

**APPLICATION OF ARTIFICIAL INTELLIGENCE
TO WASTEWATER TREATMENT PLANT OPERATION**

by

Praewa Wongburi

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Civil and Environmental Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON

2021

Date of final oral examination: 07/28/2021

The dissertation is approved by the following members of the Final Oral Committee:

Jae Park. Faculty, Professor, Civil and Environmental Engineering

Greg Harrington. Faculty, Professor, Civil and Environmental Engineering

Chin Wu. Faculty, Professor, Civil and Environmental Engineering

Andrea Hicks. Faculty, Assistant Professor, Civil and Environmental Engineering

Miaoyan Wang. Faculty, Assistant Professor, Statistics

ABSTRACT

In a wastewater treatment plant (WWTP), big data is collected from sensors installed in various unit processes, but limited data is used for operation and regulatory permit requirements. With the advancement in information technology, the data size in wastewater treatment systems has increased significantly. However, WWTPs have not used big data systematically to aid the operation and detect potential operational issues due to the lack of specialized analytical tools.

The objectives of the study were to: (1) develop analytics methods suitable for the management of big data generated in WWTPs, (2) interpret analytics results for extracting meaningful information, (3) implement a recurrent neural network (RNN) and Long Short-Term Memory (LSTM) to predict effluent water quality parameters and Sludge Volume Index (SVI), (4) apply an Explainable Artificial Intelligence (AI) algorithm to determine causes of predicted values, and (5) propose a real-time control using a predictive model to monitor and optimize the operation of WWTPs.

The predictive AI models in WWTPs were developed by applying big data analytics, statistical analysis, and RNN algorithms with an Explainable AI algorithm. The models successfully and accurately predicted the effluent water quality data and a key operational parameter, SVI. Furthermore, the Explainable AI algorithm provided insight into which influent parameters affected higher predicted effluent concentrations and SVI on a specific day, allowing operators to take corrective actions.

From a WWTP's operational data analysis, the RNN model successfully predicted the effluent concentrations of BOD₅, total nitrogen (TN) and total phosphorus (TP), and SVI. Furthermore, the

Explainable AI analysis found that higher influent NH_3N values lead to higher effluent BOD_5 , and higher influent total suspended solids (TSS) and TP values resulted in lower effluent BOD_5 , implying the importance of controlling dissolved oxygen (DO) in aeration basins. Since aeration is one of the major energy consumption sources in WWTPs, real-time prediction of the effluent water quality using the self-learning AI system developed in this study can be adopted to lower the energy cost significantly while improving effluent water quality. WWTPs must develop control methods based on the RNN prediction and Explainable AI analysis due to different operational conditions.

TABLE OF CONTENTS

1. INTRODUCTION.....	1
1.1 Background and Research Motivation	1
1.2 Problem Statement	4
1.2.1 Lack of Research on Big Data Management from WWTPs	4
1.2.2 Lack of Comprehensive Statistical Analysis of Wastewater Treatment Data.....	5
1.2.3 Lack of Predictive Models to Forecast Effluent Quality	5
1.2.4 No Logistics for a Real-Time Model for WWTPs.....	6
1.3 Research Objectives	6
1.4 Research Scope and Methodology	6
1.5 Organization of the Dissertation Proposal	8
2. LITERATURE REVIEW.....	10
2.1 Previous Studies of Big Data from WWTPs.....	10
2.1.1 Big Data Basics.....	10
2.1.2 Type of Big Data Analytics	11
2.1.3 Data from WWTPs	13
2.1.4 Big Data for Better WWTP Management.....	15
2.2 Previous Studies on Statistical Analysis in Wastewater Treatment Systems	16
2.3 Previous Studies of the Development of Predictive Models in WWTPs.....	19

2.3.1 Autoregressive Integrated Moving Average (ARIMA).....	19
2.3.2 Artificial Neural Network (ANN).....	23
2.4 Conclusions and Recommendations	27
3. BIG DATA ANALYTICS FROM A WASTEWATER TREATMENT	28
3.1 Abstract.....	28
3.2 Introduction.....	29
3.2.1 Background.....	29
3.2.2 Previous Research.....	30
3.2.3 Shortcoming of Previous Research.....	31
3.2.4 Study Objectives	32
3.3 Materials and Methods.....	32
3.3.1 Data Preprocessing.....	37
3.3.2 Statistical Analysis.....	39
3.4 Results and Discussions.....	41
3.4.1 Understanding of Data	41
3.4.2 Data Preparation.....	41
3.4.3 Data Preprocessing.....	52
3.5 Conclusions and Recommendations	67
4. RECURRENT NEURAL NETWORKS (RNN) FOR MODEL PREDICTION	69
4.1 Abstract.....	69
4.2 Introduction.....	70

