Environmental Technology & Innovation 23 (2021) 101632

Contents lists available at ScienceDirect

Environmental Technology & Innovation

journal homepage: www.elsevier.com/locate/eti

Prediction of wastewater treatment quality using LSTM neural network

Nitzan Farhi [a,*], Efrat Kohen [b], Hadas Mamane [b], Yuval Shavitt [a]

[a] School of Electrical Engineering, Tel-Aviv University, Israel
[b] School of Mechanical Engineering, Tel-Aviv University, Israel

ARTICLE INFO

Article history:
Received 26 February 2021
Received in revised form 30 April 2021
Accepted 16 May 2021
Available online 20 May 2021

Keywords:
Wastewater Treatment Plants
Activated sludge
LSTM
Sliding window
Fault prediction

ABSTRACT

Wastewater treatment (WWT) process is used to prevent pollution of water sources, improves sanitation condition, and reuse the water (mostly for agricultural purposes). One of the main goals of wastewater treatment is removal of nutrients, such as nitrogen which exists in the form of ammonia in the sewage. Excessive nitrogen concentration in the effluent is well known for eutrophication in aquatic environments and may cause a decrease of groundwater quality as a result of irrigation. However, it is not uncommon that the biological process results with undesirably high concentrations of nutrients, and therefore Wastewater Treatment Plants (WWTP) monitor nutrients to alert operators of this problem. It is desirable to identify WWT problems in the process ahead in order to achieve a better treatment. Thus, we suggest a novel machine learning method, based on Long-Short Term Memory (LSTM) architecture, that is able to predict effluent concentration of ammonia NH_4^+ and nitrate NO_3^- a few hours ahead. We used measurements from the biological reactors sampled every minute, and combine it, for the first time in the literature, with climate measurements for improved prediction accuracy. Our proposed method showed an accuracy rate of 99% and F1-Scoreof 88% when predicting ammonia concentrations and an accuracy rate of 90% and F1-Scoreof 93% when predicting nitrate concentrations.

1. Introduction

Sewage treatment is one of the main factors that allow sanitation conditions as it disrupts the transmission of infectious pathogenic vectors as deadly viruses and bacteria. On the global level, more than 80% of the sewage worldwide is dumped as-is or under-treated, thus better methods for handling sewage could benefit humanity and the environment. The growing demand for fresh clean water is one of the major global challenges, attributed mostly to population growth, climate change, rising standards of living and water quality deterioration. The benefit for reusing domestic wastewater is twofold: (1) preventing discharge of sewage to the environment, which causes pollution of water sources; and (2) decreasing the demand for fresh water mainly for agriculture.

In this paper, we used data from the Shafdan, which is the largest Wastewater Treatment Plant (WWTP) and reuse in Israel (30% of Israel's sewage). The Shafdan facility collects and treats domestic raw sewage to prevent its discharge to the environment and use its recycled water (effluent) for irrigation in the southern arid area of Israel. The process is based on activated sludge (AS) followed by soil aquifer treatment (SAT) that produces reclaimed water that is transported to the

* Corresponding author.
E-mail addresses: nitzanfarhi@mail.tau.ac.il (N. Farhi), efratkohen@mail.tau.ac.il (E. Kohen), hadasmg@tauex.tau.ac.il (H. Mamane), shavitt@eng.tau.ac.il (Y. Shavitt).

https://doi.org/10.1016/j.eti.2021.101632
2352-1864/

Negev area for irrigation. The reclaimed water has a unique water quality criteria in Israel by the health department which allows the use of the effluent for "incidental drinking", although the effluent is provided only for irrigation purposes.

Excessive nutrient concentration in the effluent adversely affects water quality and the environment (The Cadmus Group, 2009). Effluent containing high concentrations of nitrogen discharged into the aquatic environment will promote algae bloom. Algae cover the water surface and block sunlight from reaching underwater which deteriorates the ecological system. Moreover, algae bloom can promote toxic algae and bacterial growth and further result in low dissolved oxygen (DO) concentration. When effluent is used for agriculture purposes, residual nitrateinfiltrate into the ground and can pollute groundwater. From the WWTP operation point of view, high nitrateconcentration might cause a phenomenon called rising sludge (Henze et al., 1993). When high nitrateconcentration appears in the outlet of the aerobic reactor, de-nitrification completed in the bottom part of the secondary clarifies leads to accumulation of nitrogen gas. The N_2 gas rises up with the sludge and results in sludge escaping the clarifier with the effluent. The secondary effluent undergoes further treatment in the SAT basin. The amount of effluent the SAT basins can receive is based on the infiltration rate of the effluent through the soil. Infiltration rate decreases as a result of high organic load due to rising sludge or algae growth in the basin surface as a result of high nitrogen concentration.

Nitrogen removal in most WWTP is controlled by two processes named *nitrification* and *de-nitrification*. Nitrification is the process of oxidation of ammoniainto nitrate, while de-nitrification is the process of converting nitrateto nitrogen gas and this process occurs under anoxic conditions where no free oxygen is available. In the Shafdan WWTP oxygen supplied to the reactors is the key for balancing between nitrateand ammoniaconcentration, and if these concentrations could be predicted ahead of time, a change in the amount of oxygen supply ahead of time could prevent malfunctioning of the treatment process. Thus, predicting ammoniaand nitrateconcentrations ahead could indicate to the plant operators that an action is needed, and therefore prevent bad effluent quality, which influence the infiltration rates in the SAT process and eventually can cause effluent discharge to the environment.

Machine learning methods are used to train models, capable of identifying abnormal operational conditions in the WWTP operation, and thus alert on the need for controlling the concentrations of some measurements in the effluent. These models could prove to be very effective for wastewater treatment processes that are characterized by multi-variable control, high non-linearity and large time varying and complex parameter dependencies. Furthermore, temporal information in the fluctuations in the data could be used for improving predictions.

Only few studies modeled WWTP processes using neural networks. For example, Baruch et al. (2005) applied recurrent neural networks (RNNs) in modeling an adaptive control of WWTP processes. Capodaglio et al. (1991) applied two system analysis techniques, namely artificial neural systems (ANSs) and stochastic models, to analyze and predict bulking conditions which cause low effluent quality in activated sludge (AS). Qiao et al. (2019) designed a Recurrent Fuzzy Neural Network (RFNN) based approach to control the dissolved oxygen, nitrate–nitrogen (SNO) and mixed liquor suspended solids concentration in a WWTP.

