ACS ES&T Water
pubs.acs.org/estwater
Article

Novel Machine Learning-Based Energy Consumption Model of Wastewater Treatment Plants

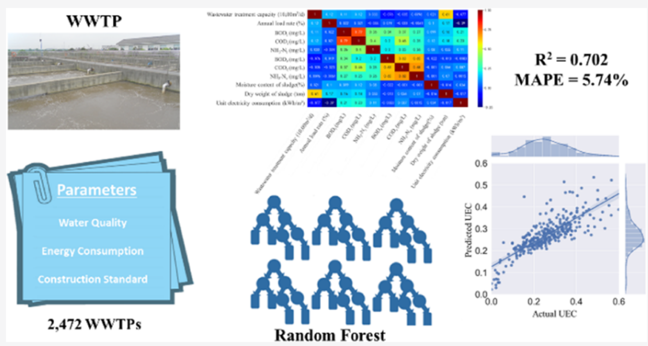
Shike Zhang, Hongtao Wang,* and Arturo A. Keller

Cite This: *ACS EST Water* 2021, 1, 2531−2540

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 increased from class I B to class I A (as 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*

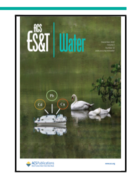
1. INTRODUCTION

Improving the energy efficiency of wastewater treatment plants (WWTPs) is receiving increasing attention as saving energy can help reduce economic costs and conserve the resources and environment.[1] With continuous development and accelerating urbanization in the society, wastewater discharge is rapidly increasing[2] 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.[3] 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.[3] What is more, the U.S. municipal wastewater treatment systems use approximately 30.2 billion kW h per year, which is about 0.8% of the total electricity use in the U.S.[4] In recent years, several energy evaluation methods have been proposed[5−7] 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.[8]

Many stakeholders have been exploring solutions to reduce the energy consumption of WWTPs, such as equipment renewal and maintenance,[9] energy recovery,[10] and technical process improvements.[11] Previous studies have 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.[3] 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.[12] 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 regard to energy efficiency[13] or WWTP operations.[14] Machine learning can be utilized in 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 become available, with the ability to adopt

Received: August 12, 2021
Revised: November 7, 2021
Accepted: November 8, 2021
Published: November 19, 2021



non-linear relationships. Once a set of inputs and the 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.[15−18]

Random forest (RF) is an ensemble learning algorithm used for classification,[19] regression,[20] and other tasks.[21] During training, numerous decision trees are generated to operate and finally obtain prediction results;[22] therefore, RF has a high prediction accuracy and is not prone to overfitting.[23] RF is one of the most popular methods in data mining[24] and big data fields;[25] it has the advantages of a fast-training speed and is suitable for processing high-dimensional data.[26] The method has been widely used in several other fields, such as medicine,[27] criminal investigation,[22] and architecture.[28] RF has also been applied to environmental engineering, such as for mapping canopy nitrogen[29] and in environmental assessment.[30] However, RF is rarely used to analyze and predict energy consumption of WWTPs.[31,32] Data envelopment analysis[33,34] and multiple linear regression[35] 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 are 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 2472 WWTPs in China, employing the RF 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 were obtained from the 2015 Urban Drainage Yearbook of China. A total of 2472 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.[36] Related studies[2,37−39] commonly use electricity intensity (kW h/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,[2,37−39] 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 plants (GB 18918-2002).

2.2. Data Cleaning. To import data to develop the model, all numerical data (including int64, float64) were converted to float64, all non-numerical data (including string and object, such as the discharge standard) were transferred into the object, and all the default data were 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 2387.

Although there were some outliers, the database is large and the RF model is good at dealing with this situation, so there was no need to eliminate them. Essentially, the basic learner of RF is robust to outliers, which makes the RF 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 of 0 was assigned. For example (Figure 1),

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

Figure 1. Matrix of discharge standard in WWTPs.

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.