4.2.1 Background	70
4.2.2 Ideal Predictive Model for a WWTP	71
4.2.3 Shortcomings of Previous Predictive Models	75
4.2.4 Study Objective.....	76
4.3 Materials and Methods.....	77
4.3.1 Data Preparation.....	77
4.3.2 Development of Recurrent Neural Networks (RNNs) Models.....	81
4.4 Results and Discussion	84
4.5 Conclusions and Recommendations	116
4.6 Future Research Plans	117
5. LOGISTICS FOR A REAL-TIME PREDICTION MODEL	118
5.1 Abstract	118
5.2 Introduction	119
5.2.1 Background	119
5.2.2 SCADA Modernization with Python	120
5.2.3 Explainable Artificial Intelligence	123
5.2.4 Literature Review of Aeration Optimization in Wastewater Treatment Process.....	124
5.2.5 Study Objectives	125
5.3 Materials and Methods.....	125
5.3.1 Logistics for a Real-Time Model in Wastewater Treatment Plant	125
5.4 Results and Discussion	133

5.5 Conclusions and Recommendations	145
5.6 Future Research	146
6. PREDICTION OF SLUDGE VOLUME INDEX (SVI) IN A WASTEWATER TREATMENT PLANT USING ARTIFICIAL INTELLIGENCE.....	148
6.1 Abstract	148
6.2 Introduction	148
6.2.1 Background	149
6.2.2 Sludge Volume Index	150
6.2.3 Activated Sludge Process	151
6.2.4 Filamentous Bulking	151
6.3 Materials and Methods	152
6.4 Results and Discussion	160
6.5 Conclusions and Recommendations	173
7. REFERENCES.....	175
8. APPENDICES.....	185

LIST OF FIGURES

<i>Figure 1.1 The growth of data from 2010 to 2020. (Source: Roser et al., 2015)</i>	<i>1</i>
<i>Figure 1.2 Number of deaths by risk factor. (Source: Ritchie & Roser, 2018)</i>	<i>3</i>
<i>Figure 1.3 Number of people with and without access to safe drinking water.</i>	<i>3</i>
<i>Figure 1.4 Organization of the preliminary dissertation proposal.</i>	<i>9</i>
<i>Figure 2.1 The three V principles of big data. (Source: Su, 2018).....</i>	<i>11</i>
<i>Figure 2.2 The iterative steps of the ARIMA approach. (Source: Box and Jenkins, 1970)</i>	<i>20</i>
<i>Figure 2.3 Frequency and trend of AI techniques applied to wastewater treatment during 1995–2019. (Source: Zhao et al., 2020)</i>	<i>21</i>
<i>Figure 2.4 Classification tree of AI technology used in wastewater treatment.....</i>	<i>22</i>
<i>Figure 2.5 Basic structure of an Artificial Neural Network (ANN). (Source: Haider et al., 2019)</i>	<i>24</i>
<i>Figure 2.6 Basic structure of a Recurrent Neural Network (RNN). (Source: Haider et al., 2019)</i>	<i>25</i>
<i>Figure 2.7 Basic structure of a Long Short-Term Memory cell (LSTM).</i>	<i>26</i>
<i>Figure 3.1 The methodology of big data analytics in WWTPs.</i>	<i>33</i>
<i>Figure 3.2 Nine Springs Wastewater Treatment Plant. (Source: McGowan & Wang, 2008).....</i>	<i>34</i>
<i>Figure 3.3 The numbers of historical Nine Springs WWTP data from 1996-2019.</i>	<i>34</i>
<i>Figure 3.4 Column name is 'MeasureCode', and 'LocationCode' contains various parameters.</i>	<i>35</i>

<i>Figure 3.5 Forms of data preprocessing. (Source: Han et al., 2012).</i>	38
<i>Figure 3.6 General layout of Nine Springs WWTP. (Source: McGowan & Wang, 2008)</i>	42
<i>Figure 3.7 The schematic of the liquid treatment process at the Nine Springs WWTP.</i>	43
<i>Figure 3.8 The relationship between influent and effluent in TSS.</i>	50
<i>Figure 3.9 The relationship between influent and effluent in TP.</i>	51
<i>Figure 3.10 The relationship between influent and effluent in TKN.</i>	51
<i>Figure 3.11 The relationship between influent and effluent in NH_3N.</i>	51
<i>Figure 3.12 The relationship between influent and effluent in BOD_5.</i>	52
<i>Figure 3.13 Correlation of effluent BOD_5 to other parameters.</i>	55
<i>Figure 3.14 Kurtosis and Skewness of normal distribution.</i>	58
<i>Figure 3.15 Box plot of yearly and quarterly effluent BOD_5.</i>	58
<i>Figure 3.16 Normal probability distribution.</i>	59
<i>Figure 3.17 The means effluent BOD_5 over a day, week, month, quarter, and year.</i>	60
<i>Figure 3.18 The average means effluent BOD_5 by year, quarter, month, and day.</i>	61
<i>Figure 3.19 The patterns effluent BOD_5 for each year.</i>	62
<i>Figure 3.20 Time series plot of effluent BOD_5 from 2015-2018.</i>	63
<i>Figure 3.21 Box plot of effluent BOD_5 by year and quarter from 2015-2018.</i>	63
<i>Figure 3.22 The mean effluent BOD_5 over a day, week, month, quarter, and year.</i>	64
<i>Figure 3.23 The average means of effluent BOD_5 over years, quarters, months, and days.</i>	65
<i>Figure 3.24 The Dickey–Fuller test with hypothesis testing.</i>	66
<i>Figure 4.1 A Recurrent Neural Networks. (Source: Olah, 2015)</i>	72

