Hybrid Model Composed of Machine Learning and ASM3 Predicts Performance of Industrial Wastewater Treatment

Boyan Xu[1,2,a], Ching Kwek Pooi[2,a], Tsuey Shan Yeap[2], Kwok Yii Leong[2], Xi Yee Soh[2], Shujuan Huang[2,3,*], Xueqing Shi[3], Giorgio Mannina[4] and How Yong Ng[1,2,*]

1 Center for Water Research, Advanced Institute of Natural Sciences, Beijing Normal University, 519087 Zhuhai, China

2 Centre for Water Research, Department of Civil and Environmental Engineering, National University of Singapore, 117576, Singapore

3 School of Environmental and Municipal Engineering, Qingdao University of Technology, Jialingjiang Road 777, 266520 Qingdao, China

4 Engineering Department, Palermo University, Viale delle Scienze, Ed. 8, 90128, Palermo, Italy

[a] Co-first authors.

[*] Corresponding authors: Shujuan Huang and How Yong Ng

Email address: e0204955@u.nus.edu (Xu Boyan); pooick@hotmail.com (Ching Kwek Pooi); e0056316@u.nus.edu (Shujuan Huang); huanghy@bnu.edu.cn (How Yong Ng)

Abstract

Conventional mechanistic models encounter difficulties in precisely predicting the treatment performance of petrochemical wastewater due to the substantial fluctuations in composition and biodegradability. In this research, a hybrid model was conceptualised, integrating machine learning models with a mechanistic model, Activated Sludge Model 3 (ASM3), to predict the treatment performance of two activated sludge reactors (ASRs) (i.e., ASR-1 and ASR-2) for the treatment of two different types of petrochemical wastewater. Herein, not surprisingly, the prediction performance of the conventional ASM3 based on average wastewater biodegradability of operation was found to be suboptimal, in terms of high MAPE (31%-47%) and low correlation coefficient (0.47-0.50) for mixed liquor suspended solids (MLSS) and chemical oxygen demand (COD). To address this, multivariate linear regression (MLR) and decision forest (DF) models were employed for the assessment and prediction of wastewater biodegradability, which was then integrated with ASM3 (referred to as hybrid models of MLR-ASM3 and DF-ASM3). The hybrid models were trained and subsequently validated using data from ASR-1 (R1), and ASR-2 (R2) which treated different compositions of petrochemical wastewater were utilised for additional validation of the developed hybrid models. Although the predicted capacity of COD by MLR-ASM3 could be accepted for the R1 operation with a correlation coefficient of 0.78, it had a poor predictive capacity for the R2 operation with a correlation coefficient of 0.48. Remarkably, DF-ASM3 exhibited superior prediction performance of both R1 and R2 as compared to MLR-ASM3 for MLSS, COD and nitrate with low MAPEs (all<25%) and high correlation coefficients (all>0.7). This study underscored the effectiveness of hybrid modeling as a strong tool in minimising the labor required for biodegradability testing. Additionally, it demonstrated the improvement of the

prediction capability of hybrid models for the treatment efficiency of petrochemical wastewater compared to mechanistic models.

Keywords: Activated sludge model 3 (ASM3); Hybrid model; Industrial wastewater treatment; Machine Learning; Oxygen uptake rate (OUR)

Nomenclature

ASR	Activated sludge reactor
ASM	Activated sludge model
COD	Chemical oxygen demand
DF	Decision Forest
IC	Ion concentration
MLR	Multivariate linear regression
NH_4^+	Ammonium
OUR	Oxygen uptake rate
sCOD	Soluble chemical oxygen demand
TDS	Total dissolved solids
TN	Total nitrogen
TOC	Total organic carbon
TSS	Total suspended solids
MAPE	Mean absolute percentage error

1. Introduction

The biological treatment of petrochemical wastewater by activated sludge treatment is a formidable task due to its markedly variable characteristics of strength and biodegradability[1]. Specifically, the sensitivity of the response of activated sludge to wastewater composition significantly impacts treatment performance, such as ammonium removal efficiency and organic removal[2]. Moreover, the dynamic concentration of toxic compounds poses a severe threat to treatment plant biomass flocculation[2]. This poses a challenge for mainstream biological treatment plants treating petrochemical wastewater, given that their incoming streams may encompass wastewater from various petrochemical industries. Consequently, the modeling of biological treatment for petrochemical wastewater is challenging. If developed successfully, highly accurate predictive models can aid plant operators by forecasting whether the plants can comply with discharge limits, providing operators ample time to optimize treatment performance.

The increasing interest in the ability of models and simulations to predict treatment plant behavior has spurred exploration into various modeling approaches, including mathematical, mechanistic and fuzzy models, dating back to the 90s[3]. Such modeling and simulation of treatment processes prove invaluable for designing, upgrading and optimising treatment processes[4]. The early construction of the Activated Sludge Model (ASM), initiated in the 1980s by Henze et al.,[5] has been further developed into Activated Sludge Model 3 (ASM3), which closely represents biological processes by incorporating two-step nitrification-denitrification[6] as well as phosphorus removal[7]. ASM3 has also shown success in the modeling and simulating petrochemical wastewater treatment in wastewater treatment plants[8]. However, ASM3 in existing industrial wastewater treatment plants encounters a key challenge: the lack of complete fluctuated quantity and quality data of influent wastewater, especially the labor-intensive wastewater biodegradability which is

cornerstone pre-required information for ASM3 development[9], although some methods based on an empirical relationship such as general patterns[10,11] and phenomenological modeling[12] might compensate for missing data points[8,13]. Hence, simulation and prediction of biodegradability information for ASM3 development can minimize the intensive labor required for biodegradability testing and help accurately forecast the treatment performance of petrochemical wastewater with varying compositions.

Multivariate linear regression (MLR) could establish correlations between independent and dependent wastewater treatment parameters, particularly in small sample sizes[14]. However, previous studies yielded contradictory results for domestic wastewater treatment prediction, with some claiming MLR models' inaccuracy and low correlation coefficient[15,16] . In contrast, others demonstrated a positive linear relationship with R^2 exceeding 0.8 [17,18]. In larger datasets, adopting machine learning stands out as a promising avenue for refining the prediction of wastewater treatment performance. This approach is robust and effective in handling complex, non-linear process parameters with intricate relationships[16,19,20]. Nevertheless, machine learning is often applied as a "black-box" method, concealing the underlying physics of the model process and lacking extrapolative capability. Additionally, industrial wastewater treatment is characterized by high fluctuation of compositions, which can lead pure machine learning models to overestimate and underestimate the treatment performance of activated sludge. Considering these challenges, a hybrid model, integrating machine learning with a mechanistic model, presents high prediction potential[21–23]. The machine learning model at the initial stage of the hybrid model could be a valuable approach to capture the variations in biodegradability information of petrochemical wastewater, and it could further complement mechanistic models to improve ASM3 prediction performance.

Currently, studies on modeling the treatment of petrochemical wastewater are very few compared to those on municipal wastewater treatment. This study introduces a serial hybrid model as an advanced predictive tool aimed at significantly improving the accuracy of forecasting performance of ASM3 for the treatment of highly variable petrochemical wastewater by activated sludge reactors (ASRs) (**Figure 1a**). Linear regression or machine learning was used to estimate biodegradability. The resulting biodegradability estimates, alongside influent and sludge characteristics, were then incorporated into the ASM3 to establish two hybrid models, namely, MLR-ASM3 and DF-ASM3. The training and subsequent validation of this hybrid model were conducted using ASR-1 (R1) data. Additionally, ASR-2 (R2), characterised by a different composition of petrochemical wastewater as an out-of-distribution dataset (**Figure 1c**), was employed for further validation of the developed model of MLR-ASM3 and DF-ASM3. The prediction performance, benefits, drawbacks and potential applications of hybrid models compared to traditional ASM3 are then presented and discussed.