When monitoring a WWTP, it is important to detect immediately any fault that occurs during the process that can lead to destructive results if not treated correctly. One method, proposed by Dairi et al. (2019) used deep learning and recurrent neural networks, to create an anomaly detection model via unsupervised learning. Based on capturing temporal auto-correlation features among multivariate time series from RNNs, abnormal events were reported from operators to check the model's accuracy. A work by Mamandipoor et al. (2020) focused on fault detection in WWTP where a series of time-steps, labeled as normal and faulty by experts, were analyzed by a neural network composed out of LSTM (further explained later) layers architecture.

Another way of keeping a WWTP from malfunction and control future values of measurements, is predicting the concentrations of effluent pollutant. For example, recent work (Pisa et al., 2019; Yaqub et al., 2020) was done in the subject of creating LSTM models that can well forecast ammoniaand total nitrogen value 4 h ahead, and developed a control strategy for reducing these concentrations. Han et al. (2018), for example, used Fuzzy Neural Networks to predict current plant ammoniaand nitrateconcentrations. Another indication of effluent quality is chemical oxygen demand (COD), that as was shown by Wang et al. (2019), could be predicted at real-time, using CNN–LSTM models. This improves current status, where measuring COD takes at least an hour and a half. Other prediction models related to the WWTP was made by Groenen (2018) where the amount of inflow to the plant was predicted using different Gated Recurrent Unit (GRU) (Cho et al., 2014) architectures. All models and architectures mentioned will be further explained later.

Most of the studies are limited and did not use a large dataset such as the Shafdan's. Since 2017, the Shafdan WWTP uses the IOSight's iGreen system as the center of operations of the facility and the main decision support system. To control and operate the WWTP efficiently, the plant data is collected from various sources including: SCADA sensors and control systems, lab results, weather reports and human observations. In the past four years, the aerobic reactors data was gathered, and currently includes 100 parameters and 850,000 usable observations.

When dealing with such great amount of data, it is important to leverage the temporal information encapsulated in the data. LSTM (Hochreiter and Schmidhuber, 1997) is an architecture that had proved its efficiency for such tasks. LSTM is based on RNN architecture, where performance decreases as a greater number of time-steps are fed into the network. However, LSTM has the ability to forget some of the less important data, and preserve the more important pieces. As can be seen in Fig. 2, an LSTM cell contains 3 gates: A forget gate (F_t in the figure and Eq. (1)) which is capable of deciding what information from the previous state C_{t-1} should not be passed into the next one C_t, an input gate (I_t in the figure and Eq. (2)) which is in charge of receiving new information and deciding how much of this information should be stored

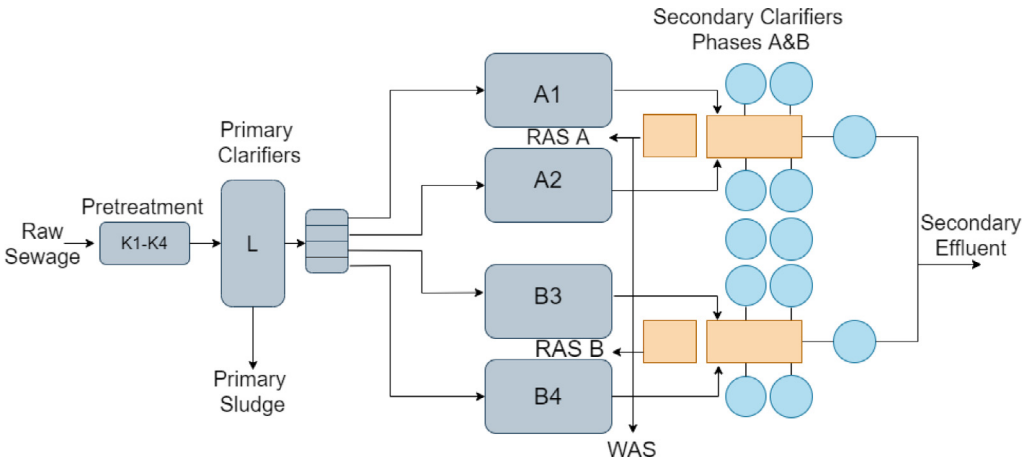


Fig. 1. Wastewater treatment process.

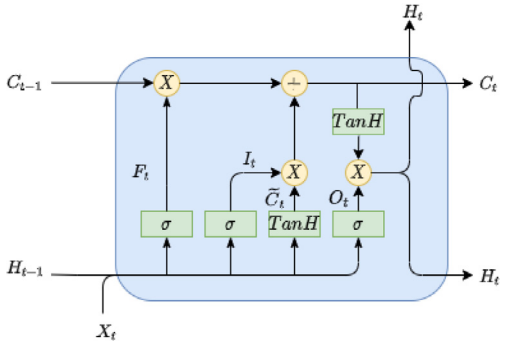


Fig. 2. The architecture of an LSTM cell.

inside the cell state and be passed to the next state C_t as defined in Eq. (5). Lastly, an output gate (O_t in the figure and Eq. (3)) which select what to output (H_t) based on the state of the cell and the last value. As can be seen in the equations, each gate ($x \in \{F, I, O\}$) contains two neural weights (W_x and B_x), which are updated in order to find the most suitable value throughout the learning process (example of a prediction made by an LSTM can be found in S1 in the supplementary material).

$$F_t = \sigma(W_f[H_{t-1}, X_t] + b_f) \tag{1}$$
$$I_t = \sigma(W_i \cdot [H_{t-1}, X_t] + b_i) \tag{2}$$
$$O_t = \sigma(W_o[H_{t-1}, X_t] + b_o) \tag{3}$$
$$\widetilde{C}_t = \tanh(W_c \cdot [H_{t-1}, X_t] + b_c) \tag{4}$$
$$C_t = F_t \cdot C_{t-1} + I_t \cdot \widetilde{C}_t \tag{5}$$
$$H_t = O_t \cdot \tanh(C_t) \tag{6}$$

In this paper, we purpose to improve the ability of predicting faults in the wastewater treatment process a few hours ahead. More specifically, predicting ammoniaand nitrateconcentrations, whose deviation from the plant's standard might have negative implications.

The contributions of this paper are as follows:

1. Improving state of the art results in the field using improved models and data processing methods with accuracy that reaches a rate of 99% and F1-Scoreof 88% when considering ammoniaconcentrations and accuracy rate of 90% and F1-Scoreof 93% when predicting nitrateconcentrations.
2. Using climate data, rain, temperature, radiation, etc., as an additional input when predicting future measurements of the WWTP. Especially rainfall is a significant factor in predicting both ammoniaand nitrateconcentrations.

Table 1
Description of the measurements we used (Mean, Std. deviation, and median were calculated for from one of the reactors) and the climate data.