<i>Figure 4.2 Long Short-Term Memory networks (LSTM). (Source: Olah, 2015).....</i>	<i>72</i>
<i>Figure 4.3 Forget gate architecture. (Source: Olah, 2015)</i>	<i>73</i>
<i>Figure 4.4 The cell state for the input gate. (Source: Olah, 2015).....</i>	<i>74</i>
<i>Figure 4.5 Update each cell state. (Source: Olah, 2015).....</i>	<i>74</i>
<i>Figure 4.6 Output gate architecture. (Source: Olah, 2015)</i>	<i>75</i>
<i>Figure 4.7 The overview of an RNN modeling process.</i>	<i>78</i>
<i>Figure 4.8 The architecture of the RNN model.....</i>	<i>79</i>
<i>Figure 4.9 Five steps to develop a simple RNN model.</i>	<i>82</i>
<i>Figure 4.10 Five steps to develop an LSTM model.</i>	<i>82</i>
<i>Figure 4.11 Steps to develop a time series model.....</i>	<i>84</i>
<i>Figure 4.12 Train and test loss over epoch for effluent BOD₅ prediction models.</i>	<i>87</i>
<i>Figure 4.13 The original data of effluent BOD₅ from 2015 to 2018.....</i>	<i>91</i>
<i>Figure 4.14 The prediction of effluent BOD₅ from 2015 to 2018 using the simple RNN model. .</i>	<i>92</i>
<i>Figure 4.15 The prediction of effluent BOD₅ from 2015 to 2018 using the LSTM model.....</i>	<i>93</i>
<i>Figure 4.16 The prediction of effluent TP from 2015 to 2018 using the simple RNN model.</i>	<i>94</i>
<i>Figure 4.17 The prediction of effluent TP from 2015 to 2018 using the LSTM model.....</i>	<i>95</i>
<i>Figure 4.18 The prediction of effluent TKN from 2015 to 2018 using the simple RNN model. ...</i>	<i>96</i>
<i>Figure 4.19 The prediction of effluent TKN from 2015 to 2018 using the LSTM model.....</i>	<i>97</i>
<i>Figure 4.20 The prediction of effluent TSS from 2015 to 2018 using the simple RNN model.....</i>	<i>98</i>
<i>Figure 4.21 The prediction of effluent TSS from 2015 to 2018 using the LSTM model.</i>	<i>99</i>
<i>Figure 4.22 The prediction of effluent NH₃N from 2015 to 2018 using the simple RNN model.</i>	<i>100</i>

<i>Figure 4.23 The prediction of effluent NH₃N from 2015 to 2018 using the LSTM model.....</i>	<i>101</i>
<i>Figure 4.24 The prediction of daily effluent BOD₅ from 2015 to 2018 using the Simple RNN model.</i>	<i>102</i>
<i>Figure 4.25 The prediction of daily effluent BOD₅ from 2015 to 2018 using the LSTM model.</i>	<i>103</i>
<i>Figure 4.26 The prediction of daily effluent TP from 2015 to 2018 using the Simple RNN model.</i>	<i>104</i>
<i>Figure 4.27 The prediction of daily effluent TP from 2015 to 2018 using the LSTM model.....</i>	<i>105</i>
<i>Figure 4.28 The prediction of daily effluent TKN from 2015 to 2018 using the Simple RNN model.</i>	<i>106</i>
<i>Figure 4.29 The prediction of daily effluent TKN from 2015 to 2018 using the LSTM model... </i>	<i>107</i>
<i>Figure 4.30 The prediction of daily effluent TSS from 2015 to 2018 using the Simple RNN model.</i>	<i>108</i>
<i>Figure 4.31 The prediction of daily effluent TSS from 2015 to 2018 using the LSTM model. ...</i>	<i>109</i>
<i>Figure 4.32 The prediction of daily effluent NH₃N from 2015 to 2018 using the Simple RNN model.</i>	<i>110</i>
<i>Figure 4.33 The prediction of daily effluent NH₃N from 2015 to 2018 using the LSTM model.</i>	<i>111</i>
<i>Figure 4.34 The prediction of daily effluent SVI from 2015 to 2018 using the Simple RNN model.</i>	<i>112</i>
<i>Figure 4.35 The prediction of daily effluent SVI from 2015 to 2018 using the LSTM model.....</i>	<i>113</i>
<i>Figure 5.1 The diagram that shows a single ordering. (Source: Lundberg & Lee, 2017)</i>	<i>124</i>
<i>Figure 5.2 Three steps of an aeration optimization method.</i>	<i>124</i>
<i>Figure 5.3 The general architecture of the SCADA system. (Source: Sosik, 2014)</i>	<i>127</i>

