INTELLIGENT RESOURCE USE IN PROCESS MANUFACTURING

Oliver Fisher, MEng

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

October, 2019

This work is split into two volumes:

Volume 1: Thesis chapters and references

Volume 2: Published works and appendices

VOLUME 1

INTELLIGENT RESOURCE USE IN PROCESS MANUFACTURING

Oliver Fisher, MEng

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

October, 2019

ABSTRACT

Economic, environmental policy and public pressure are driving businesses and Governments to reduce damage to the environment, and begin to add value to their manufacturing environment. Circular economy thinking has emerged as a model to redesign manufacturing systems to maximise resource lifetimes within the economy. Traditionally, process manufacturing is the thermal, chemical and/or biochemical conversion of resources into products, by-products and waste streams. Example industries include fast-moving consumer goods (FMCG), chemicals, plastics and oil and gas. Innovative thinking, coupled with new technologies enables process manufacturers to intelligently utilise resources not just within their internal systems, but also externally between manufacturing systems. This work aims to evaluate how process manufacturing can intelligently utilise their resources, specifically the potential of data sharing, analysis and modelling of process manufacturing systems to enhance process resilience and waste valorisation.

Cloud Manufacturing (CM) is a service-oriented business model to share manufacturing capabilities and resources on a cloud platform. Manufacturing is under pressure to achieve cost and environmental impact reductions, as manufacturing becomes more integrated and complex. Cloud manufacturing offers a solution, as it is capable of making intelligent decisions to provide the most sustainable and robust manufacturing route available. Although CM research has progressed, a consensus is still lacking on the concepts within CM as well as applications and scope beyond discrete manufacturing. This work details the definitions, characteristics, architectures and previous case studies on CM. From this, the fundamental aspects of the CM concept are identified, along with an analysis of how the concept has progressed. A new, comprehensive CM definition is formulated by combining key concepts drawn from previous definitions and for the first time emphasises CM potential to enhance process resilience and waste valorisation. Four key methods of how CM increases sustainability are identified: (1) collaborative design; (2) greater automation; (3) improved process resilience and (4) enhanced waste reduction, reuse and recovery.

Cloud manufacturing relies on the development of mathematical models to describe process manufacturing systems. The reason for this is to be able to virtualise the system onto the manufacturing cloud, which can then simulate the system to identify the most sustainable manufacturing route. The increasing availability of data, due to the adoption of low-cost industrial internet of things technologies, coupled with increasing processing power from cloud computing, is fuelling increased use of data-driven

models (DDMs) in manufacturing. Utilising case studies from the waste management industry and the food and drink industry, the considerations and challenges faced when developing DDMs for manufacturing systems are explored. Ensuring a high-quality set of model development data that accurately represents the manufacturing system is key to the successful development of a DDM. This work develops a protocol of how to develop a DDM, from defining a model's goal to evaluating a model's outputs in meaningful ways, which will enable manufacturers to build effective DDMs.

The development of a successful DDM starts with understanding the data and recognising whether the data is representative of the process manufacturing system. This is essential as process manufacturing systems contain a large level of variation. Understanding the variability defined within the data proves challenging because of the multivariate nature of process manufacturing systems. This work develops methods by which statistical and data visualisation techniques can be used to define the system described in the data. A novel visualisation tool based on parallel coordinate plot (PCP) was developed to determine if new data-points are inside or outside the manufacturing system.

Unseen data is essential to determine the model extrapolation capabilities beyond the data it was developed on. However, evaluating a DDM on unseen data is often overlooked in the literature. This work developed a novel data partitioning technique to ensure the unseen data is partitioned from across and beyond the manufacturing system, to thoroughly evaluate a DDM predictions capabilities. The methodologies developed by this work were tested by building and comparing four DDMs built for a novel bioprocess system. The bioprocess treats manufacturers' wastewaters to reduce pollutant load for water reuse, while simultaneously generating biogas. The four models' predictive capabilities were statistically evaluated on unseen data in terms of the coefficient of determination (R^2), the root of the mean squared error (RMSE), median absolute error (MAE) and mean absolute percentage error (MAPE). The random forest (RF) model was found to have the best predictive capabilities (R^2 of 0.99, RMSE of 1.4399, MAE of 2.811 and MAPE of 6.039).

Achieving sustainable process manufacturing requires being able to find the manufacturing route that maximises both the economic, environmental and social objectives. This requires the development of models that can predict multiple targets so that the relationships between a system's inputs-outputs and outputs-outputs can be established and exploited. Previous literature modelling process manufacturing systems have largely focused on building multiple independent single target models, each

predicting one output. These models fail at capturing the relationships between the system's outputs. This work develops multiple target (MT) models that are superior when describing a manufacturing system with multiple outputs.

For this study, the current leading MT models based on regression chains (RC) are studied. This work presents two novel RC models built using artificial neural networks (ANN) as a base model. The models were then compared against RC models using support vector regression (SVR) as base models and a standard multiple target ANN. Appraisal of the models through R^2, normalised root mean square error (nRMSE) and MAPE showed that the RC models built using ANN as a base model had superior predictive capabilities. The average statistical scores for the ANN based model were R^2 of 0.64, RMSE of 0.10, MAE of 2.811 and MAPE of 7.49.

The modelling methodologies developed in this work were then applied to model a new process manufacturing system, which was iron (Fe) remediation from mining wastewater. The Coal Authority is investigating new treatment methods, as the traditional passive treatment methods have a large footprint and are capital cost-intensive. The Coal Authority conducted a 40-week trial evaluating the potential of surface catalysed oxidation of ferrous iron (SCOOFI) technology to replace existing settling lagoons technologies in the treatment of mining wastewater. Three different reactor configurations were compared and the up-flow reactor was found to have the best Fe removal rates (average Fe removal during the trial of 27.58% compared to -7.85% and 7.27%). There was a high level of variability in the Fe removal rates by the reactors, so a DDM approach was utilised to investigate the variability within the system. Two different modelling approaches were trialled, one developing a DDM model using only the dataset for the up-flow reactor and the other using all three reactors datasets. The Gaussian process (GP) model trained on only the one dataset was found to be the best (R^2 of 0.86, RMSE of 3.75, MAE of 2.78).

PUBLICATIONS AND PRESENTATIONS

Publications

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2018. Cloud manufacturing as a sustainable process manufacturing route. *Journal of Manufacturing Systems*. 47; 35-68. [Impact Factor: 2.770; SNIP 2.191]

Fisher O, Watson N, Escrig J, Gomes RL, 2020. Intelligent resource use to deliver waste valorisation and process resilience in manufacturing environments. *Johnson Matthey Technology Review*. 64; 93-99. [Impact Factor: 1.27; SNIP: 0.669]

Publications accepted by journals

Fisher O, Watson N, Escrig J, Witt R, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. *Computers & Chemical Engineering* (**Accepted**, 20 Apr 2019). [Impact Factor: 3.113; SNIP: 1.613]

Works ready for publication

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Data-driven modelling of process manufacturing systems. *Chemical Engineering*. [Impact Factor: 6.735; SNIP: 1.863]

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Multiple target modelling to enable sustainable process manufacturing. *Journal of Cleaner Production* [Impact Factor: 6.395; SNIP: 2.308]

Presentations

Fisher O., Watson N.J., Porcu L., Bacon D., Rigley M., Gomes R.L. Intelligent data use for resource recovery from Small Medium Enterprises (SME) wastewaters. *Designing a Connected Future, Connected Everything*. June 25th-26th, 2019, Nottingham, UK. Poster presentation.

Fisher O., Watson N.J., Gomes R.L. Intelligent resource use in process manufacturing; *Faculty of Engineering Post Graduate Research Showcase.* May 2nd, 2019 University of Nottingham, UK. Oral Presentation;

Fisher O., Watson N.J., Porcu L., Bacon D., Rigley M., Gomes R.L.. Should SMEs have their heads in the clouds? *Nottingham Environmental Conference*. June 25th-26th, 2018, Nottingham, UK. Poster presentation.

Fisher O., Watson N.J., Porcu L., Bacon D., Rigley M., Gomes R.L.. Should SMEs have their heads in the clouds? *Faculty of Engineering Post Graduate Research Showcase.* May 2nd, 2018, University of Nottingham, UK. Oral Presentation;

Fisher, O., Watson, N.J., Gomes, R.L., (2017) Cloud manufacturing as a sustainable process manufacturing route. *10th World Congress of Chemical Engineers 2017*, 1st-5th, October, 2017, Barcelona, Spain. Oral presentation.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to Dr Rachel Gomes and Dr Nicholas Watson, my supervisors, for their guidance, support and patience during the course of my study. Without their supervisions, the completion of this work would not have been possible.

This work would not have been possible without cooperation from the team at Lindhurst Engineering Ltd. whose support and insightful advice was greatly appreciated. In particular I would like to thank Mr Martin Rigley, Mr Darren Bacon and Dr Laura Porcu who have been a pleasure to work with.

I also would like to thank everyone who supported and helped me during the past three years. Thank you to all the members of the Food, Water, Waste Research Group who have provided with technical support and have been a pleasure to work alongside. Special thanks goes out to Miss Andrea Garduno Jimenez who started this research journey with me and who has provided support, help and baked goods throughout the last three years.

Thank you to all my friends for always being there when help is needed. I am forever thankful for Dr Eleanor Hodgson, whose kindness, sympathy and care has supported me during the final push to complete this work.

Last but not least, I wish to express my deepest gratitude to my parents, grandparents, brothers and sister, for their truly unconditional and unwavering support throughout my studies, without them none of this would be possible.

TABLE OF CONTENTS

ABSTRACT ..... **i**

PUBLICATIONS AND PRESENTATIONS ..... **iv**

ACKNOWLEDGEMENTS ..... **vi**

TABLE OF CONTENTS ..... **vii**

LIST OF FIGURES ..... **xv**

LIST OF TABLES ..... **xx**

LIST OF ABBREVIATIONS ..... **xxii**

CHAPTER 1: INTRODUCTION **1**

1.1 RESEARCH BACKGROUND 1

1.2 RESEARCH PROBLEM, HYPOTHESIS AND RESEARCH QUESTIONS 3

1.3 AIM AND OBJECTIVES 4

1.4 EXPLANATION OF PUBLISHED WORKS RELATIONSHIP TO THESIS AND STATEMENT ABOUT JOINT AUTHORSHIP 5

1.4.1 PUBLISHED WORK 1 – CLOUD MANUFACTURING AS A SUSTAINABLE PROCESS MANUFACTURING ROUTE 5

1.4.2 PUBLISHED WORK 2 – INTELLIGENT RESOURCE USE TO DELIVER WASTE VALORISATION AND PROCESS RESILIENCE IN MANUFACTURING ENVIRONMENTS 5

1.4.3 PUBLISHED WORK 3 – CONSIDERATIONS, CHALLENGES AND OPPORTUNITIES FOR MANUFACTURERS WHEN DEVELOPING DATA-DRIVEN MODELS FOR PROCESS MANUFACTURING SYSTEMS 6

1.4.4 PUBLISHED WORK 4 – DATA-DRIVEN MODELLING OF PROCESS MANUFACTURING SYSTEMS 6

1.4.5 PUBLISHED WORK 5 – MULTIPLE TARGET MODELLING TO ENABLE SUSTAINABLE PROCESS MANUFACTURING 7

1.5 THESIS OVERVIEW 7

CHAPTER 2: LITERATURE REVIEW **10**

2.1 STATEMENT OF INTEGRATION OF PUBLISHED WORKS WITHIN CHAPTER 10

2.2 IMPORTANCE OF MANUFACTURING 10

2.2.1 HISTORY OF MANUFACTURING 11

2.2.2 THE DIFFERENCE BETWEEN PROCESS AND DISCRETE MANUFACTURING 13

2.3 SUSTAINABLE PROCESS MANUFACTURING 15

2.3.1 THE THREE PILLARS OF SUSTAINABILITY 15

2.3.2 DRIVER AND BARRIERS 16

2.3.3 THE CIRCULAR ECONOMY 18

2.3.4 PROCESS MANUFACTURING AND THE CIRCULAR ECONOMY 20

2.4 CLOUD MANUFACTURING 21

2.4.1 PREVIOUS CLOUD MANUFACTURING DEFINITIONS 23

2.4.2 INTRODUCING THE CLOUD MANUFACTURING ARCHITECTURE 24

2.5 MATHEMATICAL MODELS 27

2.5.1 INTRODUCTION TO MATHEMATICAL MODELLING 27

2.5.2 FIRST PRINCIPLE MODELS VS EMPIRICAL MATHEMATICAL MODELS 27

2.5.3 DATA-DRIVEN MODELS 28

2.5.4 MACHINE LEARNING 32

2.6 CASE STUDIES DEVELOPING A CLOUD MANUFACTURING SYSTEM 40

2.6.1 PREVIOUS CLOUD MANUFACTURING CASE STUDIES 40

2.6.2 INDUSTRY LEAD RESEARCH INTO CLOUD MANUFACTURING 44

2.7 BIOELECTROCHEMICAL SYSTEMS 44

2.7.1 INTRODUCING BIOELECTROCHEMICAL SYSTEMS 44

2.7.2 UNDERLYING PRINCIPLES OF BIOELECTROCHEMICAL SYSTEMS 45

2.7.3 APPLICATION OF BIOELECTROCHEMICAL SYSTEMS 48

2.7.4 TYPES OF BIOELECTROCHEMICAL SYSTEMS 48

2.7.5 PREVIOUS WORK MODELLING BIOELECTROCHEMICAL SYSTEMS 52

2.8 THE H^2AD BIOPROCESS 53

2.8.1 INTRODUCING THE H^2AD BIOPROCESS 53

2.8.2 UNDERLING PRINCIPLES OF THE H^2AD BIOPROCESS 55

2.8.3 THE H^2AD REACTOR DESIGN 59

2.8.4 H^2AD SUITABILITY AS A CLOUD MANUFACTURING CASE STUDY 59

CHAPTER 3: APPLICATION OF CLOUD MANUFACTURING TO PROCESS MANUFACTURING 61

3.1 STATEMENT INTEGRATION OF PUBLISHED WORK WITHIN CHAPTER 61

3.2 HIGHLIGHTS 61

3.3 INTRODUCTION 61

3.4 EXPANDING THE CLOUD MANUFACTURING CONCEPT 62

3.4.1 REDEFINING THE CLOUD MANUFACTURING DEFINITION 62

3.4.2 FORMULATING THE KEY CHARACTERISTICS OF CLOUD MANUFACTURING 63

3.5 CLOUD MANUFACTURING AS A ROUTE TO SUSTAINABLE MANUFACTURING 66

3.5.1 CLOUD MANUFACTURING APPLICATION FOR INCREASING SOCIAL SUSTAINABILITY 66

3.5.2 CLOUD MANUFACTURING ENABLING COLLABORATION FOR INNOVATIVE DESIGNS 67

3.5.3 GREATER AUTOMATION FROM THE ADOPTION OF CLOUD MANUFACTURING SYSTEMS 68

3.6 DEVELOPING CLOUD MANUFACTURING APPLICATION TO PROCESS MANUFACTURING 69

3.6.1 APPLICATION OF COLLABORATIVE DESIGN AND GREATER AUTOMATION TO PROCESS MANUFACTURING 69

3.6.2 CLOUD MANUFACTURING FOR WASTE VALORISATION 71

3.6.3 HOW CLOUD MANUFACTURING INCREASES PROCESS RESILIENCE 75

3.7 SYTHESISING CLOUD MANUFACTURING LIMITATIONS AND FUTURE RESEARCH DIRECTIONS FOR PROCESS MANUFACTURING ... 81

3.7.1 DATA SHARING AMONG PROCESS MANUFACTURERS 81

4.7.2 PROCESSING PLANT SECURITY 81

3.7.3 CHARACTERISATION OF PROCESS STREAMS 82

3.7.4 BARRIERS PREVENTING SMES ADOPTING NEW TECHNOLOGIES 82

3.8 CONCLUSION 83

CHAPTER 4: CONSIDERATIONS, CHALLENGES AND OPPORTUNITIES WHEN MODELLING PROCESS MANUFACTURING SYSTEMS 85

4.1 STATEMENT INTEGRATION OF PUBLISHED WORK WITHIN CHAPTER 85

4.2 HIGHLIGHTS 85

4.3 INTRODUCTION 85

4.4 OUTLINING CASE STUDIES 86

4.4.1. CASE STUDY 1: WASTE/WATER MANAGEMENT 86

4.4.2. CASE STUDY 2: CRAFT BREWERIES 86

4.5 CONSIDERATION FOR DATA-DRIVEN PROCESS MANUFACTURING MODELS 88

4.5.1 DEFINING THE MODEL'S GOAL 89

4.5.2 UNDERSTANDING THE REQUIREMENTS FROM THE MODEL 89

4.5.3 DATA CONSIDERATIONS 91

4.6 CHALLENGES FOR DATA-DRIVEN PROCESS MANUFACTURING MODELS 94

4.6.1 AVAILABILITY OF PROCESS MANUFACTURING DATA 95

4.6.2 VARIABILITY IN PROCESS MANUFACTURING SYSTEMS 96

4.6.3 RECOGNIZING THE MANUFACTURING SYSTEM'S BOUNDARIES 99

4.6.4 EVALUATION OF THE MODEL'S OUTPUT ON UNSEEN DATA 102

4.7 OPPORTUNITIES ARISING FROM DATA-DRIVEN MODELLING OF PROCESS MANUFACTURING SYSTEMS 103

4.7.1 ENABLING AFFORDABLE CHARACTERISATION OF PROCESS STREAMS 104

4.7.2 ENSURING PROCESS RESILIENCE 105

4.7.3 EVALUATING WASTE VALORISATION POTENTIAL 106

4.8 CONCLUSIONS 107

CHAPTER 5: DEVELOPING NEW PROCEDURES FOR MODELLING PROCESS MANUFACTURING SYSTEMS 109

5.1 STATEMENT INTEGRATION OF PUBLISHED WORK WITHIN CHAPTER 109

5.2 HIGHLIGHTS 109

5.3 INTRODUCTION 110

5.4 PART A: OUTLINING A NEW APPROACH FOR MODELLING PROCESS MANUFACTURING SYSTEMS 110

5.4.2 MATERIAL AND METHODS 114

5.4.3 RESULTS AND DISCUSSION 129

5.5 CONCLUSIONS FROM PART A 143

5.6 PART B: MULTIPLE TARGET MODELLING TO ENABLE SUSTAINABLE PROCESS MANUFACTURING 144

5.6.1 THE NEED FOR MULTIPLE TARGET MODELS 144

5.6.2 MATERIAL AND METHODS 144

5.6.3 RESULTS AND DISCUSSION 153

5.7 CONCLUSIONS FROM PART B 170

CHAPTER 6: APPLYING THE DEVELOPED DATA-DRIVEN MODELLING METHODOLOGY TO A REAL-WORLD PROCESS MANUFACTURING SYSTEM 172

6.1 HIGHLIGHTS 172

6.2 INTRODUCTION 172

6.3 OVERVIEW OF THE PROCESS MANUFACTURING SYSTEM MODELLED 173

6.4 THE CONSIDERATIONS AND CHALLENGES WHEN MODELLING FE REMOVAL CASE STUDY 177

6.4.1 CONSIDERATIONS WHEN MODELLING FE REMOVAL CASE STUDY 177

6.4.2 CHALLENGES WHEN MODELLING FE REMOVAL CASE STUDY 180

6.5 RESULTS AND DISCUSSION 182

6.5.1 DATA PREPARATION, PRE-PROCESSING AND PARTITIONING 182

6.5.2 DEFINING THE SYSTEM BOUNDARIES 184

6.5.3 DATA-DRIVEN MODELLING RESULTS 186

6.5.5 DETERMINING THE RELATIONSHIP BETWEEN SYSTEM PARAMETERS AND PERFORMANCE 198

6.6 CONCLUSIONS 203

CHAPTER 7: DATA-DRIVEN MODEL RELEVANCE TO MANUFACTURERS AND PLACE WITHIN A CLOUD MANUFACTURING SYSTEM 205

7.1 HIGHLIGHTS 205

7.2 INTRODUCTION 205

7.3 DATA-DRIVEN MODELS RELEVANCE TO THE MANUFACTURER 206

7.3.1 UTILISING A DATA-DRIVEN MODEL TO GAIN A DEEPER UNDERSTANDING OF A PROCESS MANUFACTURING SYSTEM 206

7.3.2 REAL-TIME MONITORING AND OPTIMISATION OF PROCESS MANUFACTURING SYSTEMS 208

7.3.3 THE RELATIONSHIP BETWEEN MODEL PREDICTION ERROR AND PROCESS MANUFACTURING SYSTEMS 210

7.3.4 SERVICE-ORIENTED BUSINESS MODEL 211

7.4 DATA-DRIVEN MODELLING ENABLING INTELLIGENT CLOUD MANUFACTURING 214

7.4.1 H^2AD CASE STUDY 214

7.5 CLOUD MANUFACTURING INCREASING A DATA-DRIVEN MODEL'S PREDICTIVE CAPABILITY 220

7.6 CONCLUSION 220

CHAPTER 8: CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH 222

8.1 CONCLUSIONS 222

8.1.1 HOW DOES CLOUD MANUFACTURING OFFER PROCESS MANUFACTURERS A NEW ROUTE TO ACHIEVE SUSTAINABLE MANUFACTURING? 223

8.1.2 WHAT CONSIDERATIONS AND CHALLENGES DO PROCESS MANUFACTURERS FACE WHEN DEVELOPING A DATA-DRIVEN MODEL THAT IS REPRESENTATIVE OF THEIR PROCESS MANUFACTURING SYSTEMS 224

8.1.3 WHAT ARE THE CURRENT STEPS WHEN DEVELOPING A DATA-DRIVEN MODEL AND WHAT ADDITIONAL STEPS ARE REQUIRED TO OVERCOME THE UNIQUE CHALLENGES WHEN MODELLING PROCESS MANUFACTURING SYSTEMS? 226

8.1.4 HOW WILL THE DEVELOPMENT OF DATA-DRIVEN MODELS SUPPORT THE TRANSITION FROM THE TRADITIONAL MANUFACTURING PARADIGM TO CLOUD MANUFACTURING? 229

8.2 RECOMMENDATIONS AND FUTURE WORK 230

8.2.1 GENERATING ADDITIONAL DATA FROM LAB-SCALE BIOELECTROCHEMICAL SYSTEMS 230

8.2.2 UTILISING A DATA-DRIVEN MODEL FOR SELF-OPTIMISATION 231

8.2.3 MANUFACTURING PRODUCTS FROM PROCESS MANUFACTURERS WASTE STREAMS 232

8.2.4 FURTHER DEVELOPING CLOUD MANUFACTURING SYSTEMS OF PROCESS MANUFACTURING ENVIRONMENTS 232

8.2.5 RESOURCE RECOVERY BY THE H^2AD BIOPROCESS 232

REFERENCES 233

APPENDIX A: PUBLICATION 1 274

APPENDIX B: PUBLICATION 2 313

APPENDIX C: PUBLICATION 3 325

APPENDIX D: PUBLICATION 4 356

APPENDIX E: PUBLICATION 5 386

APPENDIX F: H^2AD MODELLING DATA 415

APPENDIX G: COAL AUTHORITY DATA 423

LIST OF FIGURES

Figure 2.1. Timeline of industrial revolutions, adapted from [31] 13
Figure 2.2. The typical Venn diagram representation of the three pillars of sustainability 16
Figure 2.3. The nested dependencies model of sustainability 16
Figure 2.4. The circular economy model, as presented by WRAP [7]. 19
Figure 2.5. Idealised circular economy presented by Reh [55]. 21
Figure 2.6. Cloud manufacturing architecture, adapted from [64,65] 25
Figure 2.7 Documents by year across in the fields of engineering and chemical engineering. Data were taken from a Scopus search. 34
Figure 2.8. Samples from a Gaussian process prior and posterior taken from Schulz [148] . Grey lines indicate samples from the GP. Black dots mark empirical observations. The dark grey line marks the current mean of the GP. The red triangle shows the prediction for the new input point. 36
Figure 2.9. Architecture of a random forest, taken from [151]. 37
Figure 2.10. A schematic diagram of the support vector regression, taken from [154] 38
Figure 2.11. A node that receives n inputs 39
Figure 2.12. Artificial neural networks architectures (A) single layer neural network, (B) shallow neural network and (C) deep neural network 39
Figure 2.13. Fundamental bioelectrochemical process adapted from [177]. 47
Figure 2.14. Schematic representation of possible reactions occurring at biologically catalysed anodes and cathodes, taken from [176]. 48
Figure 2.15. H2AD compared to traditional anaerobic digestion (data provided by Lindhurst Engineering Ltd.) [AD = anaerobic digestion; H2AD = H2AD micro AD] 54
Figure 2.16. Lindhurst Engineering H^2AD bioprocess installed at Sutton Bonington 58
Figure 2.17. The H^2AD reactor schematic 59
Figure 3.1 Cloud manufacturing word cloud derived from [64,65,75–78]. 63
Figure 3.2 Waste hierarchy adapted from Directive 2008/98/EC [38]. 71
Figure 3.3 Integrating the cradle-to-cradle concept within the circular economy model 72

Figure 4.1 Concentration of four water quality parameters of H^2AD feedstock during 2018: (A) dissolved oxygen (DO), (B) pH, (C) total suspended solids (TSS) and (D) chemical oxygen demand (COD). 98
Figure 4.2 Amplitude, temperature and speed of sound data during fermentation of four batches of Slap in the Face beer, collected from Totally Brewed in 2018. 99
Figure 4.3 Figure showing training data that is only representative of one part of the whole manufacturing system. Where blue dots are training data, the blue box represents model boundaries and the red box represents system boundaries. 100
Figure 4.4 Scatter plot matrix for four of case study 1's feedstock's characteristics [Temp.: temperature; DO: dissolve oxygen; TSS: total suspended solids; COD: chemical oxygen demand] 101
Figure 4.5 (A) Example of unseen data only evaluating one region of the model, (B) example of unseen data evaluating the whole of the model, and (C) evaluating the whole of the model and the extrapolating capabilities of the model. Where blue dots are training data, red dots are unseen data and blue box is the model boundaries 103
Figure 5.1. Blue region represents the defined manufacturing system for: (A) Strongly represented system, weak variability is captured; (B) Moderately represented system, moderate variability captured; (C) All variability captured but system contains poorly represented regions. 112
Figure 5.2. The proposed CRISP-DM framework has been expanded to include two new steps (the green boxes) in the modelling stage. 114
Figure 5.3 Heatmaps displaying the scaled correlation coefficients between variables in (A) the original WQA data and (B) the original plus synthetic data. 121
Figure 5.4 Results from feature selection and dimensionality reduction techniques, (A) neighbourhood component analysis and (B) principal component analysis. 125
Figure 5.5 Hierarchy of data partitioning during the development of a data-driven model. 126
Figure 5.6 Spread of unseen data throughout the system. Blue markers are the model development data points, red markers are the unseen data points and the red regions are outside of the system. 127
Figure 5.7 Scatter plot matrix of variables in the model development dataset 134
Figure 5.8 (A) Heatmap displaying Pearson's correlation between the model development data and (B) boxplot displaying the distribution of the model development data 135
Figure 5.9 Traffic light system informing manufacturers if new data points are within the manufacturing system (green region), approaching the boundaries of the system (yellow region) or outside the system (red system). Where, the green line is the median

of each variable, the blue is a data point within the system and the red line is a data point outside of the system. 136
Figure 5.10 The predicted percentage of COD removal plotted against the actual value (1) and the residuals plotted against the actual values (2), for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network..... 139
Figure 5.11 Residuals of unseen predictions plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network. The red region is outside of the system. 142
Figure 5.12 Annual wastewater charge for different UK water and sewerage companies, based on 4m^3 /d of wastewater with a chemical oxygen demand of 85,000 mg/L and total suspended solids of 984 mg/L using WRAP Mogden calculator [361]. 146
Figure 5.13 The spread of unseen data throughout the system. Blue markers are the model development data points, red markers are the unseen data points and the blue box represents the model boundaries. 148
Figure 5.14 Problem transformation model, predicting three outputs. Where the model input data (X) is represented by the orange box, the model output prediction (Yi) by the green box, the single target models are within the grey box and the multiple target model is within the black box. 149
Figure 5.15 Regression chain multiple target model, predicting three outputs. Where the model input data (X) is represented by the orange box, the model output prediction (Yi) by the green box, the single target models are within the grey box and the multiple target model is within the black box. 150
Figure 5.16 Ensemble of Regression Chain (ECR model) predicting three outputs. Where the model input data (X) is represented by the orange box, the model output predictions (Yˆ_i) by the green box, the single target models are within the grey box and the ERC model is within the black box..... 152
Figure 5.17 Architecture of a Multiple Target Artificial Neural Network (MT-ANN), with 8 inputs, 2 outputs and 1 hidden layer containing 6 neurons..... 153
Figure 5.18 Evaluation metrics of the problem transformation modes. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 generation output. 155
Figure 5.19 Evaluation metrics of algorithm adaption modes. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 output. 157
Figure 5.20 Statistical analysis of algorithm adaption MT modes performance on unseen data. Where the black line is the average performance of all three outputs, the

blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 output. 158
Figure 5.21 Boxplots comparing the distributions between the actual unseen data and the models' predictions for each output. 159
Figure 5.22 Surface plots of the H^2AD process conditions impact on percentage of COD removal. Figure 5.23 Surface plots of the H^2AD process conditions impact on percentage of TSS removal. 162
Figure 5.24. Surface plots of the H^2AD process conditions impact on the generation of CH4. 163
Figure 6.1. Three different surface catalysed oxidation of ferrous iron (SCOOFI) reactor configurations. (A) Is a down-flow reactor with no packing material, (B) is a down-flow trickling reactor and (C) is an up-flow reactor. 176
Figure 6.2: Iron removal results from a trial conducted by Coal Authority. The down-flow reactor no packing material, down-flow trickling reactor and up-flow reactor results are represented by the blue, yellow and purple lines respectively. The red line represents the desludging day. 177
Figure 6.3. Percentage of iron removal by an up-flow surface catalysed oxidation of ferrous iron (SCOOFI) reactor during the trial, with desludging day represented by the red line. 178
Figure 6.4. Concentration of six characteristics of mining wastewater: (A) iron (Fe) concentration, (B) total suspended solids (TSS) concentration, (C) flowrate, (D) pH, (E) dissolved oxygen (DO) concentration and (F) conductivity. 181
Figure 6.5 Spread of unseen data throughout the system. Blue markers are the model development data points, red markers are the unseen data points and the red regions are outside of the system. 184
Figure 6.6 Timeseries data of iron removal rates (A) before and (B) after the data cleaning stage. 185
Figure 6.7 Scatter plot matrix of the mining wastewater characteristics 186
Figure 6.8 (A) Heatmap displaying Pearson's correlation between variables (B) boxplot displaying the distribution of data 186
Figure 6.9 Statistical analysis of the models' ability to fit the model development data. The number after the model algorithms states the modelling strategy. 189
Figure 6.10 Comparing the statistical analysis results of two different modelling strategies ability to fit the model development data. Where the blue bars display the analysis of modelling strategy 1's models and the red bars display the analysis of modelling strategy 2's models. 190

Figure 6.11 Residuals plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network. Where the blue dots are the residuals from modelling strategy 1's models and the red dots are the residuals from modelling strategy 2's models. 191
Figure 6.12 Statistical analysis of the models' ability to fit the unseen data (a) inside the system and (B) outside the system. The number after the model algorithms states the modelling strategy. 193
Figure 6.13 Comparing the statistical analysis results of two different modelling strategies ability to fit the unseen data. Where the blue bars display the analysis of modelling strategy 1's models and the red bars display the analysis of modelling strategy 2's models. 195
Figure 6.14 Residuals of unseen predictions plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network. The red region is outside of the system. 196
Figure 6.15 The GP model predictions plotted against the model development data (black crosses) and the unseen data (red crosses). 198
Figure 6.16 Surface plots showing the effects of mining wastewater characterises on the %Fe Removed as predicted by the GP model 200
Figure 6.17 Surface plots showing the effects of mining wastewater characterises on the %Fe Removed as predicted by the GP model 201
Figure 6.18 Surface plots showing the effects of mining wastewater characterises on the %Fe Removed as predicted by the GP model 202
Figure 7.1 Example of a GUI built for a CSTR with built-in predictive modelling capabilities. 209
Figure 7.2. The four classifications of prediction error in process manufacturing systems, where the red shading represents the region model. 210
Figure 7.3 Example of a private cloud manufacturing system built for H^2AD system. 216
Figure 7.4 Example of a community cloud manufacturing system built for H^2AD system. 217
Figure 7.5 Example of a public cloud manufacturing system built for H^2AD system. 219

LIST OF TABLES

Table 2.1 Differences between discrete and process manufacturing 14
Table 2.2. Cloud manufacturing definitions 23
Table 2.3 Cloud manufacturing case studies from academia 41
Table 2.4 Summary of the types of microbial-based bioelectrochemical systems (BES). 50
Table 3.1 Examples of cloud manufacturing's application to process manufacturing industries. 78
Table 4.1 Summary of case studies (CSs) 87
Table 4.2 Examples of data-driven models requirements for the 2 case studies, waste/water management and brewery 90
Table 4.3 Causes of variation in process manufacturers' feedstock and some industry examples 96
Table 5.1 Description of model's input variables (sourced from [331]) and relevance to H^2AD bioprocess 116
Table 5.2 Correlation strength between model input variables and COD removal efficiency, where r is Pearson's correlation coefficient. 121
Table 5.3 Statistical and visualisation techniques used to describe features within the data when defining the manufacturing system. 123
Table 5.4 The hyperparameters for each machine learning algorithm used. 128
Table 5.5 Statistical analysis of the variables in the model development dataset 132
Table 5.6 The hyperparameters values for the final machine learning models 137
Table 5.7 Statistical comparisons of the predictive capabilities of the Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural network (ANN) models 140
Table 5.8 Optimal process conditions values for each of the H^2AD bioprocess outputs 161
Table 5.9 Chemical oxygen demand (COD) and total dissolved solids (TSS) concentration in the H^2AD input and output streams for each wastewater composition scenario for both H^2AD set ups. 167
Table 5.10 Economic analysis of the wastewater disposal costs before and post H^2AD treatment and the economic benefit from the methane generated for **Configuration 1**. 168

Table 5.11 Economic analysis of the wastewater disposal costs before and post H^2AD treatment and the economic benefit from the methane generated for **Configuration 2**. 169
Table 6.1 Statistical analysis of the variables in the mining wastewater data provided by the Coal Authority. 185
Table 6.2 The hyperparameters values for the machine learning models built for modelling strategy 1 187
Table 6.3 The hyperparameters values for the machine learning models built for modelling strategy 2 188
Table 6.4 Intermediate value of each parameter. 198
Table 7.1 Examples of problems faced by overestimating or underestimating a process manufacturing output in low and high regions. 211

LIST OF ABBREVIATIONS

AM	Additive manufacturing
NH_4^+	Ammonium
AD	Anaerobic digestion
AI	Artificial intelligence
ANN	Artificial neural network
BES	Bioelectrochemical systems
COD	Chemical oxygen demand
R^2	Coefficient of determination
BES	Bioelectrochemical systems
CIP	Clean-in-place
CM	Cloud manufacturing
CBDM	Cloud-based design and manufacturing
Cu	Copper
CRM	Critical raw materials
DDM	Data-drive model
DoE	Design of experiments
ECR	Ensemble of regressor chains
ERP	Enterprise resource planner
EU	European Union
FMCG	Fast-moving consumer goods
GP	Gaussian process
GDP	Gross domestic product
HaaS	Hardware-as-a-service
HRT	Hydraulic residence time
IIoT	Industrial internet of things
IaaS	Infrastructure-as-a-service
MCS	Manufacturing cloud service
MCC	Maximum correlation chain
MAE	Mean average error
MAPE	Mean average percentage error
MEC	Microbial electrolysis cell
MES	Microbial electrosynthesis
MSC	Microbial solar cell
MFC	Microbial fuel cell
MT	Multiple target
MDC	Microbial desalination cell
NCA	Neighbourhood component analysis
NO_3^-	Nitrate
NO_2^-	Nitrite
nRMSE	Normalised root mean square error
PCC	Pair-copula construction
PO_4^{3-}	Phosphate
RMFC	Plant microbial fuel cells
PaaS	Platform-as-a-Service
PCA	Principal component analysis
PDF	Probability density function
POA	Product-oriented architecture
REM	Proton exchange membrane
RC	Regression chains
RF	Random forest

RMSE	Root mean squared error
ST	Single target
SOA	Service-oriented architecture
SME	Small and medium enterprise
SaaS	Software as a service
SO_4^2	Sulphate
SCOOFI	Surface catalysed oxidation of ferrous iron
SVR	Support vector regression
SP	System pressure
ST	System temperature
TC	Total carbon
TIC	Total inorganic carbon
tN	Total nitrogen
TSS	Total suspended solids
WH	Water hardness;
WQA	Water quality analysis
WRAP	Waste and Resources Action Program
Zn	Zinc

CHAPTER 1: INTRODUCTION

1.1 RESEARCH BACKGROUND

Manufacturing drives the economies of all countries from low to high incomes; contributing 16% to global gross domestic product and 23% of global employment in 2017 [1]. Broadly speaking, manufacturing can be split into two branches; discrete manufacturing (e.g. automobiles, computers and electronics, textiles) and process manufacturing (e.g. chemicals, food and drink, pharmaceutical, fast-moving consumer goods - FMCG). The difference being that discrete manufacturing is the cutting and assembly of a bill of materials into a final distinct product [2]; whereas, process manufacturing is the thermal, chemical and/or biochemical conversion of resources into products, by-products and waste streams [3]. Both share the core value that their products and processes must be profitable to remain in business. However, there is mounting pressure for businesses to become not just economically sustainable but also environmentally sustainable. This has been driven by government regulations and legislations, and increasing public pressure for products where the process and life cycle better considers the resources and related ecosystems. The means and ends of achieving this goal differ between discrete and process manufacturing. Discrete manufacturers have focused on redesigning their products to use sustainable resources. Whereas, process manufacturers are often able to keep the characteristics of their products the same but redesign their systems to use alternative sustainable feedstocks. This work develops ideas on how intelligent resource use within process manufacturing environments may enable waste valorisation and enhance process resilience. A comprehensive discussion of the meaning and implications of the two terms to process manufacturing is offered in chapter 3. For the purposes of this work waste valorisation and process resilience are defined as:

- Waste valorisation: refers to indsutrial activeites that add value to manufacturers proces streams that currently under utilised.
- Process resislicence: decribes a maufacturing system’s capability to migitage the effect of external and internal varaibility impacted the system’s perfromance.

2019 is the International Year of the Periodic Table of Chemical Elements, celebrating the 150th anniversary of discovering the periodic system by Russian scientist Dmitri Mendeleev. However, through unsustainable resource exploitation 23 out of the 118

elements known will have depleted reserves within 20-50 years [4]. In 2017, the EU expanded its list of critical raw materials (CRM) to 27, defined as materials considered to be of high importance to the EU economy and of high risk to their supply [5]. This recognition comes at a tipping point, where the resources required to sustain the global population exceeds those available [4]. However, often when an element/resource is consumed it is not destroyed, but dispersed through the environment as wastes. Recovering resources in discrete manufacturing relies mainly on public engagement, e.g. taking used electronics to recycling centres [4]. For process manufacturing, 50-100% of lost resources are contained in waste/water, which feeds waste management bodies (waste and/or wastewater treatment plants (WWTPs) or natural environments). [6]. Depending on the country, the waste management body can be the surrounding environment, meaning the resource is lost, or it can be a wastewater treatment process where potential exists for resource recovery beyond water treatment, for reuse.

The UK-based Waste and Resources Action Program (WRAP) charity defines the circular economy as "an alternative to a traditional linear economy (make, use, dispose) in which we keep resources in use for as long as possible, extract the maximum value from them whilst in use, then recover and regenerate products and materials at the end of each service life" [7]. The application of circular economy thinking to process manufacturing requires deeper consideration than discrete manufacturing, where it is simple to visualise a discrete product being reused, repaired and recycled (e.g. fixing a kettle). This school of thought does not lend itself to some process manufacturing products; for example, it is not possible to reuse, repair or recycle a pharmaceutical drug once consumed. However, by expanding the system to include waste produced during manufacturing and indirect waste from the consumer (in this example human waste entering wastewater treatment systems) it is possible to recapture the resource into the economy through waste recovery and valorisation.

What is a waste for one process manufacturing system may be viewed as a resource for another. The challenge of brokering relationships between manufacturers to form innovative collaborations that find ways to utilise waste from one as a raw material for another is known as industrial symbiosis [8–10]. There are examples of support networks for industrial symbioses, such as National Industrial Symbiosis Programme (UK) [11], Cleantech Ostergotland (Sweden) [12] and Kalundbord (Denmark) [13]. However, there has yet to be industry wide adoption of industrial symbiosis suggesting more needs to be done to help manufacturers overcome industrial symbiosis barriers [14].

1.2 RESEARCH PROBLEM, HYPOTHESIS AND RESEARCH QUESTIONS

Allowing resources to flow between manufacturing systems will require not only a redesign of existing systems but a framework in place which intelligently identifies a more sustainable resource recovery route and matches waste producers and consumers. Recently, cloud manufacturing (CM) has emerged as a new manufacturing paradigm. Cloud manufacturing can be defined as [3]:

> "The cloud manufacturing model is the concept of sharing manufacturing capabilities and resources on a cloud platform capable of making intelligent decisions to provide the most sustainable and robust manufacturing route available."

Although there has been extensive literature on CM application in discrete manufacturing, there has yet to be a thorough evaluation of the unique opportunities CM offers process manufacturers and how CM helps to achieve a sustainable manufacturing future. This work hypothesises that CM will enable the transformation of process manufacturing from the traditional linear business model to a sustainable, circular approach that will enhance both process resilience and waste valorisation. To achieve CM's full potential will rely on the development of mathematical models that describe the manufacturing systems virtualised on the manufacturing cloud, as these models will enable the manufacturing cloud to intelligently determine the most sustainable manufacturing route. This work proposes that data-driven models (DDMs) utilisation by process manufacturers is set to increase, being driven by a raise in the rate of manufacturing data generation by current manufacturing processes and access to affordable computing processing power from cloud computing. In order to evaluate the opportunities from CM, and underlying DDMs, for process manufacturers, several research questions were developed for this work.

1. How does CM offer process manufacturers a new route to achieve sustainable manufacturing?
2. What considerations and challenges do process manufacturers face when developing a DDM that is representative of their manufacturing system?
3. What are the current steps when developing a DDM and what additional steps are required to overcome unique challenges when modelling process manufacturing systems?

4. How will the development of DDMs support the transition from the traditional manufacturing paradigm to CM?

1.3 AIM AND OBJECTIVES

The aim of this research is to evaluate the potential of CM principles and underlying DDMs to improve process manufacturing, specifically from process resilience and waste valorisation perspectives.

The specific objectives of this research are:

1. Define and develop CM characteristics and applications to understand the opportunities for how CM may be utilised to evaluate process manufacturing systems in the collection and intelligent use of data.
2. Work with a SME for industrial relevance, collecting manufacturing data from a process manufacturing environment over a period of time to understand the system's variability.
3. To develop a protocol to aid process manufacturers to develop DDMs that are representative of their manufacturing system.
4. Develop DDMs, using data collected, in order to fit data and evaluate how the ability to predict will increase process resilience of a waste valorisation case study.
5. Develop multiple target DDMs, in order to determine the most sustainable manufacturing configuration dependant on economic and environmental constraints.
6. To prove that modelling methodology works when applied to a new process manufacturing environment.

1.4 EXPLANATION OF PUBLISHED WORKS RELATIONSHIP TO THESIS AND STATEMENT ABOUT JOINT AUTHORSHIP

Published works, and soon to be published works, have been integrated into the structure of this thesis. Sections of the published works have been extracted and integrated into the relevant thesis chapters. This method was chosen over directly inserting the published works, as it minimised unnecessary duplication and allowed the published works to be expanded on within the thesis body. All the published works can be found in their complete form in the Appendices.

The flowing section shall outline the published works relationship within the thesis and identify the principal authors of each publication.

1.4.1 PUBLISHED WORK 1 – CLOUD MANUFACTURING AS A SUSTAINABLE PROCESS MANUFACTURING ROUTE

Principal authors: Oliver Fisher

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2018. Cloud manufacturing as a sustainable process manufacturing route. *Journal of Manufacturing Systems*. 47; 35-68. [Impact Factor: 2.770; SNIP 2.191]

Relationship to thesis: This paper explores CM potential to enable waste valorisation and enhance process resilience within process manufacturing systems, two key themes of this work.

Integration within thesis structure: Sections of this paper have been integrated into the chapters 2, 3 and 8.

1.4.2 PUBLISHED WORK 2 – INTELLIGENT RESOURCE USE TO DELIVER WASTE VALORISATION AND PROCESS RESILIENCE IN MANUFACTURING ENVIRONMENTS

Principal authors: Oliver Fisher

Fisher O, Watson N, Escrig J, Gomes RL, 2019. Intelligent resource use to deliver waste valorisation and process resilience in manufacturing environments. *Johnson Matthey Technology Review*. 64; 93-99. [Impact Factor: 1.27; SNIP: 0.669]

Relationship to thesis: This paper This article aims to close the knowledge gap by exploring how an industrial digital technology (data-driven modelling) may facilitate and advance circular economy principles within process manufacturing systems, specifically waste valorisation and process resilience.

Integration within thesis structure: Sections of this paper have been integrated into the chapters 1, 2, 3 and 8.

1.4.3 PUBLISHED WORK 3 – CONSIDERATIONS, CHALLENGES AND OPPORTUNITIES FOR MANUFACTURERS WHEN DEVELOPING DATA-DRIVEN MODELS FOR PROCESS MANUFACTURING SYSTEMS

Principal authors: Oliver Fisher

Fisher O, Watson N, Escrig J, Witt R, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Considerations, challenges and opportunities for manufacturers when developing a data-driven model. *Computers & Chemical Engineering* (**Accepted**, 20 Apr 2020). [Impact Factor: 3.113; SNIP: 1.613]

Relationship to thesis: This paper proposes a new protocol when modelling process manufacturing systems. The paper outlines the considerations and challenges often overlooked when modelling process manufacturing systems. For the first time, the challenge of determining if a DDM built from manufacturing data is representative of a process manufacturing systems is addressed. This is relevant to the work in this thesis that builds DDMs from process manufacturing data.

Integration within thesis structure: Sections of this paper have been integrated into the chapters 2, 4 and 8.

1.4.4 PUBLISHED WORK 4 – DATA-DRIVEN MODELLING OF PROCESS MANUFACTURING SYSTEMS

Principal authors: Oliver Fisher

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Data-driven modelling of process manufacturing systems. *Chemical Engineering Journal* [Impact Factor: 6.735; SNIP: 1.863]

Relationship to thesis: This paper outlines a new procedure for modelling process manufacturing systems. This procedure was developed from the work conducted in this thesis when modelling the H^2AD process manufacturing system.

Integration within thesis structure: Sections of this paper have been integrated into the chapters 2, 6 and 8.

1.4.5 PUBLISHED WORK 5 – MULTIPLE TARGET MODELLING TO ENABLE SUSTAINABLE PROCESS MANUFACTURING

Principal authors: Oliver Fisher

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Multiple target modelling to enable sustainable process manufacturing. *Journal of Cleaner Production*. [Impact Factor: 6.395; SNIP: 2.308]

Relationship to thesis: This paper proposes a novel multiple target (MT) DDM approach for modelling process manufacturing systems. How a MT-DDM can then be used to increase the economic and environmental sustainability of the H^2AD process manufacturing system studied in this work is explored.

Integration within thesis structure: Sections of this paper have been integrated into the chapters 2, 6 and 8.

1.5 THESIS OVERVIEW

This work investigates how intelligent resource use in process manufacturing may lead to a sustainable manufacturing future. Chapter 2 begins with a review of the literature to explain the recently emerged manufacturing paradigm, cloud manufacturing (CM). Previous definitions, characteristic and architectures of CM are reviewed to understand how CM may enable sustainable process manufacturing. Chapter 2 outlines the procedure for developing mathematical models and gives an overview of the previous work modelling process manufacturing systems. This chapter also explains the process manufacturing system used a case study for the majority of this work, the H^2AD bioprocess.

The result chapters commence with developing new concepts on the application of CM to process manufacturing, specifically from process resilience and waste valorisation perspectives. In chapter 4 a protocol to aid process manufacturers to develop data-driven models (DDMs) is proposed. Chapter 5 is split into two parts. The first part outlines a new procedure for modelling process manufacturing systems and four DDMs are built predicting the chemical oxygen demand (COD) removal from wastewater by a novel bioprocess. The second part outlines how multiple target DDMs are built for process manufacturing system. Multiple target models are necessary, as the most sustainable manufacturing route has multiple economic and environmental considerations. Five multiple target (MT) DDMs are built predicting the COD removal rate, total suspended solids (TSS) removal rate and volume of methane (CH_4) generated by a novel bioprocess. Chapter 6 applies the modelling methodology applied in chapters

5 and 6 to a new process manufacturing environment, which is the Iron (Fe) remediation from mining wastewaters. Finally, chapter 7 explains the relevance of the developed DDMs to the process manufacturer and how the DDMs work to enable CM.

This thesis has been structured as a *Research Degree by Published Works*. The definition of published works is any research output in the public domain. This alternative format of thesis submission allows a doctoral candidate to submit a thesis for examination which incorporates:

- Sections that are in a format suitable for submission for publication in a peer reviewed journal;
- Materials which are solely and/or partly authored by the student and may be already published, not yet published, not yet published but of publishable

- standard, accepted for publication, or submitted for publication in externally refereed contexts such as journals and conference proceedings;

The published works incorporated into this thesis are included in the Appendices, and an explanation of the published works and their relevance to the thesis is given in Section 1.4.

CHAPTER 2: LITERATURE REVIEW

2.1 STATEMENT OF INTEGRATION OF PUBLISHED WORKS WITHIN CHAPTER

Integrated into this chapter is material from the published or soon to be published work:

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2018. Cloud manufacturing as a sustainable process manufacturing route. *Journal of Manufacturing Systems*. 47; 35-68. [Impact Factor: 2.770; SNIP 2.191]

Fisher O, Watson N, Escrig J, Gomes RL, 2019. Intelligent resource use to deliver waste valorisation and process resilience in manufacturing environments. *Johnson Matthey Technology Review* (**Submitted to journal,** 07 Aug 2019). [Impact Factor: 1.27; SNIP: 0.669]

Fisher O, Watson N, Escrig J, Witt R, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Considerations, challenges and opportunities for manufacturers when developing a data-driven model. *Computers & Chemical Engineering* (**Submitted to journal**, 20 Aug 2019). [Impact Factor: 3.113; SNIP: 1.613]

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Data-driven modelling of process manufacturing systems. *Chemical Engineering Journal* (**Submitted to journal**, 03 Oct 2019). [Impact Factor: 6.735; SNIP: 1.863]

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Multiple target modelling to enable sustainable process manufacturing. *Journal of Cleaner Production*. (**Submitted to journal**, 03 Oct 2019) [Impact Factor: 6.395; SNIP: 2.308]

2.2 IMPORTANCE OF MANUFACTURING

Manufacturing is the process of converting raw materials, components, or parts into finished products that meet a customer's expectations or specifications [15]. Before the first industrial revolution, production was carried out by skilled artisan (e.g. blacksmiths, cobblers, jewellers) and the main purpose was to serve the agricultural industry [16]. However, the industrial revolution led to wealthier countries and society now demands a range of products including automobiles, clean water, electronic devices, pharmaceuticals, processed food and drink, semiconductors and textiles.

In recent decades the manufacturing sectors of the USA, Japan, the UK, France and Italy have lost global market shares due to pressure from China. The exception being Germany, which has instead increased exports to China and Asia [17]. This is in part due to Germany's Government support of the application of manufacturing technology. Since the 2008 recession, policymakers in the West have focused on rebalancing the economy away from the service sector towards manufacturing [18]. Manufacturing in the USA and the UK account for 12% and 9% of the countries' gross domestic product (GDP) respectively [19]. Whereas, in Germany, the manufacturing sector accounts for 23% of GDP [19], which led to the German economy making a faster recovery. The GDP figures only show the direct impact of the manufacturing sector to the economy; however, the impacts of the sector extend more widely than the manufacturing companies themselves. A report by Oxford Economics investigated the full impact of the manufacturing sector on an economy by measuring the additional indirect and induced impacts. The indirect impacts include the economic effect of the supply chains that support the manufacturing sector. Including the indirect impacts raises the GDP impact from manufacturing in the UK from 9% to 15% and the number of people employed from 2.6 million to over five million [20]. The induced effect includes the spending by those employed directly and indirectly by the manufacturing sector. When including the induced impact the figures raised even further to 23% and seven million jobs [20].

The importance of the manufacturing sector extends beyond the economic impact. Manufacturing leads the way for innovation in the UK. In the UK, Over 70% of the business lead research and development spending comes from manufacturing companies [20]. Zaidu *et al.* claim that advanced manufacturing companies are made up of multiple interdisciplinary occupations, including: (1) production managers and directors in manufacturing; (2) biological scientists and biochemists; (3) production and process engineers; (4) metalworking production and maintenance fitters; (5) assemblers of electrical and electronic products [21]. This blend of skills facilitates a creative environment where innovative solutions are generated. Manufacturers are routinely conducting experiments and pilot studies to improve their processes. For example, the NextGen initiative, funded by the EU [22], which aims to demonstrate innovative technological, business and governance solutions for water in the circular economy, in ten high-profile, large-scale, demonstration cases across Europe.

2.2.1 HISTORY OF MANUFACTURING

The invention of the timepiece has been credited as being a large contributor to the first industrial revolution that took place in Britain in 1784, as the ability of accurately

knowing time gave birth to the desire of saving time [23]. It has been this desire to save time and money, while simultaneously increasing production capabilities, which has fuelled the majority of past advances in manufacturing. The first industrial revolution moved Britain from an agricultural-based economy to production-based economy [23].

Since the first industrial revolution, manufacturing has undergone two further industrial revolutions. Each industrial revolution has been defined by significant technological advancement, as shown in Figure 2.1. The second industrial revolution is characterised by the advancement of mass production and transportation. Rapid advancements in the creation of steel, chemicals and electricity lead to mass-produced consumer goods [23]. The third industrial revolution, occurring in the 1970s, referred to the adoption of computers and automation [24]. Manufacturing is currently undergoing a fourth industrial revolution. The fourth industrial revolution (or Industry 4.0) builds on the third by enhancing the manufacturing systems to become autonomous [25]. Fuelling Industry 4.0 is the unprecedented volume of manufacturing data generated by the adoption of cyber-physical systems, smart factories and industrial internet of things (IIoT) [26]. In 2014 it was reported that 90% of all the data in the world had been generated in the preceding two years [27].

The first, second and third industrial revolutions have led to increases in economic growth, increased productivity, and advances to welfare from high-quality goods and services [28]. However, the benefits have come at the expense of the environment, as they have been powered by non-renewable energy sources and have encouraged rapid depletion of the Earth's raw materials [29]. Prisecaru states that a challenge of the fourth industrial revolution is to make the switch to renewable energy sources and to achieve sustainable resource use [30].

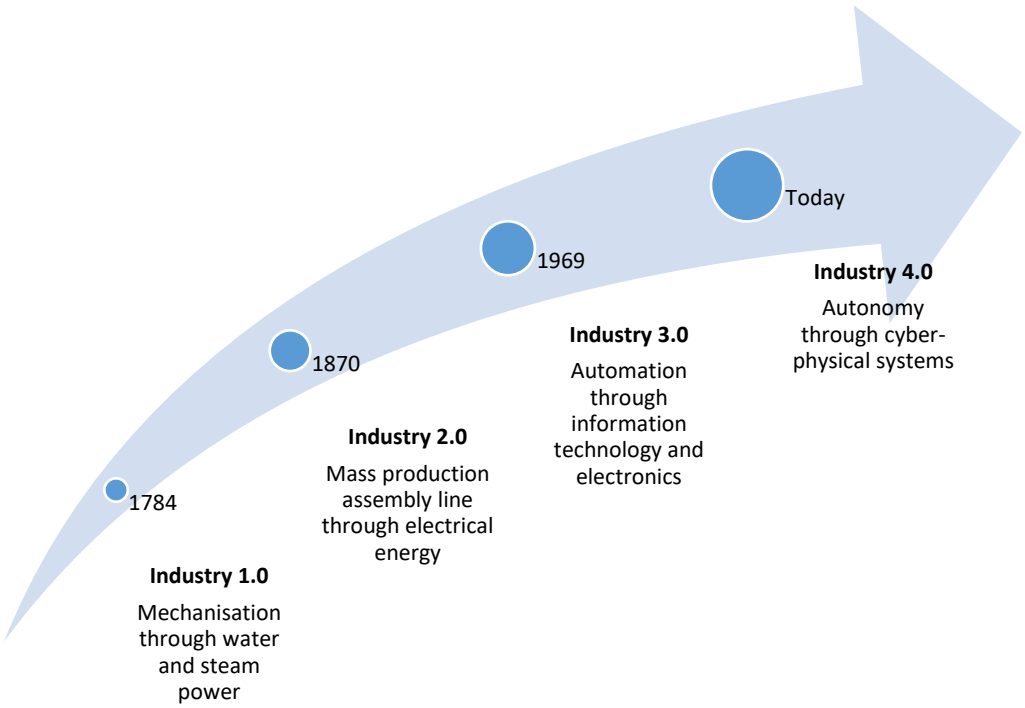


Figure 2.1. Timeline of industrial revolutions, adapted from [31]

2.2.2 THE DIFFERENCE BETWEEN PROCESS AND DISCRETE MANUFACTURING

Broadly speaking, manufacturing can be split into two branches; discrete manufacturing (automobiles, computers and electronics, textiles) and process manufacturing (chemicals, food and drink, pharmaceutical, fast-moving consumer goods - FMCG). The differences between discrete manufacturing and process manufacturing are summarised in Table 2.1. The fundamental cause of these differences is within the nature of the manufacturing process itself. In discrete manufacturing a specific quantity of material moves between different processing equipment as it is cut and assembled [2]. Whereas in process manufacturing, raw or waste materials flow through the process undergoing thermal, chemical and/or biochemical conversion. The product(s) is then separated from the process stream. There can be overlap within the two branches of manufacturing. For example, in the food industry a bar of chocolate is a distinct item or product but results from process manufacturing.

Table 2.1 Differences between discrete and process manufacturing

	Process manufacturing	**Discrete manufacturing**
Manufacturing practice	Thermal and/or biochemical/chemical conversion of raw materials into the final product and by-products, which typically are then extracted and purified.	Assembling parts and components into a final product.
Example industries	Chemicals, food and drink, paint, polymer, pharmaceutical, steel, Fast-Moving Consumer Goods (FMCG).	Automobiles, computers and electronics, furniture, textile, construction.
Quality control	Product typically must meet a specified purity and yield. During start-up or shut down products produced will not meet these specifications, and only wasteful material is produced.	Quality is assured via inspection of a statistically determined amount of products for defects. Startup and shut down is instantaneous and makes no difference to product quality.
Feedstocks	Raw materials, often in liquid/gaseous form. Variation in feedstock characteristics often present.	Discrete units/bill of materials. Feedstock characteristics are often constant.
Products	Product is typically produced as a bulk quantity item, such as chemicals, polymers, pharmaceuticals, foodstuffs, rather than distinct items. Quantified by volume/mass of product.	Distinct items, such as automobiles, toys, tools etc. Quantified by number of items produced.
Waste produced	Liquid and solid waste streams containing potentially valuable material residue, co-products, recoverable heat or water.	Defects, unused parts, cuttings etc.

Both discrete and process manufacturing share the core value that their products and processes must be profitable to remain in business. However, there is mounting pressure for businesses to become not just economically sustainable but also environmentally sustainable [32]. This has been driven by government regulations and legislation, and increasing public pressure for products where the process and life cycle better consider the resources and related ecosystems. The means and ends of achieving this goal differ between discrete and process manufacturing. Discrete manufacturers have focused on redesigning their products to use sustainable resources. Whereas, process manufacturers are often able to keep the characteristics of their products the same but redesigned their systems to use alternative sustainable feedstocks. This work develops ideas on how intelligent resource use within process manufacturing environments may be a route to achieving sustainable process manufacturing.

2.3 SUSTAINABLE PROCESS MANUFACTURING

2.3.1 THE THREE PILLARS OF SUSTAINABILITY

The idea of 'sustainability' emerged into the mainstream in the 1980s, although experts of the 17th and 18th centuries introduced the topic with regard to sustainable yield in response to dwindling forest resources across Europe [33]. Three pillars of sustainability were formalised for future sustainable development during the 2002 World Summit; these are economic, social and environment [34]. These pillars often come into conflict with one another, as seen during the 20th century where economic development for a better quality of life was favoured at the expense of the environment [29]. A Venn diagram is the typical representation of the three pillars, which is shown in Figure 2.2. The representation aims to show that sustainable development can only be achieved by balancing all the economic, social and environmental considerations. By striving to achieve sustainable development, one must endeavour to be:

- **Bearable**, developing towards the goal of a healthier environment and well-being.
- **Equitable**, aiming for a fair distribution of resources, eradicating social inequality and raising society's standard of living.
- **Viable**, achieving economic growth and development without causing harm to the environment.

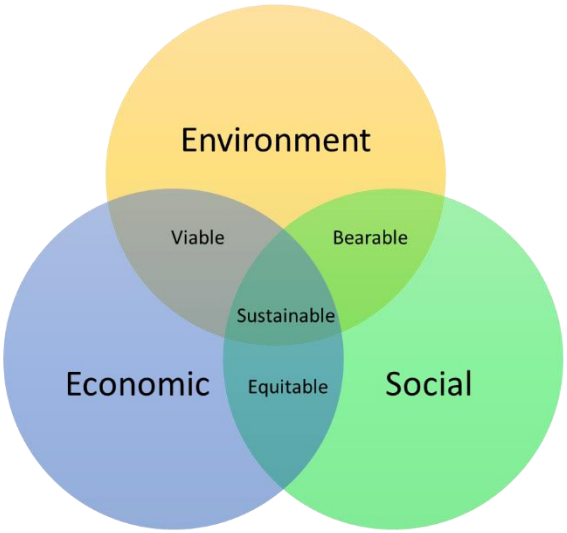


Figure 2.2. The typical Venn diagram representation of the three pillars of sustainability

Macnaghten and Jacobs claim the three pillars model of sustainability emphasises trade-offs between economic growth deteriorating environmental conditions, and a decline in the quality of life [35]. The authors argue for the nested dependencies model of sustainability [35]. In the nested dependencies model economic growth is seen as subsidiary to improving the quality of life of a society. Likewise, society development must be restricted to environmental limits.

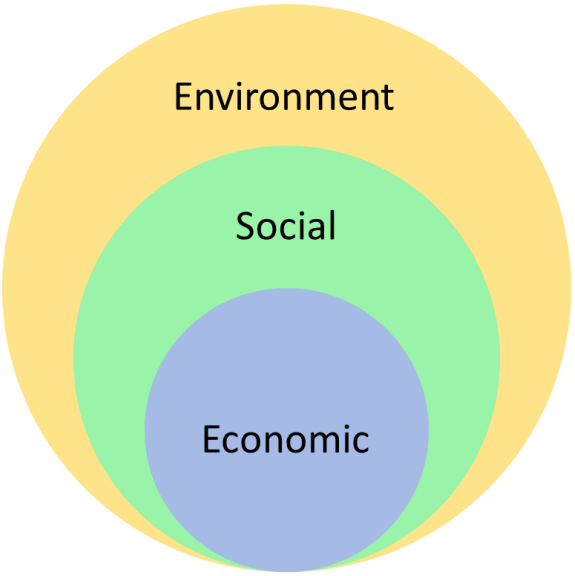


Figure 2.3. The nested dependencies model of sustainability

2.3.2 DRIVER AND BARRIERS

Manufacturers have traditionally focused on just the economic aspect of sustainable development [29], whereas external drivers are now forcing companies to adopt

environmental and societal sustainable development strategies. Five key drivers pushing manufactures towards overall sustainability have been identified as:

1. Legislation;
2. Customer demands;
3. Shortage of natural resources;
4. Population increase;
5. Business competitiveness.

Environmentally sustainable strategies were once considered the voluntary responsibility of companies. Nowadays, stricter legislation has made companies accountable and was the most influential reason manufacturing companies gave when asked what caused them to adopt environmental strategies [36]. The European Union has passed several Directives and statutes with the aim for sustainable development, with particular emphasis on environmental protection. For instance, these include Directive 2008/50/EC [37] and Directive 2008/98/EC [38] which place limits on air emissions and regulate waste disposal respectively.

Sustainability is becoming a differentiator for customers when choosing between products and was the second most popular reason given for driving companies towards environmental sustainability [36]. The UN Global Compact-Accenture CEO Study on Sustainability 2013 interviewed 1000 CEOs and found that “both the reputation and brand of their company on sustainability (81%) and the sustainability performance of products and services (77%) are important in the purchasing decisions of their consumers and customers” [39].

Fossil fuels are becoming scarcer and causing global volatility in energy prices [40]. Manufacturing is an energy-intensive process and reduction in energy consumption is one of the most important strategies for improving sustainability [29]. Other resources such as available water of appropriate quality and rare earth minerals are under pressure, and new sustainable processes are required to minimise their consumption or ideally, replace them with sustainable feedstocks. In 2017, the EU expanded its list of critical raw materials (CRM) to 27, defined as materials considered to be of high importance to the EU economy and of high risk to supply problems [5]. This recognition comes at a tipping point, where the resources required to sustain the global population exceeds those available [4].

The world’s population grew from less than 2 billion to over 6 billion during the 20th century and is expected to reach between nine and 11 billion by 2050 [41]. This is

causing an increase in consumption of water, energy, and food resources, whilst simultaneously generating more waste. These stresses are becoming more concentrated as the population continues to move towards urban areas.

Development of sustainable manufacturing processes acts as a catalyst for innovation and gives companies a competitive advantage [29,42] as well as a positive public perception. As stated, previously companies would focus on innovation and the optimisation processes at the expense of environmental concerns to gain an advantage. Legislative and regulatory drivers and the fines for breaking them have motivated the rebalance of the sustainability pillars in favour of innovation towards environmental-friendly processes. The companies that are first to redevelop processes that do not breach these regulations, as well as being more resource-efficient now gain significant economic benefits. Despite these drivers, there are still numerous barriers in the way of sustainable manufacturing [43]. These barriers can be sorted into four categories:

1. Policy and market barriers;
2. Financial and economic barriers;
3. Technical and information barriers;
4. Managerial and organisational barriers.

From these barriers, a lack of information and lack of allocated resources (time and human) are the main given responses preventing companies adopting sustainable strategies [36]. This lack of information includes areas such as environmental impact and expertise knowledge.

2.3.3 THE CIRCULAR ECONOMY

The circular economy concept emerged in response to the need for a model that would enable continued economic growth that is not limited by the finite number of natural resources [44]. As defined on Page 3, the circular economy is as an alternative to a traditional linear economy (make, use, dispose) in which we keep resources in use for as long as possible, extract the maximum value from them whilst in use, then recover and regenerate products and materials at the end of each service life (Figure 2.4) [7]. There no single origin of the circular economy concept, but the Ellen MacArthur Foundation recognise early contributions from U.S. professor John Lyle; his student William McDonough; the German chemist, Michael Braungart; and, architect and economist, Walter Stahel [45].

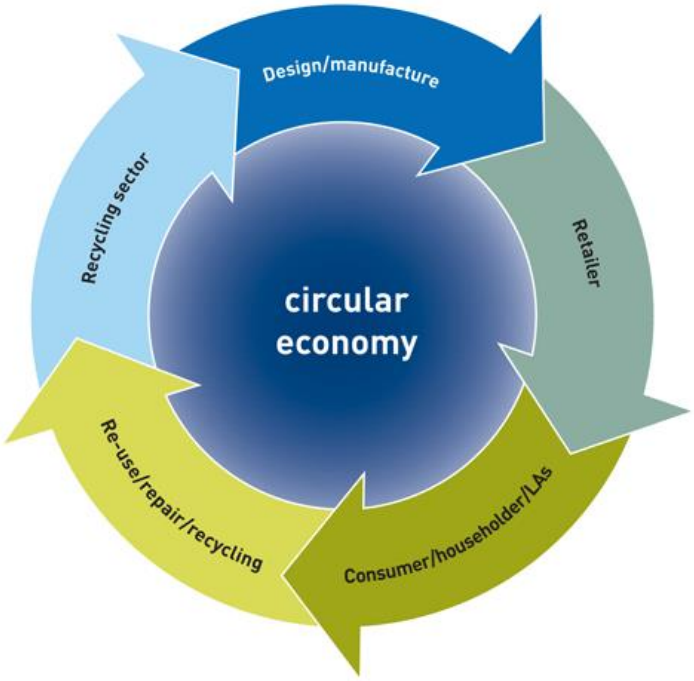


Figure 2.4. The circular economy model, as presented by WRAP [7].

The concept of a circular economy dates further back than its current incarnation. The realisation of the need to adapt manufacturing processes to become environmentally friendly was influenced by space exploration. In 1966 Boulding published an article titled “The Economics of the Coming Spaceship Earth”, which indicated that the economy should be looked at from a different perspective as the economic system at that time did not take into account the limited number of material resources [46]. In 2018, Reike *et al.* published an article detailing how there have been three distinct stages of the evolution of the circular economy concept [47]:

1. **Dealing with waste (1970-1990s)**: This period focuses on the output side, where waste is not prevented but pollution limited through principals like 'polluter pays' and ‘end-of-pipe’ treatment [48].
2. **Connection input and output in strategies for eco-efficiency (1990s-2010)**: During this period, there is a stronger integration between preventive measures and outputs measures [47]. Increasingly, environmental problems are framed as an economic opportunity so that proactive businesses can profit from efficiency gains and reputation gains [49].
3. **Maximising value retention in the age of resource depletion (2010+)**: The circular economy concept gets renewed attention, particularly from the EU, because of a fear that society cannot consume endlessly [47]. A stronger emphasis is placed on the need to retain resources within the economy as a means to combating resource depletion.

Governments and regulatory bodies attempts to encourage the adoption of circular economy concepts by industry date back to the 1990s [44]. In the 1990s, Germany and Japan were early adopters of circular economy principles, as they implemented policies to form a more closed-loop society [50]. By the late 1990s there was a recognition in China that economic growth had correlated to the amount of resource consumed [43]. The Chinese government wanted to decouple economic growth from their consumption and pollution and hence move to a more sustainable economic structure. Consequently, the circular economy concept was adopted by the Chinese Government as a main part of the national scientific development strategy [50]. This has meant that the biggest portion of research on the circular economy has come from China, with a focus on waste reduction and recycling in key industries [51]. More recently, the EU outlined a "Circular Economy Action Plan" in 2015, which aimed to boost jobs, growth and investment while protecting environment and natural capital [52]. Circular economy research is strong in Europe and is expected to increase as a result of the EU's action plan. D'Amato *et al.* found that the literature on the circular economy revolves around resource-efficiency, increasing resource productivity and decoupling resource utilisation from economic growth [53].

2.3.4 PROCESS MANUFACTURING AND THE CIRCULAR ECONOMY

The application of circular economy thinking to process manufacturing requires deeper consideration than discrete manufacturing, where it is simple to visualise a discrete product being reused, repaired and recycled (e.g. fixing a kettle). This school of thought does not lend itself to some process manufacturing products; for example, it is not possible to reuse, repair or recycle a pharmaceutical drug once consumed. However, by expanding the system to include waste produced during manufacturing and indirect waste from the consumer (in this example human waste entering wastewater treatment systems) it is possible to recapture the resource into the economy through waste recovery and valorisation.

Previous work by Li *et al.* investigated applying the circular economy concept to achieving energy conservation in China's process industries [54]. Li *et al.*'s approach to the circular economy put a higher priority on recovering energy from process manufacturing waste streams rather than material recovery, due to the process industry accounting for approximately 70% of the total annual energy consumption [54]. This approach was expanded on by Reh in 2013 to include recovering soil and water resources from residual waste (Figure 2.5) [55]. There is the possibility to recover valuable organic and inorganic resources from process manufacturers' waste streams. However, this application of circular economy encounters the biophysical limitations

of circularity such as high energy requirements of resource recovery and resource quality losses [56]. Recent research by Velenturf *et al.* broadened the scope of the circular economy so that organic and inorganic resources flows are intergated [56]. This model is more suitble for biotechnologies (e.g. bioeclectrochemical systesms, biorefineries, anaerobic digestion and thermal transformation yield by-products) that are capable of simutaneously recovering organic and inorganic resources [56].

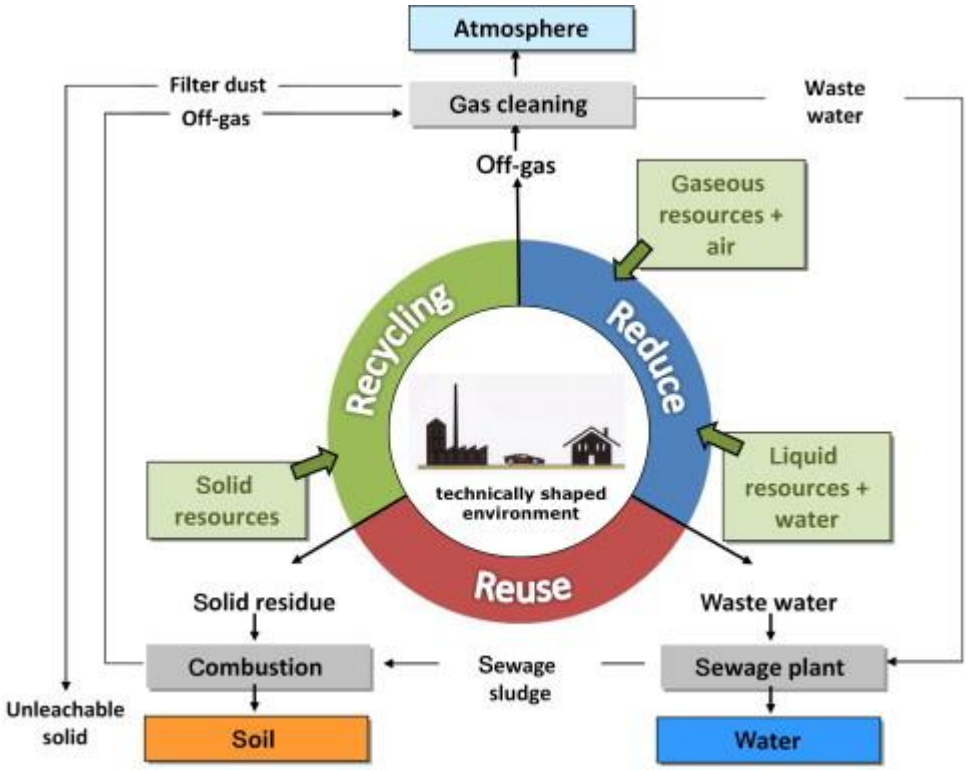


Figure 2.5. Idealised circular economy presented by Reh [55].

2.4 CLOUD MANUFACTURING

When applying the circular economy model to process manufacturing, the model is often only applied to single manufacturing systems. For example, Li outlined three case studies of the circular economy applied to three separate process industries [54]. However, this approach ignores the opportunity to increase resource recovery by sharing manufacturing resources and waste streams between multiple manufacturing systems (see Chapter 3). The current manufacturing model is not suited for this approach, as manufacturing industries are often treated as separate systems [57]. In order to create a sustainable manufacturing future, new manufacturing models need to be explored that support this circular approach.

Various advanced manufacturing models have been proposed to achieve this [58–61]. A concurrent concept across these models is the idea of an increase in collaboration,

automation, and knowledge and data sharing throughout the supply chain. As a result, the manufacturing process benefits from an increase in flexibility, resource efficiency, and customisation. Recently a new manufacturing paradigm called cloud manufacturing (CM) has emerged, which serves to act as a platform to deliver this vision [57]. Underlying the CM model is the concept of sharing manufacturing capabilities and resources on a cloud platform capable of making intelligent decisions to provide the most sustainable and robust manufacturing route available. Limitations in computer processing power, data collection and analytics, and security solutions have hindered the deployment of previous advanced manufacturing models. Cloud manufacturing, however, has the advantage of utilising new technologies such as cloud computing, big data, and the industrial internet of things (IIoT, the industrial focus of the Internet of Things), enabling CM to overcome many of these previous limitations. Cloud manufacturing also has the potential to be a key component in the future manufacturing landscape. It complements emerging manufacturing processes, such as additive manufacturing (AM), to deliver consumer customised products via sustainable processes. Where AM utilises 3D design data to build up a component in layers by depositing material, it is more commonly known as 3D printing [62].

