56-112-56 and 112-224-112 decreased the modeling accuracy. This is because as the model complexity increases, each additional unit of complexity (in this case, more nodes) contributes less to capturing useful, generalizable patterns, and more to fitting random fluctuations in the data. Batch and sequence size were chosen using the daily and weekly patterns of the combined domestic and industrial used water flowrate and nutrient composition. The measurement interval was 6 samples per hour. The batch size was chosen to represent the weekly pattern including 1008 samples and the sequence size was chosen to represent the daily pattern including 144 samples. The batch size was changed from 1008 down to 144 and sequence size was changed from 144 down to 72 which did not improve the results. Decreasing the sequence size further to 144-36 showed limiting the predicted output values in the trend figure. Increasing the learning rate from 0.001 to 0.1 decreased the model accuracy and predicted values in the trend figures were limited over 0. Decreasing the learning rate from 0.001 to 0.0001 decreased the model accuracy. Role of dynamically time-delayed input variables is significant. The model accuracy decreases from 0.71 to 0.49 for training data set and from 0.61 to 0.46 for test data set when input variables are not delayed. This can indicate that the LSTM model with the current architecture does not manage to compensate for the time delays.

Table 6. LSTM parameters: Nodes per layer, Mini Batch Size MBS, Sequence Lenght SL, Initial Learning Rate ILR, correlation coefficients for train and test Rtrain and Rtest. B bad fit to trend data.

Nodes	MBS	SL	ILR	R train	R test
7-14-7	1008	144	0.001	0.61	0.52
7-14-28 -14-7	1008	144	0.001	0.57B	0.48B
7-14 -14-7	1008	144	0.001	0.61	0.51
14-28-14	1008	144	0.001	0.65	0.56
56-112 -56	1008	144	0.001	0.66	0.48
112-224 -112	1008	144	0.001	0.61	0.49
28-56-28	1008	144	0.001	0.71	0.61
28-56-28	1008	144	0.01	0.62 B	0.43 B
28-56-28	1008	144	0.0001	0.61	0.58
28-56-28	144	72	0.001	0.67	0.49B
28-56-28	144	36	0.001	0.68	0.47B
28-56-28 w/o Td	1008	144	0.001	0.49	0.46

3.6 Summary of results

The models are compared visually for training and testing data sets in Figs. 3 and 4, and using the correlation coefficient between the real data and the model predictions in Table 7. The SVR model with highest number of parameters and very little modeling efforts (1 hour) provided excellent modeling accuracy. The LSTM model development and tuning required 10 hours of tuning and provided satisfactory modeling accuracy. The reduced ASM2d model development took over 100 hours of work and resulted in inadequate modeling accuracy. While the SVR model predictions very accurately follow the dynamic trends in the data, the LSTM model gives conservative predictions for a limited output range. SVR often provides more accurate predictions than LSTM when working with small datasets due to its ability to find a hyperplane in a high-dimensional space that best fits the data, minimizing overfitting. LSTMs, which are good in capturing long-range dependencies within large datasets, may struggle with overfitting and underperformance in scenarios with limited data due to their complex architectures. Thus, SVR is typically more suitable and reliable for small-scale data modeling where the primary goal is generalization over capturing sequential patterns.

Likewise, the reduced ASM2d model suffers from large oscillations related to the flowrate term in the mass-balance equation of the nutrient removal. When time-delayed data was used, both ML models improved prediction accuracy, the LSTM for training data 22% and testing data 15%, the SVR for training data 2% and testing data 0%. The ML model prediction accuracy increased when time-delayed data was used.

Table 7. Time consumption and correlation coefficient R for different models.

Model type	Development time [h]	R Train	R Test
SVR fine gaussian	1	0.94	0.95
SVR f.g. w/o Td	1	0.92	0.95
LSTM	10	0.71	0.61
LSTM w/o Td	10	0.49	0.46
rASM2d	100+	0.07	0.07

4. CONCLUSIONS AND FURTHER WORK

Answers to the research questions conclude the results of this study: Can the reduced ASM2d model, SVR model and LSTM model follow the dynamic trends of the effluent PO_4 data? The support vector regression with time delayed variables was very accurate in matching the dynamic trends in the data. The LSTM model had a sufficient fit. The time-delayed variables increased the ML model accuracy.The reduced ASM2d models require more tuning and development to be able to match the dynamic trends in the data. Which model gave the highest prediction accuracy? The SVR model with fine gaussian kernel function and dynamic time-delay gave the best modeling results.

Further work is suggested on control strategy design. The next step is comparison of the unit step responses of the reduced ASM2d model and ML models. If the dynamic responses are similar, the SVR model with highest model accuracy should be used further. If the ML model step responses are not similar to the reduced ASM2d model, a method of online-adaptation of the reduced ASM2d model parameters is suggested to make the model more accurate and suitable for control studies.