2.4. Random Forest. A preliminarily evaluation of the relationship between UEC (kW h/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.

https://doi.org/10.1021/acsestwater.1c00283
ACS EST Water 2021, 1, 2531−2540

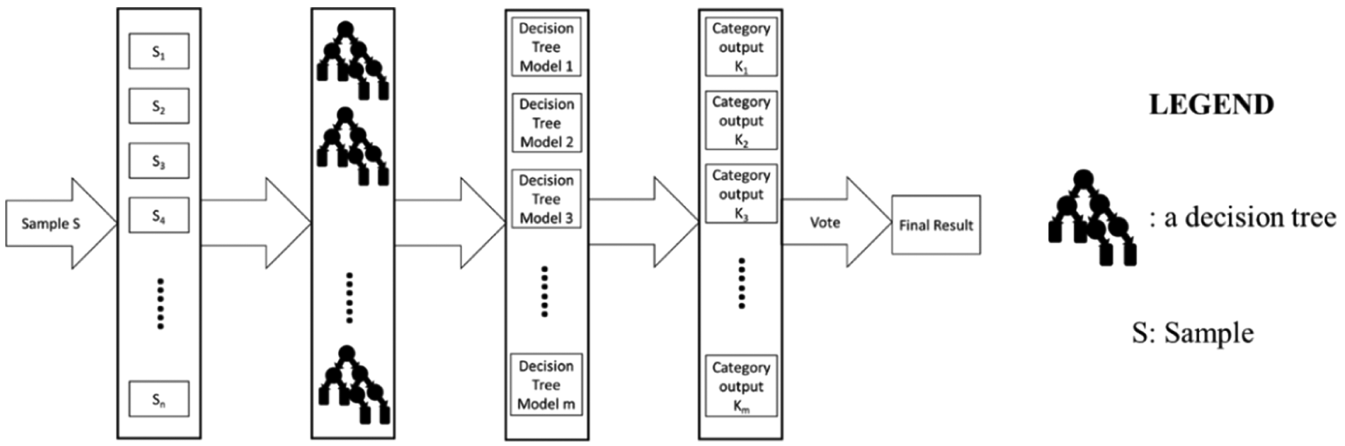


Figure 2. Process flow of the RF method.

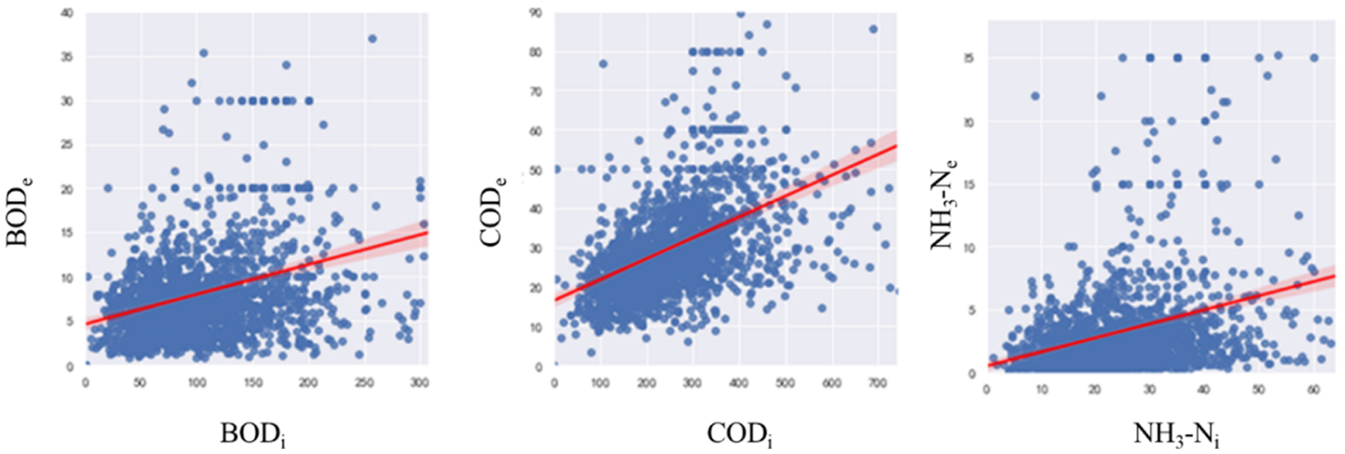


Figure 3. Scatter plot and linear regression curve of (a) BOD_i/BOD_e, (b) COD_i/COD_e, and (c) $NH_3{-}N_i/NH_3{-}N_e$.

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 RF, boosting tree, gradient boosting decision tree, and XGBoost. These algorithms are actually a combination of different algorithmic frameworks and decision trees (for details, see Table S1), so they perform quite similarly. We selected RF because it is the only algorithm that can show us the importance of each variable, which is very valuable for the subsequent analysis.

Therefore, an RF 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 RF approach are shown in Figure 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. RF 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.