Measurement	Mean	STD	Median	Comments
ammonia	6.57	5.44	5.81	Ammonia outlet concentration - mg/L as N
nitrate	5.55	3.60	5.95	Nitrate outlet concentration - mg/L as N
Reactor Energy	577.60	404.76	705.60	Rotors power consumption - KWH
Dissolved Oxygen	0.60	0.80	0.83	Dissolved oxygen in the reactor - mg/L
Dissolved Oxygen Outlet	0.61	0.93	0.37	Dissolved oxygen reactor outlet - mg/L
Rotors Depth Level	27.00	9.35	27.75	Rotor depth level - cm
Turbidity	2.69	1.52	2.57	Outlet turbidity - NTU
Outlet temperature	23.59	11.13	23.30	Outlet temperature - °C
Reactor temperature	22.20	9.08	21.93	Reactor temperature - °C
RAS TSS	5.06	3.27	4.84	Return activated sludge - total suspended solid - g/L
WAS-station	400.82	110.63	404.41	Waste activated sludge flow rate - m^3/h
FeedFlowToReactor	4342.50	1651.22	4630.00	Sewage flow rate - m^3/h
Raw sewage Conductivity	1335.43	897.76	1339.16	Raw sewage conductivity - μ S/cm
Raw sewage Tss	1343.05	3257.77	464.32	Raw sewage total suspended solid - mg/L
B11 flowmeter	826.32	209.09	848.00	Gravity thickeners centrate flow rate - m^3/h
Temperature	21.38	5.72	23.90	Outside temperature - °C
Ground Temperature	22.71	8.29	22.70	Ground temperature - °C
Relative Humidity	67.31	15.56	67.00	Relative humidity - percentage
Station Height Pressure	1008.32	4.57	1007.20	Station height pressure - hPa
Global Radiation	166.20	293.33	1.00	Global radiation - w/h^2
Direct Radiation	68.02	145.97	0.00	Direct radiation - w/h^2
Rain	0.01	0.12	0.00	Rain - mm
Wind Speed	2.39	1.52	1.60	Wind speed - m/s

3. Demonstrating aggregation techniques for compressing long time series to serve as input to the models improves prediction. Additionally, we showed how different factors as window size and prediction horizon affect prediction (Fig. 3).

2. Methods

2.1. Data description and water treatment process measurements

A general description of the biological treatment of the WWTP is described in Fig. 1. The raw sewage entering the plant flows through a pre-treatment for screening solids, wipes and grit through bar screens and cyclone grit chambers. This is followed by primary clarifiers (20 in number), where all primary sludge is transferred to anaerobic digesters. The wastewater then flows to a two stage activated sludge process. Each stage includes two bioreactors and six clarifiers of 52-meter diameter. Each bioreactor has three Zones: (1) 6000 m^3 anaerobic selector tanks, (2) 55,000 m^3 aerobic zone, and (3) an anoxic Zone. The reactor-modules, are each equipped with thirty-six horizontal rotor-aerator units that supply oxygen to the biological process. To summarize, there are 4 bioreactors, with 36 rotor-aerators per reactor, which supply the necessary oxygen to the biological process. Sensors continuously record measurements for ammonia, nitrate, flow-rates, rotors' water level depth, oxygen and turbidity. These measurements are recorded using a SCADA system, which records data every minute into the database. All measurements collected are tabulated with their mean, standard deviation and median in Table 1.

2.2. Data preparation

The automated process of measuring and recording data is not perfect and some of the measurements are missing. This is a well-known issue when analyzing datasets and can be solved by interpolating missing points (Yang et al., 2020), multiple imputation (Carpenter et al., 2019), or assigning an unused value. In our case, we choose to assign the unused value of −1 to every missing point, since the model architecture we use is capable of handling and ignoring these types of values. Parameters that lack 50% or more of the data, were not used.

Ammonia concentration depends on the hour of the day, the day in the week, and the month, thus, we added them as features (parameters) into the dataset using one-hot encoding method. In one-hot encoding, each numerical variable such as the day of the week is converted into a set of binary features (as further explained in S2 in the supplementary material).

In Israel,[1] Saturday (Shabbat) is the main rest day, and observant Jewish people are not allowed to work. Friday as the day before Saturday has also special human behavior since many people do not work, and engage in special preparation

[1] In the Tel Aviv district, which is roughly the area served by the Shafdan WWTP, about 93% of the population is Jewish.

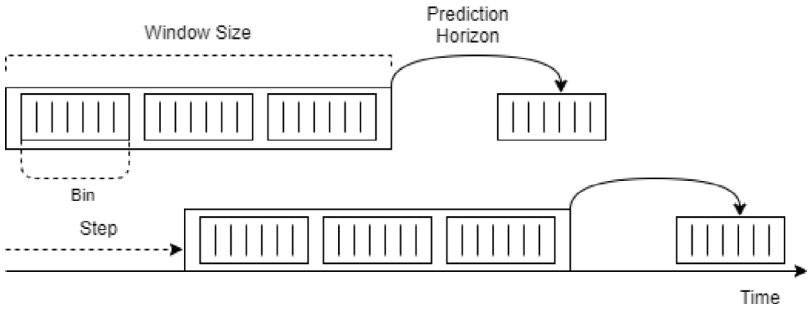


Fig. 3. Demonstration of the sliding window method.

activities towards the holy day of Shabbat. In a few major Jewish holidays we see similar human activities as in weekends, namely the inflow to the WWTP in the holiday behaves like on Saturday and the day before like a Friday. However, since there is insufficient data for holidays in our current dataset we did not add the holidays as features.

2.3. Climate dataset

Since the process of water treatment is also affected by external conditions, it is reasonable to suspect that external data can help improve prediction. Therefore, we used a climate dataset from the Israeli Meteorological Service[2] that contains environmental measurements from a station nearby the WWTP, including temperature, moisture, pressure, radiation, rain, wind speed and wind direction. Examples of known effect of climate data on the treatment process include:

- As temperature is higher, reaction kinetics is faster and bacteria are capable to treat the wastewater a shorter *hydraulic retention time* (HRT) (The Cadmus Group, 2009; Li et al., 2013). As a result, the effluent quality is much better during summer months, when the temperature is higher (Guo et al., 2010).
- Rain — since there is no sufficient separation between the sewage system and drainage system, rain water arrives with sewage to the WWTP and the *hydraulic retention time* (HRT) is much shorter.
- Radiation — since the Shafdan reactors are not very deep, 2.5 meters depth comparing to 6–8 meters when using diffusers as aeration system, bacteria in the reactors could be affected by drastic changes in radiation (Vergara et al., 2016).

Other measurements were inserted for experimental purposes.