Several projects have investigated research themes and applications of CM. In 2009, China launched the National High-Tech Research and Development Programme to promote research related to CM [63]. Two of the earliest and most cited papers on CM came from China. Tao *et al.* established and illustrated the basic ideas of CM, including architecture, definition, characteristics and benefits [64]. Xu analysed how cloud computing was changing manufacturing and defined the difference between cloud computing alongside manufacturing and CM [65]. Under the European Seventh Framework Programme the project ManuCloud was established. The ManuCloud project objective was to investigate and develop a cloud-based infrastructure to provide better support for on-demand manufacturing supply chains in the photovoltaic, organic lighting and automotive sectors [66]. There are now companies emerging which offer cloud-based services for manufacturing. For example, Plex Systems has developed the Plex Manufacturing Cloud, which is a software as a service (SaaS) enterprise resource planner (ERP) that manages the manufacturing process [67]. The software is designed to provide managers and engineers with real time access to production data. However, the full implications of CM have yet to be fully realised in industry. Cloud manufacturing has potential to work within Industry 4.0 to exploit an on-demand access to a shared collection of manufacturing resources to form temporary, reconfigurable

supply chains with enhanced efficiency, reduced production costs, and optimal resource allocation.

2.4.1 PREVIOUS CLOUD MANUFACTURING DEFINITIONS

Cloud manufacturing's first definition emerged in 2010 [57]. However, since that publication CM has generated considerable research interest and there now exists a variety of definitions. Unfortunately, there has yet to be an agreed common definition. Even the nomenclature and abbreviations can vary from author to author. A selection of some of the first CM definitions to emerge are detailed in Table 2.2; these definitions have since been commonly cited in recent publications on CM [68–74].

Table 2.2. Cloud manufacturing definitions

Definition	**References**
"Cloud-based design and manufacturing (CBDM) refer to a product development model that enables collective open innovation and rapid product development with minimum cost through social networking and crowd-sourcing platforms coupled with shared service pools of design, manufacturing resources and components."	[75]
"Cloud manufacturing (CMfg) is a computing and service-oriented manufacturing model developed from existing advanced manufacturing models (e.g., application service providers, agile manufacturing, networked manufacturing, manufacturing grids) and enterprise information technologies under the support of cloud computing, the Internet of Things (IoT), virtualization and service-oriented technologies, and advanced computing technologies."	[64]
"Cloud-based design and manufacturing (CBDM) refers to a service-oriented product development model in which service consumers are able to configure products or services as well as reconfigure manufacturing systems through Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Hardware-as-a-Service (HaaS), and Software-as-a-Service (SaaS) in response to rapidly changing customer needs. CBDM is characterised by on demand self-service, ubiquitous access to networked data, rapid scalability, resource pooling, and virtualisation. The types of deployment models include private, public, and hybrid clouds."	[76]
"Cloud manufacturing (CM) is a customer-centric manufacturing model that exploits on-demand access to a shared collection of	[77]

diversified and distributed manufacturing resources to form temporary, reconfigurable production lines with enhanced efficiency, reduced product lifecycle costs, and allow for optimal resource loading in response to variable-demand customer generated tasks."	
"Cloud manufacturing (CM) is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable manufacturing resources (e.g., manufacturing software tools, manufacturing equipment, and manufacturing capabilities) that can be rapidly provisioned and released with minimal management effort or service provider interaction."	[65]
"A new service-oriented, highly efficient and low consumptive, knowledge-based and intelligent networked agile manufacturing model and technology, allowing manufacturing resources and capabilities to be virtualized and transformed into on-demand services available to users through a product life cycle."	[78]

2.4.2 INTRODUCING THE CLOUD MANUFACTURING ARCHITECTURE

Since CM's inception in 2010, there can be seen numerous proposed architectures for its implementation. Two similar initial proposals focused on a layer hierarchal framework for a possible structure of CM's architecture [64,65]. As research has progressed, this layer structure has been varied from three layers to twelve depending on the detail and depth the authors have undertaken [76,78–80]. Common to these proposals is the issue of identifying and virtualising manufacturing resources and capabilities. These are encapsulated in the cloud as services to be managed through a centralised intelligent platform. To provide real-time control of these manufacturing services they are connected by utilising IIoT technologies. This forms the bottom part of the layered framework. The middle section is then concerned with managing the cloud services. It is the backbone of the system and has a variety of duties to maintain and operate the cloud such as: scheduling, matching users with providers, quality control of services, remote monitoring of resources, fee calculation, evaluation, optimisation of process, etc. The higher layers of the framework are the application layers, which act as a portal through which cloud users interface with the system. Cloud users can communicate their individual needs and request various manufacturing cloud resources. A crucial feature of all the proposed frameworks is that knowledge is shared across the entire system and cloud security is constant and robust. A simplified, concise model of common features is presented in Figure 2.6.

To understand how the CM architecture would be implemented, it is important to define the cloud users and their roles. There are commonly three defined roles, although the exact nomenclatural of each role varies. These are the cloud consumer, cloud provider and cloud operator [64,65,75,78]. Cloud consumers and cloud providers have obvious roles; one is utilising the cloud services for manufacturing purposes, and one is providing these services. The cloud operator manages the use, performance, and delivery of cloud services as well as negotiating the relationship between cloud providers and cloud consumers. This role is sometimes split to involve a fourth cloud user known as a cloud carrier, which acts as the intermediary that provides connectivity and transport to enable exchange of services between consumers and providers [81]. Although it has been noted that one user may perform two or more of these roles [75], the implications of this on the possible manufacturing clouds and their business models has only recently been explored [72,82]. Guo developed four possible models from different combinations of relationships between these users and the roles within the manufacturing cloud [72]. It was noted that the manufacturing cloud system design would be directly affected depending on the desired relationship between cloud users.

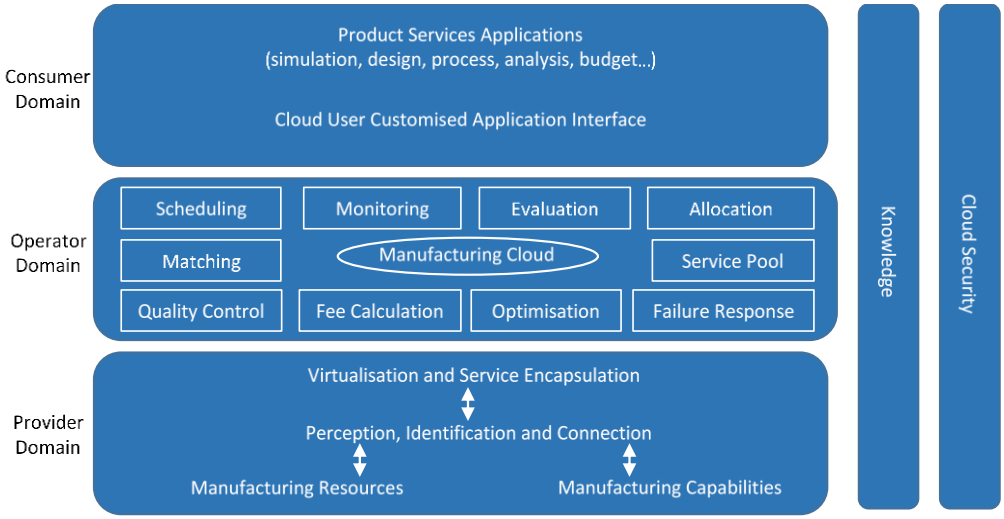


Figure 2.6. Cloud manufacturing architecture, adapted from [64,65]

The current literature on CM's architecture broadly agrees on an overview for the framework and has instead moved focus onto how the system may be implemented. Part of that involves fully defining each layer and how it functions. Recent research has tried to achieve this by first defining an architecture, containing similar elements to those previously proposed, before breaking down each layer individually into flow charts to demonstrate how they would operate in practice [79].

The mechanism through which users shall interface depends on the manufacturing cloud deployment model. The manufacturing cloud has four potential deployment

models: private, public, community and hybrid [65,76,83]. A private cloud would contain a single tenant's data and information and would provide services specific within that enterprise. By being private within an enterprise it allows the security of data, with the exception of hacking. Multiple tenants who have a common interest may share certain knowledge or resources for mutual gain through the community cloud. On the public cloud manufacturing resources and capabilities and data is shared and managed by a third party and is the platform by which tenants from various backgrounds shall be able to interact with each other. Hybrid clouds consist of multiple external and internal clouds. An enterprise may have to utilise all four versions depending on the tasks.

Early research focused on single, isolated manufacturing clouds; however, for CM to be successful, it will require the integration of multiple manufacturing clouds [84]. A hybrid manufacturing cloud allows enterprises to create multiple cloud models (public, private and community) depending on their business goals [85]. It provides a flexible and open environment that breaks down the barrier between organisations to allow for sharing of resources between enterprises, but it does not present a comprehensive method to integrate separate enterprises' manufacturing clouds. A hybrid integration framework for multiple manufacturing clouds was developed to enable users to find the most appropriate services across multiple manufacturing clouds, through a third party CM platform [70]. Similarly, a further system model has been considered where multiple 'cloud terminal subsystems' are connected to a 'cloud platform subsystem' [72]. The cloud terminal system is based on the enterprise information system. Industrial Internet of Things technologies collect information of enterpriser's information elements and provide an interconnection data interface with the cloud platform. Furthermore, it is responsible for integrating the manufacturing resources and manufacturing capabilities of the enterprise into the cloud platform for remote access and control. It would act as an enterprise's individual manufacturing cloud. The cloud platform subsystem acts as a big data management centre responsible for the convergence, integration, distribution, monitoring and operation management of the various manufacturing cloud services provided by the connected enterprises. It also provides the human-computer interface. By splitting the CM system into various cloud terminals and a cloud platform, Guo helps to demonstrate how multiple enterprises may possibly operate within the CM architecture [72].

A key feature of the CM architecture is the virtualisation of manufacturing resources and capabilities [84]. The aim of virtualisation is so that the manufacturing systems can be modelled on the manufacturing cloud. Cloud manufacturing stands apart from other

recent manufacturing paradigms, because of its ability to access data from across the supply chain to make predictions and decisions that would otherwise be overlooked [3]. To interpret this data intelligently, mathematical models built into the cloud platform are key. Therefore, the first step in creating a working CM system is the development of these models.

2.5 MATHEMATICAL MODELS

2.5.1 INTRODUCTION TO MATHEMATICAL MODELLING

Mathematical modelling originated in 30,000 BC with the counting and writing of numbers [86]. The first recorded developed models were to the fields of Architecture and Astronomy in 2,000 BC [86]. Primarily, mathematical modelling is the utilisation of analogies to help understand the behaviour of a complex system [87]. Models are used to translate our views of how the world functions into the language of mathematics. Their ability to solve complex problems increased vastly with Turing's work to develop a machine capable of simulating an algorithm's logic [88].

Mathematical modelling is utilised for four primary reasons [86]:

1. Develop scientific understanding through quantitate expression of current knowledge of a system;
2. Make predictions about how change may affect a system;
3. Aid in decision making;
4. To communicate knowledge;

Mathematical modelling is fundamental to process manufacturing [89], which utilises analogies to help understand the behaviour of a complex system [87]. A model can be used to aid in decision-making, develop scientific understanding, communicate knowledge and/or make predictions [86]. Process manufacturers have utilised mathematical modelling for four primary applications [90]:

1. Planning and design;
2. Monitoring and control;
3. Process optimisation;
4. Risk mitigation.

2.5.2 FIRST PRINCIPLE MODELS VS EMPIRICAL MATHEMATICAL MODELS

Although mathematical modelling has been considered to have existed since 2,000 BC, the turning point for process engineers utilisation of models was not until the early 20th Century with Prandtl's model on boundary layer theory [91] [92]. With Prandtl's model

followed the development of first principles models (sometimes referred to as mechanistic models) for heat and mass transfer [92]. With the application of computer simulations to process manufacturing, these calculations were first converted from manual to digital in the 1950s [93]. As personal computers became commonplace in manufacturing during the 1980s, this enabled greater empirical modelling of process manufacturing systems for optimisation [93]. Empirical modelling refers to modelling based on correlations from observational data [94].

Within the field of mathematical modelling, there are two distinct branches: first principles modelling (often referred to as mechanistic models) and empirical modelling. First principle models build a series of equations by examining the workings of the system's individual parts (Schichl, 2004). First principle models rely on system understanding to compensate for lack of data. Because of this, they have a greater potential for extrapolation, compared to empirical models which are better suited for interpolation [95]. Empirical models are mathematical equations derived from the analysis of data; therefore, requiring less knowledge of the system [96]. Empirical models rely on the assumption that the data is of sufficient granularity and/or quantity to define the system. The development of first principles models are time consuming and require detailed knowledge of the process [97]. When attempting to model a new or complex process, where little is understood of the system, empirical models can be utilised independently or in support of first principle models [98].

2.5.3 DATA-DRIVEN MODELS

Empirical models are already well utilised in process manufacturing because of the volume of data manufacturers produce [94]. Recently the capabilities of empirical models have greatly expanded due to advances in the fields of computational intelligence and machine learning; these new approaches are encompassed into the field of data-driven modelling [96]. Computational intelligence refers to nature-inspired computation approaches to problem-solving; for example, algorithms that mimic the behaviour of animals (swarm intelligence) [99] or algorithms that replicate the behaviour of how humans solve problems (artificial neural network - ANN) [100]. Machine learning focuses on the development of algorithms that can access data and use it to learn for themselves [101]. Example algorithms include support vector machines, ANN and linear regression. There is an overlap between the two fields; for example, ANN. However, there are computational intelligence algorithms (e.g. rule-based inference) [102] that are not considered suitable for machine learning.

Data-driven models (DDMs) are able to find relationships between the system state variables (input and output) without prior knowledge of the system [103]. Data-driven models derive the system's relationships by training an algorithm on manufacturing data [100]. Although data-driven algorithms have long existed [104], their use in industry has been limited because of data constraints (e.g. lack of data, data not stored in useable format) and lack of computational processing power [105]. There is an unprecedented rise in the volume of manufacturing data being generated due to the adoption of cyber-physical systems, smart factories and the IIoT [26]. In 2015, it was reported that manufacturers globally generated more than 1000 Exabyte of data, and by 2025 data generation in manufacturing will increase 20-fold [106]. This means that the volume of data available to build DDMs has never been greater. Data-driven models were not always considered suitable for enterprise-wide modelling due to the computational cost in modelling vast volumes of data [107]. However, with the introduction of cloud computing, manufacturers now have affordable access to the processing power required to model large data sets [105]. Subsequently, DDMs are becoming prevalent across industry for modelling and monitoring of plant-wide industrial processes [105].

Data-driven models have a great potential for process optimisation by being able to model the impact of process control variables on the manufacturing efficiency. Previous work has demonstrated DDMs' potential to: increase product yields of pharmaceutical processes [26], reduce resource consumption in clean-in-place systems[108], enhance oil recovery from crude oil reservoirs [109], optimise yield of a continuous flow reactor [110], maximise the harvest of active components of plants [111], develop soft-sensors for industrial processes [112], predict corrosion in pipelines [113] and control wastewater treatment processes [114]. There are several guides in the literature for developing DDMs of process manufacturing systems. Curteanu *et al.* studied three different DDM strategies modelling of electrolysis process in wastewater treatment [115]. In this work, the authors demonstrated how a DDM's prediction capabilities may be improved by developing multiple DDMs of a system stacked together [115]. Charte *et al.* [116] and Kashani et Shahhosseini [117] both independently explained the benefits of effective data partitioning to find the best configuration of a DDM. Kashani et Shahhosseini also demonstrated that when applying dimensionality reduction techniques to datasets with a large number of input variables it increases the performance of DDMs [117]. What is often overlooked in the literature is the importance of evaluating the DDM's prediction capabilities on unseen data [118]. Unseen data is data that has not been used during the development of the DDM. A

frequent flaw when building a DDM is to put too much emphasis on optimising a model's performance with the model development data [118]. This lead to overfitting, where a model too closely matches the model development data and captures relationships not present in the manufacturing system [100]. Hamid *et al.* showed how additional experiments may be performed to generate unseen data to evaluate a model extrapolation capabilities beyond the model development data [119]. However, when modelling historical manufacturing data this may not be possible. A method is required to partition unseen data from the historical data that can be used to thoroughly evaluate the DDM's model predictive capabilities throughout and beyond the system.

2.5.3.1 Multiple target data-driven models

Modelling a manufacturing system can aid decision making to identify the most sustainable manufacturing route that may have previously been unknowable [3]. For a manufacturing system to be considered truly sustainable it needs to maximise environmental and social benefits, in addition to economic benefits [120]. This is particularly important for process manufacturing, which produces a variety of outputs with different sustainability implications. To find the most sustainable manufacturing route requires the development of models that can predict multiple targets so that the relationships between a system's inputs-outputs and outputs-outputs can be established and exploited.

The majority of published research utilising DDMs within process manufacturing systems are single target (ST) models. A ST model is trained to predict the value of a single, categorical or numerical, target variable of a manufacturing system [100]. Recent uses of ST models in process manufacturing show their capability to model: adsorption [119], clean-in-place systems [108], fluid mechanics [121], heat transfer [122], mass transport [123], reaction networks [124] and synthesise design [101]. However, when modelling a manufacturing system with the aim to increase sustainability, multiple outputs are required for the model to identify the most economical and environmental solution. There are examples of multiple independent ST models predicting multiple outputs of a process manufacturing system. For example, Lesnik and Liu modelled the power density, chemical oxygen demand (COD) removal and Coulombic efficiency of a microbial fuel cell treating wastewater [125]. The authors achieved this by developing three ST models, rather than a single model predicting all three outputs [125]. Yet, building multiple ST models of a system fails to capture the relationship between the multiple outputs and are computationally expensive [126]. An alternative approach is to build a single multiple target (MT) model that captures the relationships between the output variables and are proven to use less

computational power [127]. Multiple target models are capable of predicting multiple outputs from a set of input data [127]. The development of MT models has proven successful when applied to a wide range of practices: bioinformatics [128], chemometrics [129], ecology [130], gene function prediction [131], natural language processing [132] and stock price forecasting [133]. There are a limited number of examples of MT models built for process manufacturing systems [115,134]. However, the potential of MT models has yet to be realised by the wider process manufacturing community.

There are two main approaches for using machine learning algorithms as a base method for MT learning: *problem transformation* and *algorithm adaptation* [127]. Problem transformation is the transformation of a MT model into multiple ST models, and each is solved separately using traditional machine learning procedures [127]. Whereas, algorithm adaptation methods adapt machine learning algorithms to predict all outputs simultaneously in one model [127]. Algorithm adaption methods have been shown to perform better than problem transformation methods [126,131]. This is accredited to their ability to capture not only the relationships between a model's inputs and outputs but also between multiple outputs [126]. A feature not possible in problem transformation models, because single independent models are trained for each output [127]. This work compares the capability of both MT modelling approaches to make predictions for not only the model development data but also unseen data, something not yet tested in the literature. Unseen data is data that has not been used for the training, validation or testing of the model and is used to thoroughly evaluate the final model's predictive capabilities on new data [119].

Once built, a MT model can then be used to evaluate different manufacturing routes to find the most sustainable solution. This sometimes proves challenging, as, within process manufacturing, the economic and environmental objectives are often competing [110]. There are two approaches when utilising a DDM model to ascertain the most sustainable manufacturing configuration, data exploration and self-optimisation. The first approach involves using the developed DDM to simulate output results from new synthetic input data. Variables outside of the process manufacturer's control are set at one value for each data point, whilst the process conditions are assigned values to investigate each variables' effect on the system's outputs. The process conditions values may be determined with design of experiments (DoE) techniques to systematically evaluate each process conditions effect on the outputs [119]. For the second approach, the DDM model is used within a closed-loop optimisation system to automatically determine the optimal process conditions that maximise the model outputs [110]. This

can be done either by fitting the DDM's predictions to a function that may be used to find a single optimal point of the system or by establishing the Pareto front for each process condition [110]. A Pareto front is a state from which it is impossible to optimise the process conditions further to increase a performance objective without making at least one performance objective worse off [135]. For example, Schweidtmann *et al.* were able to utilise a MT model for the self-optimisation of two chemical reactions performed in continuous flow, which determined the Pareto fronts of the four reaction conditions [110]. The use of MT models for multiple objective optimisation shows great potential for finding a solution that maximises both economic and environmental objectives.

2.5.4 MACHINE LEARNING

Data-driven models are derived by training, validating and testing artificial intelligence (AI) algorithms on a set of data, referred to as 'model development data', then evaluating the model's prediction capability on 'unseen data' [100]. The DDMs built for this work have been developed using machine learning algorithms. Machine learning is a subbranch of AI that teaches computers to do what comes naturally to humans and animals: learn from experience [136]. Machine learning is the technique that figures out the "model" out of the "data" [100]. The learning part of machine learning refers to how the process is being trained on data to find the model. There are three types of machine learning:

1. **Supervised learning**: Supervised learning takes a known set of input data and known responses to the data and seeks to build a predictor model that generates reasonable predictions for the response to new data [137].
2. **Unsupervised learning**: Unsupervised learning requires input data and no corresponding responses. The goal of unsupervised learning is to model the underlying structure or distribution in the data to learn more about the data [137].
3. **Reinforcement learning**: Reinforcement learning is concerned with the problem of finding suitable actions to take in a given situation in order to maximise a reward. Here the algorithm is not given examples of optimal outputs but must instead discover them by process of trial and error [137].

During this work, both supervised and unsupervised learning models have been developed. There are two types of supervisor learning, *classification* and *regression*. Classification models are built to produce responses that have just a few known values, such as 'true' or 'false'. Classification algorithms apply to nominal, not ordinal response

values [138]. An application of classification models to process manufacturing is for fault detection [139]. A process manufacturing fault is typically an incorrect step change to a process condition that causes the system to become unstable [139]. Kulkarni *et al.* demonstrated how support vector machines could be used to classify multiple faults in process manufacturing systems. Regression models are built when a model's response is a real number, such as miles per gallon for a particular car [140]. Regression models are widely used in process manufacturing for the prediction of continuous data. For example, predicting the removal of pollutants by novel adsorption process in the wastewater treatment sector [119].

During the development of DDMs, there are fundamental considerations around the collection and modelling of data to prevent poor generalisation and overfitting. Unless properly addressed, these are the two main sources of error in DDMs [100] and are defined as:

- *Generalisation*: is the capability of a DDM to fit and make predictions of data that was not used during the development of the model [100]. The challenge of making the model perform consistently between data used to develop the model and new input data is known as generalisation.
- *Overfitting*: is the generation of a model that corresponds too closely or exactly to a particular set of data, and may, therefore, fail to fit additional data or predict future observations reliably [141].

When developing a DDM, there are a number of machine learning algorithms to select from. The "No Free Lunch" theorem states all algorithms perform equally well when averaged over all possible datasets [142]. In deciding which algorithm to use there are two schools of thought. Either explore the literature for algorithms developed for similar manufacturing systems or start with the simplest algorithm and if that does not meet the requirements try a more complex one [142]. A TITLE-ABSTRACT-KEYWORD search of scientific documents in Scopus, related to the field of engineering and chemical engineering of machine learning algorithms was conducted. The results in Figure 2.7 show that ANN and support vector regression (SVR) have been the most published algorithms since the year 2000. Work published on deep learning algorithms, which are multiple layered ANNs with automatic feature learning [143], has increased at the fastest rate since 2014. The spike in ANN published works may be attributed to the rise in deep learning, as deep learning models are built using ANNs. Deep learning models have excellent predictive capabilities and are a fundamental part of smart manufacturing and Industry 4.0 [143]. However, they are not suitable for this work

because of their reliance on large datasets (1,000 + data points) [143]. Linear regression models were the fourth most common algorithm discussed in the published literature. Linear regression models and multiple linear regression, find linear relationships between model input(s) and a model output [137]. They are well used in industry due to their simplicity to build [144]. However, the application of linear regression to modelling process manufacturing systems is limited because of the nonlinearity of these systems [145]. Therefore, linear regression models do not feature as part of this work. The machine learning algorithms featured in this work are Gaussian process (GP), random forest (RF), SVR and ANN.

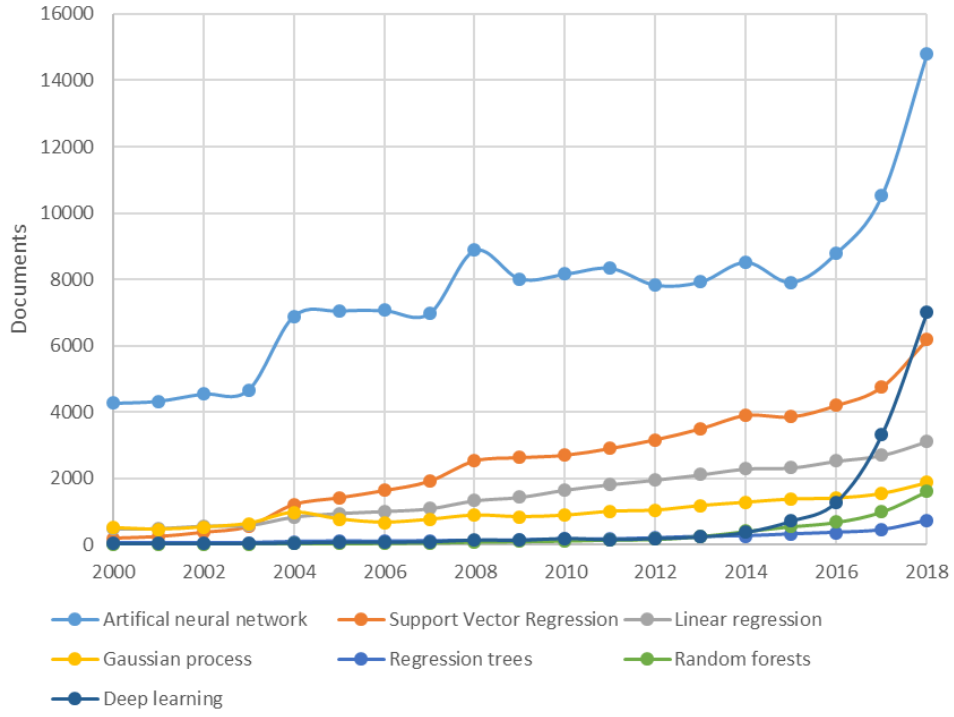


Figure 2.7 Documents by year across in the fields of engineering and chemical engineering. Data were taken from a Scopus search.

When developing a DDM, each algorithm has model parameters and hyperparameters that require tuning to find the optimal configuration. Traditionally, hyperparameters are parameters that are set before being trained on a model, compared to model parameters that are learnt during training [100]. For example, the number of neurons in an ANN's hidden layer is a hyperparameter; whereas, the weights between neurons is a model parameter. Validation data is used to tune the hyperparameters and training data is used to tune the model parameters. There are many different methods to optimise an algorithms hyperparameters, simple approaches include a grid or random search and more complex methods use RF and GP [146]. Bayesian optimisation is thought to be the most commonly used model for optimisation, as it allows exploration and

exploitation to be naturally balanced during the search [146]. The Bayesian optimiser works by constructing a conditional probabilistic model that can be updated with new evidence [146]. Simply put, the optimiser picks a prior belief about how the hyperparameters will behave, then searches the hyperparameter space by updating the prior belief based on ongoing measurements.

2.5.3.1 Gaussian process (GP)

Gaussian process extends multivariate Gaussian distributions to infinite dimensionality [147]. Gaussian process regression calculates the probability distribution over all functions that fit the data [147]. A prior probability distribution is specified, which is a range of possible distributions that might describe the system [148]. The posterior is then calculated using the model development data. The posterior is the refined distributions that fit model development data [148]. The difference between the prior and posterior is illustrated in Figure 2.8, taken from Schulz *et al.* article titled "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions" [148].

Within this GP prior, prior knowledge about functions can be incorporated through the selection of the mean and covariance functions [148]. The form of the mean function and covariance kernel function in the GP prior is chosen and tuned during model selection. The mean function is typically constant, either zero or the mean of the training dataset [148]. There are various covariance kernel functions. Some common kernel functions include constant, linear, square exponential and Matern kernel, as well as a composition of multiple kernels [148].

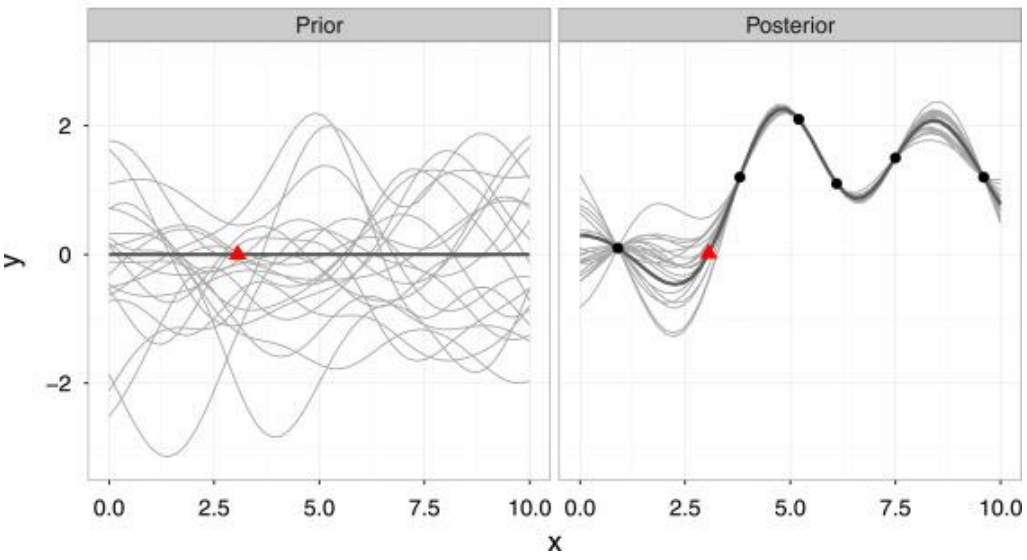


Figure 2.8. Samples from a Gaussian process prior and posterior taken from Schulz [148] . Grey lines indicate samples from the GP. Black dots mark empirical observations. The dark grey line marks the current mean of the GP. The red triangle shows the prediction for the new input point.

Gaussian process's advantage over other algorithms, is because it is a probabilistic fit. Therefore, GP models provide a belief range, including upper and lower bounds, which provides a reference to judge whether the predicted value is useful or not [147].

2.5.3.2 Random forest (RF)

Regression trees sub-divide the data into smaller datasets and further partition the sub-divisions and assigns to its nodes (leaves) [149]. Regression trees regress decisions in a tree form, starting from the root node (model inputs) down to a leaf node where the leaf node contains the response (model output) [149]. Random forest is an ensemble technique that uses multiple regression trees and a statistical technique called 'bagging' [150]. Rather than just averaging predictions from multiple trees, an RF instead randomly samples from the training data for each tree and randomly subsets the input variables at splitting nodes until the minimum node size is reached [150]. This means not all input variables are used in each individual regression tree and the regression trees are of varying sizes (Figure 2.9) [151].

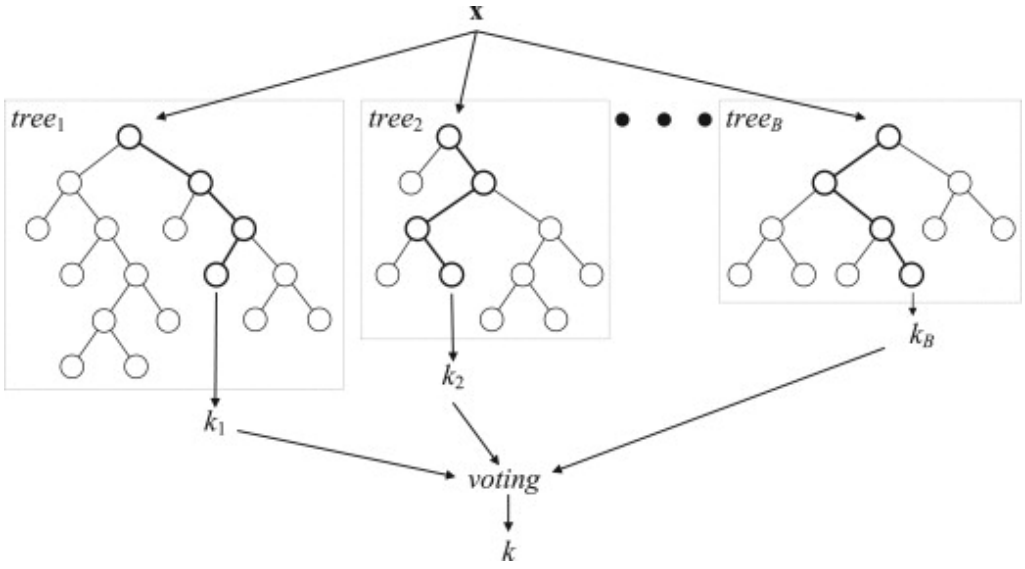


Figure 2.9. Architecture of a random forest, taken from [151].

Random forest has shown greater performance than SVR and ANN in certain applications (e.g. the pharmaceutical industry [152]); however, their use in process manufacturing has not been as widespread compared to other algorithms. Random forests ease of tuning offers an advantage for dealing with multi-dimensional complex data [150].

2.5.3.3 Support vector regression (SVR)

Support vector regression is an adaptation of the classification machine learning algorithm, support vector machines [153]. A SVR solves a non-linear problem by transforming the non-linearity between features and target using linear mapping (kernels) [149]. The goal is to find a function f(x) that deviates from the real value (y_n) by a value no greater than ε for each training point (x), and at the same time is as flat as possible, shown in Figure 2.10 [154].

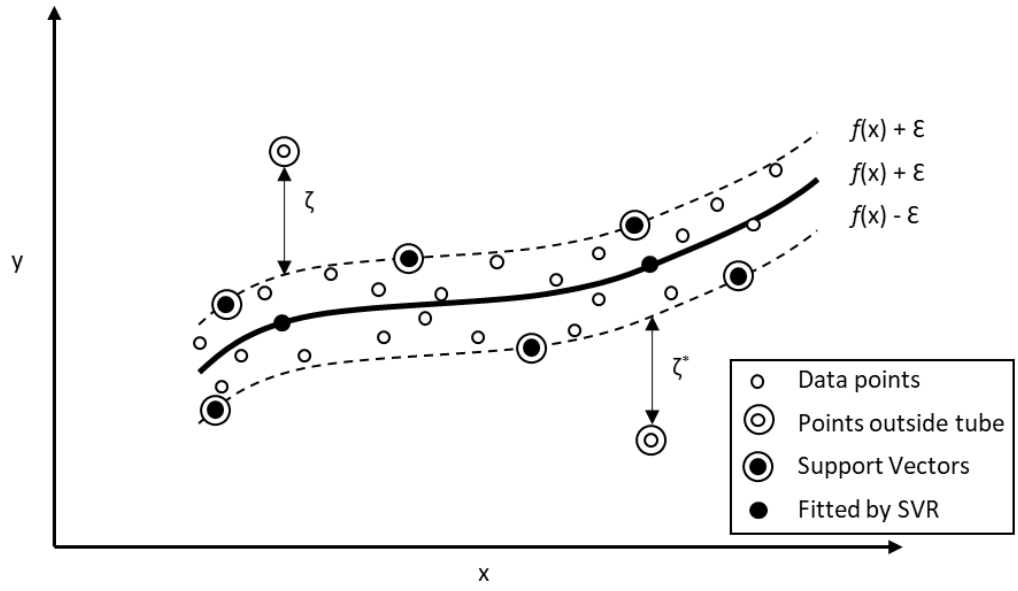


Figure 2.10. A schematic diagram of the support vector regression, taken from [154]

They are highly effective machine learning algorithms, even when presented with small quantities of data [149]. They have been applied to process manufacturing scenarios for various applications (e.g. predicting natural gas hydrate formation temperature [155], compact heat exchanger design [156], rheometry [157]). In comparing ANN and SVR, Rabbani *et al.* found SVR to be the superior model both in interpolations as well as extrapolations when trying to predict the dynamic yield stress [157].

2.5.3.4 Artificial neural networks (ANN)

Artificial neural networks are mathematical models that have been inspired by the connections made between neurons in a biological brain [158]. The neural network is composed of nodes (which represent the neurons) and the connection between these nodes are given weight values to mimic how the brain alters the association between neurons [100]. Each node has a bias which is a factor associated with the storage of information [100]. The information of the neural network is stored by the values given to weights and bias during training of the networks. An illustration of the connections around a single node is shown in Figure 2.11.

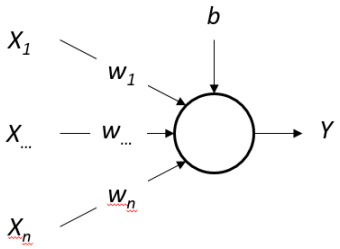


Figure 2.11. A node that receives n inputs

Within an ANN architecture nodes are connected in a series of layers. In early ANN research only networks with a simple input and output layers were used [159]. These were referred to as single-layer neural network as the nodes of the input layer merely act to transmit the input signal to the next nodes and do not calculate the weighted sum and activation function [159]. Hidden layers can be acted between the input and output layers. They are given this name as they are not accessible from the outside of the neural network [100]. Networks that contain one or more hidden layer(s) are called multi-layer neural network. A network with one hidden layer is referred to as a shallow neural network, and those with two or more hidden layers are referred to as deep neural networks, Figure 2.12.

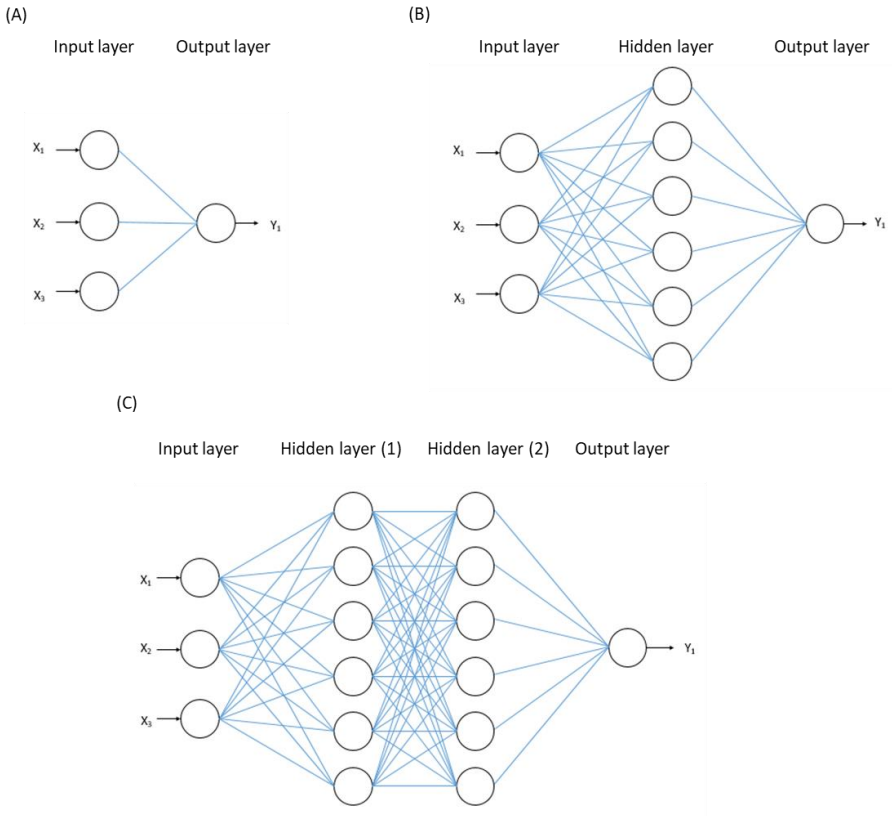


Figure 2.12. Artificial neural networks architectures (A) single layer neural network, (B) shallow neural network and (C) deep neural network

The topology of the network can also vary dependant on the direction the signal travels between the layers. There are two classifications, which are feed-forward and feedback. Feedforward allow signals to travel one way only, from input to output [160]. These networks are extensively used in pattern recognition [160]. Feedforward neural networks are ideally suitable for modelling relationships between a set of predictor or input variables and one or more response or output variables. Feedback networks can have signals travelling in both directions by introducing loops in the network [160]. Feedback networks are dynamic; their 'state' is changing continuously until they reach an equilibrium point [160]. They remain at the equilibrium point until the input changes and a new equilibrium needs to be found.

2.6 CASE STUDIES DEVELOPING A CLOUD MANUFACTURING SYSTEM

2.6.1 PREVIOUS CLOUD MANUFACTURING CASE STUDIES

As the research into CM has expanded, two directions have emerged. Broadly speaking, academia have recently been focusing on the interpretation and analysis of CM, using case studies that demonstrate different applications of CM. A selection of the recent case studies from the literature is presented in Table 2.3. These have been chosen because they demonstrate CM application in various sectors, each highlighting different CM characteristics. How and why CM is used is briefly summarised, and the main positive outcomes from the CM approach are detailed. All the case studies demonstrate how CM may achieve the optimal solution, as defined by the author. However, no one case study reached an optimal solution, and further work has been suggested for each in order to reach a more complete solution.

Table 2.3 Cloud manufacturing case studies from academia

Case study summarised and sector, where given, highlighted	**Cloud manufacturing features the study highlights**	**Outcomes from CM approach**	**Future work required suggested by this paper and *Further work suggested by original authors in italics***	**Author and year**
A proposed service optimisation configuration method for manufacturing assembly lines. Sector not specified, but model would apply to any sector with assembly lines.	**Manufacturing as a service**: Servitisation of manufacturing resources. **Intelligent on-demand manufacturing**: Optimal configuration method of services based on grey relational analysis.	CM was used for proactive discovery of manufacturing resources through servitisation. Method finds satisfying solutions with low cost and time consumption.	Method does not consider environmental impacts when determining optimal solution. *Real-time manufacturing information processing mechanism between bottom-level manufacturing machines and manufacturing cells. Batch volume of manufacturing tasks should be taken into account in the optimal selection of service composition*.	2016, [69]
A model to assess the sustainable manufacturing capability (SMC) of industrial robots (IR), used in discrete manufacturing.	**Decision making tool**: Sustainability of different manufacturing routes analysed.	Cloud platform was able to predict energy consumption of IR for assessing their capability to perform a manufacturing task.	Limited to predicting the energy consumption of a manufacturing process only to the energy consumed by the IRs. Could be expanded to include other energy outputs, such as transport of materials. *More factors should be considered to comprehensively describe the SMC of IRs, such as collaboration safety with human.*	2017, [161]

An adaptive artificial bee colony algorithm is used for decision making on the cloud service selection and composition.	**Flexibility and scalability, and decision making tool**: Feasibility of different manufacturing routes analysed.	Model includes the geographical location or services when quantifying the quality of service.	A limitation is the long time required by complex models. Extra computing power can be accessed through CM. Further work should investigate balancing cost of increased computational power with time saved. *Integration of artificial neural networks for the definition of maximum cycle number*	2015, [162]
A cloud platform is established for the polymer industry for optimising the processes of blending polymer materials.	**Decision making tool**: Data mining for optimal solution. Computer-aided process design based on CM. **Flexibility and scalability:** Flexible CM system built that responds to customers' demands.	The knowledge-based intelligent searching algorithm finished in 10 minutes compared to the 5 hours it took by traditional methods.	Case study only explored how CM can be applied to the design phase. Further work should investigate how effectively the model would respond to different customer demands during production phase. N/A	2016, [71]
Public cloud manufacturing system for SMEs in manufacturing equipment industry in South China.	**Multi-tenancy and intelligent on-demand manufacturing**: Performs intelligent multidimensional matching of consumers to providers.	Enterprises can respond to market requirements more quickly and cost effectively than before. Crowdsourcing model demonstrates how CM can	Exploring how enterprises' private clouds will integrate with public cloud. *To apply CM to more manufacturing applications.* *To investigate business models for CM.*	2015, [79]

		be used to develop innovative designs.		
A CM architecture is developed to optimise the production logistics of a paint making company.	**Manufacturing as a service:** Real-time monitoring of resources using IIoT. **Decision making tool:** Enterprise resource planning	The cloud coordinates an improved schedule resulting in logistics waiting costs and overtime working reductions of approximately 50%.	Resources monitored are all manually controlled. Therefore lacking real-time execution of management system. *Expand qualitative solution presented in paper to the quantitative stage to achieve more accurate application results.*	2016, [163]
A CM system is constructed within a conglomerate which designs and manufactures multi-disciplinary complex products in the discrete manufacturing sector.	**Manufacturing as a service and multi-tenancy:** Sharing of manufacturing recourses and capabilities through a hybrid framework integrating multiple manufacturing clouds.	Cloud manufacturing links enterprises so institutions learn to employ core competences of other institutions. This allows them to focus on their own core competences.	Proposed system does not consider what information should be private between different institutions or intellectual property issues when designing new products. *Domain-specific demands and characteristics of business collaboration to design efficient collaborative algorithms for any industry. Service-oriented architecture based integration methods of manufacturing resources and capabilities.*	2016, [70]

2.6.2 INDUSTRY LEAD RESEARCH INTO CLOUD MANUFACTURING

The focus in industry has been on developing a CM architecture, as seen from examples from the industry leaders. Siemens has developed a cloud platform called 'MindSphere' that interlinks physical products and production facilities with digital data, enabling innovative solutions and getting products onto the market faster, better and more efficiently [164]. Similarly, Plex Manufacturing Cloud delivers production control, inventory management, supply chain management, and more [67]. IBM has the IBM Watson IoT, which mines the influx of data from an IIoT manufacturing process to make businesses more cognitive [165]. Industry has directed some applications of CM, which is utilising manufacturing clouds to monitor and manage IIoT manufacturing systems within an enterprise. These examples deliver partially enabling technologies but are not true CM systems. Cloud manufacturing has further capabilities made possible from sharing of manufacturing resources and capabilities which enable innovative, sustainable solutions. In order to demonstrate CM full potential for process manufacturers, a case study with an industrial SME partner was conducted. The case study studied is a bioprocess plant built from bioelectrochemical systems technology.

2.7 BIOELECTROCHEMICAL SYSTEMS

2.7.1 INTRODUCING BIOELECTROCHEMICAL SYSTEMS

Globally, water scarcity is set to become one of the biggest challenges facing the 21st century. The World Economic Forum lists water crisis as the largest global risk in terms of potential impact [166]. One part of the solution is to become more effective at treating wastewater so it is fit for reuse. Treating wastewater consumes substantial amounts of energy. In developing countries, where water scarcity is currently a life or death issue, up to 90% of wastewater was discharged to rivers, oceans, lakes untreated [166]. If instead this could be treated and recovered, millions more people could have access to clean, safe water. How this is to be achieved is to look beyond the conventional treatment methods, that are both costly to build and energy intensive to run, and instead look at technologies which treat wastewater as a renewable resource.

Manufacturers are also facing the pressure of increasing operational costs. Effective water management is key to tackling rising manufacturing costs. This is especially true for water-intensive industries such as the food and drink industry. As population growth and climate change increase water scarcity, the development of technologies that treat wastewater efficiently, so it can be recovered and reused, is paramount. Larger manufacturers have begun utilising processes, like anaerobic digestion (AD). However, these processes have a large capital cost from an SME perspective, making them unfeasible.

In searching for the solution, research has intensified on a phenomenon that was first observed in 1911 by Potter [167]. Potter noticed that the disintegration of organic compounds by micro-organisms is accompanied by the liberation of electrical energy [167]. However, his work went largely unnoticed until 1931 when Barnet Cohen created a number of microbial half fuel cells that, when connected in series, were capable of producing over 35 volts, though only with a current of 2 milliamps [168]. In the 1960s the NASA space program attempted to apply this phenomenon for the recycling of human waste into electricity during space flight [169]. NASA's success was limited as it became apparent of the underlying complexity of the bioelectrochemical processes, and further research was needed [170]. Research into bioelectrochemical systems (BES) stagnated until the turn of the 21st century, where researchers in South Korea discovered a microbial fuel cell (MFC) milestone. In 1999 B.H. Kim et al developed the mediator-less MFC which greatly enhanced the MFC's commercial viability, by eliminating costly mediator chemicals required for electron transport [171]. Since then research into microbial based BES has grown exponentially, and is now being seen as an attractive option in the near future for sustainable wastewater treatment [172].

For BES to be seriously considered by industry, further work at the commercially sized pilot-scale is required to demonstrate that BES can operate in real-world conditions. Although there has been substantial work in the last 20 years proving BES works at laboratory scale, there have been only a handful of trials at pilot scale [173]. These trails have not been reported in papers, so a true understanding of their successes and failures is unavailable. However, secondary sources report that these trails faced two main barriers preventing a commercial viable BES plant. These are (1) affordable material and reactor design, and (2) a better understanding of the biological process [174].

Cloud manufacturing can work to support the development of new novel technologies [3]. New designs are often limited by a lack of in-house expertise [75]. The service-oriented nature of CM can work to match developers to relevant experts; whether that be material selection specialists, data analytics or developers of mathematical models for example. The models built into the cloud platform can help to gain a better understanding of the biological process, increasing the process resilience.

2.7.2 UNDERLYING PRINCIPLES OF BIOELECTROCHEMICAL SYSTEMS

Core to all BES is the transformation of chemical to electrical energy, or vice versa. They work in a similar way to a conventional fuel cell, in that they are composed of an anode and cathode separated by a semi-permeable membrane that is selective for the

passage of positively charged ions; they require the addition of a fuel to generate power [175]. However, BES differ from the conventional fuel cell in two fundamental ways:

(1) The fuel supplied is composed of organic matter and water [176];
(2) Bioelectrochemical systems utilise biocatalysts instead of precious metals [175].

There are two types of biofuel cells: direct and indirect [175].

- The direct biofuel cell is a configuration in which fuel is oxidised at the surface of the electrode. The role of the biological catalysts in this type of fuel cell is to catalyse these surface reactions.
- The indirect biofuel cell is a configuration in which fuel reacts not at the electrode, but in solution or in a separate compartment, and a redox active mediator is added to shuttle electrons between the site of reaction and the electrode.

Biological catalysts have been used in both direct and indirect configurations. Typically, enzymes have been used as catalysts in direct biofuel cells, while microorganisms have often been used in indirect biofuel cells.

2.7.2.1 Fundamental process

Outline of the bioprocess steps, with step location shown in Figure 2.13:

1. The biocatalyst on the anode decomposes organic matter and free H+ ions and electrons.
2. The electrons flow from the biocatalyst to the anode, sometimes assisted by a mediator molecule.
3. The electrons flow up from the anode, through a wire, and onto the cathode. While flowing through the wire, an electrical current is generated that can be used to perform work.
4. The H+ ions flow through the semi-permeable membrane to the cathode. This process is driven by the electrochemical gradient resulting from the high concentration of H+ ions near the anode.
5. The electrons from the cathode combine with dissolved oxygen and the H+ ions to form pure H_2O.

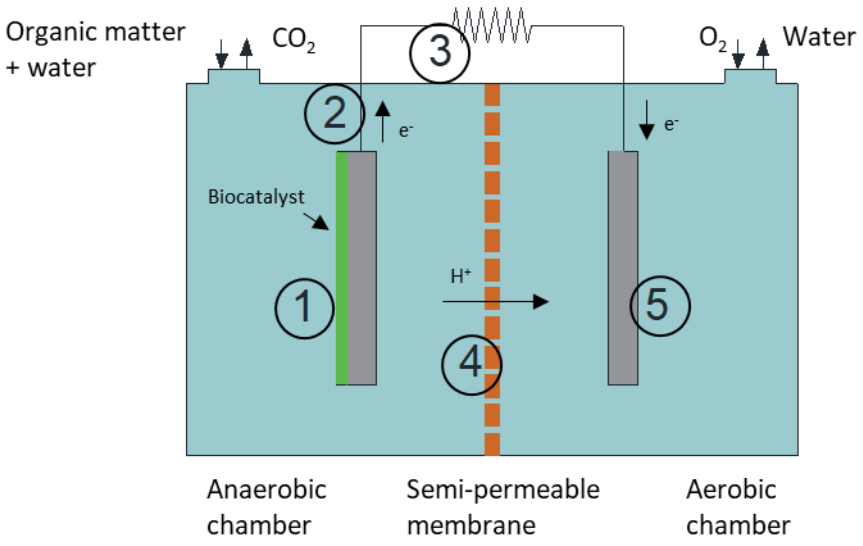


Figure 2.13. Fundamental bioelectrochemical process adapted from [177].

2.7.2.2 Microbial metabolism to electrical current

By linking up the microbial metabolism to electrodes via extracellular electron transfer, electrical current can be generated or consumed [177]. Overall, the voltage over a MFC can be described as:

$$E_{cell} = E_{emf} - (\Sigma\eta_a + |\Sigma\eta_c| + IR_\Omega)$$

Equation 2.1

With Ecell the effective cell voltage, $\Sigma\eta_a + |\Sigma\eta_c|$ electrode related losses and IR_Ω losses related to the internal resistance of the electrochemical system [178].

By linking up an oxidation reaction at the anode and a reduction reaction at the cathode results in a BES that has a high versatility for the electron donors and acceptors [178]. Where a high cell potential occurs the oxidation can break down organic pollutants and reduction can form high-value compounds at the cathode [179]. Figure 2.14 shows the examples of the possible oxidation and reduction reactions.

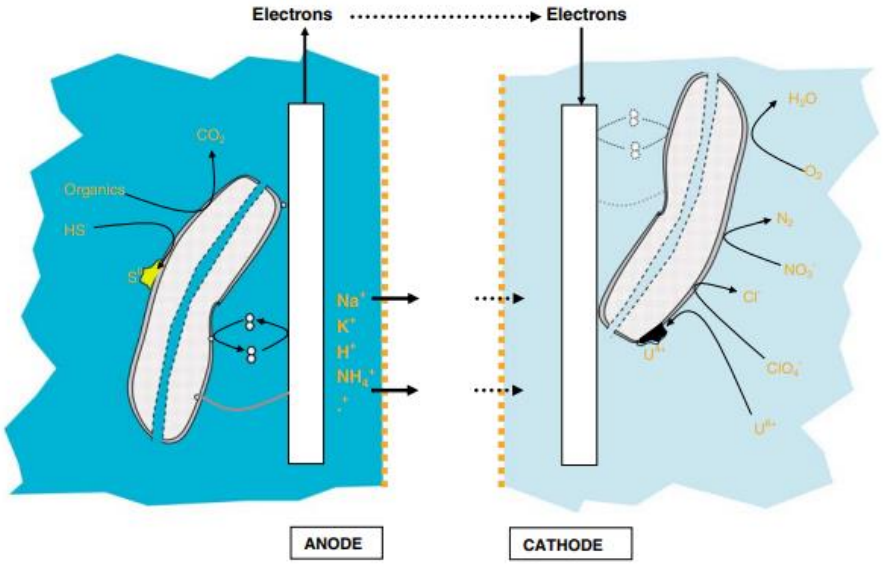


Figure 2.14. Schematic representation of possible reactions occurring at biologically catalysed anodes and cathodes, taken from [176].

2.7.3 APPLICATION OF BIOELECTROCHEMICAL SYSTEMS

Bioelectrochemical systems' various applications can be categorised into four groups [180]:

1. **Electrical power generation:** Electrical power is generated in small quantities, making MFC a suitable power source for low power systems where replacing the batteries may be impractical, such as wireless sensor networks [181,182].
2. **Bioremediation** (wastewater treatment): During the oxidation reaction pollutants in the wastewater are reduced at the anode [183]. With modifications MFC are able to treat different wastewater sources, for example, the addition of a middle chamber can be used to desalinate salt water [183].
3. **Production of valuable chemicals**: BES can harness the reducing power at the cathode to produce value-added products [184].
4. **Biosensing:** MFC can be used to detect pollutants present in water [185].

This work in interested in BES ability to treat wastewater; however, if a BES plant is to be commercially viable a combination of the above deployments will likely be required.

2.7.4 TYPES OF BIOELECTROCHEMICAL SYSTEMS

Two branches exist in BES research dependant on the type on biocatalysts used. These biocatalysts are split into two distant groups; microorganisms and enzymes/enzymes arrays [180]. Although, the first BES system to be reported used microorganisms, the majority of the research of the past 60 years was concerned with enzymatic cells [181]. This was due to enzymes being able to achieve high turnover rates, and significant current densities [180]. This was reversed in 1999 with the development of a mediator-

less MFC [171]. Up until this point microorganisms based BES had been indirect transfer and enzymes direct. Yet by removing the mediator, they became commercially viable compared to the high cost of synthesising purified enzymes. However, enzymes are still preferable for certain applications. They have stronger substrate selectivity and quantitative conversions which make these systems particularly suitable for biosensing and specific bioelectrochemical conversion applications [180]. On the other hand, whole-cell-based microbial biocatalysts are the most suitable in applications where extensive and diverse bioelectrochemical conversions of chemical compounds are required, e.g. complete oxidation of complex substrates at bioanodes for bioremediation [186]. Because of this microorganism based BES is a more suitable solution for wastewater treatment. A summary of the types of microbial-based BES is presented in Table 2.4.

Table 2.4 Summary of the types of microbial-based bioelectrochemical systems (BES).

BES Technology	Overview	Application	Variation from basic MFC
Microbial fuel cell (MFC)	Two chambers, anaerobic anode chamber and the aerobic cathode chamber, separated by a semi-permeable membrane [182].	Bioelectricity production Water treatment Biogas production	n/a
Microbial electrolysis cell (MEC)	Similar configuration to a MFC but both chambers are anaerobic and external power source provided to drive reactions at cathode [187].	Hydrogen production Water treatment	Improved hydrogen generation compared to a MFC. External power source provided.
Microbial desalination cell (MDC)	Microbial desalination cells operate in situ desalination. A third chamber is added in the middle in which salt water passes through. The cell's positive and negative electrodes attract the positive and negative salt ions in the water and, using semi-permeable membranes, filters out the salt from the seawater [188].	Desalination Water treatment Bioelectricity production	Additional third chamber for desalination.
Microbial electrosynthesis (MES)	Microbial electrosynthesis refers to the production of chemical compounds in an electrochemical cell by electricity-driven CO2 reduction as well as reduction/oxidation of other organic feedstocks using microbes as biocatalyst [189]	Synthesis of high-value chemicals.	External power source sole source of power. Cathode chamber contains CO2.

			Microbes applied to the cathode for the synthesis of high-value chemicals.
Microbial solar cell (MSC)	Microbial solar cells (MSCs) make use of photoautotrophic microbes or higher plants to entrap solar energy which is further utilised by electroactive bacteria to perform electrode-driven reactions. These reactions include generation of electric current or compounds like hydrogen, methane, ethanol, hydrogen peroxide etc [190].	Bioelectricity production Water treatment	Solar energy harnessed by the cell.
Plant microbial fuel cells (PMFC)	Plant-microbial fuel cells (PMFCs) incorporate the living plant's root system into the MFC anode so that the organic products, called as rhizodeposits, released to the soil by the roots, can be used as a substrate for electricity production by electroactive microbes [184].	Bioelectricity production Water treatment	Plant root used as the anode.

2.7.5 PREVIOUS WORK MODELLING BIOELECTROCHEMICAL SYSTEMS

Bioelectrochemical systems are complex systems, affected by various biological, physical-chemical and electrochemical factors [191,192]. Mathematical models have been developed to measure the performance of BES. The performance of a BES is dependent on the application of the BES technology, but is commonly either wastewater treatment or resource recover (either energy or material within the wastewater). The performance of any type of BES depends on a number of parameters; for example, the type of microorganisms and feedstock (wastewater) characteristics, membrane or separator properties, voltage or current supplied, mixing and diffusion phenomena and surface area of electrodes. Improving the performance of BES is still challenging and thorough understanding of the relationships among the parameters and the BES performance is important will help BES technologies move up the technology readiness level scale [174]. Both first principal models and DDM have been developed to help understand the relationship between the BES parameters and performance.

Existing first principal models built are fundamentally based on mass balances [177] and have been developed to model the: reactor hydraulics [193], biofilm [194], reaction kinetics [177], mass transport [195], ion selective membrane transport [196] and electron transfer mechanisms [197]. Notable among the first principal models is the work published by Zhang and Halme, who developed the first model to predict current as a function of substrate concentration [198]. However, the model's application was limited to due to not including mass transfer and cathode kinetics [198]. Recent research into developing first principal models have focused on trying to find the optimal reactor configuration [191]. For example, Esfandyari *et al.* extended the batch MFC model [199] to a continuous two chamber MFC model [200]. The model had a simple structure and can be used for the optimization, and design of a control system for a MFC. However, both these models are built from integrating together existing formula, which requires assuming all the model parameters are uniformly distributed and no change occurs spatially [199,200]. Many previous studies have shown that the BES components have a spatial gradient, which influences the performance of the BES [191].

Artificial neural networks algorithms are the most commonly published DDMs that predict BES performance. Tardast *et al.* developed an ANN model to predict power density generated by different MFC reactor configurations [201]. The ANN architecture had one hidden layer containing three neurons and the model had three inputs, which were temperature, pH and electron acceptor concentration [201]. The model was able to accurately predict power density and Tardast's work could be expanded to use the model to find the optimal temperature, pH and electron acceptor concentration to

maximise power density. However, the model application to other MFC systems is limited by not including any information on feedstock. Garg *et al.* compared three different DDMs (ANN, SVR and multi-gene genetic programming), to find which was superior when predicting the voltage produced by a MFC [202]. Garg *et al.* compared the three DDMs ability to fit the system during standard operation and during start-up [202]. Microbial fuel cells have a relatively long start-up period (3+ days), as the biofilm is grown on the anode and cathode [202]. Garg *et al.* work was the first attempt to model the start-up period of a MFC using DDMs [202]. The multi-gene genetic programming model was found to most accurately fit the system, both during and post start-up [202].

Other recent work modelling BES using DDMs including: predicting MFC Columbic efficiency and chemical oxygen demand (COD) removal rates using an ANN [125], predicting power density and Columbic efficiency of a MFC using neuro-fuzzy modelling [203], detecting recalcitrant organic chemicals in water with microbial fuel cells and ANNs [204], predicting COD concentration in a MFC effluent stream using ANN [185]. These models use only very limited data on the BES feedstock (e.g. temperature, pH and COD concentration). Therefore, there is an opportunity to build a DDM that is includes a greater number of feedstock characteristics (e.g. total suspends solids, dissolved oxygen, conductivity, nutrient and metal concentration and water hardness). This would help to understand the effect feedstock characteristics have on a BES performance.

2.8 THE H^2AD BIOPROCESS

The H^2AD bioprocess is the application of BES bioprocess principles to build a wastewater treatment technology that is highly scalable, compact design and economically and environmentally sustainable. For these reasons, it is an attractive option for SMEs seeking ways to treat and recover energy from their wastewaters.

2.8.1 INTRODUCING THE H^2AD BIOPROCESS

SMEs are under pressure to reduce manufacturing costs and an effective waste/water management strategy is paramount. By 2050 there will be a 400% increase in demand for water by the manufacturing sector [205]. The waste hierarchy was first introduced into European waste policy by Directive 75/442/EEC to increase the pressure on manufacturers to minimise, reuse, recycle and recover valuable resources from their waste [206]. Larger manufacturers have been utilising processes, like AD, to valorise their waste to produce bioenergy and consider the resulting solid waste as fertiliser [207]. The upfront capital costs of AD are a significant barrier for SMEs, which are generally more sensitive to additional financial costs [208]. However the adoption of

sustainable practices, like waste valorisation, are thought to present better opportunities for SMEs than to larger manufacturers, as they are able to act more nimbly to fill local or specialised markets [209].

In the UK, Lindhurst Engineering Ltd. in partnership with the University of Nottingham has developed a technology called H^{2}AD Micro AD. Lindhurst Engineering Ltd first began developing the H^{2}AD bioprocess in partnership with the University of Nottingham in 2010. It is based on Porcu *et al*. work in developing a continuous flow MFC [210]. Previously, commercial development of a plant scale MFC has been unsuccessful because of large capital costs and low generation of electricity. As part of the HORIZON 2020 project the EU funded Lindhurst in developing a hybrid of MFC and conventional anaerobic digestion (AD) to overcome these barriers [211]. Lindhurst was tasked with building five H^{2}AD 30 (3,000L volume capacity) which were trialled across Europe treating different waste sources.

The H^{2}AD bioprocess is capable of treating a variety of wastewaters, from sources including agriculture, brewing, soft drinks, foods, bio-manufacture residues, to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy. The H^{2}AD is a hybrid of anaerobic digestion and a microbial fuel cell, targeted at treating SME process manufacturer's wastewaters due to its modular, low-cost design. The advantages of the technology over traditional anaerobic digestion are summarised in Figure 2.15.

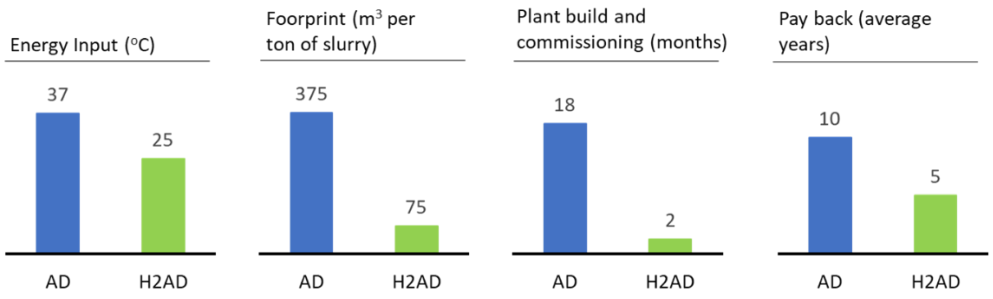


Figure 2.15. H2AD compared to traditional anaerobic digestion (data provided by Lindhurst Engineering Ltd.) [AD = anaerobic digestion; H2AD = H2AD micro AD]

The H^{2}AD technology largest potential to increase sustainability may not be from treating SME's wastewaters, but as an affordable treatment process for rural environments in low-income countries. As previously stated, up to 90% of wastewater is discharged to rivers, oceans, lakes untreated in low-income countries [166]. Conventional wastewater treatment processes, such as activated sludge, are rarely used in these environments because of a lack of electricity and financial resources [212]. The H^{2}AD technology is well suited to overcome these barriers as it is capable of generating

an independent energy supply and it is low cost compared to traditional treatments process (Figure 2.15). Furthermore, the technology has been built into a small shipping container, making it easy to transport to rural locations and quick to install on arrival. Additional affordable polishing steps could be added to the water treatment, powered partly or wholly by the bioenergy generated, to further improve the quality of treated wastewater increasing its capacity for reuse. Treating wastewater directly increases the social sustainability of rural communities by improving sanitation and public health. With appropriate polishing steps, the treated water may also provide reliable access to drinking water for rural communities.

Variations in the composition and characteristics of wastewater require management practices that can accommodate fluctuations in feedstock properties and process conditions, yet still deliver appropriate outputs without compensation in capital or operating costs. These wastewater variations, as well as potential environmental factors (e.g. seasonal), can decrease the efficiency of the bacteria to reduce pollutant load and generate biogas, which are the heart of the H^2AD bioprocess. This case study aims to develop a DDM aimed at understanding how variations in the input wastewater stream and H^2AD process conditions inform on the process performance. The performance is defined as the percentage removal of wastewater pollutants and volume of biogas generated. The model is trained on data collected from an operating H^2AD plant currently treating wastes produced by a dairy farm, located in Sutton Bonington, UK. The farm waste contains cattle slurry, bedding waste, waste milk, footbath, parlour washing and rainfall. The model will then be used to:

1. Analyse and predict wastewater variations effect on the bioprocess performance and respond to energy challenges.
2. Predict the bioprocess's ability to handle new feedstocks.

2.8.2 UNDERLING PRINCIPLES OF THE H^2AD BIOPROCESS

The H^2AD bioprocess consists of four stages: (1) pre-treatment stage, (2) MFC reaction and (3) polishing stage and (4) valorisation stage. All these stages can vary dependent on the wastewater source and the desired H^2AD outputs. The H^2AD configuration for a bioprocess treating slurry waste at Sutton Bonington is shown in Figure 2.16.

(1) Pre-treatment:
    a. The waste to be treated must be in solid-liquid phase with a particle size less than 6mm for the fluid to be able to flow through the MFC reactor. For solid biomass waste, this will require an additional stage to convert to solid-liquid phase.

b. The fluid must enter at the same temperature as the reactor. Maintaining a consistent temperature in the reactor is vital for maximising the bacteria's performance. The waste passes through a heat exchanger before entering the reactor.

The H^2AD bioprocess installed at Sutton Bonington only requires a heat exchanger for pre-treatment as the slurry is already in liquid form.

(2) MFC reaction:

a. The process conditions of the MFC reactor may be configured to stimulate methane or hydrogen generation. This decision is informed from the composition of the waste feed, as to which is preferable.

The H^2AD bioprocess installed at Sutton Bonington is configured to produce methane. This is because methane and CO_2 are already natural by-products of decomposing slurry and the H^2AD bioprocess can optimise the methane generation and harness it as a useful product.

(3) Polishing stage:

a. When the fluid leaves the H^2AD reactor it may require further treatment before it can be safely disposed of. The fluid leaving the reactor is referred to as grey water.
b. The biogas passes through an absorption column before going on to be valorised. This is to remove hydrogen sulphide (H_2S) from the biogas as it is damaging to valorising technologies downstream.

Two 1.5 m^3 tanks are installed to treat the grey water at Sutton Bonington. The first is a settlement tank where the heavy particles remaining in the tank fall out of the fluid. The second is an aerobic bio-rocks system that further breaks down the organic pollutants.

(4) Valorisation stage:

a. The biogas produced can be valorised by a variety of methods. The default configuration is for a combined heat and power (CHP) unit may be fitted to convert the biogas into heat and power. Or if heat is the desired output the biogas may be used to power a boiler. This is the preferred option for manufacturers whose main operational cost is heating. For example, a brewery may prefer this option as their processes consume large volumes of hot water.

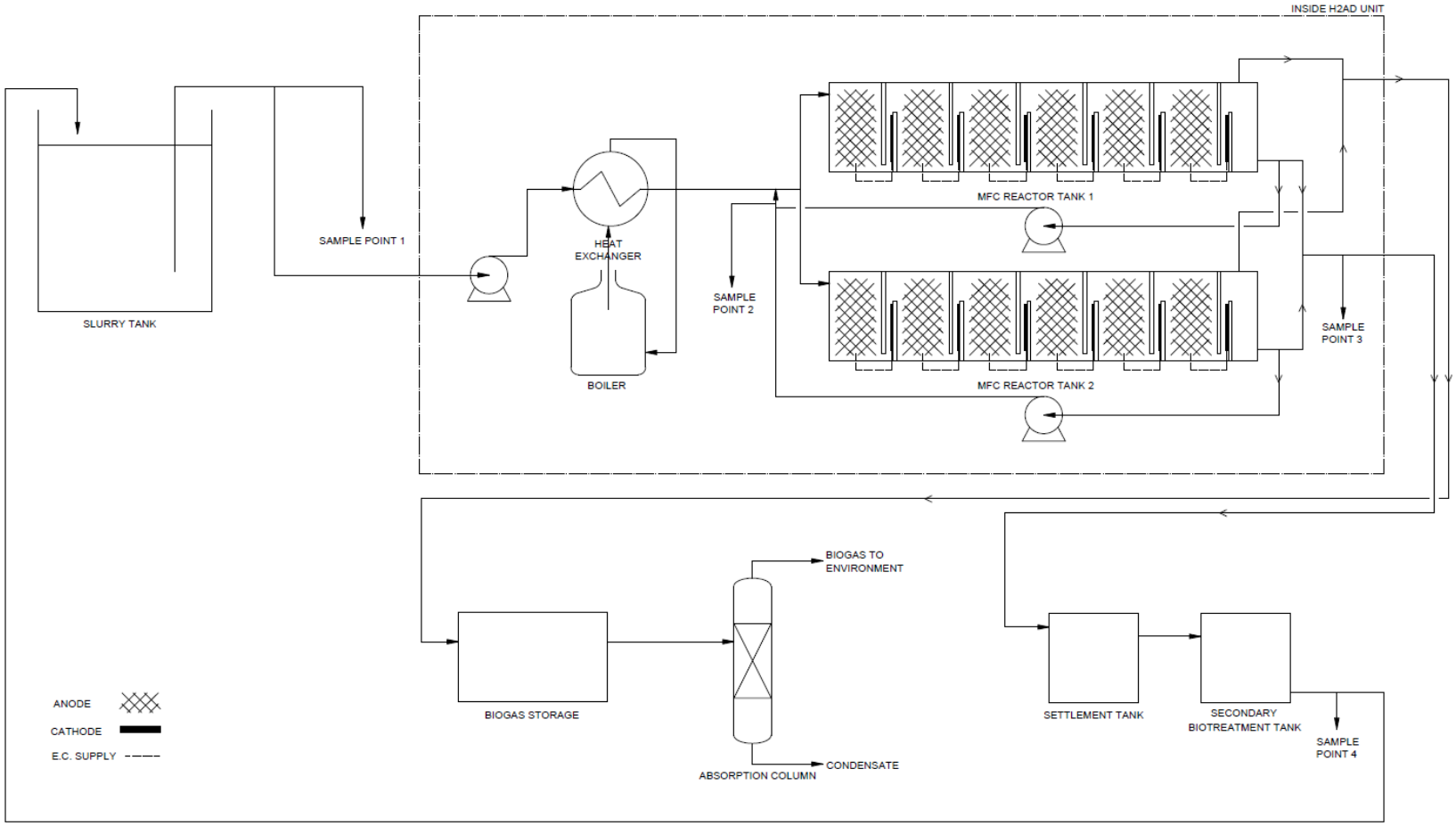


Figure 2.16. Lindhurst Engineering H^2AD bioprocess installed at Sutton Bonington

2.8.3 THE H^2AD REACTOR DESIGN

The H^2AD reactor is similar to a derivative of the MFC called microbial electrolysis (MEC). In both the H^2AD and MEC an external power is applied to the electrical circuit that drives electrons from the anode to the cathode. This supports biogas generation at the cathode. Furthermore, both operate in anaerobic conditions to promote hydrogen and methane generation instead of carbon dioxide. What makes the H^2AD reactor design unique:

- Removal of proton exchange membrane (PEM): The PEM acts as a barrier between the anodic and cathodic chambers. This is to reduce oxygen transfer to the anodic chamber while allowing for protons (H+) to transfer to the cathodic chamber. However, the PEM is very expensive and not selective to only H+ but to other cationic species such as K+, Na+, Ca+ and Mg+.
- The cathode and anode design: Convectional MFC reactors have expensive anode and cathodes that are easily degradable. Lindhurst has developed a novel design that is both low cost, robust and long-lasting.
- Semi-continuous flow: The H^2AD reactor is made up of six MFC chambers connected in series by a weir system, see Figure 2.17. Once the process fluid passed through six chambers it is recirculated. Once 300L of biogas has been produced the system will force a batch and 88L is drawn from the reactor and the same volume of fresh wastewater is added.
- Multiple H^2AD reactors operated in parallel meaning the H^2AD system can be easily scaled up to process larger volumes of waste.

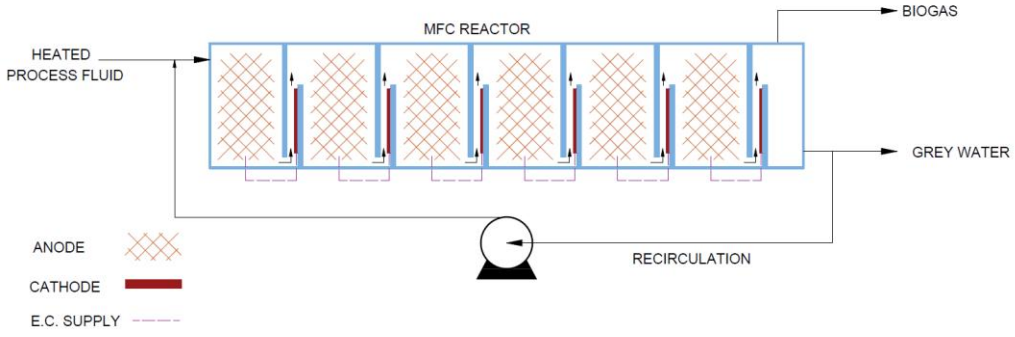


Figure 2.17. The H^2AD reactor schematic

2.8.4 H^2AD SUITABILITY AS A CLOUD MANUFACTURING CASE STUDY

How the H^2AD bioprocess matches to CM's five key characteristics have been explored to demonstrated the H^2AD bioprocess suitability as a CM case study [3]:

- **Flexibility and scalability**: the H^2AD is a modular design built into shipping containers. This means the H^2AD can be scaled to fit the demand, with more units attached if demand increases, or units turned off, if the reverse occurs. The H^2AD bioprocess can process a variety of waste feedstocks; slurry, ice cream, potatoes, brewery and fizzy drink waste has already been tested. The trend-spotting models in the cloud platform may be used to warn the manufactures if the bioprocess will need scaling up or down.
- **Multi-tenancy**: the H^2AD business is about identifying manufacturers producing wastewater who could be convinced to invest in a H^2AD unit to service their wastewater treatment requirements. The cloud platform would consist of several tenants: consumers (local wastewater producers), suppliers (wastewater treatment providers) and cloud operators (transportation, cloud managers, models developer, etc.).
- **Intelligent decision-making tool**: As part of Lindhurst H^2AD service they run a trial period where the suitability of the H^2AD bioprocess to the customer is assessed. Cloud manufacturing may utilise the models developed to predict which customers' wastes are likely to make a good match, saving time and money spent in unsuccessful trail periods.
- **On-demand manufacturing**: The H^2AD bioprocess can adapt to favour different outputs. For example, dependant on feedstock hydrogen or methane production may be favourable, and the bioprocess can adapt to maximise the desired product.
- **Manufacturing as a service**: the H^2AD bioprocess is an example of a waste treatment process sold as a service. Also, the optimisation of the bioprocess is an example of how mathematical models may be developed and sold through the cloud platform.