2. Materials and Methods

2.1. Petrochemical wastewater samples

Petrochemical wastewater was systematically collected on a weekly basis from the discharge points of three distinct local petrochemical industrial plants in Singapore — specifically, Companies 1, 2 and 3. These plants were strategically chosen for their geographical separation and employed diverse production processes, resulting in different wastewaters with varying characteristics and volumes. To ensure comprehensive homogenization, the collected wastewater was combined in a 300 L barrel and distributed in 20 L jerrycans by the specific ratios required by each reactor. The particular ratios of wastewater distribution to each reactor and their

corresponding characteristics, are outlined in **Table 1**. To standardize the pH level, acidic (HCl) and alkaline (NaOH) solutions were utilized to adjust each jerrycan's pH to a target value of 7.2. Given the absence of phosphorus content in the wastewater, potassium dihydrogen phosphate was introduced into each jerrycan to address nutrient deficiencies, maintaining a consistent C:P ratio of 100:1. Subsequently, these prepared jerrycans were stored in a cold room at 4°C, only to be returned to room temperature before usage. The proportions of wastewater allocated to each reactor were determined based on the respective companies' daily output flow rates.

2.2. Lab-scale reactors set up

The activated sludge process was adopted as the primary biological treatment for the petrochemical wastewaters and used to validate the developed predictive models. Two parallel laboratory-scale cylindrical ASRs, denoted as R1 and R2, were set up for wastewater treatment (**Figure 1b**). Each ASR featured a working volume of 6 L and a hydraulic retention time (HRT) of 18 h. The reactors were operated with a solid retention time (SRT) of 20 d. Air diffuser plates were strategically positioned at the base of each reactor for aeration and mixing. The influent wastewater was introduced into the ASR through a peristaltic pump at a controlled rate of 5.56 mL/min. Subsequently, the treated effluent from the ASR was directed downstream to the 4-L clarifier through a gravity overflow system. Settled sludge collected at the clarifier's bottom was recirculated back into the ASR at 11.1 mL/min using a peristaltic pump. The pH of reactors R1 and R2 were maintained between 7 to 8 by pH dosing pumps. The reactors were operated for 50 d (more than 2 SRTs) for acclimatization before experimental results were used for modeling in this study.

2.3. Water quality analysis

Samples of petrochemical wastewater, reactor biomass and treated effluent were collected and analyzed for water quality. The effluent sample was collected a day after the influent sample was collected. Souble portion of the sample was obtained by filtering with a 0.45-µm filter (Supor 450 membrane disc filter, Pall Corporation, USA). Total organic carbon (TOC) and total nitrogen (TN) were measured using a total organic carbon analyzer (TOC-VCSH, Shimadzu, Japan). Chemical oxygen demand (COD) was made using the closed reflux method in accordance with the Standard Methods for the Examination of Water and Wastewater[24]. NH_4^+ was quantified using the salicylate method (TNTplusTM 832, Hach, USA), while NO_2^- and NO_3^- were quantified using the colorimetric method in accordance with the Standard Methods for the Examination of Water and Wastewater[24]. Mixed liquor suspended solids (MLSS), total dissolved solids (TDS) and total suspended solids (TSS) were done in accordance with the Standard Methods for the Examination of Water and Wastewater with prepared filter papers (Whatman GF/F, Merck, Germany) and ceramic crucible [24]. Ion concentration was measured using ion chromatography (Dionex ICS-1600, ThermoFisher Scientific Inc., U.S.A.).

2.4. Kinetic and stoichiometry of sludge and wastewater biodegradability

ASM3 has been used widely for the simulation of domestic wastewater treatment. However, Kinetics and stoichiometry of biomass require further calibration for petrochemical wastewater treatment using respirometry (**Table 2** and **Figure 2**). The respirometer for these evaluations was a 1-L reactor equipped with a bottom stone aerator (BM-Advance, Surcis, Denmark). The mixed liquor underwent continuous agitation via a mechanical stirrer throughout the tests, and the temperature was maintained at 25°C. Preceding the experiments, biomass from Reactor 1 was

extracted and subjected to aeration for 24 hours to attain an endogenous state. Subsequently, the biomass underwent a triple washing process with deionized water to eliminate external COD beyond the sludge. Allylthiourea was introduced before each test to inhibit ammonium removal. Notably, anoxic tests were omitted as both reactors were consistently operated in aerobic mode. Moreover, the biodegradability of the influent wastewater was determined by oxygen uptake rate (OUR) through the respirometry analyses. The decay rate was derived from the daily measurement of the endogenous OUR over a 5-d period. A comprehensive total of 36 respirometry analyses were conducted.

2.5. Development and assessment of ASM3 and hybrid model

For conventional ASM3 development, the average biodegradability (the average value of all the measured biodegradability), biomass kinetic and stoichiometry parameters, together with the influent characteristics, were fed into the ASM3 to simulate the treatment performance of the ASRs. The simulation using average biodegradability for ASM3 was set as the benchmark for the predictive models. ASM3 modeling was coded on Python with the Pandas package installed. KEuler method was used to find the numerical approximation to the differential equations of ASM3. The assumptions used in the simulations are as follows:

1. All reactors were assumed to be well-mixed, with microorganisms to be the same for all reactors.
2. The reactors were assumed to be well-aerated.
3. The kinetic and stoichiometry parameters of R-1 and R-2 were assumed to be the same.
4. The kinetic and stoichiometry parameters were constant throughout the study.

5. Based on historical observed values, the solid removal efficiencies of the clarifiers of R1 and R2 were 95 % and 97 %, respectively.
6. Biodegradable organics in the wastewater were all assumed to be readily biodegradable.
7. The endogenous OUR profiles for the unfiltered and filtered sample are the same, hence, the particulates are non-biodegradable and inert.
8. To improve the solids simulation of the reactor, the COD/SS ratio for the influent solids used was calculated by dividing the influent particulate COD concentration by the influent SS concentration.
9. The pH of R1 and R2 were maintained between 7 to 8 by pH dosing pumps. This allowed for an excess concentration of alkalinity during modeling. The value used was 8 mmol HCO_3^- /L. However, during Day 0 – Day 30 and Day 105 – 115, the pH pump of R1 was down. Thus, the alkalinity used in the model for R1 during that period was the measured value of 3 mmol HCO_3^- /L.

In the development of the hybrid model (**Figure 1a**), the biodegradability data underwent training and prediction by using multivariate linear regression (MLR) or decision forest (DF), which were then integrated into ASM3, referred to as MLR-ASM3 and DF-ASM3, respectively. Both MLR-ASM3 and DF-ASM3 were applied to forecast crucial parameters such as MLSS, COD and nitrate. The multivariate linear regression was conducted through the regression function in Microsoft Excel's data analysis tools. For machine learning, the Microsoft Azure Machine Learning Studio was employed. The DF regression model correlated the non-biodegradable soluble organics fraction with influent characteristics. The decision forest was done using a replicate resampling method with 20 decision trees. Parameters for the decision forest model comprised a maximum depth of 32 layers per tree and a minimum of 1 sample per leaf node. The

training of the hybrid models and the subsequent validation were conducted using data from R1. R2, featuring a different composition of petrochemical wastewater as out-of-distribution dataset was, employed for further validation of the developed hybrid models (**Figure 1c**).

Statistical analysis was conducted to compare the simulated and measured results. Correlation, mean absolute percentage error (MAPE), root mean square error (RMSE) and normalized RMSE were employed to assess the accuracy of the model. The equations for calculating correlation, MAPE, RMSE and normalized RMSE are provided as follows:

$$Correlation, r = \frac{cov(y,\bar{y})}{\sigma_Y \sigma_{\bar{Y}}} \quad (1)$$

$$MAPE = \frac{absolute(\sum_{i=1}^{N} y_i - \bar{y}_i)}{N} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}{N}} \quad (3)$$

$$Normalized\ RMSE = \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}{N}}}{maximum\ value\ in\ the\ training\ dataset} \quad (4)$$

where *cov* is the covariance, y and $\bar{y}$ are measured and predicted variables, respectively, σ is the standard deviation and N is the number of samples.

3. Results and discussion

3.1. Treatment performance of petrochemical wastewater

The COD, sCOD and TOC concentrations in the influent of both R1 and R2 had standard deviation values of approximately 50% of their corresponding mean values (**Table 1**). A similar trend was observed regarding the nitrogenous content. The influent nitrogen content in R1

fluctuated from 15 mg/L to 48 mg/L, while in R2, it fluctuated from 10 mg/L to 46 mg/L. The carbonaceous organic concentrations in R1 and R2 were comparable, with average soluble COD (sCOD) concentrations of 306 mg/L and 313 mg/L, respectively (**Table 1**). However, the TN concentrations in R1 were about 50% higher than those in R2. Additionally, the ammonium concentrations in R1 exceeded those in R2, indicating a higher nitrogen content in the wastewater obtained from Company 3. This variability was expected in petrochemical wastewater due to different processes yielding distinct wastewater characteristics. It underscored the presence of a unique composition in influent petrochemical wastewater between R1 and R2.