2.4. Technical issues

2.4.1. Making the data temporal

The LSTM architecture accepts its data as a series of timestamps, each can be associated with many features. In our data, features are gathered every minute over a period of a few years, generating a time series, which is too long for the learning to be effective. To reduce the input size we employ the following process. We select a window in time as a single input and consider the data in this window to predict the value of ammoniaor nitrateconcentrations in a constant distance in the future, also called Prediction Horizon (PH). For example, an 8-h window was used to predict a value that is 4 h ahead of the end of that window (Fig. 3). Each vertical line in the bin represents a vector of feature from a given time. The selection of the window size and prediction horizon we used is detailed later in Section 3.

A window size of a few hours contains hundreds of samples, each sample is a vector of features, and is too large for an LSTM. To lower the input size we divided the windows into 1 h and 20 min bins (for ammoniaand nitrate, respectively). A bin is represented by a vector where each feature is the average value of the corresponding features of the samples. For example, if the window size is 3 h and the bin size is 1 h, a single input to the LSTM is comprised of 3 feature vectors each is an average of 1 h. The next input to the LSTM is generated by sliding the window by a certain time period, termed a step (Fig. 3). We selected the step to be equal to the bin size, but also experimented with other values. Again, detailed discussion of the model parameter selection can be found in Section 3.

The prediction result can be either a regression, namely an attempt to predict the value of a process parameter such as the ammoniaconcentration; or it can be a classification problem such as predicting that the ammoniaconcentration will rise above a certain threshold. We experimented with both type of predictions. The thresholds for ammoniaand nitrateconcentrations were supplied by the Shafdan field's experts.

2 https://ims.data.gov.il/ims/7.

2.4.2. Imbalance problem

A problem that usually occurs when trying to detect abnormal conditions is that these events are rather rare, and therefore few of them can be given to the model to learn. Therefore, the model is acquainted with normal behavior and is unable to detect anomalous events.

One possible solution for imbalanced datasets is picking specific date ranges where deviations occur more often. For example, the daily average of ammoniaconcentration in winter is 8–9 mg/L, whereas in summer they are around 5 mg/L. We set the ammoniathreshold on 15 mg/L-N as it is the maximum momentary allowed value for ammoniaby the Israeli effluent regulation. Abnormal events occur more often on winter months. Thus, we limit our training data to winter time only. For nitrate, we choose to set the threshold at 9 mg/L because on higher values there is a risk of rising sludge in the secondary clarifies (Henze et al., 1993). In this case, the data is balanced (the ratio between abnormal and normal events is 50:50) due to the fact that after primary sedimentation the C/N ratio is below 4, which restricts the de-nitrification process.

Another important system parameter is the overall sewage flow-rate to the plant. When the flow-rate to the bioreactors increases, the HRT decreases, and the bacteria do not have sufficient residence time to decrease the ammoniaconcentration. Thus, we can remove periods with low flow-rates from our training; consequently, we used HRT lower than 11 h as our threshold.

By removing data based on the two criteria above, we improved the ratio between positive (abnormal) and negative instances (normal) from a ratio of 5:95 to 20:80. It should be noted that this was not done on nitrateconcentration prediction, since the dataset is more balanced when predicting nitratemeasurements.

Finally, the data is normalized so that all variables will be scaled to the range [0, 1]. This is because LSTM models in particular and machine learning models in general are able to succeed better with that range. This is due to a problem called the vanishing gradient (Hochreiter, 1998) where neural network weights are heavily updated initially, can no longer be updated at the later stages of learning and thus are ineffective in these stages.

Another solution that helps balancing the data is random generation of samples in the minority class (the abnormal, in our case) (Lemaître et al., 2017). By taking all samples of the minority class in the training-set, and adding more samples with small random changes to it, we allow the model to learn from an even more balanced dataset, which may help the model learn better. This method was not used since the data was sufficiently balanced at this point.

2.4.3. Proposed model

We experimented with several models that cover the range of appropriate architectures for our problems. This includes CNN, GRU, and LSTM, all accept window structured data as described in Section 2.4.1. We describe below the LSTM Auto-Encoder model that was selected based on producing best results based on validation data.

The LSTM Auto-Encoder model (See Fig. 4) is composed of 7 layers, not including the input and out layers, the title above the layer describes its output shape. The first one, a Conv1D layer, extracts high-level features from the data. The layer takes 1D serial data and convolutes segment of the window size using a kernel, this layer uses Relu activation. The second layer, a dropout layer, is used to randomly select a certain percentage (in our case 30%) of the features and zero them, this layer helps reduce over-fitting as will be further explained below. The MaxPooling1D layer selects the maximum out of a few outputs of the Conv1D layer, this layer helps down-sampling the output of the Conv1D layer. These three layers are used in order to make sure that the model uses the useful parameters and ignores the less useful ones.

The fourth layer is an LSTM layer (as explained above) which outputs a single vector. To allow a connection to an additional LSTM (that accepts a 2D input), the input vector is reshaped using a Repeat Vector layer, which simply stack multiple copies of the input vector. Finally, a dense layer uses a Sigmoid activation function for classification problem or a linear activation for regression problem to output a prediction. The last layers (layers 4–6) are responsible for harnessing time information from the data in order to create the most accurate prediction.

The above neural network is trained using Mean Squared Error (MSE, see Eq. (7), but without the square root) cost function when doing a regression problem, and binary cross-entropy cost function when doing a classification problem.

2.5. Performance evaluation

When predicting a future state there are two ways to address the problem. One way is to address it as a regression problem, namely, predicting ammoniaor nitrateconcentrations. When predicting the ammoniaor nitrateconcentration value as a regression problem, we did not filter out records where the total HRT was below 11 h, as mentioned.

The second way to address the problem is a binary classification based on a threshold value. Every predicted instance that is greater than the defined threshold is labeled as ‘positive’, every other value is considered ‘negative’. As can be seen in Table 2, confusion matrix can visualize a model’s prediction efficiency, where better models will have maximal True Positive (TP) and True Negative (TN) values and minimal False Positive (FP) and False Negative (FN) values. In Table 2, the threshold is defined to be 9 mg/L of nitrateand numbers in the table are hours that are correctly/wrongly predicted to be higher or lower than the threshold.