CHAPTER 3: APPLICATION OF CLOUD MANUFACTURING TO PROCESS MANUFACTURING

3.1 STATEMENT INTEGRATION OF PUBLISHED WORK WITHIN CHAPTER

Integrated into this chapter is material from the published or soon to be published work:

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2018. Cloud manufacturing as a sustainable process manufacturing route. *Journal of Manufacturing Systems*. 47; 35-68. [Impact Factor: 2.770; SNIP 2.191]

Fisher O, Watson N, Escrig J, Gomes RL, 2019. Intelligent resource use to deliver waste valorisation and process resilience in manufacturing environments. *Johnson Matthey Technology Review*. 64; 93-99. [Impact Factor: 1.27; SNIP: 0.669]

3.2 HIGHLIGHTS

- Combines existing cloud manufacturing (CM) definitions towards a unified version.
- For the first time all of CM's characteristics have been outlined, which will aid readers' ability to synthesis new CM applications.
- Evaluates the application of CM for a sustainable manufacturing future.
- A new system-system circular economy model is proposed.
- For the first time demonstrates CM's unique potential for the process industry.

3.3 INTRODUCTION

This chapter starts by using word cloud analysis to synthesis a new comprehensive CM definition that highlights CM potential to increase manufacturing sustainability. When reviewing the published case studies on literature CM it was discovered that CM potential for process manufacturing has yet to be evaluated. This chapter presents new methods CM enables waste valorisation and enhanced process resilience of process manufacturing systems.

There are four main sections in this chapter. The first section presents a new CM definition and outlines CM characteristics. The second section focuses on how CM may be utilised to improve sustainability to both discrete and process manufacturing. The

third section explores methods by which CM enables waste valorisation and enhanced process resilience of process manufacturing systems. The fourth section discusses the specific limitations preventing CM adoption by process manufacturers.

3.4 EXPANDING THE CLOUD MANUFACTURING CONCEPT

3.4.1 REDEFINING THE CLOUD MANUFACTURING DEFINITION

Previous cloud manufacturing definitions (chapter 2.4) prioritise different aspects of CM; for example, service oriented manufacturing, product development, resource allocation or customer centric manufacturing. The chosen aspect normally reflects the possible application of CM that was being demonstrated and was therefore highlighted in the authors' definition. Despite the differences in these definitions, all contain common elements such as, network manufacturing, ubiquitous access, multi-tenancy, virtualisation, big data, everything-as-a-service, scalability, and resource pooling. These can be used to help define the fundamental principles and characteristics of CM, which in turn give thought to how it can be utilised to achieve a sustainable manufacturing route.

To draw out the main themes, the definitions in Table 2.2 have been entered into a word cloud generator [213], the result of which is presented in Figure 3.1. The word cloud highlights the following concepts: service oriented, on-demand, network, intelligent, access, knowledge based, shared pool, rapid, scalability, advance and virtualised. This work proposes a comprehensive CM definition combining the highlighted concepts from all these definitions and emphasising CM potential for sustainable manufacturing.

> "Cloud manufacturing is a service-oriented business model that virtualises industrial production capabilities and resources into on-demand services accessed over the internet. This transforms production and supply lines to become temporary, and provide greater flexibility and scalability resulting in increased resilience and sustainability throughout the large-scale production of products. Cloud manufacturing is a multi-tenant, intelligent, knowledge-based platform that can provide sustainable solutions throughout the product and process life cycle via effective collaboration, data mining, and communication across a network of developers, manufacturers and consumers."

This definition, although comprehensive, is too long to easily bring to mind when asked what CM is. Therefore, a shortened definition is offered that captures the essence of CM.

> "The cloud manufacturing model is the concept of sharing industrial production capabilities and resources on a digital platform accessed via the internet, which is capable of making intelligent decisions to provide the most sustainable and robust solution for the large scale production of industrial products."



Figure 3.1 Cloud manufacturing word cloud derived from [64,65,75–78]

3.4.2 FORMULATING THE KEY CHARACTERISTICS OF CLOUD MANUFACTURING

Cloud manufacturing differentiates from other manufacturing paradigms because of several unique characteristics. Existing papers have explained in detail a selection of CM's characteristics or listed them without explanation [64,77]. This paper complies all fundamental characteristics and explains their meaning, application within a CM environment and the benefits they bring. This is necessary because defining the characteristics helps to provide a clearer understanding on how CM could enable a sustainable manufacturing future.

Flexibility and scalability:

A major advantage of CM over traditional manufacturing systems is the scalability offered. By nature CM production lines are no longer fixed but temporary, allowing for the option of small or long production runs depending on the user and market demands [70, 206]. Users of manufacturing resources are matched to the nearest available and appropriate manufacturing resource provider depending on their need. These supply lines are highly reconfigurable, dynamic and agile giving CM the ability to adapt to unpredicted changes of circumstance. For the system to be agile when repurposing manufacturing resources across any number of sites, a high level of automation across the supply lines shall be required. Section 3.5 explains what is meant by automation in

a manufacturing context and how CM will further automatise the manufacturing process.

Multi-tenancy:

An important characteristic of CM is its ability to connect tenants intelligently for mutual benefit without human intervention. Each tenant is given some degree of control of their manufacturing cloud's architecture, e.g. adaptability of the user interface, without changing the application code [215]. The mechanisms by how multiple tenants and manufacturing clouds will interact is expanded on in section 2.4.2.

An evaluation and reviewing system for users, like what is seen on Amazon, will be required to ensure quality of service provided. MFG.com is a real world example of a multi-tenant service that aims to quickly match manufacturers to suppliers [216]. It uses consumer generated ratings to ensure consumers are matched to suppliers that have demonstrated their products meets quality requirements, are safe and delivered on time.

Knowledge intensive, intelligent decision making tool:

Increasingly vast data sets are generated from data collected throughout the manufacturing process, but with varying degrees of utilisation. Currently the tendency is for the data to be compressed and archived for record keeping, and often only used for after the fact emergency analysis [217]. Cloud manufacturing acts as a mechanism to share this data within the manufacturing cloud so that useful manufacturing knowledge is extracted to support intelligent decision-making algorithms. Big data shall play a critical role in data collection and analysis [218]. As well as enhancing the process, data and manufacturing knowledge sharing across the manufacturing cloud is required for it to operate effectively. For example, four fundamental functions of CM that require knowledge support have been outlined as examples [64]:

1. Manufacturing resource and ability perception, connection, and virtualisation;
2. Cloud service description, match and search, aggregation, and composition;
3. Optimal allocation and scheduling;
4. Business workflow management.

However, the effectiveness of utilising data to improve manufacturing is dependent on the quality of the data collected. The data collected must be appropriate to the situation; therefore, calculated decisions must be taken to identify how the data is collected and the required granularity of that data.

Intelligent on-demand manufacturing:

Network manufacturing is a paradigm that has existed since 2000 [59]. It operates as a network of distributed manufacturing resources that are integrated on a joint manufacturing process. Where CM differs is in the method of communication between enterprises. Cloud manufacturing consists of a network of manufacturing resources that may operate automatically and/or independently via the decisions taken by the cloud platform. These manufacturing resources may be distributed within a single enterprise or shared across multiple enterprises.

On the cloud there will exist a 'pool' of shared manufacturing resources that is coordinated by the cloud platform enabling on-demand manufacturing. It must act intelligently, for example not overloading one manufacturing resource but utilising required resources in the most effective way to meet the user's requirements. Wu *et al.* illustrates this point effectively:

> "If manufacturing resource "A" is more heavily utilized than others in the network, yet the desired process can be performed by combining manufacturing resources "B" and "C", the CM environment will automatically realize and capitalize upon this alternative to avoid excessive loading of manufacturing resource "A". [77]

Manufacturing resource sharing and circulation are very important objectives in CM. It is a prerequisite of achieving on-demand and collaborative configuration of resources and services [69].

Manufacturing as a service:

A recurring theme from various CM definitions is the concept of manufacturing moving from production oriented manufacturing to service oriented manufacturing. Service oriented manufacturing integrates services throughout the product lifecycle, and allows producers a greater understanding of their product leading to a more efficient production process [219]. Cloud manufacturing achieves this by packaging manufacturing resources and capabilities as services, coordinated through the cloud on a pay-as-you-go scheme. Cloud manufacturing acts as centralised management platform for these services with a key goal of linking cloud consumers to cloud providers who can meet the product requirements of the consumer while maintaining cost, time and quality constraints. Cloud computing is currently operated as a service model and has an architecture split into three layers, software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS) [65,215]. Extending this concept to CM

Tao *et al.* explains how the CM environment is enabled by the creation of manufacturing cloud services (MCSs) [64]. Manufacturing resources and capabilities are virtualised and categorised, before being sorted into an appropriate manufacturing cloud. These services are then sold through the cloud on a pay-as-you-go scheme, offering greater flexibility to manufacturers than traditional fixed contacts.

3.5 CLOUD MANUFACTURING AS A ROUTE TO SUSTAINABLE MANUFACTURING

3.5.1 CLOUD MANUFACTURING APPLICATION FOR INCREASING SOCIAL SUSTAINABILITY

As previously stated, there has been pressure from customers, investors and local communities to adopt sustainable manufacturing strategies. As a result, manufacturers in advanced economic countries have made progress at transforming their processes to become environmentally sustainable by reducing industrial energy use and CO_2 emissions [220]. Most manufacturers have focused exclusively on the environmental pillar and neglected the social pillar of sustainability [221]. However, to achieve complete sustainable manufacturing, manufacturers must also aim to increase the social sustainability of their processes. Activities that improve a manufacturer's social sustainability include ensuring good working conditions, complying with the law, treating suppliers fairly, good community relations, respecting human rights and ensuring product safety [222]. Implementing social practises has also been found to improve economic and environmental sustainability [223]. Cloud manufacturing may improve social sustainability by two means:

1. Enabling open production.
2. Integrating social performance measurements into decision making.

Open production is the concept of providing openness to the whole value creation process [224]. There are multiple social implications from open production, which include, the democratisation of production, decentralisation of production and local value creation and opportunities for bottom-up open innovation [224]. Cloud manufacturing virtualises the whole manufacturing process, from resource supplier to consumer, which would enable the transparency required for open production. Furthermore, CM's characteristics of flexibility and multi-tenancy are complementary to open production aim to decentralised manufacturing to provide more opportunity for local manufacturers and suppliers.

Cloud manufacturing integrates data from across the supply chain to intelligent identify the most sustainable manufacturing route. To include social implications into the

decision-making progress requires a method to quantify manufacturers and suppliers' social sustainability performance. There are several methodologies have been proposed to attempt to measure a manufacturer's social performance [222]. To ensure a fair judgement of a manufacturer's social performance, CM could act as an independent assessor utilising one or more of these methodologies. The most socially sustainable manufacturing route would be the route that is made up of manufacturers and suppliers that have the best social performance scores. Economic and environmental scores for the different routes would also need calculating to determine the best overall sustainable manufacturing route. Dependant on a manufacturer's requirements they may wish to weight these scores so one sustainability scores are given greater weight in the CM system's decision-making process.

3.5.2 CLOUD MANUFACTURING ENABLING COLLABORATION FOR INNOVATIVE DESIGNS

Cloud manufacturing offers various solutions to the barriers preventing sustainable manufacturing. Two of the barriers highlighted by companies is a lack of expertise knowledge available within an enterprise and ease of access to quantitative environmental impact information of manufacturing processes [36]. A characteristic of CM is that knowledge and data is collected, stored and shared throughout the supply chain. Those sharing data can determine what data, how much data, and with whom this data is shared. During the initial stages of a partnership between cloud users limited data sharing may take place, but as the relationship progresses data sharing may increase with trust.

A new concept of 'Social Manufacturing' has recently emerged where manufacturers and individuals work together in the production of physical goods [225]. Cloud manufacturing complements this concept through its ability to crowdsource information and simulate innovative designs [79]. The CM platform can be used to match developers to relevant expertise, and ubiquitous access to design information [75,76]. In the long term this should increase the amount of collaborative designs leading to innovative solutions to sustainable manufacturing problems [226]. Intellectual property barriers will arise here and deciding what information should be made available and what should remain private will be a challenge but overall, CM benefits from a share-to-gain philosophy [214].

3.5.3 GREATER AUTOMATION FROM THE ADOPTION OF CLOUD MANUFACTURING SYSTEMS

Automation is required to enable some of CM's characteristics. Automation is defined as "any activity that is performed by a machine or an algorithm with no human involvement" [227]. Automation of the production stage of manufacturing is not new. For CM to be effective it will need to integrate with other emerging technologies such as smart factories and IIoT, which work to further increase automation in the production stage. Recently the automatisation of the planning and scheduling of manufacturing activities has gained more attention as an alternative route to furthering the sustainability of manufacturing processes [227]. Lean manufacturing, just in time manufacturing, supply chain management, and other manufacturing paradigms are previous attempts at this [228]. Cloud manufacturing stands above these paradigms because of the unprecedented level of data integration across the enterprise. This will allow for the identification of previously unseen optimisation routes and implementation from a CM centralised management centre, ensuring factory floor and business decisions are simultaneously made in relationship to one another.

One advantage from the automatisation of the supply chain is greater process agility. Market turbulence, volatility and unpredictability are some of the biggest challenges facing manufacturers and agile supply chains are becoming vital in order to respond quickly to these changes [229]. Cloud manufacturing supply chains are highly reconfigurable, because of the increased emphasis on modularity [226]. This is achieved by the packaging of manufacturing resources and manufacturing capabilities as services sold through the manufacturing cloud. The flexibility and scalability offered allows CM to respond to market changes quicker due to the shorter supply contracts and provides greater customisation of products increasing the economic sustainability of manufacturing processes. For CM to be able to manipulate manufacturing resources it will rely on IIoT technologies integrated throughout the supply chain [163]. The cloud platform intelligently schedules the manufacturing process, meaning that the providers of CM services no longer must actively search for contacts. These features enable CM to become proactive rather than reactive when responding to market volatility. How CM will determine the optimal configuration and proactive resource discovery is a major challenge preventing its deployment and is the subject of much of the present research [69,230,231]. As enterprises begin to rely on a variety of service providers for different stages of the product cycle, methods to quantify and guarantee a quality of service will need to be developed [162].

Furthermore, advanced planning and scheduling shall reduce manufacturing time, energy and material waste. As seen from section 2.4 an extensive amount of the current literature has focused on developing methods by which the optimal manufacturing route can be discovered and implemented by CM. As a route to process sustainability this has great potential for minimisation of all types of manufacturing waste. However, what is yet to be explored is how CM can be harnessed for waste valorisation opportunities.

3.6 DEVELOPING CLOUD MANUFACTURING APPLICATION TO PROCESS MANUFACTURING

The literature up until now has focused on defining CM and constructing possible architectures for how CM may be applied to industry. In the literature, the limitations and benefits of CM have been discussed in context to the overall manufacturing industry. However, what has yet to be touched upon is that within manufacturing two distinct branches exist "discrete manufacturing" and "process manufacturing" [2]. Although there are several similarities between the two, there also exists significant differences. The application and benefits of CM to discrete manufacturing stand out because of discrete manufacturing's assembly lines and their high customisability. For example, the automobile industry already offers consumers the option to customise their cars during production. It is simpler to visualise a network of separate factories making unique components to be assembled at one location and being coordinated through a centralised system, or alternatively 3D industrial printers producing specialised products unique to each customer. These concepts are harder to visualise for process manufacturing, as the product produced is not an assembly of different components but rather a combination of raw materials or "ingredients".

3.6.1 APPLICATION OF COLLABORATIVE DESIGN AND GREATER AUTOMATION TO PROCESS MANUFACTURING

Traditionally manufacturing has been a linear model (make, use, dispose). However, due to diminishing resources and legislative drivers, a circular approach is being encouraged [7]. Process manufacturing has a unique opportunity for development of innovative designs because of its ability to manipulate the manufacturing route to increase the potential to valorise waste and by-product streams. One such innovative solution has been developed by Lindhurst Engineering ltd called H^2AD Micro AD. The H^2AD is a hybrid of anaerobic digestion (AD) and microbial fuel cell technology [211]. It reduces pollutants in wastewaters and generates energy that can be utilised by the manufacturer to directly reduce energy bills. Unlike traditional AD, H^2AD operates on the microscale meaning each batch takes only three days and can be retrofitted to existing small manufacturing plants making it suitable for small and medium enterprises

(SMEs). This technology may be modified to treat wastewater streams from a variety of industries. By using CM as platform to share historical H^2AD performance data, manufacturers may utilise this data when modifying the H^2AD unit for treating a previously untested industry's wastewater.

Both discrete and process manufacturing benefit from the application of CM for automated supply chains, as detailed in chapter 3.5. However, the application of automation for process manufacturing is different to discrete manufacturing. Discrete manufacturing has focused on automating manufacturing lines, by the introduction of autonomous machines used to assemble products instead of human labour [232]. Process manufacturing requires a high degree of automation, monitoring and control, and were among the first industries to be automated [233]. Traditional automation involves sophisticated process control which operates based on data collected from temperature, pressure, flow rate and level sensor measurements from an individual process plant. Cloud manufacturing could help to make more informed decisions relating to remote processing plant's operations, through the collection and analysis of data from multiple plant sources. This is facilitated by CM's characteristic of being a centralised repository for multi-site data [72,234].

Automation in the process manufacturing industry is progressing by the development of advanced, commercial sensors and analytics [235,236]. These sensors can collect data not previously available, enabling manufacturers to identify alternative sustainable manufacturing routes. An example is the development of a multi-sensor system based upon ultraviolet optical fluorescence imaging and ultrasonic acoustic sensors for clean-in-place (CIP) optimisation within the food and drink industry [237]. Existing CIP processes are time intensive and waste large amounts of energy, water, and chemicals as they operate on a worst-case scenario. This system aims to take the data recorded and use a machine learning approach to determine the optimal time to stop cleaning, saving both time and materials [237]. The savings made in time are significant, as the food industry spends on average 20% of the day cleaning which could instead be utilised for production [238]. Cloud manufacturing enables the data collected to be analysed remotely requiring no additional expertise within the enterprise. This is particularly beneficial to SMEs who may not have the resources for data analytics within the enterprise. Caution is needed that some element of human integration with safety and quality control practices is retained, rather than a complete reliance on artificial models. For instance, with the CIP example switching between producing different food products on the same manufacturing plant may result in the production of pathogens the sensors are not able to detect.

3.6.2 CLOUD MANUFACTURING FOR WASTE VALORISATION

The waste hierarchy was first introduced into European waste policy by Directive 75/442/EEC, emphasising the importance of waste minimisation, and the protection of the environment and human health [206]. It has been updated in Directive 2008/98/EC to introduce a five-step waste hierarchy that states which waste management methods have priority based on environmental impact and extended beyond waste minimisation to reuse, recycling and recovery (Figure 3.2) [38]. Preventing and reducing the amount of waste generated is always the first option in optimising the sustainability of a manufacturing process. This is an option available to both discrete and process manufacturing. However, regardless of how effective this stage is, the manufacturing environment will always produce some waste.

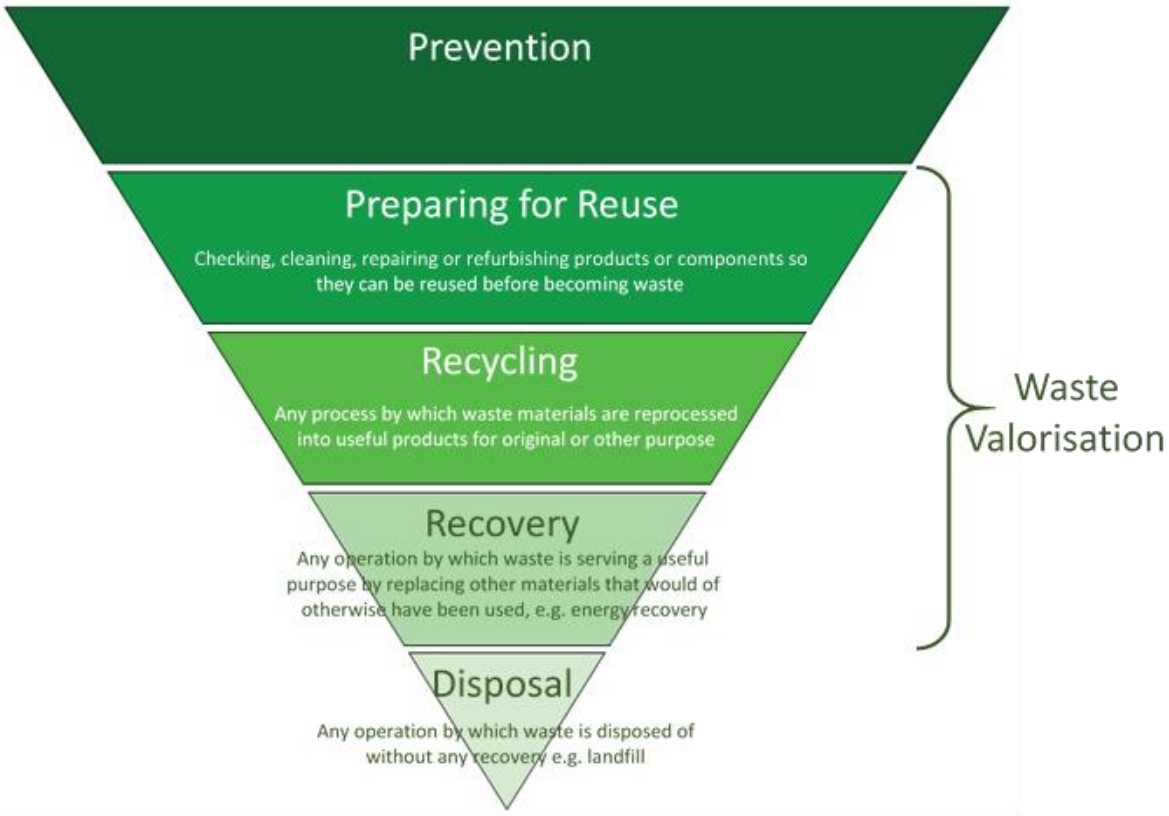


Figure 3.2 Waste hierarchy adapted from Directive 2008/98/EC [38].

The circular economy model treats this waste generated as a resource to feed back into the economy, rather than a waste to be disposed of. It takes a cyclical view where waste is constantly fed back into the economy, and there is minimal input of new material entering the system. There is a consensus among researchers that it is not possible to reach a fully circular economy, as the process of recycling will also always create waste and side-products due to increasing entropy [239]. Instead, the goal is to identify resources from renewable sources and to minimise the spent resource leaving the system

[240]. Process manufacturing systems can take this further as their waste can contain valuable material available to use as a building block for a related process manufacturing environment. This is known as the cradle-to-cradle concept, where waste is both recycled and utilised as a raw material [241]. Applying this concept expands the traditional circular economy model from a singular circular system to a series of connected circular systems, where one system's waste is anothers resource. This reimagining of the circular economy concept is shown in Figure 3.3. This will require not only a redesign of existing systems but a framework in place which intelligently identifies the most sustainable resource recovery route and matches waste producers and consumers. One example of a project attempting to do this is SHAREBOX, which aims to provide plant operators and production managers with the reliable information they need to share resources in real-time [242]. However, SHAREBOX currently limits the definition of resources to plant equipment, energy, water, residues and recycled materials. There needs to be a shift across society (inclusive from manufacturer to consumer) from viewing waste not as something that requires disposing of, but as a co-product that adds value to the same and/or related manufacturing process. The development of data-driven models to describe manufacturing systems would enable intelligent decision making on platforms like SHAREBOX that support resource sharing between systems [3].

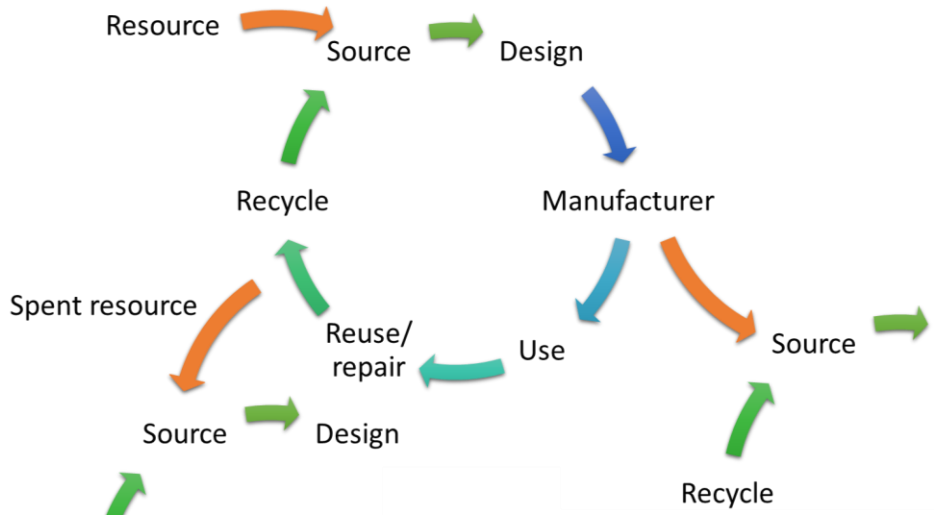


Figure 3.3 Integrating the cradle-to-cradle concept within the circular economy model

Discrete manufacturing produces mainly solid waste generated from defects, off cuttings, and unused parts during manufacturing [243]. Hence discrete manufacturing has limited waste valorisation opportunities, with instead a focus on waste minimisation and remanufacturing, which can be enhanced by CM [226]. Waste valorisation refers

to industrial processing activities aimed at reusing, recycling, or recovering useful products or sources of energy from waste.

There are currently three waste valorisation strategies, geared towards: (1) production of fuel and/or energy to replace fossil fuels; (2) production of high-value chemicals from residues; and (3) production of other useful material(s) [244]. Process manufacturing generates a wide range of solid, liquid and gas waste streams that contain significant quantities of valuable material, which may be recovered by the application of these strategies. Taking the food manufacturing industry as an example, its waste contains valuable compounds such as fats, proteins, bioethanol etc, which can be converted to soaps, rubber, fuels and more [245]. Furthermore, through biological, chemical and thermal processes, such as AD and pyrolysis, biogases and other gases can be produced as fuel or used for energy generation. Garcia-Garcia *et al.* effectively demonstrates how the waste hierarchy principles can be applied to selecting the most sustainable waste valorisation technology for treating food manufacturers' waste [245].

British Sugar's Wissington sugar production site is an excellent example of a manufacturing process that would not be economically sustainable without waste valorisation [246]. By reusing, recycling and recovering waste, British Sugar generates eleven co-products ranging from topsoil and animal feed to bioethanol and tomatoes [246]. This extensive waste valorisation has made British Sugar's Wissington factory one of the most efficient sugar factories in Europe and is an excellent example of the benefits of waste valorisation within process manufacturing [247].

Process manufacturing waste has a wide range of waste valorisation routes available [248]. As noted, a key characteristic of CM is the collection and sharing of data and expertise knowledge, which aids in the ability of CM to make intelligent, sustainable decisions. Cloud manufacturing can be used to evaluate the most sustainable valorisation waste method from this data. For example, in the production of olive-oil, solid waste pomace is produced. Traditionally pomace is a feedstock for production of pomace-olive oil and pomace wood, but there now exists several higher value products it can be converted to, including biomolecules, dyes and cosmetic products [249]. The flexible supply lines inherent within CM and real time analysis of market prices means that CM can be used to identify the most cost effective waste treatment method and co-orientate its implementation. This would be advantageous to SMEs who may not have their own waste treatment infrastructure on site but pay to have their waste taken away and treated by an external company.

Industrial symbiosis (the process of brokering relationships between manufacturers to valorise waste streams) faces the challenge of identifying local manufacturers who have the capacity to utilise another's manufacturers waste as a resource [250]. As a multi-tenancy platform, CM can be used to match different cloud users together. Connecting and enhancing communication between local manufacturers will help to identify current inefficient waste treatment methods and identify local sustainable alternatives. As an example, CM may identify an alternative feed source for a process from another local manufacturer's waste, replacing a non-renewable source. This would be a preferred treatment option because it is higher up the EU waste hierarchy. However, it is important to recognise the importance of geographical location of manufacturing and waste treatment plants when constructing the manufacturing cloud [72]. The most sustainable waste treatment option may not be local, and the transportation of waste is both environmentally and economically costly. Therefore, there must be a balance between the best treatment option and feasibility to maximise overall sustainability. Cloud manufacturing also has the capability to identify suitable transportation routes and match suppliers and customers. It has the potential to utilise wasted transportation capacity, maximising the economic return for the transporter and decreasing CO_2 transport emissions from manufacturing. Cloud manufacturing could support a business model similar to the taxi service Uber where multiple consumers with a similar journey route share a single vehicle. This is an expansion on the optimisation of supply chain scheduling CM offers, as discussed in section 3.4.

Real time monitoring of the composition of waste streams is required for CM to make intelligent decisions on how the waste should be utilised. For example, pyrolysis (the thermal decomposition of biomass in the absence of oxygen) is highly dependent on the moisture content of the feedstock [251]. At higher moisture contents, high levels of water are produced and at lower levels there is a risk that the process only produces dust instead of oil, char or gas [251]. Online measurement of the waste can feed back into the drying stage to ensure appropriate levels of drying takes place. Alternatively, in the design phase CM can be used to both access and analyse large volumes of stored process data to determine trends in moisture content of biomass waste. The conclusion may be reached to use the alternative biomass valorisation process hydrothermal carbonisation which operates in the presence of supercritical water. Hydrothermal carbonisation benefits from high moisture contents and may be judged to be the more sustainable route if a high moisture content trend is discovered [252]. Furthermore, if the waste stream is to be advertised through CM as a feedstock it will require its composition and properties to be defined to inform on the waste valorisation technology and where waste

is considered a feedstock, considerations of the variability on the waste composition is required.

Process manufacturing has the potential for a wide range of waste valorisation methods. Cloud manufacturing can support this through three main methods:

(1) It can be used as a decision tool for deciding the most sustainable waste valorisation routes, ideally the highest up the waste hierarchy to meet EU Directive;
(2) It can identify and coordinate alternative waste valorisation routes if they prove to become more sustainable;
(3) Cloud manufacturing can link and coordinate local supply chains to reuse waste streams as a feedstock, with minimal treatment.

3.6.3 HOW CLOUD MANUFACTURING INCREASES PROCESS RESILIENCE

Process manufacturing is reliant on process conditions that define how the feedstock is utilised and how equipment is operated in order to deliver a product at a predefined specification. The process conditions include but are not limited to process streams composition, temperature, pressure and flowrates. The nature of process manufacturing products having to meet a specific specification relies on a need for strong process resilience. There exist examples of process manufacturing goods being recalled because they did not meet specifications or face fines for breach of environmental limits [253]. This is also present in discrete manufacturing, though it is less common due to it being less susceptible to variations in manufacturing conditions. The current practise to ensure process resilience is to account for variability in the design stage. This section will explore how CM enhances current methods, as well as offering new solutions.

A crude but effective method of ensuring process resilience is to design the manufacturing plant to operate within the design envelope. The design envelope of a system refers to its specified requirements that, if satisfied, would enable the system to perform as intended in the operational environment [254]. This has the disadvantage of increased capital and operating costs. If the design envelope is underestimated the process plant shall fail, with a best case scenario resulting in off spec product and at worst a fatal incident [254]. Certain process conditions are outside the control of the engineers and technicians. Feedstock supply lines and composition is a critical process condition that is subject to external variables. As manufacturers move away from non-renewable sources towards waste and biomass feedstocks this will become an even greater challenge for future engineers [255]. This is again an example of how the large knowledge and data store available through CM would benefit the design stage of

manufacturing. Advanced data analysis techniques available on the manufacturing cloud would be able to predict and model to a higher degree of accuracy the future trends of feedstock variations [256]. This allows for the design envelope to be reduced while still maintaining a safe operating range and increasing economic sustainability. During the plant operation, the reconfigurable supply lines inherent within CM will allow for easy identification and implementation of alternative local feedstock sources. This can help mitigate the risk of using waste and biomass sources as feedstocks in process manufacturing operations. Furthermore, advanced data analysis of historical manufacturing data can help designers gain a better understanding of the process and products. It is more effective to ensure quality in the products by building quality into the design rather than continued testing and adjustment of critical process parameters.

It has been highlighted that a current barrier in process manufacturing that prevents enterprise-wise optimisation is the lack of data integration and supply chain coordination across the enterprise and between enterprises [87]. Qin explains how data in process manufacturing is currently collected based on a hierarchy structure [34]. The bottom levels are the monitoring of equipment via sensors that happen virtually instantaneously. At the process level there is regular monitoring of process control data (temperature, pressure and flowrate), with product quality more often sampled irregularly and maybe off site. The top level is the customer feedback data. There is a significant time lag between the data at higher levels and its implementation to the manufacturing resources. Cloud manufacturing architecture is a data-centric system that ensures data is made available across all manufacturing and business levels [23]. Rather than having a hierarchy structure, it shall be circular with data from across the supply chain feeding back into the manufacturing process. This shall reduce the time lag and better inform on the process to find the optimal configuration.

Cloud manufacturing will need to rely on big data technologies to achieve this. Big data is the collection and analysis of vast volumes of data throughout the manufacturing process to reveal trends, patterns and process optimisation routes [257]. Currently process manufacturing already collects a large amount of data from process control systems but it is not fully utilised, often only used for emergency after the event analysis [217]. Cloud manufacturing's reliance on advanced sensing technologies and widespread communication technologies will increase the amount and variety of data generated in manufacturing systems [258]. Data gathered from the manufacturing process could be used for:

1. Data mining and modelling of time series data for trend spotting and forecasting;
2. Online use for real time operations optimisation and intelligent decision making;
3. Offline troubleshooting as a valuable tool for continuous improvement.

Multivariate statistical approaches have been the recent favourite choice for data analysis, but new developments in machine learning and data mining offer opportunities to gain value from large, unclean data sets [217]. Machine learning advances in closed-loop optimisation may be used by CM to continuously optimise process control online. Houben and Lapkin summarise the recent advances in automatic optimisation of process conditions and explain how machine learning can build self-optimising algorithms [259]. By achieving tight process control the process plant will be able to operate closer to steady state process conditions despite variations in external conditions, lowering the energy and material consumption. Enterprises often lack effective analytic tools or computational power to get any real meaning out of this data. The service nature of CM allows enterprises to hire either extra computational processing power or outsource their data analysis to specialist data analysis companies on a pay-as-you-go business model.

3.6.4 EXAMPLES OF CLOUD MANUFACTURING'S APPLICATION TO PROCESS MANUFACTURING

To further reinforce how CM may be applied to process manufacturing, examples are presented in Table 3.1 for each of CM's unique characteristics applied to different industries.

Table 3.1 Examples of cloud manufacturing's application to process manufacturing industries.

Cloud manufacturing characteristic	**Application to process manufacturing**	**Industrial examples**
Flexibility and scalability	Continued monitoring of real time economic data to ensure the optimal manufacturing route.	Two adsorbents may be used for the recovery of CO_2 from flue gas by pressure swing adsorption, Zeolite and Activated Carbon [260]. During the lifetime of the plant these adsorbents may need to be replaced. Cloud manufacturing can be utilised to reassess which adsorbent is most sustainable, using the most recent economic and performance data.
Knowledge intensive	Utilise data across the supply chain to build an artificially intelligent (AI) algorithm to optimise process conditions.	The active component for all anti-malarial drugs is artemisinin and is harvested from a plant called Artemisia annua [111]. When the plant is harvested impacts the manufacturing plant's optimal process conditions to improve the artemisinin recovery [111]. Utilising IIoT technologies during growing and harvesting, CM system would integrate this data downstream to more accurately predict the required process conditions.
	More accurate pricing of feedstocks, wastes and products.	When a batch of potatoes arrives at a factory for the production of crisps, samples are taken to assess the quality of the potatoes and determine their value [261]. These tests can prove unreliable and new tests which utilise artificial neural networks are being developed [261]. By uploading these results to the manufacturing cloud, CM can be used to track various deliveries and coordinate accurate payments to each potato supplier.
Multi-tenancy	Co-ordination of multiple manufacturers' waste streams within geographical region.	Brewer spent gain (BSG) in a wet solid waste from the brewery process which has a wide variety of valorisations options [262]. Individually microbreweries do not produce enough BSG to make it

		feasible to valorise. However, CM could be used to coordinate the collection of BSG to enable one of these technologies to be successfully implemented by a third party.
	Sharing AI algorithms	An algorithm created to optimise the treatment of a wastewater by anaerobic digesters may be sold through CM to the benefit of the initial developer and other water treatment companies. The algorithm would be built upon valuable performance data that the developer would not want its competitor accessing. Cloud manufacturing acts as an impartial third party able to utilise performance data to retrain and optimise the algorithm, without disclosing eithers manufacturers' data.
On-demand	Customisation of products	The data generated by wearable 'lifestyle' technologies has the potential to revolutionise the method by which patients are diagnosed [248]. This data may also be used to determine an individual's required dosage. Cloud manufacturing would act as a platform sharing data between manufacturers, doctors and patients to enable personalised medicine.
	Responding to changes in customer's demands	Demand side management is the modification of consumer demand of energy, and is used to attempt to limit the peak demand on an electrical grid [263]. However, the energy sector is facing the challenge of trying to utilise unreliable renewable energy sources and an increase in customer demand [263]. One of the largest consumers of energy are process manufacturers. Cloud manufacturing could help limit the demand on the grid during peak times by increasing the communication between energy providers and manufacturers to schedule the most energy demanding processes when there is minimum demand on the electrical grid.

Manufacturing as a service	Outsourcing expertise	Affordable IIoT technologies are currently been developed aimed at automating the brewery process in microbreweries, generating new data sets [264]. Microbreweries have limited employees and do not have in-house data analyst to effectively make use of this data. Cloud manufacturing can be used to outsource their data analysis needs.
	Outsourcing parts of the supply chain	In the dairy industry whey is a by-product of cheese production, and has a wide variety of valorisation options [265]. Large scale cheese manufacturers normally valorise the whey on-site. Small independent cheese makers have limited capital to invest in these technologies to valorise on site, and instead the whey is treated as a waste. Through CM they can outsource the valorisation of the whey to either waste management companies or local large scale cheese manufacturers.

3.7 SYTHESISING CLOUD MANUFACTURING LIMITATIONS AND FUTURE RESEARCH DIRECTIONS FOR PROCESS MANUFACTURING

The literature to date has identified limitations that prevent widespread adoption of CM by manufacturers from the discrete manufacturing perspective. These limitations being identified as data sharing, cyber security, integration of technologies, modelling of material flow, and integrating existing systems within a CM system [85,266]. However, these limitations are also relevant to process manufacturing, for instance data sharing among process manufacturers (chapter 3.7.1). There are also limitations applicable only to process manufacturing, which are processing plant security and characterisation of process streams (chapter 3.7.2 and 3.7.3 respectively). Potential limitations may also be industry specific within process manufacturing, when the cloud user is an SME (see section 3.7.4), though CM will also afford unique opportunities to SMEs [80].

3.7.1 DATA SHARING AMONG PROCESS MANUFACTURERS

The current barriers obstructing free flow of data between enterprises have already been thoroughly discussed [98]. However, data sharing is important to elucidate on from a process manufacturing perspective. Large data sets are required to train and optimise the AI algorithms that CM will use [99]. As previously stated, the large process manufacturers were the first industries to be automated and thus have generated large quantities of data [63]. However, new companies and SMEs may not have access to enough data to build, train and validate the required mathematical models. Cloud manufacturing can overcome this by sharing either already built algorithms or manufacturing data from similar processes to act as a starting point in the construction of these algorithms. Manufacturers are naturally reluctant to share their data between competitors. A framework needs to be developed that will allow the manufacturing cloud to act as a third party to utilise another company's manufacturing data and/or models, without disclosing this information to the competing company. A pricing mechanism will also be required to assess and judge the value of the data used.

4.7.2 PROCESSING PLANT SECURITY

Unlike discrete manufacturing, process manufacturing involves the thermal, chemical and or biochemical conversion of materials to products at temperatures and pressures which can be hazardous. Some industrial reactions are exothermic and there is the potential for runaway reactions if not controlled. This is a danger unique to process manufacturing industries and there are many reported cases where this has occurred and resulted in causalities [267]. Process plant safety has evolved in past 20-30 years, moving from corrective to preventive. The emphasis is on designing out hazards and

where this is not possible, process control systems are used to monitor temperatures and pressures within a reactor ready to shut off if they reach dangerous levels. However, as with any use of online systems, the potential for hacking needs evaluating with risk mitigation. For process manufacturing this is a legitimate consideration and one not relevant to discrete manufacturing.

As manufacturing resources are increasingly controlled remotely there exist safety concerns protecting these control systems from malicious attacks. The "Aurora Project", performed by Idaho National Laboratory in 2007, first demonstrated the dangers from cyber criminals hacking into a process plant's control system to manipulate and ultimately destroy physical systems [268]. Sophisticated cyber-attacks on control systems are becoming more frequent, resulting in the need for advanced cyber-security systems [268].

3.7.3 CHARACTERISATION OF PROCESS STREAMS

As highlighted in chapter 2.2, a fundamental difference between discrete and process manufacturing is the variability in process manufacturing's process streams. Be these feedstocks, wastes or by-product streams. For CM to be able to make intelligent decisions, it will require the characterisation of these process streams in real time, compared to the current practise of infrequent sampling and analysis off line [217]. This is because the AI algorithms that will be developed to support CM's decision-making ability will rely on real time data to ensure they are continuously learning. These AI algorithms must be capable of learning, so they can continually work to optimise all aspects of the manufacturing process and adapt to variations in process streams. Process manufacturing is currently limited from further automation by the development of commercially affordable sensors capable of characterising process streams in real time.

3.7.4 BARRIERS PREVENTING SMES ADOPTING NEW TECHNOLOGIES

Small and medium enterprises are considered the backbone of a country's economy, providing the majority of all new jobs and vital for stimulating economic growth [269]. Within process manufacturing certain industries have a strong SME portfolio, for instance the food and drink industry [270]. The wealth of SMEs are seen as a benefit because of the high value artisan products they produce, for example bakery products, cheeses and craft beers. There is strong growth within these industries as customers demand quality over quantity and are prepared to pay for this. The brewery industry in the UK is an example of this, observing a large increase in microbreweries despite a downward trend in beer consumption [271]. However, SMEs can struggle to adapt to external market forces and to make the next step in expanding their business due to a

lack of expertise and resources. This is where CM shows great potential in enabling SMEs to overcome these barriers [80].

Small and medium enterprises can be slow in adopting new technologies, which are shown to improve revenue. They do not always have either the capital to invest in new technologies, or the expertise or guidance necessary to successfully implement them. The countries where SMEs have been most successful in adopting new technologies are countries which have coherent Government strategies in place to support them, for example in Germany (Industry 4.0), China (Made in China 2025), and the USA (America Makes) [272]. This has resulted in higher levels of productivity (output per worker) in these countries over others, like the UK, which have lower levels of adoption [272]. If CM is to follow this example there needs to be a strong, coherent strategy in place to advice SMEs on how best to utilise CM for their needs.

3.8 CONCLUSION

- For the first time a detailed explanation of all of CM's key characteristics has been summarised. By outlining these characteristics and how they can be applied to manufacturing, it enabled the exploration of how CM may be used to enhance sustainability.
- Cloud manufacturing has yet to be fully applied to process manufacturing. Presented are examples of applying CM characteristics to process manufacturing industries.
- Four key methods of how CM increases sustainability have been identified: (1) collaborative design; (2) greater automation; (3) increase in process resilience; and (4) enhanced waste reduction, reuse and recovery
- **Collaborative design**: A critical barrier preventing sustainable manufacturing designs is the accessibility of environmental information and expert knowledge within an enterprise. Cloud manufacturing operates on a multi-tenancy cloud platform that can crowdsource information and expertise to simulate the development of innovative sustainable designs. Process manufacturing has unique opportunities for new sustainable design from its ability to manipulate the manufacturing route to valorise waste and by-product streams.
- **Automation**: Cloud manufacturing can perform real time analyses of manufacturing and market data to intelligently schedule and plan the manufacturing process. This allows for CM to optimise the manufacturing process in real time and coordinate its implementation across the supply lines from a centralised management centre. This reduces the manufacturing costs

through minimising energy consumption, material waste and downtime. Process manufacturing will become further automated by the implementation of advanced sensors, allowing CM to develop intelligent models. These models help to identify and implement more sustainable manufacturing solutions.

- **Waste valorisation**: Process manufacturing waste has the potential for valorisation as it can contain valuable material residue, co-products opportunities, residual heat, and energy recovery. It is not always obvious which is the most sustainable waste valorisation route, and CM can be used as an intelligent decision making tool. CM can monitor functions in the composition of waste streams and market prices to identify and coordinate alternative waste valorisation routes if they prove to become significantly more sustainable. Furthermore, CM can coordinate local supply chains to reuse waste streams as a feedstock. From these methods, CM aligns with the circular economy vision perceiving waste as a valuable resource to be reused, recovered and regenerated.
- **Process resilience**: Process manufacturing systems are particularly susceptible to variations in process conditions, especially feedstock composition. Currently over engineering is the best method to ensure variations do not impact product quality. Advanced data analytic techniques offer the potential to gain better process understanding to ensure quality in the design of the system. As advanced industrial internet of things (IIoT) sensors are developed, this will provide CM with greater quantity and granularity of data meaning previously unidentifiably solutions can be implemented.

CHAPTER 4: CONSIDERATIONS, CHALLENGES AND OPPORTUNITIES WHEN MODELLING PROCESS MANUFACTURING SYSTEMS

4.1 STATEMENT INTEGRATION OF PUBLISHED WORK WITHIN CHAPTER

Integrated into this chapter is material from the published or soon to be published work:

Fisher O, Watson N, Escrig J, Witt R, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Considerations, challenges and opportunities for manufacturers when developing a data-driven model. *Computers & Chemical Engineering* (**Submitted to journal**, 20 Aug 2019). [Impact Factor: 3.113; SNIP: 1.613]

4.2 HIGHLIGHTS

- Data-driven models (DDMs) will become widespread across manufacturing.
- Paramount to DDMs is the collection of an accurate set of model development data.
- Process manufacturers face unique considerations and challenges in collecting data.
- A new protocol to aid manufacturers to develop DDMs is proposed.
- This supports the development of DDMs to meet manufacturers' requirements

4.3 INTRODUCTION

The previous chapter developed applications of CM to enhance process resilience and waste valorisation of process manufacturing systems. Cloud manufacturing relies on the development of mathematical models to describe process manufacturing systems. The reason for this is to be able to virtualise the system onto the manufacturing cloud, which can then simulate the system to identify the most sustainable manufacturing route. The increasing availability of data, due to the adoption of low-cost industrial internet of things technologies, coupled with increasing processing power from cloud computing, is fuelling increased use of data-driven models (DDMs) in manufacturing. Utilising case studies from the waste management industry and the food and drink industry, the considerations and challenges faced when developing DDMs for manufacturing systems are explored.

There are four main sections in this chapter. The first section outlines the case studies from the waste management industry and the food and drink industry used in this

chapter. The second, third and fourth chapters then outline the considerations, challenges and opportunities when modelling process manufacturing systems.

4.4 OUTLINING CASE STUDIES

4.4.1. CASE STUDY 1: WASTE/WATER MANAGEMENT

Case study one uses the H^2AD case study outlined in chapter 2.8. The H^2AD bioprocess is capable of treating a variety of wastewaters, from sources including agriculture, brewing, soft drinks, foods, bio-manufacture residues, to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy. The H^2AD is a hybrid of anaerobic digestion and a microbial fuel cell, targeted at treating SME process manufacturer's wastewaters due to its modular, low-cost design.

4.4.2. CASE STUDY 2: CRAFT BREWERIES

There are over 2,000 craft breweries within the UK, most utilising equipment and processing techniques that have not changed for over 30 years [273]. This makes it an often unpredictable and inefficient process, especially regarding water overconsumption [274]. A critical stage of the brewing process is fermentation, where yeast is added to the wort (the liquid extracted from the mashing process downstream) to convert sugar to alcohol (ethanol). The fermentation process is complete once the beer has reached the desired alcohol content and flavour profile. For SME breweries, this is currently determined by removing samples from the vessel and manually measuring the specific gravity with a hydrometer. Although the fermentation duration should be identical for each batch of a particular beer, this is rarely the case due to seasonal variability in ingredient (malts, hops, water) properties and natural fluctuations in process temperature. As specific gravity measurements are only taken every 4-10 hours (or longer if no one is working overnight) this often leads to overfermentation, affecting product quality and resource utilisation.

Ultrasonic sensors can be used to monitor industrial processes such as equipment cleaning [108] and multiphase flow [275]. Previous research has shown that ultrasonic measurements can be used to monitor changes in ethanol volume concentration during beer fermentation processes [276,277]. There is an ongoing project between the University of Nottingham and an SME craft brewery in Nottingham, UK called Totally Brewed, to develop a low-cost ultrasonic sensor designed for craft brewers. This project records ultrasonic and temperature measurements during fermentation and uses supervised machine learning algorithms, such as artificial neural networks (ANNs), to predict the alcohol concentration from these measurements.

These case studies were chosen because they contain characteristics unique to process manufacturing systems (Table 4.1). Examples of the considerations and challenges resulting from these characteristics are referred to throughout this paper to support the topics discussed in each section.

Table 4.1 Summary of case studies (CSs)

	Waste/water management CS1	Brewery CS2
Manufacturing system	The H^2AD bioprocess is a circular economy technology, which treats wastewater to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy.	Fermentation is a critical process in beer production where the sugar in the wort is converted to alcohol.
Project's aim	To develop a model to analyse and predict the effect of wastewater variability (e.g. suspended solid content) has on the bioprocess performance. Utilise model to make predictions on bioprocess's ability to treat new feedstocks.	To develop a model to predict alcohol concentration from temperature and ultrasonic measurements during a fermentation process.
Feedstock characteristics	Farm waste which has variability due to changes in farm practices, the welfare of cattle, rainfall, seasons and more.	Raw ingredients are cereal grain, hops, water and brewer's yeast. Each ingredient contains inherent variation which can affect flavour, colour, carbonation, alcohol content and other subtle changes in the beer.
Product specification	Wastewater pollutants are reduced so water is fit for reuse. And biogas generated contains a minimum of 70% by volume of methane.	'Slap in the Face' beer containing 4% alcohol concentration.
Process description	The H^2AD is a semi-batch bioprocess, meaning that X volume of liquid enters and leaves the system at predefined time intervals. When entering the system the feedstock is heated to 30°C. The	The main stages in the brewing process are: wort production, alcoholic fermentation and maturation. It is a batch process with the fermentation lasting between 3-6 days.

	process stream is recirculated through the system.	
Model inputs	Water quality analysis parameters of feedstock (e.g. pH, dissolved oxygen, chemical oxygen demand, total suspended solids). H^2AD process conditions (e.g. temperature, system pressure).	Temperature, ultrasonic velocity and received ultrasonic signal amplitude, during fermentation.
Model outputs	Chemical oxygen demand (mg/L) and total suspended solids (mg/L) of the output stream. Daily biogas production (L) and methane concentration of biogas (% vol).	Alcohol concentration during fermentation.

4.5 CONSIDERATION FOR DATA-DRIVEN PROCESS MANUFACTURING MODELS

When developing a model, there are considerations, which must be made before the data collection and model building stages. Primarily, these considerations are

1. Defining the model's goal;
2. Understanding what is required from the model and,
3. Data considerations (what data is already collected, what additional data points are required, how much data is needed, how the data is to be collected and how often).

Detailed discussions between the manufacturer and modeller must take place to address these considerations. The modeller (or team of modellers) may be an internal employee(s) or manufacturers may sub-contract data analysists and software specialists to model their processes. Digital companies, like Microsoft's Azure software, are now offering these services through the cloud [278]. This service-orientated approach is often more affordable as they operate on a pay-as-you-go business model [3]. This means they are more readily available, especially for SMEs, who are understandably unlikely to have the required expertise within their current workforce or be able to afford the cost of developing, running and maintaining the models themselves.

4.5.1 DEFINING THE MODEL'S GOAL

When constructing a mathematical model, the first stage is to identify and define the model's goal, as this will define the model's outputs [87]. A model may have multiple goals; these may change over time and may come into conflict with each other. For example, in the waste/water management case study (CS1), the model's aim is to understand and predict how the system' variability affects the H^2AD bioprocess performance and make predictions how the H^2AD bioprocess will perform on new feedstocks. The performance of the process is defined as the percentage removal of wastewater pollutants and volume of biogas generated. The model's goal is to make predictions using feedstock and the H^2AD process data to predict the removal of key pollutants and volume of biogas generated. These two goals may come into conflict with one another; for example, if pollutant removal is favoured over biogas generation. In the brewery case study (CS2), the model's aim is to predict when fermentation is complete and the beer has reached desired alcohol concentration and flavour profile. By doing so it will minimise over-fermentation, which wastes resources and results in an inferior product. This defined the model's goal of developing a model that could predict the alcohol concentration in real-time from affordable ultrasonic and temperature measurements.

4.5.2 UNDERSTANDING THE REQUIREMENTS FROM THE MODEL

A model's goal will inform on what function the model is to perform and what degree of accuracy is required in the model outputs. A DDM may perform two functions:

1. **Fit data** to identify relationships between input, output, process and material properties.
2. **Predict** process performance and resource use.

Dependant on the goal, a DDM may perform only function (1) or it may perform both functions (1) and (2). Statistical tools (linear and multiple regression, curve fitting, etc.) are used to fit equations to data to identify cause-effect relationships between variables [279]. However, their practical application to large-scale problems is limited due to the high dimensionality and nonlinearity of manufacturing systems [137]. Data-driven models are able to model highly complex, nonlinear systems to identify these relationships [279]. Dependant on the manufacturer's requirement, DDMs can also be utilised for predictions [280]. However, the complexity of the model may increase, which will be dependant on the required accuracy of the model's predictions [280]. Table 4.2 gives examples of both functions (1) and (2).

Table 4.2 Examples of data-driven models requirements for the 2 case studies, waste/water management and brewery

	Identify relationships	Make predictions
Waste/water management CS1	To understand how variations in both the feedstock's characteristics and process conditions affect the H^2AD bioprocess. In addition, identify which key variables in both strongly affect performance.	(1) To predict how variations in feedstock will affect the bioprocess performance and optimise the process conditions to maximise performance. (2) To predict the H^2AD's performance when treating new feedstocks to determine if they are financially viable
Brewery CS2	Models data to identify the relationship between temperature and ultrasonic measurements during fermentation to alcohol. Furthermore, to understand how variations in these measurements affect alcohol generation.	(1) Utilises the model to make a prediction of fermentation endpoint for the current batch. (2) To evaluate feedstocks (hops, malt, water, yeast) influence the fermentation time and adjust process conditions to accommodate.

An important consideration for the manufacturer is assessing the required accuracy between the model's output and the true value. Is it important to the manufacturer whether the model is able to make predictions to 99.9% accuracy or is a model with reduced accuracy sufficient to complete the task? According to Dean Abbott, president of Abbott Analytics, it is more desirable to a manufacturer to build a model that meets the requirements for the task at a cheaper cost than to invest more money in unnecessary accuracy [281]. The accuracy of a DDMs is widely defined by the norm of the differences ("residuals") between estimates and observations [282]. The required accuracy will be determined by any one or a combination of the following factors:

1. **Product specification:** when DDMs are utilised to predict a product's specification, the model's accuracy is influenced by the acceptable tolerance around the product specification. The tolerance may be determined by regulations, customer demands and/or economic feasibility.

2. **Regulation(s):** manufacturers are subject to regulations regarding the accuracy to which they state the composition of their product, co-product and waste streams. For example, beers of an alcoholic strength not exceeding 5.5% volume that are sold within the European Union are subject to a 0.5% tolerance [283].
3. **Economic:** when predicting an output that has a strong correlation to the economic feasibility of the process. Dependant on how precarious the process is between profit or loss making will influence how accurate the DDM is required to be. For example, when predicting the product yield of yield-driven processes. Yield-driven processes are productions processes whose economic feasibility are substantially influenced by-product yield.
4. **Safety:** when making predictions that influence the safety of the manufacturing system the DDM's accuracy is influenced by the severity and likelihood of the hazard. For example, when predicting hazardous gas dispersion under complex terrain conditions a high level prediction accuracy is required [284].

4.5.3 DATA CONSIDERATIONS

Data-driven models can be harnessed to support the development and implementation of new technologies and processes [285]]. However, this requires the generation of a set of data, which can be costly and time-consuming [26]. The data consists of collected input and output data. The data is split into "model development data" and "unseen data". As the name suggests, the model development data is data used in training, validating and testing DDM, and unseen data is data used to evaluate the predictive capabilities of the final model [100]. The model development data is further partitioned into training, validation and, in the case of artificial neural networks (ANN), testing data [137].

There are multiple methods by which this is done dependant on the field [286]. The most popular methods in machine learning are holdout validation and cross validation [286]. In holdout validation, the data is split 50-70% to training data, while the remaining is validation data [116]. Training data is the data fitted to the model's algorithm and validation data provides an unbiased evaluation of the model's fit to the training data and is used for tuning the algorithm's hyperparameters. A hyperparameter is a parameter set before the training stage; for example, the number of hidden nodes in an ANN hidden layer [100]. Holdout validation is used when a large amount of data is available and the computational time would be too great to perform multiple testing. Cross validation is for medium and small sized datasets, the data is split into 80% (or 90%) training data and 20% (or 10%) validation data 5 (or 10) times (folds) [116]. The

model training and validation stages are then repeated for each fold and the average accuracy of all models is given as a final performance result. This helps to avoid any potential bias that could be introduced by using holdout validation [116]. The definition of what is a large, medium or small dataset varies between disciplines. In the field of image recognition a dataset of approximately 28,000 data points was considered small [287]. Whereas, holdout validation has been the common approach in process manufacturing datasets with smaller than 100 data points [115,117].

Future process manufacturers should consider partitioning their model development data by cross validation, as those who have achieved positive results [113,116]. Models built from artificial neural networks (a sub-field of machine learning) traditionally split the model development data into three categories; training, validation and testing data [137]. The majority of approaches split the data randomly 70, 15 and 15% respectively [137]. The testing data is used to provide an unbiased evaluation of the final model's fit on the training data. The unseen data is data that the algorithm has never seen during the development of the model [100], and its purpose is to measure how well the model fits new data and how accurate the model's predictions are. Evaluating the model on unseen data is essential to avoid overfitting and ensure good generalisation, yet this stage has often been overlooked [119].

Ensuring the model development data and unseen data is balanced and representative of the system is paramount to the performance of DDMs [288]. To achieve this there are five questions to consider when generating a set of data:

1. What data already exists, to what granularity (the scale and level of detail in the data) and how representative of the system is the data?
2. What volume of data is required for model development?
3. How was the data collected?
4. What, if any, additional data will the model require and how will it be collected?

Manufacturing companies today are collecting vast troves of process data but typically use them only for monitoring purposes and after the event analysis, and not as a basis for improving process resilience [26]. As part of the data consideration stage, the modeller will need to decide on the optimal number of input and output variables the model will require. This is achieved through feature selection and dimensionality reduction techniques, which improve the model's performance by identifying and removing noisy and/or irrelevant variables from the data [26]. Irrelevant variables are variables that have little to no impact on the model's output. Additional data of variables not currently measured may be required for the model. For example, in CS1 the H^2AD

bioprocess was already collecting 7 process performance variables but no data on the characteristics of the incoming or outgoing wastewater streams. Nineteen water quality variables were identified for initial collection, though with appreciation that not all of these would be truly relevant to modelling the output data. This made for a total of 26 input variables that was reduced to 9 by feature selection followed by principal component analysis. The volume of data is determined by three factors.

1. The data needs to extend to, and preferably beyond, the system boundaries. Testing the model's capability to fit and predict beyond the system's boundaries is important to establish the confidence boundaries of the model.
2. The duration period of data collection must be sufficient to capture any temporal variations observed in the system.
3. There must be an equal distribution of the data within the system boundaries, to ensure the model is capable of fitting data and making predictions throughout the system.

If it is possible to collect data that achieves these three factors efficiently, then the volume of data required will be less. This can be done using design of experiments (DoE) techniques, such as central composite design and Box-Behnken design, to systematically collect data from across the system [119]. However, collecting data in industrial environments has numerous challenges and a less intrusive method may be required.

Data is collected either manually or automatically. Data collected manually is time-consuming and introduces the possibility of human error [289]. Whereas, data collected automatically generally increases the volume of data recorded and removes human error [289]. However there is a capital cost in purchasing and installing the required systems to collect the data automatically. Data-driven models often benefit from the automatic collection of data, as they can react in real-time to changes in the input variables. The DDM's response can feed into a larger online system creating a feedback loop. When automatically collecting data on certain variables is not possible, DDMs may utilise cheap sensors to predict the value of the variable. This is shown in CS2 which uses a DDM to predict alcohol concentration from automatic ultrasonic and temperature measurements.

However, automatic data collection comes with challenges associated with handling large datasets. When handling a large dataset, the five V's are often employed to characterise the data [290]:

1. **Volume:** refers to the size of the datasets. Manufacturers have an ever-growing amount of data generated across the company, in the terabytes or even petabytes size, and require the infrastructure to process the large volume of data [217]. A multinational manufacturer is capable of producing terabytes of data every second [291]. For example, in a consumer packaged goods company that produces a personal care product, a single machine alone generates 5000 data samples every 33 ms, resulting in four trillion (i.e., 4 Tera) samples per year [106].
2. **Velocity:** is the speed to which data is generated, collected and analysed.
3. **Variety:** is defined as the different types of data incorporated into the datasets. Structured data (data that is recorded in tables) is being replaced with unstructured data and 95% of the world's data is now unstructured data [292]. Currently, process manufacturing data is still highly structured [217]. However, with the introduction of new data sources from novel sensors integrated across the supply chain, the volume of unstructured data (texts, image, audio, video, log files, consumer feedback, etc.) will continue to increase.
4. **Value:** refers to the value extracted from the data. Collecting and storing data can be expensive [293]. It is not economical for manufacturers to measure variables that have little to no impact on the process performance or ensuring product specifications are met.
5. **Veracity:** perhaps the most important characteristic from a modelling point of view, the trustworthiness and quality of the data. If a DDM has been trained on untrustworthy, data all outputs from that model are also untrustworthy.

4.6 CHALLENGES FOR DATA-DRIVEN PROCESS MANUFACTURING MODELS

Process manufacturers face unique challenges when collecting a set of data that is representative of the manufacturing system. The machine learning techniques utilised by DDMs have increased capabilities to fit data and make predictions when more data is provided [100]. However, when ensuring there is sufficient data, and that it is representative of the system, there are a number of challenges in:

1. The availability of manufacturing data;
2. The variability of manufacturing systems;
3. Establishing the system boundaries;
4. Validating the model's output.

4.6.1 AVAILABILITY OF PROCESS MANUFACTURING DATA

Data collection from real industrial processes faces a plethora of well-known challenges: sampling frequency, spatial representation of process environments, incomplete data, working conditions, sensor malfunction, communication exception or database shutdown, accuracy, etc. [294,295]. This can result in poor quality data for model development that may require cleaning. Cleaning may be necessary to remove outliers within the data that are not representative of the system; however, the challenge is in knowing what data is and isn't representative of the system. Although DDMs developed from machine learning algorithms can overcome these challenges, large volumes of data are required [217]. This is not always possible from process manufacturing systems, as they may be subject to additional challenges concerning the availability and variability of data. An established process manufacturing system may already have a wealth of historical data available for the initial training stage. However, this may not always be the case and the reasons for this may include:

1. Modelling a new or adapted process, meaning no historical data representative of the current process exists;
2. Variables (e.g. process conditions, external influences, new or changing feedstock) required for the model were not previously measured;
3. Manufacturers may not have stored their data or not stored their data in a useable format;
4. Manufacturers may produce a variety of products, meaning data on any one product is limited;
5. Manufacturers may not keep potentially damaging data to the business if not required to.

Furthermore, industries that produce a variety of products from the same set of processing equipment (common in the pharmaceutical and food and drinks industries) will require the creation of a dataset for each of their products; as the process conditions, feedstocks and resources will be specific to that product. This data may not be immediately available, as these industries tend to produce a batch of one product then switch to another. This means the time between batches of the same product can be extensive delaying the development of the model. The CS2 faced these challenges, where craft breweries tend to ferment a range of different beers and a model would need to be developed/adapted for each beer. The beer in question for this study is currently only brewed once a month so this means only 12 data sets could be recorded each year, which may not be sufficient for the machine learning models. Equally, for CS1 a new

dataset will be required for optimising the H^2AD bioprocess's performance on new feedstocks (e.g. a different farm's waste, food waste).

4.6.2 VARIABILITY IN PROCESS MANUFACTURING SYSTEMS

All manufacturing systems contain inherent variation (e.g. fluctuations in process temperatures, pressure, and flowrates, human operators, leakages) which can affect the performance of a DDM. This variability may be overcome by collecting sufficient quantity of relevant data as the variability will be captured within the model [296]. However, this is not always possible for process manufacturers (see section 3.1). Process manufacturers face the further challenge from the variations present in the resource flow, particularly when the resource flow is a waste stream. The feedstock physical and compositional variability can have a significant impact on the biochemical and thermochemical conversion to the final product [297]. This variation can occur for a number of reasons, as summarised in Table 4.3.

Table 4.3 Causes of variation in process manufacturers' feedstock and some industry examples

Cause of variation	Industry example
Processes upstream	Water is an essential brewing ingredient. Inherent variability in the water treatment methods employed upstream has a direct impact on the water's characterises, and thus the final beer's characteristics [298].
Changes in operating practices upstream	Farming practices vary throughout the year. For example, the practices of spreading the farm waste stored in a slurry tank increases between the months of April and September, as per the EU waste framework Directive 91/676/EEC [299]. This affects the slurry's characteristics (including total suspended solids, metal content and nutrient and organic load) and impacts processes utilising it as a feedstock, e.g. the H^2AD bioprocess.
Changes in supplier	Anaerobic digestion is an attractive option for the valorisation of food waste and other wastes with high organic load and calorific value. However, instability of anaerobic digesters is a common problem that can be exasperated by changes to the feedstock [300]. Food waste characteristics vary hugely dependant on its source, as well as being inherently variable [300]. Introducing a new feedstock to a digester can reduce biogas generation as the system takes time to adapt [301].

Local Agronomic Conditions	Geographic location affects feedstock characteristics through variations in local agronomic conditions [297]. The structure of corn harvested in the US has been shown to have a stronger correlation to geographic location than genetic variety [302].
Cultivation and harvesting practises	There is extensive exploitation of plants for active ingredients for drug development [303]. The active ingredient derives their therapeutic effects from secondary metabolites, which is influenced by numerous natural factors [303]. The timing of harvesting and/or handling of the plant material also has an impact on plant quality [303].
Seasonal	Sewage sludge generated by wastewater treatment plants is often used for agriculture as it recycles nutrients and organic matter to land. However, there is an environmental threat from the heavy metal content in the sludge [304]. The heavy metal content has been shown to vary by seasons and understanding this variability is important to minimise any negative environmental impact from using the sludge as a fertiliser [304].
Storage and transportation	Potatoes degrade after they are harvested, at a rate determined by the storage: temperature, relative humidity, air circulation and gas composition [305]. During transportation, the potatoes' quality is further reduced by bruising [305]. Food manufacturers receive potatoes from multiple suppliers. Therefore, classification is required to determine if a potato quality is sufficient for further food processing [261].

During CS1, the H^2AD bioprocess was affected from variations in the feedstock supply, caused by changes to farm practice. There is an inherent variation present in the feedstock, as shown by the fluctuations in concentrations of four key water quality parameters in Figure 4.1. The feedstock is stored in a slurry tank open to the environment, meaning during the winter months the total suspended solids (TSS) would decrease as the feedstock became diluted by increased rainfall.

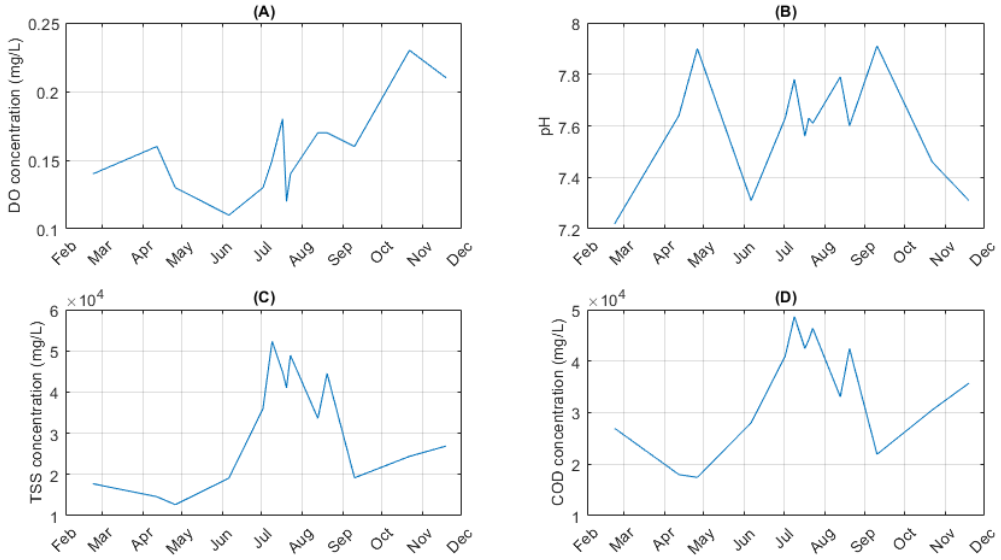


Figure 4.1 Concentration of four water quality parameters of H^2AD feedstock during 2018: (A) dissolved oxygen (DO), (B) pH, (C) total suspended solids (TSS) and (D) chemical oxygen demand (COD).

There are further problems that arise from unexpected variability between batches of the same product. The brewery CS2 helps to illustrate the variability within supposedly identical processes. Figure 4.2 shows the ultrasonic velocity, received ultrasonic signal amplitude and temperature as a function of fermentation time for four batches of a particular beer. Each of these measurements was recorded during different months in 2018, and are the input variables for a DDM predicting alcohol concentration during fermentation. The results indicate that the majority of variation within the signal is caused by temperature variations, which is typical of ultrasonic measurements. However, although the ultrasonic results show the same overall trends there are other variations caused by feedstock variability and marginally different brewing conditions (e.g. volume in the fermenter). These will affect the performance of the machine learning models and affect their accuracy. In Figure 4.2, batch 4 shows a totally different temperature and speed of sound profile to the other batches. This was the result of a failure of the temperature control system during fermentation. Therefore, this batch was not representative of the system and is not suitable for training the model, further limiting the data available for model development.

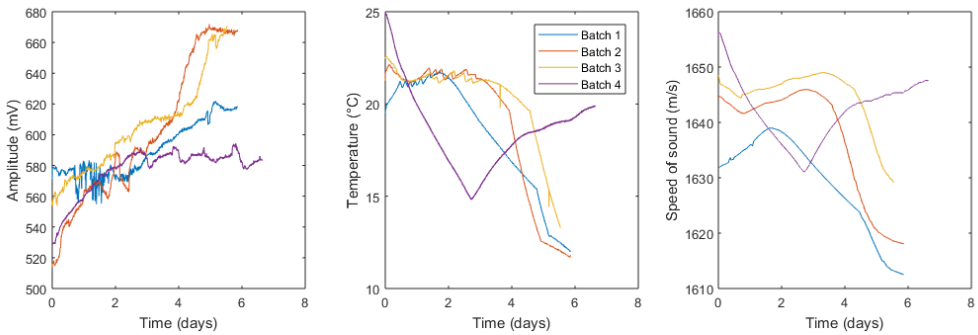


Figure 4.2 Amplitude, temperature and speed of sound data during fermentation of four batches of Slap in the Face beer, collected from Totally Brewed in 2018.

4.6.3 RECOGNIZING THE MANUFACTURING SYSTEM'S BOUNDARIES

Predictions are a key feature of DDMs for process optimisation and intelligent decision-making. Data-driven models' predictive capabilities are generally strongest within the boundaries of the data used to develop them [100]. If the training data is only representative of a subsection of the manufacturing system, the model will fail at making accurate predictions of the entire system. Figure 4.3 helps to visualise this problem. The model will be able to make accurate predictions for data inputted that is within the blue box but its prediction capability will likely decrease in accuracy the further away from the model boundaries. There are two methods to ensure the manufacturer can be confident the in model's results:

1. Knowing the system boundaries and ensuring the training data extends to these boundaries;
2. Knowing the model boundaries and ensuring that only predictions made from this within the training data range are acted upon.

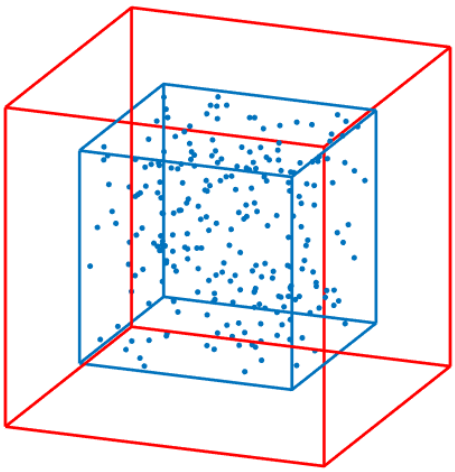


Figure 4.3 Figure showing training data that is only representative of one part of the whole manufacturing system. Where blue dots are training data, the blue box represents model boundaries and the red box represents system boundaries.

Knowing the system boundaries of a process usually requires prior knowledge elicited from process industry practitioners [295]. However, when modelling a new or changed system this knowledge may need reinforcement from trends identified within the training data. Data visualisation plays an important role when understanding the space the system operates within [306]. It is simple to visualise the boundaries' of a system of two or three dimensions, as demonstrated in Figure 4.3. However, process manufacturing models typically contain data of higher dimensionality that will be impossible to visualise using conventional two-dimensional Euclidean space [307]. There exist a number of approaches to plotting high dimensional data [308]. Approaches like the scatter plot matrix can be utilised to plot the individual relationships between the variables alongside one another [308]. The parallel coordinates system representation is particularly useful in visualising process manufacturing data in one plot [307]. Wang *et al.* developed the Scan Circle algorithm that is able to identify regions of interest within the data that can be used to help describe the feasible region a DDM may operate successfully [309]. Once the modeller is confident in the boundary conditions the training data is selected from the historical data, ensuring that the training data extends to these boundaries. A scatter plot matrix was used to understand the model boundaries for the feedstock characteristics in CS1, a section is shown in Figure 4.4. The diagram clearly shows that TSS is between 10,000 and 20,000 mg/L for the majority of the training data, implying that predictions for when TSS is greater than 20,000 are likely to be less accurate. It also indicates that there is a strong positive correlation between TSS and chemical oxygen demand (COD).

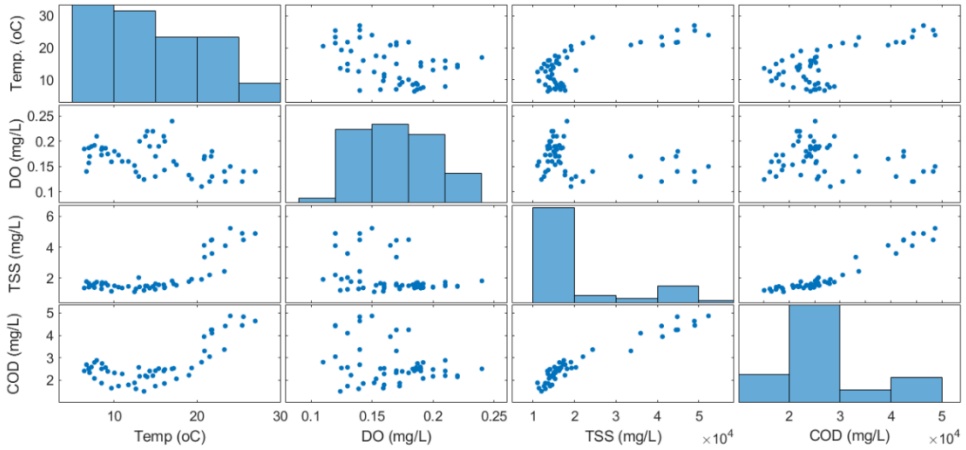


Figure 4.4 Scatter plot matrix for four of case study 1's feedstock's characteristics [Temp.: temperature; DO: dissolve oxygen; TSS: total suspended solids; COD: chemical oxygen demand]

If no historical data exists, then data needs to be collected and it is likely that the system boundaries are yet to be defined. This can be overcome by collecting data over a suitable number of batches and/or timeframe of a continuous process, to be confident that the system is modelled within this data. This may be a time-intensive approach, as data may need to be collected for up to or beyond a year. For example, the process may be influenced by seasonal variations. The performance of DDMs may heavily deteriorate because of the outliers that industrial processes commonly involve [295]. Therefore, data visualisation techniques are required to define the boundaries of the model. An alternative method is to conduct a DoE, to determine a robust set of experiments that will define the system limits and generate data to extend to these boundaries [119]. By using DoE the data collection time and volume of data may be greatly reduced; however, the manufacturer may face additional costs from process disruptions necessary for experiment execution to achieve a robust dataset [26]. Dependant on the potential economic benefit of the model, it may be worth the additional cost or it may prove more feasible to conduct pilot-scale work to generate the data.

