A Novel Machine Learning-based Energy Consumption Model of Wastewater Treatment Plants

Shike Zhang[1], Hongtao Wang[1,2*], Arturo A. Keller[3]

1 College of Environmental Science and Engineering, Key Laboratory of Yangtze River Water Environment, Ministry of Education, UNEP-Tongji Institute of Environment for Sustainable Development, Tongji University, 1239 Siping Rd, Shanghai, 200092, China

2 Shanghai Institute of Pollution Control and Ecological Security, Tongji University, 1239 Siping Rd, Shanghai, 200092, China.

3 Bren School of Environmental Science and Management, UC Santa Barbara, Santa Barbara, California 93106, United States

Corresponding author: Email: hongtao@tongji.edu.cn (Hongtao Wang)

Abstract: Wastewater treatment plants (WWTPs) can account for up to 1% of a country's energy consumption. Meanwhile, WWTPs have high energy-saving potential. To achieve this, it is necessary to establish appropriate energy consumption models for WWTPs. Several recent models have been developed using logarithmic, exponential, or linear functions. However, the behavior of WWTPs is non-linear, and difficult to fit with simple functions particularly for non-numerical variables. Thus, traditional modeling methods cannot effectively describe the relationship between water and energy in WWTPs. Therefore, a machine learning method was adopted in this study to investigate the energy consumption in WWTPs; a novel energy consumption model with a non-numerical variable (discharge standard) for WWTPs was developed using the random forest algorithm. The model can also predict the energy consumption of WWTPs after upgrading discharge standards. We found that the unit electricity consumption of WWTPs exhibited an average increase of 17% after the effluent

discharge standard was raised from Class I B to Class I A (per China's classification). The correlation coefficient of the model was 0.702. Thus, the developed model can provide a better understanding of energy efficiency in WWTPs.

Keywords： Machine learning; random forest; energy efficiency; energy consumption model; wastewater treatment plant

1. Introduction

Improving energy efficiency of WWTPs is receiving increasing attention, as saving energy can help reduce economic costs and conserve the resources and environment (E.Açıkkalp, 2018). With continuous development and accelerating urbanization in society, wastewater discharge is rapidly increasing (Habib et al. 2020) and water quality requirements are more stringent; therefore, the total energy consumption of WWTPs is also increasing. WWTPs are the primary energy-consuming units of the urban water cycle (Sabia et al. 2020). Thus, the high energy consumption of WWTPs has become a global concern. It has been estimated that in 2018 the energy demand of WWTPs in some European countries accounted for 1% of the energy consumption of the entire country (Sabia et al. 2020). What's more, the U.S. municipal wastewater treatment systems use approximately 30.2 billion kWh per year, which is about 0.8% of the total electricity use in the U.S. (EPRI, 2013). In recent years, several energy evaluation methods have been proposed (e.g., Hernández-Sancho et al. 2009, Mizuta et al. 2010, Molinos-Senante et al. 2014) to investigate the energy consumption of WWTPs. In these methods, the energy consumption of WWTPs is commonly related to factors such as the capacity and influent and effluent concentrations of pollutants (Torregrossa et al. 2016).

Many stakeholders have been exploring solutions to reduce the energy consumption of WWTPs, such as equipment renewal and maintenance (Daw et al. 2012), energy recovery (Behera et al. 2020), and technical process improvements (Farahbakhsh et al. 2020). Previous studies have been mostly focused on technical processes. However, with the use of new equipment, technologies, and new standards, the change in energy consumption has gradually attracted significant attention (Sabia et al. 2020). With increasing urban wastewater discharge and high requirements of clean water, new WWTPs are regularly established, and the discharge standards of the old WWTPs have gradually improved (Smith et al. 2019). However, a larger process capacity and higher standards may lead to higher energy consumption. Therefore, while considering water quality, one should also focus on the energy efficiency of WWTPs.

Machine learning is an important and relatively novel method in environmental modeling, particularly with regards to energy efficiency (e.g., Wang et al. 2019) or WWTP operations (e.g. Hernandez-del-Olmo et al. 2019). Machine learning can be utilized in a real-time agent modeling, employing real-time data so that operators can forecast WWTP’s future operating status; the model itself can be improved continuously as new data becomes available, with the ability to adopt non-linear relationships. Once a set of inputs and corresponding outputs are presented to the model, it learns the relationship between the inputs and outputs. Accordingly, for a new set of inputs, the trained model can generalize this relationship to produce the corresponding outputs (e.g., Heslot et al. 2014, Song et al. 2017, Xing et al. 2019, Zhu et al. 2020). Random forest is an ensemble learning algorithm used for classification (e.g., Duro et al. 2012), regression (e.g., Wei et al. 2019), and other tasks (e.g., Chen et al. 2018). During training, numerous decision trees are generated to operate and finally obtain prediction results (e.g.,Tian et al. 2020); therefore, random forest has a high prediction accuracy, and is not prone to overfitting (e.g., Breiman 2001). Random forest is one of the most popular methods in data mining (e.g., Wylie et al. 2019) and big data fields (e.g., Pamulaparty et al. 2017); it has the advantages of a fast-training speed and is suitable for processing high-dimensional data (e.g., Belgiu and Dragut 2016). The method has been widely used in several other fields, such as medicine (Yeşilkanat 2020), criminal investigation (Tian et al. 2020), and architecture (Cheng et al. 2020). Random forest has also been applied to environmental engineering, such as for mapping canopy nitrogen (Loozen et al. 2020) and in environmental assessment (Paul et al. 2020). However, random forest is rarely used to analyze and predict energy consumption of WWTPs (e.g., Bagherzadeh et al. 2021, Perez et al. 2021). Data Envelopment Analysis (e.g., Yang et al. 2021, Huang et al. 2021) and Multiple Linear Regression (e.g., Xu et al. 2018) are commonly used methods to analyze the energy efficiency of WWTPs. The main principle of Data Envelopment Analysis is using the method of linear programming and Multiple Linear Regression is a kind of generalized linear model. However, Data Envelopment Analysis cannot be used when some data is missing, and it cannot predict the future trend in a statistical way. Multiple Linear Regression cannot achieve the accuracy we need.

This study aims to develop an energy consumption model of WWTPs through machine learning, using data from 2,472 WWTPs in China, employing the random forest approach. This model is expected to provide a better understanding of energy consumption in WWTPs.

2. Data and Methods

2.1 Data Sources

The data of this study was collected from the *2015 Urban Drainage Yearbook of China*. A total of 2,472 entries were selected from this yearbook with relatively complete and reliable data. For the energy consumption of WWTPs, the energy consumed by processing 1 m^3 wastewater is often used as an evaluation indicator of the energy intensity of WWTPs (Scott et al. 2011). Related studies (e.g., Mjalli et al. 2007, Gazendam et al. 2016, Trenouth et al. 2018, Habib et al. 2020) commonly use electricity intensity (kWh/m^3) to indicate the energy consumption of WWTPs. In this study, the following parameters from the Yearbook, which have also been used in related studies (Mjalli et al. 2007, Gazendam et al. 2016, Trenouth et al. 2018, Habib et al. 2020), have been adopted as the primary factors affecting the energy consumption of WWTPs: influent BOD_5 concentration (BOD_i), influent COD concentration (COD_i), influent NH_3-N concentration (NH_3-N_i), effluent BOD_5 concentration (BOD_e), effluent COD concentration (COD_e), effluent NH_3-N concentration (NH_3-N_e), effluent discharge standards, wastewater treatment capacity, annual load rate (actual treatment capacity divided by designed treatment capacity), moisture content of sludge, and dry weight of sludge. The discharge standards primarily include Class I A, Class I B, and Class II, referring to the Chinese National Standard, *Discharge standard of pollutants for municipal wastewater treatment plant* (GB 18918-2002).