2.5. Model Validation. RF 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 used, RF 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)[40,41] as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n}(\mathrm{actual}\ (t) - \mathrm{predicted}\ (t))^2 \tag{1}$$

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

https://doi.org/10.1021/acsestwater.1c00283
ACS EST Water 2021, 1, 2531−2540

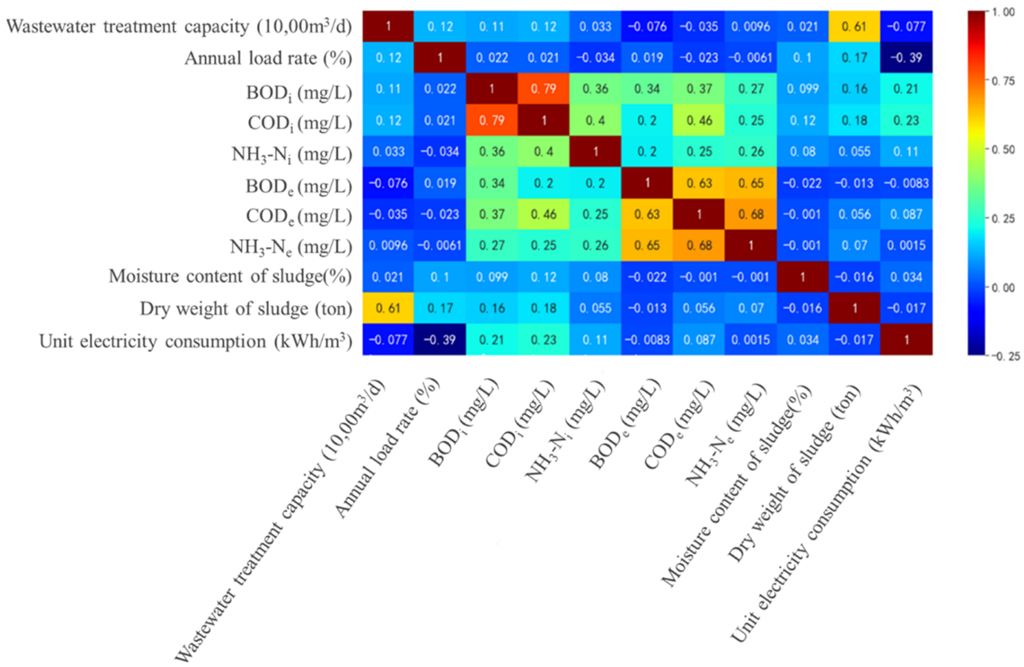


Figure 4. Correlation thermodynamic diagram of variables.

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

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

$$\mathrm{MedAE} = \mathrm{median}(|\mathrm{actual}\ (t_1) - \mathrm{predicted}\ (t_1)|, ..., |\mathrm{actual}\ (t_n) - \mathrm{predicted}\ (t_n)|) \tag{5}$$

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

where n is the number of decision tree models, 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.

In Figure 3, each point in the figure represents a WWTP data set, and it may be noted that the BOD$_i$ and BOD$_e$, COD$_i$ and COD$_e$, and 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$_e$, and 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$_e$, and 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. Importance of Features. The importance of a feature X in an RF was calculated as follows:

A For each decision tree in an RF, 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 RF, then the importance of the feature is given as

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

This expression can be used as a measure of the importance of the 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 Figure 4.

In Figure 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 the influent and effluent (Figure 4).

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

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

As shown in Table 1 and Figure 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.[42] This is followed by the annual load rate, which is also expected to be a major factor.[42] 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.

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 chemical oxygen demand (COD) and biological oxygen demand (BOD) highly affected the energy consumption of WWTPs.[43] This is consistent with other studies, and the pollution load is consistent with the energy consumption load of WWTPs.[42] COD_i/COD_e is significantly more important than BOD_i/BOD_e since the pollutants measured by BOD are a 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 belong to both 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 that of 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; this 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[44] 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.[45] Considering the fact that the smaller the number of highly correlated variables, the more convenient is the practical application of the model, and 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 Figure 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 are 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, 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[a]

parameter	class I A	class I B
COD_e (mg/L)	50	60
BOD_e (mg/L)	10	20
NH3−N_e (mg/L)	5(8)	8(15)

[a]Note: the value in the bracket means the standard at temperature ≤ 12 °C, 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 $NH_3{-}N_i/NH_3{-}N_e$) 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 Figure 5, it is evident that the wastewater treatment capacity is negatively related to UEC, and for wastewater treatment capacities from 10,000 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 a minimal decrease in UEC, which is a consideration for the design of WWTPs. The overall trend is consistent with the finding of previous studies.[33,34] What is more, this finding also follows the scale economy of WWTPs.[46]