The treatment performance of R1 and R2 alongside their corresponding simulated outcomes are presented in **Table 3**, **Figures 3a and 3b**. Although R1 had COD concentrations similar to municipal wastewater, it only attained an average COD removal of 65%. In contrast, R2 demonstrated slightly better organic removal, with an average COD removal of 72%. Of note, the diverse characteristics of petrochemical wastewater variables led to variations in biodegradability, impacting the treatment performance of two ASRs. Hence, the COD removal efficiencies of ASRs fluctuated significantly, ranging from 17% to 87% and 24% to 90% for R1 and R2, respectively. In view of the suboptimal ASR treatment performance, additional downstream treatments such as advanced oxidation processes or activated carbon adsorption may be necessary.

Solids content is important in the context of petrochemical wastewater treatment, where analyzing the solids content in the ASRs provides insights into how the wastewater influences microbial activities. **Table 4** displays the solids content, which is generally within an acceptable range, and simulation of solids could be done with the ASM3 and hybrid models (**Figure 4**), which will be discussed in the following section.

The measured NO_3^--N concentration, simulated NO_3^--N concentration and ammonium removal attributable to nitrification in R1 and R2 are shown in **Figure 5**. Notably, during the initial period with low influent ammonium concentration (Day 0–10), R1 achieved 100% ammonium removal with an alkalinity of 3 mmol HCO_3^-/L. However, as influent ammonium concentration increased, the ammonium removal declined, primarily due to inadequate alkalinity in the wastewater leading to a pH decrease[25]. For example, the pH of R1 dropped to approximately 5.5 during Day 10–30 and Day 105–115, resulting in a reduced ammonium removal, with the lowest ammonium removal recorded at 18% on Day 131. Similar observations were noted during Day 105–115 and Day 122–131. Notably, ammonium removal due to nitrification improved when pH was regulated. Both reactors achieved 100% ammonium removal when the pH was maintained between 7 and 8.

3.2. Prediction of organics removal of petrochemical wastewater

The kinetic and stoichiometry parameters determined by the 36 respirometry analyses using different batches of R1's influent wastewater are shown in Error! Reference source not found.. The fitted exogenous graphs treating 2 representative different batches (Day 1 and Day 7) of R1's influent wastewater using the determined kinetic and stoichiometry parameters are shown in **Figure 2a** and **2b**. Despite different respirometer analyses with different F/M ratios and different batches of wastewater, the kinetic and stoichiometry parameters did not differ much. The kinetic and stoichiometry parameters are similar to the reference values except for b_H, K_S and K_{STO}. To obtain a good fit (**Figure 2c**), b_H had to be reduced to 0.15 d^{-1}, with the K_S value increased to 6.0 g COD/m^3 and K_{STO} value decreased to 0.2 g COD/g COD. The high K_S value indicates that the maximum rate of substrate removal was lower than that of domestic wastewater. This suggests

that the organics in the petrochemical wastewater were more difficult to treat than those of domestic wastewater. The b_H value of the biomass is also lower than that of domestic activated sludge, indicating their lower activity. This is expected, as the HRT and SRT of the system in this study are significantly longer than ASRs for domestic wastewater treatment.

Using the average biodegradability obtained from the respirometry method is common for the simulation of wastewater treatment. Thus, in this study, the simulation using average biodegradability was set as the benchmark for the predictive models. After obtaining the average biodegradability results from the first 50 d of operation of R1, the mechanistic model was then used to predict the treatment performance of R1. The statistical results of the measured and simulated parameters are tabulated in Error! Reference source not found.. The COD removal simulation yielded acceptable results, with MAPE of 30.8% and normalized RMSE of 0.33. However, simulation using average biodegradability is not recommended, as it is unable to accurately predict the removal trend with a correlation coefficient of 0.5. This was mostly due to the fluctuation in influent wastewater characteristics, which could not be simulated using average biodegradability values. From Day 52 to Day 70, the measured COD removal efficiencies were consistently below 50% (Error! Reference source not found.**a**). The predicted simulation results, however, were still maintained at ~65%. This resulted in an underestimation of effluent COD concentration, which may lead to the plant's inability to meet the discharge limit.

Multivariate linear regression for prediction of biodegradability was constructed before the development of MLR-ASM3. As aforementioned, the MLR model development was based on the data obtained from the first 50 d of operation of R1. The MLR equation to determine the inert fraction of COD using the Excel data analysis regression function is shown in Eq. 5 as follows.

$$Inert\ fraction\ of\ COD = \frac{9.060 + 0.062[COD] + 0.091\ [COD_S] + 0.629\ [TOC] - 1.213\ [TN] + 1.426\ [NH_4{}^+ - N] - 0.024\ [Br^-]}{COD} \quad (5)$$

By correlating the inert fraction to the influent characteristic, the relationship between biodegradability information could be modeled by MLR. As a result, the simulation of MLR-ASM3 yielded better prediction performance as compared to the simulation using average biodegradability (**Figure 2a and 3c**). Although it had similar RMSE value and MAPE results as compared to ASM3 (Error! Reference source not found.), the correlation coefficient was much higher with a value of 0.78 by using MLR-ASM3 as compared to 0.5 achieved by using average biodegradability. During Days 50–70, the MLR-ASM3 model was able to predict the low COD removal, despite low COD removal data not being used in the determination of the regression model (Error! Reference source not found.**a**). Of note, TOC has a significant weightage in Eq. 5 with a coefficient of 0.63, which suggests that COD/TOC could played an important role in COD removal of ASRs. Indeed, when the COD/TOC ratio of the wastewater was low, the inert fraction of COD increased according to Eq. 5, which thus could result in a lower COD removal. From the perspective of mechanisms of activated sludge treatment, a low COD/TOC ratio suggests that the organics are non-oxidizable, as a result, the ASR performance would decrease. Hence, the results of the MLR model could aid in suggesting that the low COD removal during Day 50–70 of R1 was attributed to the low COD/TOC ratio (data not shown).

The MLR-ASM3 model, employing the MLR equation derived from the data of R1, was utilized to forecast the performance of R2 (**Figure 2b**). The correlation coefficient between measured COD removal and simulated COD removal was 0.48 with a MAPE value of 35.5%

(Error! Reference source not found.). The reason for the poor prediction is most likely due to the difference in wastewater characteristics between R1 and R2 (**Figure 2d**). The application of MLR-ASM3 for predicting petrochemical wastewater treatment performance appears to be limited to the reactor on which the hybrid model was trained. Additionally, there is a possibility that biodegradability has a non-linear relationship with influent characteristics of R2.

The hybrid model could be further improved by replacing the MLR with a machine-learning model. DF-ASM3 model reduced R1's MAPE from 29.6% to 25.3% and increased the correlation coefficient from 0.78 to 0.83, as compared to the MLR-ASM3 model (**Table 4**). The simulated COD removal closely followed the measured COD removal (Error! Reference source not found.**a**). As seen in Day 15–25, the MLR-ASM3 model was unable to simulate the fluctuation in COD removal, but the DF-ASM3 model followed the measured removal closely. Of note, the DF-ASM3 model improved the simulation of R2 significantly (Error! Reference source not found.**b**). When predicting R2's performance, DF-ASM3 reduced the MAPE from 35.5% to 22.2% and improved the correlation coefficient from 0.48 to 0.74, as compared to the MLR-ASM3 model. Similar to R1, the DF-ASM3 model was able to follow the measured COD removal closely in R2. In addition, as seen in Day 33 of R2, the DF-ASM3 model was able to accurately predict the low COD removal, a significant improvement over the MLR-ASM3 model (Error! Reference source not found.**b**). A possible reason is that the machine learning model could establish the non-linear relationship between biodegradability and influent characteristics. As a result, the machine learning is able to predict the biodegradability of R2's influent wastewater (Error! Reference source not found.**d**). This also shows that the machine learning model was successful in correlating the biodegradability to the influent characteristics of various types of petrochemical wastewater.