2.2 Data Cleaning

To import data to develop the model, all numerical data (including int64, float64) was converted to float64, all non-numerical data (including string and object, such as the discharge standard) was transferred into the object, and all the default data was converted to NaN (Not a Number).

Since a few of the WWTPs did not have the “unit electricity consumption” (UEC) parameter in the Yearbook, to ensure the reliability of the model, we removed the data for those 85 WWTPs, so that the number of the remaining WWTPs was 2,387.

Although there were some outliers, the data base is large and Random Forest model is good at dealing with this situation, so there was no need to eliminate them. Essentially, the basic learner of random forest is robust to outliers, which makes the random forest algorithm robust to outliers. Unlike linear regression, the entire space in linear regression has the same equation, so a very simple model can be locally fitted to each subspace.

In the case of regression, it is usually a very low-order regression model. Therefore, for regression, extreme values do not affect the entire model because they are averaged locally.

2.3 Preprocessing of Regression Variables

The numerical variables can be directly applied to the regression. For the object type variables, such as discharge standards, their classification scheme was transformed into a matrix with 0 and 1 values, such that, the row of the matrix represents the different WWTPs, and the column represents the different discharge standards. A value of 1 indicated that the WWTP represented by this row used the discharge standard of this column. Otherwise, a value 0 was assigned. For example (Fig. 1), the discharge standards of WWTP A, WWTP B, and WWTP C are Class I A, Class II, and Class I B, respectively. Therefore, the sum of each row in the matrix is 1, and the sum of each column equals the total number of WWTPs using the discharge standard of this column. In this study, numerical values (e.g., 1, 2, 3, etc.) were not used to represent the different discharge standards. This is because the values themselves include the potential relationship of size or numerical operation, and substituting a relationship which is unrelated to the statistical content into the regression model will lead to model deviation.

$$
\begin{array}{r c}
 & \begin{array}{ccc} \textbf{Class I A} & \textbf{Class I B} & \textbf{Class II} \end{array} \\
\begin{array}{r} \textbf{WWTP A} \\ \textbf{WWTP B} \\ \textbf{WWTP C} \end{array} & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}
\end{array}
$$

Fig. 1 Matrix of discharge standard in WWTPs

2.4 Random Forest

A preliminarily evaluation of the relationship between UEC (kWh/m^3) and other parameters was performed using Python, to conduct Multiple Linear Regression analysis between UEC and BOD_i, COD_i, NH_3-N_i, BOD_e, COD_e, NH_3-N_e, wastewater treatment capacity, annual load rate, moisture content of sludge, and dry weight of sludge. It was found that $R^2 \leq 0.2$, which is too small, thus the regression equation was not sufficiently reliable. Moreover, the discharge standard is a character-type variable that cannot be included in the statistics and effectively predict its influence by Multiple Linear Regression. Owing to the high correlation between the parameters, a large fitting deviation occurred when using the Multiple Linear Regression method to obtain the relationship between UEC and dependent variables. In addition, the use of a Multiple Linear Regression model is limited in this case because of the existence of non-numerical variables, such as the discharge standard.

Machine learning algorithms like TensorFlow or Keras require very large databases, which are not available for this study. While there are machine learning algorithms like Lars, Lasso or Support Vector Machine (SVM) which only need small databases, they cannot reach the accuracy needed for this study. Therefore, we considered a subset of machine learning algorithms, including Random Forest (RF), Boosting Tree, Gradient Boosting Decision Tree (GBDT) and XGBoost. These algorithms are actually combination of different algorithmic frameworks and decision trees (Detail see Table S1), so they perform quite similarly. We selected Random Forest because it is the only algorithm that can show us the importance of each variable, which is very valuable for the subsequent analysis.

Therefore, a random forest algorithm was introduced to extract the relationship between UEC and the different variables, including non-numerical ones. Simultaneously, the factors indicating the influence of each variable on UEC were calculated, and then the factors that significantly affected UEC were selected for further analysis and to develop a model for evaluating the UEC. Finally, the change in UEC was calculated using the model after a simulated improvement of the discharge quality to meet a higher standard, which can help in future management of WWTPs.

The steps conducted for the random forest approach were showed in Fig. 2.

From a mathematical perspective, a complex functional relationship exists between independent variables and dependent variables, which is composed of the basic operations of independent variables. Random forest approximates the coefficients before each dependent variable by learning from a large amount of data. All the models in this study were coded in Python 3.7.3, and the prediction curves were plotted from the Python data using MATLAB R2018a.

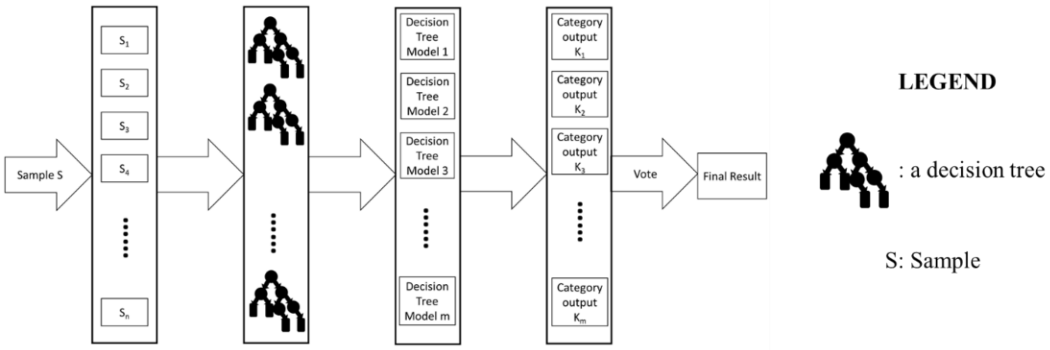

Fig. 2 Process flow of the random forest method

2.5 Model Validation

Random forest uses a bootstrapping algorithm for sampling. As the bootstrapping algorithm returns samples after sampling, some data are not extracted. By calculating the limit, it was observed that approximately 1/3 of the data were not extracted.

Because out-of-bag (OOB) data were not been used, random forest can use these data for model validation. Moreover, as each sample obtained by bootstrap trains a small model S_n, the OOB data can be tested for each model of the sample.