The construction scale of WWTPs in China can be divided into five categories [according to the *Ministry of Construction of China (2001). Construction standard of urban sewage treatment project. no. 77*], as shown in Table 3:

From the predicted data (Figure 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

https://doi.org/10.1021/acsestwater.1c00283
ACS EST Water 2021, 1, 2531−2540

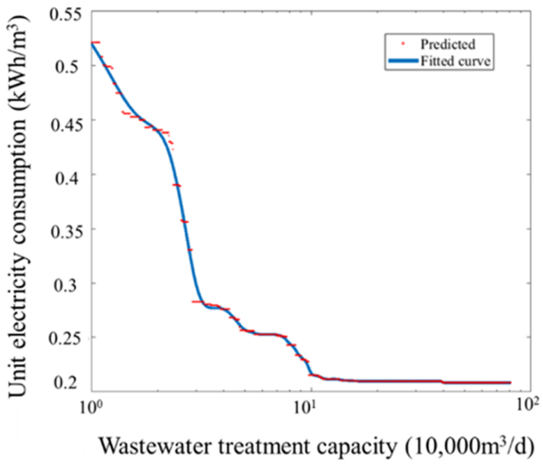


Figure 5. Predictive model of UEC as a function of wastewater treatment capacity with other variables constant.

Table 3. Standard of the 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

is 9.91% of the total WWTPs considered in this study (Figure 6).

The ordinate of a point on the curve in Figure 6a shows the proportion of the scale that is reflected by the abscissa of total WWTPs in China. The ordinate of a point on the curve in Figure 6b 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 Figure 6a, the curve peaks in category V, which means that 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 Figure 6b, the tangent slope of the curve also shows that category V includes most of the WWTPs in China. In short, Figure 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, and it reflects the divergence of design and actual usage of WWTPs. As shown in Figure 7, the annual load rate has a

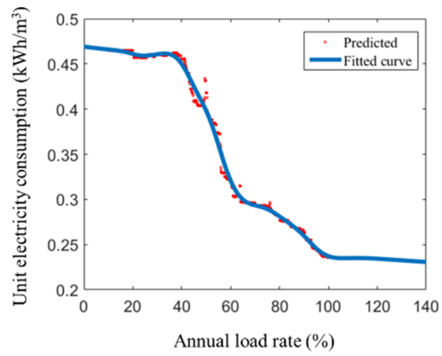


Figure 7. Predictive model of UEC as a function of annual load rate with other variables constant.

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.,[34] and it also follows the rule of extensive models of different factory managements.[47] This finding indicates that it will be better to do more studies on the amount of wastewater needed to be treated in one area before designing the treatment capacity of the WWTP.

3.3.3. Reduction Ratios. In this section, the effects of COD_i/COD_e, BOD_i/BOD_e, and $NH_3{-}N_i/NH_3{-}N_e$ 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 are generally in line with Huang et al.,[34] while the

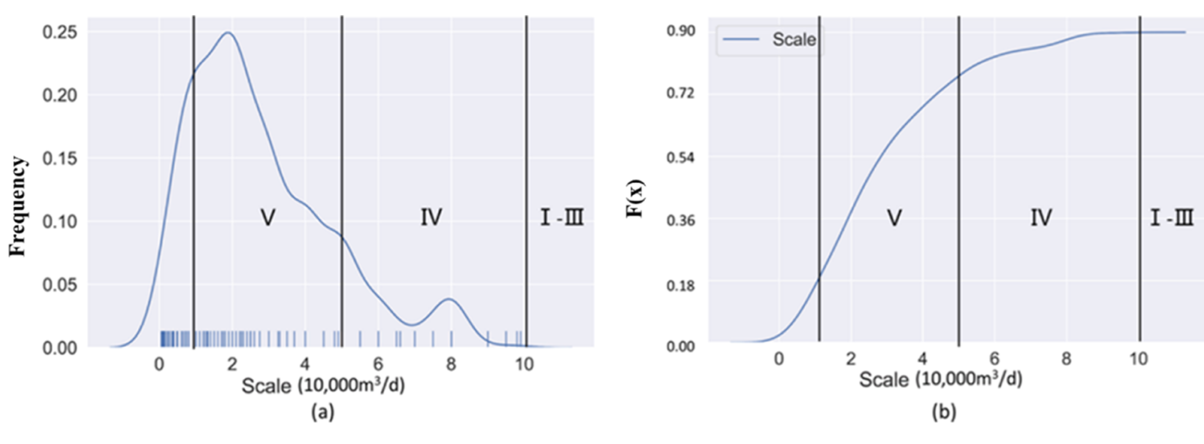


Figure 6. (a) Kernel frequency distribution and (b) probability distribution of wastewater treatment capacity of WWTPs in the model (to make the figure clearer, all the WWTPs with a wastewater treatment capacity above 100,000 m^3/d were not counted in the figure).