<i>Figure 5.4 The main components of the SCADA system. (Modified from Manda et al., 2018) .</i>	<i>128</i>
<i>Figure 5.5 General WWTP layout. (Source: Richards, 2020).....</i>	<i>129</i>
<i>Figure 5.6 The flow chart of the real-time logistics system.....</i>	<i>130</i>
<i>Figure 5.7 The data collection diagram in a wastewater treatment facility.</i>	<i>131</i>
<i>Figure 5.8 The automation control system of a WWTP. (Source: Du et al., 2019)</i>	<i>132</i>
<i>Figure 5.9 The structure of the modern SCADA. (Modified from Manda et al., 2018).....</i>	<i>132</i>
<i>Figure 5.10 The modern SCADA with Python.....</i>	<i>133</i>
<i>Figure 5.11 The SHapley summary plot of the model that includes flow rate.....</i>	<i>135</i>
<i>Figure 5.12 The SHapley summary plot of the model that includes organic loading.</i>	<i>135</i>
<i>Figure 5.13 The SHapley summary plot of the model that includes flow rate and organic loading.</i>	<i>136</i>
<i>Figure 5.14 The force plot of the first observation including flow rate.</i>	<i>136</i>
<i>Figure 5.15 The force plot of the first observation including organic loading.</i>	<i>136</i>
<i>Figure 5.16 The force plot of the first observation including flow rate and organic loading....</i>	<i>137</i>
<i>Figure 5.17 The collective force plot of all input variables.</i>	<i>137</i>
<i>Figure 5.18 The inputs values in the collective force plot with different inputs.....</i>	<i>139</i>
<i>Figure 5.19 The mean of inputs values in the collective force plot by selecting one input value.</i>	<i>140</i>
<i>Figure 5.20 SHAP dependence plot of influent NH_3N to influent BOD_5.....</i>	<i>141</i>
<i>Figure 5.21 SHAP dependence plot of influent TKN to influent BOD_5.....</i>	<i>142</i>
<i>Figure 5.22 The diagram of the real-time logistics for wastewater treatment operation.</i>	<i>143</i>

<i>Figure 5.23 The real-time logistics for the prediction models in WWTPs.</i>	144
<i>Figure 6.1 Sludge Volume Index data from 1996 to 2020.</i>	154
<i>Figure 6.2 Box plot of yearly SVI from 1996 to 2020.</i>	154
<i>Figure 6.3 Sludge Volume Index data from 2001 to 2020.</i>	155
<i>Figure 6.4 Box plot of yearly SVI from 2001 to 2020.</i>	155
<i>Figure 6.5 Sludge Volume Index data from 2010 to 2020.</i>	156
<i>Figure 6.6 Box plot of yearly SVI from 2010 to 2020.</i>	157
<i>Figure 6.7 The Recurrent Neural Networks model prediction in Python.</i>	159
<i>Figure 6.8 Correlation coefficient between each parameter.</i>	160
<i>Figure 6.9 Normal distribution of the first dataset from 1996 to 2020.</i>	161
<i>Figure 6.10 Normal distribution of the second dataset from 2001 to 2020.</i>	162
<i>Figure 6.11 Normal distribution of the third dataset from 2010 to 2020.</i>	162
<i>Figure 6.12 Normal probability plot of the first dataset.</i>	163
<i>Figure 6.13 Normal probability plot of the second dataset.</i>	164
<i>Figure 6.14 Normal probability plot of the third dataset.</i>	164
<i>Figure 6.15 Mean Absolute Error between each epoch of the model.</i>	165
<i>Figure 6.16 The Sludge Volume Index prediction model of the first set of data (1996 to 2020).</i>	167
<i>Figure 6.17 The Sludge Volume Index prediction model of the second set of data (2001 to 2020).</i>	168
<i>Figure 6.18 The Sludge Volume Index prediction model of the third set of data (2010 to 2020).</i>	169

<i>Figure 6.19 SHapley interpretation plots for the first model.</i>	<i>170</i>
<i>Figure 6.20 SHapley interpretation plots for the second model.</i>	<i>171</i>
<i>Figure 6.21 SHapley interpretation plots for the third model.</i>	<i>172</i>
<i>Figure 8.1 A simple RNN model in Python.</i>	<i>188</i>
<i>Figure 8.2 A simple RNN model in Python (Continued).</i>	<i>189</i>
<i>Figure 8.3 A simple RNN model in Python (Continued).</i>	<i>190</i>
<i>Figure 8.4 An LSTM model in Python.</i>	<i>191</i>
<i>Figure 8.5 An LSTM model in Python (Continued).</i>	<i>192</i>
<i>Figure 8.6 An LSTM model in Python (Continued).</i>	<i>193</i>