When solving time series regression problems, one of the most popular metrics is Root Mean Squared Error (Eq. (7)) where t_i is the true ith value, p_i is the predicted ith value, and n is the number of values. This metric captures how different

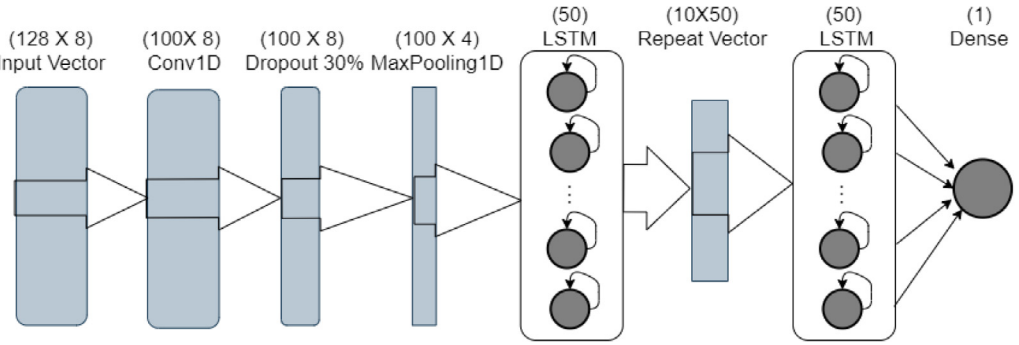


Fig. 4. The LSTM Auto-Encoder model.

Table 2
A confusion matrix of binary classification of nitrate concentrations.

		Prediction	
		Negative	Positive
Actual	Negative	*TN* - 2125	*FP* - 103
	Positive	*FN* - 78	*TP* - 2140

is the predicted curve from the real measurement. We did not use Mean Absolute Percentage Error like (Pisa et al., 2018) suggest, since zero values might occur. Also, negative predictions are penalized more in MAPE, which will cause models to be less likely to predict deviations.

When solving binary classification problems, accuracy as defined in Eq. (8) is usually used. Another common performance evaluation method for binary classifiers is the ROC curve. The area under the ROC curve (AUC) gives a numerical grade between 0 and 1 to the classifier, where 1 is a perfect classification. Other important metrics are precision (Eq. (10)), recall (Eq. (9)), and F1-Score(Eq. (11)), which is calculated from precision and recall. These metrics are important when classes are imbalanced or the prediction of one class is more important than the other. In this paper we will mainly look at the problem as a classification problem, because ammoniaor nitrateconcentrations are less interesting when they do not deviate from the threshold.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(t_i - p_i\right)^2} \tag{7}$$

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{8}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{11}$$

3. Results

Our main purpose is to understand the capabilities of machine learning models for predicting a future state of the WWTP, hours ahead. We will also explore how different parameters affect the prediction quality. For the experiments, we split the dataset into three parts: 60% of the data is the training set, which the model learns from; 20% is the validation set, which we use to understand which parameters help to achieve the best results, and 20% is the test set, which is used to test the final model. We used the validation set to understand the models' timely parameters and hyper-parameters as described in the next few sections. The test set was used to determine the model performance, in particular to derive the results in Tables 3 and 4.

The proposed model was implemented using the Keras (Chollet, 2015) library with TensorFlow backend (Abadi et al., 2015) using Python. Training the model was done on an Intel(R) Xeon(R) Platinum 8171M CPU in the Microsoft Azure cloud, where a single prediction takes about 5 ms. As mentioned above, different models were used, such as the one in Fig. 4. The data that was fed into the model was composed out of different reactors' measurements, climate data and one-hot encoding of hours, months, and weekdays.

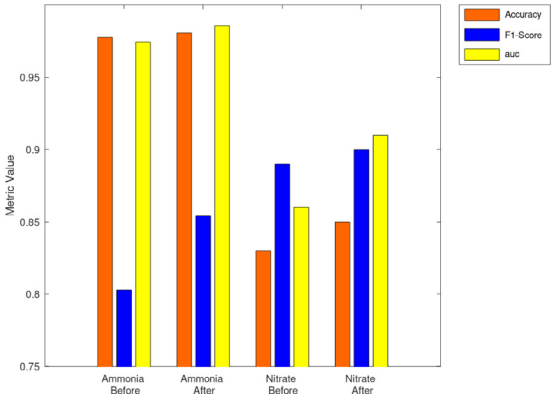


Fig. 5. Metrics before and after climate data was added.