The self-detection of the model uses the mean squared error (MSE), average absolute percentage (MAPE), root mean square error (RSME), mean absolute error (MAE), median absolute error (MedAE) and mean squared logarithmic error (MSLE) (Zhong et al. 2021, Gupta et al. 2021) as follows:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(actual(t) - predicted(t)\right)^2 \quad (1)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|actual(t)-predicted(t)|}{actual(t)} \times 100\% \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(actual(t) - predicted(t)\right)^2} = \sqrt{MSE} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|actual(t) - predicted(t)| \quad (4)$$

$$MedAE = median\ (|actual(t_1) - predicted(t_1)|, \dots, |actual\ (t_n)\ - predicted(t_n)|) \quad (5)$$

$$MSLE = \frac{1}{n}\sum_{t=1}^{n}\left(\log_e\left(1 + actual(t)\right) - \log_e\left(1 + predicted(t)\right)\right)^2 \quad (6)$$

where n is the number of decision tree model, $actual(t)$ is the actual UEC of a WWTP, and $predicted(t)$ is the predicted UEC of a WWTP.

2.6 Data Preprocessing

Certain evident linear relationships exist between some variables in the Yearbook, which were removed before modeling to obtain a model with improved accuracy.

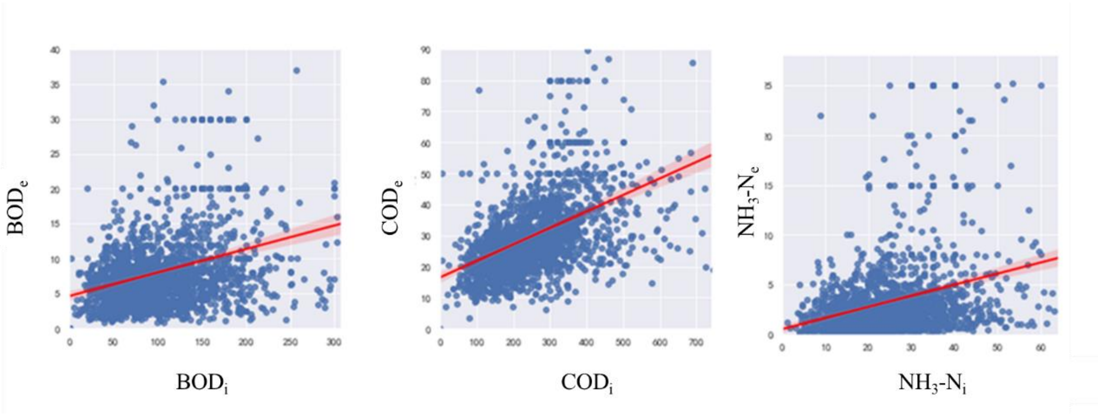

Fig. 3 Scatter plot and linear regression curve of

(a) BOD_i/BOD_e, (b) COD_i/COD_e, and (c) NH_3-N_i/NH_3-N_e.

In Fig. 3, each point in the figure represents a WWTP dataset, and it may be noted that the BOD_i and BOD_e, COD_i and COD_e, NH_3-N_i and NH_3-N_e, for most WWTPs are evenly distributed along a straight line. The light-red area around the regression curve represents the confidence interval. Therefore, we performed a linear regression between BOD_i and BOD_e, COD_i and COD_e, NH_3-N_i and NH_3-N_e, which showed that the linear correlation between BOD_i and BOD_e, COD_i and COD_e, NH_3-N_i and NH_3-N_e was high. At the same time, the correlation between each variable was reduced to the lowest possible limit to improve the accuracy of machine learning. In this study, the removal ratios BOD_i/BOD_e, COD_i/COD_e, and NH_3-N_i/NH_3-N_e were used instead of a single variable for analysis, which practically represent the reduction multiple of BOD, COD, and NH_3-N_e of treated wastewater.

2.7 The Importance of Features

The importance of a feature X in a random forest was calculated as follows:

A. For each decision tree in a random forest, the corresponding OOB data were used to calculate the OOB data error, which was recorded as errOOB1.

B. Random noise interference was added to the characteristic X of all samples of the OOB data, and the OOB data error was calculated again, which was recorded as errOOB2.

C. If there are N trees in the random forest, then the importance of the feature is given as

$$X_{Importance} = \sum (errOOB2 - errOOB1) / N \quad (7)$$

This expression can be used as a measure of the importance of corresponding features because if a feature was randomly added with noise, the accuracy rate outside the bag was highly reduced, which indicated that this feature had a high influence on the classification results of samples; in other words, it was of high importance.

3. Results and Discussion

3.1 Correlation between Variables

To accurately analyze the relationship between UEC and the different variables, we calculated the correlation between these variables, as shown in Fig. 4.

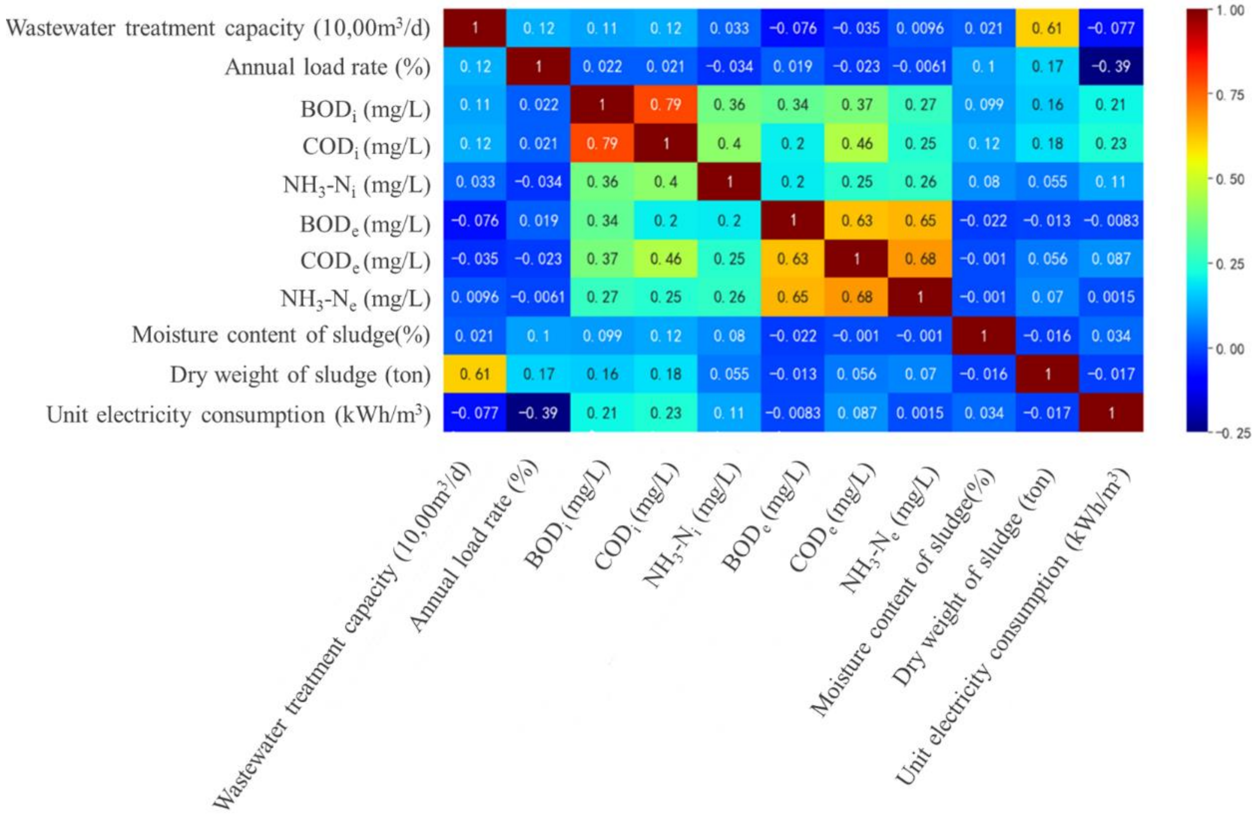

Fig. 4 Correlation thermodynamic diagram of variables

In Fig. 4, the correlation of UEC, wastewater treatment capacity, annual load rate, moisture content of sludge, dry weight of sludge, BOD_i, BOD_e, COD_i, COD_e, NH_3-N_i, and NH_3-N_e are described by the thermodynamic diagram. The number on the color block represents the correlation between the corresponding variables of the abscissa and ordinate. A darker red implies a higher correlation, and a darker blue indicates a lower correlation. The UEC is highly correlated with BOD, COD, and NH_3-N of influent and effluent (Fig. 4).