LIST OF TABLES

<i>Table 2.1 The definitions of big data related to four themes. (Source: Riahi & Riahi, 2018)</i>	<i>12</i>
<i>Table 2.2 Three stages for gaining confidence in sensors and analyzers. (Source: Shaw, 2017)14</i>	
<i>Table 2.3 The classification of concentration in domestic wastewater.</i>	<i>16</i>
<i>Table 3.1 The selection of parameters from 'SiteCode' and 'LocationCode'.</i>	<i>47</i>
<i>Table 3.2 The influent table from 'MTR VLT' location.</i>	<i>48</i>
<i>Table 3.3 The effluent table from 'EFF BLDG' location.</i>	<i>49</i>
<i>Table 3.4 Number of parameters in 'MeasureCode'.</i>	<i>49</i>
<i>Table 3.5 The clean dataset from the Nine Springs WWTP big data.</i>	<i>50</i>
<i>Table 3.6 The table of descriptive statistics.</i>	<i>53</i>
<i>Table 3.7 Correlation coefficient between parameters.</i>	<i>54</i>
<i>Table 3.8 Heatmap of the correlation coefficient.</i>	<i>55</i>
<i>Table 3.9 Program for the result after removing missing values.</i>	<i>56</i>
<i>Table 3.10 Program for processing normal distribution test.</i>	<i>57</i>
<i>Table 4.1 Applications of AI models for operation management in WWTPs.</i>	<i>77</i>
<i>Table 4.2 The result from data scaling.</i>	<i>80</i>
<i>Table 4.3 The values of water quality parameters in wastewater treatment from 2015 to 2018.</i>	<i>85</i>
<i>Table 4.4 Performances of the effluent BOD₅ prediction models.</i>	<i>86</i>
<i>Table 4.5 The summary of the accuracy of effluent parameters models.</i>	<i>89</i>
<i>Table 4.6 General descriptive statistics of the daily data from 2015 to 2018.</i>	<i>90</i>

<i>Table 4.7 Comparison of the model accuracy between the discrete big data and the daily dataset.</i>	114
<i>Table 4.8 The average R2 score between the discrete big data and the daily dataset.</i>	115
<i>Table 5.1 The dataset of the model including output and input parameters.</i>	134
<i>Table 6.1 Nine Springs Wastewater Treatment Dataset</i>	152
<i>Table 6.2 Dataset from 1996 to 2020.</i>	153
<i>Table 6.3 Normalization inputs and output values of the models.</i>	158
<i>Table 6.4 General statistics for the dataset from 2010 to 2020.</i>	160
<i>Table 8.1 Application of AI technologies to pollutant removal in WWTPs.</i>	185
<i>Table 8.2 Application of AI technologies to pollutant removal in WWTPs (Continued).</i>	186
<i>Table 8.3 Application of AI models for operation management during wastewater treatment.</i>	187

1. INTRODUCTION

1.1 Background and Research Motivation

Due to the demands for lower operation and maintenance cost, energy cost, stringent compliance requirements of water quality parameters, and lower greenhouse gas emission, WWTPs (WWTPs) must be in a smart management mode. Due to the extensive use of water quality sensors, big data is generated every day or even every second. In the wastewater treatment sector and many other fields such as finance, marketing, stocks, health care, and so on, big data plays an essential role. Data generation has been tremendously increasing since 2010, and 90% of the world's data has been created in the past two years (Figure 1.1).