3.3 Prediction of concentration of solids in ASRs

While it is crucial to predict the COD content in the effluent, it is equally vital not to neglect simulating the solids content. Precisely, simulating solids content can also serve as a reliable indicator of the reactor's performance. This enables users to assess whether the reactor is capable of handling shock-loading scenarios effectively. Herein, the simulated values of conventional ASM3 differed significantly from the measured reactor's MLSS (**Table 4**). Hence, it had a high MAPE of 47% and a low correlation coefficient of 0.47. Both the MLR-ASM3 model and DF-ASM3 model were able to significantly improve the prediction capacity of the reactors' solids concentration. The correlation coefficients of solids simulation using both models for both R1 and R2 were 0.79 and 0.73, respectively. The reactor solids concentrations and simulated solids concentrations of R1 and R2 were further plotted in Error! Reference source not found.**a and 4b**. It is worth noting that both models had very similar solids simulation results. The similar solids simulation results between the MLR-ASM3 model and the DF-ASM3 model could mainly be attributed to the solids removal rate, and the kinetic and stoichiometry parameters used in both hybrid models were similar. Overall, MLR-ASM3 and DF-ASM3 showed better prediction capacities of MLSS of ASRs treating petrochemical wastewater as compared to conventional ASM3.

3.4 Prediction of effluent nitrate of ASRs

Simulation of effluent nitrate concentrations of R1 achieved a MAPE of 22.9% by using ASM3 (**Table 4**). R1's correlation coefficient by using ASM3, however, was only 0.67. MLR-ASM3 and DF-ASM3 could achieve simulation of effluent nitrate concentration of R1 with MAPEs of 21.5% and 19.0%, respectively. The correlation coefficients were only slightly

improved by both hybrid models as compared to ASM3. This suggests that better prediction of biodegradability by using MLR or DF played a minor role in effluent nitrate concentration prediction. The possible reason for the low prediction capacity of all three models could be mainly due to the pH fluctuation of R1, which affected the nitrification. For example, when pH control failed from Day 105-115 for R1, the ammonium removal by nitrification fluctuated from 17% to 96% from Day 100–130 (Error! Reference source not found.**a**), and models were unable to simulate the change, thus resulting in a low correlation coefficient value. On the other hand, both hybrid models had good results in the effluent nitrate simulation for R2 with MAPEs less than 20%, normalized RMSE less than 0.2 and a correlation coefficient of more than 0.8 (**Table 4**). This is most likely attributed to the well pH control of R2, achieving 100% ammonium removal by nitrification most of the time (Error! Reference source not found.**b**). Overall, the results suggest that the prediction performance of effluent nitrate largely relies on the content of alkalinity of wastewater instead of the model types. Hence, nitrate simulation would be better if the alkalinity was constantly maintained. Alternatively, the simulation would improve with proper pH control in the ASRs, ensuring the ASRs had sufficient alkalinity for complete ammonium removal during petrochemical wastewater treatment.

While the primary function of the model is to simulate treatment performance (effluent nitrate here), it also could assess the status of the pH controllers of ASRs. As observed in R1, when the pH controller was operational, the simulated nitrate concentration closely mirrored the measured nitrate concentration (Error! Reference source not found.**a**). Conversely, during Days 100–130, a significant disparity between the simulated and measured nitrate concentrations suggested a malfunction in the pH controller. This positioned the model as a potential virtual sensor for monitoring the condition of the pH controllers.

4. Conclusion

Three distinct strategies were adopted in developing predictive models to forecast petrochemical wastewater treatment by ASRs: the mechanistic ASM3, MLR-ASM3 and DF-ASM3. The mechanistic model exhibited low MAPE, low normalized RMSE and hard to accurately predict COD removal of ASRs due to significant fluctuations in influent wastewater characteristics. The hybrid model MLR-ASM3, while able to capture the trend of COD removal in R1, fell short in predicting R2's COD removal. It suggests that the application of MLR-ASM3 for predicting petrochemical wastewater treatment performance lacks extrapolative capability. Furthermore, the hybrid model was enhanced by using the Decision Forest model followed by ASM3, yielding the most accurate prediction of COD removal. This suggests that the machine learning model could capture non-linear information from residuals, compensating for the mechanistic model's inaccuracies and thereby enhancing the hybrid model's extrapolation ability to reactors treating diverse compositions of petrochemical wastewater. Furthermore, we found that both hybrid models showed better prediction capacities of MLSS of ASRs as compared to mechanistic ASM3. Moreover, hybrid models not only could simulate the nitrate concentration of ASRs but can be potentially serve as soft sensors for monitoring the condition of the pH controllers. Hence, this lab-scale study demonstrated the potential of using hybrid modeling as a digital tool to effectively simulate the performance of full-scale petrochemical wastewater treatment plants under the variations of influent loading. Nevertheless, model interpretation (i.e., the understanding of hidden kinetics) is still a concern. One potential approach involves more detailed influent characteristics such as omics data to infer the hidden kinetics and further improve the hybrid model

accuracy. This refinement might help to ultimately transform the machine-learning-based models from "black boxes" to "white boxes".

Additional information

The authors declare no competing interests.

Data availability

Data will be made available on request.

References

(1) Zaffaroni, C.; Daigger, G.; Nicol, P.; Lee, T. W. Wastewater Treatment Challenges Faced by the Petrochemical and Refinery Industry, and Opportunities for Water Reuse. *Water Pract. Technol.* **2016**, *11* (1), 104–117. https://doi.org/10.2166/wpt.2016.012.

(2) Zou, S.; Yan, N.; Zhang, C.; Zhou, Y.; Wu, X.; Wang, J.; Liu, Y.; Zhang, Y.; Rittmann, B. E. Acclimation of Nitrifying Biomass to Phenol Leads to Persistent Resistance to Inhibition. **2019**. https://doi.org/10.1016/j.scitotenv.2019.133622.

(3) Asthana, P.; Hazela, B. Applications of Machine Learning in Improving Learning Environment; 2020; pp 417–433. https://doi.org/10.1007/978-981-13-8759-3_16.

(4) Drewnowski, J. Advanced Supervisory Control System Implemented at Full-Scale WWTP-A Case Study of Optimization and Energy Balance Improvement. *Water (Switzerland)* **2019**, *11* (6). https://doi.org/10.3390/w11061218.

(5) Henze, M.; Design, I. T. G. for M. M. for; Treatment, O. of B. W. *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*; IWA Publishing: London, 2000; Vol. no. 9.

(6) Iacopozzi, I.; Innocenti, V.; Marsili-Libelli, S.; Giusti, E. A Modified Activated Sludge Model No. 3 (ASM3) with Two-Step Nitrification-Denitrification. *Environ. Model. Softw.* **2007**, *22* (6), 847–861. https://doi.org/10.1016/j.envsoft.2006.05.009.

(7) Rieger, L.; Koch, G.; Kühni, M.; Gujer, W.; Siegrist, H. The Eawag Bio-p Module for Activated Sludge Model No. 3. *Water Res.* **2001**, *35* (16), 3887–3903. https://doi.org/10.1016/S0043-1354(01)00110-5.

(8) Melcer, H.; Dold, P. L.; Jones, R. M.; Bye, C. M.; Takacs, I.; Stensel, H. D.; Wilson, A. W.; Sun, P.; Bury, S. *Methods for Wastewater Characterization in Activated Sludge Modeling*; 2003.

(9) Rieger, L.; Takács, I.; Villez, K.; Siegrist, H.; Lessard, P.; Vanrolleghem, P. A.; Comeau, Y. Data Reconciliation for Wastewater Treatment Plant Simulation Studies-Planning for High-Quality Data and Typical Sources of Errors. *Water Environ. Res.* **2010**, *82* (5), 426–433. https://doi.org/10.2175/106143009X12529484815511.

(10) De Keyser, W.; Gevaert, V.; Verdonck, F.; De Baets, B.; Benedetti, L. An Emission Time Series Generator for Pollutant Release Modelling in Urban Areas. *Environ. Model. Softw.* **2010**, *25* (4), 554–561. https://doi.org/10.1016/J.ENVSOFT.2009.09.009.