When implementing a new model into a system, careful monitoring is initially required to ensure the model is reflective of the system. Particularly if there is variability in the feedstock and other external factors. Once successfully installed, the model requires monitoring to prevent a gradual degradation in performance. The decrease of the prediction quality is caused by the gradual and abrupt changes in the process [310]. This

can be avoided by retraining the model periodically as the availability and collection of data increases.

4.6.4 EVALUATION OF THE MODEL'S OUTPUT ON UNSEEN DATA

Model evaluation is necessary to determine how accurately the model reflects the system [100]. As proving absolute validity is a non-attainable goal, achieving a high degree of *face validity* is the aim. Face validity means that, from statistical analysis, the model appears to be an accurate representation of the system. If the accuracy of the model does not meet the requirements defined by the manufacturer the model must undergo redevelopment either through optimisation of the algorithm's architecture or the collection of additional data [311].

Data-driven models can overfit the data, if the model too closely matches the training data, meaning that they can identify and exploit apparent relationships in the training data that do not hold in general. To determine a DDM's face validity, the model is tested against unseen data [100]. Unseen data is data that has not been used at any stage in model development. Like the training data, it consists of a set of known input and output variables. Unseen data has two sources:

1. **Partitioning model development data:** split data into model development data (training, validation and testing data) and unseen data.
2. **Experimental data:** an experiment is performed to generate unseen data that covers the range of the model boundaries. Design of experiments can be used to ensure the parameter range for the experiment is comprehensive.

Partitioning data is the most common approach [286]. The data partitioning can be performed randomly or by using a fixed method. However the data is partitioned, the unseen validation data must extend evenly to cover the boundaries of the model. This is to ensure the model is able to both fit and predict data through the manufacturing system. The unseen data should also aim to go beyond the limits of the model. By using data beyond the model boundaries the modeller gains knowledge of the model's capability to fit and predict outside of the training data. An example of good and poor selection of unseen data is shown in Figure 4.5.

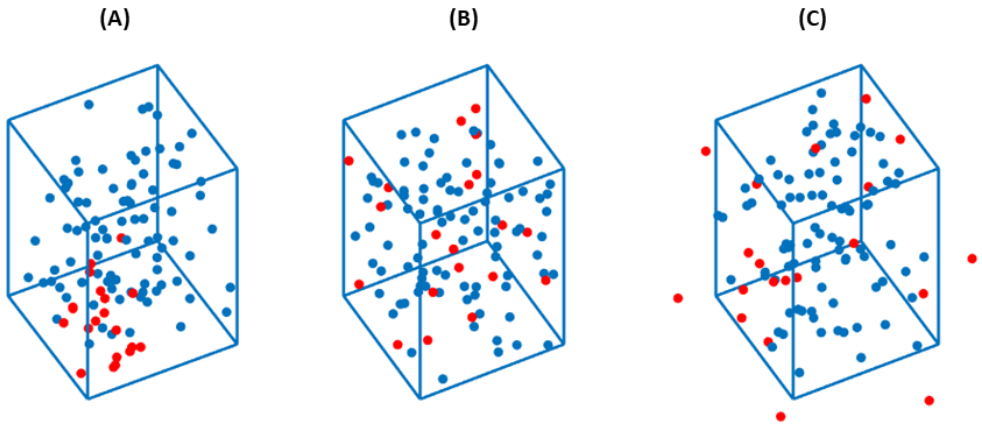


Figure 4.5 (A) Example of unseen data only evaluating one region of the model, (B) example of unseen data evaluating the whole of the model, and (C) evaluating the whole of the model and the extrapolating capabilities of the model. Where blue dots are training data, red dots are unseen data and blue box is the model boundaries

4.7 OPPORTUNITIES ARISING FROM DATA-DRIVEN MODELLING OF PROCESS MANUFACTURING SYSTEMS

The application of DDMs to support and improve traditional modelling tasks has been well established. Examples exist of their use for design and planning, control and monitoring, optimisation and safety [107,312–314]. Furthermore, techniques are being designed to integrate data-driven and first principal models into a hybrid model structure [97]. Hybrid models are first-principal models supported by DDMs to describe part of the system that is not understood, and impossible to model from a first principle approach. These hybrid models have shown to yield better results than the single modelling approach [97]. While it is important to note these applications of DDMs, what is of interest in this study is the new opportunities that may arise from a data-driven modelling approach. Especially those that are unfeasible from a solely mechanistic modelling approach. In an age where manufacturing is under pressure to reduce cost and environmental impact, while maintaining product quality, the utilisation of data-driven modelling is key. When the Aberdeen Group surveyed 223 global manufacturing organisations, 47% believed they needed to become more data-driven to remain competitive [315].

The complexity and variability of feedstocks and processes within a process manufacturing system present challenges in creating consistent high-quality products. However, their complex multi-component nature, particularly in the waste streams, presents a variety of different valorisation and cost reduction opportunities. To enable DDMs to determine the best route to increase process resilience and/or waste valorisation in real-time, it will require the successful online characterisation of these process streams.

4.7.1 ENABLING AFFORDABLE CHARACTERISATION OF PROCESS STREAMS

One of the main benefits of DDMs is their ability to enable real-time decisions by the collection of real-time manufacturing data [316]. However, process manufacturing's feedstock, product and waste streams are not a set of specific discrete components. Instead, they are a multi-component and/or multi-phase matrix subject to inherent variation. Therefore, real-time online characterisation is expensive and sometimes unfeasible. This results in important process variables, like product quality, being measured infrequently offline [317]. Because of the infrequency in which these measurements are performed, continuous direct monitoring of process streams' composition is not possible [112].

Industrial processes are often described as *data rich but information poor*. There is often a large quantity of process data from conventional measurements (pressure, temperature, flow rate, etc.) but a lack of data on stream compositions. This is due to a lack of suitable and affordable technologies capable of performing these measurements non-invasively, online and in real-time. [318]. To overcome this, process manufacturers rely on soft sensors, which are able to model data collected from conventional measurements and used to predict key variables [319]. As more novel industrial internet of things (IIoT) technologies are introduced, the variety of data available to train soft sensors shall expand. One such example is a multi-sensor system designed to optimise the clean-in-place (CIP) process, which utilises both visual and ultrasonic data [237,238]. By utilising DDMs, soft sensors are able to compute not only a wider range of data but also unclean data [217]. This will allow for the development of cheaper sensors, as the requirement for high-quality data shall decrease, instead being replaced by volume of data [217]. However, the data must always be representative of the system.

Being able to track the relevant composition and physical characteristics of a process stream in real-time, will become even more important to process manufacturers as regulations may change so that manufacturers have more responsibility to record and share data for compliance. Combating fraud in certain process manufacturing industries is an ongoing issue, with a particular focus on the food and drinks sector [320]. The introduction of a greater number of affordable soft sensors would help build trust across the supply chain and with consumers. If this is to be successful, ensuring that these sensors are not susceptible to malicious interference or corrupted with false data is paramount.

4.7.2 ENSURING PROCESS RESILIENCE

As with all suppliers, process manufacturers are required to meet certain targets to be commercially successful and adhere to regulations. Paramount among these, from a manufacturing process viewpoint, is meeting minimum product specification (e.g. yield and quality) [321]. Ensuring consistency for these targets is challenging as process manufacturers are subjected to a higher level of variability than other industries [321]. For example, in the biopharmaceutical industry yields can vary from 50% to 100% for no immediately discernible reason [26]. In case study 2, the unexplained variation between batches, shown in Figure 4.2 (see section 4.6.2), results in disparities in the final alcohol concentration in the product. By taking observational data (meaning data routinely collected by a manufacturing process) and developing a DDM, the variables that are having a significant effect on the industry yields can be identified and optimised [26]. The advantage of using data-driven techniques this way is that the DDM can intelligently identify significant variables responsible for the variability. This is achieved by using unsupervised machine learning techniques for exploratory data analysis to find hidden patterns or grouping in data [137].

Although it is possible to model and limit the impact of some variability within a processing plant [26], a substantial challenge for process manufacturers is limiting the impact of variables outside of an engineer's control. Paramount of these variables is the variability in the process's feedstock. This variability can have a significant impact on product quality and yield. For example, the production of anti-malarial drugs which is reliant on the active component artemisinin harvested from a plant called *Artemisia annua* [111]. The plant's artemisinin concentration varies dependant on when it is harvested during its lifecycle and the storage conditions [111]. By fitting historical harvesting, transportation and process data, a DDM would determine the optimal harvesting and storage point to maximise artemisinin recovery. The model could also be used to predict the feedstock's potential to produce artemisinin and whether to adjust the price for the supplier as not all feedstock is created equally.

Variation in the feedstock is not the only concern for a process manufacturer. The waste produced by process manufacturing is equally variable, in regards to the volume produced, the concentration of valuable components and pollutants [322]. Governments and regulatory bodies around the world set maximum limits on what manufacturers may emit to the environment [323]. Manufacturers may choose to treat their waste onsite, so it is within these limits, pay a third party to handle their waste or partially treat their waste to reduce cost. There is a cost involved with all options and this cost may fluctuate with waste characteristics. Being able to predict and accommodate these variations

ahead of time may mean measures can be taken to reduce this cost. Process manufacturing waste is unique from discrete manufacturing in that there exists the potential for waste valorisation [244]. However, one of the barriers to waste valorisation is the detrimental effect variation in the waste has on the technologies in terms of performance. Case study 1 is an example of a waste valorisation technology that utilises a DDM to increase the resilience of the bioprocess. By collecting a set of training data over a 1-year period and fitting it to a model, it was possible to capture the variability of the feedstock due to seasons, changes in farm practices and inherent variation. The model is able to identify which key feedstock characteristics variations were detrimental to the bioprocess's performance. The bioprocess's process conditions were also varied throughout the year, in order to investigate their effect on performance. By combining and fitting this data to a model, it was possible to predict how to optimise the bioprocess conditions dependant on the incoming feedstock characteristics.

4.7.3 EVALUATING WASTE VALORISATION POTENTIAL

Increased competition for access to critical resources is a major concern for the manufacturing industry. In 2017, the EU expanded its list of critical raw materials (CRM) to 27, defined as materials considered to be of high importance to the EU economy and of high risk to their supply [5]. Greater focus has been placed on developing solutions towards implementing the circular economy into manufacturing [243]. The circular economy is being pursued by Governments, with Europe leading the way by implementing the 2018 Circular Economy Action Plan to create a set of measures to help manufacturers to adopt circular economy systems [324]. Encouraging manufacturers to stop viewing waste as a *waste* but instead as *co-product* is an important step towards this goal.

Process manufacturing generates a wide range of solid, liquid and gas waste streams that have excellent potential for waste valorisation within a circular economy. Waste valorisation refers to industrial processing activities aimed at reusing, recycling, or recovering useful products or sources of energy from waste [325]. There are currently three waste valorisation strategies, geared towards (1) production of fuel and/or energy to replace fossil fuels; (2) production and/or extraction of high-value chemicals from residues; and (3) production of other useful material(s) [244].

Process manufacturers have a variety of options when evaluating which waste valorisation route to take. Choosing which route to take depends on a number of factors: composition, volume and variability of the waste stream, capital and operational costs of onsite treatment versus valorisation by third party, pre-treatment costs, other waste

being produced by nearby manufacturers, handling and transportation costs. Due to the large number of factors involved, it can be time and cost-intensive for a manufacturer to employ someone to evaluate the most sustainable route. There exists a number of enterprise resource planning (ERP) programs, that use DDMs, capable of simulating the manufacturing process to intelligently determine the most sustainable waste valorisation route [326]. These platforms are available via the cloud and systems like IBM Watson IoT are capable of integrating the manufacturer's existing ERP programs into their software [165]. However, while these platforms are capable of planning the logistics and costs, it has not been the focus to predict the performance of these different waste valorisation technologies at treating a waste stream. There is a great opportunity to have models incorporated into the process to take a holistic view, considering waste minimisation and/or valorisation alongside the key product objectives.

The sharing of data via a cloud platform could be used to support DDMs' ability to predict how a valorisation technology may perform on a new waste stream. Manufacturers are naturally reluctant to share their valuable data with competitors. To overcome this barrier, a third party would utilise a model built from data collected from another's manufacturing process to make predictions. Although, it should be noted that the data would have to be relevant to the system under investigation to provide meaning results. For example, manufacturer A is producing 'X' as a waste stream and wants to know if any technology is available to add value to the waste stream that would otherwise cost to manage. Manufacturers B and C are valorising waste streams with similar characteristics to X but using different technologies. Both B and C are also using DDMs uploaded into a cloud platform to monitor the performance of their processes. The cloud platform may utilise B and C's models to make an estimation of how the respective technologies may perform when processing X. This prediction can be used by manufacturer A to inform which technology would be most suitable to valorise A. This service would be available through a cloud platform sold as a service with a proportion of the income going towards manufacturers B and C as an incentive to share their models.

4.8 CONCLUSIONS

- The increasing availability of data, due to the adoption of low-cost industrial internet of things technologies, coupled with increasing processing power from cloud computing, is fuelling increase use of data-driven models in manufacturing. Utilising case studies from the food and drink industry and

waste management industry, the considerations and challenges faced when developing data-driven models for manufacturing systems are explored. The considerations and challenges outlined in this work can then be used as a protocol for future modelling of process manufacturing systems.

- This work identified three consideration for process manufacturers before modelling their systems. The model's goal must be clearly defined and should aim to solve the manufacturing problem that has been identified. The manufacturer should also consider what is required from the model. Is it enough simply to fit the data, or will the model be required to make predictions? Being able to make predictions increases the model's value but also its complexity. The accuracy in which the model is required to perform these functions should be defined by the product specification, regulations, economic value and safety requirements.
- This work identified four challenges faced by process manufacturers when modelling their systems. Challenges arise when ensuring a representative data set is used for model development. Process manufacturers face unique challenges regarding this point because of the limited availability and variability of industrial data. Even when data is available, knowing if it representative of the system and relevant to the model can be challenging. Furthermore, the boundaries of a process manufacturing system are hard to classify. The difficulty occurs in knowing when enough data has been collected to be representative of the whole system.
- Once a manufacturer is able to overcome the considerations and challenges involved when modelling a process manufacturing system, DDMs will allow the process manufacturers to characterise process streams, enhance process resilience and evaluate waste valorisation.

CHAPTER 5: DEVELOPING NEW PROCEDURES FOR MODELLING PROCESS MANUFACTURING SYSTEMS

5.1 STATEMENT INTEGRATION OF PUBLISHED WORK WITHIN CHAPTER

Integrated into this chapter is material from the published or soon to be published work:

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Data-driven modelling of process manufacturing systems. *Chemical Engineering Journal* (**Submitted to journal**, 03 Oct 2019). [Impact Factor: 6.735; SNIP: 1.863]

Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL, 2019. Multiple target modelling to enable sustainable process manufacturing. *Journal of Cleaner Production*. (**Submitted to journal**, 03 Oct 2019) [Impact Factor: 6.395; SNIP: 2.308]

5.2 HIGHLIGHTS

Highlights from part A:

- Data-driven models (DDMs) can successfully optimise process manufacturing systems.
- Evaluating DDMs extrapolation abilities on unseen data evaluates a DDM's suitability.
- Collecting evenly distributed unseen data from a running process may be infeasible.
- A method for analytically partitioning unseen data from historical data is offered.
- Four DDMs are built to optimise a novel waste valorisation bioprocess.

Highlights from part B:

- Sustainable manufacturing entails maximising economic and environmental objectives.
- Within process manufacturing, these objectives are often conflicting
- Multiple target models can predict multiple process manufacturing system's outputs.
- Two novel multi-target artificial neural network models are proposed.

- Methods to utilise a multiple target model to increase sustainability is then explored.

5.3 INTRODUCTION

The development of a successful DDM starts with understanding the relationships within the model data and recognising whether the data is representative of the process manufacturing system. Chapter 5 builds on previous work to develop a modelling procedure that accounts for the considerations and challenges outlined in chapter 4. Achieving sustainable process manufacturing requires being able to find the manufacturing route that maximises both the economic and environmental objectives. This requires the development of models that can predict multiple targets so that the relationships between a system's inputs-outputs and outputs-outputs can be established and exploited. Chapter 5 is split into two parts. Part A outlining a new approach for modelling process manufacturing systems and Part B outlining how multiple target DDMs can be utilised to increase economic and environmental sustainability.

5.4 PART A: OUTLINING A NEW APPROACH FOR MODELLING PROCESS MANUFACTURING SYSTEMS

Manufacturers are reliant on mathematical models to describe the complex behaviour of their processes [87]. Models have been heavily utilised in process manufacturing, to describe the thermal and bio/chemical processes at the heart of process manufacturing systems [327]. There are fundamentally different two categories of models: first-principal models and empirical/data-driven models. First-principal modelling is the construction of a sequence of equations describing a system's individual components that, when assembled, describe the complete manufacturing system [86]. Data-driven modelling combines the fields of computational intelligence and machine learning, to fit data to algorithms that learn hidden relationships within manufacturing data to make predictions [96]. Deciding which to use is dependent on the level of knowledge known about the manufacturing system and if the manufacturing data is representative of the system. First principal models require a greater level of understanding of the system than data-driven models (DDMs) and are more time consuming to build [97]. However, the accuracy of a DDM's predictions relies on the modelling data being representative of the manufacturing system [101].

5.4.1 THE NEED FOR A NEW APPROACH

The application of data-driven modelling in manufacturing has progressed dramatically over the last two decades [136]. This rise can be attributed to two trends of the 21st Century; (1) cloud computing enabling a substantial increase in computer processing

power and (2) increase in volume and variety of data being generated. Data generation by the manufacturing industry is overwhelming; in 2015 more than 1000 Exabytes of manufacturing data was produced worldwide and by 2025 this number is expected to increase 20-fold [106]. However, there are sectors of manufacturing where data generation and collection are inherently limited. For example, in the food and drink industry, the same manufacturing equipment tends to be used to produce a range of products. This means that time between batches of the same product can be extensive, reducing the data available to develop a model for the production of that specific product. Furthermore, the use of legacy equipment that do not feature as many sensors is limiting data collection.

The widely accepted framework for developing a DDM is the cross-industry standard process for data mining (CRISP-DM) [328]. The CRISP-DM is composed of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The CRISP-DM is an iterative process and each phase may be revisited as more information is made available. Following the CRIPS-DM methodology aims to ensure a representative DDM is developed that has industry relevance and impact [329]. A stated advantage of the CRISP-DM is that it is applicable across various industries [328]. However, process manufacturers face some unique challenges from data availability and variability that the current CRISP-DM method fails to address, as it is has been designed to handle large-scale data mining problems [328]. These challenges may present problems in developing a DDM that is representative of the process manufacturing system.

There is no single definition of a manufacturing system. For this work, a manufacturing system is defined as input streams which pass through a process changing their nature into output streams, which may consist of multiple products and/or waste materials. The input and output streams and the manufacturing process can be monitored by measuring various parameters and these measurements can be used to develop a DDM of the system. Defining the boundaries of the system is challenging because of the variability present within manufacturing systems. When defining the system boundaries, a decision needs to be made on how much of the variability is required to be captured within the system. For example, Figure 5.1 shows a manufacturing system whose boundaries have been defined to capture different levels of variability within the system. By extending the boundaries to capture more of the system's variability results in regions that are not as well defined by the data, which may negatively impact on the predictive capabilities of the DDM. Therefore, how much variability is necessary to

capture varies between manufacturing environments and the level of accuracy required of the DDM.

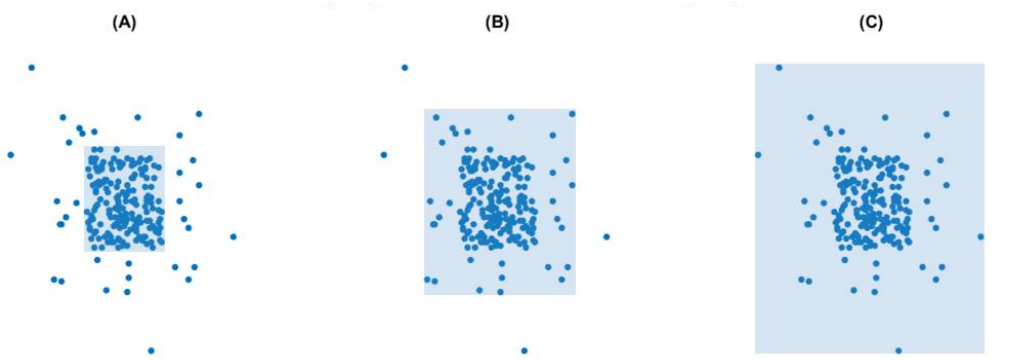


Figure 5.1. Blue region represents the defined manufacturing system for: (A) Strongly represented system, weak variability is captured; (B) Moderately represented system, moderate variability captured; (C) All variability captured but system contains poorly represented regions.

For the model's data to be representative of the system, the data needs to not only extend to the system boundaries but also describe the space within. The majority of DDMs in the literature are developed for experimental or pilot-scale systems [119]; therefore, making it possible to control the system parameters to ensure data is evenly distributed across the whole system. This can be achieved by conducting a design of experiments [119]. However, when attempting to model from an operational manufacturing system it is not always possible to interrupt production and a non-intrusive method is required. An option often used is to train the model on historical manufacturing data [330]. Previous work often assumes that due to the large volume of historical data it is inherently representative of the current manufacturing system and little attempt is given to establish the boundaries of the system within the historical data [330]. Not only is this not possible for manufacturing systems where data is limited, but it also ignores the fact that manufacturing systems are continuously evolving. Manufacturing systems are generally described by high dimensional data increasing the complexities of extracting the key features of the data or visualising the system boundaries [307]. Exploratory data analysis may be performed to better understand the manufacturing system and to assist in defining the boundaries of the system.

There are several guides in the literature for developing DDMs of process manufacturing systems. Curteanu et al. studied three different modelling strategies to study the electrolysis process in wastewater treatment [115]. In this work, the authors demonstrated how a DDM's prediction capabilities may be improved by developing multiple DDMs of a system stacked together [115]. The separate work presented by Charte et al. [116] and Kashani et Shahhosseini [117] both explained the benefits of

effective data partitioning to find a DDM best configurations. Kashani et Shahhosseini also demonstrated that applying dimensionality reduction techniques to datasets with a large number of input variables can improve the performance of DDMs [117]. What is often overlooked in the literature is the importance of evaluating the DDM's prediction capabilities on unseen data [118]. Unseen data is data that has not been used during the development of the DDM. Traditionally testing data is partitioned from the model development data and used to evaluate the model's performance on future data and to avoid overfitting [100]. Overfitting is the generation of a model that corresponds too closely or exactly to the noise (error) within the dataset, which negatively impacts future predictions [141]. However, using only testing data to evaluate a model's prediction capability on future data is not sufficient for process manufacturing systems that face data availability and variability challenges. Unseen data is also required to evaluate what portioning of the manufacturing system does the model represent and the model's extrapolating capabilities beyond the model development data. Hamid et al. showed how additional experiments may be performed to generate unseen data to evaluate a model extrapolation capabilities beyond the model development data [119]. However, when modelling historical manufacturing data this may not be possible. A method is required to partition unseen data from the historical data that can be used to thoroughly evaluate the DDM's model predictive capabilities throughout and beyond the system.

This research proposes a methodology for modelling process manufacturing systems that face data availability constraints. To achieve this, it is proposed to add two new steps to the modelling phase of the CRISP-DM methodology, as shown in Figure 5.2. The first is to use exploratory data analysis to clearly define the system boundaries and the spread of data within the boundaries. The second is to use unseen data to evaluate how well the developed model captures the variability of the process manufacturing system. The proposed methodology is demonstrated by developing a DDM for an industrial case study that faces data availability and variability challenges. The case study aims to develop a DDM describing the novel H^2AD bioprocess developed by Lindhurst Engineering Ltd., which is described in section 5.4.2.1.

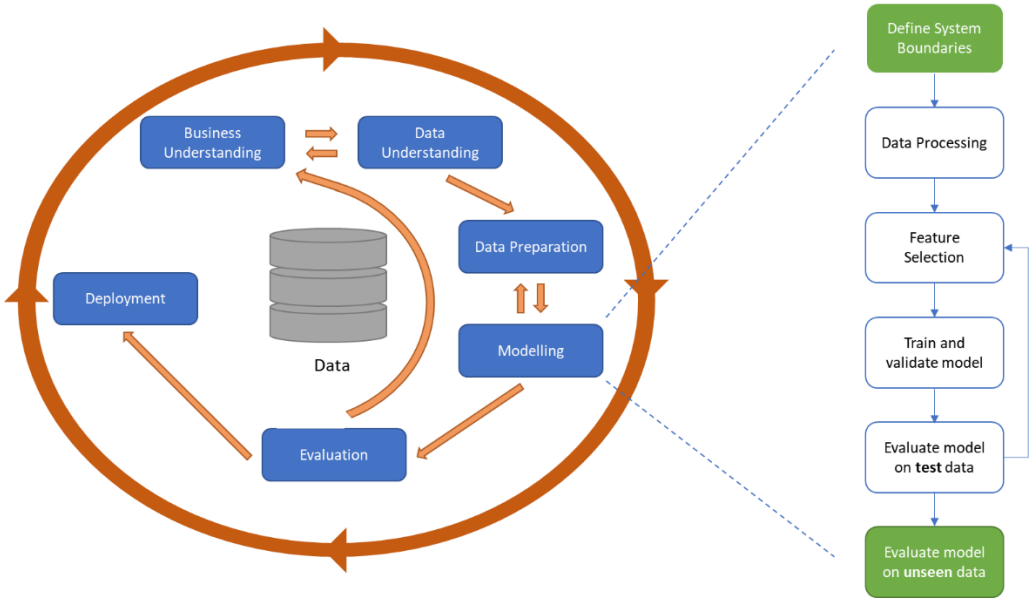


Figure 5.2. The proposed CRISP-DM methodology has been expanded to include two new steps (the green boxes) in the modelling stage.

5.4.2 MATERIAL AND METHODS

5.4.2.1 Process manufacturing system

The manufacturing system studied in this work is a bioprocess developed by Lindhurst Engineering Ltd targeted at treating a SMEs manufacturer's wastewaters, due to its modular, low-cost design. This case study is summarised in section 2.8.

The performance of the H^2AD bioprocess varies with changes in the wastewater characteristics (temperature, pH, composition, etc.). The performance is defined as the percentage removal of wastewater pollutants and volume of biogas generated. This project aims to develop a data-driven model to understand how variations in the input wastewater stream and H^2AD process conditions inform on the bioprocess's performance. The model is trained on data collected from an operating H^2AD plant currently treating wastes produced by a 300-cow dairy farm. The farm waste contains cattle slurry, bedding waste, waste milk, footbath, parlour washing and rainfall. The model may then be used to:

1. Analyse and predict the effect that varying wastewater characteristics (e.g. pH, chemical oxygen demand, total suspended solids) has on the bioprocess performance.
2. Predict the bioprocess's ability to handle new feedstocks.

The H^2AD system was selected to be utilised to demonstrate the methodology developed in this paper because of the unpredictable variability in the H^2AD's feedstock (farm waste). This uncertainty poses challenges in defining the boundaries of the H^2AD system and collecting data that is representative of the system. When this is not possible,

it is imperative to determine what portion of the system is represented in the data in order to define the model's prediction boundaries. Furthermore, the input and output data are collected manually from weekly samples limiting the available data to develop a DDM on**.**

5.4.2.2 Modelling data

The DDM is trained on a supervised machine learning algorithm. Supervised learning takes a known set of input data with known responses to the data (output data) and seeks to build a predictor model that generates reasonable predictions for the response to new data [137]. When describing the H^2AD system, the model input data is collected from two sources; water quality analysis (WQA) performed on H^2AD feedstock samples and H^2AD process data. Eighteen water quality variables are manually measured from samples taken to assess the slurry and 4 H^2AD process variables, these are described in Table 5.1. The H^2AD system automatically collects data on hydraulic residence time, pump speed, system temperature and pressure. These variables have been incorporated into the model to investigate their effect on the H^2AD performance.

Table 5.1 Description of model's input variables (sourced from [331]) and relevance to H^2AD bioprocess

Input variable	Unit	Definition	Relevance to the H^2AD bioprocess
pH	--	The negative logarithm of the hydrogen ion concentration.	The pH of the wastewater significantly affects the performance of BES. For pH ranges pH < 3 and pH > 10 the BES performance deteriorates [332].
Dissolved oxygen (DO)	mg/L	The amount of free, non-compound oxygen in the water. The presence of DO in water is desirable to prevent the formation of noxious odours and sustain aerobic lifeforms.	Increases in DO concentration in the feedstock can increase the efficiency of BES to remove some water pollutants (nitrogen and phosphorus) [333].
Conductivity (Cond.)	mS/mc	The ability of water to conduct an electrical current.	Increases in conductivity can increase the power generation of BES [172]. However, because of the low conductivity in many wastewaters, the ohmic losses can reduce the BES's efficiency [172].
Total suspended solids (TSS)	mg/L	A measurement of particles that are larger than 2 microns.	The H^2AD reactor can become blocked when TSS is greater than 40,000 mg/L for a sustained period (from discussions with Lindhurst Engineering Ltd.)
Chemical oxygen demand (COD)	mg/L	The amount of dissolved oxygen required to oxidise and stabilise the given sample of water.	An advantage of BES is their ability to operate at very low COD concentrations (≈ 20mg/L) [334]. However, dues to the scale of the H^2AD system, the bioprocess requires COD approximately greater than 15,000 mg/L in the feedstock to feed the bacteria in the bioprocess (from discussions with Lindhurst Engineering Ltd.).

Total carbon (TC)	mg/L	A measurement of the total amount of carbon in the water.	Carbon is the source of the electrogenesis phenomenon that powers BES [335]. Carbon can be sourced from different organic and inorganic feedstocks, which affect how the BES perform [335].
Total inorganic carbon (TIC)	mg/L	Measurement of the amount of carbon bound to inorganic compounds in the water.	Inorganic sources of carbon are less beneficial to BES; for example, Feng *et al.* found that BES powered by an inorganic substrate had a poorer nitrogen removal performance [336].
Total organic carbon (TOC)	mg/L	Measurement of the amount of carbon bound to organic compounds in the water.	The TOC describes the amount of organic material present in the wastewater and can be used with COD to predict the amount of CH_4 content of the final biogas (18.75 x COD x TOC = CH_4 (%)) [177].
Total nitrogen (tN)	mg/L	The sum amount of nitrogen within nitrite, nitrate, ammonia and organically bonded nitrogen.	Nitrogen sources can affect BES performance by inhibiting effects on microbes, adjusting pH, and competition for electron donors/acceptors [337].
Ammonium (NH_4^+)	mg/L	Ammonia is either present in gas form or as ammonium ion in water. At certain concentrations is toxic to fish.	See tN.
Nitrite (NO_2^-)	mg/L	Nitrite is relatively unstable and easily oxidised to nitrate from. Nitrite is highly toxic to fish and other aquatic species.	See tN.
Nitrate (NO_3^-)	mg/L	Nitrate is the most oxidised form of nitrogen found in wastewater. It can have serious and fatal effects on infants.	See tN.

Water hardness (WH)	mg/L	A measurement of the amount of dissolved calcium and magnesium in the water.	There are fewer studies on the effect of water hardness on BES, but BES should great potential to reduce water hardness, which can cause scaling in hot water pipes, kitchens, bathtubs and other units [338].
Chloride (Cl^-)	mg/L	Higher than usual chloride concentrations are taken as an indication that the body of water is being used for waste disposal.	Chloride ions and adversely affect cathode catalysts present in BES [339].
Phosphate (PO_4^{3-})	mg/L	Phosphate discharged to surface water causes noxious algal blooms.	There are fewer studies on the effect of phosphate on BES, but BES show great potential of recovering phosphate from wastewater [337].
Sulphate (SO_4^{2-})	mg/L	Sulphate presences in water can lead to the formation of hydrogen sulphide, which is corrosive to gas piping and damage engines.	Sulphate presences in water can lead to the formation of hydrogen sulphide in the biogas produced by H^2AD bioprocess [331]. Hydrogen sulphide is corrosive to gas piping and damages engines.
Zinc (Zn)	mg/L	Zinc measurements are used to determine if biosolids are suitable for lad application.	Metals can substitute for oxygen to act as an electron acceptor to possible inhibiting BES [340].
Copper (Cu)	mg/L	Copper measurements are used to determine if biosolids are suitable for lad application	See Zinc.
Hydraulic residence time (HRT)	Days	The average length of time that a soluble compound remains in the system.	Length of time the wastewater remains in the H^2AD system being treated, before being discharged.

Pump speed	L/minute	The pump speed controls the flow rate of the water being displaced.	The recirculation speed within the H^2AD reactor.
System temperature	^oC	A temperature reading of a system.	The average temperature across the H^2AD reactor.
System pressure	Bar	A pressure reading of a system.	The average pressure across the H^2AD reactor.

In discussions with the SME partner, it was determined that the chemical oxygen demand (COD) removal was the most important measurement of the H^2AD bioprocesses performance. The COD removal was chosen because UK wastewater management companies use COD concentrations, as part of the in the Mogden formula, when calculating the wastewater disposal cost [341]. That rate wastewater managers charge for COD concentrations varies between companies. This means that the H^2AD bioprocess treatment becomes an attractive option in regions that charge more for COD's presence in wastewater; for example, in the Southwest. By fitting feedstock and H^2AD process data to make predictions of the COD removal rate, the model may be utilised to evaluate the H^2AD system economic suitability to treat different waste/wastewater streams in manufacturing environments.

The H^2AD plant has been operating since March 2019; however, comprehensive study measuring the feedstock and output wastewater streams is yet to occur. This is due to ongoing farm practises disrupting the supply of feedstock to the unit. In order to have enough data to demonstrate the methodologies developed in this paper, it was decided to generate synthetic data. In situations where real data may be difficult to obtain, due to budget, time or privacy concerns, generating synthetic data poses a viable alternative [342]. Synthetic data is utilised to address imbalances within the data, for proof of concept and privacy protection [342]. Developing a DDM from synthetic data is unlikely to accurately portray the H^2AD system, but it will provide data sufficient to demonstrate the techniques developed in this work.

5.4.2.2.1 Generation of model input data

The model has input data that originates from two sources. From previous work, there are 35 data-points available of WQA performed on farm waste samples, taken from the slurry tank feeding the H^2AD system. In addition, there is one year's worth of H^2AD process data automatically collected by the unit. Resende *et al*. showed that the bacteria community within the cattle slurry can vary throughout the year due to seasonal variations [343]. For the model to try and capture this seasonal variability, data collection over a minimum of one year is required. This informed the decision to generate an additional 17 WQA data points bringing the total to 52 data points, representing one year's worth of weekly sampling. When generating synthetic data the standard method is to randomly sample from a probability density function (PDF) fitted to existing data [342]. Pair-copula constructions (PCCs) are a popular tool for modelling multivariate data, due to their simple structure and high flexibility [344]. Several different classes of PCCs exist [345]. The Clayton copula produced synthetic data that most closely matched the original data, demonstrated by the similarity of the

correlations' coefficients in the original dataset and synthetic dataset, shown in Figure 5.3.

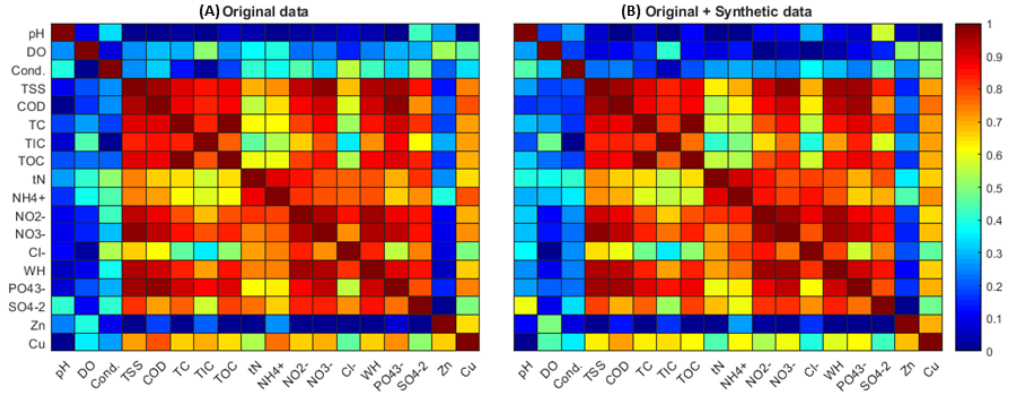


Figure 5.3 Heatmaps displaying the scaled correlation coefficients between variables in (A) the original WQA data and (B) the original plus synthetic data.

5.4.2.2.2 Generation of model output data

Because of insufficient data relating input variables to the model's output variable, it was not possible to follow the same PDF method. There have been several attempts to model bioelectrochemical systems (BES) via first principal and data-driven methods [346]; however, none of them is representative of a plant scale BES system. Furthermore, the H^2AD bioprocess is a unique system configuration, which combines aspects of both anaerobic digesters and BES, and utilises a novel, low-cost electrode design [210]. This means that synthetic data generated from existing BES first-principal equations would not be representative of the H^2AD system. Instead, a complex non-linear equation was developed to generate synthetic COD removal percentages from the input data. Trends identified from historical H^2AD data governed the development of the non-linear equation. For confidentiality reasons, it is not possible to disclose the historical H^2AD bioprocess data; however, the trends highlighted from the data and between discussion with SME partner are given in Table 5.2.

Table 5.2 Correlation strength between model input variables and COD removal efficiency, where r is Pearson's correlation coefficient.

	Positive	Negative
Strong $\lvert 0.7 < r < 1.0 \rvert$	HRT, System temperature	DO, Cond., pH, WH
Moderate	--	TIC, Cl-, TC, Zn, NO2-, NH4+, TOC, SO4-2, tN, NO3-, TSS, Cu,

|0.35 < r <.7|		PO43-, Pump speed, system pressure
Weak	--	COD
|0 < r <.0.35|		

The output equations are physically based, and their forms are not of great importance since the goal is only to create synthetic data with characteristics similar to the historical data. This approach to synthetic modelling is not new, especially within the field of hydrology [347]. This approach is justified because the synthetic data is not being employed to show the success or failures or the H^2AD bioprocess, rather it is utilised for proof of concept for the modelling methodologies developed in this work. The dataset generated is included in Appendix F.

5.4.2.3 Model development

Following the CRISP-DM framework, the model's development comes after the data preparation step. The model development stage is normally split into 4 steps: data processing, feature selection, model training and validation, and evaluation of the model on test data [105,348,349]. When evaluating a model's ability to fit the data and make predictions, previous work has only evaluated the model on data used during the development of model [113,350]. It is important to evaluate the model on unseen data both inside and outside of the system to get a truer assessment of the model's extrapolation capabilities. However, in order for this to be achieved, first, the data must be explored to define what part of the system is represented by the data and the boundaries of the system. Therefore, two new steps have been added (steps 1 and 6 below) so that the modelling methodology prosed by this work is:

1. Define system's boundaries.
2. Data processing.
3. Feature selection.
4. Train and validate the model.
5. Evaluate the mode on **test** data.
6. Evaluate the model on **unseen** data.

The core requirement of a DDM is to ensure the model's predictive capability meets the requirements of the manufacturer. When the dataset is fixed the potential for prediction improvement lies with either data processing and feature selection (steps 2 and 3) or more sophisticated algorithms (stage 4) [142]. In the literature, the majority

of DDM development methodologies for manufacturing have been for experimental data [116,119,351]. Therefore, the focus is on finding the optimal structure and/or topology for the selected machine-learning algorithm, rather than on stages 2 and 3. This is because the experiments have been designed to collect data from across the system and only relevant variables are measured. In contrast, those papers that developed a DDM from manufacturing data often had access to large datasets, requiring a greater focus on stages 2 and 3 in order to format the dataset ready for modelling [330,352]. However, these methods also neglected to define the system boundaries, relying instead on the likelihood that a large dataset is inherently representative of the system.

5.4.2.3.1 Defining the system boundaries

Exploratory data analysis was performed to understand the manufacturing system represented in the data, using a combination of statistical and data visualisation techniques. The model development data was analysed to determine the system's features: central tendency, correlations, distribution, spread, modality, and extents. Of particular interest was defining the feedstock characteristics data and COD removal, as compared to the H^2AD process data, these are subject to variability that is beyond the manufacturer's ability to control. The statistical measures employed were: mean, medium, standard deviation, Pearson's correlation coefficient, skewness, range, interquartile points, maximum, and minimum. Data visualisation plays an important role in detecting patterns, trends, and outliers in groups of data [307]. Visualising manufacturing systems' data is made challenging because of their high dimensional nature [307]. The data visualisation techniques capable of plotting high dimensional data utilised were: scatter plot matrix, box plots, heatmaps, and parallel coordinate plots.

This article used a mixture of statistical and visualisation techniques to identify and communicate key features of the H^2AD input and output data, these are detailed in Table 5.3.

Table 5.3 Statistical and visualisation techniques used to describe features within the data when defining the manufacturing system.

	Importance to model	Statistical technique	Visualisation technique
Central tendency	Identifies system's centre	Mean Medium	Box plot Histogram
Spread of data	Identify if data is well spread out across the	Variance Standard deviation Skewness	Box plot Scatter plot matrix

	system or concentrated around the centre		Parallel coordinate plot
Correlations	Identify relationships between the system variables.	Pearson's coefficient	Scatter plot matrix Heatmap Parallel coordinate plot
Distribution	Identify if one part of the system is more represented	Skewness	Multidistributions surface plots Histogram
Modality	Identify different regions of activity within the system	--	Scatter plot matrix Histogram
Data extents	Identify the boundaries of the system	Range Interquartile points Maximum Minimum	Box plot Parallel coordinate plot
Outliers	Identify data points unlikely to be representative of the system	--	Box plot

5.4.2.3.2 Data processing

For this study data was integrated from two sources, WQA performed of H^2AD input and output streams samples and H^2AD process data automatically collected remotely by the unit. Data normalisation was undertaken to ensure the impact of variables of lower magnitudes are given the same weight by the model's algorithm as the variables of larger magnitudes. The COD and TSS data have magnitudes within the order of 10,000, while for DO is the magnitude is much lower and in the order of 0.1. To normalise the data the minimax function was applied, as this normalises the variables without any loss of information [330]. The equation to normalise the data is given below:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Equation 5.1

Where x is the variable, x_{norm} is the normalised variable and x_{min} a*nd* x_{max} is the minimum and maximum values of the variable being normalised respectively.

5.4.2.3.3 Feature selection and dimensionality reduction

When attempting to develop a model with limited data points one risks the *curse of dimensionality*. The *curse of dimensionality* is the name given to the phenomena observed that as the number of input variables increases the volume of data required to avoid overfitting increases exponentially [330]. Which in turn degrades the prediction capabilities of a DDM on new data [353]. Two methods exist to avoid this problem; *feature selection* and *dimensionality reduction*. Not all the input variables measured have a significant effect on the output variables [354]. Feature selection reduces the number of variables in the dataset, keeping only the most relevant variables [354]. The number of dimensions in the remaining data can be further reduced by extracting the key features of variables and transposing these into a new smaller set of variables [113]. This is known as dimensionality reduction. There a variety of feature selection and dimensionality reduction techniques available [330,355]. This work utilised a neighbourhood component analysis (NCA) feature selection and principal component analysis (PCA) to reduce the number of dimensions in the data. The NCA is a machine learning technique that learns the feature weights of the input data on the output data [356], weights close to zero are classified irrelevant. An NCA for regression identified 7 input variables (TSS, COD, TC, TIC, NO_3^-, Cl^- and PO_4^{-3}) to remove, displayed in Figure 5.4 (A), so that the H^2AD input variables are reduced from 22 to 15.

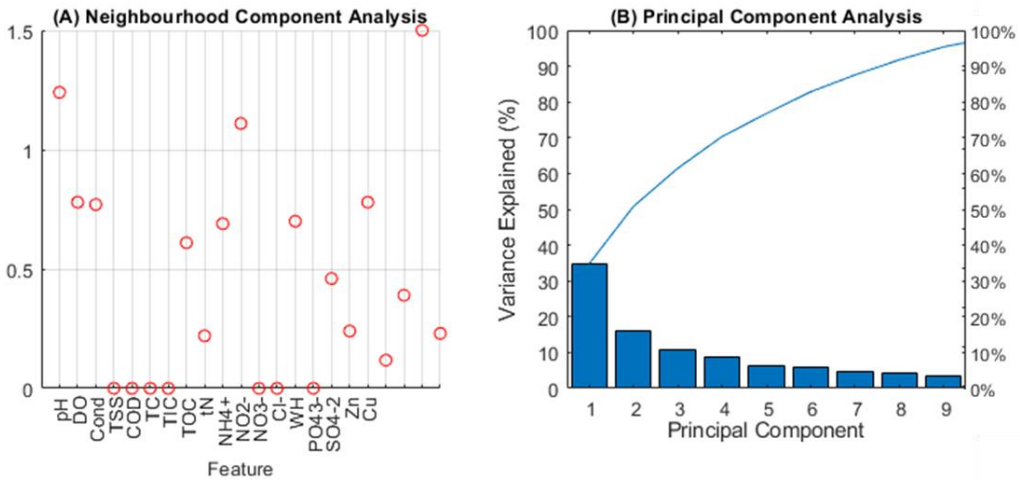


Figure 5.4 Results from feature selection and dimensionality reduction techniques, (A) neighbourhood component analysis and (B) principal component analysis.

Principal component analysis is one of the most widely used and understood dimensionality reduction techniques, being first used in 1901 [137]. Principal component analysis works on the principle that in data sets with many variables, groups of variables often move together as they are measuring the same driving force within a system [137]. A PCA is a mathematical procedure that transforms potentially correlated

data into an orthogonal system of linearly uncorrelated principal components (PC) [113]. All the PC are orthogonal to each other and ordered, so that the maximum variance is captured in PC1 and the variance decreases as one moves down the PCs. The standard practice is to reduce the number of variables so that 95% of the variance within the dataset is captured by the PC [113]. By applying PCA to the H^2AD dataset, the input variables were further reduced from 15 to 9 PCs, displayed in Figure 5.4(B).

5.4.2.3.4 Data partitioning

The data used to develop a DDM is then partitioned at several levels, as independent data is required to train, validate, test and evaluate the model, see Figure 5.5. The data is first partitioned into model development data and unseen data. The model development data is further partitioned into training, validation and testing data [137]. Unseen data-points are kept apart until the end to evaluate the prediction capabilities of the model on new data, not used during the model's development. While most studies will evaluate the model using testing data, evaluating the model on unseen data has often been overlooked [119]; although, this is changing with some recent published DDM using unseen data [113,119]. The difference between testing and unseen data is that testing data is used to evaluate the DDM during the model development to prevent overfitting on the training data, meaning it only evaluates the model's prediction capability within the system. Whereas, unseen data evaluate the final model's ability to make predictions on new data, both inside and outside of the system.

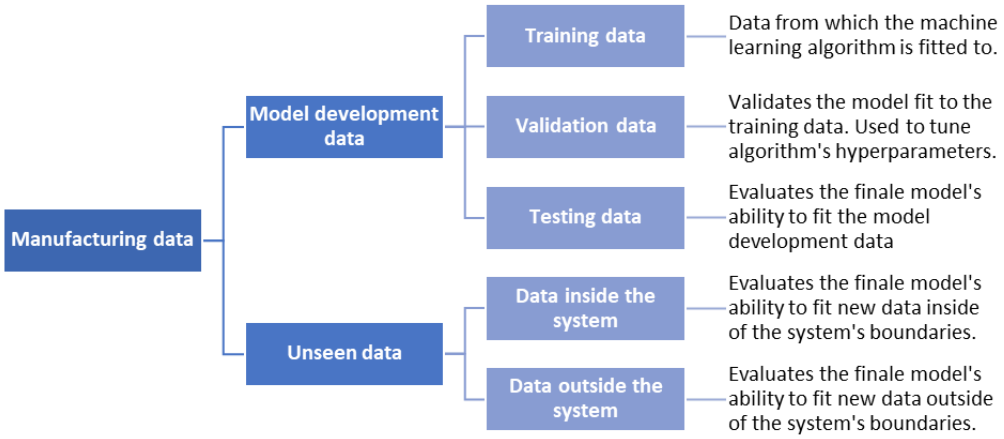


Figure 5.5 Hierarchy of data partitioning during the development of a data-driven model.

There are two means of acquiring unseen data. The first is to generate data through experiments so that the data extends evenly across and beyond the system boundaries. This can be achieved by performing a design of experiments [119]. When this is not possible the manufacturing data is partitioned into model development data and unseen.

This was the case for this study and is the most common method of the two [286]. In previous work, the data is randomised and split 60-90% model development data and 40-10% unseen data, the percentages are dependant on the size of the dataset [116]. In larger datasets (greater 1,000 data points), this method is likely to include data points from across the system. However, in smaller datasets (less than 50 data points) the likelihood of having regions that are unevaluated by unseen data is higher. This was the case when this method was applied to the 45 H^2AD system data points. To overcome this, a new systemic method to partition the data that ensures unseen data extends evenly within and beyond the system boundaries was developed. The method developed in this paper splits the manufacturing data into 80% model development data (36 data point) and 20% unseen data (9 data points). Rather than sampling the unseen randomly from the data, 40% of the unseen data is taken from outside of the system and 60% taken inside the system. To ensure that the data taken from within the system is distributed evenly across the system, unseen data was sampled at set percentiles throughout the system. The 10th and 90th percentile were sampled to evaluate the model close to the system boundaries, and 25th, 50th and 75th percentiles were sampled to evaluate the model evenly across the system. In Figure 5.6 the final spread of unseen data throughout and beyond the system is shown.

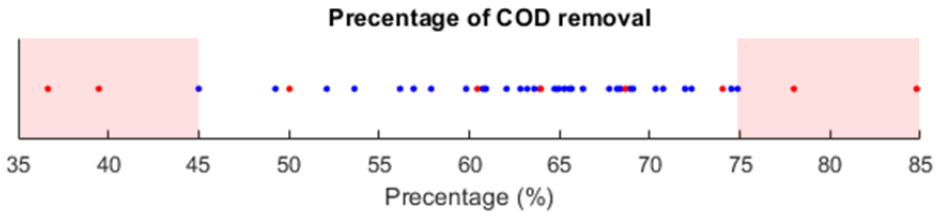


Figure 5.6 Spread of unseen data throughout the system. Blue markers are the model development data points, red markers are the unseen data points and the red regions are outside of the system.

5.4.2.3.5 Algorithm selection, training, validation and testing

When developing a DDM, there are a number of machine learning algorithms to select from. The "No Free Lunch" theorem states all algorithms perform equally well when averaged over all possible datasets [142]. In deciding which algorithm to use there are two schools of thought. Either explore the literature for algorithms developed for similar manufacturing systems or start with the simplest algorithm and if that does not meet the requirements try ones that are more complex [142]. For this work, four machine learning algorithms suitable for regression were investigated: Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural networks (ANN). These algorithms were chosen as they have been shown to have stronger predictive powers over other algorithms like logistic regression [142,349].

When developing a DDM, each algorithm has model parameters and hyperparameters that require tuning to find the optimal configuration. Hyperparameters are parameters that are set before being trained on a model, compared to model parameters that are learnt during training [100]. For example, the number of neurons in an ANN's hidden layer is a hyperparameter; whereas, the weights between neurons is a model parameter. Validation data is used to tune the hyperparameters and training data is used to tune the model parameters. There are many different methods to optimise an algorithm's hyperparameters, simple approaches include a grid or random search, to more complex methods using RF and GP [146]. Bayesian optimisation is thought to be the most commonly used model for optimisation, as it allows exploration and exploitation to be naturally balanced during the search [146]. The Bayesian optimiser works by constructing a conditional probabilistic model that can be updated with new evidence [146]. Simply put, the optimiser picks a prior belief about how the hyperparameters will behave, then searches the hyperparameter space by updating the prior belief based on ongoing measurements. The training, validation and testing of the DDMs and hyperparameter optimisation were carried out by MATLAB R2017.b software developed by MathWorks. The hyperparameters optimised are displayed in Table 5.4, all other hyperparameters not listed were left on their default setting by MATLAB R2017.b.

Table 5.4 The hyperparameters for each machine learning algorithm used.

Algorithm	**Hyperparameters**
Gaussian process (GP)	Covariance function, kernel function, kernel scale, sigma
Random forest (RF)	Number of trees, depth of trees, the minimum node size.
Support vector regression (SVR)	Epsilon, box constraint, kernel function, kernel scale, polynomial order
Artificial neural network (ANN)	Number of hidden layers, number of nodes in hidden layers, training function, learning rate.

5.4.2.3.6 Model evaluation

The prediction capabilities of the various ANN models were statistically evaluated in terms of the coefficient of determination (R^2) the root of the mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The R^2 is the square of the sample correlation coefficient between observed values and

predicted values, and is a measure of the explained variance by the model [113]. The RMSE measures the mean square magnitude of the error [348], while MAE measures the absolute magnitude of the errors [357]. Both are useful because RMSE puts a greater influence on large errors than smaller ones, which is sometimes useful if large errors are important to be avoided [357]. While MAE is easily interpretable, a model with a low MAE can have the occasional very high error [357]. The MAPE is similar to MAE but is based on percentage error, being that it is scaled independently and can be used to compare model performance across different data sets [357]. The statistically best model is the one that has an R^2 closest to 1 while minimising RMSE, MAE and MAPE. The R^2, RMSE, MAE and MAPE are defined as follows:

$$R^2 = \frac{(\sum_{i=1}^{n}(p_i - \bar{p}_i)(a_i - \bar{a}_i))^2}{\sum_{i=1}^{n}(p_i - \bar{p}_i)^2(a_i - \bar{a}_i)^2}$$

Equation 5.2

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(a_i - p_i)^2}{n}}$$

Equation 5.3

$$MAE = \frac{\sum_{i=1}^{n}|a_i - p_i|}{n}$$

Equation 5.4

$$MAPE = 100 \times \frac{\sum_{i-1}^{n}\left|\frac{a_i - p_i}{a_i}\right|}{n}$$

Equation 5.5.5

Where n is the number of points, p_i is the predicted value, a_i is the actual observed value, and the symbol $\bar{}$ is the average of the related values.

5.4.3 RESULTS AND DISCUSSION

5.4.3.1 System boundaries

To understand the manufacturing system represented in the data, a combination of statistical and data visualisation techniques were employed (Table 5.5, Figures 5.7 and 5.8). The model development data was analysed to determine the system's features: central tendency, correlations, distribution, spread, modality, and extents. Of particular interest was defining the feedstock characteristics data and COD removal, as compared to the H^2AD process data, these are subject to variability that is beyond the manufacturer's ability to control.

The central tendency can be used to describe the centre point of each variable within the system. Statistically, this is achieved with either the mean or median and has been calculated for each variable in Table 5.5. The variables TOC, tN and system pressure

show greater than 10% difference between the mean and median. This can indicate the mean is being influenced by extreme scores. When visualising the data for TOC and tN in Figures 5.7 and 5.8(B) there are noticeable clusters towards the lower values for both. On inspecting the histograms in Figure 5.7, over half of the TOC values are clustered around the 0.2 value. This is similar but less extreme for tN, which shows the cluster is more spread out around the bottom values. The feedstock variables Nh_4^+, NO_2^-, WH and SO_4^- all have a cluster of measurements lowering the central tendencies. The pH, DO, conductivity, Zn and Cu, have similar values for mean and medium and the histograms and box plots, show a central tendency in the middle of the range values. Defining the mean for H^2AD process variables that are tightly controlled is less informative because these values operate at set values. For example, the system pressure has three values it operates at (Figure 5.7), so a mean between these values does represent the value we expect for new data-points. Instead, the median is more informative in representing the centre of the system. Defining the central tendencies of the system informs what values we expect to observe in new data points collected. If the central tendencies of the system shift as new data is collected, it is possible there has been a change to the system and the model needs retraining.

The relative standard deviation informs on the spread of the data around the mean. The pH and conductivity have small relative standard deviations, implying they are closely clustered around the mean. Whereas, the remaining feedstock variables have high relative standard deviations, indicating that the data is more spread out. This shows these variables are subject to greater levels of variability. The scatter plot matrix visualised the distribution and spread of the data. The feedstock variables have data that is distributed across the system. Although there are variables that contain clusters (e.g. tN) and there more strongly defined, there appear to be no poorly defined regions. Contrasting this the H^2AD process variables, pump speed and system pressure are only defined in 4-3 exact regions respectively. As these are control variables, this will only become an issue if the operator sets the H^2AD process conditions outside of those the model was trained on. If the H^2AD system does move to operate at different conditions, new data will need to be collected to retrain the model.

The correlations within the dataset are shown in Figure 5.8(A). Instantly recognisable in the heatmap, is that the TOC, tN, NH_4^+, NO_2^-, WH and SO_4^{-2} all show strong positive correlations to one another and Cu is moderately correlated with this set as well. The strong correlation between tN, NH_4^+ and NO_2^- is expected, as they are all measuring nitrogen's different states within the feedstock. The pH has strong negative correlations with Cu and percentage of COD removal. Knowing the strong correlations within the

dataset will help to identify new data-points that behave unexpectantly. This is another indicator that the system has evolved beyond the data that the model was developed on.

Table 5.5 Statistical analysis of the variables in the model development dataset

	Feature	Mean	Medium	Maximum	Minimum	Range	Relative Standard Deviation	Skewness
Feedstock characteristics	pH	7.54	7.52	7.90	7.28	0.623	2.24	0.391
	DO	0.182	0.190	0.280	0.110	0.170	22.9	-0.0864
	Cond.	11.7	11.6	12.8	10.3	2.54	5.33	-0.0784
	TOC	8020	6650	12200	5060	7140	27.2	0.750
	tN	1830	1560	3240	770	2470	33.0	0.612
	NH4+	944	880	1490	604	888	24.4	0.794
	NO2-	12.0	11.1	16.9	8.40	8.51	24.0	0.411
	WH	209	191	281	160	120	18.1	0.461
	SO32-	5460	5200	8170	3480	4700	25.8	0.313
	Zn	24.4	23.9	35.8	10.3	25.5	27.7	-0.310
	Cu	62.0	64.1	101	22	79	29.4	-0.620

H^2AD process conditions	HRT	3.71	3.48	6.09	2.25	3.84	25.8	1.20
	Pump speed	9.72	9.94	9.94	8.84	1.10	3.94	-1.76
	Proces Temp.	32.9	33.5	39.2	29.8	9.40	5.93	0.612
	System pressure	0.00333	0.00300	0.00500	0.00300	0.00200	16.0	1.32
H^2AD output	% COD removal	63.6	64.3	74.9	45.0	29.9	10.6	-0.731

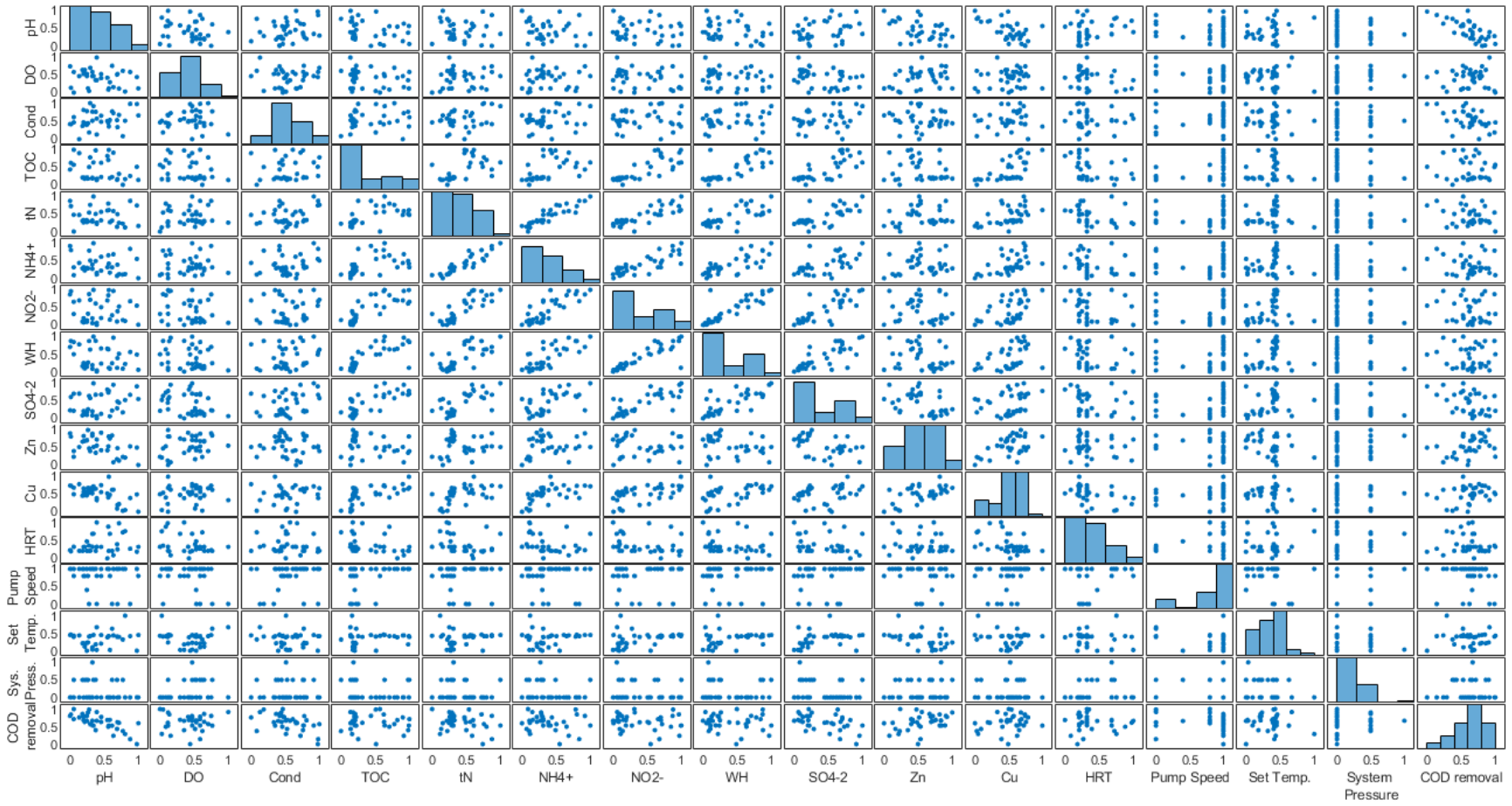


Figure 5.7 Scatter plot matrix of variables in the model development dataset

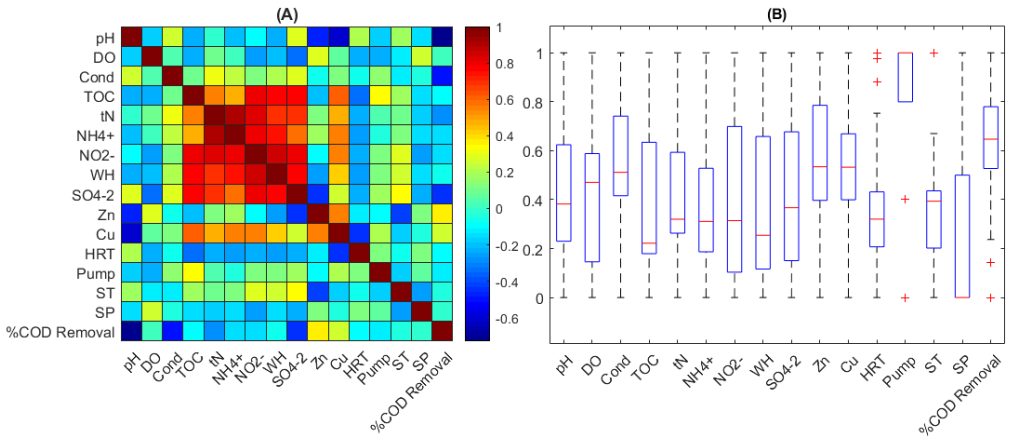


Figure 5.8 (A) Heatmap displaying Pearson's correlation between the model development data and (B) boxplot displaying the distribution of the model development data

5.4.3.1.1 Determining if new data points are within the system

A useful tool for visualising all the systems variables in one figure is the parallel coordinate plot (PCP). The PCP is able to visualise high dimensional data by, setting each variable along as a point on the x-axis [307]. Data-points are normalised, so all values can be displayed on a single y-axis, and plotted as a series of lines, displaying a data-point normalised value for each variable [307]. By adapting the PCP, a traffic light classification was developed that and can quickly inform manufacturers if variables of a new sample are within the boundaries of the model development data (green region), approaching the system boundaries (yellow region) or outside the system (red region), see Figure 5.9. The red boundaries are defined as the maximum and minimum of the model development data and the yellow boundaries the 0.1 and 0.9 quantiles. In Figure 5.9 two examples are given of a data-point that remained in the system, and a data-point that was well outside of the system for certain variables. The advantage of this traffic light classification is that it provides a simple visualisation that can be displayed on a handheld device and provide the operators with real-time information that the system is currently operating within the boundaries the model was developed on. Therefore, informing them which model predictions can be trusted.

While the red boundaries will always be the maximum and minimum, the yellow boundaries may vary depending on the requirements of the model predictions and how accurate the model is at predicting outside of the system. For example, when the RF model was tested on unseen data outside of the system (see section 5.4.3.2), it performed relatively well (MAE 4.96) and when predicting the percentage of COD removal, it meets the manufacturer's requirement. However, if this model was built for a predicting an output where a greater level of accuracy is required (e.g. predicting a variable critical to preventing a runaway reaction), the yellow boundaries may have to be increased to

give an earlier warning to the operators that the system is approaching the point outside the boundaries of the data the model was trained on.

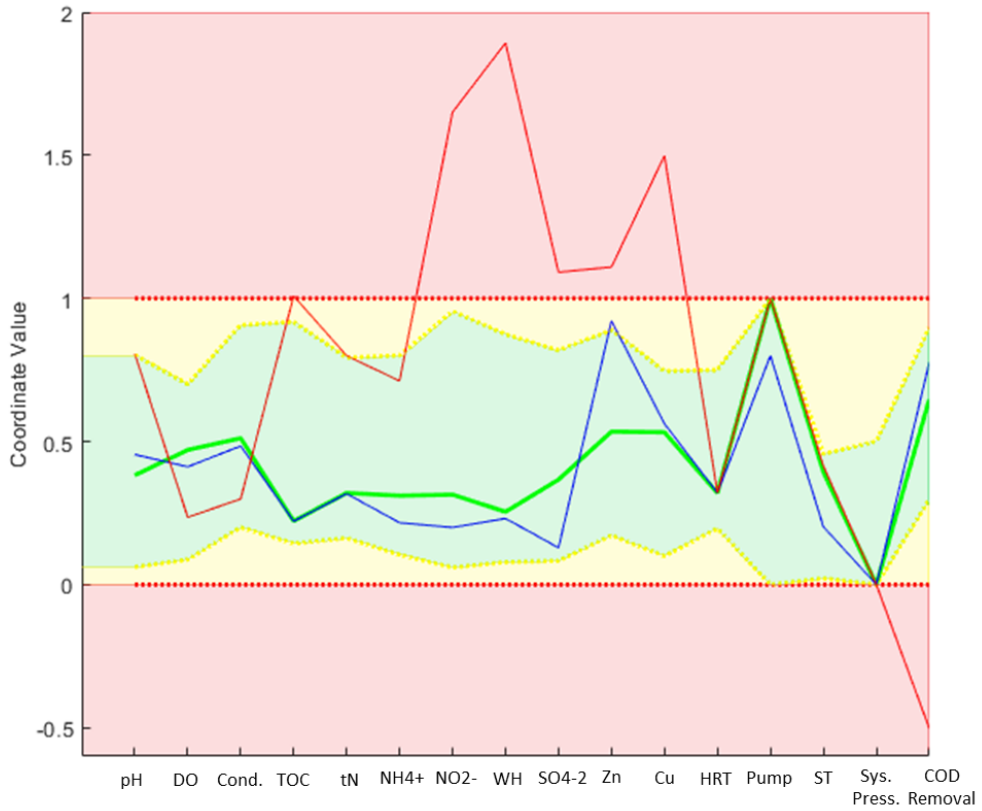


Figure 5.9 Traffic light system informing manufacturers if new data points are within the manufacturing system (green region), approaching the boundaries of the system (yellow region) or outside the system (red system). Where, the green line is the median of each variable, the blue is a data point within the system and the red line is a data point outside of the system.

5.4.3.2 Results from data-driven models

With respect to predicting the H^2AD bioprocess to remove COD from wastewaters, several models were built to fit the manufacturing data, including the GP model, RF model, SVR model and ANN model. The model development data was used to determine the optimal hyperparameters and model parameters of the DDMs. The 45 model development data points were randomised and 85% (38 data points) were partitioned for the training and validation of the model and the remaining 15% of data (7 data points) were partitioned for testing the final model. A Bayesian optimisation was performed to optimise the hyperparameters, using k-fold cross-validation to evaluate each configuration's ability to fit the validation data. For the k-fold cross-validation, a k value of 10 was chosen because of the limited number of data points [116]. The models' ability to fit new data was then evaluated using the testing data points. The optimised hyperparameter values for each model are displayed in Table 5.6.

Table 5.6 The hyperparameters values for the final machine learning models

Gaussian Process		**Random Forest**		**Support Vector Regression**		**Artificial neural network**	
H-parameter	Value	H-parameter	Value	H-parameter	Value	H-parameter	Value
Sigma	0.0203	Trees	301	Epsilon	0.88	Hidden layers	1
Basic function	Linear	Min no. of predictors	3	Box constraint	937.84	Neurons	6
Kernel function	Squared exp.	Min leaf size	1	Kernel function	Gaussian	Training function	Levenberg-Marquardt
Kernel scale	0.0562			Kernel scale	5.87	Learning rate	0.001

It has been common practice within process manufacturing to use regression plots, of actual values against a model's predictions when determining how well the model fits the data [113,119]. In Figure 5.10 the regression plots for the developed DDMs indicates that all the models were able to achieve good agreement between the actual and predicted values. From visual inspection, the GP, SVR and ANN models were slightly superior to the RF model and this is supported by the statistical analysis in Table 5.7. The GP model outperformed the other models in the MAE and MAPE error metrics. Both these scores present the magnitude of error around the predicted value and show that the average error of the GP model was the lowest; however, the GP model RMSE was higher The RMSE squares the error magnitude, giving a greater value to outliers. A high RMSE score indicates that the GP model's predictions included some unexpected high errors between the actual percentage of COD removal and the predicted. The full results, showing the model's predictions and residuals are provided in the supplementary material.

Regression plots are useful for a quick guide in determining how accurately a model's predicting match the actual data but are limited when analysing how a DDM performs throughout the system. This is because they are less informative in explaining how the model performs in different region of the system. Instead, residuals plots can be used to better understand where a model overestimates or underestimates the actual value

and determine if there are fundamental flaws within the model. The residuals are calculated as the actual value minus the predicted percentage removal of COD, and the plot for each model are presented in Figure 5.10. The RF model appears to be overestimating at values lower than approximately 60% COD removal and overestimating above this. Furthermore, it is possible to observe a linear relationship in the residual error. Similar, the residual plot for the SVR model shows that the model appears to consistently overestimate below 60% COD removal by the same amount. Above 60% COD removal, the model appears to both overestimate and underestimate by consistently the same amount (with exceptions). Future testing on unseen data is required to see if these patterns are restricted only to the model development data. If not, it may be possible to define this relationship and using it to calibrate the models' predictions.

All four of the developed DDM are on a similar performance level to other DDMs built to study a BES. Lesnik and Liu built an ANN model predicting the COD removal by a microbial fuel cell and when tested on a validation data achieved MAPE 0.57% and 1.06% [125]. However, the authors did not evaluate the model on testing data or unseen data, reducing the confidence of the MAPE score. Garg *et al.* built and evaluated three DDMs of a BES; utilising GP, SVR and ANN machine learning algorithms [202]. After evaluating the DDMs on testing data, Garg *et al.* also observed the GP model to have the best overall performance [202]. What neither of these studies has done nor, to the best knowledge of the authors, other BES DDMs, is evaluate the DDMs' prediction capabilities on unseen data. Without evaluating on unseen data, it is harder to assess the prediction capabilities of a DDM on new data. This is vital if the model's predictions are to be trusted by a manufacturer.

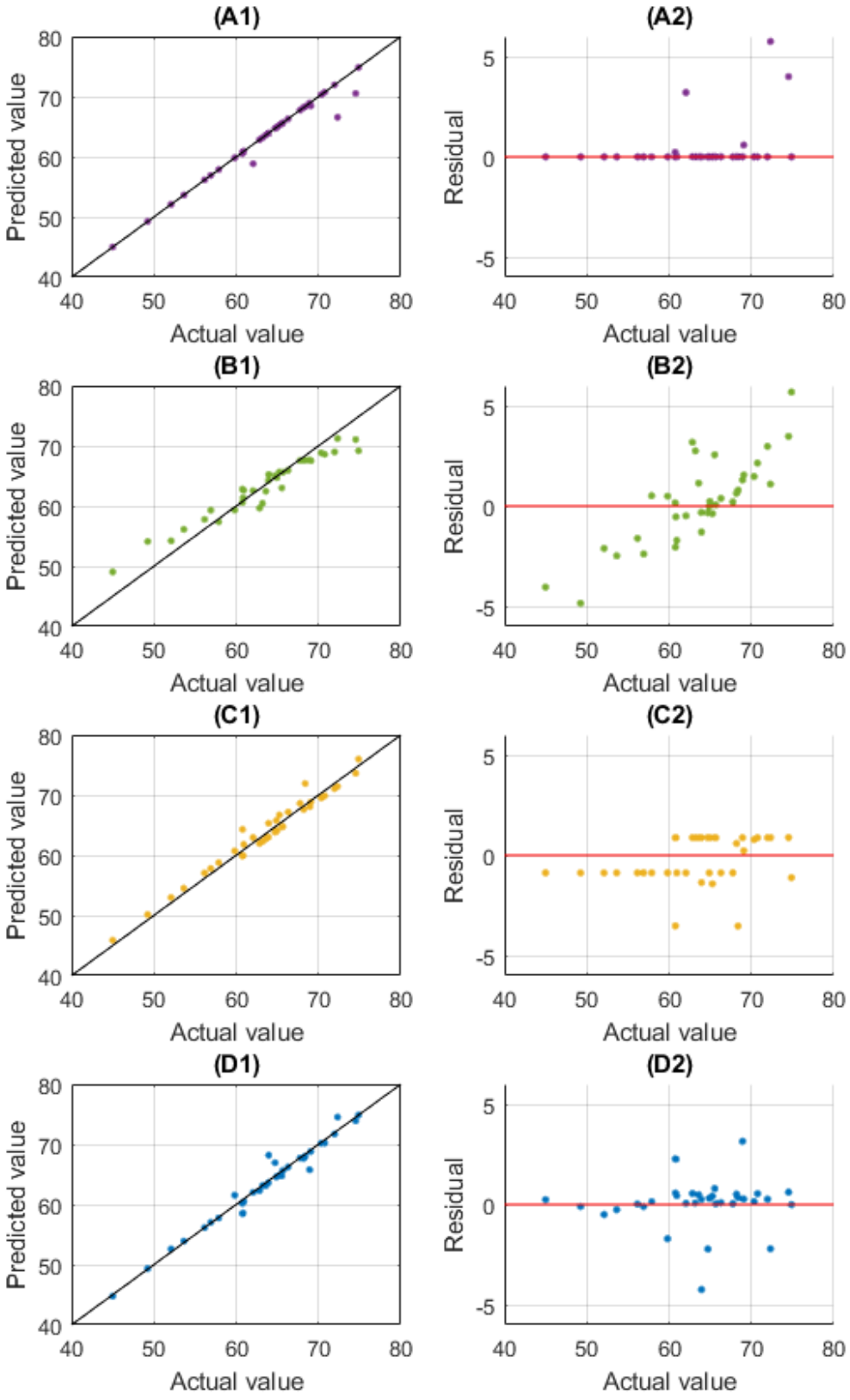


Figure 5.10 The predicted percentage of COD removal plotted against the actual value (1) and the residuals plotted against the actual values (2), for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network

Table 5.7 Statistical comparisons of the predictive capabilities of the Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural network (ANN) models

Data	R^2				RMSE				MAE				MAPE			
index	GP	RF	SVR	ANN	GP	RF	SVR	ANN	GP	RF	SVR	ANN	GP	RF	SVR	ANN
Training	0.966	0.945	0.969	0.967	1.202	1.226	1.166	1.262	0.383	1.595	1.031	0.737	0.549	2.604	1.638	1.150
Unseen	0.650	0.989	0.540	0.429	5.328	1.439	7.421	8.715	8.879	2.811	8.450	9.245	16.868	6.039	15.736	17.286
Inside	0.850	0.986	0.995	0.954	3.629	1.075	0.783	2.981	3.999	1.092	3.171	3.478	6.628	1.949	4.900	5.632
Outside	0.834	0.995	0.527	0.416	4.276	1.745	9.259	9.288	14.980	4.960	15.050	16.454	29.668	11.152	29.281	31.853

5.4.3.3 Evaluating the models' predictive capability on unseen data

Applying George Box's aphorism "all models are wrong, but some are useful" to manufacturing, helps to illustrate that while it is not possible to perfectly model the manufacturing system, a model may aid in decision-making, develop the scientific understanding about the system, communicate knowledge and/or make predictions [358]. However, the challenge is deciding which models are useful and which are not, to avoid giving the manufacturing false or misleading information. Part of the solution is to evaluate how well the model performs on new unseen data. Nine unseen data points were partitioned from the data before model development to evaluate the DDMs' predictive capabilities. These unseen data points are further partitioned into points inside and outside of the system to give a comprehensive assessment of the models' predictive capabilities across and beyond the system.