3.2 Regression

The analysis of the importance of the independent variables is presented in Table 1 and Fig. S1. The regression model had an $R^2 = 0.702$, which was significantly higher than that of the Multiple Linear Regression, implying higher accuracy.

As shown in Table 1 and Fig. S1, the most important variable was the wastewater treatment capacity, which is expected since this determines the sizing of pumps, air blowers and other equipment that consumes electricity (Torregrossa et al. 2018). This is followed by the annual load rate, which is also expected to be a major factor (Torregrossa et al. 2018). Wastewater treatment capacity and annual load rate can reflect the influence of the design and practical operation of WWTPs on energy consumption, with a total importance of 0.38. The high importance indicated that the design of a

WWTP was very important, so a clear treatment target would significantly affect the energy consumption of WWTPs.

Table 1. Variables and their importance

Variable	Importance
Wastewater treatment capacity (m^3/d)	0.2130
Annual load rate (%)	0.1758
COD_i /COD_e	0.1655
BOD_i /BOD_e	0.1170
Moisture content of sludge (%)	0.1134
NH_3-N_i/NH_3-N_e	0.0846
Dry weight of sludge (ton)	0.0747
Discharge standard	0.0560

The removal efficiency of COD and BOD had a significant impact on the energy consumption of WWTPs, which also verified that the level of removal of COD and BOD highly affected the energy consumption of WWTPs (Longo et al. 2016). This is consistent with other studies, and the pollution load is consistent with the energy consumption load of WWTPs (Torregrossa et al. 2018). CODi/CODe is significantly more important than BODi/BODe since the pollutants measured by BOD are subset of pollutants measured by COD, so COD contains some pollutants that do not belong to

BOD. Second, the model may divide the pollutants that both belong to BOD and COD into the importance of BOD_i/BOD_e and COD_i/COD_e. These two factors may contribute to the finding that the importance of COD_i/COD_e is significantly higher than BOD_i/BOD_e.

However, the removal efficiency of NH_3-N, one of the primary pollutants in sewage, is of relatively low importance to UEC that is because the primary function of most WWTPs in China is to remove organic matter rather than denitrification, which leads to the lower importance of NH_3-Ni/NH_3-N_e. Some studies have shown that COD, BOD, and NH_3-N are correlated (Luo et al. 2019) and the energy to power blower fans are actually the main factor in electricity consumption for the removal of COD, BOD, and NH_3-N (Piotrowski and Ujazdowski 2020). Considering the fact that the smaller the number of highly correlated variables, the more convenient is the practical application of the model, we assigned the importance of the overlap between variables to the high correlation variable, which led to the low importance of NH_3-N_i/NH_3-N_e. Depending on the request to the accuracy of the model, NH_3-N_i/NH_3-N_e can be neglected during practical usage, but to analyze the model more clearly and completely we will still take NH_3-N_i/NH_3-N_e into consideration in the following discussion.

There are limits to the moisture content of sludge, so there will be energy consumption to separate water from sludge. However, from Fig. 4 it appears that moisture content of sludge has low correlation with other variables, so its importance will be higher. During sludge conditioning, drying, and incineration, a large amount of energy is required; however, in the current statistical yearbook of WWTPs, there is no data on energy consumed for sludge disposal. Therefore, we could not further analyze the importance of sludge treatment.

The results in Table 1 show that compared with the data type variables, the importance of the discharge standard (Table 2) was low because BOD_i, COD_i, and NH_3-N_i are limited by discharge standard, so discharge standard is highly correlated to them. Since the current model is built to minimize the influence of this correlation, so the importance of discharge standard is low. Table 1 indicates that the discharge standard is low in importance; hence, in the following analysis, the discharge standard was not used to forecast the UEC of WWTPs.

Table 2. Discharge standard of COD, BOD, and NH_3-N

Parameter	Class I A	Class I B
COD_e (mg/L)	50	60
BOD_e (mg/L)	10	20
NH_3-N_e (mg/L)	5(8)	8(15)

Note: The value in the bracket means standard at temperature ≤ 12 ℃, which was not modeled in this study.

3.3 Model Demonstration and Prediction of Energy Consumption

An energy consumption model for WWTPs was established through training using a large amount of data. By changing the input variables, we can predict the change in the energy consumption of a WWTP. We selected the following variables with high importance: design treatment capacity, annual average load rate, and removal ratios (BOD_i/BOD_e, COD_i/COD_e, and $NH_3\text{-}N_i/NH_3\text{-}N_e$) to obtain the prediction function. The model can be directly presented through this curve and it can be applied to the management of energy efficiency of real WWTPs.

3.3.1 Wastewater Treatment Capacity

From the predictive model shown in Fig. 5, it is evident that the wastewater treatment capacity is negatively related to UEC, and for wastewater treatment capacities from 10,000 m^3/d to 100,000 m^3/d, the UEC decreases rapidly with an increase in the design treatment capacity. Above 100,000 m^3/d there is minimal decrease in UEC, which is a consideration for the design of WWTPs. The overall trend is consistent with the finding of previous studies (e.g., Yang et al. 2021, Huang et al. 2021). What's more, this finding also follows the scale economy of WWTPs (Hernández-Chover et al. 2018).

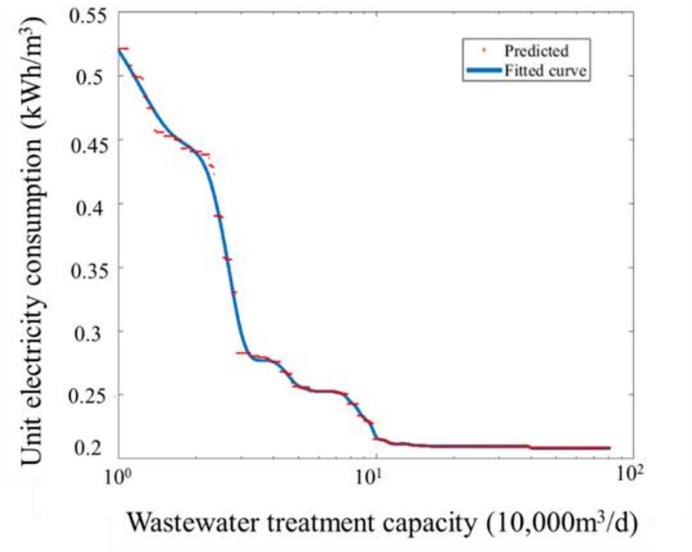

Fig. 5 Predictive model of UEC as a function of wastewater treatment capacity with other variables constant.