(11) Harmel, R. D.; Smith, D. R.; King, K. W.; Slade, R. M.; Smith, P. *Data Uncertainty Estimation Tool for Hydrology and Water Quality (DUET-H/WQ): Estimating Measurement Uncertainty for Monitoring and Modelling Applications*; 2008.

(12) Butler, D.; Friedler, E.; Gatt, K. Characterising the Quantity and Quality of Domestic Wastewater Inflows. *Water Sci. Technol.* **1995**, *31* (7), 13–24. https://doi.org/10.1016/0273-1223(95)00318-H.

(13) Trojanowicz, K.; Styka, W.; Baczynski, T. Experimental Determination of Kinetic Parameters for Heterotrophic Microorganisms in Biofilm under Petrochemical Wastewater Conditions. *Polish J. Environ. Stud.* **2009**, *18* (5), 913–921.

(14) Razi, M. A.; Athappilly, K. A Comparative Predictive Analysis of Neural Networks (NNs), Nonlinear Regression and Classification and Regression Tree (CART) Models. *Expert Syst. Appl.* **2005**, *29*, 65–74. https://doi.org/10.1016/j.eswa.2005.01.006.

(15) Hamada, M.; Zaqoot, H. A.; Jreiban, A. A. Application of Artificial Neural Networks for the Prediction of Gaza Wastewater Treatment Plant Performance-Gaza Strip. *J. Appl. Res. Water Wastewater* **2018**, *9* (1), 399–406.

(16) Zare Abyaneh, H. Evaluation of Multivariate Linear Regression and Artificial Neural Networks in Prediction of Water Quality Parameters. *J. Environ. Heal. Sci. Eng.* **2014**, *12* (1), 40. https://doi.org/10.1186/2052-336X-12-40.

(17) Lee, S.-P.; Min, S.-Y.; Kim, J.-S.; Park, J.-U.; Kim, M.-S.; Lee, S.-P.; Min, S.-Y.; Kim, J.-S.; Park, J.-U.; Kim, M.-S. A Study on the Influence of a Sewage Treatment Plant's Operational Parameters Using the Multiple Regression Analysis Model. *Environ. Eng. Res.* **2014**, *19* (1), 31–36. https://doi.org/10.4491/eer.2014.19.1.031.

(18) Razif, M.; Yanuwiadi, B.; Rachmansyah, A.; Fajarindra Belgiawan, P. Implementation of Regression Linear Method to Predict WWTP Cost for EIA: Case Study of Ten Malls in Surabaya City. *Procedia Environ. Sci.* **2015**, *28*, 158–165.

https://doi.org/10.1016/j.proenv.2015.07.022.

(19) Guo, H.; Jeong, K.; Lim, J.; Jo, J.; Kim, Y. M.; Park, J. pyo; Kim, J. H.; Cho, K. H. Prediction of Effluent Concentration in a Wastewater Treatment Plant Using Machine Learning Models. *J. Environ. Sci. (China)* **2015**, *32*, 90–101. https://doi.org/10.1016/j.jes.2015.01.007.

(20) Hamed, M. M.; Khalafallah, M. G.; Hassanien, E. A. Prediction of Wastewater Treatment Plant Performance Using Artificial Neural Networks. *Environ. Model. Softw.* **2004**, *19* (10), 919–928. https://doi.org/10.1016/j.envsoft.2003.10.005.

(21) Anderson, J. S.; Mcavoy, T. J.; Hao, O. J. Use of Hybrid Models in Wastewater Systems. **2000**. https://doi.org/10.1021/ie990557r.

(22) Lee, D. S.; Vanrolleghem, P. A.; Park, J. M. Parallel Hybrid Modelling Methods for a Full-Scale Cokes Wastewater Treatment Plant. In *IWA Conference on Environmental Biotechnology*; 2003.

(23) Pooi, C. K.; Loka, V.; Ng, H. Y. Treatment and Hybrid Modeling of Domestic Reverse Osmosis Concentrate Using Biological Activated Carbon. *Desalination* **2019**, *468* (October), 114047. https://doi.org/10.1016/j.desal.2019.06.013.

(24) APHA. *Standard Methods for the Examination of Water and Wastewater*; 2012. https://doi.org/ISBN 9780875532356.

(25) Tarre, S.; Green, M. High-Rate Nitrification at Low PH in Suspended- and Attached-Biomass Reactors. **2004**, *70* (11), 6481–6487. https://doi.org/10.1128/AEM.70.11.6481.

Table 1. Characteristic and ratio of wastewater fed to reactor 1 and reactor 2

Parameter	Reactor 1 (R1)	Reactor 2 (R2)
Company 1: 2: 3	2: 1: 1	4: 2: 1
COD (mg/L)	357 ± 213	395 ± 215
sCOD (mg/L)	306 ± 166	313 ± 188
TOC (mg/L)	63.7 ± 31.8	75.2 ± 29.0

TN (mg/L)	48.7 ± 22.1	32.2 ± 11.6
NH_4^+-N (mg/L)	40.5 ± 16.8	24.4 ± 12.2
NO_2^--N (mg/L)	0.0 ± 0.0	0.0 ± 0.0
NO_3^--N (mg/L)	4.2 ± 5.8*	3.2 ± 3.4*
TSS (mg/L)	207 ± 280	209 ± 192
TDS (mg/L)	5572 ± 2600	5320 ± 1586
COD/SS of solids	0.580 ± 0.857	0.567 ± 0.483
Alkalinity (mmol/L)	3.0	6.0

* The standard deviation value is larger than the mean value due to the strong positive skewness.

Table 2. Kinetic and stoichiometric parameters of heterotroph used in this study.

Parameters	**Reference values at 20 °C**	**Best fit values at 25 °C**
Heterotroph max. growth rate, μ_H (d^{-1})	2.0	2.0

Aerobic yield of heterotrophic biomass, Y_H (g COD/ g COD)	0.63	0.6
Aerobic yield of stored product per S_S, $Y_{STO,O2}$ (g COD/g COD)	0.85	0.83
Heterotroph decay rate, b_H (d^{-1})	0.2	0.15
Storage rate constant, k_{STO} (g COD/ g COD · d)	5.0	5.0
Substrate saturation constant, K_s (g COD/m^3)	2.0	6.0
X_{STO} saturation constant, K_{STO} (g COD/ g COD)	1.0	0.2

Table 3. Treatment performance of reactor R1 and reactor R2.

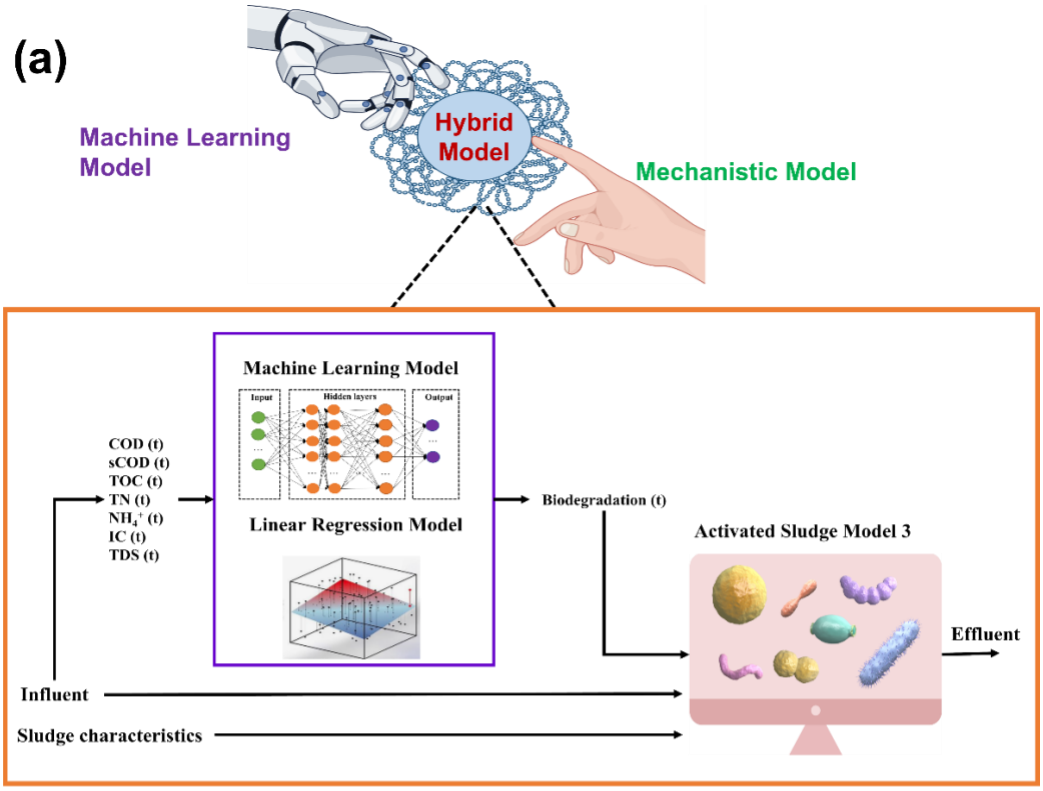
	Reactor R1			**Reactor R2**		
	Min.	**Max.**	**Average**	**Min.**	**Max.**	**Average**
COD removal (%)	17.3	87.3	65.2	23.9	90.2	71.7
TOC removal (%)	23.8	91.6	68.4	26.1	91.4	69.3
Ammonium removal due to nitrification (%)	17.3	100	81.5	39.4	100	97.6