The model abilities to predict the percentage of COD removal in wastewater were statistically evaluated using, R^2, RMSE, MAE and MAPE (Table 5.7). Although all models showed good quality predictions on the model development data ($R^2 > 0.95$), the GP, SVR and ANN models' performance all decreased when evaluated on unseen data; whereas, the RF model's performance increased (0.99). The increase between model development data and unseen data of RF model's R^2 value is not reflected by the other error metrics. However, the RMSE, MAE and MAPE still show that the RF model is clearly superior to the other models when predicting on unseen data. In Figure 5.11, the residual plot for the model's predictions on unseen data are displayed. The RF and SVR models are able to match the performance between model development data and unseen data within the system. However the SVR model performance significantly decreased outside of the system; whereas, for the RF model the decrease was not as extreme. This reflects the strong generalising capabilities that RF algorithms are known for [150]. The RF model predictions on percentage COD removals were more accurate for high values of COD removal outside the system than they were for low values, being able to accurately predict one of the high COD values outside of the system.

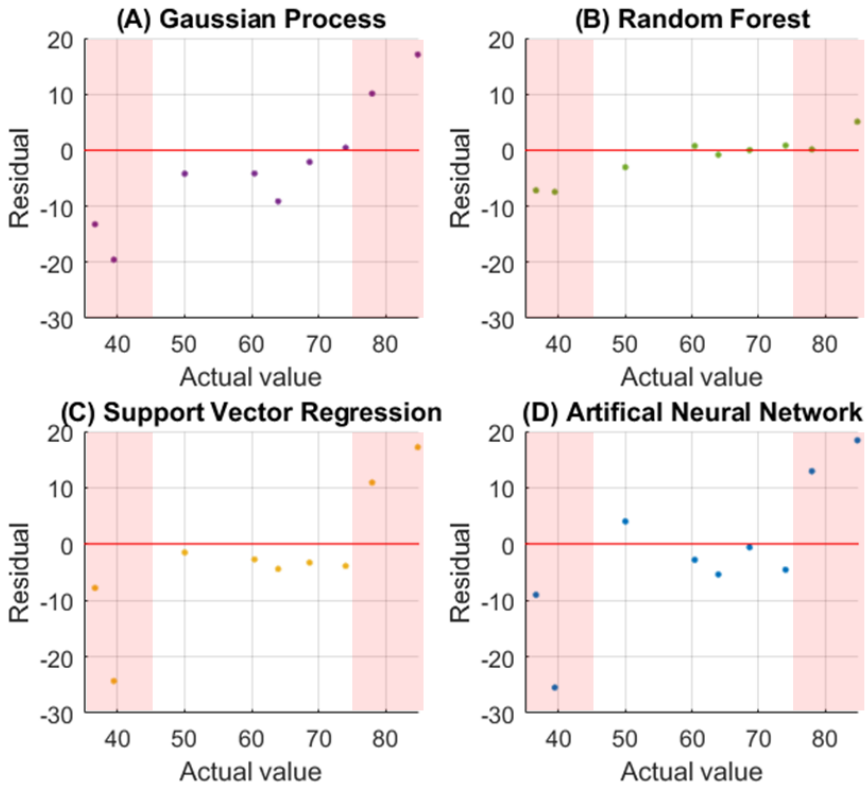


Figure 5.11 Residuals of unseen predictions plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network. The red region is outside of the system.

When selecting which model to use, it is important to consider not only how the model performs on the model development data, but also unseen data. The RF model was deemed to be the best overall model for predicting the percentage of COD removal by the H^2AD bioprocess, because of its consistent performance on both datasets. The RF model is an example of an ensemble model and this may be the reason the RF model performed better for predictions on unseen data. Ensemble models are built from integrating the predictions made by various models into one output [359]. Ensemble modelling has shown great potential for modelling bioprocesses as a way to account for uncertainty in the model structures and parameters [359]. Ensemble models, like stacked ANN, split the modelling space, by a means similar to cross-validation, so each model is a percentage of the overall space [360]. The 'random' in RF sets it apart from previous work, like stacked ANN, because it not only integrating models that have split the modelling space but has also split the number of input variables available for each model randomly [150]. When modelling systems that face data limitations, ensemble models can overcome the risk of overfitting and reduce the effect of the *curse of dimensionality*.

5.5 CONCLUSIONS FROM PART A

- To avoid overfitting, the developed DDM's prediction capabilities should be evaluated on unseen data, taken both inside and outside of the system. This work developed a methodology to determine the manufacturing system boundaries using both statistical and data visualisation techniques. Once the system boundaries are defined, it allows for: determining the DDM's interpolation and extrapolation capabilities and knowing when to retrain the model.
- By defining the system boundaries, manufacturers can utilise visualisation techniques, like PCP, to monitor when the system moves beyond the system boundaries. The model will then need retraining, incorporating the new data into the model development data. A traffic light protocol for easily classifying if new data collected is inside or outside the system was developed for this work.
- This study compares the performance of four machine learning algorithms, GP, RF, SVR and ANN. The models' capability to predict the percentage of COD removal from wastewater by the H^2AD bioprocess was statistically evaluated using R^2, RMSE, MAE and MAPE.
- The GP model had the best performance when evaluating the model's ability to fit the model development data (R^2 of 0.97, RMSE of 1.20, MAE of 0.38 and MAPE of 0.55). However, the model performance was the second worst of the four DDMs when evaluated using unseen data (R^2 of 0.65, RMSE of 5.33, MAE of 8.88 and MAPE of 16.87).
- The random forest (RF) model was found to have the best predictive capabilities on unseen data (R^2 of 0.99, RMSE of 1.44, MAE of 2.814 and MAPE of 6.04). Because of the RF model ability to extrapolate beyond the data it was trained on, it was deemed the best model at describing the COD removal by the H^2AD bioprocess.

5.6 PART B: MULTIPLE TARGET MODELLING TO ENABLE SUSTAINABLE PROCESS MANUFACTURING

5.6.1 THE NEED FOR MULTIPLE TARGET MODELS

Process manufacturing sectors such as food and drink, pharmaceutical and waste processing are constantly striving to make their processes more sustainable from an environmental and economic perspective. A powerful tool to achieve this are data-driven models (DDMs). Data-driven models are developed using algorithms that learn hidden relationships within manufacturing data to make predictions. Previous attempts at modelling process manufacturing systems, have largely focused on building multiple independent single target models, each predicting one output. These models face limitations when capturing the relationships between the system's outputs. This paper develops Multiple Target (MT) models that are superior when describing a manufacturing system with multiple outputs.

The aim of this work is to investigate the best techniques to develop a MT model of a process manufacturing system and demonstrate how that MT model may be utilised to increase sustainability within the system. The main contributions presented in this work include:

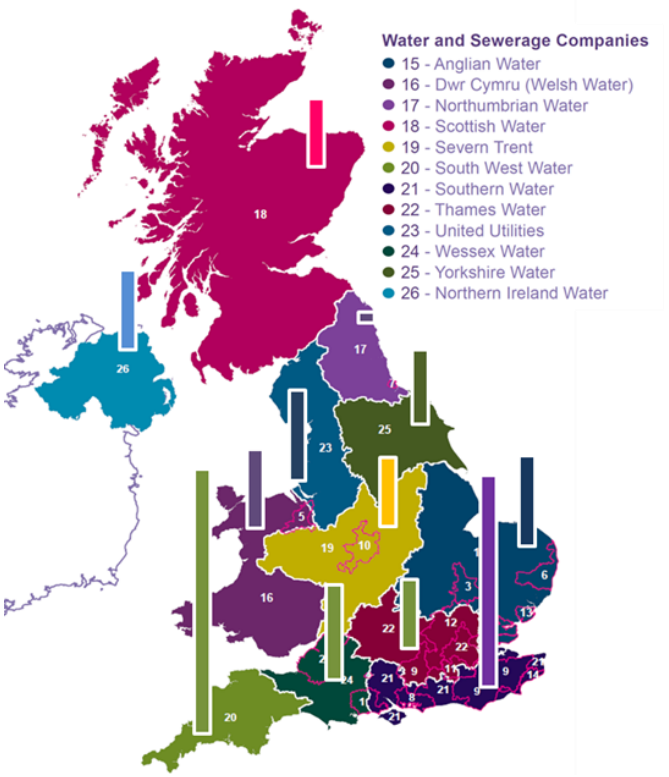
- Evaluating the performance problem transformation and leading algorithm adaptation models when modelling a manufacturing system. Including evaluating their abilities to make predictions on unseen data, both inside and outside of the manufacturing system.
- Proposing a novel MT model by adapting state-of-the-art algorithm adaptation models to use artificial neural networks (ANN) as a base model for the first time. The new model is then evaluated against existing MT models techniques.
- Utilising the final MT model to investigate the most sustainable manufacturing route of a novel bioprocess manufacturing system.

5.6.2 MATERIAL AND METHODS

The models developed are trained on data collected from an operating H^2AD plant currently treating farm wastes produced by a 300 cow dairy farm. The farm waste contains cattle slurry, bedding waste, waste milk, footbath, parlour washing and rainfall. To characterise the waste entering the H^2AD bioprocess, Water Quality Analysis (WQA) was performed on samples collected from the H^2AD plant. A total of 18 WQA parameters were measured (although not all of these would be truly relevant to modelling the output data). These were: pH, dissolved oxygen (DO), conductivity, total suspended solids (TSS), chemical oxygen demand (COD), total carbon (TC), total

inorganic carbon (TIC), total organic carbon (TOC), total nitrogen (tN), ammonium (NH_4^+), nitrite (NO_2^-), nitrate (NO_3^-), water hardness (WH), chloride (Cl^-), phosphate (PO_4^{3-}), sulphate (SO_4^{2-}), zinc (Zn) and copper (Cu). In addition to this, the H^2AD unit automatically collected data on four H^2AD process conditions via sensors connected by Internet of Things (IoT) technologies to a spreadsheet hosted on a cloud platform. These were HRT, pump speed, System Temperature (ST) and System Pressure (SP).

The H^2AD system has two objectives: improve the wastewater quality for reuse and energy generation from waste streams. This was represented in the model by 3 outputs: percentage of COD removal (%COD removal), percentage of TSS removal (%TSS removal) and daily volume of methane produced by the H^2AD bioprocess (CH_4 generation). The pollutants COD and TSS were chosen because UK wastewater management companies use COD and TSS concentrations in the wastewater, as part of the Mogden formula, when calculating wastewater disposal cost [361]. The rate wastewater companies charge for wastewater disposal differs between companies, see Figure 5.12. This will have implications when determining whether it is more sustainable to maximise pollutant removal or energy generation. The energy generation potential of the H^2AD bioprocess is represented by the volume of CH_4 produced. The H^2AD bioprocess produces biogas, made up of CH_4, carbon dioxide, oxygen and hydrogen sulphide components. Of these, only CH_4 has significant calorific value and can be used to estimate the energy potential from the biogas produced [362].



Annual sewerage charge

Southwest	£	205,378
Southern	£	157,373
Wessex	£	74,967
United Utilities	£	70,693
Anglian	£	69,138
Northern Ireland	£	61,251
Dwr Cymru	£	60,313
Yorkshire Water	£	57,776
Scotland	£	52,889
Thames Water	£	51,450
Severn Trent	£	50,980
Northumbrian	£	1,880

Figure 5.12 Annual wastewater charge for different UK water and sewerage companies, based on 4m^3 /d of wastewater with a chemical oxygen demand of 85,000 mg/L and total suspended solids of 984 mg/L using WRAP Mogden calculator [361].

5.6.2.2 Modelling data

The H^{2}AD plant has been operating since March 2019; however, a comprehensive study measuring the input and output streams is yet to occur. This is due to ongoing farm practises disrupting the supply of wastewater to the unit. In order to have enough data to demonstrate the methodologies developed in this work, it was decided to generate additional synthetic data following the same produced outline in section 5.4.2.2. The dataset generated is included in Appendix F.

5.6.2.3 Multiple target model development

For this work, 3 problem transformation MT models and 5 algorithm adaptation MT models were developed from the dataset describing the H^{2}AD system. Each model was then used to predict the values for the model development data and unseen data. The prediction capabilities of the models were then statistically evaluated. Before developing the MT models, data preparation, pre-processing and partitioning tasks were performed on the H^{2}AD dataset.

5.6.2.3 Data preparation, pre-processing and partitioning

The data's preparation is considered one of the most important and time-consuming tasks [330]. Data preparation is aimed at transforming raw data into high-quality data

that is formatted for the modelling processes [354]. For this study, the dataset was normalised as part of the data preparation stage. This was to ensure variables of higher magnitudes are not given greater weight by the DDM. The COD and TSS data have magnitudes within the order of 10,000, while for DO is the magnitude is much lower and in the order of 0.1. To normalise the data the minimax function was applied, as this normalises the variables without any loss of information [330].

Effective data pre-processing is essential for any data-driven approach and if done correctly can significantly improve a DDM's predictive capabilities by avoiding the *curse of dimensionality* [349]. The curse of dimensionality is the name given to the phenomena observed, that as the number of input variables increases the volume of data required to avoid overfitting increases exponentially [330]. Overfitting is the generation of a model that corresponds too closely or exactly to a particular set of data, and may, therefore, fail to fit unseen data or predict future observations reliably [141]. The risk of overfitting can be reduced by feature selection and dimensionality reduction. Reducing the number of variables also increases computational efficiency, while also potentially increasing the performance of the machine learning algorithm by removing non-discriminative features [349].

This work utilised a neighbourhood component analysis (NCA) feature selection method and principal component analysis (PCA) to reduce the number of dimensions in the data. The NCA is a machine learning technique that learns the feature weights of the input data on the output data [356], weights close to zero are classified irrelevant. An NCA for regression identified 2 input variables (conductivity and water hardness) to remove so that the H^2AD input variables are reduced from 22 to 20. PCA is a mathematical procedure that transforms potentially correlated data into an orthogonal system of linearly uncorrelated principal components (PCs) [113]. All the PCs are orthogonal to each other and ordered so that the maximum variance is captured in PC1 and the variance decreases as one moves down the PCs. The standard practice is to reduce the number of variables so that 95% of the variance within the dataset is captured by the PCs [113]. By applying PCA to the H^2AD dataset, the input variables were further reduced from 20 to 9 PCs.

The data used to develop a DDM is then partitioned at several levels, as independent data is required to train, validate, test and evaluate the model. The data is first partitioned into model development data and unseen data. The model development data is further partitioned into training, validation and testing data [137]. Unseen data-points are kept apart until the models are complete to evaluate the prediction capabilities of

the model on data not used during the model's development. The systematic approach developed in section 5.5 was used to partition the manufacturing data into 70% model development data and 30% unseen data. This ensured that unseen data was selected from across and beyond the model boundaries for each output, as shown in Figure 5.13.

The model development data is then portioned into training, validation and testing data. When developing a DDM, each algorithm has model parameters and hyperparameters that require tuning to find the optimal configuration. Hyperparameters are parameters that are set before being trained on a model, compared to model parameters that are learnt during training [100]. For example, the number of neurons in an ANN's hidden layer is a hyperparameter; whereas, the weights between neurons is a model parameter. Validation data is used to tune the hyperparameters and training data is used to tune the model parameters. Testing data is utilised to evaluate the final model's ability to fit new data.

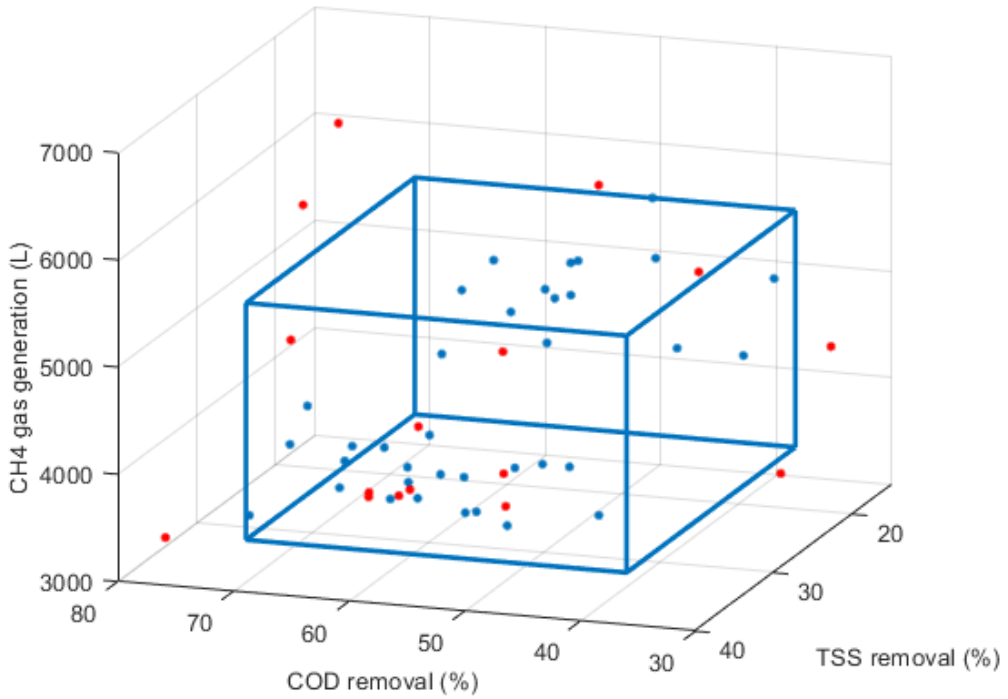


Figure 5.13 The spread of unseen data throughout the system. Blue markers are the model development data points, red markers are the unseen data points and the blue box represents the model boundaries.

5.6.2.4 Problem transformation model development

Problem transformation MT models work by stacking independent DDMs, to predict the output variables simultaneously, as shown in Figure 5.14. As mentioned previously, the problem transformation techniques have proven less effective as algorithm adaption modelling techniques, as the technique does not capture any possible relationships between model outputs [127]. However, by developing problem transformation models,

it is possible to evaluate the suitability of different machine learning algorithms to model the H^{2}AD system. This may reduce the number of further algorithms adaption models produced, as it may be possible to eliminate underperforming machine learning algorithms.

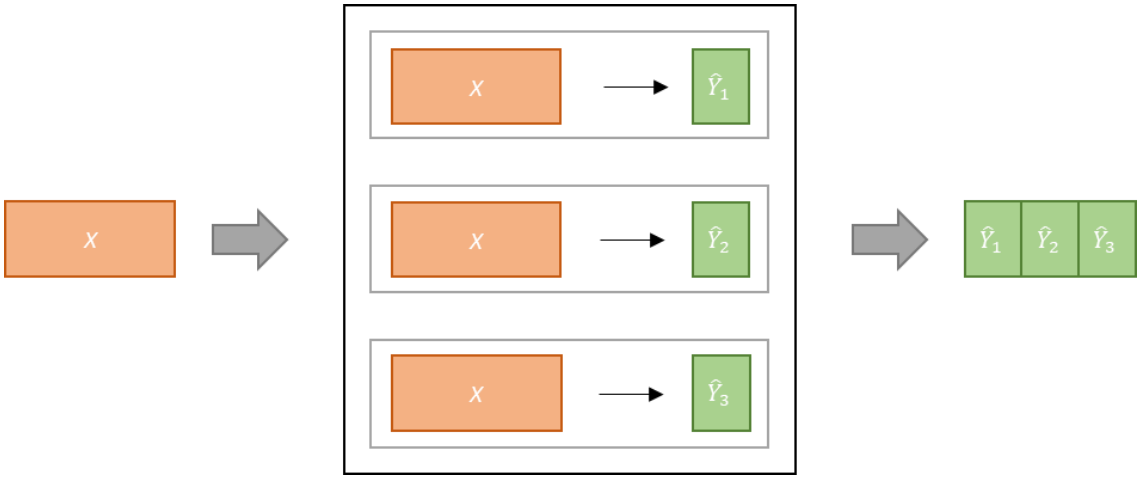


Figure 5.14 Problem transformation model, predicting three outputs. Where the model input data (X) is represented by the orange box, the model output prediction ($\hat{Y}_i$) by the green box, the single target models are within the grey box and the multiple target model is within the black box.

The abilities of three machine learning models to model the H^{2}AD system were evaluated: random forest (RF), support vector regression (SVR) and artificial neural network (ANN). Each of these algorithms has proven effective at modelling various process manufacturing systems [149,152,350,363,364].

The RF and SVR were chosen as they are used as the base models in the leading MT modelling techniques developed by Spyromitros-Xioufis *et al.* and Melki *et al.* [126,365]. An ANN algorithm was also used, as ANNs have excellent capabilities when modelling non-linear systems and are the most extensively utilised DDM in process manufacturing [366]. Furthermore, the ANN architecture already contains the capability to be simply modified to predict multiple outputs [100]. As of yet, ANN models have not been utilised as a base model in either Spyromitros-Xioufis *et al.* or Melki *et al.* techniques. Consequently, it is currently not known how a multiple target artificial neural network (MT-ANN) compares to one built using either Spyromitros-Xioufis *et al.* or Melki *et al.* techniques.

5.6.2.5 Algorithm adaption model development

There are serval approaches to use when building MT models, including linear target combinations for MT regression [367], multi-objective random forests [368], multiple output ANN [115], boosted-neural network ensemble [369] and multi-task deep learning [370]. Recently, Spyromitros-Xioufis *et al.* investigated whether the

techniques of the multi-target (label) classification techniques could be successfully transformed for MT regression problems [365]. Spyromitros-Xioufis *et al.* explored the possibility of stacking or chaining regression models together, with Regression Chains (RCs) proving the most effective [365]. In building a RC model a random chain, or sequence, of the output variables is selected and for each output in the chain, ST models are built sequentially by using the output of the previous model as input for the next (Figure 5.15). There are two main problems with using a single RC:

1. The relationships between the output variables at the start of the chain and the end of the chain are not exploited [126].
2. Prediction error at the start of the chain will propagate through the rest of the chain [365].

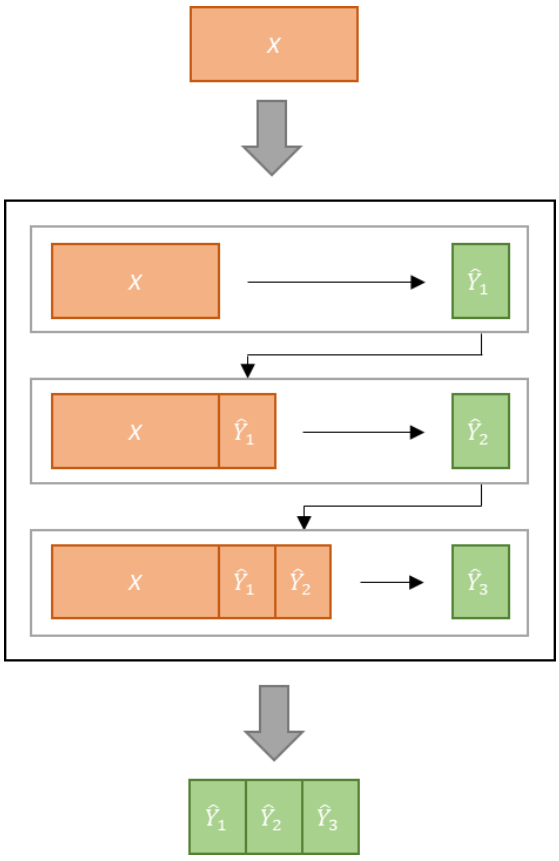


Figure 5.15 Regression chain multiple target model, predicting three outputs. Where the model input data (X) is represented by the orange box, the model output prediction ($\hat{Y}_i$) by the green box, the single target models are within the grey box and the multiple target model is within the black box.

To overcome this an ensemble of regressor chains (ERC) was proposed [365]. The ECR works by creating multiple RCs, with every possible permutation of output variable sequence order, and the final prediction values are obtained by taking the mean of the predicted values for each output variable, as shown in Figure 5.16. Because of the computational time for building models with outputs that exceed 10 permutations, the

sequence order is randomised, and 10 RCs are selected for constructing the ERC. When developing an ERC model, the training stage follows a different procedure than outlined in Figure 5.16 Instead of using the predicted values for the outputs as model inputs, as one progresses down each chain, the actual values are used. This is to minimise error propagation during training and increase the accuracy of the model. If the predicted values were used, each prediction error in the preceding model in the chain is propagated down the chain during training.

As the number of output variables increases, the number of possible chains increases by a factorial function. Therefore, there is no guarantee that the 10 random chains generated will truly capture the relationships among the outputs. Additionally, building an ERC becomes computationally more expensive as the number of outputs increases. Melki *et al.* proposed an adaption to the ERC, which instead used a single chain based on the maximisation of the correlations among the output variables [126]. Melki *et al.* referred to their techniques as SVR correlation chain (SVRCC), as it was built using a SVR as a base model. For this work, it is referred to as the maximum correlation chain (MCC) for when applying the technique to other base models. To construct a MCC the first step is to calculate the outputs' correlation coefficient matrix, which will describe the linear relationship among the output variables. This can then be sorted into descending order, creating the MCC. The MCC then follows the same training procedure as the ERC, but only one chain is created with the maximum correlation order. On following this procedure, the MCC order for the H^2AD system was:

$$\%\text{TSS removal} \geq \%\text{COD removal} \geq \text{Generation of } CH_4 \text{ gas}$$

The final MT model built for this work was a multiple target artificial neural network (MT-ANN) model. An ANN model is highly suited to MT modelling as its architecture can be quickly adaptedfor multiple outputs [365], as shown in Figure 5.17. Because of this, MT-ANNs are among some of the earliest examples of MT modelling [371,372]. More recently Curteanu *et al.* modelled the electrolysis process in wastewater treatment and found that MT-ANN outperformed multiple individual ANN models [115]. A MT-ANN was constructed of the H^2AD system and compared against the RC techniques.

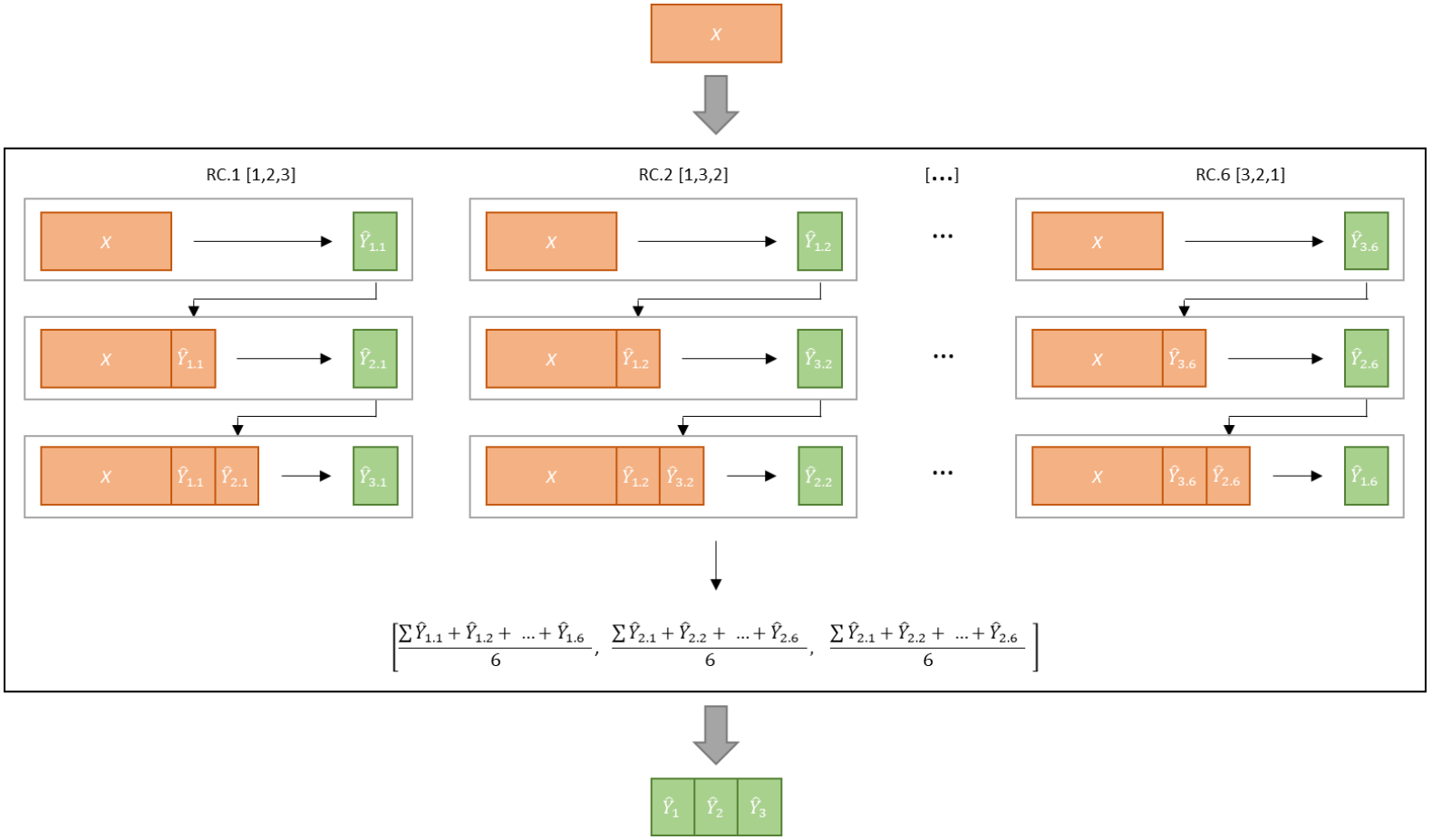


Figure 5.16 Ensemble of Regression Chain (ECR model) predicting three outputs. Where the model input data (X) is represented by the orange box, the model output predictions (Y^_i) by the green box, the single target models are within the grey box and the ERC model is within the black box.

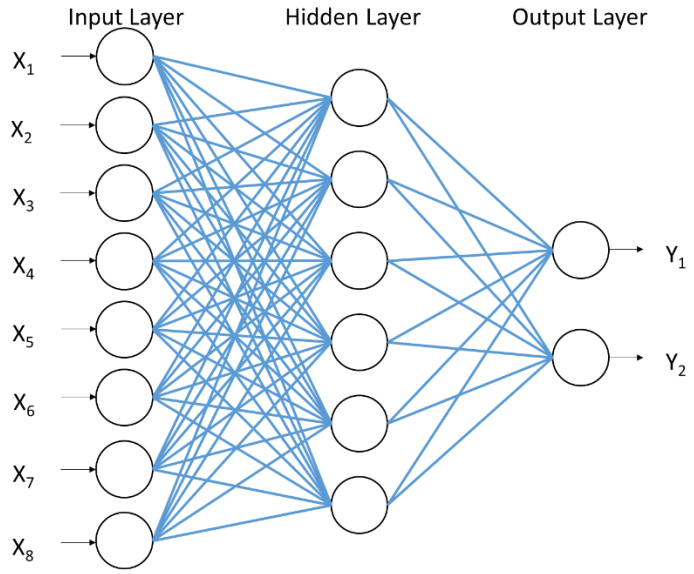


Figure 5.17 Architecture of a Multiple Target Artificial Neural Network (MT-ANN), with 8 inputs, 2 outputs and 1 hidden layer containing 6 neurons

5.6.2.6 Model evaluation

The prediction capabilities of the various models were statistically evaluated in terms of the coefficient of determination (R^2), the normalised root of mean squared error (nRMSE) and the mean absolute percentage error (MAPE). The R^2 and MAPE have the same form as described in Equations 5.2 and 5.5 respectively. The RMSE and mean average error (MAE) were the error scores are scaled in the same units as the predicted value. In order to compare the models' performance when predicting different output the normalised root of the mean squared error was utilised [357].

$$nRMSE = \frac{1}{Y_{max} - Y_{min}} \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n}}$$

Equation 5.6

Where n is the number of data points, $\hat{Y}$ is the predicted value, Y_i is the actual observed value, and the symbol $\bar{}$ is the average of the related values.

5.6.3 RESULTS AND DISCUSSION

5.6.3.1 Results from data-driven models

5.6.3.1.1 Problem transformation multiple target models

With respect to modelling the H^2AD bioprocess's capability to remove COD and TSS from wastewaters and generate CH4 gas, 8 MT DDMs were built to fit the manufacturing data. Three of the models were examples of problem transformation and used RF, SVR and ANN as base models. The model development data was used to determine the optimal hyperparameters and model parameters of the DDMs. The model development data points were randomised and 85% were partitioned for the training

and validation of the model and the remaining 15% of data were partitioned for testing the final model. A Bayesian optimisation was performed to optimise the hyperparameters, using k-fold cross-validation to evaluate each configuration's ability to fit the validation data. For the k-fold cross-validation, a k value of 10 was chosen because of the limited number of data points [116]. Bayesian optimisation is one of the more common approaches when searching for machine learning hyperparameters [146]. The Bayesian optimiser works by constructing a conditional probabilistic model that can be updated with new evidence [146]. Simply put, the optimiser picks a prior belief about how the hyperparameters will behave, then searches the hyperparameter space by updating the prior belief based on ongoing measurements. The model ability to fit new data was then evaluated using the testing data points.

The RF, SVR and ANN models' capabilities to predict the %COD removal, %TSS removal and CH_4 generation were statistically evaluated using R^2, nRMSE and MAPE, the results are displayed in Figure 5.18. In terms of R^2 and MAPE, the SVR and ANN models show, on average, a clear superiority over the RF model. This indicates that the RF model's error between actual and predicted values is larger. When breaking down the models' predictive capabilities for each output, it can be seen that the RF model performs consistently worst when predicting the %COD removal. The SVR and ANN show similar performance at predicting the %COD removal, though the ANN model has a higher nRMSE score, suggesting that the ANN model is subject to a greater number of outliers within the predictions. This trend between the SVR and ANN model is repeated when predicting CH_4 generation, but when predicting %TSS removal the SVR model performs consistently worse than the ANN model. Overall, the SVR and ANN show similar predictive capabilities, with the ANN marginally outperforming the SVR model.

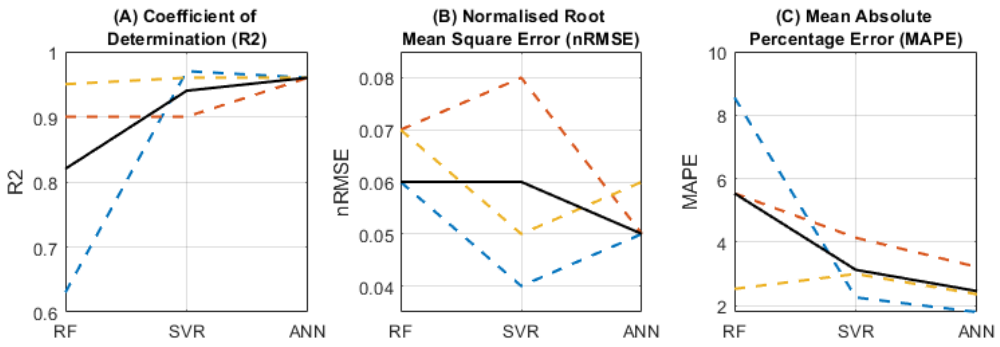


Figure 5.18 Evaluation metrics of the problem transformation modes. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 generation output.

Problem transformation models have been shown to be less accurate when predicting multiple outputs, as they are unable to capture the relationship between model outputs [127]. However, they can be effective at determining the suitability of various machine learning algorithms as a base model for algorithm adaption models [126]. After statistical analysis of the predictive capabilities of the RF, SVR and ANN models, the decision was taken to use only SVR and ANN algorithms as base models for the algorithm adaption MT models. The main advantages of using RF models stem from their simplicity to tune and the fact they are ensemble models, less prone to predicting outliers [150]. However, previous work shows they frequently perform worse than SVR and ANN methods [150], as is the case in this work.

5.6.3.1.2 Algorithm adaption multiple target models

Five algorithm adaption MT models were built to describe the H^2AD system. The first two models were built following ERC techniques using SVR and ANN as base models, these are referred to as ERC-SVR and ERC-ANN respectively. Models 3 and 4 were built following Melki *et al.* MCC technique using SVR and ANN as base models, these are referred to as MCC-SVR and MCC-ANN respectively. The final MT model built was a MT-ANN. All the base models were built following the Bayesian optimisation and cross-validation procedures.

The ERC-SVR, ERC-ANN, MCC-SVR, MCC-ANN and MT-ANN models' capabilities to predict the % COD, %TSS removal and CH_4 generation were statistically evaluated using R^2, nRMSE and MAPE. The results, displayed in Figure 5.19, indicate all the models' statistical metrics follow a similar trend and that ERC-ANN is a clearly superior model. The MCC-ANN and ERC-ANN have a similar MAPE score, indicating that on average they have similar prediction errors; however, the MCC-ANN has a higher nRMSE error. That is the result of the model containing a larger number of outlier predictions. The ERC-ANN is an example of an ensemble model, which are able

to mitigate the influence of outlier predictions. This is because an ensemble model outputs are an average of multiple predictions; an outlier may raise the value of the prediction, but by averaging the prediction an ensemble model is capable of reducing the outliers' impact on model performance.

Overall, all of the models showed good ability to fit the H^2AD manufacturing data (R^2>0.91, nRMSE < 0.08 and MAPE < 4.3). The RC models built using an ANN as a based model were capable of more accurate predictions and the ERC MT modelling technique was superior to the MCC technique in terms of R^2 and nRMSE. Particularly interesting was comparing the ANN RC models to the MT-ANN model, as these techniques have never been compared before. The MT-ANN model was outperformed by both the ECR-ANN and MCC-ANN models. The reason for this may be that the models were constructed on only 36 data-points. Multiple target ANN models are capable of achieving the best results in fields that have access to 10,000+ data points [118]. However, they are known to be less capable when modelling systems where data is limited [118]. This is because, as an ANN architecture increases its requirement for data increases to avoid overfitting [330]. The RC ANN models built have smaller architectures (only one neuron in outer layer), which may explain their better performance. Before drawing any final conclusions on the MT models, it is important to evaluate the models' performance when making prediction for unseen data.

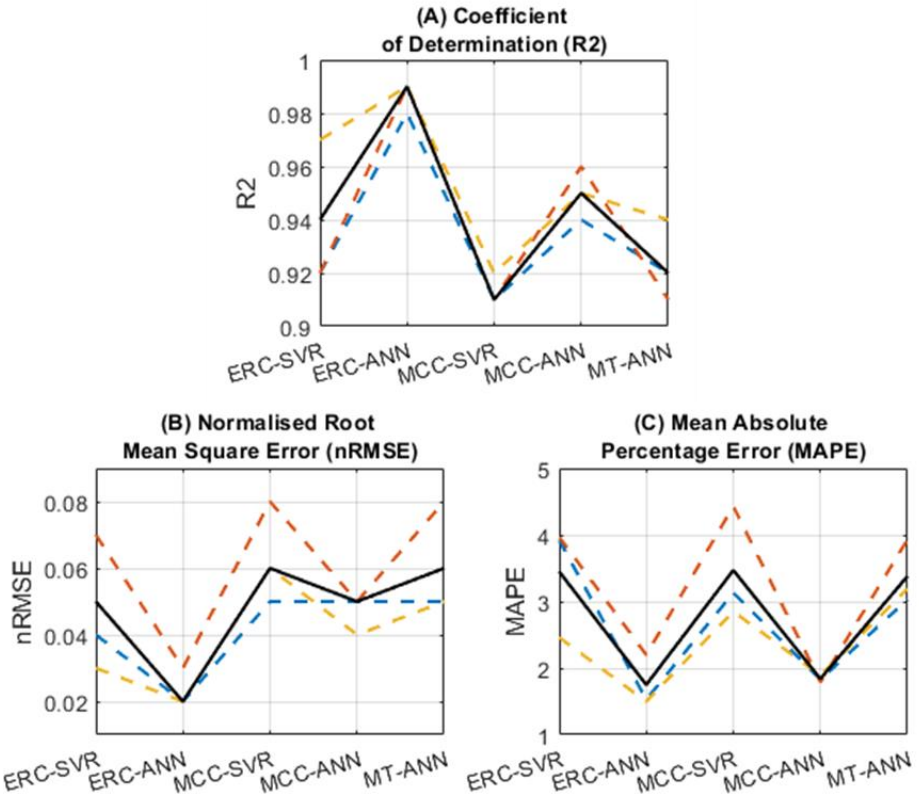


Figure 5.19 Evaluation metrics of algorithm adaption modes. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 output.

5.6.3.1.3 Comparison of the models' performance on unseen data

A common error when developing a DDM, is to put too much emphasis on optimising the model's capability to fit the model development data [118]. The best DDMs are able to extrapolate beyond the data they were developed from, to make accurate predictions on unseen data. Before commencing work on developing the MT models, 16 data points were partitioned from the data to be used to thoroughly evaluate the models' predictive capabilities. Ten points were taken from within the model development boundaries and 6 points outside the model boundaries. The statistical metrics results for each model, inside and outside the system, are shown in Figure 5.20.

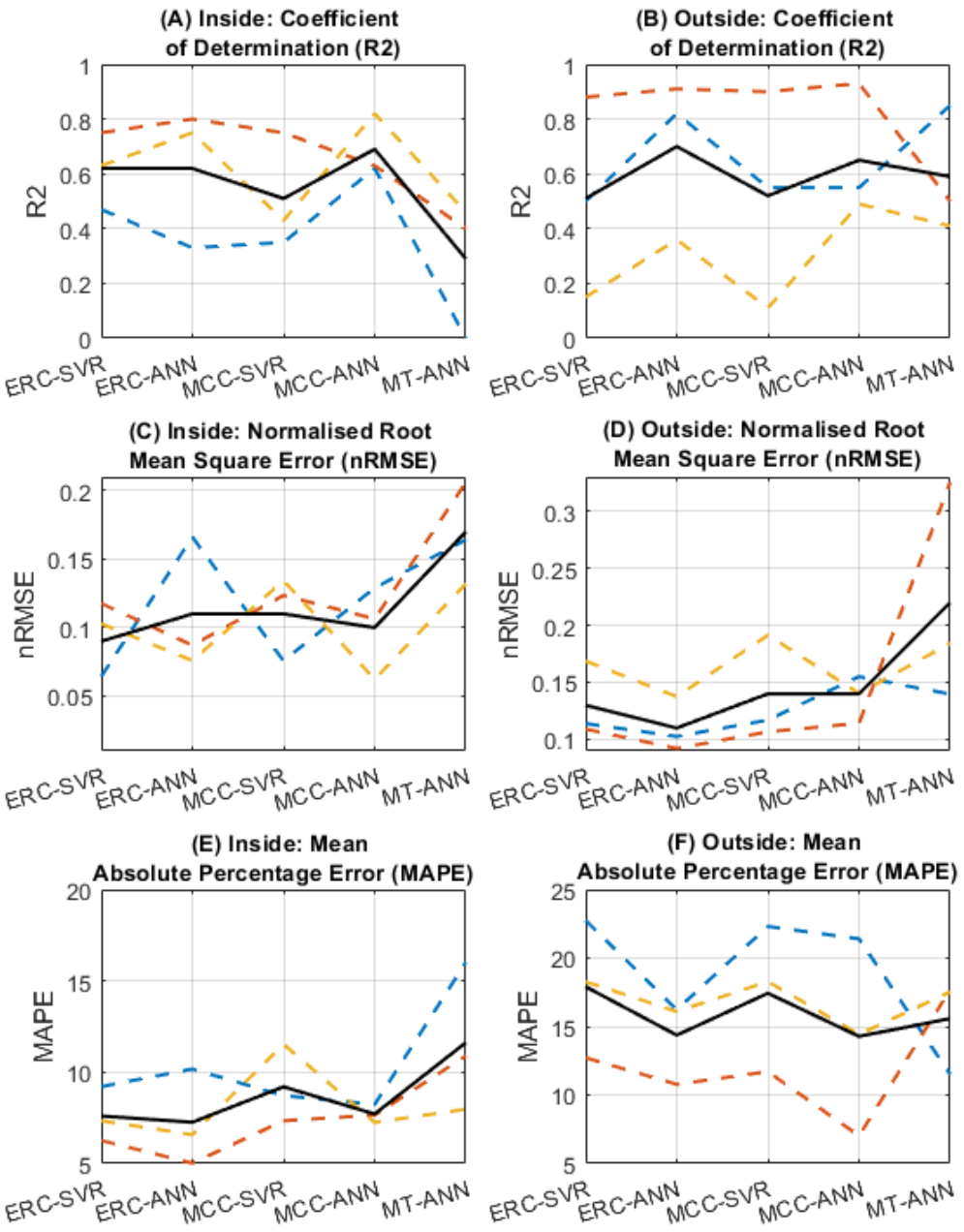


Figure 5.20 Statistical analysis of algorithm adaption MT modes performance on unseen data. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 output.

When evaluating the models' abilities to make predictions on unseen data within the system, it is immediately noticeable that the RC chain models have a much closer performance than when comparing their performance on the model development data. Although, the MT-ANN model proves to be significantly worse than all RC models, emphasising the point raised earlier about the tendencies or larger ANN architectures to overfit. This highlights the advantages of the algorithm adaptions techniques developed by Spyromitros-Xioufis *et al.* and Melki *et al.* when building a MT model [126,365].

The ERC-ANN model, is no longer a clearly superior model, due to its relatively poor performance when predicting the %COD removal. The ERC-SVR, ERC-ANN and MCC-ANN average performance are all relatively equal when making predictions for unseen data inside the system, each performing slightly better for a different statistical metric, ERC-SVR (nRMSE = 0.9), ERC-ANN (MAPE = 7.5) and MCC-ANN (R^2 = 0.64).

Comparing the models' performance when making predictions on unseen data outside of the system serves two purposes. Not only does it evaluate the models' capabilities to extrapolate beyond the data they were trained on, but also to extrapolate beyond boundaries of the model development data. This is necessary as process manufacturing systems are continuously evolving (e.g. changes in feedstock, suppliers, equipment, operating conditions); therefore, it is necessary to understand the model capabilities to adapt to shifts in the system operating region. As expected, there is decrease in performance when making predictions on unseen data, represented by an increase in the average nRMSE and MAPE errors for all models. The ERC and MCC perform similarly for their respective SVR and ANN base models. The RC models that used ANN as a base model were superior to those that used SVR for the R^2 and MAPE statistic metrics. In order to determine which out of ERC-ANN model and MCC-ANN model has the strongest predictive capabilities, boxplots produced comparing the models' predictions against the unseen data, shown in Figure 5.21.

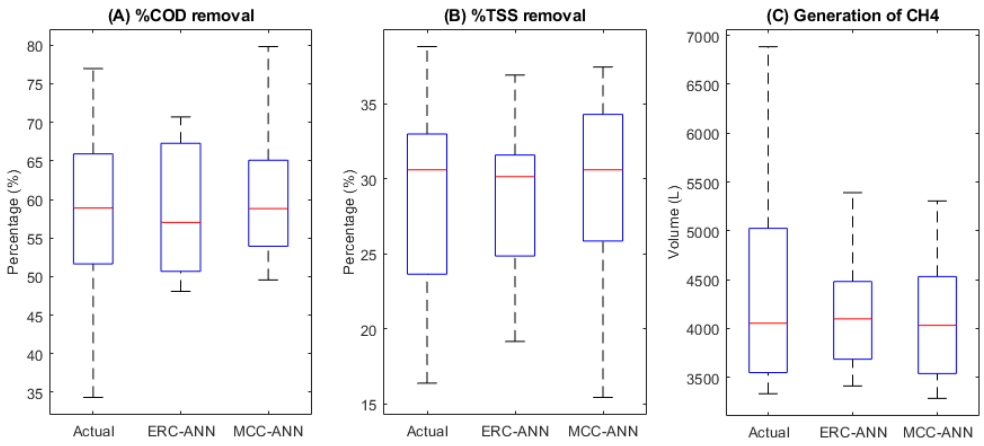


Figure 5.21 Boxplots comparing the distributions between the actual unseen data and the models' predictions for each output.

Overall, the MCC-ANN is judged to be the best model when predicting %COD removal, %TSS removal and CH_4 generation. This conclusion was reached by the fact that MCC-ANN, was among the best performing models for both the unseen data inside and outside the system. Though the average performance of the ERC-ANN model was

similar and sometimes better, than the MMC-ANN, it was not chosen because of its inferior predictions capabilities of %COD removal inside the system. When comparing the distributions in Figure 5.21, the MCC-ANN predictions are aligned closer to that of the actual data than the ERC-ANN predictions are. Furthermore, Melki *et al.* demonstrated the increased run times when training an ERC model compared to an MCC model [126]. This may prove a significant factor to process manufacturers monitoring their systems with a MT model, as DDM model requires monitoring and retraining to prevent a gradual degradation in performance. The decrease of the prediction quality is caused by the gradual and abrupt changes in the process [310]. This can be avoided by retraining the model periodically as the availability and collection of data increases.

5.6.3.2 Utilising a model to optimise H^2AD bioprocess sustainability

Once built and evaluated, the selected MT model (MCC-ANN) can be utilised to assess the impact the H^2AD process conditions has on the H^2AD bioprocess. This was achieved by supplying the MCC-ANN model with a matrix of process data, containing constant values for the wastewater characteristic but various process condition values. Figures 5.22 – 5.24 shows the effect of the four investigated process conditions on the removal of COD, removal of TSS and generation of CH_4.

5.6.3.2.1 Understanding the relationship between process conditions and pollutant remediation and energy generation

It may be discovered through analyses of Figures 5.22 and 5.23, that the optimal process conditions are different for the %COD removal and %TSS removal. Observations from Figures 5.22(A-C), show that the lowest value of HRT (3.52 days) achieves best % COD removal rates. The CH_4 generation displays the same relationship with HRT as %COD removal in Figures 5.24 (A-C). Whereas, in Figures 5.23(A-C) the relationship between %TSS removal and HRT is more complex. The optimal %TSS removal occurs at the highest HRT value (6.09 days), the %TSS removal then decreases with HRT until a midpoint before increasing again. If this was real data, rather than synthetic, it could be hypothesized that because at low HRT values the H^2AD is frequently provided with wastewater providing more fuel for the bioprocess, hence high %COD removal and CH_4 generation, with good %TSS removal. However, as the HRT increases less fuel is provided decreasing the %COD removal and CH_4 generation and also %TSS removal up until a point. After this point %TSS removal increases as the TSS in the system begins to settle out.

When trying to identify the optimal pump speed for %COD removal and %TSS removal, a conflicting relationship was observed in the surface plots. Figures 5.22(A, D, E) show the best %COD removals are observed above pump speeds of 9.45 (l/min); whereas, in Figures 5.23 (A, D, E) the best %TSS removals are observed above pump speeds of 9.45 (l/min). Both values appear to show that pump speed has little effect to either output above 9.45(l/min). The results of pump speeds above 9.45 (L/min) produce the optimal configurations to generate CH_4. The results in Figure 5.22(F) and Figure 5.24(F), indicate a maximum point for both %COD removal and CH_4 generation, is achieved at a system temperature of 34.5^{o}C. The performance of each output decreases on temperatures above and below this. However, Figure 5.23 (F) shows that better %TSS removal rates can be achieved at increasing the temperature above this point. The optimal process condition value for each output is displayed in Table 5.8. While, these values can be useful to determine the optimal process conditions, the relationships within the surface plots are also useful when models outputs come into conflict (e.g. %COD and %TSS removal), to find values that maximise both outputs.

Table 5.8 Optimal process conditions values for each of the H^2AD bioprocess outputs

Process condition	**%COD removal**	**%TSS removal**	**CH_4 generation**
HRT (days)	3.52	6.09	3.52
Pump speed (l/min)	9.45	8.95	9.94
ST (^{o}C)	34.5	39.20	34.5
SP (bar)	0.003	0.005	0.003

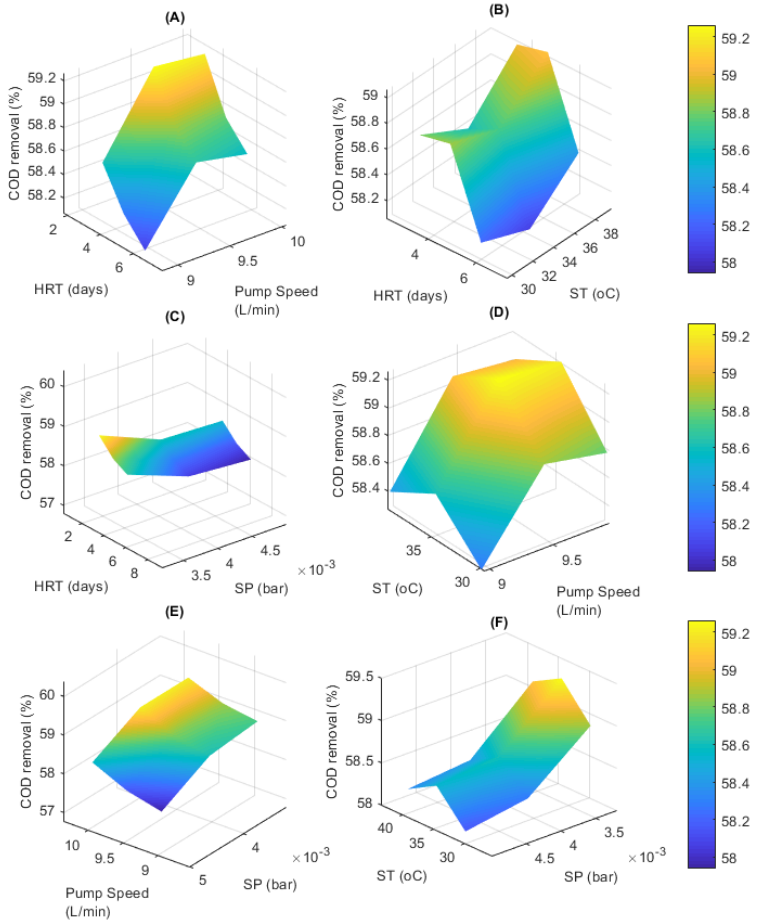


Figure 5.22 Surface plots of the H^{2}AD process conditions impact on percentage of COD removal.

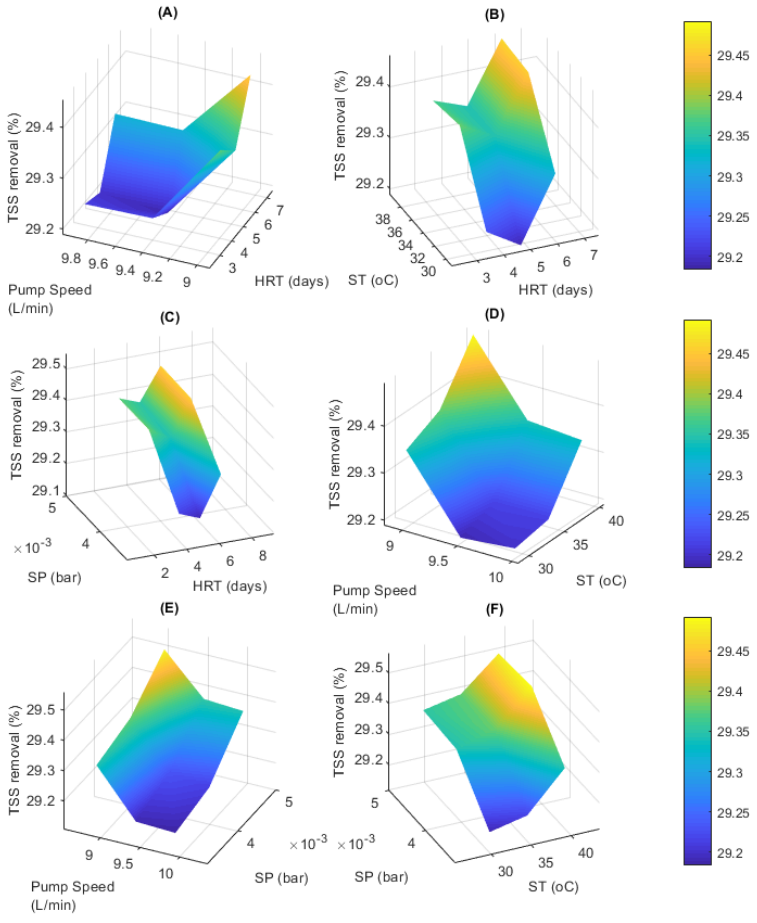


Figure 5.23 Surface plots of the H^{2}AD process conditions impact on percentage of TSS removal.

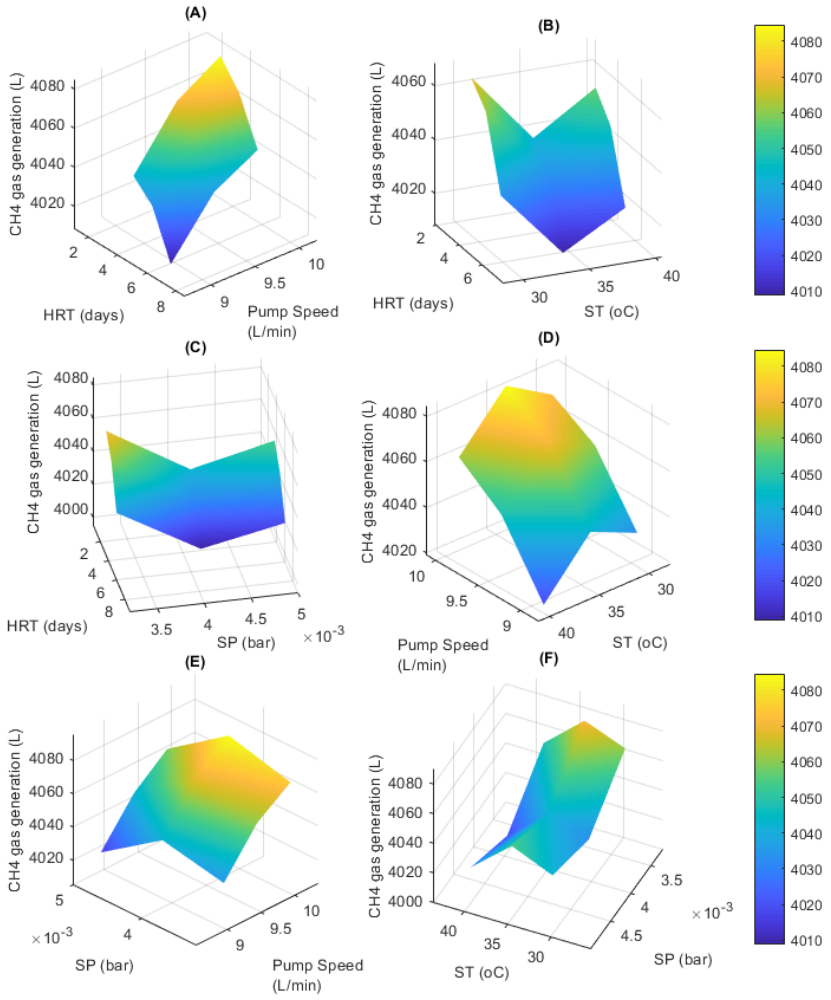


Figure 5.24. Surface plots of the H^{2}AD process conditions impact on the generation of CH4.

5.6.3.2.2 Determining the most sustainable configuration of the H^{2}AD system

After analysing the impacts of the process conditions on the H^{2}AD bioprocess, it became apparent when attempting to optimise for %COD removal and CH_4 generation, %TSS removal will decrease. Finding the pre-eminent solution will depend on maximising the economic and environmental benefits when treating a manufacturer's wastewater stream.

The environmental objective of the system is to minimise the pollutants leaving the system and maximise the energy recovered from the waste. This closely matches the economic objective of the H^{2}AD system to (A) reduce waste disposal costs associated with manufacturing process by reducing the pollutant load in wastewater and (B) reduce the cost of energy by recovering energy from the waste stream. However, there are also economic costs associated with operating the H^{2}AD. These are largely from the required energy to heat the incoming wastewater to the required system temperature and maintaining system temperature. These may come into conflict with each other, as

operating the H^2AD system on higher temperatures was found to increase both pollutant removal and energy generation. The extra cost associated with at operating at higher temperatures may be greater than the economic benefits from the savings generated from the reduced pollutant load and onsite energy generation.

The most sustainable solution will differ between the manufacturing environments and the wastewater characteristics. When the H^2AD effluent is being emitted to a waste management body that charges for pollutant load, is the driving economic force of the system, the goal would be to maximise pollutant removal regardless of bioenergy generation. As the results from section 3.2.1 indicate that %COD removal and %TSS removal have different optimal conditions, the process conditions configuration would depend on the characteristics of the wastewater being treated by the H^2AD system. However, if pollutant load charges in the effluent are lower the optimal economic solution may be to maximize energy recovery from the waste stream.

The MCC-ANN model was used to determine the economic benefit of optimising the H^2AD configuration dependant on wastewater composition and further treatment costs. Four different wastewater composition scenarios were proposed using different high and low values of COD and TSS concentrations, the values are presented in Table 5.9. The low and high values were taken as the maximum and minimum values observed in the model development data. The remaining wastewater characteristics were held constant and an intermediate value. The two H^2AD configuration were (1) optimising for COD% removal and CH_4 generation (configuration 1) and (2) optimising for TSS% removal. The H^2AD process conditions were taken from for each configuration were taken from results of the surface plot modelling in Table 5.8.

The MCC-ANN model predicted the COD and TSS removal rates and the volume of biogas generated for each configuration for the four scenarios. The MAPE error metric was used to estimate error boundaries for the model's predictions. The removal rates were then used to calculate the pollutant concentration in the treated wastewater, using Equation 5.7. The maximum and minimum pollutant concentrations were calculated using the MAPE metric in Equation 5.8. The results for each scenario and H^2AD configuration are presented in Table 5.9.

$$x_p = x - x \times r_p$$

Equation 5.7

$$x_p = x - x \times r_p \times (1 \pm MAPE)$$

Equation 5.8

Where x_p is the predicted pollutant concentration in the treated wastewater, x is the pollutant concentration in the untreated wastewater, r_p is the predicted pollutant removal rate by the H^{2}AD bioprocess and *MAPE* is the mean absolute percentage error associated with the prediction.

Economic analysis was conducted to identify which H^{2}AD configuration was the most sustainable for each scenario dependant on if the wastewater was being sent to a wastewater treatment company that charge a high amount (United Utilities) or a low amount (Wessex Water). The disposal charges were calculated for each scenario using the Mogden formula, assuming 4m^3/day of wastewater was being treated [361]. The energy value from the methane gas was calculated by using Equation 5.9 from Bauer *et al.* [362] and then converted to kilowatt hours (kWH). A biogas generator is not capable of converting all this energy to electricity, as approximately 65% of the energy is lost as heat and other mechanical losses [362]; therefore, the predicted energy generated by the H^{2}AD was reduced by 65% to account for inefficiencies. The average UK price of £0.124 per kWH at time of writing was used to determine the economic benefit from the biogas [373].

$$1\ m^3\ of\ CH_4 = 39.79\ MJ$$

Equation 5.9

The economic analysis of the MCC-ANN model results for configuration 1 and 2 are presented in Tables 5.10 and 5.11 respectively. The analysis indicated that for all but one of the manufacturing environments, the optimal H^{2}AD configuration is to maximise the removal of COD and generation of CH_4. The largest economic benefit (£30.00 per day) due to optimising the H^{2}AD configuration was for high COD and low TSS concentrations in the wastewater being treated by United Utilities. Optimising the H^{2}AD configuration for this manufacturing environment would save the process manufacturer £10,950.00 per annum. The predictive analysis indicates that there is only one manufacturing environment where the H^{2}AD process should be configured to maximise TSS removal (configuration 2); that is when wastewater is being treated by United Utilities and has a high TSS concentration and a low COD concentration.

Optimising the H^2AD configuration would save the process manufacturer £1,825.00 per annum in this scenario.

Table 5.9 Chemical oxygen demand (COD) and total dissolved solids (TSS) concentration in the H^2AD input and output streams for each wastewater composition scenario for both H^2AD set ups.

			Configuration 1		Configuration 2	
Scenario	**Pollutant**	**Pollutant conc. (mg/L)**	**Pollutant conc. (mg/L)**	**Methane generated (L/day)**	**Pollutant conc. (mg/L)**	**Methane generated (L/day)**
1	COD	40,000	11900 (+/-19.5%)	5270 (+/-7.3%)	20500 (+/-7.9%)	4040 (+/-7.3%)
	TSS	30,000	24100 (+/-1.9%)		21000 (+/-3.3%)	
2	COD	40,000	12300 (+/-18.6%)	5250 (+/-7.3%)	20500 (+/-8.2%)	4130 (+/-7.3%)
	TSS	13,000	10600 (+/-1.8%)		9060 (+/-3.3%)	
3	COD	17,000	5070 (+/-19.3%)	5240 (+/-7.3%)	8740 (+/-7.8%)	4010 (+/-7.3%)
	TSS	30,000	24100 (+/-1.9%)		21000 (+/-3.3%)	
4	COD	17,000	5070 (+/-19.4%)	5120 (+/-7.3%)	8740 (+/-8.1%)	4000 (+/-7.3%)
	TSS	13,000	10600 (+1.8%, -1.7%)		9060 (+/-3.3%)	

Table 5.10 Economic analysis of the wastewater disposal costs before and post H^2AD treatment and the economic benefit from the methane generated for **Configuration 1**.

Wastewater company	**Scenario**	**Disposal charge with no treatment** £/day	**Disposal charge post H2AD** £/day	**Saving** £/day	**Value from methane** £/day	**Total economic benefit** £/day
United Utilities	1	415	240 (+/-5.8%)	175 (+/-8.1%)	10 (+/-7.3%)	185 (+/-8.1%)
	2	285	140 (+/-8.6%)	145 (+/-8.2%)	10 (+/-7.3%)	155 (+/-8.1%)
	3	310	210 (+/-3.8%)	100 (+/-8.0%)	10 (+/-7.3%)	110 (+/-7.9%)
	4	180	105 (+/-5.6%)	75 (+8.1%, -8.0%)	10 (+/-7.3%)	85 (+8.0%, -7.9%)
Wessex Water	1	160	85 (+/-7.0%)	75 (+/-8.1%)	10 (+/-7.3%)	85 (+8.1%, -8.0%)
	2	120	55 (+/-9.8%)	65 (+/-8.2%)	10 (+/-7.3%)	75 (+/-8.1%)
	3	110	70 (+/-4.5%)	40 (+/-8.0%)	10 (+/-7.3%)	50 (+/-7.9%)
	4	70	40 (+/-6.5%)	30 (+8.1%, -8.0%)	10 (+/-7.3%)	40 (+8.0%, -7.9%)

Table 5.11 Economic analysis of the wastewater disposal costs before and post H^2AD treatment and the economic benefit from the methane generated for **Configuration 2**.

Wastewater company	**Scenario**	**Disposal charge with no treatment** £/day	**Disposal charge post H2AD** £/day	**Saving** £/day	**Value from methane** £/day	**Total economic benefit** £/day
United Utilities	1	415	255 (+/-4.9%)	160 (+/-8.0%)	5 (+/-7.3%)	165 (+/-8.0%)
	2	285	165 (+/-6.0%)	120 (+/-8.1%)	5 (+/-7.3%)	125 (+8.0%, -8.1%)
	3	310	205 (+/-4.2%)	110 (+/-7.9%)	5 (+/-7.3%)	115 (+/-7.8%)
	4	180	110 (+/-5.0%)	70 (+/-8.0%)	5 (+/-7.3%)	75 (+/-7.9%)
Wessex Water	1	160	95 (+5.3%, -5.6%)	65 (+8.6%, -8.0%)	5 (+/-7.3%)	70 (+8.5%, -8.0%)
	2	120	65 (+/-6.3%)	50 (+/-8.1%)	5 (+/-7.3%)	55 (+/-8.0%)
	3	110	70 (+/-4.3%)	40 (+/-7.9%)	5 (+/-7.3%)	45 (+/-7.8%)
	4	70	45 (+5.2%, -5.1%)	25 (+8.0%, -8.1%)	5 (+/-7.3%)	30 (+/-7.9%)

5.7 CONCLUSIONS FROM PART B

- When searching for the most sustainable manufacturing solution, multiple economic and environmental objectives must be considered. The objectives often come into conflict with one another. By developing a MT model of the manufacturing system, one would be able to investigate the impact of various manufacturing factors on the objectives and understand how the objectives relate to one another.
- In this study, two state-of-the-art MT modelling approaches were applied to a manufacturing system for the first time. These were RC based techniques called ERC and MCC. The RC technique treats the outputs of the system as inputs for subsequent outputs, thus capturing the relationship between model outputs. Previous work has used only RF and SVR algorithms as base models for RC models. This study was the first example of an ANN base model being used for either an ERC or a MCC MT model.
- Utilising the MCC-ANN model to develop surface plots of %COD removal, %TSS removal and CH_4 generation, it was found that the %COD and CH_4 generation favoured closely similar configurations of the H^2AD system's process conditions. However, the %TSS removal was found to have the opposite relationship with all process conditions apart from the system temperature.
- The optimal H^2AD process conditions for COD% removal were determined as HRT of 3.52 days, pump speed of 9.45 l/min, a system temperature of 34.5^{o}C and a system pressure of 0.003 bars. The optimal H^2AD process conditions for TSS% removal were determined as HRT of 6.09 days, pump speed of 8.95 l/min, a system temperature of 39.2^{o}C and a system pressure of 0.005 bars. The optimal H^2AD process conditions for CH_4 generation were determined as HRT of 3.52 days, pump speed of 9.94 l/min, a system temperature of 34.5^{o}C and a system pressure of 0.003 bars.
- What is sustainable for one manufacturing environment may not be for another, as different environments have different economic and environmental objectives. The MCC-ANN model was used to perform economic analysis for four different wastewater composition operating in two different manufacturing environments.
- An economic benefit up of £10,950 was achieved from using the MCC-ANN model to find the optimal H2AD configuration for a specific manufacturing environment.

CHAPTER 6: APPLYING THE DEVELOPED DATA-DRIVEN MODELLING METHODOLOGY TO A REAL-WORLD PROCESS MANUFACTURING SYSTEM

6.1 HIGHLIGHTS

- The treatment of acid mine drainage from abandoned mines in the UK is the responsibility of the Coal Authority.
- Current treatments methods are deemed to be capital cost-intensive, so alternative treatment methods are explored as a route to economic sustainability.
- Three different treatment methods were trialled by the Coal Authority were found to have a high level of variability in iron (Fe) removal rates.
- A data-driven modelling approach was utilised to investigate the variability within the system.
- Two different modelling strategies were trialled, one using only the dataset from the specific treatment method, the other using all three datasets.
- The GP model trained on only the one dataset was found to be the best (R^2 of 0.86, RMSE of 3.75, MAE of 2.78).

6.2 INTRODUCTION

The previous two chapters described the considerations and challenges when attempting to model a process manufacturing system, before developing a methodology for how data-driven models (DDMs) may be developed for process manufacturing systems. The work was based on developing a DDM model to describe Lindhurst Engineering Ltd.'s H^2AD system. However, each process manufacturing system has its own set of considerations and challenges when attempting to model the system. Therefore, in order to validate the proposed methodology, the modelling techniques developed were applied to a new process manufacturing system. Furthermore, moving from modelling the H^2AD system to a new case study, demonstrates the flexibility of the developed modelling approach.

This chapter is divided into four sections. The first section describes the real-world process manufacturing case study: the treatment of mining wastewater to remediate iron (Fe) pollutants by the Coal Authority. A modelling approach aims to understand

variability in Fe removal rates observed in the system. The second section discusses the considerations and challenges in modelling Fe removal from wastewater using the historical data available. The third section explores the historical manufacturing data, to understand the nuances contained within the system to establish the system boundaries. The fourth section provides the results of DDM and compares two different modelling strategies when attempting to accurately portray the process manufacturing system. The final section describes how the DDM model may be used to understand how variations in the system's feedstock affect the efficiency of the system.

6.3 OVERVIEW OF THE PROCESS MANUFACTURING SYSTEM MODELLED

The mining industry is one of the oldest documented human activities and provides the raw material for various process manufacturing industries (e.g. energy, chemical, pharmaceutical) [374]. Until mine closure regulations were introduced in the UK, mines that have had their mineral wealth extracted were often abandoned. There are abandoned mines located across the globe that are sources of environmental hazards, such as concealed entrances, workings that underlie buildings, streets and dams and acid mine drainage (AMD) [375]. Of these the most persistent and damaging environmental hazards is AMD [375], which occurs as sulphide mineral ores that have been exposed through mining activities, are oxidised [376]. This leads to high concentrations of potentially toxic dissolved metals, metalloids and sulphates, as well as acidification of the water [376]. This causes serious environmental threats, particularly on soil, water resources and aquatic ecosystems [376]. Among the dissolved metals, iron in +2 oxidation state (Fe(II)) is the most copious and commonly found metal in AMD. The Fe(II) in AMD reacts with DO to produce Fe oxide precipitates, which harms aquatic life by embedding on stream or ocean beds [376]. The EU Water Framework Directive (WFD, 2000/60/EC) required European countries to achieve good chemical and ecological water quality status by 2015. Though this ambitious target was not met [377], it has led to recognition of the need to treat the AMD being produced by abandoned mines before it is realised to the surrounding environment [378].

There are two approaches to treating AMD; chemical neutralisation (lime addition) resulting in precipitation of metal hydroxides or biological processes [379]. The most commonly used method is chemical neutralisation; however, this method has high operational costs, large solid generation and hard water disposal disadvantages, so alternative biological processes are being explored [379]. In the UK, the Coal Authority was established as a non-departmental government body responsible for abandoned coal mines across the UK [380]. The Coal Authority typical employ a passive biological

process for treating mining wastewater: aeration cascades, settling lagoons and aerobic wetlands [380]. The aeration cascades encourages the oxidation of Fe(II) to Fe(III) (Equation 7.1) followed by the hydrolysis of Fe(III) into $Fe(OH)_3$ (Equation 7.2), which precipitates and gradually settles out of the wastewater in the settling lagoons [381]. The remaining finer Fe particles are filtered out when the water passes through a series of reed beds.

$$Fe^{2+} + 0.25O_2 + H^+ = Fe^{3+} + 0.5H_2O$$

Equation 6.1

$$Fe^{3+} + 3H_2O = Fe(OH)_3 + 3H^+$$

Equation 6.2

The main advantage of this passive treatments is that they have lower operationally cost, due to requiring no electrical energy or external additives and a smaller amount of manual supervision [381]. Furthermore, a passive treatment system has operational lifespan and management period that makes it suitable for remote sites such as abandoned mines [381]. However, a passive treatment system's remediation efficiency is proportionally dependent on the size of the site [381], which can make them unfeasible when there is limited affordable land available.

Recent studies have evaluated alternative treatment technologies; for example, adsorption, inorganic and organic ion exchange, microfiltration, reverse osmosis, solvent extraction and ultra-filtration [382]. However, these technologies are currently too expensive to be applied at a large scale [383]. Development of bioprocess reactors utilising sulphate reducing bacteria (SRB) under anaerobic conditions has emerged as a feasible method for treating AMD [379]. Hurtado *et al.* demonstrated how a SRB reactor may be integrated into a biological treatment process to improve economic performance when removing sulphate and metals from AMD [383]. The SRB process removes metals through metal sulphide precipitation.

There is an opportunity for waste valorisation by recovering valuable metals by separating the insoluble metal sulphide from the liquid phase [379]. Reactors with two or more stages allow for the selective recovery of metals from AMD [379]. Acid mine drainage contains various metals (e.g. Zn), some of which are on the EU's list of critical raw materials [5].