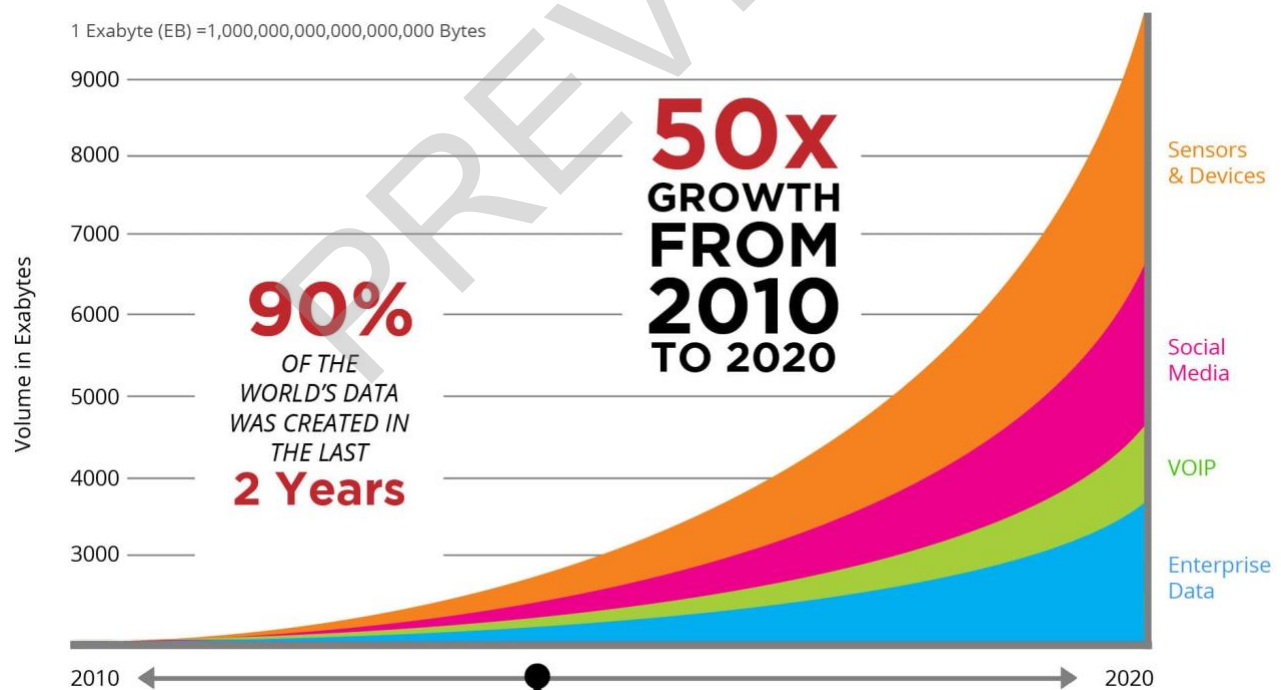


Figure 1.1 The growth of data from 2010 to 2020. (Source: Roser et al., 2015)

Big data is currently most preferably exploited in organizations, companies, and businesses. However, a massive amount of data results in the difficulties of storing, monitoring, analyzing, and visualizing data for further processing. Therefore, data is often stored and then underutilized.

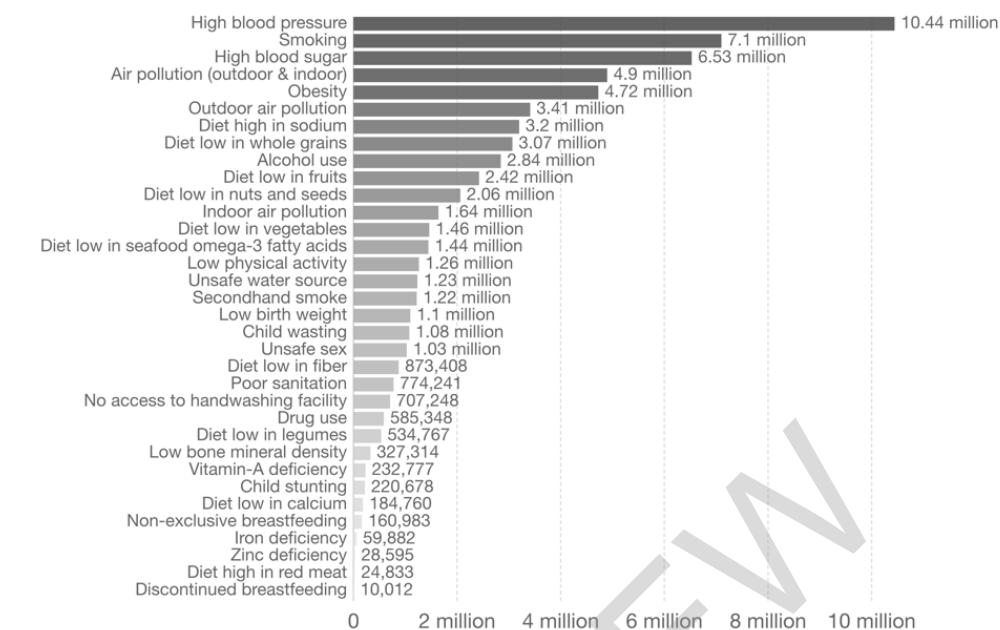
A good understanding of the data dynamics in WWTPs is vital for reliable monitoring and control activities. However, the dynamical behavior of the data is usually complicated and uncertain due to nonlinearity, variations from the environmental conditions, strong interactions between the process variables involved, and changes in the flow rate and concentration of the composition of the influent (Harrou et al., 2018). Finding insight from historical and real-time data can directly improve traditional operational systems in a better direction.

A series of wastewater treatment processes remove pollutants from wastewater to be safely reused or discharged into natural water resources. Treated water can be recycled and redistributed for agricultural, industrial, and other purposes or safely released back into the natural resources without causing any adverse effects (Grant et al., 2012). The effluent from a WWTP must meet the National Pollutant Discharge Elimination System (NPDES) permit to protect the environment and public health (Siegrist, 2017). Lack of access to safe water leads to a risk factor for infectious diseases such as cholera, diarrhea, and dysentery. According to the Global Burden of Disease study, 1.2 people died prematurely in 2017 due to unsafe water (Figure 1.2). This number was three times the number of homicides in 2017 and approximately equal to the amount that died in road accidents globally. Besides, only 71% of the world population has access to safe drinking water, which means that 29% of the world population does not have access to safe water. It equates to 2.1 billion people globally (Figure 1.3).