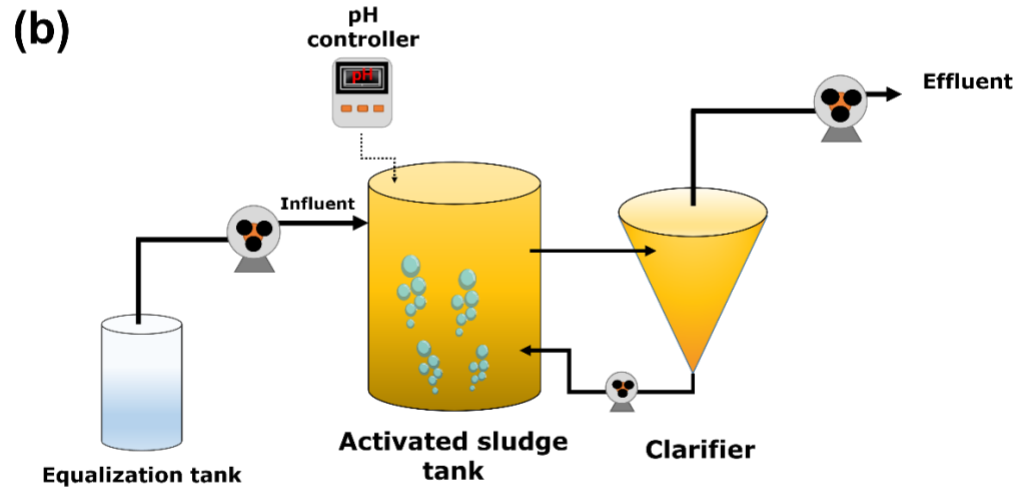
MLSS (mg/L)	830	2680	1488	1840	4660	3245

Table 4. Statistical results of average biodegradability model, MLR-ASM3 model & DF-ASM3 model for Reactor R1 & R2.

Model	**Reactor R1**			**Reactor R2**	
	Average biodegradability	MLR-ASM3 Model	DF-ASM3 Model	MLR-ASM3 Model	DF-ASM3 Model
COD removal					
MAPE (%)	30.8	29.6	25.3	35.5	22.2
RMSE	30.2	30.7	27.4	28.6	24.1

Normalized RMSE	0.33	0.34	0.30	0.32	0.27
Correlation coefficient	0.50	0.78	0.83	0.48	0.74
Reactor's MLSS concentration					
MAPE (%)	47	24.4	24.7	15.4	15.4
RMSE	830	397	412	605	610
Normalized RMSE	0.56	0.27	0.28	0.19	0.19
Correlation coefficient	0.47	0.79	0.79	0.73	0.73
Effluent NO_3^--N concentration					
MAPE (%)	22.9	21.5	19.0	16.4	16.7
RMSE	8.59	8.40	8.44	5.63	5.61
Normalized RMSE	0.29	0.28	0.28	0.18	0.18
Correlation coefficient	0.67	0.69	0.70	0.83	0.83


(a)
Machine Learning Model
Hybrid Model
Mechanistic Model
Machine Learning Model
Input
Hidden layers
Output
Linear Regression Model
COD (t)
sCOD (t)
TOC (t)
TN (t)
NH_4^+ (t)
IC (t)
TDS (t)
Biodegradation (t)
Activated Sludge Model 3
Effluent
Influent
Sludge characteristics


(b)
pH controller
Effluent
Influent
Equalization tank
Activated sludge tank
Clarifier

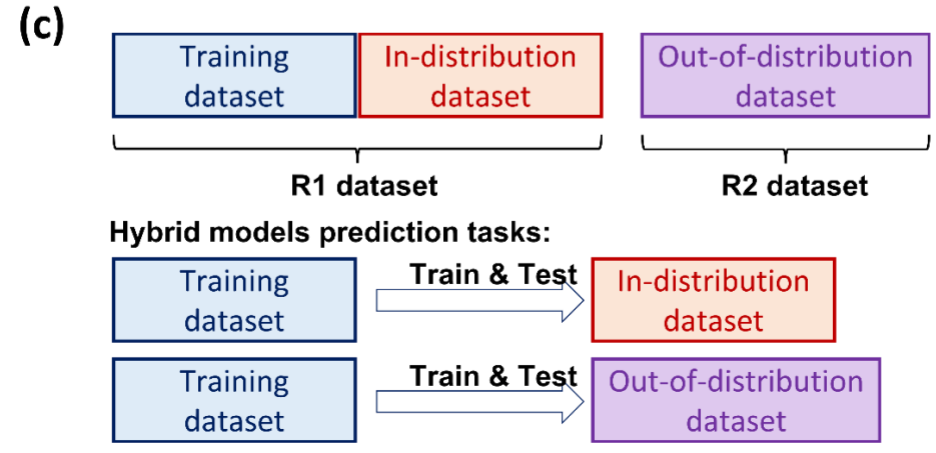

(c)
Training dataset
In-distribution dataset
Out-of-distribution dataset
R1 dataset
R2 dataset
Hybrid models prediction tasks:
Training dataset
Train & Test
In-distribution dataset
Training dataset
Train & Test
Out-of-distribution dataset

Figure 1. (a) Flow diagram of the proposed serial hybrid model. **(b)** Schematic diagram of lab scale reactor. **(c)** Prediction tasks to evaluate hybrid model performance.