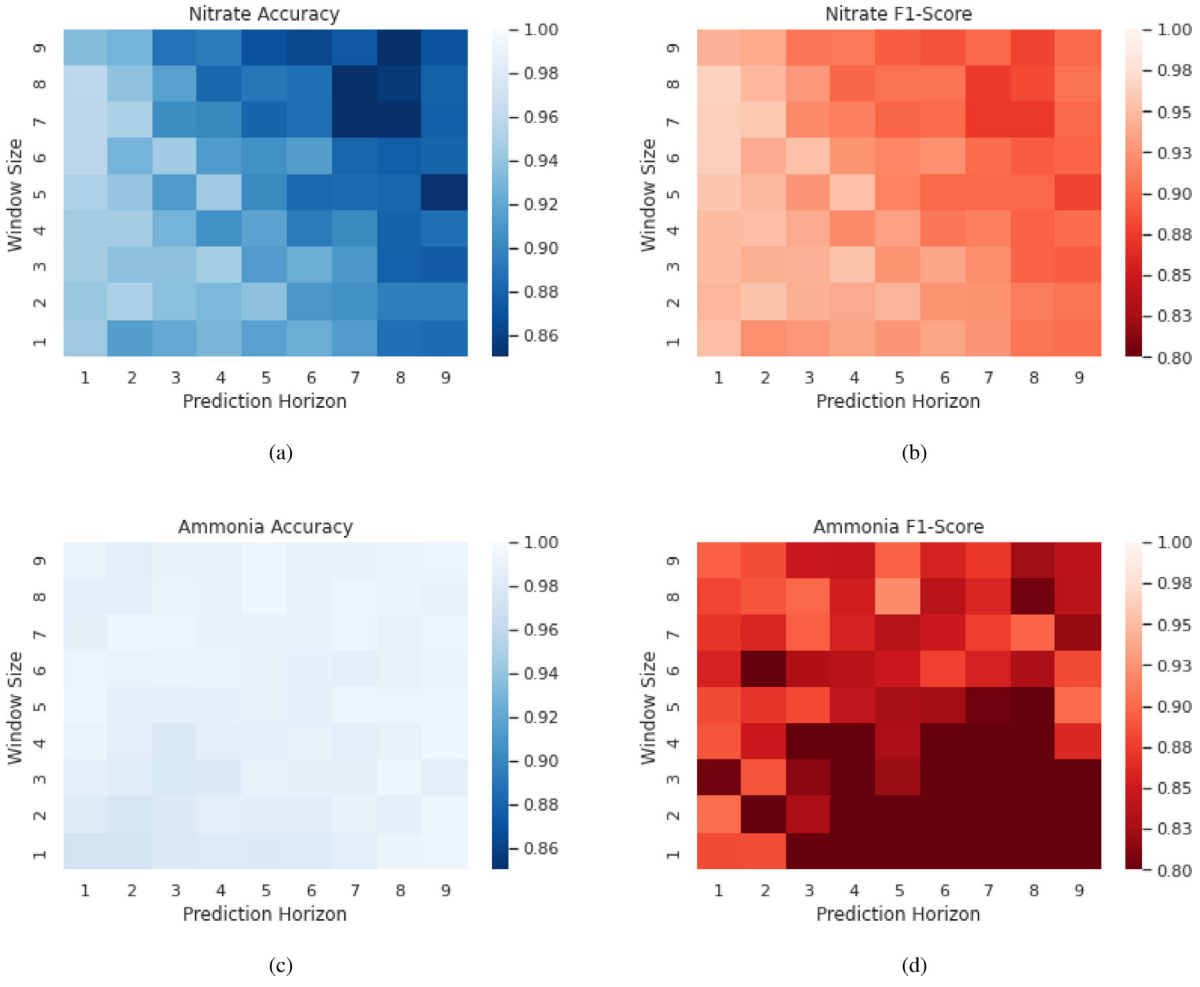


Fig. 6. Heat-maps Displaying the Performance of the Selected Model when using Different Window Size and Prediction Horizon.

Two important timely parameters of a prediction model that should alert when crossing from the threshold are: (1) how far ahead in the future one can predict (predicting horizon), and (2) how far in the past one should go to accumulate data to make the prediction. Fig. 6 shows how these two parameters affect the accuracy and F1-Scoremetrics (see Eqs. (8) and (11)). When looking at ammoniaconcentration prediction, both metrics show that the prediction horizon is very good until 4 h ahead, predicting more hours ahead will result in good accuracy and AUC (not shown in the figure), but not as

Table 3
Results of different models when looking at different metrics of ammoniaconcentration prediction (the best results are highlighted in bold). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Model\Metrics	Regression	Classification		
	RMSE	Accuracy	F1-Score	AUC
Yaqub et al. (2020)	0.068	X	X	X
Pisa et al. (2018)	0.12	X	X	X
Simple LSTM	0.042	0.97	0.83	0.98
Stacked LSTM	0.045	0.98	0.8	**0.99**
Bidirectional LSTM	**0.040**	0.98	0.84	0.98
Stacked GRU	0.043	0.98	0.85	0.92
GRU CNN	0.046	0.98	0.85	0.97
CNN	0.07	0.98	0.8	0.96
LSTM Auto-Encoder	**0.040**	**0.98**	**0.88**	0.98
LSTM Auto-Encoder Without Climate Data	0.05	0.91	0.82	0.94

Table 4
Results of different models when looking at different metrics of nitrateconcentration prediction (the best results are highlighted in bold).

Model \Metrics	Regression	Classification		
	RMSE	Accuracy	F1-Score	AUC
Yaqub et al. (2020)	0.12	X	X	X
Pisa et al. (2018)	0.4	X	X	X
Simple LSTM	0.101	0.89	**0.93**	0.93
Stacked LSTM	**0.095**	0.88	0.92	0.93
Bidirectional LSTM	0.097	0.90	**0.93**	0.94
Stacked GRU	0.116	0.87	0.91	0.92
GRU CNN	0.107	**0.98**	0.85	**0.97**
CNN	0.127	0.90	**0.93**	0.92
LSTM Auto-Encoder	0.097	0.90	0.91	0.91
LSTM Auto-Encoder Without Climate Data	0.100	0.85	0.88	0.90

good F1-Score. When predicting nitrateconcentrations, both F1-Scoreand accuracy are similarly high for up to 4 h ahead. The Shafdan plant operators verified that 4 h is sufficient time in order to change the process's strategy and reduce the concentrations of ammoniaor nitrate, the same amount of hours was also used by Pisa et al. (2019).

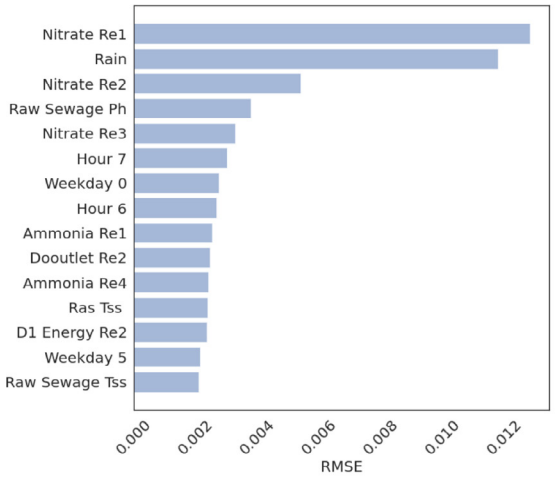
We are the first to include climate data in our prediction model. To evaluate the contribution of climate data to our prediction, we run the model also without climate data. Fig. 5 shows that all metrics were improved when the model was using the climate dataset as well. Both accuracy and AUC improved by 1% and F1-Scorewas improved by 5% when predicting ammoniaconcentrations and in nitrateconcentrations accuracy and F1-Scoreimproved by 2%, and AUC improved by 5%. Thus, WWTP should either get a live feed of climate data or install climate sensors when a decision-making software operates the plant.

A useful method to extract important information after creating a model is understanding the importance of each feature in the model, which can be done by replacing a feature by random data and checking how it affects predictions by calculating the difference in the RMSE (Eq. (7)) between the regular prediction and the one after the randomization. Fig. 7 depicts the RMSE of different features.

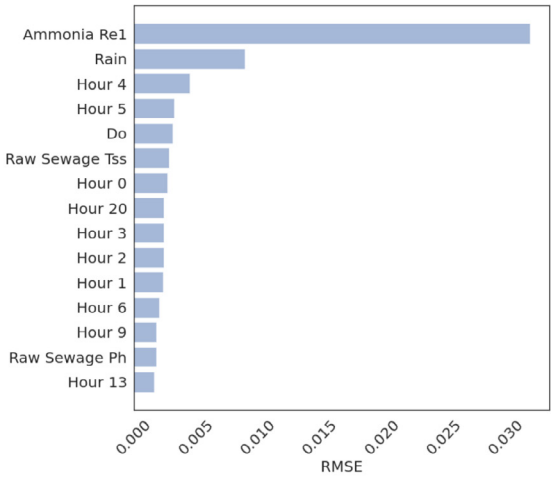
Rain, which is extracted from the Climate dataset, and the time in the day have a great impact on the model. For example, Hour_20 depicts the importance of the one-hot encoding of the hour 20:00 in the model, Weekday 0 depicts the one-hot encoding of the first day of the week (Sunday, in Israel). Other measurements from the reactors also appear in the figure, for example measurement from the first reactor is denoted as Re1.