The construction scale of WWTPs in China can be divided into five categories (Ministry of Construction of China, 2001), as shown in Table 3:

Table 3 Standard of construction scale of WWTPs in China

Category	Construction Scale
I	500,000~1,000,000 m^3/d
II	200,000~500,000 m^3/d
III	100,000~200,000 m^3/d
IV	50,000~100,000 m^3/d
V	10,000~50,000 m^3/d

From the predicted data（Fig. 5）, we found that the UEC of WWTPs with a scale of I, II, and III was relatively low. Therefore, we can conclude that the WWTPs larger than 100,000 m^3/d have effectively reduced energy consumption, and there are 245 WWTPs in this range in the database, which is 9.91% of the total WWTPs considered in this study (Fig. 6).

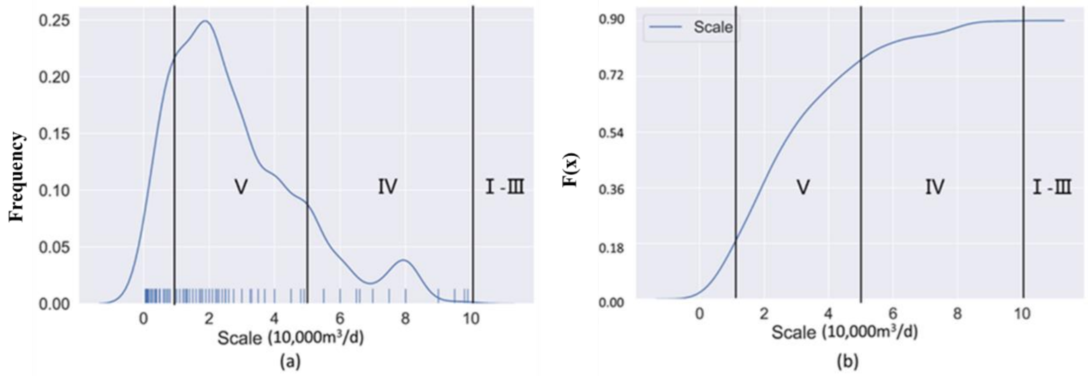

Fig. 6 (a) Kernel frequency distribution (b) Probability distribution of wastewater treatment capacity of WWTPs in the model (to make the figure clearer, all the WWTPs with a wastewater treatment capacity above 100,000 m^3/d were not counted in the figure).

The ordinate of a point on the curve in Fig.6 (a) shows the proportion of the scale that is reflected by abscissa of total WWTPs in China. The ordinate of a point on the curve in Fig.6 (b) shows the accumulated proportion of the scale smaller than abscissa of total WWTPs in China and the slope of the point shows the frequency. In Fig.6 (a) the curve peaks in Category V which means the scale of WWTPs concentrated in Category V and, after the peak, the number of WWTPs decreases with the scale in an overall trend. In Fig.6 (b), the tangent slope of the curve also shows that Category V includes most of the WWTPs in China. In short, Fig.6 shows the scale distribution of WWTPs in China and we can find that most WWTPs in China are small-scale.

3.3.2 Annual Load Rate

Annual Load Rate means the percentage usage of wastewater treatment capacity over the year, it reflects the divergence of design and actual usage of WWTPs. As shown in Fig. 7, the annual load rate has a significant impact on the UEC. The UEC remained high when the annual load rate was less than 40%, but the UEC decreased significantly when the annual load rate was between 40% and 100%; meanwhile, the UEC remained stable in a low range after the annual load rate was more than 100%. (i.e., overload). However, as overload may damage the instruments and equipment, a load rate between 60% and 100% should be maintained in the design and operation of WWTPs. The trend in this study corresponds well the results of Huang et al. 2021 and it also follows the rule of extensive models of different factory managements (Gerami et al. 2021). This finding indicates that it will be better to do more study on the amount of wastewater needed to be treated in one area before designing the treatment capacity of the WWTP.

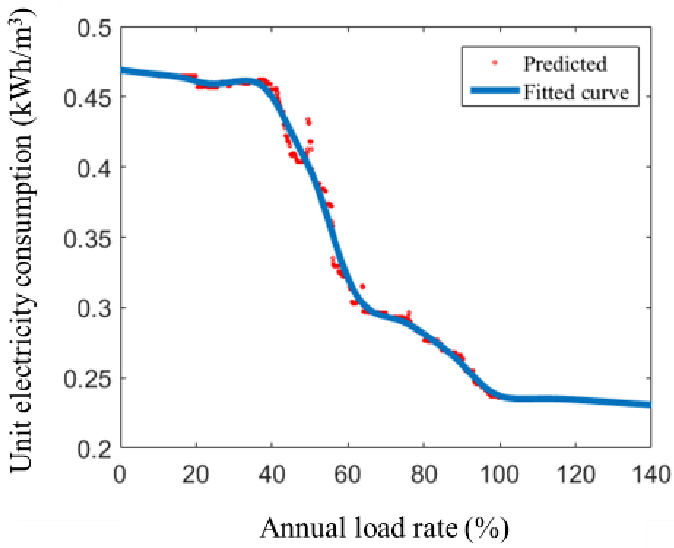

Fig. 7 Predictive model of UEC as a function of annual load rate with other variables constant.

3.3.3 Reduction Ratios

In this section, the effects of COD_i/COD_e, BOD_i/BOD_e, and $NH_3\text{-}N_i/NH_3\text{-}N_e$ on UEC are analyzed. As shown in Table 1, the importance of the UEC of COD removal was significantly greater than that of BOD and NH_3-N, and UEC was primarily affected by COD removal. To achieve a comprehensive study, the influence of BOD removal and NH_3-N removal on UEC is also discussed herein. However, the low importance is reflected on the ordinate of BOD and NH_3-N. In general, the trends of COD and BOD is generally in line with Huang et al. 2021, while the trend of NH3-N is a little conflict with the common sense, we are going to explain it in the following discussion.