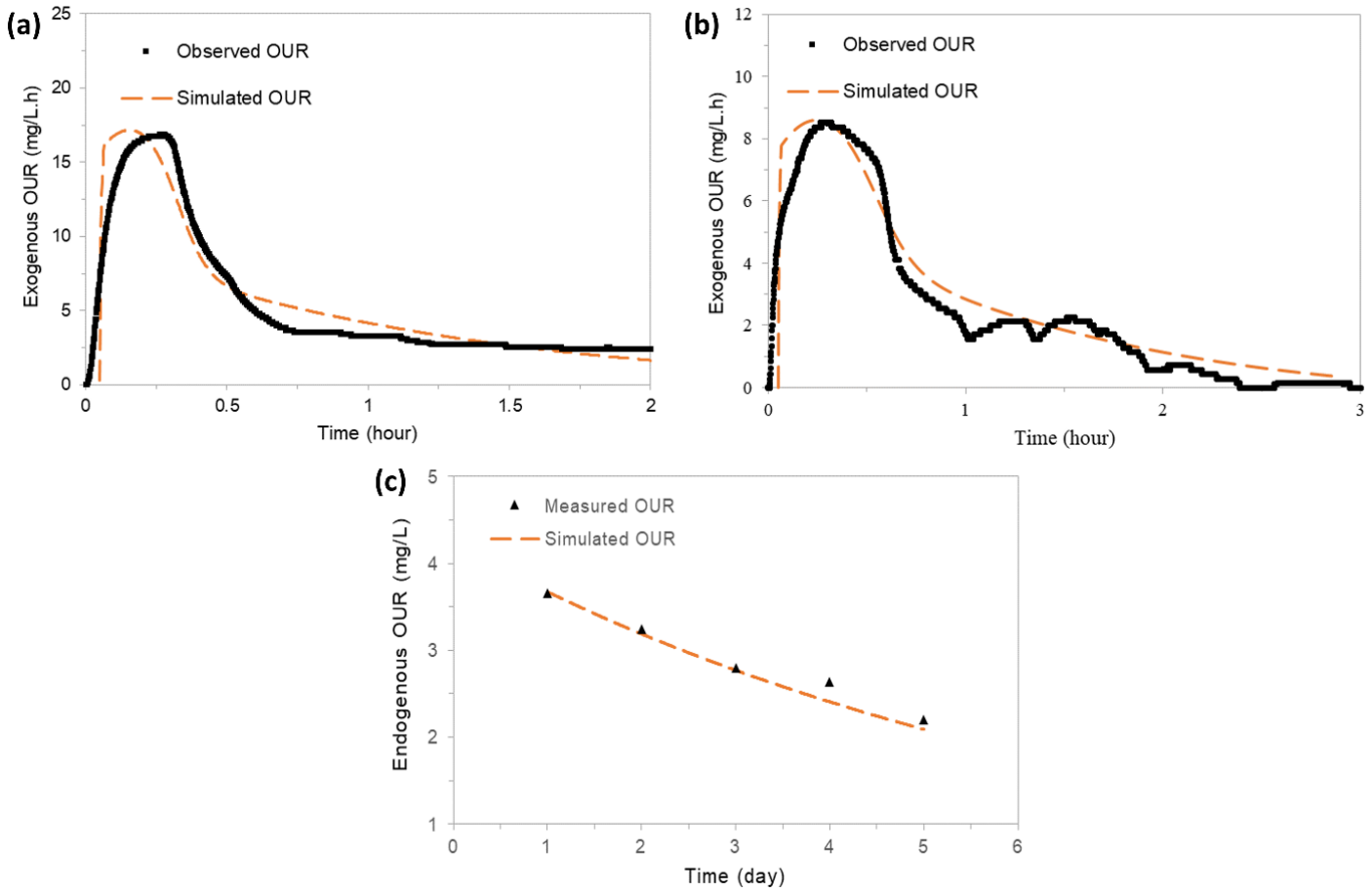


Figure 2. **(a)** Respirometry analysis of R1's influent wastewater obtained in Day 1, $X_{H,ini}$= 700 gCOD/m^3; **(b)** Respirometry analysis of R1's influent wastewater obtained in Day 7, $X_{H,ini}$= 750 gCOD/m^3; **(c)** Endogenous OUR of reactor R1 over a period of 5 days.

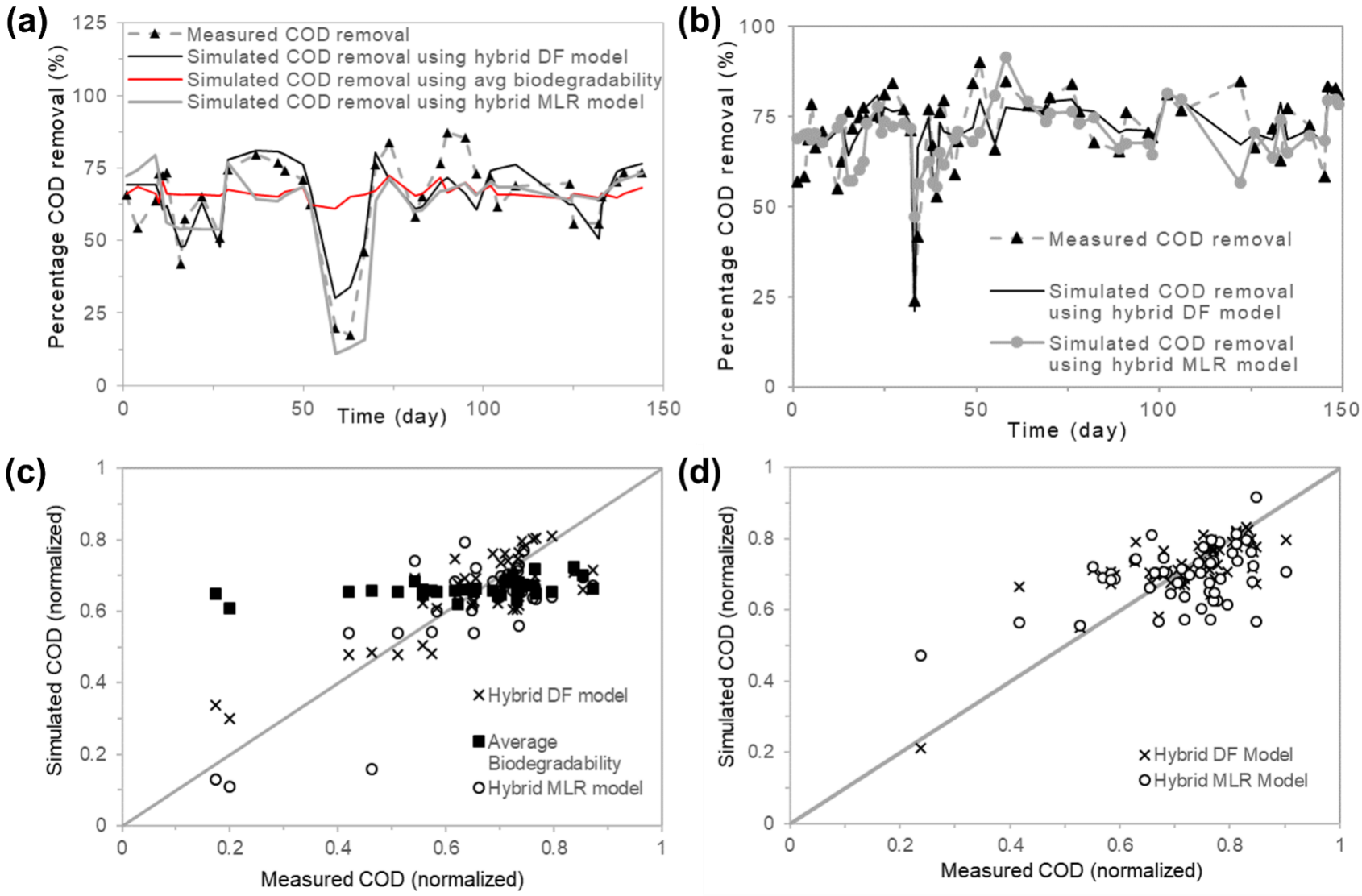

(a)
Measured COD removal
Simulated COD removal using hybrid DF model
Simulated COD removal using avg biodegradability
Simulated COD removal using hybrid MLR model
Percentage COD removal (%)
Time (day)
(b)
Measured COD removal
Simulated COD removal using hybrid DF model
Simulated COD removal using hybrid MLR model
Percentage COD removal (%)
Time (day)
(c)
Hybrid DF model
Average Biodegradability
Hybrid MLR model
Simulated COD (normalized)
Measured COD (normalized)
(d)
Hybrid DF Model
Hybrid MLR Model
Simulated COD (normalized)
Measured COD (normalized)

Figure 3. (a) Measured COD removal of R1 against simulated COD removal using average biodegradability, MLR-ASM3 model & DF-ASM3 model; **(b)** Measured COD removal of R2 against simulated COD removal using average biodegradability, MLR-ASM3 model & DF-ASM3 model; **(c)** Simulated COD (normalized) against measured COD (normalized) of R1; **(d)** Simulated COD (normalized) against measured COD (normalized) of R2.