Predictably, current ammoniaconcentrations of a reactor is the most important feature when predicting future ammonia, and current nitrateconcentration is the most important when predicting future nitrateconcentrations. This is reasonable, since models heavily rely on the using the delta between a current value and the predicted value when making a prediction. The model also relies heavily on the rain feature both for nitrateand ammoniaprediction, this can be explained by the fact that rain water causes the hydraulic retention time (HRT) to be much shorter and bacteria is less capable of treating the wastewater. Furthermore, weekdays and different reactor measurements were also used by the model when making a prediction.

When looking at ammoniaconcentrations, morning hours have great impact on the model whereas nitrateprediction uses various hours and weekdays for the prediction. Also, the predicting model of nitrateconcentrations is greatly influenced by current ammoniaconcentrations, but there is no strong influence of nitratewhen predicting ammoniaconcentrations. We assume this is due to the filtering of data periods of high HRT when predicting ammoniaconcentrations. During high HRT, excessive oxidation will cause low ammoniaconcentrations and high nitrateconcentrations and vice versa.

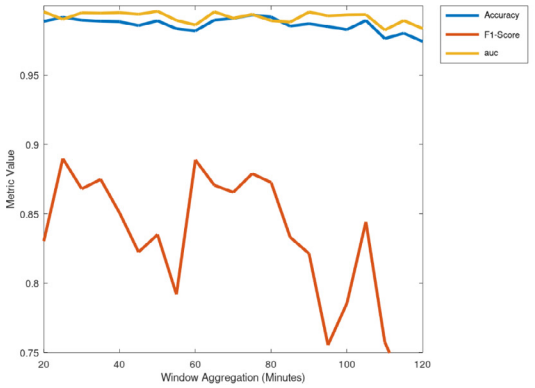


(a) 15 Nitrate Features that their contribution to the prediction was the greatest

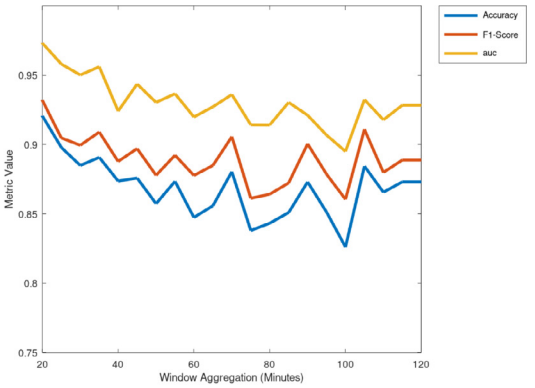


(b) 15 Ammonia Features that their contribution to the prediction was the greatest

Fig. 7. Feature Importance.



(a) Ammonia prediction metrics as a function of window aggregation



(b) Nitrate predictions metrics as a function of window aggregation

Fig. 8. Window Aggregation.

Another technique that helped achieve better accuracy is aggregating timestamps. For example, when no aggregation is done, a model that learns from an 8 h of accepts as input 480 samples, one per minute. This results with too many parameters to learn, and the model's performance decreases or fails completely. The large amount of data with small differences between consecutive time-steps distract the model from the larger scope. When trying to find the best aggregation size, we found that aggregation of 60 min into 1-h timestamp using a mean function achieved the best result when predicting ammoniaconcentrations and aggregation of 20 min are optimal for predicting nitrateconcentrations, as can be seen in Fig. 8, respectively. Other aggregation functions were considered, such as maximum value or sum of all measurements in the window, but none of them were beneficial to the learning process as mean did. Note that climate data was given in a frequency of 10 min measurements and the treatment process measurements' frequency is 1 min.

We also tested the size of the steps between consecutive windows, looking at the range of 1 bin size to 20 bin size. Results showed that when the number of steps was larger than 1, large volumes of data were ignored, which made the learning sub-optimal and the amount of test instances smaller. Therefore, step size was selected to be 1.

Fig. 6 explores the accuracy and F1-Scorefor both ammoniaand nitrateas a function of the window size and prediction horizon. The figure uses the same color scale for both ammoniaand nitratefor each metric. When predicting ammonia-concentration, accuracy and F1-Scoreare reaching maximum value at window size of 8 h (when each single instance is an aggregation of 60 min). Note that for ammoniaaccuracy is extremely high for all combination depicted, thus the Fig. 6(c) appears bleary.

Models that used a window size of 5 h (15 instances, since 20 min were aggregated into one instance) were able to predict nitrateconcentrations better. This could be explained by the fact that LSTM models are confused by information given to them prior to the 5 h as was also mentioned in the work of Mamandipoor et al. (2020).

Window size and prediction horizon are used in every LSTM model (Tian, 2020). These parameters affect the accuracy of the model although not always could be rationally explained. For example, as can be seen in Fig. 6, accuracy and F1-Scoredecreases as prediction horizon increases, which is expected. On the other hand, while the expectation is that larger window size will result in better predictions, we do not see this for all three cases (ammoniaaccuracy as mentioned above is always good regardless of the parameters). Deviation from the expected behavior may be explained by the possibility that LSTM models are sometimes confused by older information instead of taking older timestamps into account only when they help to increase performance.

In preparation for the final comparisons, we used the optimal parameters of window size and window aggregation. We maximized the success rate of the validation set by using a method called hyper-parametering; with this method, an exhaustive search of different parameters and models is used for training the model and the ones who achieve the best results at the validation-set is selected for the testing with the test set. The parameters that were tuned:

- Number of epochs $\in \{20, 50, 100, 200, 500\}$ - The number of times the model ran over the training set.
- Optimizer $\in$ {Adam, SGD} - In charge of changing the model's weights throughout the training stage.
- Batch Size $\in \{20, 32, 64\}$ - The number of instances that the model learned before changing its weights.

When considering an optimizer, SGD, unlike Adam, helped some of the models avoid the exploding gradient problem, which is common on neural networks based on LSTM. It is helpful to view the loss value of the model as a function of the number of epoch when trying to understand when the model starts memorizing samples instead of understanding the data (over-fitting), this occurs at around epoch 100 where validation loss remains about the same whereas training loss keeps decreasing.

Finally, Tables 3 and 4 compare the various models we tested with previous works (Yaqub et al., 2020; Pisa et al., 2018). Clearly, our results are better than state of the art previous methods, which did not report all the metrics we used (marked with X on both tables). Furthermore, even when comparing to our most successful model, the LSTM auto-encoder, without the use of climate data, it still performs better than previous works. We attribute this to the usage of different time processing such as different window aggregation, steps and window sizes.