https://doi.org/10.1021/acsestwater.1c00283
ACS EST Water 2021, 1, 2531−2540

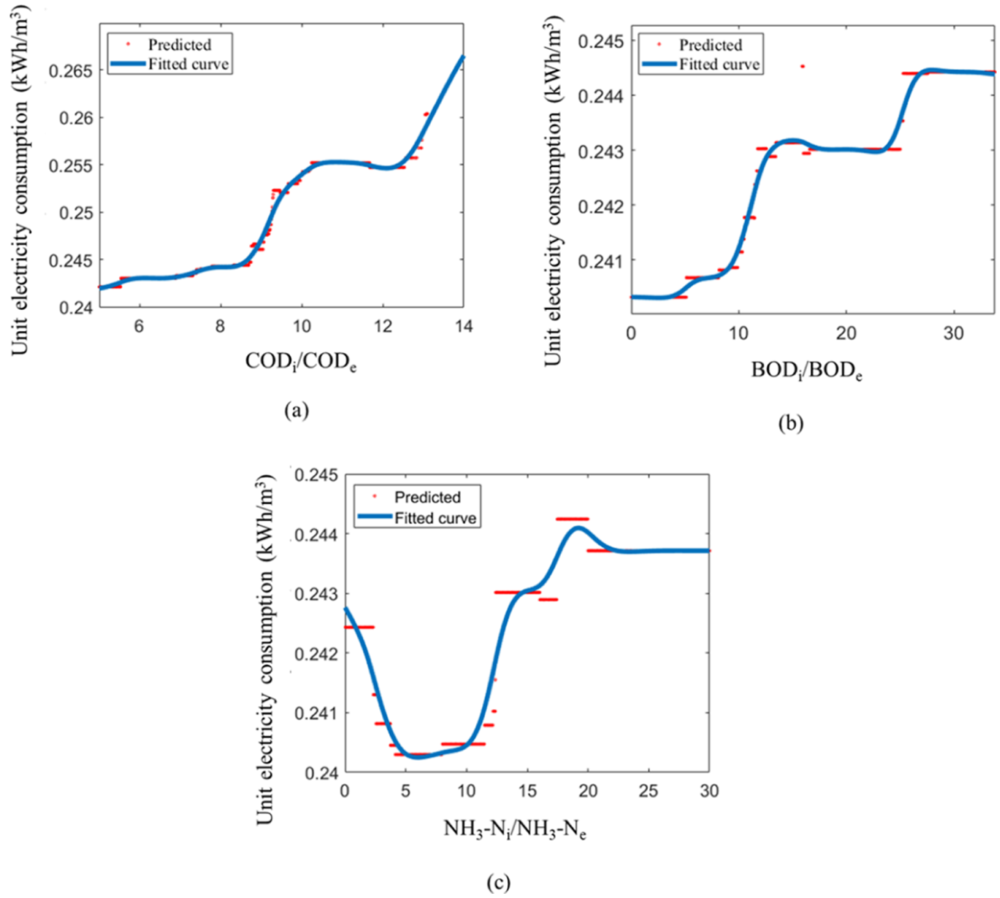


Figure 8. Predictive model of UEC as a function of (a) COD_i/COD_e, (b) BOD_i/BOD_e, and (c) NH_3-N_i/NH_3-N_e with other variables constant.

trend of NH_3-N is a little conflict with the common sense, and we are going to explain it in the following discussion.

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-N_i/NH_3-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. However, 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 Figure 8a, 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 a 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 Figure 8b, 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 of 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 of almost 0, implying that WWTPs had a high energy efficiency in the treatment of BOD but require further improvement.