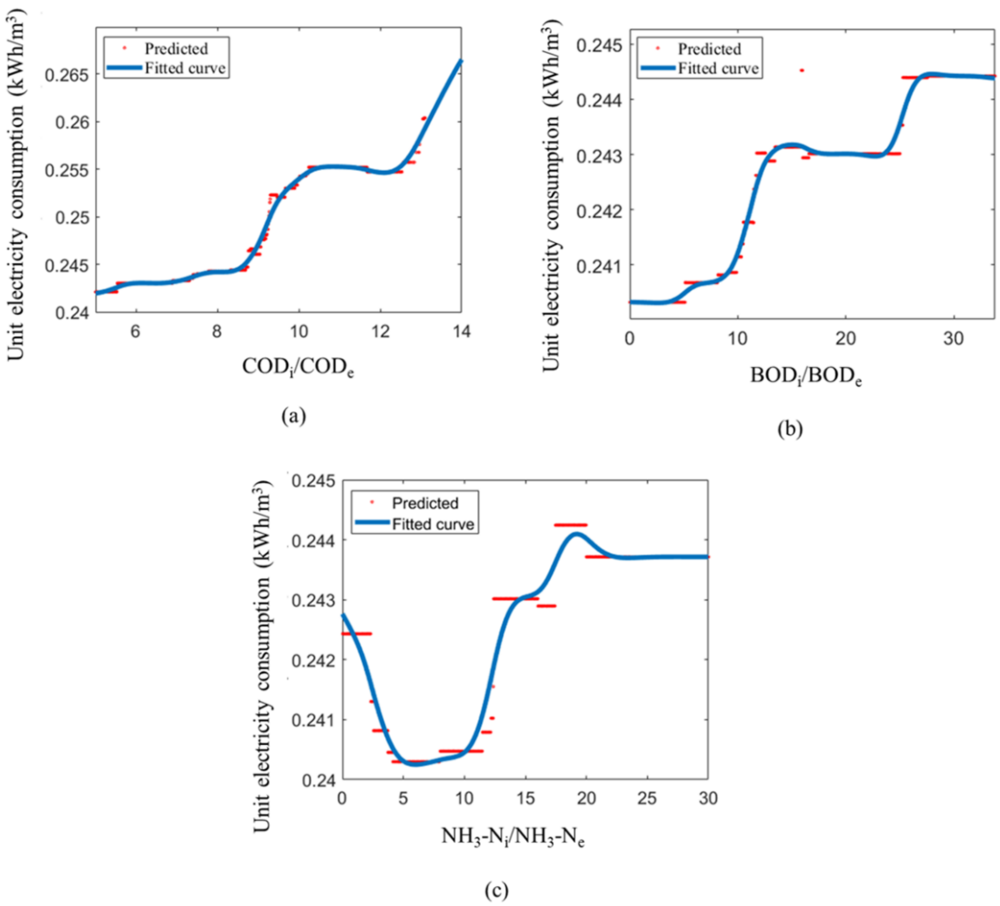

Fig. 8 Predictive model of UEC as a function (a) COD_i/COD_e, (b) BOD_i/BOD_e, and (c) $NH_3\text{-}N_i/NH_3\text{-}N_e$ with other variables constant.

When the independent variable was too large, a flat response occurred in this region of the predictive model due to the lack of data, in other words, when BOD_i/BOD_e or $NH_3\text{-}N_i/NH_3\text{-}N_e$ was too large there will be insufficient data in the database to train the model at this level and the prediction will be the average of these data; therefore, these regions are not discussed in the following analysis. But with the construction of huge WWTPs in China, there will be more data in the future, reducing this issue.

From the predictive model shown in Fig. 8 (a), it is seen that COD_i/COD_e and UEC are positively correlated; that is, the higher the COD reduction ratio, the higher the energy consumption. As expected, a WWTP that seeks to have higher removal efficiency requires more energy. In the predictive model, a relatively flat response occurs when the reduction multiple is less than 9. From the available data, the average COD_i/COD_e was 9.23, near the edge of the region with a minimal slope, indicating that WWTPs had a high energy efficiency in the removal of COD.

As shown in the predictive model in Fig. 8 (b), BOD_i/BOD_e and UEC are positively correlated for BOD_i/BOD_e in the range of 0–15 and 25–30, with a minimal slope (except for the platform) at 0–10, and a region with a slope almost 0 in the 15–25 range. From the available data, the average BOD_i/BOD_e was 18.59, in the middle of the region with a slope almost 0, implying that WWTPs had a high energy efficiency in the treatment of BOD, but requires further improvement.

As shown in Fig. 8 (c), the overall trend of the predictive model of $NH_3\text{-}N_i/NH_3\text{-}N_e$ indicates that the UEC decreases monotonically when $NH_3\text{-}N_i/NH_3\text{-}N_e$ is lower than 5, increases monotonically after $NH_3\text{-}N_i/NH_3\text{-}N_e$ is greater than 10, and finally tends to a constant value. A minimum slope is observed between 5 and 10. Therefore, the optimal value of ammonia nitrogen reduction should be between 5 and 10. When $NH_3\text{-}N_i/NH_3\text{-}N_e$ is less than 10, it is negatively correlated with UEC, which is contrary to the common understanding that a larger reduction multiple, leads to a higher energy consumption. The specific reasons for this require further analysis. However, the possible reasons are as follows: 1. As the importance of $NH_3\text{-}N_i/$ $NH_3\text{-}N_e$ in UEC is low, which causes the difference between the maximum and minimum values of the final prediction result to be ≤ 0.04 kWh, the measuring instrument may not be highly accurate. 2. When NH_3-N_i/NH_3-N_e is 5–10, the reduction multiple is easily achieved. For a lower value, energy consumption may be required to limit the reduction multiple. This study aimed to predict the UEC of WWTPs if the plant upgrades to a higher standard, thus improving the removal ratios (COD_i/COD_e, BOD_i/BOD_e, and NH_3-N_i/ NH_3-N_e). The number of WWTPs applied Class I A were 1,041 and 1,184 in Class I B, with only 162 in Class II. Therefore, we primarily considered the improvement of the discharge standard from Class I B to Class I A. The specific discharge standards are listed in Table 2.

The ratios COD_i/COD_e, BOD_i/BOD_e, and NH_3-N_i/ NH_3-N_e were used as variables in the model, since the discharge standard restricts COD_e, BOD_e, NH_3-N_e. Therefore, the following modifications were adopted in this study:

$$\mathrm{y}_{COD} = \frac{COD_i/COD_e}{x_{COD}} \times COD_e \tag{8}$$

$$\mathrm{y}_{BOD} = \frac{BOD_i/BOD_e}{x_{BOD}} \times BOD_e \tag{9}$$

$$\mathrm{y}_{N} = \frac{NH_3-N_i/\mathrm{NH_3-N}_e}{x_N} \times \mathrm{NH_3-N}_e \tag{10}$$

where y_{COD}: COD_I/COD_E after upgrading to a higher class, x_{COD}: COD_E of Class I A, y_{BOD}: BOD_I/BOD_E after upgrading to a higher class, x_{BOD}: BOD_E of Class I A, y_N: $(NH_3 - N_I/NH_3 - N_E)$ after upgrading to a higher class, and x_N: $(NH_3 - N_E)$ of Class I A.

From equation 4-6, a new value of $\mathrm{COD_i/COD_e}$, $\mathrm{BOD_i/BOD_e}$, $\mathrm{NH_3 - N_i/NH_3 - N_e}$

after the improvement of the discharge standard from Class I B to Class I A was calculated. The model predicted the UEC of WWTPs with the new data.

The results showed that when the discharge quality of WWTPs was upgraded from Class I B to Class I A, the increase in the UEC of WWTPs varied due to the various effluent qualities. The UEC of WWTPs had an average increase of 17%, obtained from Equations 4-6.

3.3.4 Model Validation

As previously mentioned, the R^2 of the model was 0.702. MSE = 0.00662 $(kWh/m^3)^2$, MAPE = 5.74%, RSME = 0.106 kWh/m^3, MAE = 0.0416 kWh/m^3, MedAE = 0.0416 kWh/m^3 and MSLE = 0.00327 (obtained from equations (1) to (6)), which are very low (Weber et al. 2020). These low evaluation metrics indicate that the model for UEC of WWTPs developed in this study was quite accurate. As shown in Fig. 9 and Fig. S2, the actual and predicted UEC exhibit the same trend when the UEC is not too high or too low. The predicted UEC was not accurate when the corresponding actual value was too high or too low because of insufficient data for fully developing the model. In fact, in practice, there are not many cases of too large or too small WWTPs, so the effect of these WWTPs are not significant.