Number of deaths by risk factor, World, 2017

Total annual number of deaths by risk factor, measured across all age groups and both sexes.

Our World
in Data



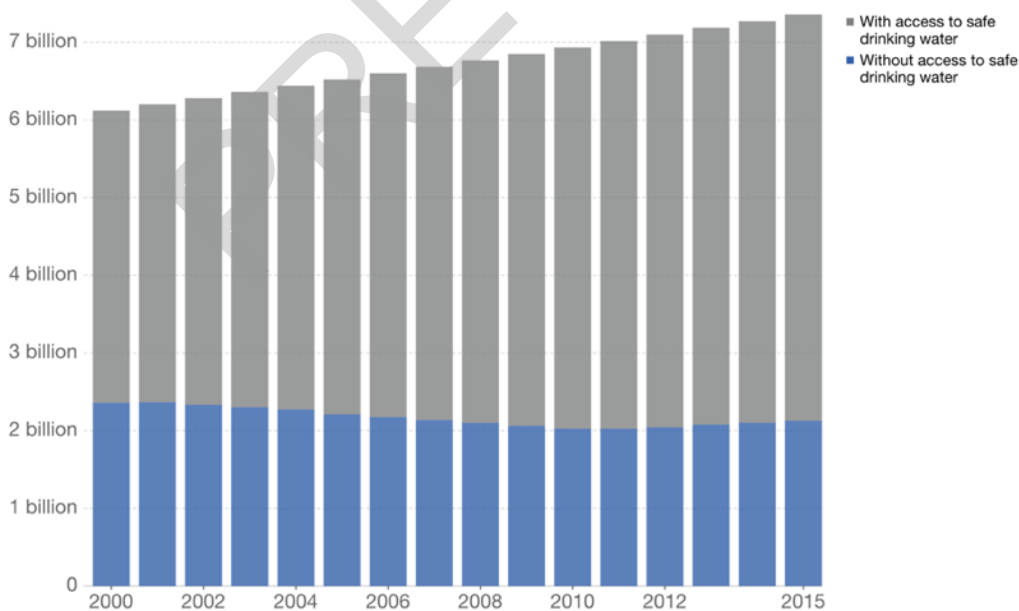
Source: IHME, Global Burden of Disease (GBD)

CC BY

Figure 1.2 Number of deaths by risk factor. (Source: Ritchie & Roser, 2018)

Number of people with and without access to safe drinking water, World

Our World
in Data



Source: Our World in Data based on WHO, WASH

OurWorldInData.org/water-access • CC BY

Figure 1.3 Number of people with and without access to safe drinking water.

(Source: Ritchie & Roser, 2018)

Even though there is an increasing amount of historical data in WWTPs, most information in the data will remain unexploited. According to the operator's point of view, the main reason for this is the high dimensionality of the data, where traditional analysis tools cannot be used (Durrenmatt, 2011). As a result, a large amount of data is lost. Various methods and tools such as big data analytics, statistical analysis, deep learning, and artificial intelligence (AI) applications extract information hidden in the data. Furthermore, this would assist the operator in further optimization of the WWTP, improve the effluent quality, reduce the human errors in operating processes, foster the operator's knowledge of the plant processes, and provide other supporting information. Finally, the main goal of a WWTP is to ensure the efficiency of wastewater treatment operation, which is tremendously essential for community health and the environment. Advanced tools and techniques to improve wastewater treatment operations were introduced in this study for analyzing, modeling, optimizing, and forecasting wastewater treatment quality.

1.2 Problem Statement

A literature review on big data management in wastewater treatment operation indicates four principle wastewater management problems that need to be solved in this study: (1) lack of research on big data management from WWTP, (2) lack of comprehensive statistical analysis of the data, and (3) lack of real-time predictive models to forecast effluent quality.

1.2.1 Lack of Research on Big Data Management from WWTPs

After researching publications in big data management in wastewater treatment facilities from various resources such as Google Scholar®, Scopus®, and Web of Science®, there have been limited studies (Gheraout et al., 2018; Romero et al., 2017). A large amount of data is generated from various wastewater treatment operations every day. Big data should be exploited to enhance the operating systems. Unfortunately, due to the lack of specialized tools, operators and engineers

cannot extract meaningful and valuable information from the massive amount of high-dimensional data.

1.2.2 Lack of Comprehensive Statistical Analysis of Wastewater Treatment Data

Several studies have analyzed energy flow and influent in wastewater treatment facilities to monitor, assess, and model the WWTPs (Martin & Vanrolleghem, 2014). In addition, many publications have illustrated the usefulness of statistical analysis models for WWTP optimization (Cheng et al., 2019; Newhart et al., 2019); operation (Garbowski et al., 2018); analysis (Pantsar-Kallio et al., 1999; Taheriyoun & Moradinejad, 2015; Zhang et al., 2019) and control (Harrou et al., 2018; Maiza et al., 2013). However, few studies have been performed on finding patterns, determining the relationship of each parameter, and selecting meaningful information from big data in wastewater treatment operations with the use of advanced analytics. As a result, big data is underutilized.