As shown in Figure 8c, the overall trend of the predictive model of NH_3-N_i/NH_3-N_e indicates that the UEC decreases monotonically when NH_3-N_i/NH_3-N_e is lower than 5, increases monotonically after NH_3-N_i/NH_3-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-N_i/NH_3-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-N_i/NH_3-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 kW h, 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 was 1041 in class I A and 1184 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, and NH_3-N_e. Therefore, the following modifications were adopted in this study:

https://doi.org/10.1021/acsestwater.1c00283
ACS EST Water 2021, 1, 2531−2540

$$y_{COD} = \frac{COD_i/COD_e}{x_{COD}} \times COD_e \tag{8}$$

$$y_{BOD} = \frac{BOD_i/BOD_e}{x_{BOD}} \times BOD_e \tag{9}$$

$$y_N = \frac{NH_3 - N_i/NH_3 - N_e}{x_N} \times NH_3 - N_e \tag{10}$$

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 eqs 4−6, a new value of COD_i/COD_e, BOD_i/BOD_e, $NH_3{-}N_i/NH_3{-}N_e$ 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 eqs 4−6.

3.3.4. Model Validation. As previously mentioned, the R^2 of the model was 0.702. MSE = 0.00662 (kW h/m^3)^2, MAPE = 5.74%, RSME = 0.106 kW h/m^3, MAE = 0.0416 kW h/m^3, MedAE = 0.0416 kW h/m^3, and MSLE = 0.00327 (obtained from 1), which are very low.[48] These low evaluation metrics indicate that the model for UEC of WWTPs developed in this study was quite accurate. As shown in Figure 9 and Figure S2, the actual and predicted UECs 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 effects of these WWTPs are not significant.

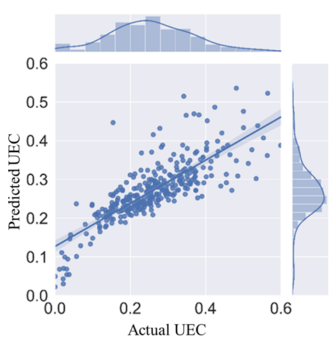


Figure 9. QQPlot and simplified Kernel frequency distribution representing actual and predicted UEC for 347 WWTPs.

3.4. Comparison with Other Approaches. Compared to the data envelopment analysis, RF is more stable when some input data are missing, which means that a unified model can be made without considering special cases that one or more variables are missing. Considering the fact that it is 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 an RF model is relatively low. In this study, the multiple linear regression model was considered, but the R^2 (0.147) was too low. In comparison, the RF model can achieve a much higher R^2 (0.702). Therefore, RF is more suitable to build the model than data envelopment analysis or multiple linear regression.

4. CONCLUSIONS

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 RF 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 did not 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.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestwater.1c00283.

Detailed steps of the RF algorithm, algorithm principle, variables and their importance, detailed Kernel frequency distribution representing actual and predicted UEC for 347 WWTPs, and biological processes of WWTPs in China (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Hongtao Wang − *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, Shanghai 200092, China; Shanghai Institute of Pollution Control and Ecological Security, Tongji University, Shanghai 200092, China;* orcid.org/0000-0002-8066-475X; Email: hongtao@tongji.edu.cn

Authors

Shike Zhang − *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, Shanghai 200092, China*

Arturo A. Keller − *Bren School of Environmental Science and Management, UC Santa Barbara, Santa Barbara, California 93106, United States;* orcid.org/0000-0002-7638-662X