Previous work has demonstrated the application of DDM to predict the performance of different AMD treatment systems [384,385]. Sahinkaya built an ANN model to predict concentrations of sulphate, chemical oxygen demand (COD), acetate and Zn a SBR

continuous stirred tank reactor effluent [384]. By developing a DDM, Sahinkaya aimed to automate the alkalinity addition to maintain good process performance in a Zn recovering SBR reactor system [384]. Atasoy *et al.* developed an ANN model to predict the performance of both up-flow and down-flow fluidised bed reactors treating AMD [385]. The reactors performance by effluent sulphate, COD, alkalinity, and sulphide concentrations [385]. The authors compared two modelling strategies, the first using both influent and effluent concentrations as model inputs, and the second using only influent concentrations as model inputs. By using effluent parameters as model inputs, the authors aimed to predict reactor performance at different operational conditions without conducting additional experiments [385]. Both these studies were able to obtain good fitting between the model development data and model's predictions, all the models developed had a coefficient of determination (R^2) less than 0.9. However, none of the models was tested on unseen data, meaning that their ability to extrapolate beyond the model development data was not evaluated.

7.3.2 PROPOSED TECHNOLOGY

The Coal Authority has conducted studies to evaluate the potential of surface catalysed oxidation of ferrous iron (SCOOFI) technology to replace existing settling lagoons technologies in the treatment of mining wastewater. New treatment technologies are being considered because of the high footprint cost associated with settling lagoons. A SCOOFI reactor contains a porous plastic media with a very high surface area ratio, as oxygenated mine wastewater flows through the media it precipitates onto the surface [380]. A SCOOFI reactor is an example of a packed-bed reactor, which are commonly used for wastewater treatment [331]. Previous work by the Coal Authority showed that the SCOOFI achieved Fe removal rates up to 139 g/m^3 per day, suggesting the system could be potentially used in areas where space limits the open lagoons or settlement ponds [380]. However, the system was deemed not suitable because of two maintenance issues; first the unstable nature of the precipitate and second the system quickly became blocked by precipitate [380]. Further trials were commissioned to investigate different system configurations to overcome these barriers.

A trial conducted between May 2018 and March 2019, compared two different SCOOFI reactor configurations. The reactors were made from Enduramaxx 2.7 m^3 degree cone bottom tanks. The packing material Veolia Cascade Filterpak, which is a high-efficiency plastic media that provides the optimum effective surface area for Fe precipitation [386]. The configurations differ based upon how the wastewater enters and leaves the reactor. Packed bed reactors can be operated in either down-flow or up-

flow mode and the dosing can be continuous or intermittent (e.g. trickling filter) [331]. Both down-flow and up-flow configurations were compared against each other and a down-flow reactor with no packing material. The three reactor configurations are displayed in Figure 6.1.

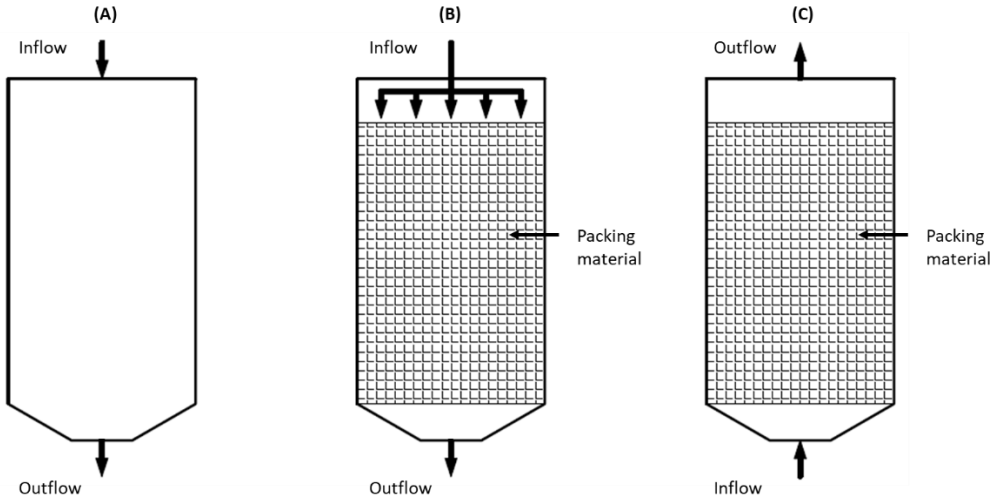


Figure 6.1. Three different surface catalysed oxidation of ferrous iron (SCOOFI) reactor configurations. (A) Is a down-flow reactor with no packing material, (B) is a down-flow trickling reactor and (C) is an up-flow reactor.

Down-flow trickling reactors have been extensively used by industry for many years [387]. However, when plotting the results of the trial in Figure 6.2, the up-flow SCOOFI reactor appears to have the best overall performance at treating the mining wastewater for the remediation of Fe. This was confirmed by measuring the average Fe removals for each reactor (27.58% compared to -7.85% for down-flow no packing and 7.27% for down-flow packing). This increase in performance may be attributed to the fact that up-flow reactors are known to present advantages from the point of view of liquid distribution, wetting efficiency and thermal stability [387].

Comparing the SCOOFI reactors with packing material to one without, shows that the reactors with packing material achieve higher rates of Fe removal. From inspection of the time series data in Figure 6.2, all three reactors exhibit periods of similar peaks and troughs with respect to Fe removal. This demonstrates the Fe removal rates are influenced not only by the reactor configuration but also, the variability present in the mining wastewater stream. A DDM was built to both understand and fit data to predict the impact the mining wastewater characteristics have on the SCOOFI reactor's efficiency at remediating Fe.

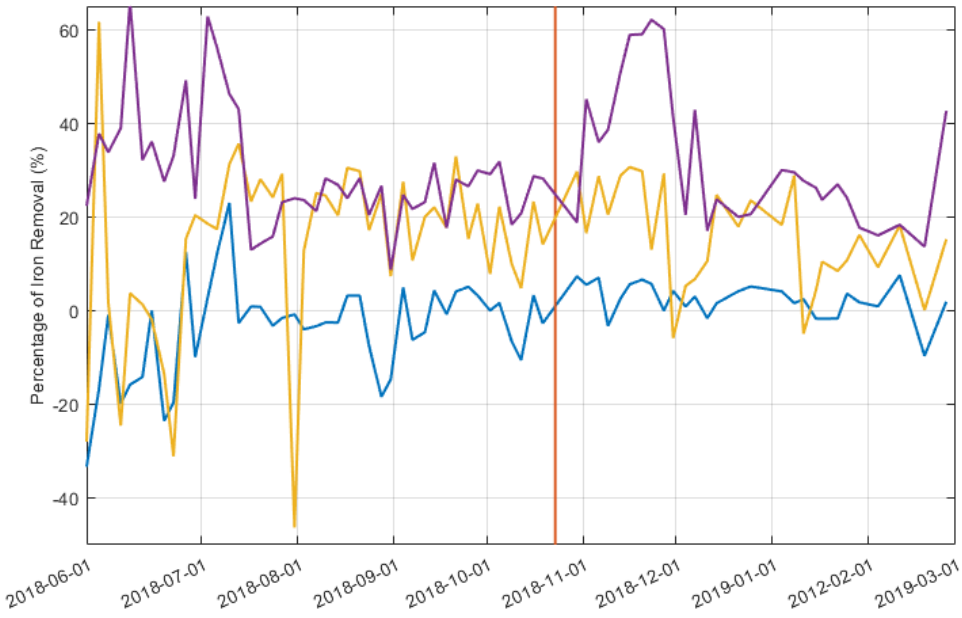


Figure 6.2: Iron removal results from a trial conducted by Coal Authority. The down-flow reactor no packing material, down-flow trickling reactor and up-flow reactor results are represented by the blue, yellow and purple lines respectively. The red line represents the desludging day.

6.4 THE CONSIDERATIONS AND CHALLENGES WHEN MODELLING FE REMOVAL CASE STUDY

6.4.1 CONSIDERATIONS WHEN MODELLING FE REMOVAL CASE STUDY

6.4.1.1 Model goal

The aim of the trial was to determine which of the three SCOOFI reactors is a viable alternative to open lagoons treatment methods for the remediation of Fe in mining wastewater. The Coal Authority supplied data from the trial that ran for 40 weeks and after plotting the percentage of Fe removed (Fe% Removed) a high level of variability in the removal rates was observed, displayed in Figure 6.3. By modelling the system, the goal was to understand the causes of the observed variability and to make predictions on how the incoming mine water characteristics affect the Fe% Removed. Three different SCOOFI reactors were investigated as part of the trial. The up-flow reactor was found to have the highest Fe removal rates (STATS), which informed the decision to develop a DDM modelling that reactor.

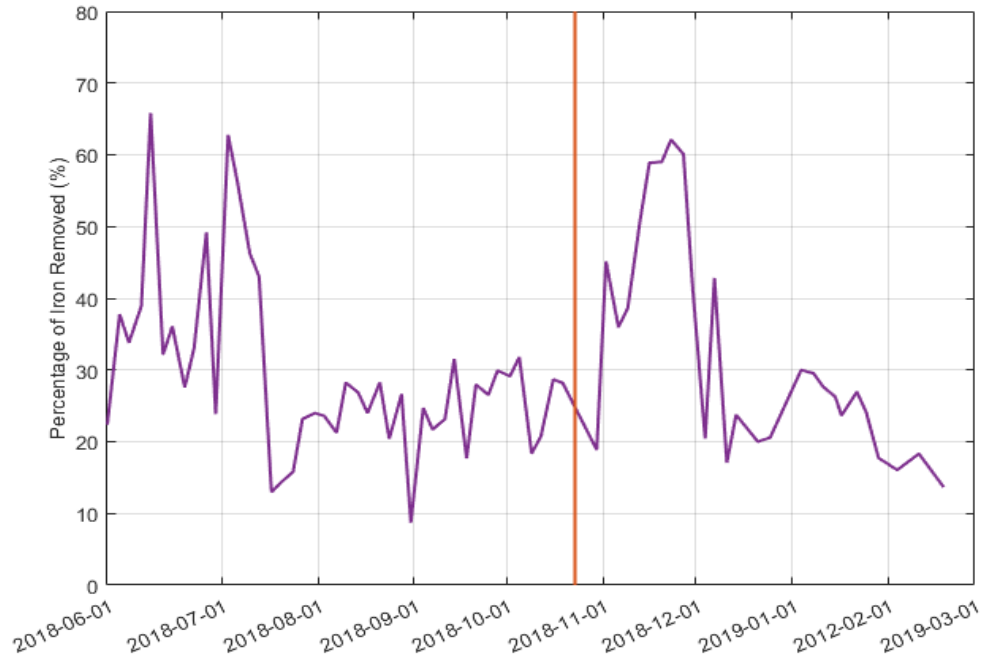


Figure 6.3. Percentage of iron removal by an up-flow surface catalysed oxidation of ferrous iron (SCOOFI) reactor during the trial, with desludging day represented by the red line.

6.4.1.2 Model requirements

The fundamental requirement of a DDM built for a process manufacturing system is to fit manufacturing data to understand relationships within the system. In addition, the DDM may also be used to make predictions about a systems performance and resource use. The DDM developed for this work is required to perform both tasks to:

1. **Fit data** to understand how relationships between mining wastewater characteristics and the system's ability to remediating Fe. Furthermore, to identify which characteristics strongly affect the efficiency.
2. **Predict** how variations in the mining wastewater will affect the system's efficiency to remediating Fe.

A model's required accuracy is determined by product specification, regulation, economic and safety factors. The priority of these factors is influenced by the industry's needs. The model produced by this would not used to make predictions to monitor a final product's quality but to explore relationships within the manufacturing system. Therefore, the accuracy required is lower; however, it must still be high enough to have confidence in the relationships identified by the model.

6.4.1.3 Data considerations

The collection of a high-quality set of data that is representative of the manufacturing system is the most important task in developing an accurate DDM [288]. In chapter 5,

four questions were identified for consideration for collecting a representative set of data:

1. What data already exists, to what granularity (the scale and level of detail in the data) and how representative of the system is the data?
2. What volume of data is required for model development?
3. How was the data collected?
4. What, if any, additional data will the model require and how will it be collected?

There are 74 historical data points of describing the up-flow SCOOFI reactor, though this number may decrease after data preparation, pre-processing and partitioning tasks are performed on the data. This is a greater volume of data available than for either of the models developed in chapter 5, implying that there is a sufficient volume of data to develop a DDM. However, this may not be the case, as the volume of data required is unique to each manufacturing system depending on the volume of data needed to accurately represent the system. In order to try and capture the system's variability, the data was collected over a 40-week period. At this point it is not known whether this is satisfactory to capture the variability or if sufficient volume of data has been collected. Evaluating the final DDM's predictive capabilities on unseen data will inform if more data is required to capture the relationships within the system.

In addition to the data points collected from the up-flow SCOOFI reactor, there are data points collected from the other two SCOOFI reactors trialled. The data collected from both reactors will contain knowledge about how the wastewater's variability affected the SCOOFI system. It may be possible to incorporate this knowledge into the DDM built for up-flow SCOOFI reactor system. To investigate whether the prediction accuracy would improve by incorporating the knowledge of the other two reactors, two modelling strategies were trialled.

1. Developing a DDM with only data relating to Reactor 3 (up-flow configurations)
2. Developing a DDM with data from all three reactors.

The data was collected manually by taking samples of the system's input and output streams. Therefore, it is important to remember during the data preparation stage that there is the potential for human error in the results.

As the model is to be developed from historical data, this limits the possibility of the collection of additional data. However, one additional variable was identified to be required by the model; this was the *number of days since desludging*. During the trial,

the system required desludging, which is performed to reduce the build of solid deposits in the system that reduce performance. This is similar to the backwashing required to clean traditional industrial filter systems [388]. There was a significant increase in Fe% Removed (28.2% up to 62.1%) following desludging, displayed in Figure 6.3. Without including this variable in the model data, the DDM would contribute this 33.9% increase in Fe% Removed wholly to changes in mining wastewater characteristics.

6.4.2 CHALLENGES WHEN MODELLING FE REMOVAL CASE STUDY

6.4.2.1 Availability of data

There is limited data available to develop a DDM for the system. This is because the system being modelled is a trial of a new approach for the remediation of Fe; therefore, there is no additional historical data representative of the current system outside of the data collected from the trial. Moreover, the dataset is further limited because of missing values, incorrect sensor measurements and human error.

6.4.2.2 Variability of data

In addition to the discussed variability in the Fe% removed, the mining wastewater to be treated is also highly variable. Six characteristics (iron (Fe) concentration, total suspended solids (TSS) concentration, flowrate, pH, dissolved oxygen (DO) concentration and conductivity) of the mining wastewater were measured, all of which displayed evidence of variability when plotted in Figure 6.4. This adds the challenge of deciding how much variability should the model aim to capture. For, as discussed in chapter 5, the model risks containing regions within the system that are poorly represented by the data the greater variability is captured within the data. In Figure 6.4 (A), it can be seen that the majority of the Fe concentration in the wastewater is within the range of 10 and 15 mg/l; however, there are two points outside of this range. Removing these points will likely increase the accuracy of the model but limit the boundaries of the system that the model represents. This may reduce the extrapolative ability of the model.

Deciding on the level variability within the historical data to include is one challenge. Another is knowing how much of the system's variability is captured within the data. The trial was run for 40-weeks, so while this may capture 77% of the seasonal variability the additional 13% is not represented. Plus, this is on the assumption that one year's worth of data will capture the seasonal variability, when in fact multiple years are likely to be required. There is also variability between geographical locations. The model's prediction capability will be limited to the mining wastewater produced from the site it the data was collected from. It may; however, be possible to expand the

model's predictions capabilities by including data taken from other manufacturing environments. These ideas are further developed in chapter 8.

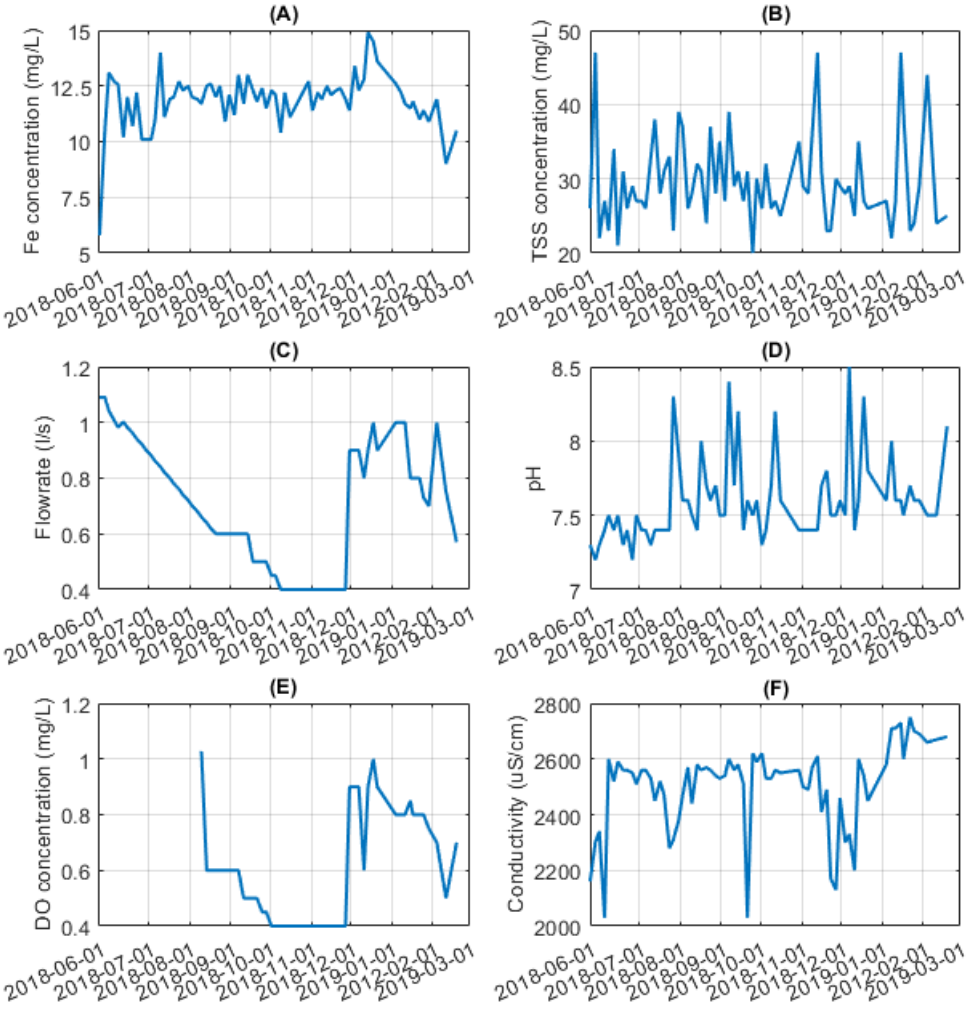


Figure 6.4. Concentration of six characteristics of mining wastewater: (A) iron (Fe) concentration, (B) total suspended solids (TSS) concentration, (C) flowrate, (D) pH, (E) dissolved oxygen (DO) concentration and (F) conductivity.

6.4.2.3 Establishing the system boundaries

Defining the boundaries of the system informs on how much variability will be captured by the model. Figure 6.3 displays the Fe removal rates during the trial. The system appears to operate between four phases across the trial, each with different level of variability:

1. Start until 15/07/18: Fe removal rates vary between 20-65% suddenly, indicating an unstable period during start-up.
2. 15/07/18 – 01/11/18: variability is still present within the system but overall the system operates relatively steadily between Fe removal rates between 10-30%.

3. 01/11/18 – 10/12/18: the system is desludged increasing the Fe removal rates to between 35-60%. The system is more stable at the higher Fe removal rates than during phase one, but there is one occasion where the removal rate dramatically drops.
4. 10-12/18 – end: The system operates steadily between 15-30% Fe removed. Although after the 01/01/2019 there is a gradual decrease in %Fe removed, possible indicating further desludging is required.

This level of variability in the system and the various stages of the system reflect the complexity of the system being modelled. As previously stated, DDM predictions capabilities are generally limited to within the boundaries of the system the model is trained on [98]. Therefore, the boundaries of the system have been defined using statistical and visualisation techniques.

6.4.2.4 Validating the model's prediction capability

A model's predictive capability is validated on unseen data. There are two ways to produce unseen data, either performing additional experiments to generate unseen data or partitioning historical manufacturing data into model development data and unseen data. As the DDM is to be built from historical manufacturing data, the second option was selected for this work. However, feedback from this work could aid the Coal Authority in future pilot scale trials and treatment development. To ensure that unseen data is partitioned from evenly across and beyond the system boundaries, the partitioning method developed in chapter 5 was followed.

6.5 RESULTS AND DISCUSSION

6.5.1 DATA PREPARATION, PRE-PROCESSING AND PARTITIONING

The manufacturing data provided by the Coal Authority (Appendix G) required the execution of several data preparation techniques for the data to be in a condition suitable for modelling. Data is considered suitable when there is no missing values, incorrect measurements or undefined regions. The data contained several challenges common to manufacturing data, which were: sampling frequency, incomplete data, senor malfunctions and accuracy. The immediately noticeable challenge was developing a modelling strategy to handle the missing data points. The conventional techniques to handle missing values within time series data are [389]:

1. *Ignoring*: the missing values are completely ignored, and the analysis carries on.
2. *Deletion*: delete the missing variable or sampling point.

3. *Mean/mode imputation*: the mean or mode of the variable is inputted to replace the missing value.
4. *Last observation carried forward*: this method replaces the missing value with the last observed value.

The last observation carried forward is a commonly used method; although, it introduces bias to the dataset and should be used sparingly, if at all [389]. The DO measurements contained a large section of missing values between 25/05/2018 to 07/08/2018. Since the number of data points missing represented a significant portion of the overall dataset (29.0%), the last observation carried forward was not possible. Instead, the datapoints were deleted. For the one missing Fe input value and the four other missing DO data points, the last observation carried forward was used, as these only represented 1.3% and 5.3% of the data respectively.

The other big challenge within the data was the flowmeter malfunction between 12/06/2018 and 17/08/2018. The same flowrate result of 1.002 was used for all data points. However, the Coal Authority observed a decreasing flowrate during this period due to pipe scaling. Therefore, the values were replaced based upon a monotonic decrease in flow between the 08/06/18 and 21/08/2018, as advised by the Coal Authority. The majority of these data points were deleted as they corresponded to missing DO values and only two values were added.

Up to two additional variables were added to the datasets for each modelling strategy. The first variable added to both datasets was the time since desludging (days). This variable was added to capture the influences the time between desludging has on the system. The second variable, added to only second modelling strategy, was a categorical variable stating whether the data was from reactor configuration 1, 2 or 3. The mine wastewater entering each reactor is virtually identical as they have come from the same source. However, the effluent leaving the reactor is different based upon the configuration of the reactor. Without including this variable, the model will not know which treatment method the wastewater has undergone and is likely to give predictions that are somewhere in between the outputs of all the reactors. The final data preparation technique was to normalise the data based on the minimax function used in chapter 5.

To avoid the *curse* of *dimensionality* the dimensionality reduction technique principal component analysis (PCA) was performed on both datasets, reducing the number of dimensions in each dataset by one. Feature selection was not used for two reasons. The first was that the Coal Authority wanted to investigate how all the mine wastewater characteristics measured effected the remediation of Fe and second, the number of

dimensions is not excessive, as DDM developed with 6/7 input parameters with a similar volume of data have been sufficiently accurate [115,116].

The data was portioned using the method developed in section 5.4. The same unseen data was portioned for both modelling strategies, rather than partition unseen data from across the whole dataset used in modelling strategy 2. This decision was informed by the model's goal to develop a DDM describing the up-flow reactor. It is only the predictions made for the up-flow reactor that are relevant for this work. In Figure 6.5 the final spread of unseen data throughout and beyond the system is shown.

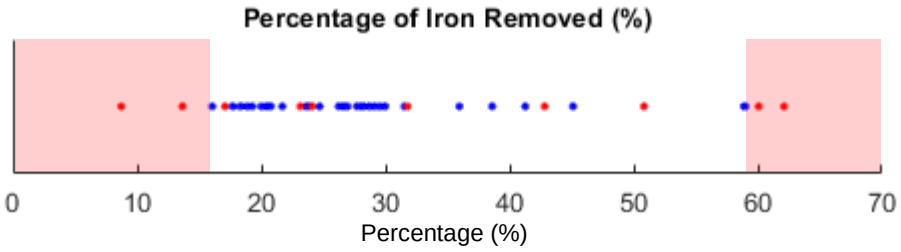


Figure 6.5 Spread of unseen data throughout the system. Blue markers are the model development data points, red markers are the unseen data points and the red regions are outside of the system.

6.5.2 DEFINING THE SYSTEM BOUNDARIES

Data visualisation techniques were used to identify the system boundaries and describe the space the data occupies within the boundaries. These figures were then used to determine the system's features: central tendency, correlations, distribution, spread, modality, and extents. The manufacturing data post cleaning stage was analysed as this is the data that the DDM shall be developed from. However, Figure 6.6 illustrates the data lost during cleaning, from the missing DO values. It is immediately noticeable that the cleaning stage has significantly reduced the number of data points describing the system above the 35% Fe removed. The system appears to perform best during the initial month after start-up and post desludging, average Fe% removed during these periods is 40.59% and 44.50% respectively, compared to 23.51% and 22.35%). This is likely because the reactor has become partially blocked by sludge.

However, as a result of cleaning this section of the system is less represented by the data. When evaluating the final model's predictive capability, special attention should be given to these predictions in this section, in order to establish whether the reduced volume of data has impacted the model's predictive capability in this section. If the model has been dramatically impacted, the option to increase the volume of data available by removing the DO variable from the model should be explored.

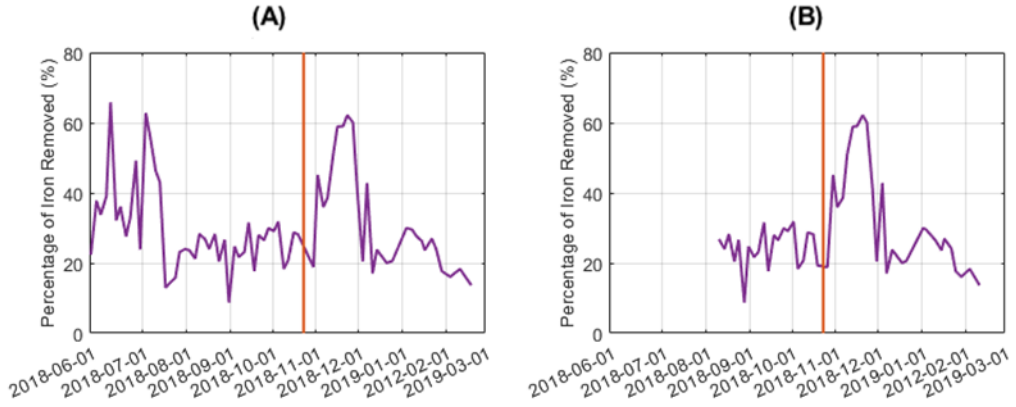


Figure 6.6 Timeseries data of iron removal rates (A) before and (B) after the data cleaning stage.

The central tendency can be used to describe the centre point of each variable within the system. Statistically, this is achieved with either the mean or median and has been calculated for each variable in Table 6.1. For all the variables the mean and medians are within 10% of each other, indicating central tendency are not being skewed by extreme scores. Defining the central tendencies of the system informs what values we expect to observe in new data points collected.

The spread of data around the mean is statistically evaluated by the standard deviation (relative standard deviation was used to compare between variables). The flowrate, DO and TSS have the highest relatives distributions, in that order, and indicate that the data is spread out around the mean. This spread of data is apparent in the boxplots produced in Figure 6.8 (B). The lack of a bottom whisker to the DO and flowrate boxplots indicate that all the 1st quartile values are the same. The scatter plot matrix, in Figure 6.7, reproduces this observation. The heatmap produced displaying Pearson's correlation values between the variables, show that only the DO and flowrate variables are correlated.

Table 6.1 Statistical analysis of the variables in the mining wastewater data provided by the Coal Authority.

Feature	Mean	Median	Maximum	Minimum	Range	Relative Standard Deviation	Skewness
Fe	12.01	12.10	14.50	9.01	5.49	7.16	-0.40
TSS	29.69	28.50	47.00	20.00	27.00	19.98	1.29
pH	7.66	7.60	8.50	7.30	1.20	3.68	1.46
DO	0.61	0.60	1.03	0.40	0.63	31.74	0.51
Conductivity	2538.13	2560.00	2750.00	2030.00	720.00	6.09	-1.52
Flowrate	0.63	0.60	1.00	0.40	0.60	32.90	0.49

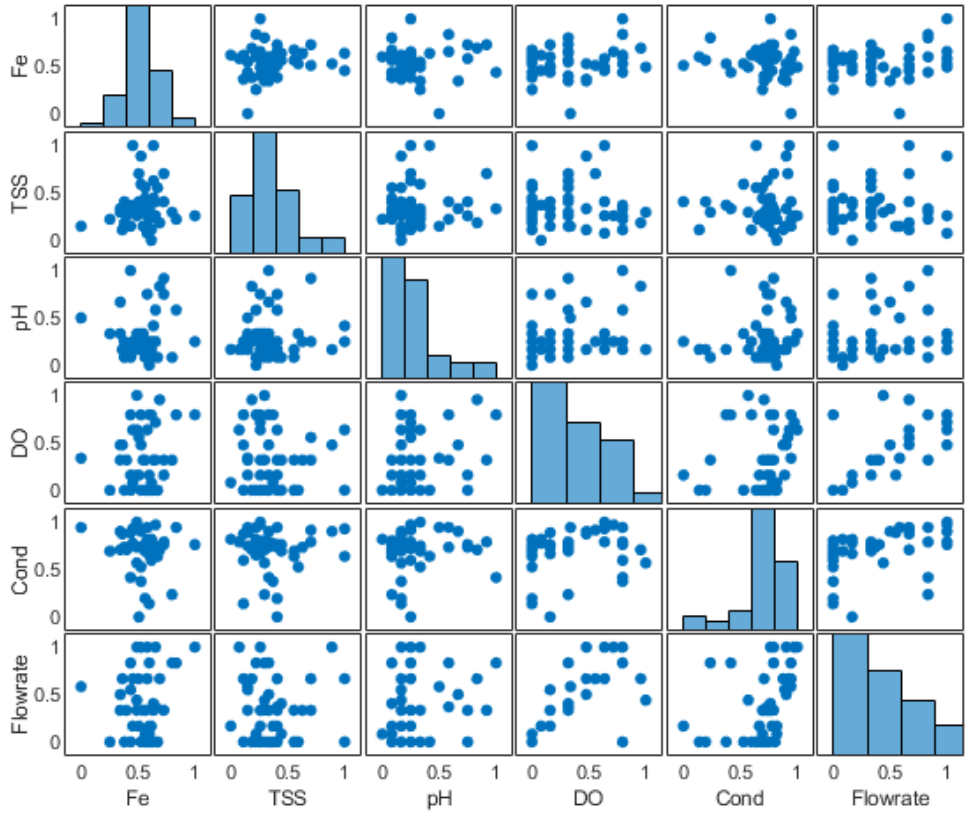


Figure 6.7 Scatter plot matrix of the mining wastewater characteristics

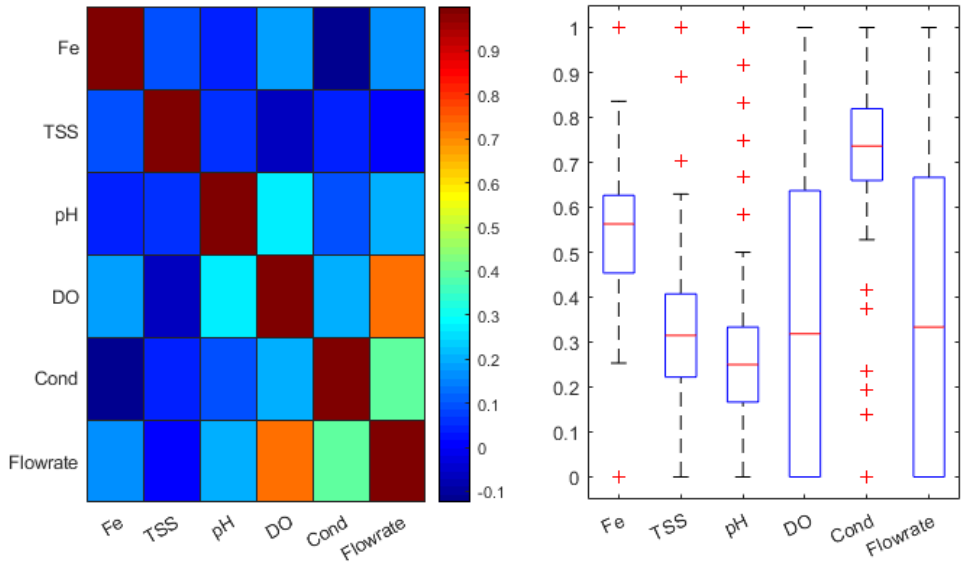


Figure 6.8 (A) Heatmap displaying Pearson's correlation between variables (B) boxplot displaying the distribution of data

6.5.3 DATA-DRIVEN MODELLING RESULTS

Two different modelling strategies were evaluated, with respect to predicting the up-flow SCOOFI reactor's ability to remove Fe from the mining wastewater:

1. Developing a DDM with only data relating to Reactor 3 (up-flow configurations)
2. Developing a DDM with data from all three reactors.

Four DDMs were built for each strategy using different machine learning algorithms, these were Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural network (ANN) algorithms. These algorithms were chosen as they have been shown to have stronger predictive powers over other algorithms like logistic regression [142,349].The model development data was used to determine the optimal hyperparameters and model parameters of the DDMs. The model development data points were randomised and 85% were partitioned for the training and validation of the model and the remaining 15% of data were partitioned for testing the final model. A Bayesian optimisation was performed to optimise the hyperparameters, using k-fold cross-validation to evaluate each configuration's ability to fit the validation data. For the k-fold cross-validation, a k value of 10 was chosen because of the limited number of data points [116]. The models' ability to fit new data was then evaluated using the testing data points. The optimised hyperparameter values for the models developed for modelling strategy 1 and 2 are displayed in Tables 6.2 and 6.3 respectively.

Table 6.2 The hyperparameters values for the machine learning models built for modelling strategy 1

Gaussian Process		**Random Forest**		**Support Vector Regression**		**Artificial neural network**	
H-parameter	Value	H-parameter	Value	H-parameter	Value	H-parameter	Value
Sigma	0.03377	Trees	498	Epsilon	0.004818	Hidden layers	1
Basic function	Linear	Min no. of predictors	4	Box constraint	113.46	Neurons	10
Kernel function	Matern32	Min leaf size	1	Kernel function	Linear	Training function	Levenberg-Marquardt
Kernel scale	1.3566			Kernel scale	NaN	Learning rate	0.007

Table 6.3 The hyperparameters values for the machine learning models built for modelling strategy 2

Gaussian Process		**Random Forest**		**Support Vector Regression**		**Artificial neural network**	
H-parameter	Value	H-parameter	Value	H-parameter	Value	H-parameter	Value
Sigma	0.005888	Trees	507	Epsilon	0.0401	Hidden layers	1
Basic function	Linear	Min no. of predictors	3	Box constraint	118.02	Neurons	12
Kernel function	Rational quadratic	Min leaf size	1	Kernel function	Linear	Training function	Levenberg-Marquardt
Kernel scale	0.001785			Kernel scale	NaN	Learning rate	0.002

The model's ability to fit the data were statistically evaluated using coefficient of determination (R2), the root of the mean squared error (RMSE) and mean absolute error (MAE). The statistically best model is the one that has an R^2 closest to 1 while minimising RMSE and MAE. For equations for R^2, RMSE and MAE see section 5.4. The results are visualised, in Figures 6.9 and 6.10, using a parallel coordinate plot (PCP) and multiple bar graphs. The PCP is suited to identifying the model that has the best performance determined across all the R^2, RMSE and MAE scores, while the bar charts aid with comparing the two modelling strategies. It was not possible to use the MAPE statistical analysis measure used in chapter 5. This was because the model output data included a value of 0 that meant the MAPE measure was undefined [357]. The statistical analysis indicates that the GP model built using data from all three reactor configurations is clearly superior at predicting the Fe removal by an up-flow SCOOFI reactor. The information displayed by Figure 6.9 clearly shows that the model has the highest R^2 score and lowest RMSE and MAE scores. The models built using SVR are statistically the worst models (R^2 0.65, RMSE, 5.94, MAE 4.24), indicating the the SVR machine learning algorithms are not suited to modelling this manufacturing system. The RF and ANN models have a similar performance to one another, and the GP model built from only reactor 3 data.

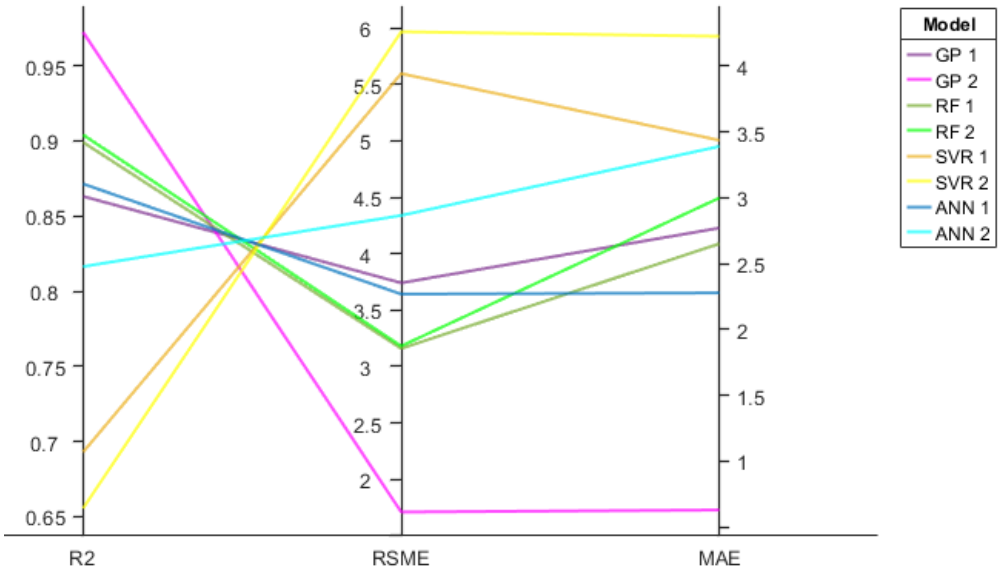


Figure 6.9 Statistical analysis of the models' ability to fit the model development data. The number after the model algorithms states the modelling strategy.

Utilising the statistical analysis to compare the two modelling strategies produced an unclear result. If the modelling strategies are compared only by their respective R^2 scores, the second modelling strategy has clearly benefited the GP model, has made no difference to the RF model and has had a negative effect on the SVR and ANN models. This conclusion is mirrored by the RMSE and MAE analysis; although, the differences between the two scores is more pronounced in the MAE scores. Overall, modelling strategy 1 proved slightly superior, based on the fact it benefited both SVR and ANN models, whilst the modelling strategy benefitted only the GP model. However, before any definitive conclusions are made, the modelling strategies need to be compared on unseen data to see if either modelling strategy increases the models' extrapolating capability.

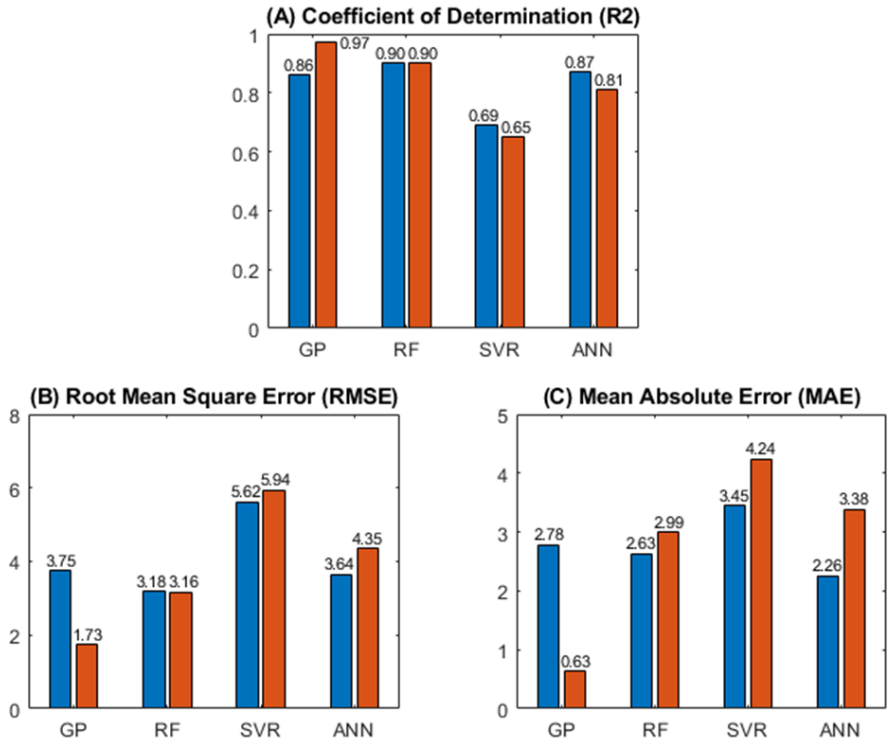


Figure 6.10 Comparing the statistical analysis results of two different modelling strategies ability to fit the model development data. Where the blue bars display the analysis of modelling strategy 1's models and the red bars display the analysis of modelling strategy 2's models.

Plotting the residuals in Figure 6.11 allows for a better understanding of where a model overestimates or underestimates the actual value, as well as highlighting the fundamental flaws within a model. In the region of approximately 16-30 Fe% Removed, where the density of data is greatest, both modelling strategies had a similar distribution of residuals. The exception is the GP model, which shows signs of overfitting on the training and validation data. The overfitting to the training and validation indicated by the zero residual on all but a few data points in Figure 6.11(A). Comparing the modelling strategies on regions that are not well represented by the data, it a similarly unclear picture as the statistical results. The GP model built using data from all three reactors is superior to the one built using data from only the up-flow reactor; although, this may be because the model has overfitted the data. Overfitting decreases a model's ability to make prediction on new data. The RF model built from just the up-flow reactor data is clearly superior to alternative RF model. While the residual plots for the SVR and ANN models appear to show small statistical difference between the two modelling strategies.

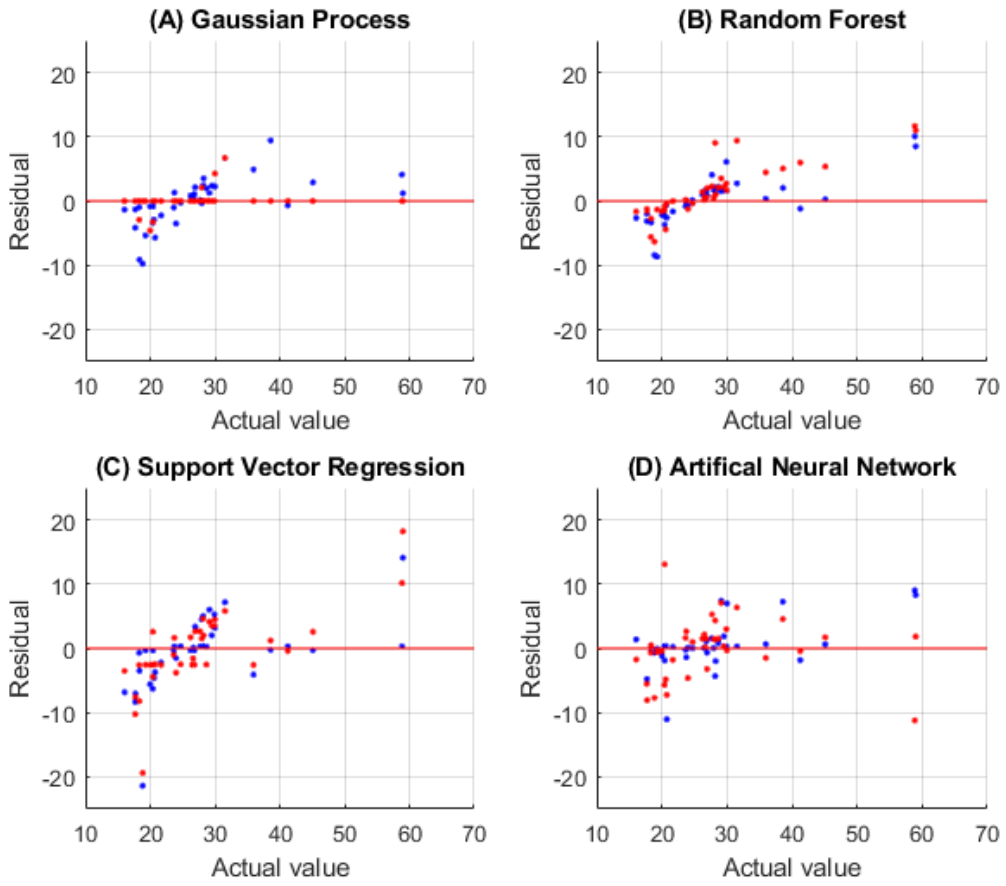


Figure 6.11 Residuals plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network. Where the blue dots are the residuals from modelling strategy 1's models and the red dots are the residuals from modelling strategy 2's models.

The GP model built using modelling strategy 2 compared favourably to other DDMs built studying the treatment of mine wastewater. The ANN model developed by Sahinkaya to model the biotreatment of zinc-containing wastewater had R^2 scores in between 0.827 and 0.998 for the model's outputs [384]. While the ANN model developed by Atasoy *et al.* to model performance of up-flow and down-flow fluidized bed reactors treating AMD, has R^2 scores between 0.92 and 0.98 [385]. Neither of these studies has nor, to the best knowledge of the authors, other AMD treatment DDMs, has evaluated the DDMs' prediction capabilities on unseen data. Without evaluating using unseen data, it is not possible to assess the prediction capabilities of a DDM beyond the data they were trained on data.

6.5.3.1 Comparison of models using unseen data

A common error when developing a DDM is to put too much emphasis on optimising the model's capability to fit the model development data [118]. This does not evaluate the model's ability to extrapolate beyond the model development data. The best DDMs are able to extrapolate beyond the data they were developed from, to make accurate

predictions on unseen data. Before commencing work on developing the MT models, 10 data points were partitioned from the data to be used to thoroughly evaluate the models' predictive capabilities. Six points were taken from within the model development boundaries and four points outside the model boundaries. The statistical metrics results for each model, inside and outside the system, are shown in Figures 6.12 and 6.13.

Comparing the models' performance between the model development data and the unseen data, the decrease in performance observed in chapter 6 is repeated by the Coal Authority manufacturing data. This is reflected by a decrease in R^2 scores and an increase in RMSE and MAE scores between the model development data and the unseen data. When evaluating the models' abilities to make predictions on unseen data within the system, it is immediately noticeable that the GP model built following modelling strategy 1 had the best prediction accuracy. When statistically evaluating the model's predictive accuracy the model had the best scores (R^2 0.74, RMSE, 7.42, MAE 5.00). The GP model built following modelling strategy 2 had performed best on the model development data; however, it had an average R^2 and RMSE scores, of 0.49 and 10.33 respectively, second best MAE score of 7.30. The SVR models poor performance on the model development data was likewise mirrored by a poor performance on unseen data inside the system. In addition, the ANN model following modelling strategy 1 shows a poor ability to extrapolate beyond the data it was trained, having close to the worst R^2 and RMSE scores, of 0.19 and 12.99 respectively and the worst MAE score of 10.24. Normally ANN is known for their ability to model complex non-linear systems [118]. This result highlights the truth expressed by the No-Lunch Theorem that no one machine learning algorithm is best for all scenarios [142].

An interesting result, concerning the R^2 scores, was observed when it came to evaluating the models on unseen data outside of the system. The RMSE and MAE scores for all the developed models were higher when making predictions on unseen data outside of the system compared to predictions made on unseen data inside the system, reflecting a poorer prediction accuracy. However, the R^2 scores increased, implying an increase in prediction accuracy. When studying the residual plots in Figure 6.14, it can be seen that the models were generally less accurate at predicting unseen data outside of the system. This apparent error in the R^2 scores maybe because the R^2 only agrees if the predicted values come from an ordinary regression (or similar method) on the observed values [390]. This may not be the case when applying a model to some non-linear systems and may lead to misleading R^2 scores [390]. Therefore, the

RMSE and MAE scores have been given more attention when deciding on the best model.

The model's ability to extrapolate beyond boundaries of the model development data was investigated by evaluating the models on unseen data outside of the system. As stated the prediction capabilities decreased when making predictions outside of the system, reflected by an increase in RSME and MAE scores. The exception to this is the SVR model built on just the reactor 3 dataset, who's RMSE scores actually decreased. Overall, both the SVR and GP models built on just reactor 3 dataset performed best on the unseen data outside the system, as the SVR model had the lowest RSME score while the GP models had the lowest MAE score.

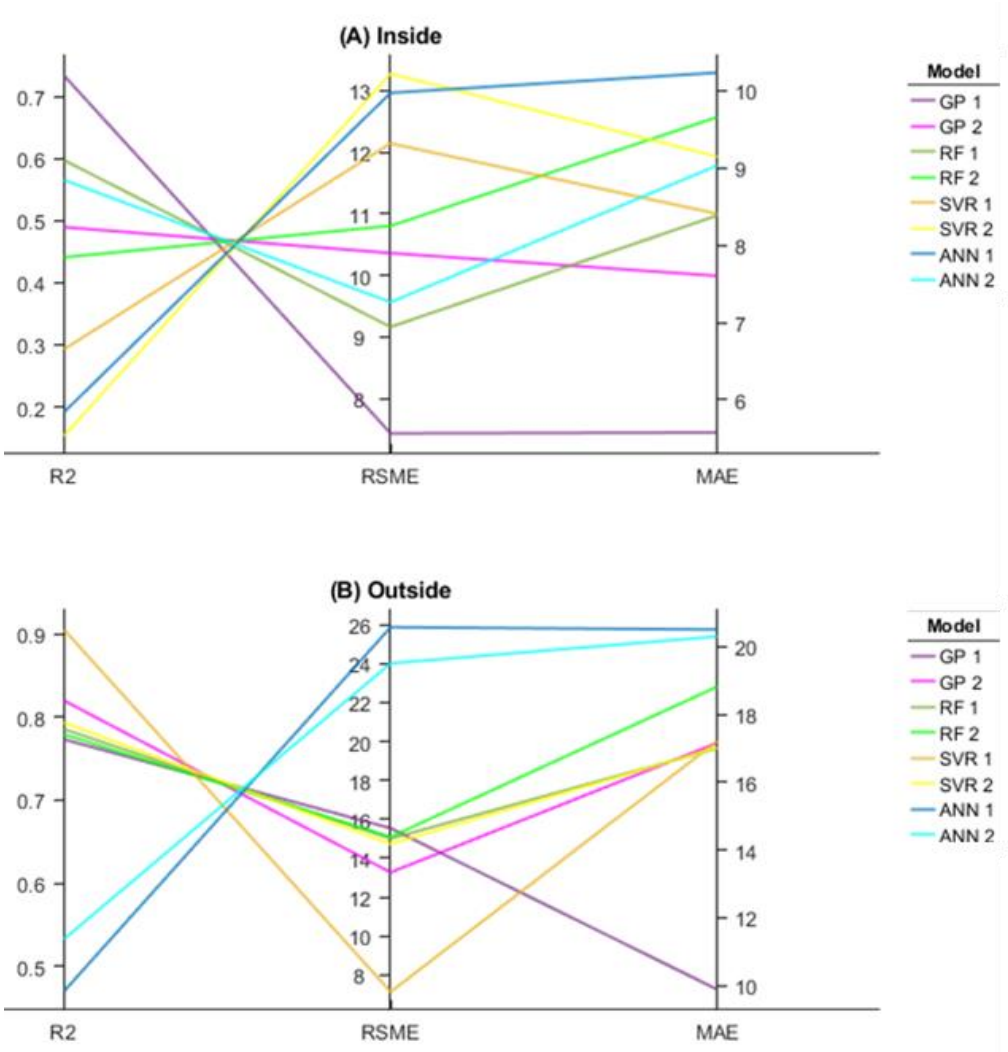


Figure 6.12 Statistical analysis of the models' ability to fit the unseen data (a) inside the system and (B) outside the system. The number after the model algorithms states the modelling strategy.

When comparing the two different modelling strategies (using only reactor 3 dataset or using reactor 1, 2 and 3 datasets) on the model development data, the models built using

only reactor 3's dataset had a slight advantage. This was to test whether the model's predictive capability would increase by expanding the data available to learn from. This was reasoned by the fact the model would have more examples of the inherent variability within the mining wastewater. Furthermore, previous works have shown that the easiest and quickest method to improve a model's accuracy is to increase the volume of data available to learn from [142]. When comparing the modelling strategies on unseen data, it was found that this hypothesis proved incorrect for this particular manufacturing system. There were examples, such as the ANN model, that benefitted from being built using all three datasets (R^2 0.19 → 0.56, RMSE 12.99 → 9.54, MAE 10.24 → 8.90). Overall, however, the models built using only the reactor 3 dataset were shown to be more accurate (Figure 6.13 and Figure 6.14).

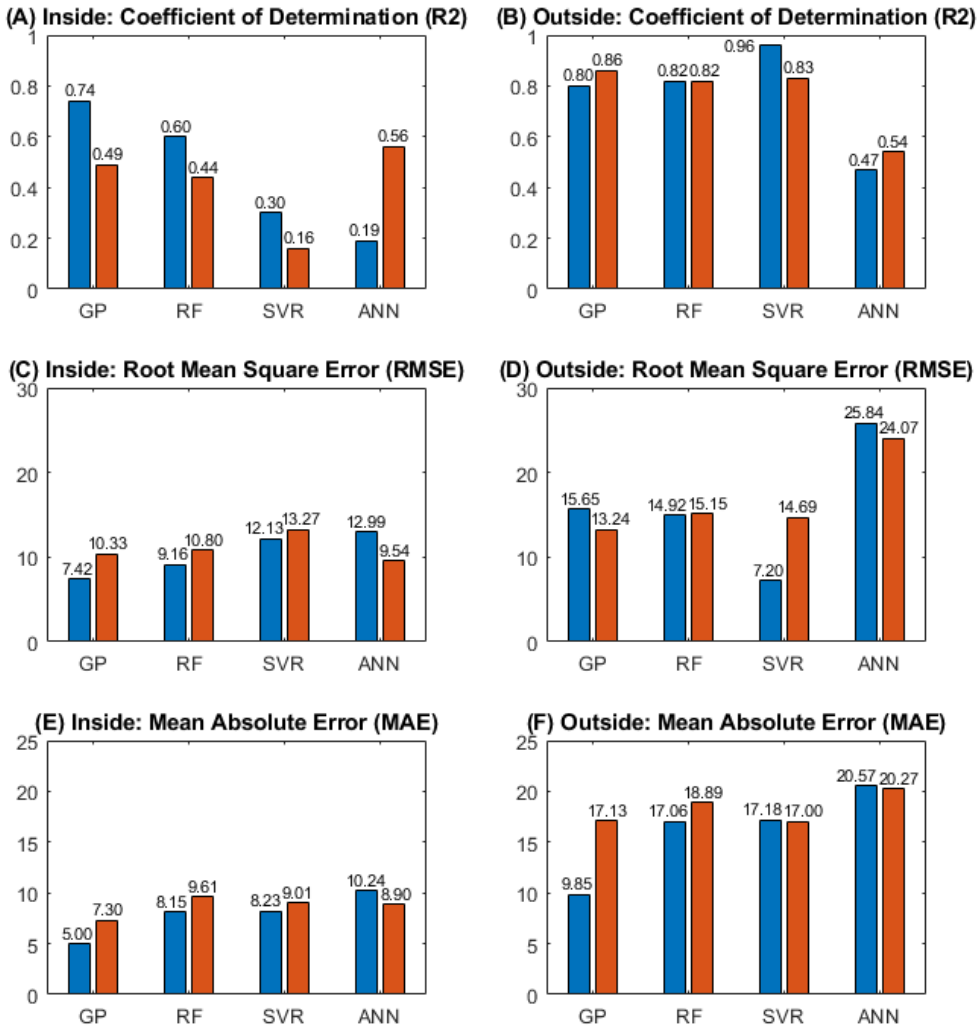


Figure 6.13 Comparing the statistical analysis results of two different modelling strategies ability to fit the unseen data. Where the blue bars display the analysis of modelling strategy 1's models and the red bars display the analysis of modelling strategy 2's models.

The GP model built from just reactor 3 dataset was deemed to be the best overall model for predicting the percentage of Fe removal by the SCOOFI up-flow reactor. The GP was chosen because of its superior prediction capability on unseen data, indicating that the model is the best at extrapolating beyond that data it was trained on. Previous work Samuelsson *et al.* has demonstrated the GP based models are powerful tools when monitoring wastewater treatment process [391]. This finding is reflected in the outcomes of this work when modelling mining wastewater treatment.

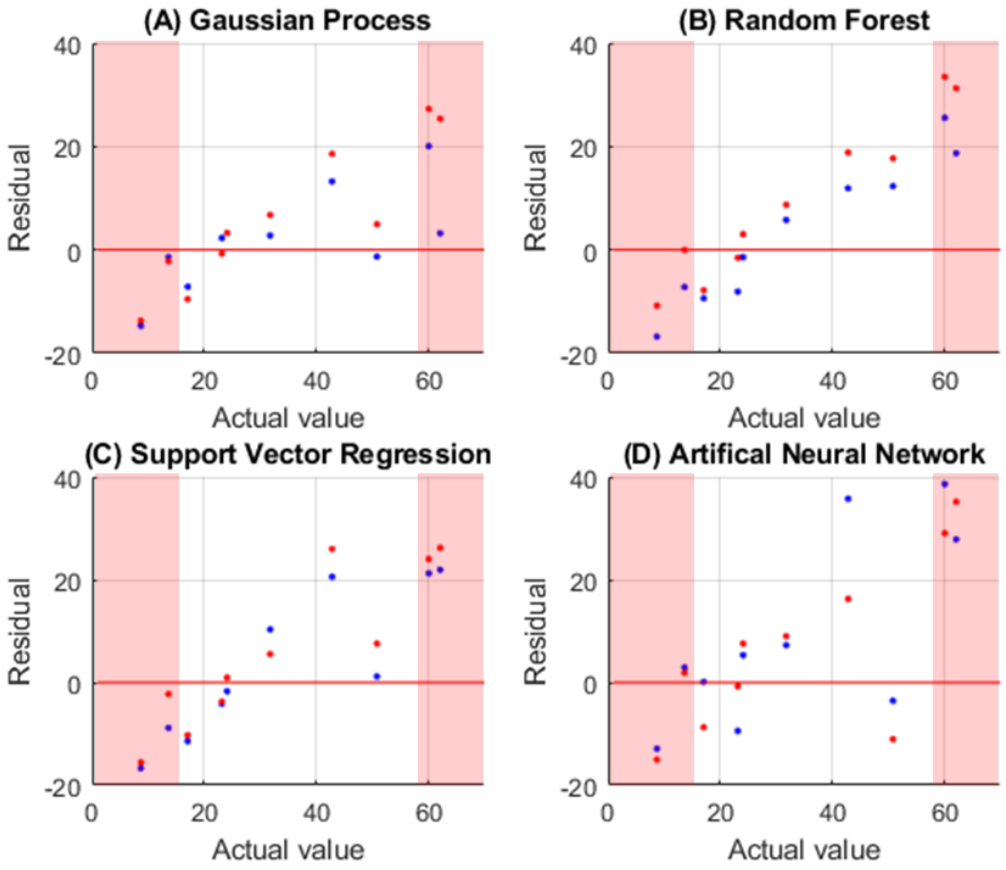


Figure 6.14 Residuals of unseen predictions plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network. The red region is outside of the system.

6.5.3.1 The GP model's ability to describe the manufacturing system

In section 6.4.2.3, the system was observed to have four phases of operation:

1. Start until 15/07/18: Fe removal rates vary between 20-65% suddenly, indicating an unstable period during start-up.
2. 15/07/18 – 01/11/18: variability is still present within the system but overall the system operates relatively steadily between Fe removal rates between 10-30%.
3. 01/11/18 – 10/12/18: the system is desludged increasing the Fe removal rates to between 35-60%. The system is more stable at the higher Fe removal rates than during phase one, but there is one occasion where the removal rate dramatically drops.
4. 10-12/18 – end: The system operates steadily between 15-30% Fe removed. Although after the 01/01/2019 there is a gradual decrease in %Fe removed, possibly indicating that further desludging is required.

The first stage was removed from the model during cleaning; the model's ability to describe the other three phrases is shown in Figure 6.15. The model is able to accurately represent the majority of the second phase (between 15/07/2018 – 01/11/2018). The

model was able to make reasonable predictions model development data and even two of the unseen data point in this phase. The model did, however, display signs of overestimating the Fe% removed between 10/01/2018 and end of the phase and was unable to predict the unseen datapoint of 8% Fe removed. There was a concern that the model prediction capability would decrease for the third phase, because of the lack of data points above 35% Fe removed that were available to learn from. This did not seem to be entirely the case though, the model was able to accurately fit the model development data and make good predictions for two of the unseen data points in this region. The two biggest model errors did occur during this region. However, they were mitigated by the fact the model was underestimating the SCOOFI reactors ability to removal Fe. Of the two possible errors, overestimating or underestimating, the second is the favourable, as it is not a problem for the Coal Authority if the system performs better than expected. However, if the model was to overestimate by a large amount, mining wastewater above the regulatory limits may be emitted into the watercourse, which would have negative environmental impacts. The GP model clearly performs best when making predictions for the final phase of the system, as it was able to accurately match the model development data and the unseen data. Overall, the GP model was sufficiently accuracy when describing the Fe removal rates from mining wastewater by the up-flow SCOOFI reactor. The developed GP model can now be utilised to discover how the mining wastewater characteristics affect the SCOOFI reactor.

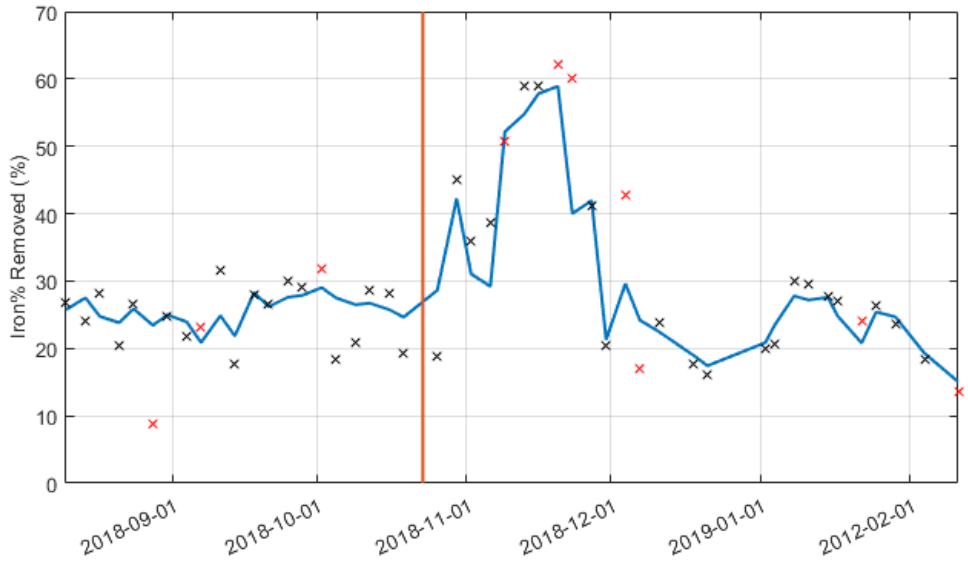


Figure 6.15 The GP model predictions plotted against the model development data (black crosses) and the unseen data (red crosses).

6.5.5 DETERMINING THE RELATIONSHIP BETWEEN SYSTEM PARAMETERS AND PERFORMANCE

The GP model was used to produce three-dimensional response surface plots to analyse how the system parameters affect the Fe removal rate by SCOOFI up-flow reactor. Figures 6.16 -6.18 shows the effect of the seven investigated parameters on the removal of Fe, with five of the seven parameters held constant at its intermediate value for each plot. The intermediate value for each parameter is communicated in Table 6.4.

Table 6.4 Intermediate value of each parameter.

Parameter	Value
Iron concentration in wastewater (Fe conc.)	11.76 mg/l
Total suspended solids concentration in wastewater (TSS conc.)	33.50 mg/l
pH	7.90
Dissolved oxygen (DO)	0.71 (mg/l)
Conductivity	2390.00 (uS/mc)
Flowrate	0.70 (l/s)
Time since last desludging (Desludging)	75.50 (days)

Figures 6.16(A-F) shows the response of the Fe% removed when varying the initial Fe concentration against each of the other parameters. Generally, the GP-based analysis indicated a higher Fe concentration in the wastewater increases the Fe removal rate. Of note was Figure 6.16 (D), which displayed the effect of Fe concentration and

conductivity on the Fe removal rate. Unlike Figures 6.16 (A-C, E), which had a maximum at an edge or corner, Figure 6.16 (D) maximum was a less defined plateau. The surface plot displayed high Fe removal across all wastewater Fe concentrations and conductivity values apart from the lowest conductivity values and when Fe concentration and conductivity were at their lowest and highest respectively. Figure 6.16 (F) indicates that the time since desludging has a much large effect on the Fe% Removed than that of the concentration of Fe in the wastewater.

Figures 6.16 (A, G, H) and 6.17 (I-K) mostly implied that higher concentrations of TSS in the wastewater improve the Fe% Removed. A similar relationship is observed in Figure 11 (D) is repeated in Figure 6.17 (J) between the wastewater TSS concentration and flowrate. A maximum point is determined at higher TSS concentrations; however, high Fe removal rates are achieved at lower TSS concentrations and low flowrates. Figures 6.16 (B, G) and 6.17 (L-O) show that pH effect on the Fe% Removed was less clear and was dependant on the other parameters being varied. When pH was low, high Fe removal rates were achieved at high Fe and TSS concentrations and when pH was low, high Fe removal rates were achieved at low flowrates and minimum time since desludging. Whereas, when only pH and DO were varied, the Fe removal was driven almost entirely by the DO concentration.

Repeating this analysis (Figures 6.16-6.18) for the other variables drew the following conclusions from the GP-bases data analysis:

- For the majority of wastewater compositions, increasing the Fe and TSS concentrations increases the Fe removal rates.
- Desludging should occur approximately every 55 days to achieve peak Fe removal and has one of the largest impacts on the removal of Fe.
- Parameters whose effect are overshadowed are:
    - Fe concentration by time since desludging.
    - pH by DO concentration.
    - Do by time since desludging.
- pH, DO, conductivity and flowrate effect on the Fe removal rates is dependent on the other parameters values.

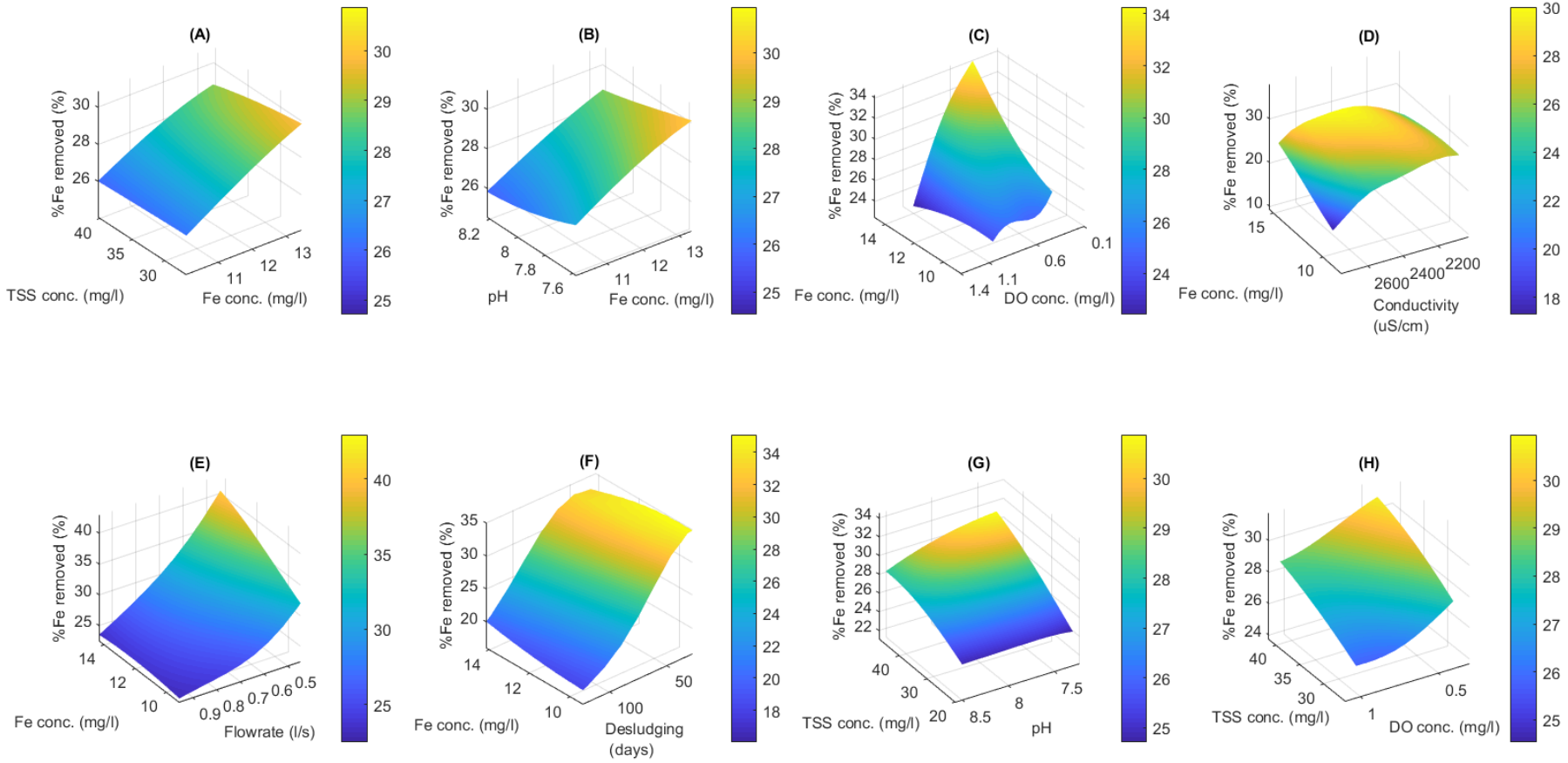


Figure 6.16 Surface plots showing the effects of mining wastewater characterises on the %Fe Removed as predicted by the GP model.

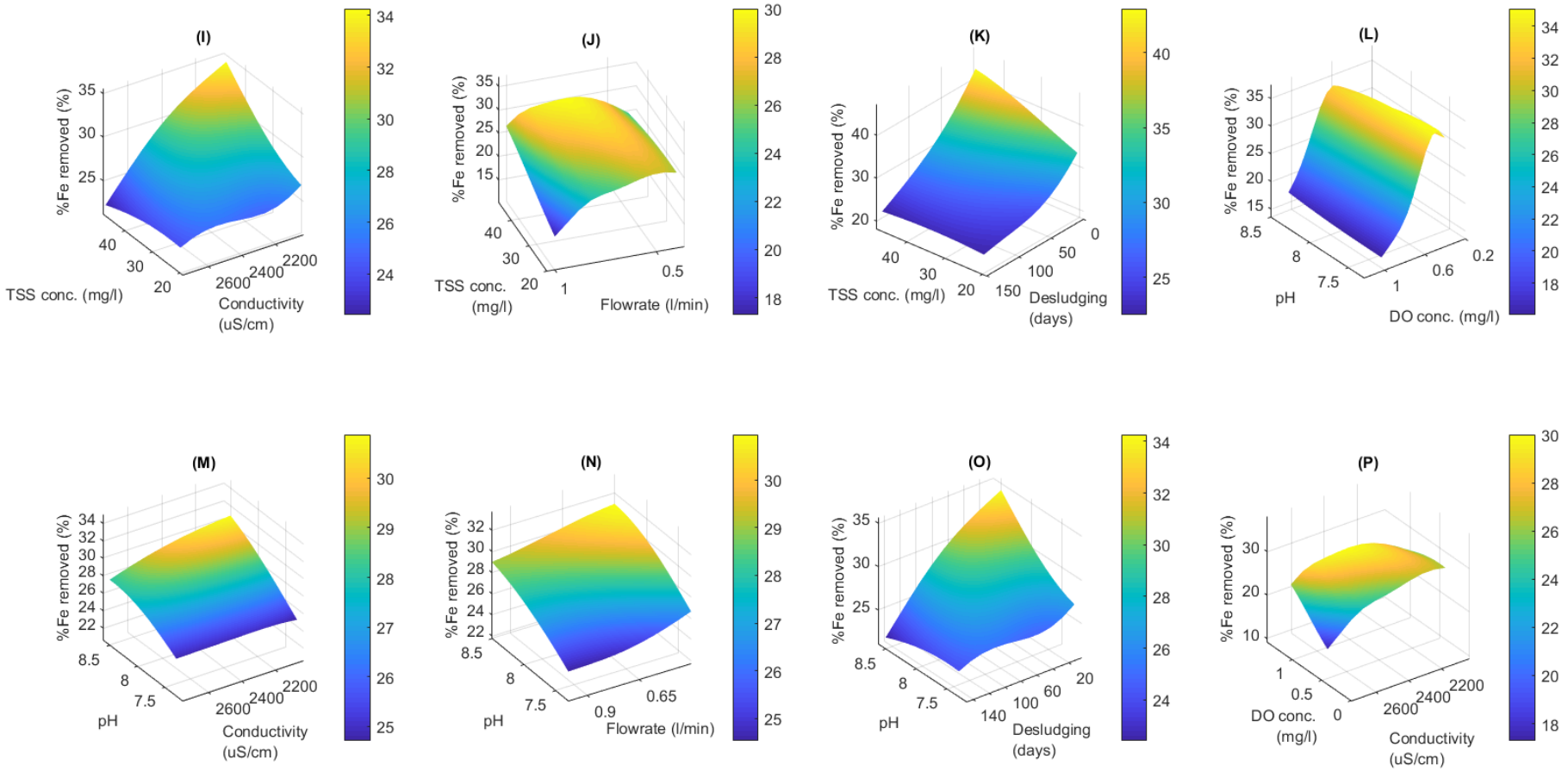


Figure 6.17 Surface plots showing the effects of mining wastewater characterises on the %Fe Removed as predicted by the GP model.

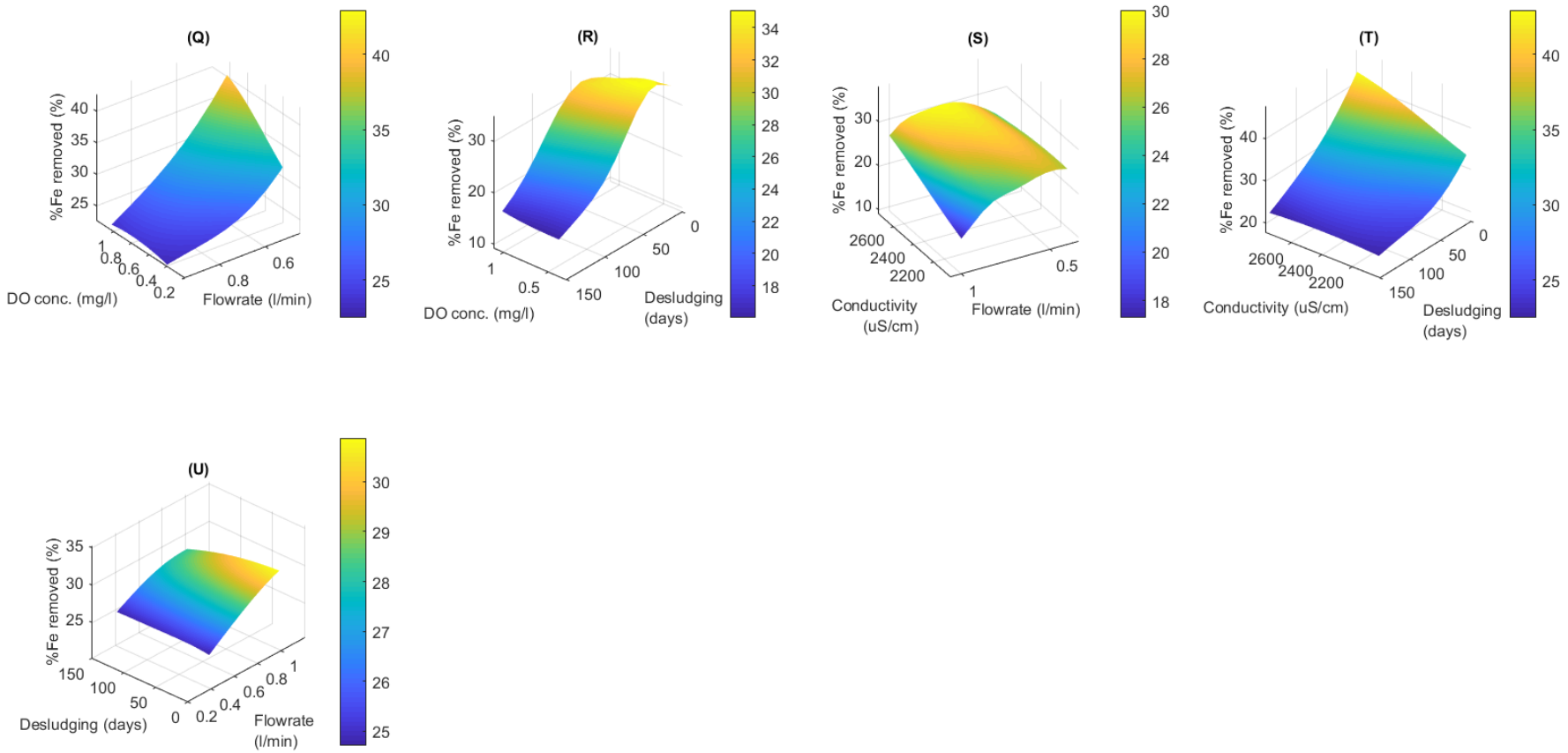


Figure 6.18 Surface plots showing the effects of mining wastewater characterises on the %Fe Removed as predicted by the GP model.