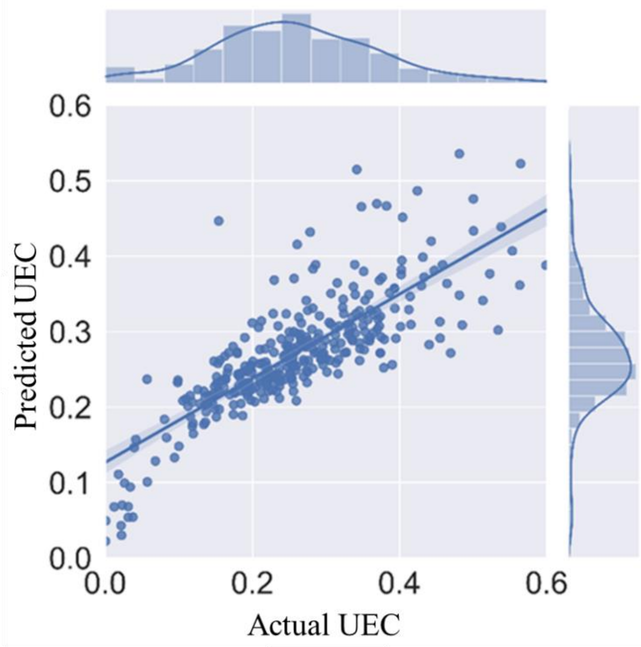

Fig. 9 QQPlot and simplified kernel frequency distribution representing actual and predicted UEC for 347 WWTPs

3.4 Comparison with other approaches

Compared to the Data Envelopment Analysis, Random Forest is more stable when some input data is missing, which means a unified model can be made without considering special cases that one or more variables are missing. Considering the fact that it’s hard to set up a monitoring system that would include thousands of WWTPs across China with exactly the same variables, it would be impossible to build a normalized model using Data Envelopment Analysis.

As mentioned in Section 3.1, the WWTP variables are correlated, so the accuracy of a Multiple Linear Regression model compared to a Random Forest model is relatively low. In this study, Multiple Linear Regression model was considered, but the R^2 (0.147) was too low. In comparison, the random forest model can achieve a much higher R^2 (0.702). Therefore, Random Forest is more suitable to build the model than Data Envelopment Analysis or Multiple Linear Regression.

4. Conclusion

In this study, an energy consumption model for WWTPs was developed using machine learning. The UEC of a WWTP can be predicted with a few key parameters by the model using the random forest algorithm. It can also predict the UEC of a WWTP for policy formulation and improvement of sewage treatment standards. This model can be a useful tool for investigating the water-energy nexus in WWTPs. Although the particular model in this study is based on data from Chinese WWTPs, it can be easily applied to WWTPs worldwide by changing the input data. In this study, we didn't investigate the influence of local climate and treatment technologies due to insufficient data, which are also very important and deserve further research in the future.

References

Bagherzadeh, F., M. J. Mehrani, M. Basirifard and J. Roostaei (2021). "Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance." Journal of Water Process Engineering **41**: 102033.

Behera, C. R., R. Al, K. V. Gernaey and G. Sin (2020). "A process synthesis tool for WWTP – An application to design sustainable energy recovery facilities." Chemical Engineering Research and Design **156**: 353-370.

Belgiu, M. and L. Dragut (2016). "Random forest in remote sensing: A review of applications and future directions." Isprs Journal of Photogrammetry and Remote Sensing **114**: 24-31.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Chen, W., S. Zhang, R. W. Li and H. Shahabi (2018). "Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naive Bayes tree for landslide susceptibility modeling." Science of the Total Environment **644**: 1006-1018.

Cheng, L., J. De Vos, P. Zhao, M. Yang and F. Witlox (2020). "Examining non-linear built environment effects on elderly's walking: A random forest approach." Transportation Research Part D: Transport and Environment **88**: 102-552.

Construction Standard of Urban Sewage Treatment Project: Ministry of Construction of the People's Republic of China; 2001.

Daw, J., K. Hallett, J. Dewolfe and I. Venner (2012). "Energy Efficiency Strategies for Municipal Wastewater Treatment Facilities." Office of Scientific & Technical Information Technical Reports.

Duro, D. C., S. E. Franklin and M. G. Dube (2012). "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery." Remote Sensing of Environment **118**: 259-272.

E. Açıkkalp ve S. Yerel Kandemir (2018). "Optimum insulation thickness of the piping system with combined economic and environmental method." Energy Sources Part A-Recovery Utilization and Environmental Effects, **40** (23): 2876-2885.

EPRI, Electricity use and management in the municipal water supply and wastewater industries, 2013, 5-16.

Gazendam, E., B. Gharabaghi, J. D. Ackerman and H. Whiteley (2016). "Integrative neural networks models for stream assessment in restoration projects." Journal of Hydrology: 339-350.

Habib, R. Z., T. Thiemann and R. A. Kendi (2020). "Microplastics and Wastewater Treatment Plants—A Review." Journal of Water Resource and Protection **12**(1): 1-35.

Hernandez-del-Olmo, F., E. Gaudioso, N. Duro and R. Dormido (2019). "Machine Learning Weather Soft-Sensor for Advanced Control of Wastewater Treatment Plants." Sensors **19**(14).

Heslot, N., D. Akdemir, M. E. Sorrells and J. L. Jannink (2014). "Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions." Theoretical and Applied Genetics **127**(2): 463-480.

Hernández-Sancho, F. and R. Sala-Garrido (2009). "Technical efficiency and cost analysis in wastewater treatment processes: A DEA approach." Desalination **249**(1): 230-234.

Li, W., X. Shi, S. Zhang and G. Qi (2019). "Modelling of ammonia recovery from wastewater by air stripping in rotating packed beds." Science of The Total Environment **702**: 134-971.

Longo, S., B. M. d'Antoni, M. Bongards, A. Chaparro, A. Cronrath, F. Fatone, J. M. Lema, M. Mauricio-Iglesias, A. Soares and A. Hospido (2016). "Monitoring and diagnosis of energy consumption in wastewater treatment plants. A state of the art and proposals for improvement." Applied Energy **179**: 1251-1268.

Loozen, Y., K. T. Rebel, S. M. de Jong, M. Lu, S. V. Ollinger, M. J. Wassen and D. Karssenberg (2020). "Mapping canopy nitrogen in European forests using remote sensing and environmental variables with the random forests method." Remote Sensing of Environment **247**: 111933.

Luo, L., M. Dzakpasu, B. Yang, W. Zhang, Y. Yang and X. C. Wang (2019). "A novel index of total oxygen demand for the comprehensive evaluation of energy consumption for urban wastewater treatment." Applied Energy **236**: 253-261.

Ministry of Construction of China (2001). Construction standard of urban sewage treatment project. No.77.

Mizuta, K. and M. Shimada (2010). "Benchmarking energy consumption in municipal wastewater treatment plants in Japan." Water Scienceence & Technology A Journal of the International Association on Water Pollution Research **62**(10): 2256-2262.