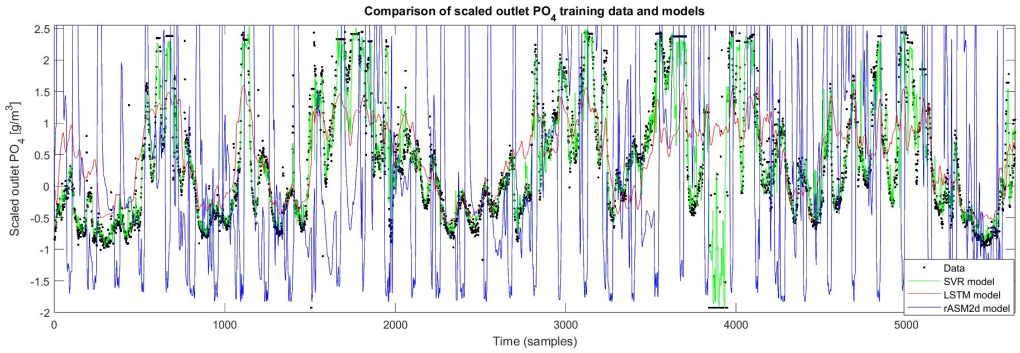

Fig. 3. Training data set. Outlet PO_4 data in black dots, and model predictions with SVR in green, LSTM in red and rASM2d in blue. Time in samples.

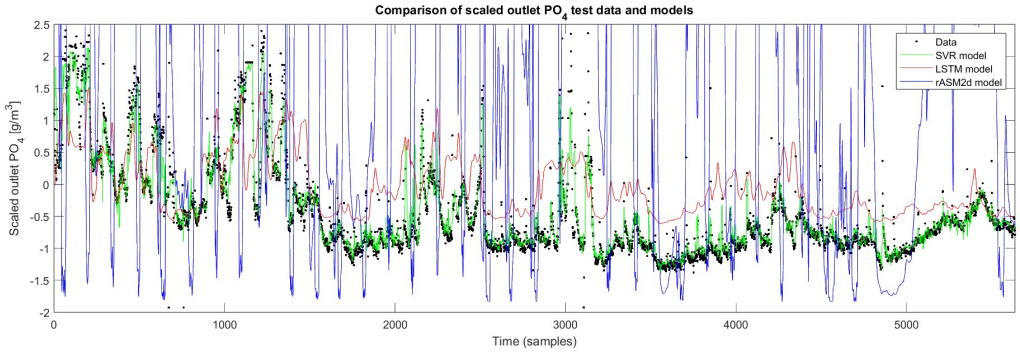

Fig. 4. Testing data set. Outlet PO_4 data in black dots, and model predictions with SVR in green, LSTM in red and rASM2d in blue.Time in samples.

5. CONTRIBUTIONS

Gjermund Sørensen, Katrine Marsteng Jansen (Hias IKS WRRF) and Torgeir Saltnes (Hias How2O) provided valuable insights in the Hias Process. Tobias Korten and Katrine Marsteng Jansen from (Hias IKS), and Axel Tveiten Bech and Ola Solli Grønningsæter (Digitread Connect) collected the data. Professor Tiina Komulainen developed the reduced ASM2d models, input-output design, final development and testing of the ML models, and prepared the article draft. Master student Malik Baqeri developed the data pre-processing methods and initial development of the LSTM and SVR models. Katrine Marsteng Jansen provided facilitated data collection and provided feedback on the research work and manuscript. Professor Arvind Keprate supervised development of the ML models.

REFERENCES

Baqeri, M. (2024). *Machine Learning and Control for Phosphorous Removal*. Master thesis, Oslo Metropolitan University. URL `https://hdl.handle.net/11250/3163016`.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20, 273–297. doi:10.1007/BF00994018.

EuropeanCommission (2021). Eu strategy on energy system integration. URL `https://energy.ec.europa.eu/topics/energy-system-integration/eu-strategy-energy-system-integration_en`.

Henze, M., Gujer, W., Mino, T., Matsuo, T., Wentzel, M.C., v.R. Marais, G., and Loosdrecht, M.C.V. (1999). Activated sludge model no.2d, asm2d. *Water Science Technology*, 39(1), 165–182.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Komulainen, T.M., Baqeri, A.M., Nermo, E., Keprate, A., Saltnes, T., Jansen, K.M., and Korostynska, O. (2023). Estimation of effluent nutrients in municipal mbbr process. In *Proceedings of The 64th International Conference of Scandinavian Simulation Society*. Linköping University Electronic Press, Västerås, Sverige 26-27.9.2023. doi:10.3384/ecp200037.

Komulainen, T.M., Baqeri, M., Jansen, K.M., Saltnes, T., Bech, A.T., and Korostynska, O. (2024). Virtual sensors for the hias process. *Water Practice and Technology*, wpt2024176. doi:10.2166/wpt.2024.176.

Nair, A.M., Fanta, A., Haugen, F.A., and Ratnaweera, H. (2019). Implementing and extended kalman filter for estimating nutrient composition in a sequential batch mbbr pilot plant. *Water Science Technology*, 80(2), 317–328. doi:10.2166/wst.2019.272.

Nermo, E. (2023). *MPC of nutrient removal in wastewater treatment process*. Oslo Metropolitan University. URL `https://hdl.handle.net/11250/3101078`.

Paepae, T., Bokoro, P.N., and Kyamakya, K. (2021). From fully physical to virtual sensing for water quality assessment: A comprehensive review of the relevant state-of-the-art. *Sensors*, 21(21). doi:10.3390/s21216971.

Rudi, K., Goa, I.A., Saltnes, T., Sørensen, G., Angell, I.L., and Eikås, S. (2019). Microbial ecological processes in mbbr biofilms for biological phosphorus removal from wastewater. *Water Science Technology*, 79(8), 1467–1473. doi:10.2166/wst.2019.149.