6.6 CONCLUSIONS

- There are different active-passive and chemical-biological approaches to treating mining wastewater. The Coal Authority has traditionally used a passive biological approach for the remediation of Fe from mining wastewater. However, they are interested in exploring new methods with a lower capital cost and one technology being trialled are SCOOFI reactors.
- A SCOOFI reactor is an example of a packed bed reactor to remove Fe from wastewaters. Three different reactor configurations were trialled: no packing material, down-flow with packing and up-flow with packing. The up-flow with packing reactor was found to be superior at removal of Fe from wastewater.
- However, there was a large amount of variability in Fe removal rates. A DDM approached was used to investigate if the variability is significantly influenced by key mine wastewater characteristics.
- Two modelling strategies were trialled to investigate the effect of expanding the data available to DDM to learn from. It was found that the DDMs' generally performed best when learning from data directly describing the system being modelled. The expectation to this was the ANN model, whose performance increased when trained on all three reactors datasets.
- Four different machine learning algorithms were built and compared using statistical evaluations. To determine which model best described the process manufacturing system, the models were evaluated based on their accuracy when making predictions on unseen data.
- The GP model was determined to best describe the removal of Fe from wastewater by an up-flow SCOOFI reactor (R^2 =0.74, RMSE = 7.42, MAE = 5.00).
- The GP model was then used to produce surface plots to investigate each of the system parameters effects on Fe% removal. From the GP-based analysis the highest Fe removal occurred at high concentrations of Fe and TSS in the wastewater and up to 55 days since desludging the reactor. The other parameters had conflicting impacts on Fe% Removed, dependent on the other parameters. This illustrates the complex, non-linearity of the system and why a DDM analysis is required.
- Further work may look at optimising desludging strategy to enhance Fe% Removed by the system, akin to the work performed by González *et al.* when enhancement of peak flux capacity in membrane bioreactors by controlling the backwashing strategy [388]. If the dataset was expanded to include more

examples of the effects of desludging it may be possible to utilise the model for predicting the optimal desludging frequency.

CHAPTER 7: DATA-DRIVEN MODEL RELEVANCE TO MANUFACTURERS AND PLACE WITHIN A CLOUD MANUFACTURING SYSTEM

7.1 HIGHLIGHTS

- This chapter focuses on the need and relevance of data-driven models (DDMs) to process manufacturers.
- How a DDM may be used by a process manufacturer to understand and enhance their systems is explained.
- An example of how a DDM may provide real-time prediction within a graphical user interface (GUI) is outlined.
- Two different business models for commissioning and developing a DDM is proposed.
- Finally, how a DDM would work within a cloud manufacturing (CM) system built for process manufacturing environments is outlined.

7.2 INTRODUCTION

In chapters 4 and 5, a methodology was developed for the modelling process manufacturing system using DDM techniques. These chapters focussed on the unique considerations and challenges faced by process manufacturers and developed methods on how these may be overcome. However, a process manufacturer is unlikely to be interested in the technical details of how the model is made but rather what economic and/or environmental benefits the manufacture will get out of the model. This chapter addresses this issue as well as three other practical implications when developing a DDM for a process manufacturer (contracting out model development, dangers of overestimating and underestimating for different scenarios, commercialisation of manufacturing data)

The title of this work is ‘Intelligent Resource Use in Process Manufacturing”. Cloud manufacturing is a new manufacturing paradigm that would enable intelligent resource use, through the sharing of manufacturing resources and capabilities on a manufacturing cloud platform. Chapter 3, for the first time, explored how cloud manufacturing could be applied to process manufacturing to increases sustainability. A key characteristic of cloud manufacturing is that the manufacturing cloud platform is a knowledge-intensive, decision-making tool. If this characteristic is to be realised, it will require the

development of mathematical models to describe process manufacturing systems. The second focus of this chapter is utilising the models developed in the previous chapters to illustrate intelligent decision-making, process manufacturing CM system.

7.3 DATA-DRIVEN MODELS RELEVANCE TO THE MANUFACTURER

A process manufacturer is not going to be interested in which machine learning algorithm the DDM uses or the data preparation, pre-processing and partitioning techniques employed in developing the DDM. Instead, what is more likely to be of interest to the manufacturer after the model is built is:

- What does the model tell the manufacturer about their system?
- How can the model be used to enhance the manufacturing process?
- Can the manufacturer trust the model's predictions?
- Will the model require retraining and, if so, how often?
- How does the manufacturing pay for the development of the model and future servicing costs?

7.3.1 UTILISING A DATA-DRIVEN MODEL TO GAIN A DEEPER UNDERSTANDING OF A PROCESS MANUFACTURING SYSTEM

The work in chapters 5 and 6 showed how a DDM could be utilised to analyse the effect the system parameters have on key process objectives. There may be a singular process objective, such as the Fe removal rate in chapter 6, or multiple objectives, as demonstrated in chapter 5 in understanding the influence the process conditions had on the pollutant removal rates and bioenergy generation using the H^2AD bioprocess and agricultural wastewaters as a feedstock. This is achieved by generating data where the parameters of interest are varied incrementally over a range of values, while the remaining parameters are held constant at an intermediate value. The data is then inputted into the developed DDM, which predicts the effect of the parameters on the outputs of the system. It may be that the manufacturer wishes to understand the parameter(s) effect beyond the boundaries of the data that the model was originally developed for. For example, the DDM model built in section 6.1, predicting the COD removal by the H^2AD bioprocess was built using HRT data between 2-6 days. If a manufacturer wanted to predict the effect of increasing the HRT to 7+ days, it must be remembered that DDMs have a limited ability to extrapolate beyond the boundaries of the data on which they were trained [98].

The clear advantage of using DDM-analysis to a process manufacturer is not needing to perform pilot studies or interrupting the manufacturing process to investigate their

effect, which is often costly to the process manufacturer [26]. Traditionally manufacturers built response surface models, which is an approach that has seen wide use by process manufacturers, due to its ability to optimise a process with a minimal amount of data [111]. Response surface modelling is a ordinary differential equation model that can model the impact of various variables, both individually and through their cumulative interactions, on a system response [392]. However, studies that have compared a response surface model (RSM) and to a DDM show that the DDM consistently outperforms the RSM in its ability to fit the data and make predictions [111,119,393,394].

As previously discussed, process manufacturing systems and waste streams are subject to a higher level of variability, particularly in the process and raw material streams [395]. This can negatively affect key process manufacturing variables; for example, product yield and quality [395]. Traditionally the impacts from variability have been mitigated by either over-engineering a process manufacturing plants design window, or by process control [90]. The difficulty in trying to mitigate the effects of variability is that it is not always possible to know the cause of the variability or what impact it is having on the process manufacturing system [26]. If the variability effect on the system is unknown, it is not possible to model or simulate its effect on the system using traditional first-principal models. The advantage of DDM models is that they are a *black-box* approach, meaning that a model can be built with no knowledge of its internal workings [396]. This approach is capable of capturing the variability present in the system and its effect on the system's outputs, with no guidance from the modeller. A constructed DDM may then be used for two applications:

1. Predicting the impact variability has on the system;
2. Predict/forecast the occurrence of future variability and influence of future process performance.

To understand the impact the variability has on the system, the first application uses the surface plot modelling techniques previously described. The advantage of the second method is that future variability in the composition and supply of raw materials may be predicted before it has occurred. This will increases the resilience of process manufacturing system as measurers can be put in place ahead of time to minimise the variability effect. There are examples of forecasting being used in process manufacturing [397]; however, there is limited work focussed on predicting the composition and supply of raw materials.

7.3.2 REAL-TIME MONITORING AND OPTIMISATION OF PROCESS MANUFACTURING SYSTEMS

If a developed DDM is to be used for ongoing monitoring and optimisation of process streams, the development of a graphical user interface (GUI) will be required, allowing the process manufacturer to interact with the model to make predictions in real-time. Recently, there has been a big rise in the number of companies offering to design GUI systems to visualise and display a manufacturer's data in real-time. Examples of these companies include Accenture, Cognizant, Intelligent Plant and Splunk [398–401]. While some of these companies, such as Cognizant, have started to build DDMs into their GUI systems the full potential of DDM to self-optimise a process manufacturing system has yet to be realised. What these companies tend to offer instead under their "Process Optimisation" feature is to integrate processes and data to enable visualisation across the system for informed decision making [398].

An outline of a GUI to visualise a continuous stirred tank reactor (CSTR) system, with built-in predictive modelling capabilities is displayed in Figure 7.1. This GUI performs the standard GUI tasks of visualising real-time data. Further levels could be added giving detailed analysis of the manufacturing data and the implications for the manufacturing key performance indicators. An additional feature, is that the purposed GUI can offer the process manufacturer predictions of what the effect would be of changing a process condition on the product quality. For example, the process manufacturer trying to lower operational costs wants to know the effect of lowering the reactor temperature by 7°C. The manufacturer would have to enter a new temperature setpoint into the GUI, and they would get an instant response informing the manufacturer that this is predicted to cause an 8.85% decrease in product composition. Further capabilities could be added comparing the economic saving of operating at a lower temperature to the economic impact of producing a lower quality product. This GUI could be downloaded onto process manufacturer's phones, tablets and laptops to offer instant predictions to the user.

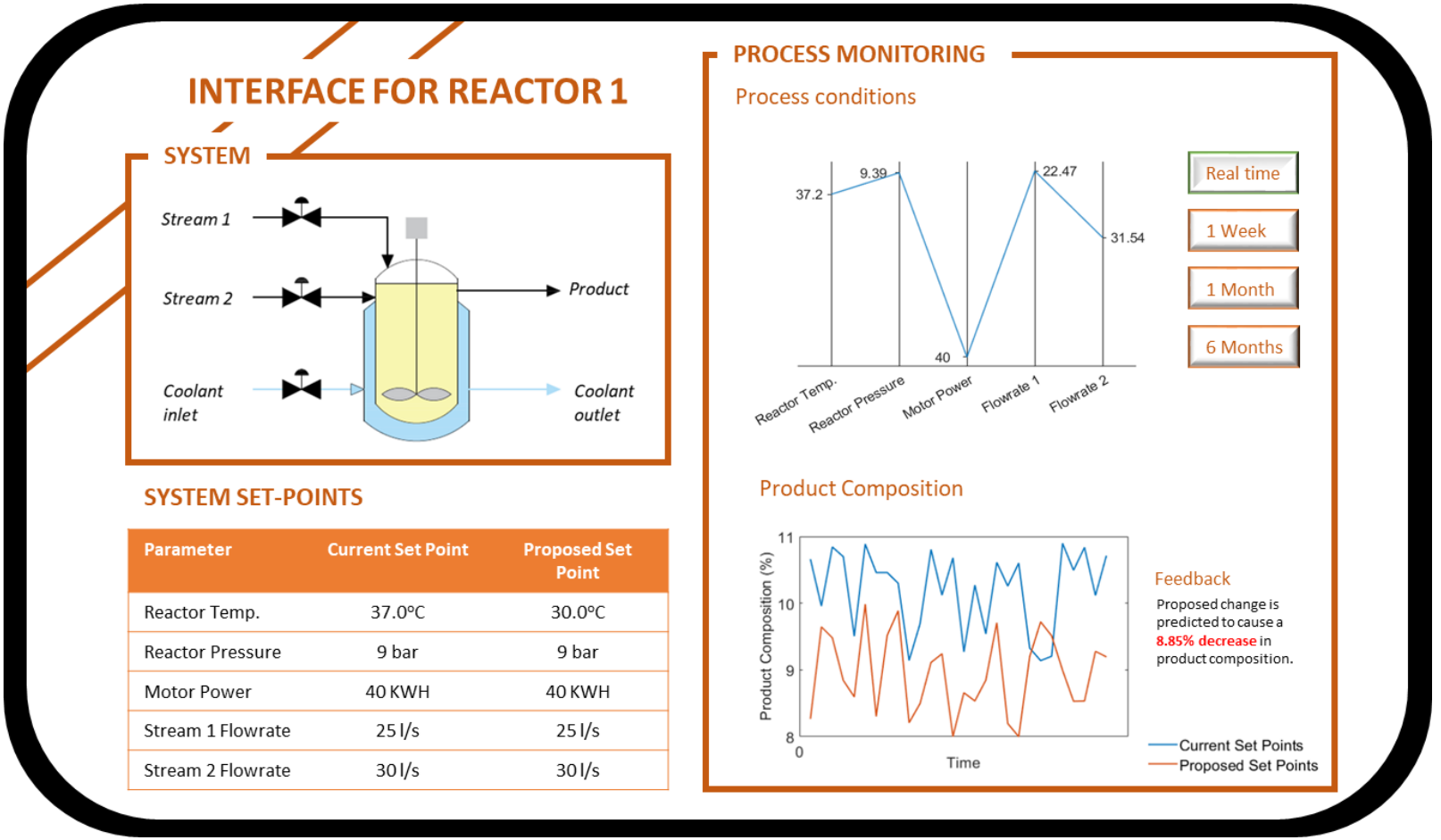


Figure 7.1 Example of a GUI built for a CSTR with built-in predictive modelling capabilities.

7.3.3 THE RELATIONSHIP BETWEEN MODEL PREDICTION ERROR AND PROCESS MANUFACTURING SYSTEMS

Data-driven models built for regression problems may systematically overestimate or underestimate the actual value if the relationship between model inputs and outputs is non-linear [402]. Process manufacturing systems are often complex and non-linear [402] and are susceptible to this issue. This has implications when modelling process manufacturing systems, as there are ranges where prediction error may have economic, environmental and safety impacts on the manufacturing process. In order to understand the dangers of overestimating and underestimating for different parts of the manufacturing system, four classifications have been proposed. These are visualised in Figure 7.2. The classifications are:

1. Underestimating for low values region;
2. Underestimating for high values region;
3. Overestimating for low values region;
4. Overestimating for high values region.

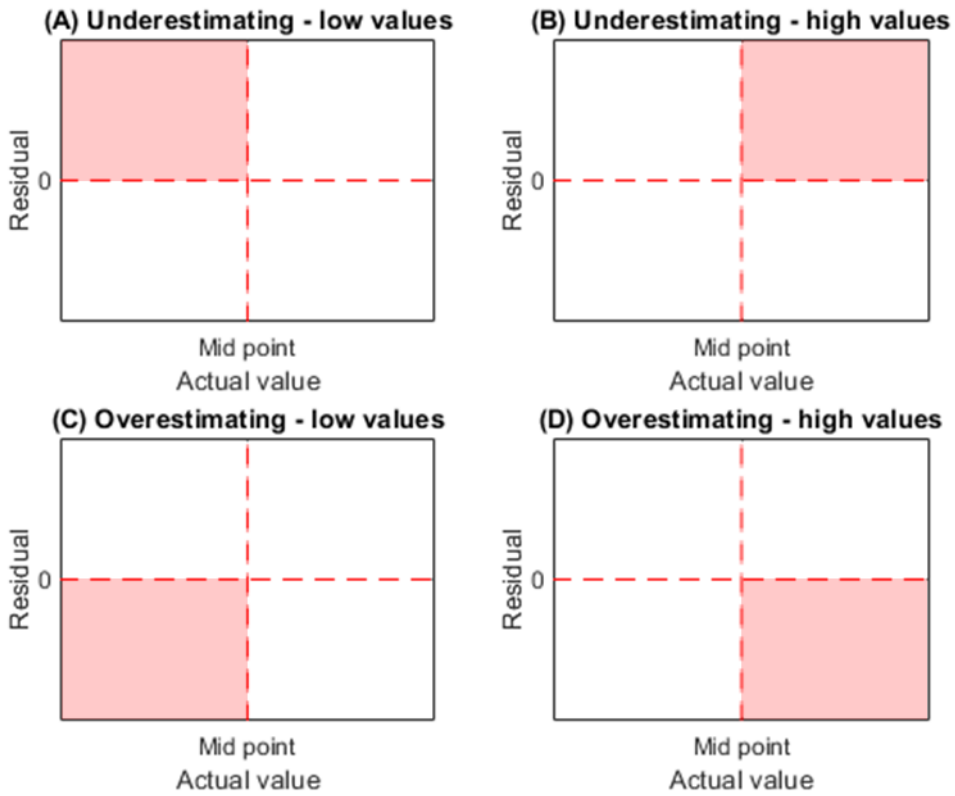


Figure 7.2. The four classifications of prediction error in process manufacturing systems, where the red shading represents the region model.

Underestimating occurs when the model's predicted value is lower than the actual value and vice-versa for overestimating. The regions are the possible ranges for the process

manufacturing system output. For example; the H^2AD bioprocess in chapter 5 had a range of percentages of COD removal rates between 45.0% and 74.9%. By this classification system the low-value region would be between 45.0% and 60.0%, and the high value between 60.0% and 74.9%. For this process manufacturing system the prediction error region that would cause most problems would be when overestimating %COD removed for low values. This is because it may result in the process manufacturer believing the system is achieving the minimum removal rates to meet economic and environmental objectives. Further examples of the implications of underestimating and overestimating in process manufacturing systems are given in Table 7.1.

Table 7.1 Examples of problems faced by overestimating or underestimating a process manufacturing output in low and high regions.

	Low-value region	**High-value region**
Under-estimating	The International Energy Agency have set industry the target of reducing CO_2 emissions to help tackle climate change [403]. When predicting the CO_2 emission by an industrial process, underestimating the CO_2 emitted will lead to an increase in environmental damage and potential fines.	Hydrogen gas forms combustible or explosive mixtures with atmospheric oxygen over a wide range of concentrations [404]. Underestimating the volume of hydrogen generated during production will have safety implications — especially when operating near the top boundaries of the system.
Over-estimating	When treating process manufacturing wastewaters, pollutants have to be removed to below a certain limit to avoid environmental damage and potential fines [323]. Overestimating pollutant removal rates, when the actual removal rates are low, may result in pollutants not being removed to below the limits.	The pharmaceutical industry demands high purity in their products [405]. Overestimating the purity of products will lead to mislabelling of the products purity and possibly products which are not as pure as required by the consumer.

7.3.4 SERVICE-ORIENTED BUSINESS MODEL

Process manufacturers, especially SMEs, are unlikely to have access to the required expertise to build a DDM that accurately represents their system. There exists software

packages that allow users to develop a DDM quickly [278,406,407]. However, without an appreciation of the considerations and challenges when modelling manufacturing systems, or the modelling techniques to overcome these challenges, the resulting DDM is unlikely to be representative of the process manufacturing system. Therefore, this work proposes there will be two approaches when commissioning, developing, building and servicing a DDM:

1. Product-oriented architecture (POA)
2. Service-oriented architecture (SOA)

These two business models are based on how software companies commercialise their products. In the POA business model, the selling of the product and any required future servicing of a product are separate business transitions [219]. The consumers own the product and are responsible for arranging future servicing if necessary. Applying this model to developing a DDM for a process manufacturing system would mean a process manufacturer commissions a modeller to develop a DDM to be built for their system. Once built the DDM is owned by the manufacturer. The manufacturer is responsible for the retraining of the model, as the model predictive capabilities may deteriorate with time. When retraining is required, the manufacturer pays a one-off fee for this service. This business model would be suitable when developing a DDM model for a process manufacturing system that

1. Has been stable for an extended period (more than 2-3 years);
2. Is not predicted to change in the future and
3. Has accessed to a large volume of historical data (more than 2-3 years) to develop the model with.

The reason for this is that the deteriorating rate of the model's predictive capabilities is likely to be slow if the system remains largely unchanged; the model will therefore not need regular retraining. Furthermore, having a large volume of historical manufacturing data available to develop the DDM will increase the likelihood that the DDM is representative of the manufacturing system and capturers the system's variability [142]. Although, this is not the case if there is changing feedstocks/ raw materials, process conditions and requirement of outputs. The DDM may need regular servicing at the start of its conception to ensure that its predictions are accurate. The service rate should decrease as trust in the model's predictions is established.

For process manufacturers that will be subjected to frequent changes to their systems (e.g. changes to feedstock supply and composition, site location, updating legacy

equipment, introducing new technologies), a SOA business model may be more appropriate. The SOA business model gained lots of attention in the early 2000s as the software industry moved from selling their software as a product to selling it as a service [408]. More recently, manufacturers are shifting from selling single-use products to selling their products as services [409], and SOA a key characteristic of CM systems [3]. Sprott and Wilkes definition of SOA is [408]:

> "The policies, practices, frameworks that enable application functionality to be provided and consumed as sets of services published at a granularity relevant to the service consumer. Services can be invoked, published and discovered, and are abstracted away from the implementation using a single, standards-based form of interface."

Applying SOA to a DDM business model would mean, instead of commissioning a model to be built for a process manufacturing system, a process manufacturer would hire the modelling services of a modelling company. The manufacturer would not pay a single fee for a developed DDM, but pay a subscription fee to use a DDM developed of their system. It would then be the responsibility of the modelling company to service the DDM model, to ensure it remains representative of the process manufacturing system. Different subscription pay scales would be necessary dependant on the nature of the system being modelled. A system that will be subjected to more changes will require the more to be serviced (retrained) more frequently than one with less frequent changes to the system. Therefore, there would be different pay scales dependant on the frequency of servicing. As a system becomes more stable, and the model requires less frequent servicing, the process manufacturer may change to a lower pay scale.

There is the possibility that complications may arise due to defining which stakeholders own what part of the intellectual property. The final DDM would be jointly owned, the process manufacturer would own the manufacturing data used to build the model, and the modelling company would own the architecture of the developed model. The modelling company could not take the developed DDM and use it to make predictions for another, similar manufacturing process. Likewise, the process manufacturer could not use the developed DDM without paying a subscription fee to the modelling company. In the event of terminating the contract, there is the option that the model could be sold to the process manufacturer for continued use. The process manufacturer would now be responsible for the servicing of the model as described in the POA business model.

7.4 DATA-DRIVEN MODELLING ENABLING INTELLIGENT CLOUD MANUFACTURING

In chapter 3, new concepts were developed of how cloud manufacturing (CM) may be applied to process manufacturing systems to increase collaboration, automation, process resilience and waste valorisation. A novel CM definition was proposed, which is:

> "The cloud manufacturing model is the concept of sharing manufacturing capabilities and resources on a cloud platform capable of making intelligent decisions to provide the most sustainable and robust manufacturing route available."

It was realised that for *intelligent* CM systems to be built it would require the development of mathematical models to describe the process manufacturing system. Without the development of mathematical models, it would not be possible to virtualise the process manufacturing system onto a cloud platform, and the platform would not be able to analyse or simulate the system to make intelligent decisions.

The following section will explain how the DDMs developed in chapters 5, to describe the H^2AD bioprocess, may be utilised in a CM system to enhance process resilience and waste valorisation. The H^2AD was chosen as a case study because, with the addition of a DDM, it would be able to effectively demonstrate the five CM characteristics of flexibility and scalability, multi-tenancy, intelligent decision-making, on-demand manufacturing and manufacturing as a service.

7.4.1 H^2AD CASE STUDY

Cloud manufacturing systems have three types of users; cloud consumers, cloud providers and cloud operators [3]. Cloud consumers are utilising the cloud services to acquire manufacturing services, and the cloud provider is providing these services. The cloud operator manages the use, performance, and delivery of cloud services as well as negotiating the relationship between cloud providers and cloud consumers. Applying these definitions to the H^2AD case study, the cloud consumers are the process manufacturers who are searching for the best method to valorise the waste produced by their processes. The cloud providers are Lindhurst Engineering Ltd. whose H^2AD bioprocess performs two process manufacturing services. The first is that the bioprocess treats the process manufacturer's waste/wastewater streams to reduce the pollutant load and make the wastewater fit for reuse. The second is that the bioprocess recovers energy from waste streams by generating biogas.

As previously discussed, a CM system has four potential deployment models: private, public, community and hybrid [3]. A private cloud would contain a single tenant's data and information and would provide services specific within that enterprise. By being private within an enterprise, it allows greater security of data. Multiple tenants who have a common interest may share certain knowledge or resources for mutual gain through the community cloud. On the public cloud manufacturing resources and capabilities and data is shared and managed by a third party and is the platform by which tenants from various backgrounds shall be able to interact with each other. Hybrid clouds consist of multiple external and internal clouds. An outline for how the developed DDM would work within all four CM models built for a H^2AD system is proposed below.

7.4.1.1 Private manufacturing cloud

A private CM system, built for the H^2AD system, would have one cloud consumer (waste producer) and one cloud provider (waste valorisation service provider). In Figure 7.3 the interactions between the different cloud users are displayed. A DDM would work within the manufacturing cloud to intelligently:

- Increase the bioprocess's resilience to variability within the feedstock

As already discussed the H^2AD bioprocess performance is influenced by the characteristics of the waste it is treating. The performance is defined as multiple pollutant removal rate and volume of biogas generated. A DDM model uploaded into the manufacturing cloud would be able to make predictions on how the waste's characteristics affect the H^2AD performance. The manufacturing cloud could advise the cloud provider to optimise the H^2AD configuration to account for this variability in the cloud consumer's waste.

As more data is collected (1+ years), the DDM may be able to make predictions on waste characteristics and prepare the H^2AD bioprocess in advance. For example, the model may learn to expect the period of the year where the process manufacturer produces larger than normal volumes of waste. The manufacturing cloud could start making preparations to configure the H^2AD bioprocess to a lower HRT to accommodate increased volume of waste before this occurs. The H^2AD bioprocess is a living system and does not respond well to sudden large changes to the system. Being able to forecast and prepare for major events would increase resilience of the bioprocess. The manufacturing cloud would work to communicate foreseen events to cloud users and coordinate a strategic response.

- Optimise the H^2AD bioprocess for the desired product

The H^2AD system has multiple objectives; pollutant removal and energy recovery from waste streams. In chapter 5, it was demonstrated how a multiple target DDM might be built to optimise the system's sustainability which is dependant on waste characteristics and environmental parameters (e.g. site location, cost of further treatment). The manufacturing cloud would inform the cloud consumer of the most sustainable objective (e.g. maximise pollutant removal), as predicted by the DDM, and if the cloud consumer agrees it would then inform the cloud provider how best to configure the H^2AD to achieve the objective.

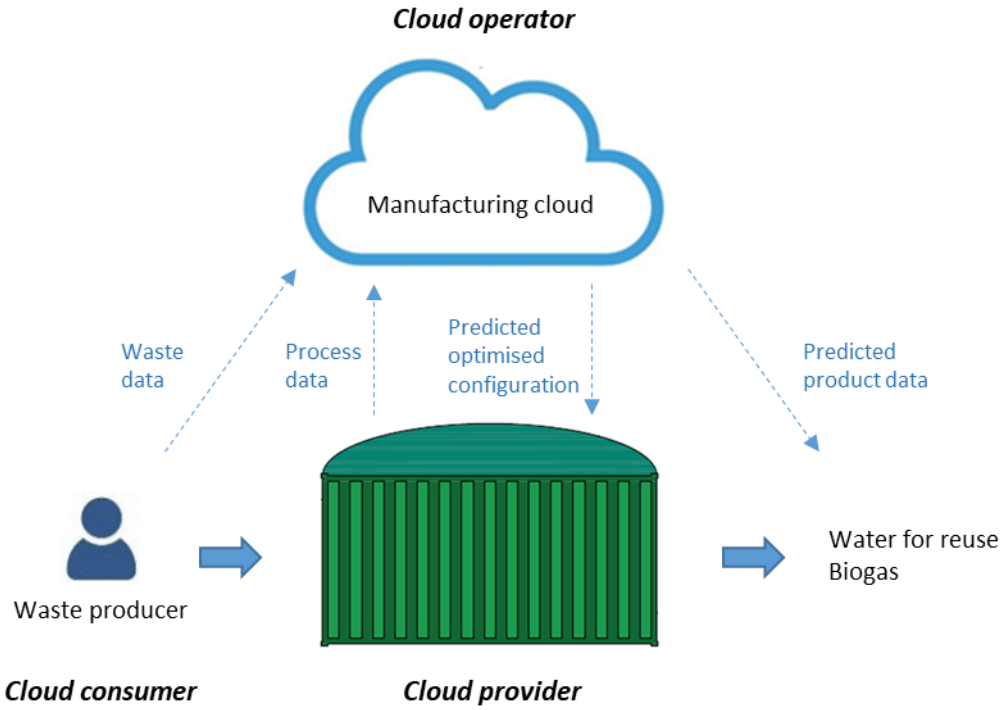


Figure 7.3 Example of a private cloud manufacturing system built for H^2AD system.

7.4.1.2 Community manufacturing cloud

A community CM system, built for the H^2AD system, could have multiple cloud consumers and multiple cloud providers. In Figure 7.4 the interactions between the multiple cloud consumers, single cloud provider and cloud operator are displayed. A community CM system may be suitable for process manufacturers that want to recover value from their waste streams but do not produce a sufficient volume of waste to make utilising the H^2AD system feasible. In this instance, a manufacturing cloud may foster and develop relationships between a group of process manufacturers to combine their waste streams into one of sufficient volume to be treated by the H^2AD bioprocess. In

this system the cloud operator would have an active role in coordinating the collection and processing of the process manufacturers' waste.

A DDM would work within the manufacturing cloud to perform the same services as outlined in the private manufacturing cloud system. In addition to these tasks a DDM would work to intelligently:

- Evaluate what proportion of the products each cloud consumer is owed

The cloud consumers would each be owed a different proportion of the H^2AD products depending on the volume and quality of waste provided as a feedstock for the H^2AD system. Assessing the value of the waste is difficult due to the fact that different pollutants within the waste can benefit or harm the bioprocess. This is further complicated as the pollutants may be beneficial for one objective and detrimental for another. A DDM model can be used to evaluate the waste characteristics effect on the H^2AD, which the cloud operator can use in deciding what proportion of the products is each consumer owed.

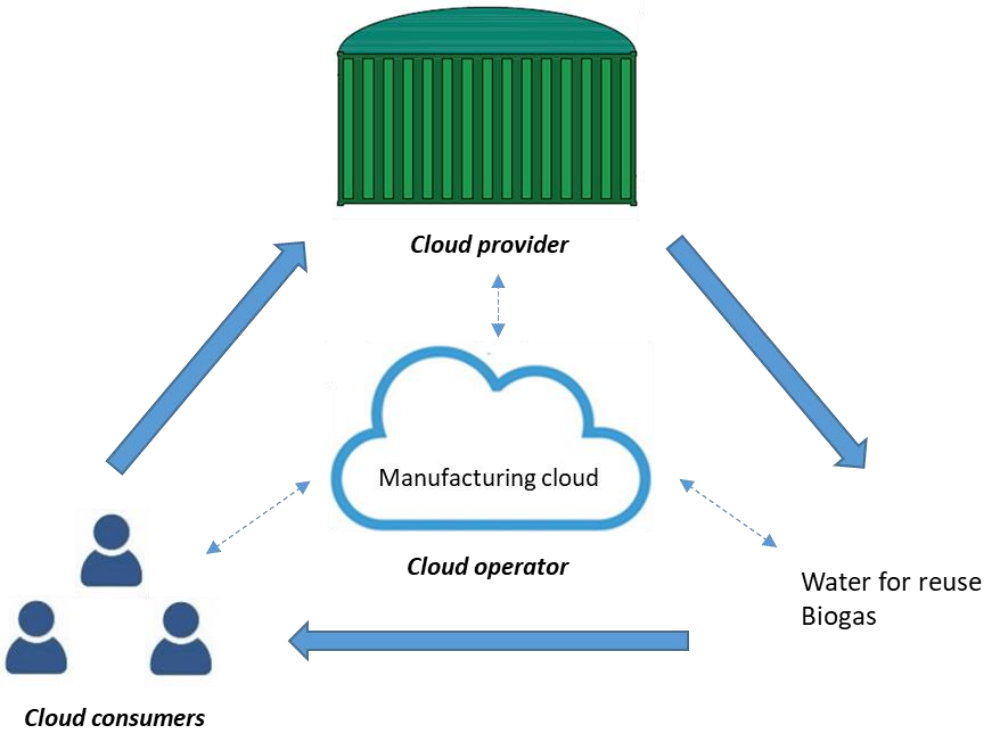


Figure 7.4 Example of a community cloud manufacturing system built for H^2AD system.

7.4.1.3 Public manufacturing cloud

An advantage of the H^2AD bioprocess is that it is capable of treating a variety of process manufacturing sectors waste [211]. Likewise, a process manufacturer, who has value in their waste, has a variety of waste valorisation routes available to choose. The purpose

of a public cloud manufacturing system would be to match waste producers to manufacturers whose processes could recover resources from the waste or view the waste as a feedstock. There would be multiple cloud consumers, and cloud providers and the cloud operator would manage the relationships formed between cloud users. An example of a public cloud manufacturing system built to match waste producers and waste managers is displayed in Figure 7.5.

In chapter 3 a new circular economy model was proposed that explains how resource transfer between manufacturing systems will further extend a resource value and use. Building a public manufacturing cloud that encompasses multiple process manufacturers would help to realise system-to-system circular economy, as the manufacturing cloud would identify manufacturing systems that could use one system's waste as a resource. A DDM model would work within a public cloud manufacturing system to intelligently:

- Predict the most sustainable manufacturing route to recover resources from a process manufacturer's waste stream

A process manufacturer would upload their waste characteristic data to the manufacturing cloud. A DDM within the manufacturing cloud would assess the economic and environmental performance of multiple waste valorisation technologies to find the most sustainable manufacturing route. The DDM model built in chapter 2 would be used to evaluate new waste streams performance when valorised by the H^2AD bioprocess. Further models DDM would be required to evaluate the waste stream on other waste valorisation technologies; for example, anaerobic digestion, incineration or conventional waste/ wastewater treatment plants.

- Predict a waste valorisation technology potential to treat new waste streams

The second use of DDM is a similar application to CM but is from a cloud provider perspective. Instead of uploading data about a waste to be valorised, a process manufacturer who is seeking a waste as a feedstock uploads a DDM describing their system to the manufacturing cloud. The manufacturing cloud then uses the DDM to evaluate the process manufacturer's system suitability to treat the cloud consumers' waste streams.

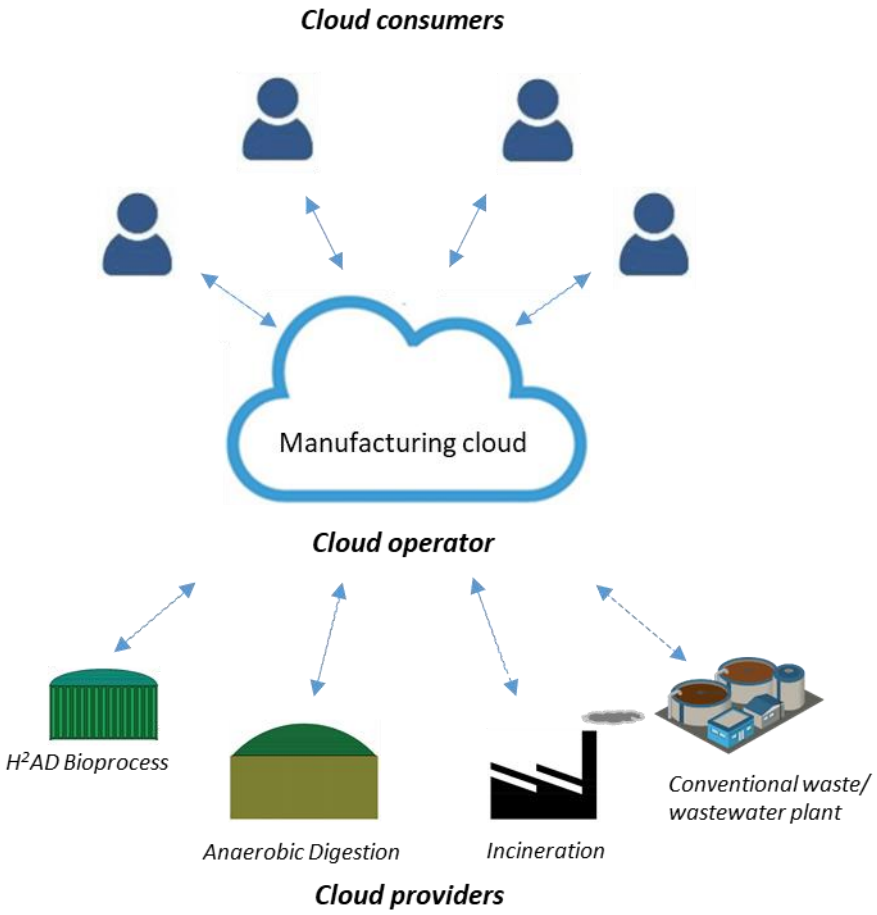


Figure 7.5 Example of a public cloud manufacturing system built for H²AD system.

7.4.1.3 Hybrid manufacturing cloud

The final type of deployment model is the hybrid manufacturing cloud. The hybrid model consists of multiple private, community and public manufacturing clouds. They are built for process manufacturers who want to take advantage of all the possible applications of CM. The reason for having multiple manufacturing clouds, instead of one manufacturing cloud, is data security [70]. There are applications of cloud manufacturing for process resilience and optimisation that require modelling process manufacturing data. A manufacturer's data is highly valuable and if leaked could offer an edge to competitors [410]. Further, there are safety implications as there is the possibility of cybercriminals hacking into a process plant's control system to manipulate and ultimately destroy physical systems [268]. Manufacturing clouds that are being used to optimise and manage the internal workings of a manufacturing process are better suited to private clouds.

7.5 CLOUD MANUFACTURING INCREASING A DATA-DRIVEN MODEL'S PREDICTIVE CAPABILITY

Gómez *et al.* state that the best way to increase a DDM predictive capability is to increase the data available for it to learn from [142]. There is the opportunity for CM to expand the predictive capabilities of DDM by providing a greater volume and variety of manufacturing data to train a DDM. However, chapter 6 showed that this is not always the case, as only one of the four DDM (the ANN model) prediction capabilities increased by increasing the data available to learn from. This was likely to have been because the difference between the manufacturing systems. Increasing the model development data did improve the prediction capabilities of one of the models though, so there is potential that this technique may be more successful in the future. If CM was to be used to develop DDM from multiple process manufacturer's data, there are data confidentially issues that must be addressed.

As previously stated, a process manufacturer's data is valuable, and manufacturers do not readily share their data outside of their organisations. However, manufacturers who have similar systems stand to gain from developing a DDM with greater predictive capabilities. To overcome the data sharing concerns, a cloud operator could act as an impartial third party to develop the DDM. For example, process manufacturers A and B are both using the H^2AD bioprocess to treat farm waste. The data generated from manufacturer A and manufacturer B are shared with a cloud operator. The cloud operator then uses the combined datasets to develop a DDM that describes both manufacturers A and B systems. The model is then returned to both process manufacturers. At no point is manufacturer A's data shared with manufacturer B or vice-versa. Both process manufacturers would contribute to the cost of developing the DDM, and their contribution would be determined by what percentage their data made up the final model development dataset.

7.6 CONCLUSION

- Process manufacturers are not likely to be interested in how a DDM is developed, but in what are the real benefits the models have on their manufacturing process, how can the model be used by the manufacturer, can the manufacturer trust the model and how does the manufacturer pay to develop and maintain the model.
- How a DDM may be used by a process manufacturer to understand and optimise their systems has been with a specific example of how a DDM may provide real-time prediction within a GUI is outlined.

- A DDM model may systematically overestimate or underestimate a prediction. Prediction errors may have a greater negative impact for certain ranges for specific process manufacturing environments. This chapter illustrated the problems this may cause for four process manufacturing scenarios.
- This chapter proposes two different business models for the development and servicing of a DDM. The business models were adapted from how the software industry approach of selling their products as either POA or SOA. For a POA the process manufacturer commissions and owns the DDM. The process manufacturer then pays for the DDM to be serviced as required. For the SOA the developed DDM is jointly owned between the modelling company and the process manufacturer. The process manufacturer pays a subscription to use the model and have the model serviced as needed. The subscription cost is dependent on the frequency of servicing required.
- Cloud manufacturing systems have been proposed for discrete manufacturing systems, but there is limited research on how CM systems that enable features specialised to process manufacturers would work. This chapter addresses this by outlining four different CM systems that would work for the H^2AD case study: private, community, public and hybrid.
- Finally, a new idea of how CM could be used to expand the prediction capabilities of a DDM is presented. This is achieved by sharing data used to develop a DDM between process manufacturers who have similar manufacturing systems. The cloud operator in the manufacturing cloud acts as an impartial third party to develop the model, so no sharing of confidential manufacturing data between competitors occurs.

CHAPTER 8: CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

8.1 CONCLUSIONS

During the 20th Century, process manufacturing systems have been driven by economic incentives to expand the production of goods and services, with little to no regard to the environmental impacts this caused. In the last couple of decades, there has been a recognition by the international community that this is unsustainable, and manufacturers have been pressured to introduce sustainable strategies. However, the progress made has not been sufficient and the resources required to support the global population now exceeds those available. The challenge facing both society and manufacturers is to become not only "carbon-neutral" but also "resource-neutral". In playing their part process manufacturers must redesign their systems around three sustainable manufacturing principals:

1. Replace non-renewable resources with renewable resources where sustainable;
2. Minimise resource waste during production;
3. Utilise waste from one system as a renewable resource for another system.

The main aim of this work was to investigate how process manufacturers can utilise their manufacturing data to intelligently manage resources within their system. It was hypothesised that cloud manufacturing (CM) will enable the transformation of process manufacturing from the traditional linear business model to a sustainable, circular approach. Furthermore, that the implementation of CM to process manufacturing will require development of data-driven models (DDMs) to virtualise process manufacturing systems. The development of DDMs that are representative of process manufacturing systems will face some considerations and challenges that are unique to process manufacturers. To investigate this hypothesis, four research questions have been answered:

1. How does CM offer process manufacturers a new route to achieve sustainable manufacturing?
2. What considerations and challenges do process manufacturers face when developing a DDM that is representative of their manufacturing system?

3. What are the current steps when developing a DDM and what additional steps are required to overcome unique challenges when modelling process manufacturing systems?
4. How will the development of DDMs support the transition from the traditional manufacturing paradigm to CM?

8.1.1 HOW DOES CLOUD MANUFACTURING OFFER PROCESS MANUFACTURERS A NEW ROUTE TO ACHIEVE SUSTAINABLE MANUFACTURING?

Recent developments in technologies that are digitalising manufacturing have seen a rush of research into investigating the opportunities they offer. However, a framework for the infrastructure these different technologies shall operate within is currently lacking. Cloud manufacturing has the potential to act as an intelligent platform to interact remotely with the increasingly digitalised supply chain. This complete monitoring across the supply chain allows CM to identify, coordinate and implement opportunities for increased sustainability.

Cloud manufacturing application has mainly been theorised from a discrete manufacturing point of view. Therefore, the full potential of CM when applied to process manufacturing systems had yet to be considered. This work developed four key methods of how CM increases process manufacturing sustainability:

- **Collaborative design**: A critical barrier preventing sustainable manufacturing designs is the accessibility of environmental information and expert knowledge within an enterprise. Cloud manufacturing operates on a multi-tenancy cloud platform that can crowdsource information and expertise to simulate the development of innovative sustainable designs. Process manufacturing has unique opportunities for new sustainable design from its ability to manipulate the manufacturing route to valorise waste and by-product streams.
- **Automation**: Cloud manufacturing can perform real time analyses of manufacturing and market data to intelligently schedule and plan the manufacturing process. This allows for CM to optimise the manufacturing process in real time and coordinate its implementation across the supply lines from a centralised management centre. This reduces the manufacturing costs through minimising energy consumption, material waste and downtime. Process manufacturing will become further automated by the implementation of advanced sensors, allowing CM to develop intelligent models. These models help to identify and implement more sustainable manufacturing solutions.

- **Waste valorisation**: Process manufacturing waste has the potential for valorisation as it can contain valuable material residue, co-products opportunities, residual heat, and energy recovery. It is not always obvious which is the most sustainable waste valorisation route, and CM can be used as an intelligent decision-making tool. CM can monitor functions in the composition of waste streams and market prices to identify and coordinate alternative waste valorisation routes if they prove to become significantly more sustainable. Furthermore, CM can coordinate local supply chains to reuse waste streams as a feedstock. From these methods, CM aligns with the circular economy vision perceiving waste as a valuable resource to be reused, recovered and regenerated.
- **Process resilience**: Process manufacturing systems are particularly susceptible to variations in process conditions, especially feedstock composition. Currently over engineering is the best method to ensure variations do not impact product quality. Advanced data analytic techniques offer the potential to gain better process understanding to ensure quality in the design of the system. As advanced industrial internet of things (IIoT) sensors are developed, this will provide CM with greater quantity and granularity of data meaning previously unidentifiably solutions can be implemented.

8.1.2 WHAT CONSIDERATIONS AND CHALLENGES DO PROCESS MANUFACTURERS FACE WHEN DEVELOPING A DATA-DRIVEN MODEL THAT IS REPRESENTATIVE OF THEIR PROCESS MANUFACTURING SYSTEMS

The use of DDMs is on the rise across manufacturing primarily due to the following three points:

1. There is an increasing amount of data generated in manufacturing as IIoT technologies become more widespread.
2. Manufacturers have easier access to the computational power required to harness this data due to cloud computing enabling parallel processing.
3. New machine learning techniques are enabling the utilisation of new sources of data, such as texts, image, audio, video, log files.

A protocol was written for process manufacturers to follow before developing process manufacturing. Before attempting to develop a DDM there are considerations and challenges the process manufacturer must overcome. The considerations identified were:

1. The model's goal must be clearly defined and should aim to solve the manufacturing problem that has been identified.
2. The manufacturer should also consider what is required from the model. Is it enough simply to fit the data, or will the model be required to make predictions? Being able to make predictions increases the model's value but also its complexity. The accuracy in which the model is required to perform these functions should be defined by the product specification, regulations, economic value and safety requirements.
3. Paramount to the success of DDMs is the generation/collection of a representative set of data used to develop the model. This data is split into training data, used to train the model, and unseen validation data to evaluate the prediction capability of the model. Before generating/collecting this data the manufacturer should consider what data is already available and if it is representative of the system.

Challenges arise when ensuring a representative data set is used for model development. Process manufacturers face unique challenges regarding this point because of the limited availability and variability of industrial data. Even when data is available, knowing if it representative of the system and relevant to the model can be challenging. Furthermore, the boundaries of a process manufacturing system are hard to classify. The difficulty occurs in knowing when enough data has been collected to be representative of the whole system, an area not explored in the literature.

Once a manufacturer is able to overcome the considerations and challenges involved when modelling a process manufacturing system, DDMs will allow the process manufacturers to:

1. **Characterise process streams**: Through the development of soft sensors that utilise recent advances in machine learning able to utilise cheaper, unclean data. This will allow process manufacturers to develop a deeper understanding of their processes, providing optimisation opportunities in a cost-effective manner.
2. **Enhance process resilience**: By fitting data to the DDM, manufacturers are able to make accurate predictions on the effect that variability across the system has on their process and take measures to optimise them.
3. **Evaluate waste valorisation routes**: Existing DDMs may be used as a tool to evaluate new waste streams' compatibility with an existing waste valorisation

technology. Alternatively, use DDMs as a tool for manufacturers to see which technology is the most sustainable option for their waste.

8.1.3 WHAT ARE THE CURRENT STEPS WHEN DEVELOPING A DATA-DRIVEN MODEL AND WHAT ADDITIONAL STEPS ARE REQUIRED TO OVERCOME THE UNIQUE CHALLENGES WHEN MODELLING PROCESS MANUFACTURING SYSTEMS?

The development of DDMs for manufacturing systems plays a significant role in understanding relationships within the system and making predictions. However, in order for a model to be trusted, the model's predictions capabilities must be thoroughly evaluated. This is of greater importance to manufacturing systems where data is limited because models built from limited data are in danger of overfitting to the model development data. To avoid overfitting, the developed DDM's prediction capabilities should be evaluated on unseen data, taken both inside and outside of the system. Defining the boundaries of a manufacturing system and understanding the space inside the system is made more challenging because manufacturing systems are often highly dimensional. This work developed a methodology to determine the manufacturing system boundaries using both statistical and data visualisation techniques. Once the system boundaries are defined, it allows for:

1. Determining the DDM's interpolation and extrapolation capabilities: knowing the system boundaries within the data allows unseen data to be sampled from inside and outside the system. This can then be used to assess the model's interpolation and extrapolation capabilities. As part of this work, a systematic method for partitioning unseen data was proposed. This method ensures unseen data extends evenly across and beyond the system.
2. Knowing when to retrain the model: Manufacturing systems are continuously evolving and a DDM built from one set of data will likely deteriorate with time, as the system moves beyond the original defined system. By defining the system boundaries, manufacturers can utilise visualisation techniques, like parallel coordinate plot (PCP), to monitor when the system moves beyond the system boundaries. The model will then need retraining, incorporating the new data into the model development data. A traffic light protocol for easily classifying if new data collected is inside or outside the system was developed for this work.

In the past, manufacturers have often chosen their machine learning algorithm based on previous research done in their field. Recent advancements in the field of machine

learning have led to the development of many accurate and advanced machine learning techniques (algorithms, optimisation procedures, data pre-processing). Furthermore, the No Free Lunch Theorem states that the optimal machine learning technique cannot be chosen a priori, and different algorithms need to be evaluated to find their suitability for a given system. This study compares the performance of four machine learning algorithms, Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural network (ANN). The models' capability to predict the percentage of COD removal from wastewater by the H^2AD bioprocess was evaluated statistically. The statistical techniques used were the coefficient of determination (R^2), the root mean squared error (RMSE), the mean absolute error (MAE) and mean absolute percentage error (MAPE). It was found that all the models performed well at fitting the model development data, with the GP model performing the best (R^2 of 0.97, RMSE of 1.202, MAE of 0.383 and MAPE of 0.549). However, when evaluated on unseen data the GP model prediction capability decreased (R^2 of 0.65, RMSE of 5.328, MAE of 8.879 and MAPE of 16.686), indicating that it had overfitted the model development data. Based on their ability to fit both the model development data and the unseen data, the RF model showed the greatest overall performance (R^2 of 0.99, RMSE of 1.4399, MAE of 2.811 and MAPE of 6.039).

The modelling methodology capability to develop MT models was then explored.

When searching for the most sustainable manufacturing solution, multiple economic and environmental objectives must be considered. The objectives often come into conflict with one another. By developing a multiple target (MT) model of the manufacturing system, one would be able to investigate the impact of various manufacturing factors on the objectives and the objectives relate to one another. This will support manufacturers to determine the most sustainable manufacturing route. Despite recent advances in MT modelling, there are limited examples of MT models being developed to process manufacturing systems. The MT models that have been developed frequently use a problem transformation approach, where multiple independent ST models are developed for each output. This fails to capture the relationship between the model's outputs, resulting in models with inferior predictive capabilities.

In this study, two state-of-the-are MT modelling approaches were applied to a manufacturing system for the first time. These were regression chain (RC) based techniques called ensemble of regressor chains (ERC) and maximum correlation chain (MCC). The RC technique treats the outputs of the system as inputs for subsequent

outputs, thus capturing the relationship between model outputs. Previous work has used only RF and SVR algorithms as base models for RC models. This study was the first example of an ANN algorithm being used for either an ERC or a MCC MT model. Previous work has shown the versatility of ANNs when modelling process manufacturing systems [111, 118]. This was reflected in this study, as the ANN was able to consistently outperform the ERC and MCC built from SVR base models. After statistical analysis of the model predictions on unseen data and after comparing the models' predictions distributions with that of the real data, the MCC-ANN model was found to be the superior model.

Utilising the MCC-ANN model to develop surface plots of %COD removal, %TSS removal and CH_4 generation, it was found that the %COD and CH_4 generation favoured closely similar configurations of the H^2AD system's process conditions. However, the %TSS removal was found to have the opposite relationship with all process conditions apart from the system temperature. This demonstrated the complex task when identifying the most sustainable solution, balancing then economic and environmental considerations. What is sustainable for one manufacturing environment may not be for another, as different environments have different economic and environmental objectives. An economic benefit up of £10,950 was achieved from using the MCC-ANN model to find the optimal H^2AD configuration for a manufacturing environment that charges more for wastewater disposal.

The modelling methodology proposed by in the work was tested to model the treatment of mining wastewater by an up-flow surface catalysed oxidation of ferrous iron reactor (SCOOFI) to remediate Fe. The data was supplied by the Coal Authority who were interested in understanding the variability in the Fe removal rates. Three different datasets were supplied from three SCOOFI reactors with different configurations.

Two modelling strategies were trialled to investigate the effect of expanding the data available to DDM to learn from. The statistical performance of the two strategies found that the average R^2, RMSE and MAE for the models built using one dataset was R^2 of 0.46, RMSE of 10.43 and MAE of 7.91, and those built with three datasets were R^2 of 0.41, RMSE of 10.99 and MAE of 8.71. Therefore, it was found that the DDMs' generally performed best when learning from data directly describing the system being modelled. The expectation to this was the ANN model, whose performance increased when trained on all three reactors datasets (R^2 0.19 → 0.56, RMSE 12.99 → 9.54, MAE 10.24 → 8.90). Four different machine learning algorithms were built and compared. To determine which model best described the process manufacturing system, the

models were evaluated based on their accuracy when making predictions on unseen data.

The GP model was determined to best describe the removal of Fe from wastewater by an up flow SCOOFI reactor. The model's statistical capabilities on unseen data were R^2 of 0.74, RMSE of 7.42 and MAE of 5.00. The GP model was then used to produce surface plots to investigate each of the system parameters effects on Fe% removal. From the GP-based analysis the highest Fe removal occurred at high concentrations of Fe and TSS in the wastewater and up to 55 days since desludging the reactor. The other parameters had conflicting impacts on Fe% Removed, dependent on the other parameters. This illustrates the complex, non-linearity of the system and why a DDM analysis is required.

8.1.4 HOW WILL THE DEVELOPMENT OF DATA-DRIVEN MODELS SUPPORT THE TRANSITION FROM THE TRADITIONAL MANUFACTURING PARADIGM TO CLOUD MANUFACTURING?

The process of making a DDM that is representative of the process manufacturing is complex, with a number of consideration and challenges to address. However, the process manufacturer is unlikely to be interested in the process of developing a DDM or what is the architecture of the final model. This work outlined how DDM analysis may be employed so a process manufacturer can gain knowledge and insights about the relationship between the system parameters. Furthermore, a DDM prediction capability for forecasting analysis was discussed and the potential use of a DDM to forecast composition and supply of raw materials. How a DDM may be used by a process manufacturer to understand and optimise their systems is explained in this work. An example of how a DDM may provide real-time prediction within a graphical user interface (GUI) is outlined.

A DDM model may systematically overestimate or underestimate a prediction. The previous work modelling process manufacturing systems use one or more statistical metrics to evaluate prediction error. However, it is rare for residual plots to be produced because regression plots are favoured instead due to regression being simple to understand. This work demonstrates how residual plots can be used to better understand where a DDM is overestimating or underestimating. Knowing this is important, as there are certain ranges that have a greater negative impact for when predicting certain ranges for specific process manufacturing environments.

This work proposes two different business models for the development and servicing of a DDM. In the product-oriented architecture (POA) business model the process

manufacturer commissions and owns the DDM. Whereas, in the service-oriented architecture (SOA) the DDM is jointly owned by the modelling company and the manufacturer. The suitable of each business model depends on the variability within the process manufacturing system. A system with higher levels of variability, which will mean the model requires frequent retraining, will favour the SOA approach; while, the POA approach will be favoured by stable systems.

Cloud manufacturing systems have been proposed for discrete manufacturing systems, but there is limited research on how CM systems that enable features specialised to process manufacturers would work. This chapter addresses this by outlining four different CM systems, private, community, public and hybrid would work for the H^2AD process manufacturing system. Finally, a new idea of how CM could be used to expand the prediction capabilities of a DDM is presented. This is achieved by sharing that data used to develop a DDM between process manufacturers who have similar manufacturing systems. The cloud operator in the manufacturing cloud acts as an impartial third party to develop the model, so no sharing of confidential manufacturing data between competitors occurs.

8.2 RECOMMENDATIONS AND FUTURE WORK

On reflection there are two main areas w

1. Generating additional data from lab-scale bioelectrochemical systems;
2. Utilising a data driven model for self-optimisation;

Four areas have been identified for future work to expand and increase the impact of the research presented within this body of work:

1. Manufacturing products form process manufacturers waste streams;
2. Further developing cloud manufacturing systems of process manufacturing environments;
3. Resource recovery by the H^2AD bioprocess.

8.2.1 GENERATING ADDITIONAL DATA FROM LAB-SCALE BIOELECTROCHEMICAL SYSTEMS

The modelling techniques developed in this work were formed from the experience of modelling the H^2AD bioprocess. However, the H^2AD system presented data availability challenges. The H^2AD bioprocess has a minimum 2 weeks start-up period during which the biofilms on the anodes and cathodes are grown. During this period the H^2AD performance is limited and data collected from the system does not describe the bioprocess operating at steady-state conditions. Due to several factors, the H^2AD

bioprocess was forced to shut down and start-up multiple times, reducing the volume of data collected at steady-state conditions. This limited the data available, making it challenging to produce meaningful conclusions from the modelling techniques. This challenge was overcome by generating synthetic H^2AD output data from trends observed in historical H^2AD data. The synthetic data was used to develop and demonstrate the data-driven modelling techniques proposed for this work. The modelling techniques were validated by modelling another process manufacturing system, the wastewater treatment of acid mine water. However, additional data could have been generated from lab-scale microbial fuel cell experiments that attempted to mimic the H^2AD configuration. This would have allowed for additional validation of the modelling techniques developed.

8.2.2 UTILISING A DATA-DRIVEN MODEL FOR SELF-OPTIMISATION

The aim of developing DDMs was to demonstrate CM capability for intelligent decision making to find the most sustainability manufacturing route. The DDMs built for this work have been used to optimise process manufacturing systems by simulating different system configurations and analysing the predicted results. This application of DDM-analysis is beneficial to the process manufacturer because it does not require costly experimental work. However, this method still required human input in determining the most sustainable manufacturing route. There is the potential to completely automatise the optimisation procedure using DDMs. Schweidtmann et al. demonstrated how self-optimisation could be achieved by either fitting the DDM's predictions to a function that may be used to find a single optimal point of the system or by establishing the Pareto front for each process condition [110]. Another approach is reinforcement learning. Reinforcement learning is a field of machine learning concerned with the problem of finding suitable actions to take in a given situation in order to maximise a reward. Here the algorithm is not given examples of optimal outputs but must instead discover them by process of trial and error [137]. By including the process conditions in the model outputs would enable a reinforcement learning algorithm to find the optimal process conditions. These two methods may provide alternative optimal solutions for the H^2AD system and the iron wastewater treatment system that are unbiased by human interpretation of results. If this work was to be repeated, it would prove interesting to compare the self-optimisation approaches to the simulation approach. The self-optimisation would still require human intervention to validate the final decision; however, it would greatly increase CM's ability to automate the decision-making process."

8.2.3 MANUFACTURING PRODUCTS FROM PROCESS MANUFACTURERS WASTE STREAMS

There needs to be a shift across society (inclusive from manufacturer to consumer) from viewing waste not as something that requires disposing of, but as a co-product that adds value to the same and/or related manufacturing process. Process manufacturing systems are in an ideal situation as their waste can contain valuable material available to use as a building block for their or a related process manufacturing environment. Technologies to recover resource from process manufacturers in the form of heat and water are well established in process manufacturing. Recently there have been new technologies developed to recover valuable and scarce resources. The next step is going beyond recourse recovery and directly manufacturer products from a process manufacturer's waste stream. This work recommends further research on how process manufacturing systems can be designed so that products are manufactured from waste streams.

8.2.4 FURTHER DEVELOPING CLOUD MANUFACTURING SYSTEMS OF PROCESS MANUFACTURING ENVIRONMENTS

Cloud manufacturing has great potential to help achieve the goal of resource-neutral process manufacturing. Cloud manufacturing works enables this goal by acting as the manufacturing paradigm through which resources can move between process manufacturers. A CM system may intelligently match resources leaving one system as a waste to systems which view this waste as a resource. Collaborative work between industry and academia is needed to develop case studies demonstrating CM being used to manage resource transfer between manufacturing systems.

8.2.5 RESOURCE RECOVERY BY THE H^2AD BIOPROCESS

The H^2AD bioprocess is configured to reduce pollutant load in a process manufacturer's wastewater to make it fit for reuse, whilst simultaneously recovering energy. The H^2AD system is one of the first examples of a plant scale BES. However, the resource recovery potential of the H^2AD bioprocess could be expanded to include other valuable materials. For example, previous work has shown BES capability to recover nutrients (phosphorous and nitrogen) [337] and heavy metals [411] from wastewaters. Further work is required to develop systems that could separate and purify the valuable materials contained with the H^2AD output streams. The DDM would require redevelopment to predict not only chemical oxygen demand (COD) and total suspended solids (TSS) in the outputs stream, but also the quantity of resources recovery. This would expand the capabilities of the DDM to determine the most sustainable H^2AD configuration.

REFERENCES

[1] The World Bank. World Bank Open Data | Data 2017. https://data.worldbank.org/ (accessed July 11, 2019).

[2] Brandl D. 1. Manufacturing Control. Des Patterns Flex Manuf 2007.

[3] Fisher O, Watson N, Porcu L, Bacon D, Rigley M, Gomes RL. Cloud manufacturing as a sustainable process manufacturing route. J Manuf Syst 2018;47. https://doi.org/10.1016/j.jmsy.2018.03.005.

[4] Supanchaiyamat N, Hunt AJ. Conservation of critical elements of the periodic table. Chem Sus Chem 2019;12:397–403. https://doi.org/10.1002/cssc.201802556.

[5] European Commission. 2017 List of critical raw materials for the EU. 2017.

[6] Puyol D, Batstone DJ, Hülsen T, Astals S, Peces M, Krömer JO. Resource recovery from wastewater by biological technologies: opportunities, challenges, and prospects. Front Microbiol 2017;7:2106. https://doi.org/10.3389/fmicb.2016.02106.

[7] Waste and Resources Action Program. WRAP and the circular economy | WRAP UK 2018. http://www.wrap.org.uk/about-us/about/wrap-and-circular-economy; (accessed November 15, 2018).

[8] Charles RG, Douglas P, Baker JA, Carnie MJ, Douglas JO, Penney DJ, Watson TM. Platinized counter-electrodes for dye-sensitised solar cells from waste thermocouples: A case study for resource efficiency, industrial symbiosis and circular economy. J Clean Prod 2018;202:1167–78. https://doi.org/10.1016/j.jclepro.2018.08.125.

[9] Álvarez R, Ruiz-Puente C. Development of the tool SymbioSyS to support the transition towards a circular economy based on industrial symbiosis strategies. Waste Biomass Valori 2017;8:1521–30. https://doi.org/10.1007/s12649-016-9748-1.

[10] Fraccascia L, Albino V, Garavelli CA. Technical efficiency measures of industrial symbiosis networks using enterprise input-output analysis. Int J Prod Econ 2017;183:273–86. https://doi.org/10.1016/j.ijpe.2016.11.003.

[11] National Industrial Symbiosis Programme - International Synergies n.d. https://www.international-synergies.com/projects/national-industrial-

symbiosis-programme/ (accessed November 27, 2019).

[12] Cleantech Östergötland | Nätverket för miljöteknikföretagen i Östergötland n.d. https://cleantechostergotland.se/ (accessed November 27, 2019).

[13] Home | Kalundborg Symbiose - n.d. http://www.symbiosis.dk/en/ (accessed November 27, 2019).

[14] Taddeo R, Simboli A, Ioppolo G, Morgante A. Industrial symbiosis, networking and innovation: The potential role of innovation poles. Sustainability 2017;9:169. https://doi.org/10.3390/su9020169.

[15] Business Dictionary. Manufacturing n.d. http://www.businessdictionary.com/definition/manufacturing.html (accessed September 28, 2019).

[16] Stott R. Artisans and capitalist development. J Early Repub 1996;16:257. https://doi.org/10.2307/3124249.

[17] World Trade Organization. WTO | Trade Statistics - World Trade Statistical Review 2016 n.d. https://www.wto.org/english/res_e/statis_e/wts2016_e/wts16_toc_e.htm (accessed September 29, 2019).

[18] HM Treasury. Budget 2011. London: 2011.

[19] The World Bank. Manufacturing value added (% GDP) - 2017 n.d. http://data.worldbank.org/indicator/NV.IND.MANF.ZS; (accessed April 10, 2017).

[20] Oxford Economics. The true impact of UK manufacturing. Oxford: 2018.

[21] Ali Zaidi, Colin Howat, Hayley Limmer. Sector Insights: skills and performance challenges in the advanced manufacturing sector. London: 2015.

[22] Towards a next generation of water systems and services for the circular economy. | NextGen Project | H2020 | CORDIS | European Commission n.d. https://cordis.europa.eu/project/rcn/216646/factsheet/en (accessed July 23, 2019).

[23] Schmenner RW. Looking ahead by looking back: swift, even flow in the history of manufacturing. Prod Oper Manag 2009;10:87–96. https://doi.org/10.1111/j.1937-5956.2001.tb00069.x.

[24] Greenwood J. Revolution: technology, productivity, and income inequality. Econ Rev 1999:2–12.

[25] Lasi H, Fettke P, Kemper HG, Feld T, Hoffmann M. Industry 4.0. Bus Inf Syst Eng 2014;6:239–42. https://doi.org/10.1007/s12599-014-0334-4.

[26] Sadati N, Chinnam RB, Nezhad MZ. Observational data-driven modeling and optimization of manufacturing processes. Expert Syst Appl 2018;93:456–64. https://doi.org/10.1016/j.eswa.2017.10.028.

[27] Eastin MS, Brinson NH, Doorey A, Wilcox G. Living in a big data world: Predicting mobile commerce activity through privacy concerns. Comput Human Behav 2016;58:214–20. https://doi.org/10.1016/j.chb.2015.12.050.

[28] Morrar R, Arman H, Mousa S. The fourth industrial revolution (Industry 4.0): A social innovation perspective. Technol Innov Manag Rev 2017;7:12–20.

[29] Giret A, Trentesaux D, Prabhu V. Sustainability in manufacturing operations scheduling: A state of the art review. J Manuf Syst 2015;37:126–40. https://doi.org/10.1016/J.JMSY.2015.08.002.

[30] Prisecaru P. Challenges of the fourth industrial revolution. Knowl Horizons Econ 2016;8:57.

[31] Ang J, Goh C, Saldivar A, Li Y. Energy-efficient through-life smart design, manufacturing and operation of ships in an Industry 4.0 environment. Energies 2017;10:610. https://doi.org/10.3390/en10050610.

[32] Ilgin MA, Gupta SM. Environmentally conscious manufacturing and product recovery (ECMPRO): A review of the state of the art. J Environ Manage 2010;91:563–91. https://doi.org/10.1016/J.JENVMAN.2009.09.037.

[33] Purvis B, Mao Y, Robinson D. Three pillars of sustainability: in search of conceptual origins. Sustain Sci 2019;14:681–95. https://doi.org/10.1007/s11625-018-0627-5.

[34] Nations United. Report of the world summit on sustainable development, Johannesburg, South Africa, 26 August-4 September 2002. New York: 2002.

[35] Macnaghten P, Jacobs M. Public identification with sustainable development: Investigating cultural barriers to participation. Glob Environ Chang 1997;7:5–24. https://doi.org/10.1016/S0959-3780(96)00023-4.

[36] Bey N, Hauschild MZ, McAloone TC. Drivers and barriers for implementation of environmental strategies in manufacturing companies. CIRP Ann 2013;62:43–6. https://doi.org/10.1016/J.CIRP.2013.03.001.

[37] European Commission. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Off J Eur Communities 2008;51.

[38] European Commission. Directive 2008/98/EC of the European Parliament and of the Council of 19 November 2008 on waste and repealing certain Directives. Off J Eur Union 2008;34.

[39] Rob Hayward, Jennifer Lee, Justin Keeble, Robert McNamara, Carrie Hall, Sean Cruse, Peter Lacy. The UN global compact-Accenture CEO study on sustainability 2013. 2013.

[40] García-Martos C, Rodríguez J, Sánchez MJ. Modelling and forecasting fossil fuels, CO2 and electricity prices and their volatilities. Appl Energy 2013;101:363–75. https://doi.org/10.1016/J.APENERGY.2012.03.046.

[41] Siemieniuch CE, Sinclair MA, Henshaw MJ deC. Global drivers, sustainable manufacturing and systems ergonomics. Appl Ergon 2015;51:104–19. https://doi.org/10.1016/J.APERGO.2015.04.018.

[42] Nidumolu R, Prahalad CK, Rangaswami MR. Why sustainability is now the key driver of innovation. IEEE Eng Manag Rev 2013;41:30–7. https://doi.org/10.1109/EMR.2013.6601104.

[43] Shi H, Peng SZZ, Liu Y, Zhong P. Barriers to the implementation of cleaner production in Chinese SMEs: government, industry and expert stakeholders' perspectives. J Clean Prod 2008;16:842–52. https://doi.org/10.1016/j.jclepro.2007.05.002.

[44] Winans K, Kendall A, Deng H. The history and current applications of the circular economy concept. Renew Sustain Energy Rev 2017;68:825–33. https://doi.org/10.1016/J.RSER.2016.09.123.

[45] What is a Circular Economy? | Ellen MacArthur Foundation n.d. https://www.ellenmacarthurfoundation.org/circular-economy/concept (accessed September 28, 2019).

[46] Boulding K. The economics of the coming spaceship earth. In: Jarrett H, editor.

Sixth Resour. Futur. Forum Environ. Qual. a Grow. Econ. Washingt., Washington: 1966, p. 3–14.

[47] Reike D, Vermeulen WJV, Witjes S. The circular economy: New or refurbished as CE 3.0? — exploring controversies in the conceptualization of the circular economy through a focus on history and resource value retention options. Resour Conserv Recycl 2018;135:246–64. https://doi.org/10.1016/J.RESCONREC.2017.08.027.

[48] Gertsakis J, Lewis H. Sustainability and the Waste Management Hierarchy: a discussion paper on the waste management hierarchy and its relationship to sustainability. RMIT Univ Melb 2003:1–15.

[49] Blomsma F, Brennan G. The emergence of circular economy: A new framing around prolonging resource productivity. J Ind Ecol 2017;21:603–14. https://doi.org/10.1111/jiec.12603.

[50] Geng Y, Doberstein B. Developing the circular economy in China: Challenges and opportunities for achieving "leapfrog development." Int J Sustain Dev World Ecol - INT J Sustain DEV WORLD ECOL 2008;15:231–9. https://doi.org/10.3843/SusDev.15.3:6.

[51] Murray A, Skene K, Haynes K. The circular economy: An interdisciplinary exploration of the concept and application in a global context. J Bus Ethics 2017;140:369–80. https://doi.org/10.1007/s10551-015-2693-2.

[52] European Commission. European Commission's "Circular Economy Action Plan" - United Nations Partnerships for SDGs platform n.d. https://sustainabledevelopment.un.org/partnership/?p=29808 (accessed September 28, 2019).

[53] D'Amato D, Droste N, Allen B, Kettunen M, Lähtinen K, Korhonen J, Leskinen P, Matthies BD, Toppinen A. Green, circular, bio economy: A comparative analysis of sustainability avenues. J Clean Prod 2017;168:716–34. https://doi.org/10.1016/J.JCLEPRO.2017.09.053.

[54] Li H, Bao W, Xiu C, Zhang Y, Xu H. Energy conservation and circular economy in China's process industries. Energy 2010;35:4273–81. https://doi.org/10.1016/J.ENERGY.2009.04.021.

[55] Reh L. Process engineering in circular economy. Particuology 2013;11:119–33. https://doi.org/10.1016/J.PARTIC.2012.11.001.

[56] Velenturf APM, Archer SA, Gomes HI, Lag-brotons AJ, Purnell P. Circular economy and the matter of integrated resources. Sci Total Environ 2019. https://doi.org/10.1016/j.scitotenv.2019.06.449.

[57] Li B, Zhang L, Wang S-L, Tao F, Cao J, Jiang X-D, Song X, Chai X-D. Cloud manufacturing: A new service-oriented networked manufacturing model. Comput Integr Manuf Syst CIMS 2010;16:1-7+16.

[58] Mahesh M, Ong SKK, Nee AYCYC, Fuh JYHYH, Zhang YFF. Towards a generic distributed and collaborative digital manufacturing. Robot Comput Integr Manuf 2007;23:267–75. https://doi.org/10.1016/j.rcim.2006.02.008.

[59] Montreuil B, Frayret J-MM, D'Amours S. A Strategic framework for networked manufacturing. Comput Ind 2000;42:299–317. https://doi.org/10.1016/S0166-3615(99)00078-0.

[60] Mehrabi MG, Ulsoy AG, Koren Y, Mehrabi, M. G. M, Ulsoy, A. G., Koren, Y. Reconfigurable manufacturing systems: Key to future manufacturing. J Intell Manuf 2000;11:403–19. https://doi.org/10.1023/A:1008930403506.

[61] Da Silveira G, Borenstein D, Fogliatto FS. Mass customization: Literature review and research directions. Int J Prod Econ 2001;72:1–13. https://doi.org/10.1016/S0925-5273(00)00079-7.

[62] Gao W, Zhang Y, Ramanujan D, Ramani K, Chen Y, Williams CB, Wang CCL, Shin YC, Zhang S, Zavattieri PD. Computer-aided design the status, challenges, and future of additive manufacturing in engineering. Comput Des 2015;69:65–89. https://doi.org/10.1016/j.cad.2015.04.001.

[63] Ministry of Science and Technology of the People's Republic of China. National High-tech R&D Program 2016.

[64] Tao F, Zhang L, Venkatesh VC, Luo Y, Cheng Y. Cloud manufacturing: A computing and service-oriented manufacturing model. Proc Inst Mech Eng Part B J Eng Manuf 2011;225:1969–76. https://doi.org/10.1177/0954405411405575.

[65] Xu X. From cloud computing to cloud manufacturing. Robot Comput Integr Manuf 2012;28:75–86. https://doi.org/10.1016/j.rcim.2011.07.002.

[66] Manucloud-project.EU. Project Overview - The ManuCloud Project 2016. http://www.manucloud-project.eu/; (accessed December 5, 2016).

[67] The Plex Manufacturing Cloud n.d. https://www.plex.com/products.html

(accessed July 17, 2019).

[68] Liu K, Zhong P, Zeng Q, Li D, Li S. Application modes of cloud manufacturing and program analysis. J Mech Sci Technol 2017;31:157–64. https://doi.org/10.1007/s12206-016-1215-1.

[69] Zhang Y, Xi D, Li R, Sun S. Task-driven manufacturing cloud service proactive discovery and optimal configuration method. Int J Adv Manuf Technol 2016;84:29–45. https://doi.org/10.1007/s00170-015-7731-9.

[70] Yang C, Shen W, Lin T, Wang X. A hybrid framework for integrating multiple manufacturing clouds. Int J Adv Manuf Technol 2016;86:895–911. https://doi.org/10.1007/s00170-015-8177-9.

[71] Qiu X, He G, Ji X. Cloud manufacturing model in polymer material industry. Int J Adv Manuf Technol 2016;84:239–48. https://doi.org/10.1007/s00170-015-7580-6.

[72] Guo L. A system design method for cloud manufacturing application system. Int J Adv Manuf Technol 2016;84:275–89. https://doi.org/10.1007/s00170-015-8092-0.

[73] Zhou J, Yao X. Hybrid teaching–learning-based optimization of correlation-aware service composition in cloud manufacturing. Int J Adv Manuf Technol 2017;91:3515–33. https://doi.org/10.1007/s00170-017-0008-8.

[74] Jin H, Yao X, Chen Y. Correlation-aware QoS modeling and manufacturing cloud service composition. J Intell Manuf 2017;28:1947–60. https://doi.org/10.1007/s10845-015-1080-2.

[75] Schaefer D, Thames L, Wu D, Yim S, Rosen D. Distributed collaborative design and manufacture in the cloud - motivation, infrastructure, and education. Comput Educ J 2012;3:1–16.