1.2.3 Lack of Predictive Models to Forecast Effluent Quality

The water quality predictions in WWTPs were attempted using advanced machine learning tools and techniques. Nonetheless, without the pretreatment of big data, the forecast may not be accurate. Also, previous studies focused on traditional deterministic modeling methodology (Boyd et al., 2019; Huang et al., 2016; Khademikia et al., 2016; Pisa et al., 2019). The novel deep learning method, an RNN, is rarely applied. The conventional approaches might be accurate in predicting if the system is fixed and all the parameters are determined. Furthermore, it is time-consuming, intrusive, and limited by small data analysis.

1.2.4 No Logistics for a Real-Time Model for WWTPs

Real-time data reflects the status of an operational system. The common characteristic of the data is the strict time constraint (Wu et al., 2006). Real-time information is associated with a timestamp and life cycle, and they are only valid for the responding sampling time. Deep learning algorithms with real-time modeling would be a reliable tool for early warning of potential operational upsets and subsequent effluent quality permit violations. There is no study on the development of logistics for real-time modeling using RNNs to help operators for monitoring, detecting fault operation systems in WWTPs.

1.3 Research Objectives

The main objective of this paper is to enhance the operation and performance of WWTPs through big data management and model prediction with AI applications. The study has the following four specific objectives:

- (1) To develop analytics for big data from a WWTP;
- (2) To interpret analytics results for extracting meaningful information;
- (3) To develop deep learning models for forecasting effluent quality from historical wastewater treatment data, which include train and evaluate the models to acquire the most effective algorithms; and
- (4) To establish logistics for a real-time self-learning AI system for monitoring and detecting problems during WWTP operation.

1.4 Research Scope and Methodology

Four principal tasks in this study are discussed below.

Task 1: Big data analytics frameworks in wastewater treatment operation.

This task involves collecting data, understanding the processes in wastewater treatment, visualizing data, selecting meaningful information, and developing big data analytics procedures.

Task 2: Statistical analysis techniques to obtain a pattern and meaningful information.

This task applies various statistical methods such as descriptive statistical analysis, correlation coefficient, box plot, normal distribution, and hypothesis testing to find a relationship between parameters, a pattern of data, and insight of information.

Task 3: DNNs model development for prediction.

This task is to develop a predictive model by implementing advanced modeling techniques, Recurrent Neural Networks (RNNs). The methods include preparing the data, selecting the train and test dataset, build a simple RNNs model and the Long Short-Term Memory (LSTM) model with a different number of hidden layers, train the models to predict an output result, and compare and evaluate the models. The parameters to be predicted will be Total Phosphorus (TP), Total Suspended Solids (TSS), and NH_3 /Total Nitrogen (TN) in addition to BOD_5 (Biochemical Oxygen Demand) and lastly, Sludge Volume Index (SVI). Control of dissolved oxygen (DO), sludge and energy management logistics may be attempted from the big data.

Task 4: Propose logistics for a real-time prediction model to detect fault errors in wastewater treatment operations.

Unless a WWTP allows access to real-time big data and incorporates the AI model into their Supervisory Control and Data Acquisition (SCADA) system, it is impossible to implement a real-time prediction model to the existing system. Therefore, this task proposes logistics for

real-time model development in WWTPs to help a WWTP's operator for failure detections beforehand.

1.5 Organization of the Dissertation Proposal

The thesis proposal is organized as follows:

Chapter 2 presents a comprehensive review of previous studies on big data in the wastewater treatment sector, statistical analysis in wastewater treatment systems, and the development of prediction models in WWTPs.

Chapter 3 presents big data management with statistical analysis. Big data analytics processes include data collection, data understanding, data preparation, data mining, evaluation, and deployment. Data was collected from the Nine Springs WWTP in Madison Metropolitan Sewerage District (MMSD), Wisconsin. This data is comprehensively studied through data preprocessing techniques contain data cleaning, data integration, data transformation, and data reduction. Finally, statistical analysis techniques extracted meaningful patterns and information to obtain the appropriate dataset.

Chapter 4 presents the development of model predictions using RNNs. After the big data is extracted to a manageable size and the statistical preprocessing technique is implemented to select meaningful information, the prediction efficiency of wastewater effluent quality will improve significantly. The results of traditional RNN models and RNN-LSTM models were compared and evaluated to choose an optimization algorithm for implementation.

Chapter 5 proposes the subsequent work plan to develop an RNN model using different parameters, develop logistics for monitoring, detecting, and proactive maintenance, assist better decision making, and optimize wastewater treatment facilities.