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestwater.1c00283

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Yerel Kandemir, S.; Açıkkalp, E. "Optimum insulation thickness of the piping system with combined economic and environmental method". *Energy Sources, Part A* **2018**, *40*, 2876−2885.
(2) Habib, R. Z.; Thiemann, T.; Kendi, R. A. "Microplastics and Wastewater Treatment Plants—A Review". *J. Water Resour. Protect.* **2020**, *12*, 1−35.
(3) Sabia, G.; Petta, L.; Avolio, F.; Caporossi, E. "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 Convers. Manage.* **2020**, *220*, 113067.
(4) EPRI. *Electricity Use and Management in the Municipal Water Supply and Wastewater Industries*, 2013; pp 5−16.
(5) Hernández-Sancho, F.; Sala-Garrido, R. "Technical efficiency and cost analysis in wastewater treatment processes: A DEA approach". *Desalination* **2009**, *249*, 230−234.
(6) Mizuta, K.; Shimada, M. Benchmarking energy consumption in municipal wastewater treatment plants in Japan. *Water Sci. Technol.* **2010**, *62*, 2256−2262.
(7) Molinos-Senante, M.; Hernandez-Sancho, F.; Sala-Garrido, R. Benchmarking in wastewater treatment plants: a tool to save operational costs. *Clean Technol. Environ. Policy* **2014**, *16*, 149−161.
(8) Torregrossa, D.; Schutz, G.; Cornelissen, A.; Hernández-Sancho, F.; Hansen, J. Energy saving in WWTP: Daily benchmarking under uncertainty and data availability limitations. *Environ. Res.* **2016**, *148*, 330−337.
(9) Daw, J.; Hallett, K.; Dewolfe, J.; Venner, I. *Energy Efficiency Strategies for Municipal Wastewater Treatment Facilities*; Office of Scientific & Technical Information Technical Reports, 2012.
(10) Behera, C. R.; Al, R.; Gernaey, K. V.; Sin, G. A process synthesis tool for WWTP - An application to design sustainable energy recovery facilities. *Chem. Eng. Res. Des.* **2020**, *156*, 353−370.
(11) Tamjidi Farahbakhsh, M.; Chahartaghi, M. "Performance analysis and economic assessment of a combined cooling heating and power (CCHP) system in wastewater treatment plants (WWTPs)". *Energy Convers. Manage.* **2020**, *224*, 113351.
(12) Smith, K.; Guo, S.; Zhu, Q.; Dong, X.; Liu, S. An evaluation of the environmental benefit and energy footprint of China's stricter wastewater standards: Can benefit be increased? *J. Clean. Prod.* **2019**, *219*, 723−733.
(13) Wang, H.; Lei, Z.; Zhang, X.; Zhou, B.; Peng, J. A review of deep learning for renewable energy forecasting. *Energy Convers. Manage.* **2019**, *198*, 111799.
(14) Hernandez-del-Olmo, F.; Gaudioso, E.; Duro, N.; Dormido, R. "Machine Learning Weather Soft-Sensor for Advanced Control of Wastewater Treatment Plants". *Sensors* **2019**, *19*(), 3139. DOI: 10.3390/s19143139
(15) Heslot, N.; Akdemir, D.; Sorrells, M. E.; Jannink, J.-L. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* **2014**, *127*, 463−480.
(16) Song, R.; Keller, A. A.; Suh, S. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environ. Sci. Technol.* **2017**, *51*, 10777−10785.
(17) Xing, L.; Xue, M.; Hu, M. Dynamic simulation and assessment of the coupling coordination degree of the economy-resource-environment system: Case of Wuhan City in China. *J. Environ. Manage.* **2019**, *230*, 474−487.
(18) Zhu, Y.; Xu, W.; Luo, G.; Wang, H.; Yang, J.; Lu, W. Random Forest enhancement using improved Artificial Fish Swarm for the medial knee contact force prediction. *Artif. Intell. Med.* **2020**, *103*, 101811.
(19) Duro, D. C.; Franklin, S. E.; Dubé, M. G. 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 Sens. Environ.* **2012**, *118*, 259−272.
(20) Wei, J.; Huang, W.; Li, Z.; Xue, W.; Peng, Y.; Sun, L.; Cribb, M. Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. *Remote Sens. Environ.* **2019**, *231*, 111221.
(21) Chen, W.; Zhang, S.; Li, R.; Shahabi, H. Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naive Bayes tree for landslide susceptibility modeling. *Sci. Total Environ.* **2018**, *644*, 1006−1018.
(22) Tian, H.; Bai, P.; Tan, Y.; Li, Z.; Peng, D.; Xiao, X.; Zhao, H.; Zhou, Y.; Liang, W.; Zhang, L. A new method to detect methylation profiles for forensic body fluid identification combining ARMS-PCR technique and random forest model. *Forensic Sci. Int.: Genet.* **2020**, *49*, 102371.
(23) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(24) Wylie, B. K.; Pastick, N. J.; Picotte, J. J.; Deering, C. A. Geospatial data mining for digital raster mapping. *GIScience Remote Sens.* **2019**, *56*, 406−429.
(25) Pamulaparty, L.; Rao, C. V. G.; Rao, M. S. Critical review of various near-duplicate detection methods in web crawl and their prospective application in drug discovery. *Int. J. Biomed. Eng. Technol.* **2017**, *25*, 212.
(26) Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogrammetry Remote Sens.* **2016**, *114*, 24−31.
(27) Yeşilkanat, C. M. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos, Solit. Fractals* **2020**, *140*, 110−210.
(28) Cheng, L.; De Vos, J.; Zhao, P.; Yang, M.; Witlox, F. Examining non-linear built environment effects on elderly's walking: A random forest approach. *Transport. Res. Transport Environ.* **2020**, *88*, 102552.
(29) Loozen, Y.; Rebel, K. T.; de Jong, S. M.; Lu, M.; Ollinger, S. V.; Wassen, M. J.; Karssenberg, D. Mapping canopy nitrogen in European forests using remote sensing and environmental variables with the random forests method. *Remote Sens. Environ.* **2020**, *247*, 111933.
(30) Paul, G. C.; Saha, S.; Ghosh, K. G. Assessing the soil quality of Bansloi river basin, eastern India using soil-quality indices (SQIs) and Random Forest machine learning technique. *Ecol. Indicat.* **2020**, *118*, 106804.
(31) Bagherzadeh, F.; Mehrani, M.-J.; Basirifard, M.; Roostaei, J. Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. *J. Water Process Eng.* **2021**, *41*, 102033.
(32) Pérez, V. M.; Fernández, J. M. M.; Balsera, J. V.; Álvarez, C. A. A Random Forest Model for the Prediction of FOG Content in Inlet Wastewater from Urban WWTPs. *Water* **2021**, *13*, 1237.
(33) Yang, J.; Chen, B. Energy efficiency evaluation of wastewater treatment plants (WWTPs) based on data envelopment analysis. *Appl. Energy* **2021**, *289*, 116680.
(34) Huang, R.; Shen, Z.; Wang, H.; Xu, J.; Ai, Z.; Zheng, H.; Liu, R. Evaluating the energy efficiency of wastewater treatment plants in the Yangtze River Delta: Perspectives on regional discrepancies. *Appl. Energy* **2021**, *297*, 117087.
(35) Xu, J.; Luo, P.; Lu, B.; Wang, H.; Wang, X.; Wu, J.; Yan, J. Energy-water nexus analysis of wastewater treatment plants (WWTPs) in China based on statistical methodologies. *Energy Procedia* **2018**, *152*, 259−264.
(36) Scott, C. A.; Pierce, S. A.; Pasqualetti, M. J.; Jones, A. L.; Montz, B. E.; Hoover, J. H. Policy and institutional dimensions of the water−energy nexus. *Energy Pol.* **2011**, *39*, 6622−6630.
(37) Mjalli, F. S.; Al-Asheh, S.; Alfadala, H. E. Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. *J. Environ. Manage.* **2007**, *83*, 329−338.
(38) Gazendam, E.; Gharabaghi, B.; Ackerman, J. D.; Whiteley, H. Integrative neural networks models for stream assessment in restoration projects. *J. Hydrol.* **2016**, *536*, 339−350.
(39) Trenouth, W. R.; Gharabaghi, B.; Farghaly, H. Enhanced roadside drainage system for environmentally sensitive areas. *Sci. Total Environ.* **2018**, *610−611*, 613−622.