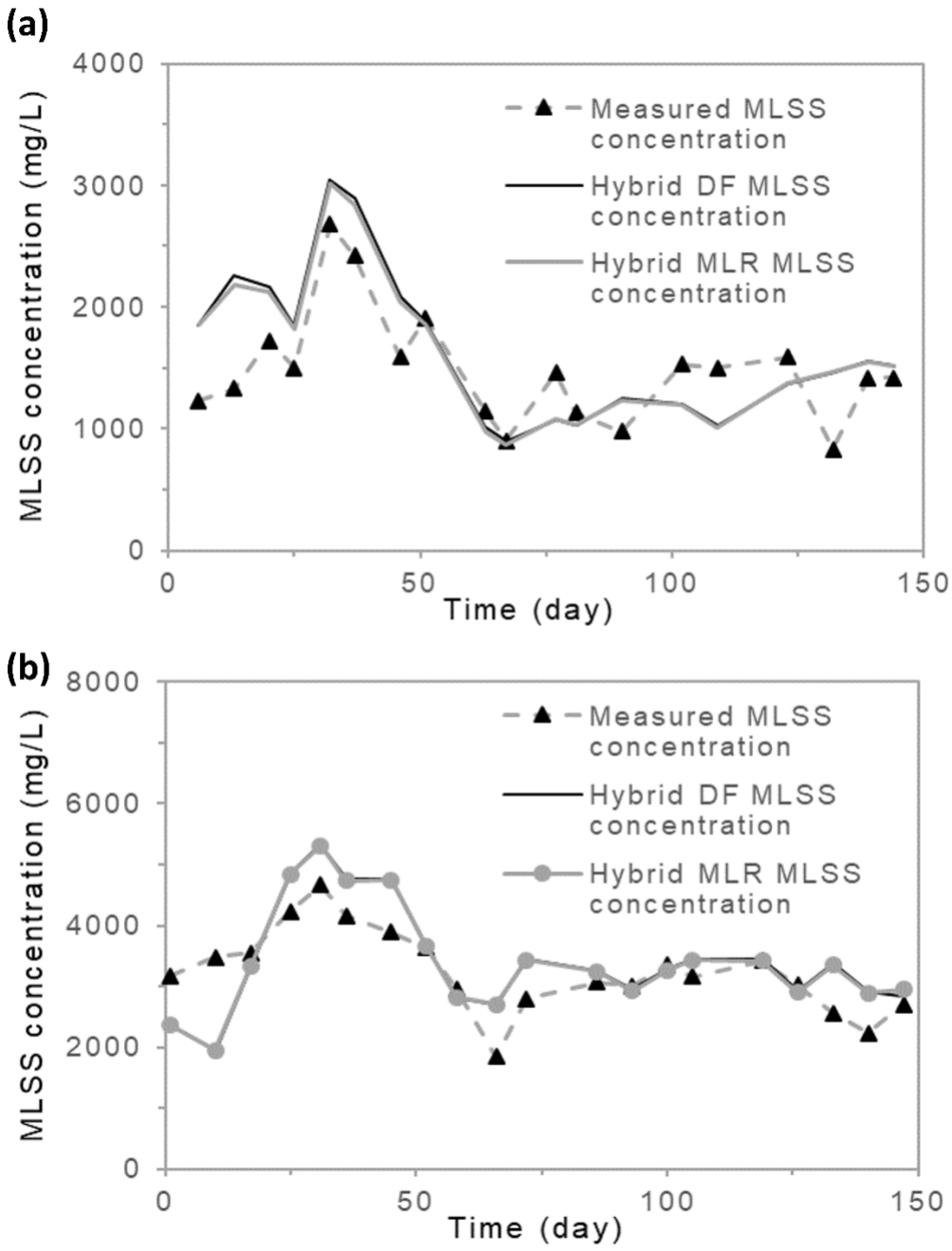


Figure 4. (a) Measured MLSS concentration of R1 against simulated MLSS concentration using MLR-ASM3 model & DF-ASM3 model; **(b)** Measured MLSS concentration of R2 against simulated MLSS concentration using MLR-ASM3 model & DF-ASM3 model.

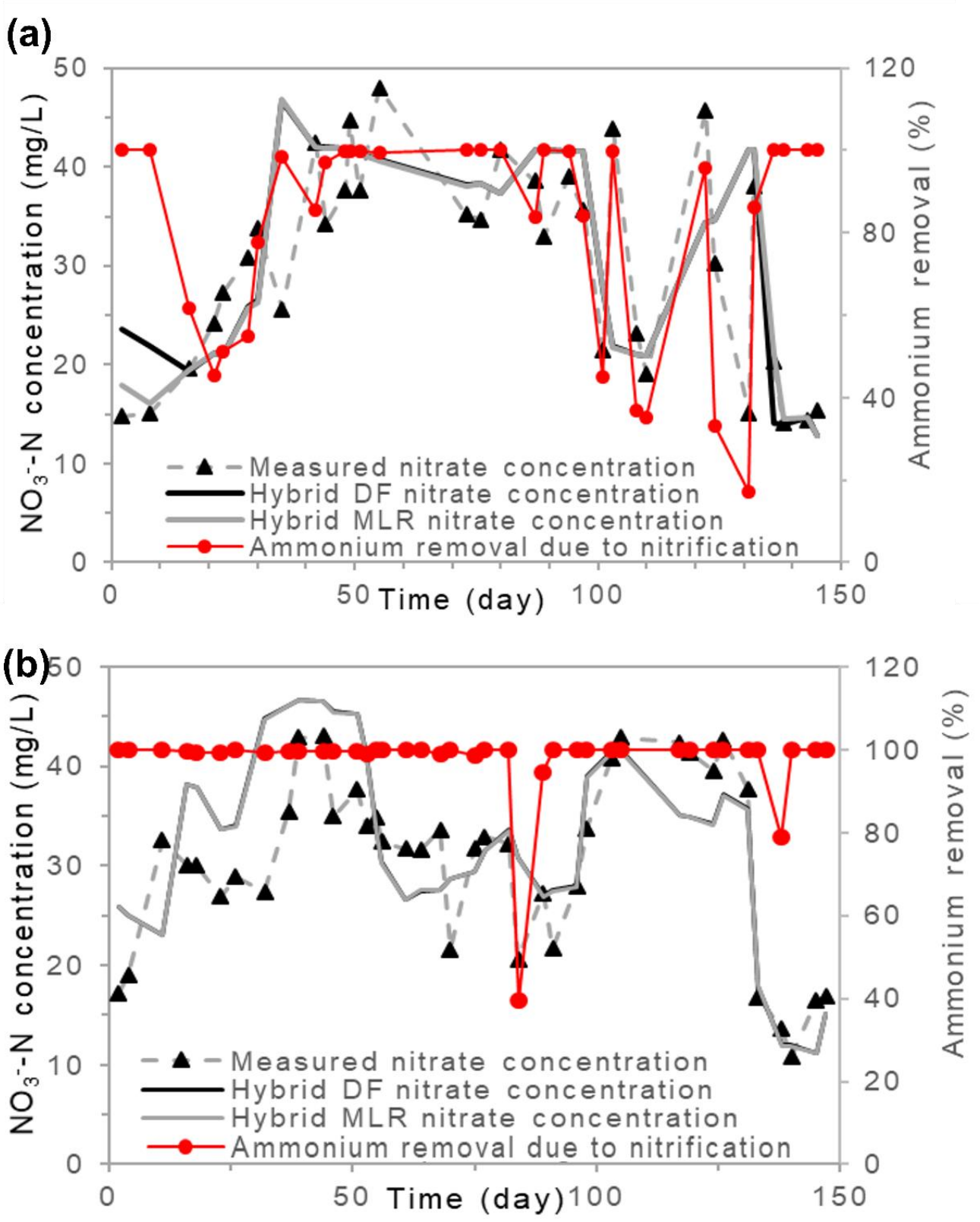

(a)
NO_3^--N concentration (mg/L)
Ammonium removal (%)
Time (day)
Measured nitrate concentration
Hybrid DF nitrate concentration
Hybrid MLR nitrate concentration
Ammonium removal due to nitrification
(b)
NO_3^--N concentration (mg/L)
Ammonium removal (%)
Time (day)
Measured nitrate concentration
Hybrid DF nitrate concentration
Hybrid MLR nitrate concentration
Ammonium removal due to nitrification

Figure 5. (a) Measured effluent nitrate concentration, simulated effluent nitrate concentration, and ammonium removal due to nitrification of reactor R1 using MLR-ASM3 model & DF-ASM3 model; **(b)** Measured effluent nitrate concentration, simulated effluent nitrate concentration, and ammonium removal due to nitrification of reactor R2 using MLR-ASM3 model & DF-ASM3 model. Note: The pH of reactor R1 and R2 were maintained between pH 7 – 8 with the installation of pH dosing pumps. However, during Day 0 – Day 30 and Day 105 – 115, the pH pump of R1 was down.