Mjalli, F. S., S. Al-Asheh and H. E. Alfadala (2007). "Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance." Journal of Environmental Management **83**(3): 329-338.

Molinos-Senante, M., F. Hernandez-Sancho and R. Sala-Garrido (2014). "Benchmarking in wastewater treatment plants: a tool to save operational costs." Clean Technologies and Environmental Policy **16**(1): 149-161.

Pamulaparty, L., C. V. G. Rao and M. S. Rao (2017). "Critical review of various near-duplicate detection methods in web crawl and their prospective application in drug discovery." International Journal of Biomedical Engineering and Technology **25**(2-4): 212-226.

Paul, G. C., S. Saha and K. G. Ghosh (2020). "Assessing the soil quality of Bansloi river basin, eastern India using soil-quality indices (SQIs) and Random Forest machine learning technique." Ecological Indicators **118**: 106-804.

Perez, V. M., J. M. M. Fernandez, J. V. Balsera and C. A. Alvarez (2021). "A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs." Water **13**(9): 1237.

Piotrowski, R. and T. Ujazdowski (2020). "Designing Control Strategies of Aeration System in Biological WWTP." Energies **13**(14): 1-17.

Sabia, G., L. Petta, F. Avolio and E. Caporossi (2020). "Energy saving in wastewater treatment plants: A methodology based on common key performance indicators for the evaluation of plant energy performance, classification and benchmarking." Energy Conversion and Management **220**: 113-067.

Scott, C. A., S. A. Pierce, M. J. Pasqualetti, A. L. Jones, B. E. Montz and J. H. Hoover (2011). "Policy and institutional dimensions of the water–energy nexus." Energy Policy **39**(10): 6622-6630.

Siddiqi, A. and L. D. Anadon (2011). "The water–energy nexus in Middle East and North Africa." Energy Policy **39**(8): 4529-4540.

Smith, K., S. Guo, Q. Zhu, X. Dong and S. Liu (2019). "An evaluation of the environmental benefit and energy footprint of China's stricter wastewater standards: Can benefit be increased?" Journal of Cleaner Production **219**: 723-733.

Song, R. S., A. A. Keller and S. Suh (2017). "Rapid Life-Cycle Impact Screening Using Artificial Neural Networks." Environmental Science & Technology **51**(18): 10777-10785.

Tamjidi Farahbakhsh, M. and M. Chahartaghi (2020). "Performance analysis and economic assessment of a combined cooling heating and power (CCHP) system in wastewater treatment plants (WWTPs)." Energy Conversion and Management **224**: 113-351.

Tian, H., P. Bai, Y. Tan, Z. Li, D. Peng, X. Xiao, H. Zhao, Y. Zhou, W. Liang and L. Zhang (2020). "A new method to detect methylation profiles for forensic body fluid identification combining

ARMS-PCR technique and random forest model." Forensic Science International: Genetics **49**: 102-371.

Torregrossa, D., U. Leopold, F. Hernandez-Sancho and J. Hansen (2018). "Machine learning for energy cost modelling in wastewater treatment plants." Journal of Environmental Management **223**(OCT.1): 1061-1067.

Torregrossa, D., G. Schutz, A. Cornelissen, F. Hernández-Sancho and J. Hansen (2016). "Energy saving in WWTP: Daily benchmarking under uncertainty and data availability limitations." Environmental Research **148**: 330-337.

Trenouth, W. R., B. Gharabaghi and H. Farghaly (2018). "Enhanced roadside drainage system for environmentally sensitive areas." Science of The Total Environment **610-611**: 613-622.

Wang, H. Z., Z. X. Lei, X. Zhang, B. Zhou and J. C. Peng (2019). "A review of deep learning for renewable energy forecasting." Energy Conversion and Management **198**: 111799.

Weber, V. A. M., F. d. L. Weber, A. d. S. Oliveira, G. Astolfi, G. V. Menezes, J. V. de Andrade Porto, F. P. C. Rezende, P. H. d. Moraes, E. T. Matsubara, R. G. Mateus, T. L. A. C. de Araújo, L. O. C. da Silva, E. Q. A. de Queiroz, U. G. P. de Abreu, R. da Costa Gomes and H. Pistori (2020). "Cattle weight estimation using active contour models and regression trees Bagging." Computers and Electronics in Agriculture **179**: 105-804.

Wei, J., W. Huang, Z. Q. Li, W. H. Xue, Y. R. Peng, L. Sun and M. Cribb (2019). "Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach." Remote Sensing of Environment **231**: 111221.

Wylie, B. K., N. J. Pastick, J. J. Picotte and C. A. Deering (2019). "Geospatial data mining for digital raster mapping." Giscience & Remote Sensing **56**(3): 406-429.

Xing, L., M. G. Xue and M. S. Hu (2019). "Dynamic simulation and assessment of the coupling coordination degree of the economy-resource-environment system: Case of Wuhan City in China." Journal of Environmental Management **230**: 474-487.

Yeşilkanat, C. M. (2020). "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm." Chaos, Solitons & Fractals **140**: 110-210.

Zhu, Y. A., W. Y. Xu, G. L. Luo, H. L. Wang, J. J. Yang and W. Lu (2020). "Random Forest enhancement using improved Artificial Fish Swarm for the medial knee contact force prediction." Artificial Intelligence in Medicine **103**: 101811.

Yang, J., B. Chen (2021). “Energy efficiency evaluation of wastewater treatment plants (WWTPs) based on data envelopment analysis” Applied Energy **289**: 116680

Huang, R., Z. Shen, H. Wang, J. Xu, Z. Ai, H. Zheng, R. Liu (2021). “Evaluating the energy efficiency of wastewater treatment plants in the Yangtze River Delta: Perspectives on regional discrepancies” Applied Energy **297**: 117087

Xu, J., P. Luo, B. Lu, H. Wang, X. Wang, J. Wu, J. Yan (2018). “Energy-water nexus analysis of wastewater treatment plants (WWTPs) in China based on statistical methodologies” Energy Procedia **152**: 259-264.

Hernández-Chover, V., Á. Bellver-Domingo and F. Hernández-Sancho (2018). "Efficiency of wastewater treatment facilities: The influence of scale economies." Journal of Environmental Management **228**: 77-84.

Gerami, N., A. Ghasemi, A. Lotfi, L. G. Kaigutha and M. Marzband (2021). "Energy consumption modeling of production process for industrial factories in a day ahead scheduling with demand response." Sustainable Energy, Grids and Networks **25**: 100420.

Zhong, S., K. Zhang, M. Bagheri, J. G. Burken, A. Gu, B. Li, X. Ma, B. L. Marrone, Z. J. Ren, J. Schrier, W. Shi, H. Tan, T. Wang, X. Wang, B. M. Wong, X. Xiao, X. Yu, J.-J. Zhu and H. Zhang (2021). "Machine Learning: New Ideas and Tools in Environmental Science and Engineering." Environmental Science & Technology **55**(19): 12741-12754.

Gupta, S., D. Aga, A. Pruden, L. Zhang and P. Vikesland (2021). "Data Analytics for Environmental Science and Engineering Research." Environmental Science & Technology **55**(16): 10895-10907.