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

(41) Gupta, S.; Aga, D.; Pruden, A.; Zhang, L.; Vikesland, P. Data Analytics for Environmental Science and Engineering Research. *Environ. Sci. Technol.* **2021**, *55*, 10895−10907.

(42) Torregrossa, D.; Leopold, U.; Hernández-Sancho, F.; Hansen, J. Machine learning for energy cost modelling in wastewater treatment plants. *J. Environ. Manage.* **2018**, *223*, 1061−1067.

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

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

(45) Piotrowski, R.; Ujazdowski, T. Designing Control Strategies of Aeration System in Biological WWTP. *Energies* **2020**, *13*, 3619.

(46) Hernández-Chover, V.; Bellver-Domingo, Á.; Hernández-Sancho, F. Efficiency of wastewater treatment facilities: The influence of scale economies. *J. Environ. Manage.* **2018**, *228*, 77−84.

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

(48) Weber, V. A. M.; de Lima Weber, F.; da Silva Oliveira, A.; Astolfi, G.; Menezes, G. V.; de Andrade Porto, J. V.; Rezende, F. P. C.; de Moraes, P. H.; Matsubara, E. T.; Mateus, R. G.; de Araújo, T. L. A. C.; da Silva, L. O. C.; de Queiroz, E. Q. A.; de Abreu, U. G. P.; da Costa Gomes, R.; Pistori, H. Cattle weight estimation using active contour models and regression trees Bagging. *Comput. Electron. Agric.* **2020**, *179*, 105−804.

https://doi.org/10.1021/acsestwater.1c00283
ACS EST Water 2021, 1, 2531−2540