For nitrate, all the models we tested and also (Yaqub et al., 2020) performed quite well for regression, and most were performing well for classification. For ammonia, the differences were larger and the LSTM auto-encoder model was a clear winner both for regression and classification.

The importance of the classification results is that they test the prediction of a process failure at the WWTP, as well as the importance of the F1-Scoredue to the imbalance of the data. Although steps were taken to minimize the imbalance effect on the results (Section 2.4.2) we still noticed that a single mistake at predicting the positive class damages the F1-Scorebadly. The threshold was chosen to maximize the F1-Scoreas it is the balance between precision and recall.

4. Conclusion and future work

We showed the possibility of accurately predicting a future concentration of ammoniaand nitrate, using the temporal measurements of the plant combined with climate data. In a data rich environment of a WWTP, we studied how to extract the most out of the data in order to feed it to a deep learning model whose performance is dictated to a large degree by the number of samples it is fed with.

In the future, we would like to better understand the effect that an input of long-term time series has on the model performance. For example, our results show that when the window is increased beyond a certain size, we do not gain improvement in prediction. However, old data may still hold important information, and we would like to understand how can we utilize it in LSTM models.

Another limitation of our models is explainability: an operator may want to know why the model reached its conclusion. For this end, we suggest to study *attention* mechanisms (Vaswani et al., 2017) that are capable to export the key input parameters that led to their decision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This publication was supported by AI for Earth grant from Microsoft, United States and a grant from TAU, Israel Center for Data Science. We would also like to thank the Shafdan for sharing their WWTP data and knowledge.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eti.2021.101632.

References

Abadi, Martín, et al., 2015. Tensorflow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, Software available from tensorflow.org.

Baruch, Ieroham S., Georgieva, Petia, Barrera-Cortes, Josefina, de Azevedo, Sebastiao Feyo, 2005. Adaptive recurrent neural network control of biological wastewater treatment. Int. J. Intell. Syst. 20 (2), 173–193.

Capodaglio, Andrea G., Jones, Harold V., Novotny, Vladimir, Feng, Xin, 1991. Sludge bulking analysis and forecasting: application of system identification and artificial neural computing technologies. Water Res. 25 (10), 1217–1224.

Carpenter, Corey M.G., Wong, Lok Yee J., Gutema, Danyeh L., Helbling, Damian E., 2019. Fall creek monitoring station: using environmental covariates to predict micropollutant dynamics and peak events in surface water systems. Environ. Sci. Technol. 53 (15), 8599–8610.

Cho, Kyunghyun, Van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, Bengio, Yoshua, 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chollet, François, 2015. Keras. https://github.com/fchollet/keras,

Dairi, Abdelkader, Cheng, Tuoyuan, Harrou, Fouzi, Sun, Ying, Leiknes, TorOve, 2019. Deep learning approach for sustainable WWTP operation: A case study on data-driven influent conditions monitoring. Sustainable Cities Soc. 50, 101670.

Groenen, Inske, 2018. Representing Seasonal Patterns in Gated Recurrent Neural Networks for Multivariate Time Series Forecasting (Ph.D. thesis), Master thesis.

Guo, Jianhua, Peng, Yongzhen, Huang, Huijun, Wang, Shuying, Ge, Shijian, Zhang, Jingrong, Wang, Zhongwei, 2010. Short-and long-term effects of temperature on partial nitrification in a sequencing batch reactor treating domestic wastewater. J. Hard Mater. 179 (1–3), 471–479.

Han, Honggui, Zhu, Shuguang, Qiao, Junfei, Guo, Min, 2018. Data-driven intelligent monitoring system for key variables in wastewater treatment process. Chin. J. Chem. Eng. 26 (10), 2093–2101.

Henze, Mogens, Dupont, Rene, Grau, Petr, De La Sota, Alejandro, 1993. Rising sludge in secondary settlers due to denitrification. Water Res. 27 (2), 231–236.

Hochreiter, Sepp, 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 6 (02), 107–116.

Hochreiter, Sepp, Schmidhuber, Jürgen, 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Lemaître, Guillaume, Nogueira, Fernando, Aridas, Christos K., 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. 18 (17), 1–5.

Li, Hongyan, Zhang, Yu, Yang, Min, Kamagata, Yoichi, 2013. Effects of hydraulic retention time on nitrification activities and population dynamics of a conventional activated sludge system. Front. Environ. Sci. Eng. 7 (1), 43–48.

Mamandipoor, Behrooz, Majd, Mahshid, Sheikhalishahi, Seyedmostafa, Modena, Claudio, Osmani, Venet, 2020. Monitoring and detecting faults in wastewater treatment plants using deep learning. Environ. Monit. Assess. 192 (2), 148.

Pisa, Ivan, Santín, Ignacio, Lopez Vicario, José, Morell, Antoni, Vilanova, Ramon, 2018. A recurrent neural network for wastewater treatment plant effuents' prediction. Actas XXXIX Jornadas Automática Badajoz.

Pisa, Ivan, Santin, Ignacio, Morell, Antoni, Vicario, Jose Lopez, Vilanova, Ramon, 2019. LSTM-based wastewater treatment plants operation strategies for effluent quality improvement. IEEE Access 7, 159773–159786.

Qiao, Jun-Fei, Han, Gai-Tang, Han, Hong-Gui, Yang, Cui-Li, Li, Wei, 2019. Decoupling control for wastewater treatment process based on recurrent fuzzy neural network. Asian J. Control 21 (3), 1270–1280.

The Cadmus Group, 2009. Nutrient Control Design Manual: State of Technology Review Report. Technical Report EPA/600/R-09/012, US Environmental Protection Agency.

Tian, Zhongda, 2020. Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM. Eng. Appl. Artif. Intell. 91, 103573.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia, 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008.

Vergara, C., Muñoz, R., Campos, J.L., Seeger, M., Jeison, D., 2016. Influence of light intensity on bacterial nitrifying activity in algal-bacterial photobioreactors and its implications for microalgae-based wastewater treatment. Int. Biodeterioration Biodegrad. 114, 116–121.

Wang, Z., Man, Y., Hu, Y., Li, J., Hong, M., Cui, P., 2019. A deep learning based dynamic COD prediction model for urban sewage. Environ. Sci.: Water Res. Technol. 5 (12), 2210–2218.

Yang, Kun, Yu, Zhenyu, Luo, Yi, 2020. Analysis on driving factors of lake surface water temperature for major lakes in Yunnan-Guizhou Plateau. Water Res. 184, 116018.

Yaqub, Muhammad, Asif, Hasnain, Kim, Seongboem, Lee, Wontae, 2020. Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network. J. Water Process Eng. 37, 101388.