[76] Wu D, Thames JL, Rosen DW, Schaefer D. Enhancing the product realization process with cloud-based design and manufacturing systems. J Comput Inf Sci Eng 2013;13:041004. https://doi.org/10.1115/1.4025257.

[77] Wu D, Greer MJ, Rosen DW, Schaefer D. Cloud manufacturing: Strategic vision and state-of-the-art. J Manuf Syst 2013;32:564–79. https://doi.org/10.1016/j.jmsy.2013.04.008.

[78] Zhang L, Luo Y, Tao F, Li B, Ren L, Zhang X, Guo H, Cheng Y, Hu A, Liu Y.

Cloud manufacturing: A new manufacturing paradigm. Enterp Inf Syst 2012;8:1–21. https://doi.org/10.1080/17517575.2012.683812.

[79] Ren L, Zhang L, Tao F, Zhao C, Chai X, Zhao X. Cloud manufacturing: From concept to practice. Enterp Inf Syst 2015;9:186–209. https://doi.org/10.1080/17517575.2013.839055.

[80] Huang B, Li C, Yin C, Zhao X. Cloud manufacturing service platform for small-and medium-sized enterprises. Int J Adv Manuf Technol 2013;65:1261–72. https://doi.org/10.1007/s00170-012-4255-4.

[81] Wu D, Thames JL, Rosen DW, Schaefer D. Towards a cloud-based design and manufacturing paradigm: looking backward, looking forward 2013:315–28. https://doi.org/10.1115/detc2012-70780.

[82] Ren L, Cui J, Wei Y, LaiLi Y, Zhang L. Research on the impact of service provider cooperative relationship on cloud manufacturing platform. Int J Adv Manuf Technol 2016;86:2279–90. https://doi.org/10.1007/s00170-016-8345-6.

[83] Wu D, Rosen DW, Wang L, Schaefera D. Cloud-based manufacturing: Old wine in new bottles? Procedia CIRP 2014;17:94–9. https://doi.org/10.1016/j.procir.2014.01.035.

[84] He W, Xu L. A state-of-the-art survey of cloud manufacturing. Int J Comput Integr Manuf 2015;28:239–50. https://doi.org/10.1080/0951192X.2013.874595.

[85] Lu Y, Xu X, Xu J. Development of a hybrid manufacturing cloud. J Manuf Syst 2014;33:551–66. https://doi.org/10.1016/j.jmsy.2014.05.003.

[86] Schichl H. Models and the history of modeling. In: Josef K, editor. Model. Lang. Math. Optim. 1st ed., Boston: Springer US; 2004, p. 25–36. https://doi.org/10.1007/978-1-4613-0215-5.

[87] Cross M, Moscardini AO. Learning the art of mathematical modelling. Chichester, NY, USA: John Wiley & Sons, Inc.; 1985.

[88] Žák S. A Turing machine time hierarchy. Theor Comput Sci 1983;26:327–33. https://doi.org/10.1016/0304-3975(83)90015-4.

[89] Hangos KM, Cameron IT, editors. 1 - The role of models in process systems engineering. Process Model. Model Anal., vol. 4, Academic Press; 2001, p. 3–18. https://doi.org/https://doi.org/10.1016/S1874-5970(01)80025-4.

[90] Perry RH, Green DW. Perry's chemical engineers' handbook. New York: McGraw-Hill; 2008.

[91] Zhang S, Jiang G, Guo H, Shi C, Xia H. Apply of synergy drilling fluid system in WeiBei oilfield. Drill Fluid Complet Fluid 2013;30:20–2.

[92] Levenspiel O. Modeling in chemical engineering. Chem Eng Sci 2002;57:4691–6. https://doi.org/10.1016/S0009-2509(02)00280-4.

[93] Baghirov KM, Abbasova NM, Valiyeva FM, Aghamaliyev ZZ. Manufacturing capability of mathermatical modeling of chemical processes. Process Petrochemistry Oil-Refining 2016;17:144–8.

[94] Rasmuson A, Andersson B, Olsson L, Andersson R, Olisson L, Andersson R. Empirical model building. Math. Model. Chem. Eng. First, Cambridge: Cambridge Univerisity Press; 2014, p. 40–52. https://doi.org/10.1017/CBO9781107279124.005.

[95] Mathews P. Design of xxperiments language and concepts. Des. Exp. With MINITAB, Milwaukee: ASQ Quality Press; 2004, p. 93–142.

[96] Solomatine D, See LM, Abrahart RJ. Data-driven modelling: Concepts, approaches and experiences. In: Abrahart RJ, See LM, Solomatine DP, editors. Pract. Hydroinformatics. Water Sci. Technol. Libr., Berlin: Springer; 2008, p. 17–31. https://doi.org/https://doi.org/10.1007/978-3-540-79881-1.

[97] von Stosch M, Oliveira R, Peres J, Feyo de Azevedo S. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Comput Chem Eng 2014. https://doi.org/10.1016/j.compchemeng.2013.08.008.

[98] Cheung H, Wang S. A comparison of the effect of empirical and physical modeling approaches to extrapolation capability of compressor models by uncertainty analysis: A case study with common semi-empirical compressor mass flow rate models. Int J Refrig 2018;86:331–43. https://doi.org/10.1016/J.IJREFRIG.2017.11.020.

[99] Saka MP, Doğan E, Aydogdu I. Analysis of Swarm Intelligence–Based Algorithms for Constrained Optimization. Swarm Intell Bio-Inspired Comput 2013:25–48. https://doi.org/10.1016/B978-0-12-405163-8.00002-8.

[100] Kim S. MATLAB Deep learning with machine learning, neural networks and artificial intelligence. 1st ed. Apress; 2017. https://doi.org/10.1007/978-1-4842-

2845-6.

[101] Coley CW, Green WH, Jensen KF. Machine learning in computer-aided synthesis planning. Acc Chem Res 2018;51:1281–9. https://doi.org/10.1021/acs.accounts.8b00087.

[102] Reshmidevi TV, Eldho TI, Jana R. A GIS-integrated fuzzy rule-based inference system for land suitability evaluation in agricultural watersheds. Agric Syst 2009;101:101–9. https://doi.org/10.1016/J.AGSY.2009.04.001.

[103] Angria S L, Sari YD, Zarlis M, Tulus. Data-driven modelling for decision making under uncertainty. IOP Conf. Ser. Mater. Sci. Eng., vol. 300, 2018, p. 12013.

[104] Ojha VK, Abraham A, Snášel V. Metaheuristic design of feedforward neural networks: A review of two decades of research. Eng Appl Artif Intell 2017;60:97–116. https://doi.org/10.1016/j.engappai.2017.01.013.

[105] Ge Z. Review on data-driven modeling and monitoring for plant-wide industrial processes. Chemom Intell Lab Syst 2017;171:16–25. https://doi.org/10.1016/J.CHEMOLAB.2017.09.021.

[106] Yin S, Kaynak O. Big data for modern industry: challenges and trends. Proc IEEE 2015;103:143–6. https://doi.org/10.1109/JPROC.2015.2388958.

[107] Boukouvala F, Li J, Xiao X, Floudas CA. Data-driven modeling and global optimization of industrial-scale petrochemical planning operations. 2016 Am. Control Conf., 2016, p. 3340–5. https://doi.org/10.1109/ACC.2016.7525433.

[108] Escrig J, Woolley E, Rangappa S, Simeone A, Watson NJ. Clean-in-place monitoring of different food fouling materials using ultrasonic measurements. Food Control 2019;104:358–66. https://doi.org/10.1016/J.FOODCONT.2019.05.013.

[109] Hemmati-Sarapardeh A, Ghazanfari M-H, Ayatollahi S, Masihi M. Accurate determination of the CO2-crude oil minimum miscibility pressure of pure and impure CO2 streams: A robust modelling approach. Can J Chem Eng 2016;94:253–61. https://doi.org/10.1002/cjce.22387.

[110] Schweidtmann AM, Clayton AD, Holmes N, Bradford E, Bourne RA, Lapkin AA. Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. Chem Eng J

2018;352:277–82. https://doi.org/10.1016/J.CEJ.2018.07.031.

[111] Pilkington JL, Preston C, Gomes RL. Comparison of response surface methodology (RSM) and artificial neural networks (ANN) towards efficient extraction of artemisinin from Artemisia annua. Ind Crops Prod 2014;58:15–24. https://doi.org/10.1016/J.INDCROP.2014.03.016.

[112] Slišković D, Grbić R, Hocenski Ž. Methods for Plant Data-Based Process Modeling in Soft-Sensor Development. Automatika 2011;52:306–18. https://doi.org/10.1080/00051144.2011.11828430.

[113] Abbas MH, Norman R, Charles A. Neural network modelling of high pressure CO 2 corrosion in pipeline steels. Process Saf Environ Prot 2018;119:36–45. https://doi.org/10.1016/j.psep.2018.07.006.

[114] Güçlü D, Dursun Ş. Artificial neural network modelling of a large-scale wastewater treatment plant operation. Bioprocess Biosyst Eng 2010;33:1051–8. https://doi.org/10.1007/s00449-010-0430-x.

[115] Curteanu S, Piuleac CG, Godini K, Azaryan G. Modeling of electrolysis process in wastewater treatment using different types of neural networks. Chem Eng J 2011. https://doi.org/10.1016/j.cej.2011.05.104.

[116] Charte F, Romero I, Pï¿½rez-Godoy MD, Rivera AJ, Castro E. Comparative analysis of data mining and response surface methodology predictive models for enzymatic hydrolysis of pretreated olive tree biomass. Comput Chem Eng 2017. https://doi.org/10.1016/j.compchemeng.2017.02.008.

[117] Kashani MN, Shahhosseini S. A methodology for modeling batch reactors using generalized dynamic neural networks. Chem Eng J 2010. https://doi.org/10.1016/j.cej.2010.02.053.

[118] Panerati J, Schnellmann MA, Patience C, Beltrame G, Patience GS. Experimental methods in chemical engineering: Artificial neural networks–ANNs. Can J Chem Eng 2019;97:2372–82. https://doi.org/10.1002/cjce.23507.

[119] Hamid HA, Jenidi Y, Somerfield C, Gomes RL. Predicting the capability of carboxylated cellulose nanowhiskers for the remediation of copper from water using response surface methodology (RSM) and artificial neural network (ANN) models. Ind Crops Prod 2016;93:108–20. https://doi.org/10.1016/J.INDCROP.2016.05.035.

[120] Stock T, Seliger G. Opportunities of sustainable manufacturing in Industry 4.0. Procedia CIRP 2016;40:536–41. https://doi.org/10.1016/j.procir.2016.01.129.

[121] Ling J, Jones R, Templeton J. Machine learning strategies for systems with invariance properties. J Comput Phys 2016;318:22–35. https://doi.org/10.1016/J.JCP.2016.05.003.

[122] Baghban A, Kahani M, Nazari MA, Ahmadi MH, Yan W-M. Sensitivity analysis and application of machine learning methods to predict the heat transfer performance of CNT/water nanofluid flows through coils. Int J Heat Mass Transf 2019;128:825–35. https://doi.org/10.1016/J.IJHEATMASSTRANSFER.2018.09.041.

[123] Kolouri S, Park SR, Thorpe M, Slepcev D, Rohde GK. Optimal mass transport: Signal processing and machine-learning applications. IEEE Signal Process Mag 2017;34:43–59. https://doi.org/10.1109/MSP.2017.2695801.

[124] Ulissi ZW, Medford AJ, Bligaard T, Nørskov JK. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. Nat Commun 2017;8:14621. https://doi.org/10.1038/ncomms14621.

[125] Larson Lesnik K, Liu H. Predicting microbial fuel cell biofilm communities and bioreactor performance using artificial neural networks 2017. https://doi.org/10.1021/acs.est.7b01413.

[126] Melki G, Cano A, Kecman V, Ventura S. Multi-target support vector regression via correlation regressor chains. Inf Sci (Ny) 2017;415–416:53–69. https://doi.org/10.1016/J.INS.2017.06.017.

[127] Borchani H, Varando G, Bielza C, Larrañaga P. A survey on multi-output regression. Wiley Interdiscip Rev Data Min Knowl Discov 2015;5:216–33. https://doi.org/10.1002/widm.1157.

[128] Liu Q, Xu Q, Zheng VW, Xue H, Cao Z, Yang Q. Multi-task learning for cross-platform siRNA efficacy prediction: an in-silico study. BMC Bioinformatics 2010;11:181. https://doi.org/10.1186/1471-2105-11-181.

[129] Burnham AJ, MacGregor JF, Viveros R. Latent variable multivariate regression modeling. Chemom Intell Lab Syst 1999;48:167–80. https://doi.org/10.1016/S0169-7439(99)00018-0.

[130] Kocev D, Džeroski S, White MD, Newell GR, Griffioen P. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecol Modell 2009;220:1159–68. https://doi.org/10.1016/J.ECOLMODEL.2009.01.037.

[131] Kocev D, Ceci M. Ensembles of xxtremely randomized trees for multi-target regression, Springer, Cham; 2015, p. 86–100. https://doi.org/10.1007/978-3-319-24282-8_9.

[132] Jeong M, Lee GG. Multi-domain spoken language understanding with transfer learning. Speech Commun 2009;51:412–24. https://doi.org/10.1016/j.specom.2009.01.001.

[133] Xiong T, Bao Y, Hu Z. Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting. Knowledge-Based Syst 2014;55:87–100. https://doi.org/10.1016/J.KNOSYS.2013.10.012.

[134] Liu Y, Pan Y, Huang D, Wang Q. Fault prognosis of filamentous sludge bulking using an enhanced multi-output gaussian processes regression. Control Eng Pract 2017;62:46–54. https://doi.org/10.1016/J.CONENGPRAC.2017.02.003.

[135] Wang Z, Rangaiah GP. Application and analysis of methods for selecting an optimal solution from the Pareto-Optimal Front obtained by multiobjective optimization. Ind Eng Chem Res 2017;56:560–74. https://doi.org/10.1021/acs.iecr.6b03453.

[136] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science (80-) 2015;349:255 LP – 260. https://doi.org/10.1126/science.aaa8415.

[137] Bishop CM. Pattern recognition and machine learning (information science and statistics). Berlin, Heidelberg: Springer-Verlag; 2006.

[138] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerg Artif Intell Appl Comput Eng 2007;160:3–24.

[139] Kulkarni A, Jayaraman VK, Kulkarni BD. Knowledge incorporated support vector machines to detect faults in Tennessee Eastman Process. Comput Chem Eng 2005;29:2128–33. https://doi.org/10.1016/J.COMPCHEMENG.2005.06.006.

[140] Ge Z, Song Z, Ding SX, Huang B. Data mining and analytics in the process

industry: The role of machine learning. IEEE Access 2017;5:20590–616. https://doi.org/10.1109/ACCESS.2017.2756872.

[141] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.

[142] Gómez D, Rojas A. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. Neural Comput 2015;28:216–28. https://doi.org/10.1162/NECO_a_00793.

[143] Wang J, Ma Y, Zhang L, Gao RX. Deep learning for smart manufacturing: Methods and applications. J Manuf Syst 2018;48:144–56. https://doi.org/10.1016/J.JMSY.2018.01.003.

[144] Wang Z, Srinivasan RS. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renew Sustain Energy Rev 2017;75:796–808. https://doi.org/10.1016/j.rser.2016.10.079.

[145] Gong J, You F. Resilient design and operations of process systems: Nonlinear adaptive robust optimization model and algorithm for resilience analysis and enhancement. Comput Chem Eng 2017. https://doi.org/10.1016/j.compchemeng.2017.11.002.

[146] Snoek J, Rippel O, Swersky K, Kiros R, Satish N, Sundaram N, Patwary MMA, Prabhat P, Adams RP. Scalable bayesian optimization using deep neural networks. Proc. 32Nd Int. Conf. Int. Conf. Mach. Learn. - Vol. 37, JMLR.org; 2015, p. 2171–80.

[147] Wang C, Liu Y, Zheng S, Jiang A. Optimizing combustion of coal fired boilers for reducing NOx emission using Gaussian Process. Energy 2018;153:149–58. https://doi.org/10.1016/J.ENERGY.2018.01.003.

[148] Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. J Math Psychol 2018;85:1–16. https://doi.org/10.1016/J.JMP.2018.03.001.

[149] Yildiz B, Bilbao JI, Sproul AB. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renew Sustain Energy Rev 2017;73:1104–22. https://doi.org/10.1016/j.rser.2017.02.023.

[150] Ahmad MW, Mourshed M, Rezgui Y. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy Build 2017;147:77–89. https://doi.org/10.1016/J.ENBUILD.2017.04.038.

[151] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests. Pattern Recognit 2011;44:330–49. https://doi.org/10.1016/J.PATCOG.2010.08.011.

[152] Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO. Random forest models to predict aqueous solubility. J Chem Inf Model 2006;47:150–8. https://doi.org/10.1021/CI060164K.

[153] Cuentas S, Peñabaena-Niebles R, Garcia E. Support vector machine in statistical process monitoring: a methodological and analytical review. Int J Adv Manuf Technol 2017;91:485–500. https://doi.org/10.1007/s00170-016-9693-y.

[154] Lahiri SK, Ghanta KC. The Support Vector Regression with the parameter tuning assisted by a differential evolution technique: Study of the critical velocity of a slurry flow in a pipeline. Chem Ind Chem Eng Q 2008;14:191–203. https://doi.org/10.2298/CICEQ0803191L.

[155] Mesbah M, Soroush E, Rezakazemi M. Development of a least squares support vector machine model for prediction of natural gas hydrate formation temperature. Chinese J Chem Eng 2017;25:1238–48. https://doi.org/10.1016/J.CJCHE.2016.09.007.

[156] Peng H, Ling X. Predicting thermal–hydraulic performances in compact heat exchangers by support vector regression. Int J Heat Mass Transf 2015;84:203–13. https://doi.org/10.1016/J.IJHEATMASSTRANSFER.2015.01.017.

[157] Rabbani Y, Shirvani M, Hashemabadi SH, Keshavarz M. Application of artificial neural networks and support vector regression modeling in prediction of magnetorheological fluid rheometery. Colloids Surfaces A Physicochem Eng Asp 2017;520:268–78. https://doi.org/10.1016/J.COLSURFA.2017.01.081.

[158] Basheer I., Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods 2000;43:3–31. https://doi.org/10.1016/S0167-7012(00)00201-3.

[159] Sanger TD. Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Networks 1989;2:459–73. https://doi.org/10.1016/0893-

6080(89)90044-0.

[160] Chiang Y-M, Chang L-C, Chang F-J. Comparison of static-feedforward and dynamic-feedback neural networks for rainfall–runoff modeling. J Hydrol 2004;290:297–311. https://doi.org/10.1016/J.JHYDROL.2003.12.033.

[161] Zhao Y, Liu Q, Xu W, Wu X, Jiang X, Zhou Z, Pham DT. Dynamic and unified modelling of sustainable manufacturing capability for industrial robots in cloud manufacturing. Int J Adv Manuf Technol 2017;93:2753–71. https://doi.org/10.1007/s00170-017-0634-1.

[162] Lartigau J, Xu X, Nie L, Zhan D. Cloud manufacturing service composition based on QoS with geo-perspective transportation using an improved Artificial Bee Colony optimisation algorithm. Int J Prod Res 2015;53:4380–404. https://doi.org/10.1080/00207543.2015.1005765.

[163] Qu T, Lei SP, Wang ZZ, Nie DX, Chen X, Huang GQ. IoT-based real-time production logistics synchronization system under smart cloud manufacturing. Int J Adv Manuf Technol 2016;84:147–64. https://doi.org/10.1007/s00170-015-7220-1.

[164] MindSphere – Siemens Cloud for Industry - Industry Services - Siemens n.d. https://w3.siemens.co.uk/services/uk/en/portfolio/plant-data-services/cloud-for-industry/pages/default.aspx#; (accessed July 19, 2019).

[165] International Business Machines. IBM Watson Internet of Things (IoT) 2016. https://www.ibm.com/internet-of-things (accessed January 10, 2019).

[166] World Economic Forum. The Global Risks Report 2018; 13th Edition. Geneva: 2018.

[167] Potter MC. Electrical effects accompanying the decomposition of organic compounds. Proc. R. Soc. B Biol. Sci., vol. 84, The Royal Society London; 1911, p. 260–76. https://doi.org/10.1098/rspb.1911.0073.

[168] Cohen B. The bacterial culture as an electrical half-cell. J Bacteriol 1931;21:18–9.

[169] Del Duca MG, Fuscoe JM. Thermodynamics and applications of bioelectrochemical energy conversion systems. 6th Agard Colloq. Energy Sources Energy Convers., Cannes: 1967, p. 811–39.

[170] Schröder U. Anodic electron transfer mechanisms in microbial fuel cells and

their energy efficiency. Phys Chem Chem Phys 2007;9:2619–29. https://doi.org/10.1039/b703627m.

[171] Kim HJ, Park HS, Hyun MS, Chang IS, Kim M, Kim BH. A mediator-less microbial fuel cell using a metal reducing bacterium, Shewanella putrefaciens. Enzyme Microb Technol 2002;30:145–52. https://doi.org/10.1016/S0141-0229(01)00478-1.

[172] Rozendal RA, Hamelers HVM, Rabaey K, Keller J, Buisman CJN. Towards practical implementation of bioelectrochemical wastewater treatment. Trends Biotechnol 2008;26:450–9. https://doi.org/10.1016/j.tibtech.2008.04.008.

[173] Logan BE. Scaling up microbial fuel cells and other bioelectrochemical systems. Appl Microbiol Biotechnol 2010;85:1665–71. https://doi.org/10.1007/s00253-009-2378-9.

[174] European Commision. Future brief: Bioelectrochemical systems. Wastewater treatment, bioenergy and valuable chemicals delivered by bacteria. 2013.

[175] Palmore GTR, Whitesides GM. Microbial and enzymatic biofuel cells. Enzym. Convers. Biomass Fuels Prod., Washington: ACS Books, American Chemical Society; 1994, p. 271–90.

[176] Rabaey K, Rodríguez J, Blackall LL, Keller J, Gross P, Batstone D, Verstraete W, Nealson KH. Microbial ecology meets electrochemistry: Electricity-driven and driving communities. ISME J 2007;1:9–18. https://doi.org/10.1038/ismej.2007.4.

[177] Rabaey K, Angenent L, Schroder U, Keller J. Bioelectrochemical systems : from extracellular electron transfer to biotechnological application. London: IWA Publishing; 2010.

[178] Rabaey K. Bioelectrochemical systems: From extracellular electron transfer to biotechnological application. Water Intell Online 2009;8. https://doi.org/10.2166/9781780401621.

[179] Szpyrkowicz L, Kaul SN, Neti RN, Satyanarayan S. Influence of anode material on electrochemical oxidation for the treatment of tannery wastewater. Water Res 2005;39:1601–13. https://doi.org/10.1016/J.WATRES.2005.01.016.

[180] Freguia S, Virdis B, Harnisch F, Keller J. Bioelectrochemical systems: Microbial versus enzymatic catalysis. Electrochim Acta 2012;82:165–74.

https://doi.org/10.1016/J.ELECTACTA.2012.03.014.

[181] Logan BE, Hamelers B, Rozendal R, Schröder U, Keller J, Freguia S, Aelterman P, Verstraete W, Rabaey K. Microbial fuel cells: methodology and technology. Environ Sci Technol 2006;40:5181–92. https://doi.org/10.1021/es0605016.

[182] Madiraju KS, Lyew D, Kok R, Raghavan V. Carbon neutral electricity production by Synechocystis sp. PCC6803 in a microbial fuel cell. Bioresour Technol 2012;110:214–8. https://doi.org/10.1016/j.biortech.2012.01.065.

[183] Luo H, Xu P, Roane TM, Jenkins PE, Ren Z. Microbial desalination cells for improved performance in wastewater treatment, electricity production, and desalination. Bioresour Technol 2012;105:60–6. https://doi.org/10.1016/J.BIORTECH.2011.11.098.

[184] Bajracharya S, Sharma M, Mohanakrishna G, Dominguez X, Strik DPBTBBTB, Sarma PM, Dominguez Benneton X, Strik DPBTBBTB, Sarma PM, Pant D. An overview on emerging bioelectrochemical systems (BESs): Technology for sustainable electricity, waste remediation, resource recovery, chemical production and beyond. Renew Energy 2016;98:153–70. https://doi.org/10.1016/j.renene.2016.03.002.

[185] Feng Y, Harper WF. Biosensing with microbial fuel cells and artificial neural networks: Laboratory and field investigations. J Environ Manage 2013;130:369–74. https://doi.org/10.1016/J.JENVMAN.2013.09.011.

[186] Sleutels THJA, Ter Heijne A, Buisman CJN, Hamelers HVM. Bioelectrochemical systems: An outlook for practical applications. ChemSusChem 2012;5:1012–9. https://doi.org/10.1002/cssc.201100732.

[187] Jeremiasse AW, Hamelers HVM, Buisman CJN. Microbial electrolysis cell with a microbial biocathode. Bioelectrochemistry 2010;78:39–43. https://doi.org/10.1016/J.BIOELECHEM.2009.05.005.

[188] Cao X, Huang X, Liang P, Xiao K, Zhou Y, Zhang X, Logan BE. A new method for water desalination using microbial desalination cells. Environ Sci Technol 2009;43:7148–52. https://doi.org/10.1021/es901950j.

[189] Bajracharya S, Ter Heijne A, Dominguez Benetton X, Vanbroekhoven K, Buisman CJN, Strik DPBTB, Pant D. Carbon dioxide reduction by mixed and pure cultures in microbial electrosynthesis using an assembly of graphite felt and stainless steel as a cathode. Bioresour Technol 2015;195:14–24.

https://doi.org/10.1016/j.biortech.2015.05.081.

[190] Santoro C, Arbizzani C, Erable B, Ieropoulos I. Microbial fuel cells: From fundamentals to applications. A review. J Power Sources 2017;356:225–44. https://doi.org/10.1016/j.jpowsour.2017.03.109.

[191] Gadkari S, Gu S, Sadhukhan J. Towards automated design of bioelectrochemical systems: A comprehensive review of mathematical models. Chem Eng J 2018;343:303–16. https://doi.org/10.1016/j.cej.2018.03.005.

[192] Oliveira VBB, Simões M, Melo LFF, Pinto AMFRMFR. Overview on the developments of microbial fuel cells. Biochem Eng J 2013;73:53–64. https://doi.org/10.1016/j.bej.2013.01.012.

[193] Vilà-Rovira A, Puig S, Balaguer MD, Colprim J. Anode hydrodynamics in bioelectrochemical systems. RSC Adv 2015;5:78994–9000. https://doi.org/10.1039/C5RA11995B.

[194] Beyenal H, Babauta J. Biofilms in bioelectrochemical systems. Hoboken, NJ, USA: John Wiley & Sons, Inc; 2015. https://doi.org/10.1002/9781119097426.

[195] Pham TH, Aelterman P, Verstraete W. Bioanode performance in bioelectrochemical systems: recent improvements and prospects. Trends Biotechnol 2009;27:168–78. https://doi.org/10.1016/J.TIBTECH.2008.11.005.

[196] Harnisch F, Warmbier R, Schneider R, Schröder U. Modeling the ion transfer and polarization of ion exchange membranes in bioelectrochemical systems. Bioelectrochemistry 2009;75:136–41. https://doi.org/10.1016/J.BIOELECHEM.2009.03.001.

[197] Picioreanu C, Katuri KP, van Loosdrecht MCM, Head IM, Scott K. Modelling microbial fuel cells with suspended cells and added electron transfer mediator. J Appl Electrochem 2010;40:151–62. https://doi.org/10.1007/s10800-009-9991-2.

[198] Zhang X-C, Halme A. Modelling of a microbial fuel cell process. Biotechnol Lett 1995;17:809–14. https://doi.org/10.1007/BF00129009.

[199] Esfandyari M, Fanaei MA, Gheshlaghi R, Akhavan Mahdavi M. Mathematical modeling of two-chamber batch microbial fuel cell with pure culture of Shewanella. Chem Eng Res Des 2017;117:34–42. https://doi.org/10.1016/J.CHERD.2016.09.016.

[200] Esfandyari M, Fanaei MA, Gheshlaghi R, Akhavan Mahdavi M. Dynamic modeling of a continuous two-chamber microbial fuel cell with pure culture of Shewanella. Int J Hydrogen Energy 2017;42:21198–202. https://doi.org/10.1016/J.IJHYDENE.2017.07.042.

[201] Tardast A, Rahimnejad M, Najafpour G, Ghoreyshi A, Premier GC, Bakeri G, Oh S-E. Use of artificial neural network for the prediction of bioelectricity production in a membrane less microbial fuel cell. Fuel 2014;117:697–703. https://doi.org/10.1016/J.FUEL.2013.09.047.

[202] Garg A, Vijayaraghavan V, Mahapatra SS, Tai K, Wong CH. Performance evaluation of microbial fuel cell by artificial intelligence methods. Expert Syst Appl 2014;41:1389–99. https://doi.org/10.1016/J.ESWA.2013.08.038.

[203] Esfandyari M, Fanaei MA, Gheshlaghi R, Mahdavi MA. Neural network and neuro-fuzzy modeling to investigate the power density and Columbic efficiency of microbial fuel cell. J Taiwan Inst Chem Eng 2016;58:84–91. https://doi.org/10.1016/J.JTICE.2015.06.005.

[204] King ST, Sylvander M, Kheperu M, Racz L, Harper WF. Detecting recalcitrant organic chemicals in water with microbial fuel cells and artificial neural networks. Sci Total Environ 2014;497–498:527–33. https://doi.org/10.1016/J.SCITOTENV.2014.07.108.

[205] OECD. OECD Environmental Outlook to 2050. OECD Publishing; 2012. https://doi.org/10.1787/9789264122246-en.

[206] European Union. Directive 75/442/EEC of the European Parliament and of the Council of 15July 1975 on waste. Off J Eur Union 1975;L 442:39–41. https://doi.org/doi:10.3000/19770677.L_2013.124.eng.

[207] Lin CSK, Pfaltzgraff LA, Herrero-Davila L, Mubofu EB, Abderrahim S, Clark JH, Koutinas AA, Kopsahelis N, Stamatelatou K, Dickson F, Thankappan S, Mohamed Z, Brocklesby R, Luque R. Food waste as a valuable resource for the production of chemicals, materials and fuels. Current situation and global perspective. Energy Environ Sci 2013;6:426–64. https://doi.org/10.1039/c2ee23440h.

[208] Rizos V, Behrens A, Kafyeke T, Hirschnitz-Garbers M, Ioannou A. The circular economy: barriers and opportunities for SMEs. CEPS Work Doc 2015:1–25. https://doi.org/0006-8993(94)90176-7 [pii] ET - 1994/02/04.

[209] Moore SB, Manring SL. Strategy development in small and medium sized enterprises for sustainability and increased value creation. J Clean Prod 2009;17:276–82. https://doi.org/10.1016/j.jclepro.2008.06.004.

[210] Zhang X, Porcu L, Andresen JM. Sustainable energy from dairy farm waste using a microbial fuel cell (MFC). Proc World Renew Energy Congr – Sweden, 8–13 May, 2011, Linköping, Sweden 2011;57:247–50. https://doi.org/10.3384/ecp11057247.

[211] H2AD - Innovative and scalable biotechnology using Microbial Fuel Cell and Anaerobic Digestion for the treatment of micro-scale industrial and agriculture effluents to recover energy from waste 2015. https://cordis.europa.eu/project/rcn/199494/factsheet/en (accessed July 19, 2019).

[212] Mahon] J [Mac, Gill LW. Sustainability of novel water treatment technologies in developing countries: Lessons learned from research trials on a pilot continuous flow solar water disinfection system in rural Kenya. Dev Eng 2018;3:47–59. https://doi.org/https://doi.org/10.1016/j.deveng.2018.01.003.

[213] WordArt.com - Word Cloud Art Creator n.d. https://wordart.com/ (accessed July 19, 2019).

[214] Wu D, Greer MJ, Rosen DW, Schaefer D. Cloud manufacturing: drivers, current status, and future trends 2013:V002T02A003. https://doi.org/10.1115/msec2013-1106.

[215] Wang P, Gao RX, Fan Z. Cloud computing for cloud manufacturing: benefits and limitations. J Manuf Sci Eng 2015;137:044002. https://doi.org/10.1115/1.4030209.

[216] MFG n.d. https://www.mfg.com/ (accessed July 19, 2019).

[217] Qin SJ. Process data analytics in the era of big data. AIChE J 2014;60:3092–100. https://doi.org/10.1002/aic.14523.

[218] Zhong RY, Lan S, Xu C, Dai Q, Huang GQ. Visualization of RFID-enabled shopfloor logistics Big Data in Cloud Manufacturing. Int J Adv Manuf Technol 2016;84:5–16. https://doi.org/10.1007/s00170-015-7702-1.

[219] Gao J, Yao Y, Zhu VCY, Sun L, Lin L. Service-oriented manufacturing: A new product pattern and manufacturing paradigm. J Intell Manuf 2011;22:435–46.

https://doi.org/10.1007/s10845-009-0301-y.

[220] Griffin PW, Hammond GP, Norman JB. Industrial energy use and carbon emissions reduction: a UK perspective. WIREs Energy Environ 2016;5:684–714. https://doi.org/10.1002/wene.212.

[221] Liew W Te, Adhitya A, Srinivasan R. Sustainability trends in the process industries: A text mining-based analysis. Comput Ind 2014;65:393–400. https://doi.org/10.1016/j.compind.2014.01.004.

[222] Ciccullo F, Pero M, Caridi M, Gosling J, Purvis L. Integrating the environmental and social sustainability pillars into the lean and agile supply chain management paradigms: A literature review and future research directions. J Clean Prod 2018;172:2336–50.
https://doi.org/https://doi.org/10.1016/j.jclepro.2017.11.176.

[223] Wang Z, Subramanian N, Gunasekaran A, Abdulrahman MD, Liu C. Composite sustainable manufacturing practice and performance framework: Chinese auto-parts suppliers′ perspective. Int J Prod Econ 2015;170:219–33. https://doi.org/https://doi.org/10.1016/j.ijpe.2015.09.035.

[224] Basmer S, Buxbaum-Conradi S, Krenz P, Redlich T, Wulfsberg JP, Bruhns F-L. Open production: Chances for social sustainability in manufacturing. Procedia CIRP 2015;26:46–51. https://doi.org/https://doi.org/10.1016/j.procir.2014.07.102.

[225] Hamalainen M, Karjalainen J. Social manufacturing: When the maker movement meets interfirm production networks. Bus Horiz 2017;60:795–805. https://doi.org/10.1016/J.BUSHOR.2017.07.007.

[226] Bi ZM, Wang L. Manufacturing paradigm shift towards better sustainability, Springer, London; 2013, p. 99–119. https://doi.org/10.1007/978-1-4471-4935-4_5.

[227] Monfared MAS, Yang JB. Design of integrated manufacturing planning, scheduling and control systems: A new framework for automation. Int J Adv Manuf Technol 2007;33:545–59. https://doi.org/10.1007/s00170-006-0476-8.

[228] Esmaeilian B, Behdad S, Wang B. The evolution and future of manufacturing: A review. J Manuf Syst 2016;39:79–100. https://doi.org/10.1016/J.JMSY.2016.03.001.

[229] Jassbi J, di Orio G, Barata D, Barata J. The impact of cloud manufacturing on supply chain agility. 2014 12th IEEE Int. Conf. Ind. Informatics, IEEE; 2014, p. 495–500. https://doi.org/10.1109/INDIN.2014.6945563.

[230] Yu C, Zhang W, Xu X, Ji Y, Yu S. Data mining based multi-level aggregate service planning for cloud manufacturing. J Intell Manuf 2018;29:1351–61. https://doi.org/10.1007/s10845-015-1184-8.

[231] Cao Y, Wang S, Kang L, Li C, Guo L. Study on machining service modes and resource selection strategies in cloud manufacturing. Int J Adv Manuf Technol 2015;81:597–613. https://doi.org/10.1007/s00170-015-7222-z.

[232] Martinez Lastra JL, Delamer IM. Ontologies for production automation. Adv. web Semant. I Ontol. web Serv. Appl. wemantic web, Springer, Berlin, Heidelberg; 2008, p. 276–89. https://doi.org/10.1007/978-3-540-89784-2_11.

[233] Kockmann N. Digital methods and tools for chemical equipment and plants. React Chem Eng 2019;4:1522–9. https://doi.org/10.1039/C9RE00017H.

[234] Helo P, Suorsa M, Hao Y, Anussornnitisarn P. Toward a cloud-based manufacturing execution system for distributed manufacturing. Comput Ind 2014;65:646–56. https://doi.org/10.1016/j.compind.2014.01.015.

[235] Saguy IS, Singh RP, Johnson T, Fryer PJ, Sastry SK. Challenges facing food engineering. J Food Eng 2013;119:332–42. https://doi.org/10.1016/J.JFOODENG.2013.05.031.

[236] Ming-Li Yeh, Chuei-Tin Chang. An automaton-based approach to evaluate and improve online diagnosis schemes for multi-failure scenarios in batch chemical processes. Chem Eng Res Des 2011.

[237] Simeone A, Watson N, Sterritt I, Woolley E. A multi-sensor approach for fouling level assessment in clean-in-place processes. Procedia CIRP 2016;55:134–9. https://doi.org/10.1016/J.PROCIR.2016.07.023.

[238] Simeone A, Deng B, Watson N, Woolley E. The circular economy: Swings and roundabouts? Sensors 2018;18. https://doi.org/10.3390/s18113742.

[239] Millar N, Mclaughlin E, Börger T. The Circular Economy : Swings and Roundabouts ? 2019;158:11–9. https://doi.org/10.1016/j.ecolecon.2018.12.012.

[240] Korhonen J, Honkasalo A, Seppälä J. Circular economy: The concept and its limitations. Ecol Econ 2018;143:37–46.

https://doi.org/10.1016/J.ECOLECON.2017.06.041.

[241] McDonough W, Braungart M. Cradle to cradle: remaking the way we make things. 1st. London: MacMillan; 2010.

[242] Sharebox - Secure sharing n.d. sharebox-project.eu/#overview (accessed July 10, 2019).

[243] Lieder M, Rashid A. Towards circular economy implementation: a comprehensive review in context of manufacturing industry. J Clean Prod 2016;115:36–51. https://doi.org/10.1016/J.JCLEPRO.2015.12.042.

[244] Arancon RAD, Lin CSK, Chan KM, Kwan TH, Luque R. Advances on waste valorization: new horizons for a more sustainable society. Energy Sci Eng 2013;1:53–71. https://doi.org/10.1002/ese3.9.

[245] Garcia-Garcia G, Woolley E, Rahimifard S. A framework for a more efficient approach to food waste management. ETP Int J Food Eng 2016;1:65–72. https://doi.org/10.18178/ijfe.1.1.65-72.

[246] British Sugar. How our factory operates n.d. http://www.britishsugarlearningzone.com/how-our-factory-operates (accessed February 16, 2017).

[247] Short SW, Bocken NMP, Barlow CY, Chertow MR. From refining sugar to growing tomatoes. J Ind Ecol 2014;18:603–18. https://doi.org/10.1111/jiec.12171.

[248] Ferrasse J-H, Seyssiecq I, Roche N. Thermal gasification: A feasible solution for sewage sludge valorisation? Chem Eng Technol 2003;26:941–5. https://doi.org/10.1002/ceat.200301813.

[249] Valta K, Aggeli E, Papadaskalopoulou C, Panaretou V, Sotiropoulos A, Malamis D, Moustakas K, Haralambous K-J. Adding value to olive oil production through waste and wastewater treatment and valorisation: The case of greece. Waste Biomass Valori 2015;6:913–25. https://doi.org/10.1007/s12649-015-9373-4.

[250] Tseng ML, Tan RR, Chiu ASF, Chien CF, Kuo TC. Circular economy meets industry 4.0: Can big data drive industrial symbiosis? Resour Conserv Recycl 2018;131:146–7. https://doi.org/10.1016/j.resconrec.2017.12.028.

[251] Roel J. M. Westerhof, Norbert J. M. Kuipers, Sascha R. A. Kersten, Swaaij

WPM van. Controlling the water content of biomass fast pyrolysis oil. Ind Eng Chem Res 2007. https://doi.org/10.1021/IE070684K.

[252] Funke A, Ziegler F. Hydrothermal carbonization of biomass: A summary and discussion of chemical mechanisms for process engineering. Biofuels, Bioprod Biorefining 2010;4:160–77. https://doi.org/10.1002/bbb.198.

[253] Agency FS. Food law prosecutions 2017. www.food.gov.uk/enforcement/prosecutions (accessed April 10, 2017).

[254] Towler GP, Sinnott RK. Chapter 1 - Introduction to design. Chem. Eng. Des. Princ. Pract. Econ. Plant Process Des., Butterworth-Heinemann; 2013, p. 3–32.

[255] Kenney KL, Smith WA, Gresham GL, Westover TL. Understanding biomass feedstock variability. Biofuels 2013;4:111–27. https://doi.org/10.4155/bfs.12.83.

[256] Waller MA, Fawcett SE. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. J Bus Logist 2013;34:77–84. https://doi.org/10.1111/jbl.12010.

[257] Grossmann I. Enterprise-wide optimization: A new frontier in process systems engineering. AIChE J 2005;51:1846–57. https://doi.org/10.1002/aic.10617.

[258] Bao Y, Ren L, Zhang L, Zhang X, Luo Y. Massive sensor data management framework in cloud manufacturing based on Hadoop. IEEE 10th Int. Conf. Ind. Informatics, IEEE; 2012, p. 397–401. https://doi.org/10.1109/INDIN.2012.6301192.

[259] Houben C, Lapkin AA. Automatic discovery and optimization of chemical processes. Curr Opin Chem Eng 2015;9:1–7. https://doi.org/10.1016/j.coche.2015.07.001.

[260] Chue KT, Kim JN, Yoo YJ, Cho SH, Yang RT. Comparison of activated carbon and Zeolite 13X for CO2 recovery from flue gas by pressure swing adsorption. Ind Eng Chem Res 1995;34:591–8. https://doi.org/10.1021/ie00041a020.

[261] Lopez-Juarez I, Rios-Cabrera R, Hsieh SJ, Howarth M. A hybrid non-invasive method for internal/external quality assessment of potatoes. Eur Food Res Technol 2018;244:161–74. https://doi.org/10.1007/s00217-017-2936-9.

[262] Xiros C, Christakopoulos P. Biotechnological potential of brewers spent grain and its recent applications. Waste Biomass Valori 2012;3:213–32.

https://doi.org/10.1007/s12649-012-9108-8.

[263] Gonzalez Viejo C, Fuentes S, Li G, Collmann R, Condé B, Torrico D. Development of a robotic pourer constructed with ubiquitous materials, open hardware and sensors to assess beer foam quality using computer vision and pattern recognition algorithms: RoboBEER. Food Res Int 2016;89:504–13. https://doi.org/10.1016/j.foodres.2016.08.045.

[264] Ben-Hamed U, Seddighi H, Thomas K. Economic returns of using brewery's spent grain in animal feed. Int J Econ Manag Eng 2011;5:142–5.

[265] Smithers GW. Whey and whey proteins—From 'gutter-to-gold.' Int Dairy J 2008;18:695–704. https://doi.org/10.1016/J.IDAIRYJ.2008.03.008.

[266] Tao F, Zuo Y, Xu L Da, Zhang L. IoT-Based intelligent perception and access of manufacturing resource toward cloud manufacturing. IEEE Trans Ind Informatics 2014;10:1547–57. https://doi.org/10.1109/TII.2014.2306397.

[267] Mannan MS, Reyes-Valdes O, Jain P, Tamim N, Ahammad M. The evolution of process safety: current status and future direction. Annu Rev Chem Biomol Eng 2016;7:135–62. https://doi.org/10.1146/annurev-chembioeng-080615-033640.

[268] Liu C-C, Stefanov A, Hong J, Panciatici P. Intruders in the grid. IEEE Power Energy Mag 2012;10:58–66. https://doi.org/10.1109/MPE.2011.943114.

[269] A A, H L, P A, A R. Statistics on small and medium-sized enterprises n.d. http://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_small_and_medium-sized_enterprises (accessed January 5, 2018).

[270] Department for Enviroment Food & Rural Affairs. Food statistics pocketbook 2016. London: 2016.

[271] Ellis V, Bosworth G. Supporting rural entrepreneurship in the UK microbrewery sector. Br Food J 2015;117:2724–38. https://doi.org/10.1108/BFJ-12-2014-0412.

[272] J. M. Made smarter - Review 2017. London: 2017.

[273] Simmonds J. Number of breweries in the UK breaks through the 2,000 barrier for the first time since the 1930s 2017. https://www.uhy-uk.com/news-events/news/number-of-breweries-in-the-uk-breaks-through-the-2000-barrier-

for-the-first-time-since-the-1930s/ (accessed December 19, 2018).

[274] Edmonds D. Before the taps run dry: Incentivizing water sustainability in America's craft breweries. Georg Washingt J Energy Environ Law 2016;7:164–76.

[275] Al-Aufi YA, Hewakandamby BN, Dimitrakis G, Holmes M, Hasan A, Watson NJ. Thin film thickness measurements in two phase annular flows using ultrasonic pulse echo techniques. Flow Meas Instrum 2019;66:67–78. https://doi.org/10.1016/J.FLOWMEASINST.2019.02.008.

[276] Resa P, Elvira L, de Espinosa F. Concentration control in alcoholic fermentation processes from ultrasonic velocity measurements. Food Res Int 2004;37:587–94. https://doi.org/10.1016/j.foodres.2003.12.012.

[277] Krause D, Schöck T, Hussein MA, Becker T. Ultrasonic characterization of aqueous solutions with varying sugar and ethanol content using multivariate regression methods. J Chemom 2011;25:216–23. https://doi.org/10.1002/cem.1384.

[278] Microsoft. Azure Machine Learning Studio 2018. https://azure.microsoft.com/en-gb/services/machine-learning-studio/ (accessed October 24, 2018).

[279] Pasini A. Artificial neural networks for small dataset analysis. J Thorac Dis 2015;7:953–60. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61.

[280] Almeida JS. Predictive non-linear modeling of complex data by artificial neural networks. Curr Opin Biotechnol 2002;13:72–6. https://doi.org/10.1016/S0958-1669(02)00288-4.

[281] Garment V. 3 Ways to Test the Accuracy of Your Predictive Models 2014. https://www.kdnuggets.com/2014/02/3-ways-to-test-accuracy-your-predictive-models.html (accessed November 15, 2018).

[282] Duarte B, Saraiva P. M, C. PC. Combined mechanistic and empirical modelling. Int J Chem React Eng 2004;2:1–19. https://doi.org/10.2202/1542-6580.1128.

[283] European Union. Regulation (EU) No 1169/2011 on the provision of food information to consumers. Off J Eur Union 2011:18–63. https://doi.org/10.1109/60.911397.

[284] Wang R, Chen B, Qiu S, Zhu Z, Wang Y, Wang Y, Qiu X. Comparison of

machine learning models for hazardous gas dispersion prediction in field cases. Int J Environ Res Public Health 2018;15. https://doi.org/10.3390/ijerph15071450.

[285] Qiao G, Riddick F, McLean C. New manufacturing modeling methodology: Data driven design and simulation system based on XML. Proc. 35th Conf. Winter Simul. Driv. Innov., Winter Simulation Conference; 2003, p. 1143–8.

[286] Liu H, Cocea M. Semi-random partitioning of data into training and test sets in granular computing context. Granul Comput 2017;2:357–86. https://doi.org/10.1007/s41066-017-0049-2.

[287] Ng H-W, Nguyen VD, Vonikakis V, Winkler S. Deep learning for emotion recognition on small datasets using transfer learning. Proc. 2015 ACM Int. Conf. Multimodal Interact. - ICMI '15, New York, New York, USA: ACM Press; 2015, p. 443–9. https://doi.org/10.1145/2818346.2830593.

[288] Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl 2004;6:20–9. https://doi.org/10.1145/1007730.1007735.

[289] Skoogh A, Perera T, Johansson B. Input data management in simulation – Industrial practices and future trends. Simul Model Pract Theory 2012;29:181–92. https://doi.org/10.1016/J.SIMPAT.2012.07.009.

[290] Addo-Tenkorang R, Helo PT. Big data applications in operations/supply-chain management: A literature review. Comput Ind Eng 2016;101:528–43. https://doi.org/10.1016/J.CIE.2016.09.023.

[291] Dijcks JP. Oracle: Big Data for the Enterprise. 2013.

[292] Blazquez D, Domenech J. Big Data sources and methods for social and economic analyses. Technol Forecast Soc Change 2018;130:99–113. https://doi.org/10.1016/J.TECHFORE.2017.07.027.

[293] Kaisler S, Armour F, Espinosa JA, Money W. Big data: Issues and challenges moving forward. 2013 46th Hawaii Int. Conf. Syst. Sci., 2013, p. 995–1004. https://doi.org/10.1109/HICSS.2013.645.

[294] Souza FAA, Araújo R, Mendes J. Review of soft sensor methods for regression applications. Chemom Intell Lab Syst 2016;152:69–79. https://doi.org/10.1016/j.chemolab.2015.12.011.

[295] Shang C, Yang F, Huang D, Lyu W. Data-driven soft sensor development based on deep learning technique. J Process Control 2014;24:223–33. https://doi.org/10.1016/j.jprocont.2014.01.012.

[296] Kay JW, Titterington DM, Kay SLSJW. Statistics and neural networks: advances at the interface. Oxford University Press; 1999.

[297] Williams CL, Westover TL, Emerson RM, Tumuluru JS, Li C. Sources of biomass feedstock variability and the potential impact on biofuels production. BioEnergy Res 2016;9:1–14. https://doi.org/10.1007/s12155-015-9694-y.

[298] Simate GS. Water treatment and reuse in breweries. Brew Microbiol 2015:425–56. https://doi.org/10.1016/B978-1-78242-331-7.00020-4.

[299] European Union. Directive 91/676/CEE concerning the protection of waters against pollution caused by nitrates from agricultural sources. Off J Eur Communities 1991:1–8. https://doi.org/10.1017/CBO9781107415324.004.

[300] Fisgativa H, Tremier A, Dabert P. Characterizing the variability of food waste quality: A need for efficient valorisation through anaerobic digestion. Waste Manag 2016;50:264–74. https://doi.org/10.1016/J.WASMAN.2016.01.041.

[301] Zhang C, Su H, Baeyens J, Tan T. Reviewing the anaerobic digestion of food waste for biogas production. Renew Sustain Energy Rev 2014;38:383–92. https://doi.org/10.1016/J.RSER.2014.05.038.

[302] Templeton DW, Sluiter AD, Hayward TK, Hames BR, Thomas SR. Assessing corn stover composition and sources of variability via NIRS. Cellulose 2009;16:621–39. https://doi.org/10.1007/s10570-009-9325-x.

[303] Ncube B, Finnie JF, Van Staden J. Quality from the field: The impact of environmental factors as quality determinants in medicinal plants. South African J Bot 2012;82:11–20. https://doi.org/10.1016/J.SAJB.2012.05.009.

[304] García-Delgado M, Rodríguez-Cruz MS, Lorenzo LF, Arienzo M, Sánchez-Martín MJ. Seasonal and time variability of heavy metal content and of its chemical forms in sewage sludges from different wastewater treatment plants. Sci Total Environ 2007;382:82–92. https://doi.org/10.1016/J.SCITOTENV.2007.04.009.

[305] Eltawil MA, Samuel DVK, Singhal OP. Potato storage technology and store design aspects. Agric Eng Int CIGR J 2006;8:1–18.

https://doi.org/10.1093/hmg/ddw317.

[306] Lee H-Y, Ong H-L. Visualization support for data mining. IEEE Expert Syst Their Appl 1996;11:69–75. https://doi.org/10.1109/64.539019.

[307] Wang K, Salhi A, Fraga ES. Process design optimisation using embedded hybrid visualisation and data analysis techniques within a genetic algorithm optimisation framework. Chem Eng Process Process Intensif 2004;43:657–69. https://doi.org/10.1016/J.CEP.2003.01.001.

[308] Carr DB, Littlefield RJ, Nicholson WL, Littlefield JS. Scatterplot matrix techniques for large N. J Am Stat Assoc 1987;82:424–36. https://doi.org/10.1080/01621459.1987.10478445.

[309] Wang K, Salhi A, Fraga E. Cluster identification using a Parallel Coordinate System for knowledge discovery and nonlinear optimization. 12th Eur. Symp. Comput. Aided Process Eng., 2002.

[310] Kadlec P, Gabrys B. Soft sensors: Where are we and what are the current and future challenges? vol. 2. IFAC; 2009. https://doi.org/10.3182/20090921-3-TR-3005.00098.

[311] Keviczky L, Banyasz C. 10. Process identification. Two-Degree-of-Freedom Control Syst. - Youla Parameterization Approach, Elsevier; 2015, p. 309–15.

[312] Wang J, Chang Q, Xiao G, Wang N, Li S. Data driven production modeling and simulation of complex automobile general assembly plant. Comput Ind 2011;62:765–75. https://doi.org/10.1016/J.COMPIND.2011.05.004.

[313] Krenczyk D. Data-driven modelling and simulation for integration of production planning and simulation systems. Sel Eng Probl 2012:119–22.

[314] Pan Y, Hu M. A data-driven modeling approach for digital material additive manufacturing process planning. 2016 Int. Symp. Flex. Autom., 2016, p. 223–8. https://doi.org/10.1109/ISFA.2016.7790165.

[315] Geiger D. Data-Driven Manufacturing – Monetizing the Analytical Edge 2017. https://www.aberdeen.com/opspro-essentials/data-driven-manufacturing-monetizing-analytical-edge/ (accessed November 7, 2018).

[316] Chaturvedi AR, Hutchinson GK, Nazareth DL. Supporting complex real-time decision making through machine learning. Decis Support Syst 1993;10:213–33. https://doi.org/10.1016/0167-9236(93)90039-6.

[317] Yan W, Tang D, Lin Y. A data-driven soft sensor modeling method based on deep learning and its application. IEEE Trans Ind Electron 2017;64:4237–45. https://doi.org/10.1109/TIE.2016.2622668.

[318] Dong D, Mcavoy TJ. Nonlinear principal component analysis - Based on principal curves and neural networks. Comput Chem Eng 1996;20:65–78. https://doi.org/10.1016/0098-1354(95)00003-K.

[319] Kleinert T, Schladt M, Muehlbeyer S, Schocker A. Combination of process analytical technology with soft sensors for online process data and advanced process information. TM-TECHNISCHES Mess 2011;78:589–602. https://doi.org/10.1524/teme.2011.0165.

[320] Manning L, Soon JM. Food safety, food fraud, and food defense: A fast evolving literature. J Food Sci 2016;81:R823–34. https://doi.org/10.1111/1750-3841.13256.

[321] Gani R. Chemical product design: challenges and opportunities. Comput Chem Eng 2004;28:2441–57. https://doi.org/10.1016/J.COMPCHEMENG.2004.08.010.

[322] Parlikar U, Bundela PS, Baidya R, Ghosh SK. Effect of variation in the chemical constituents of wastes on the co-processing performance of the cement kilns. Procedia Environ Sci 2016;35:506–12. https://doi.org/10.1016/J.PROENV.2016.07.035.

[323] European Union. Directive 2010/75/EU of the European Parliament and of the Council of 24 November 2010 on industrial emissions (integrated pollution prevention and control), OJ L 334. 2010.

[324] Springer NP, Schmitt J. The price of byproducts: Distinguishing co-products from waste using the rectangular choice-of-technologies model. Resour Conserv Recycl 2018;138:231–7. https://doi.org/10.1016/J.RESCONREC.2018.07.034.

[325] Kabongo JD. Waste Valorization. In: Idowu SO, Capaldi N, Zu L, Gupta A Das, editors. Encycl. Corp. Soc. Responsib., Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, p. 2701–6. https://doi.org/10.1007/978-3-642-28036-8_680.

[326] The Access Group. Access ERP systems for all industries 2018. https://www.theaccessgroup.com/supply-chain/industries/ (accessed January 11, 2019).

[327] Himmelblau DM. Accounts of experiences in the application of artificial neural betworks in chemical engineering. Ind Eng Chem Res 2008;47:5782–96. https://doi.org/10.1021/ie800076s.

[328] Shearer C. The CRISP-DM model: the new blueprint for data mining. J Data Warehous 2000;5:13–22.

[329] Shafique U, Qaiser H. A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). Int J Innov Sci Res 2014;12:217–22.

[330] Al-Fattah SM, A Al-Naim H, Al-Naim HA. Artificial-intelligence technology predicts relative permeability of giant carbonate reservoirs. SPE Reserv Eval Eng 2009;12:96–108. https://doi.org/10.2118/109018-PA.

[331] Tchobanoglous G, Stensel DH, Tsuchihashi R, Burton F, Abu-Orf M, Bowden G, Pfrang W. Wastewater engineering: treatment and resource recovery. 5th ed. New York: New York : McGraw-Hill; 2014.

[332] Li S, Chen G, Li S, Chen G. Factors affecting the effectiveness of bioelectrochemical system applications: Data synthesis and meta-analysis. Batteries 2018;4:34. https://doi.org/10.3390/batteries4030034.

[333] Tao Q, Luo J, Zhou J, Zhou S, Liu G, Zhang R. Effect of dissolved oxygen on nitrogen and phosphorus removal and electricity production in microbial fuel cell. Bioresour Technol 2014;164:402–7. https://doi.org/10.1016/J.BIORTECH.2014.05.002.

[334] Pant D, Singh A, Van Bogaert G, Irving Olsen S, Singh Nigam P, Diels L, Vanbroekhoven K. Bioelectrochemical systems (BES) for sustainable energy production and product recovery from organic wastes and industrial wastewaters. RSC Adv 2012;2:1248–63. https://doi.org/10.1039/C1RA00839K.

[335] Huang B, Feng H, Wang M, Li N, Cong Y, Shen D. The effect of C/N ratio on nitrogen removal in a bioelectrochemical system. Bioresour Technol 2013;132:91–8. https://doi.org/10.1016/J.BIORTECH.2012.12.192.

[336] Feng H, Huang B, Zou Y, Li N, Wang M, Yin J, Cong Y, Shen D. The effect of carbon sources on nitrogen removal performance in bioelectrochemical systems. Bioresour Technol 2013;128:565–70. https://doi.org/10.1016/j.biortech.2012.11.004.

[337] Kelly PT, He Z. Nutrients removal and recovery in bioelectrochemical systems: A review. Bioresour Technol 2014;153:351–60. https://doi.org/10.1016/j.biortech.2013.12.046.

[338] Arugula MA, Brastad KS, Minteer SD, He Z. Enzyme catalyzed electricity-driven water softening system. Enzyme Microb Technol 2012;51:396–401. https://doi.org/10.1016/J.ENZMICTEC.2012.08.009.

[339] Wang X, Cheng S, Zhang X, Li X, Logan BE. Impact of salinity on cathode catalyst performance in microbial fuel cells (MFCs). Int J Hydrogen Energy 2011;36:13900–6. https://doi.org/10.1016/J.IJHYDENE.2011.03.052.

[340] Nancharaiah Y V, Mohan SV, Lens PNL. Bioresource technology metals removal and recovery in bioelectrochemical systems : A review. Bioresour Technol 2015;195:102–14. https://doi.org/10.1016/j.biortech.2015.06.058.

[341] Ofwat. How non-household customers are charged 2017. http://www.ofwat.gov.uk/nonhouseholds/yourwaterbill/hownonhousehold/ (accessed January 10, 2019).

[342] Albuquerque G, Lowe T, Magnor M. Synthetic generation of high-dimensional datasets. IEEE Trans Vis Comput Graph 2011;17:2317–24. https://doi.org/10.1109/TVCG.2011.237.

[343] Resende JA, Godon J-J, Bonnafous A, Arcuri PB, Silva VL, Otenio MH, Diniz CG. Seasonal variation on microbial community and methane production during anaerobic digestion of cattle manure in Brazil. Microb Ecol 2016;71:735–46. https://doi.org/10.1007/s00248-015-0647-y.

[344] Hobæk Haff I, Aas K, Frigessi A. On the simplified pair-copula construction - Simply useful or too simplistic? J Multivar Anal 2010;101:1296–310. https://doi.org/10.1016/j.jmva.2009.12.001.

[345] Stöber J, Joe H, Czado C. Simplified pair copula constructions—Limitations and extensions. J Multivar Anal 2013;119:101–18. https://doi.org/10.1016/J.JMVA.2013.04.014.

[346] Recio-Garrido D, Perrier M, Tartakovsky B. Modeling, optimization and control of bioelectrochemical systems. Chem Eng J 2016;289:180–90. https://doi.org/10.1016/j.cej.2015.11.112.

[347] Brissette FP, Khalili M, Leconte R. Efficient stochastic generation of multi-site

synthetic precipitation data. J Hydrol 2007;345:121–33. https://doi.org/10.1016/J.JHYDROL.2007.06.035.

[348] Dalmau M, Atanasova N, Gabarrón S, Rodriguez-Roda I, Comas J. Comparison of a deterministic and a data driven model to describe MBR fouling. Chem Eng J 2015. https://doi.org/10.1016/j.cej.2014.09.003.

[349] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 2018;81:1192–205. https://doi.org/10.1016/J.RSER.2017.04.095.

[350] Feng Y, Cui N, Chen Y, Gong D, Hu X. Development of data-driven models for prediction of daily global horizontal irradiance in Northwest China. J Clean Prod 2019;223:136–46. https://doi.org/10.1016/J.JCLEPRO.2019.03.091.

[351] Xia C, Zhang D, Pedrycz W, Zhu Y, Guo Y. Models for microbial fuel cells: A critical review. J Power Sources 2018;373:119–31. https://doi.org/10.1016/j.jpowsour.2017.11.001.

[352] Perrotta F, Parry T, Neves LC. Application of machine learning for fuel consumption modelling of trucks. Proc - 2017 IEEE Int Conf Big Data, Big Data 2017 2018;2018-Janua:3810–5. https://doi.org/10.1109/BigData.2017.8258382.

[353] Hastie T, Tibshirani R, Friedman J. Overview of supervised learning. Elem. Stat. Learn., 2009, p. 9–41.

[354] Ramírez-Gallego S, Krawczyk B, García S, Woźniak M, Herrera F. A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing 2017;239:39–57. https://doi.org/10.1016/J.NEUCOM.2017.01.078.

[355] Abualigah LM, Khader AT, Al-Betar MA, Alomari OA. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Syst Appl 2017;84:24–36. https://doi.org/10.1016/j.eswa.2017.05.002.

[356] Yang W, Wang K, Zuo W. Neighborhood component feature selection for high-dimensional data image enhancement and restoration. J Comput 2012;7:161–8. https://doi.org/10.4304/jcp.7.1.161-168.

[357] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J

Forecast 2006;22:679–88. https://doi.org/10.1016/j.ijforecast.2006.03.001.

[358] Box GEP. Science and statistics. J Am Stat Assoc 1976;71:791–9.

[359] Liu Y, Gunawan R. Bioprocess optimization under uncertainty using ensemble modeling. J Biotechnol 2017;244:34–44. https://doi.org/10.1016/J.JBIOTEC.2017.01.013.

[360] Piuleac CG, Rodrigo MA, Cañizares P, Curteanu S, Sáez C. Ten steps modeling of electrolysis processes by using neural networks. Environ Model Softw 2010. https://doi.org/10.1016/j.envsoft.2009.07.012.

[361] Tools: Water Efficiency in the Food and Drink industry | WRAP UK n.d. http://www.wrap.org.uk/node/31479 (accessed August 13, 2019).

[362] Bauer A, Leonhartsberger C, Bösch P, Amon B, Friedl A, Amon T. Analysis of methane yields from energy crops and agricultural by-products and estimation of energy potential from sustainable crop rotation systems in EU-27. Clean Technol Environ Policy 2010;12:153–61. https://doi.org/10.1007/s10098-009-0236-1.

[363] Le H, Kabbur S, Pollastrini L, Sun Z, Mills K, Johnson K, Karypis G, Hu W-S. Multivariate analysis of cell culture bioprocess data—Lactate consumption as process indicator. J Biotechnol 2012;162:210–23. https://doi.org/10.1016/J.JBIOTEC.2012.08.021.

[364] Bradford E, Schweidtmann AM, Zhang D, Jing K, del Rio-Chanona EA. Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes. Comput Chem Eng 2018;118:143–58. https://doi.org/10.1016/J.COMPCHEMENG.2018.07.015.

[365] Spyromitros-Xioufis E, Grigorios Tsoumakas ·, Groves W, Vlahavas I, Tsoumakas G. Multi-target regression via input space expansion: treating targets as inputs. Mach Learn 2016;104:55–98. https://doi.org/10.1007/s10994-016-5546-z.

[366] Himmelblau DM. Applications of artificial neural networks in chemical engineering. Korean J Chem Eng 2000;17:373–92. https://doi.org/10.1007/BF02706848.

[367] Tsoumakas G, Spyromitros-Xioufis E, Vrekou A, Vlahavas I. Multi-target regression via random linear target combinations, Springer, Berlin, Heidelberg;

2014, p. 225–40. https://doi.org/10.1007/978-3-662-44845-8_15.

[368] Kocev D, Vens C, Struyf J, Džeroski S. Ensembles of multi-objective decision trees. Mach. Learn. ECML 2007, Berlin, Heidelberg: Springer Berlin Heidelberg; 2007, p. 624–31. https://doi.org/10.1007/978-3-540-74958-5_61.

[369] Hadavandi E, Shahrabi J, Shamshirband S. A novel Boosted-neural network ensemble for modeling multi-target regression problems. Eng Appl Artif Intell 2015;45:204–19. https://doi.org/10.1016/J.ENGAPPAI.2015.06.022.

[370] Ranjan R, Patel VM, Chellappa R. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans Pattern Anal Mach Intell 2019;41:121–35. https://doi.org/10.1109/TPAMI.2017.2781233.

[371] Baxter J, Jonathan. Learning internal representations. Proc. eighth Annu. Conf. Comput. Learn. theory - COLT '95, New York, New York, USA: ACM Press; 1995, p. 311–20. https://doi.org/10.1145/225298.225336.

[372] Caruana R. Learning many related tasks at the same time with backpropagation. In: Tesauro G, Touretzky DS, Leen TK, editors. Adv. Neural Inf. Process. Syst. 7, MIT Press; 1995, p. 657–64.

[373] UK Power. Compare Energy Prices Per kWh | Gas & Electric Per Unit | UKPower n.d. https://www.ukpower.co.uk/home_energy/tariffs-per-unit-kwh (accessed September 30, 2019).

[374] Dubiński J. Sustainable development of mining mineral resources. J Sustain Min 2013;12:1–6. https://doi.org/10.7424/JSM130102.

[375] Scott Fields. The Earth's open wounds: Abandoned and orphaned mines. Environ Health Perspect 2003;111:154–61.

[376] Kefeni KK, Msagati TAM, Mamba BB. Acid mine drainage: Prevention, treatment options, and resource recovery: A review. J Clean Prod 2017;151:475–93. https://doi.org/10.1016/J.JCLEPRO.2017.03.082.

[377] Voulvoulis N, Arpon KD, Giakoumis T. The EU Water Framework Directive: From great expectations to problems with implementation. Sci Total Environ 2017;575:358–66. https://doi.org/10.1016/J.SCITOTENV.2016.09.228.

[378] Mighanetara K, Braungardt CB, Rieuwerts JS, Azizi F. Contaminant fluxes from point and diffuse sources from abandoned mines in the River Tamar catchment,

UK. J Geochemical Explor 2009;100:116–24. https://doi.org/10.1016/J.GEXPLO.2008.03.003.

[379] Godoi LAG de, Foresti E, Damianovic MHRZ. Down-flow fixed-structured bed reactor: An innovative reactor configuration applied to acid mine drainage treatment and metal recovery. J Environ Manage 2017;197:597–604. https://doi.org/10.1016/J.JENVMAN.2017.04.027.

[380] Moorhouse AML, Wyatt LM, Watson IA, Hill S. A high surface area media treatment trial of a circum-neutral, net alkaline coal mine discharge in the South Derbyshire Coal Field (UK) using hydrous ferric oxide. In: Brown A, editor. Int. Mine Water Assoc. Annu. Conf. 2013 Reliab. Mine Water Technol., Golden: 2013, p. 667–72. https://doi.org/9781627487757.

[381] Oh C, Yu C, Cheong Y, Yim G, Song H, Hong J, Ji S. Efficiency assessment of cascade aerator in a passive treatment system for Fe(II) oxidation in ferruginous mine drainage of net alkaline. Environ Earth Sci 2015;73:5363–73. https://doi.org/10.1007/s12665-014-3791-7.

[382] Oyewo OA, Agboola O, Onyango MS, Popoola P, Bobape MF. Current methods for the remediation of acid mine drainage including continuous removal of metals from wastewater and mine dump. In: Prasad MNV, Favas PJ de C, Maiti SK, editors. Bio-Geotechnologies Mine Site Rehabil. 1st ed., Elsevier; 2018, p. 103–14. https://doi.org/10.1016/B978-0-12-812986-9.00006-3.

[383] Hurtado C, Viedma P, Cotoras D. Design of a bioprocess for metal and sulfate removal from acid mine drainage. Hydrometallurgy 2018;180:72–7. https://doi.org/10.1016/J.HYDROMET.2018.07.006.

[384] Sahinkaya E. Biotreatment of zinc-containing wastewater in a sulfidogenic CSTR: Performance and artificial neural network (ANN) modelling studies. J Hazard Mater 2009;164:105–13. https://doi.org/10.1016/J.JHAZMAT.2008.07.130.

[385] Atasoy AD, Babar B, Sahinkaya E. Artificial neural network prediction of the performance of upflow and downflow fluidized bed reactors treating acidic mine drainage water. Mine Water Environ 2013;32:222–8. https://doi.org/10.1007/s10230-013-0232-x.

[386] Cascade Filterpak™ | Veolia n.d. https://www.veoliawatertechnologies.com/en/solutions/products/cascade-

filterpak (accessed September 18, 2019).

[387] Saroha AK, Khera R. Hydrodynamic study of fixed beds with cocurrent upflow and downflow. Chem Eng Process Process Intensif 2006;45:455–60. https://doi.org/10.1016/J.CEP.2005.11.005.

[388] González E, Díaz O, Segredo-Morales E, Rodríguez-Gómez LE, Vera L. Enhancement of peak flux capacity in membrane bioreactors for wastewater reuse by controlling the backwashing strategy. Ind Eng Chem Res 2019;58:1373–81. https://doi.org/10.1021/acs.iecr.8b05650.

[389] Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013;64:402–6. https://doi.org/10.4097/kjae.2013.64.5.402.

[390] Alexander DLJ, Tropsha A, Winkler DA. Beware of R2: Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. J Chem Inf Model 2015;55:1316–22. https://doi.org/10.1021/acs.jcim.5b00206.

[391] Samuelsson O, Björk A, Zambrano J, Carlsson B. Gaussian process regression for monitoring and fault detection of wastewater treatment processes. Water Sci Technol 2017;75:2952–63. https://doi.org/10.2166/wst.2017.162.

[392] Box GEP, Hunter JS, Hunter WG. Statistics for experimenters : design, innovation, and discovery. Wiley-Interscience; 2005.

[393] Alara OR, Abdurahman NH, Afolabi HK, Olalere OA. Efficient extraction of antioxidants from Vernonia cinerea leaves: Comparing response surface methodology and artificial neural network. Beni-Suef Univ J Basic Appl Sci 2018;7:276–85. https://doi.org/10.1016/J.BJBAS.2018.03.007.

[394] Patel KA, Brahmbhatt PK. A comparative study of the RSM and ANN models for predicting surface roughness in roller burnishing. Procedia Technol 2016;23:391–7. https://doi.org/10.1016/J.PROTCY.2016.03.042.

[395] Vargas JM, Nielsen S, Cárdenas V, Gonzalez A, Aymat EY, Almodovar E, Classe G, Colón Y, Sanchez E, Romañach RJ. Process analytical technology in continuous manufacturing of a commercial pharmaceutical product. Int J Pharm 2018;538:167–78. https://doi.org/10.1016/J.IJPHARM.2018.01.003.

[396] Houben C, Peremezhney N, Zubov A, Kosek J, Lapkin AA. Closed-loop multitarget optimization for discovery of new emulsion polymerization recipes. Org Process Res Dev 2015;19:1049–53.

https://doi.org/10.1021/acs.oprd.5b00210.

[397] Chiang L, Lu B, Castillo I. Big data analytics in chemical engineering. Annu Rev Chem Biomol Eng 2017;8:63–85. https://doi.org/10.1146/annurev-chembioeng-060816-101555.

[398] Upstream Consulting | Accenture n.d. https://www.accenture.com/gb-en/services/energy/upstream (accessed September 26, 2019).

[399] Cognizant. The symbiosis of humans and machines | artificial intelligence n.d. https://www.cognizant.com/en-uk/artificial-intelligence-for-business?utm_source=google&utm_medium=ppc&utm_campaign=EU_COG_UKI_SEA_BT_AI_Generic_EN_IE_BMM%257c%257cEU%257cUKI%257ccrsIND%257ccrsSERV%257ccrsHZ%257cGeneric&utm_term=machine%2520learning&utm_co (accessed September 26, 2019).

[400] Intelligent Plant n.d. https://www.intelligentplant.com/index.html (accessed July 2, 2019).

[401] Splunk. The Data-to-Everything Platform n.d. https://www.splunk.com/ (accessed September 26, 2019).

[402] Lemonick S. Is machine learning overhyped? Chemists weigh in on the technique's possibilities and its pitfalls. Chem Eng News 2018.

[403] Vooradi R, Anne SB, Tula AK, Eden MR, Gani R. Energy and CO2 management for chemical and related industries: issues, opportunities and challenges. BMC Chem Eng 2019;1:7. https://doi.org/10.1186/s42480-019-0008-6.

[404] Najjar YSH. Hydrogen safety: The road toward green technology. Int J Hydrogen Energy 2013;38:10716–28. https://doi.org/10.1016/J.IJHYDENE.2013.05.126.

[405] Pollard DJ, Woodley JM. Biocatalysis for pharmaceutical intermediates: the future is now. Trends Biotechnol 2007;25:66–73. https://doi.org/10.1016/J.TIBTECH.2006.12.005.

[406] ML Engine | Cloud Machine Learning Engine (Cloud ML Engine) | Google Cloud n.d. https://cloud.google.com/ml-engine/ (accessed May 23, 2019).

[407] Artificial Intelligence | Dell EMC United Kingdom n.d. https://www.dellemc.com/en-gb/solutions/artificial-intelligence/index.htm

(accessed May 23, 2019).

[408] David Sprott, Lawrence Wilkes. Understanding service-oriented architecture. Archit J 2004;1:10–7.

[409] Wen X, Zhou X. Servitization of manufacturing industries based on cloud-based business model and the down-to-earth implementary path. Int J Adv Manuf Technol 2016;87:1491–508. https://doi.org/10.1007/s00170-014-6348-8.

[410] Esposito C, Castiglione A, Martini B, Choo K-KR. Cloud manufacturing: security, privacy, and forensic concerns. IEEE Cloud Comput 2016;3:16–22. https://doi.org/10.1109/MCC.2016.79.

[411] Wang H, Ren ZJ. Bioelectrochemical metal recovery from wastewater: A review. Water Res 2014;66:219–32. https://doi.org/10.1016/J.WATRES.2014.08.013.

VOLUME 2

PUBLISHED WORKS AND APPENDICES

Oliver Fisher, MEng

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

October, 2019

APPENDIX A: PUBLICATION 1

Cloud manufacturing as a sustainable process manufacturing route

Oliver FISHER, Nicholas WATSON, Rachel L GOMES

FWW - Food, Water, Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom

Corresponding author
Author: Rachel L Gomes
Tel.: +44 115 846 8883.
E-mail address: rachel.gomes@nottingham.ac.uk

Abstract

Cloud Manufacturing (CM) is a service oriented business model to share manufacturing capabilities and resources on a cloud platform. Manufacturing is under pressure to achieve cost and environmental impact reductions, as manufacturing becomes more integrated and complex. Cloud manufacturing offers a solution, as it is capable of making intelligent decisions to provide the most sustainable and robust manufacturing route available. Although CM research has progressed, a consensus is still lacking on the concepts within CM as well as applications and scope beyond discrete manufacturing.

The aim of this paper is to demonstrate how CM offers a more sustainable manufacturing future to the industry as a whole, before focusing specifically on the application to process manufacturing (e.g. food, pharmaceuticals and chemicals). This paper details the definitions, characteristics, architectures and previous case studies on CM. From this, the fundamental aspects of the CM concept are identified, along with an analysis of how the concept has progressed. A new, comprehensive CM definition is formulated by combining key concepts drawn from previous definitions and emphasizes CM potential for sustainable manufacturing.

Four key methods of how CM increases sustainability are identified: (1) collaborative design; (2) greater automation; (3) improved process resilience and (4) enhanced waste reduction, reuse and recovery. The first two key methods are common to both discrete and process manufacturing, however key methods (3) and (4) are more process manufacturing specific and application of CM for these has yet to be fully realised. Examples of how CM's characteristics may be utilised to solve various process manufacturing problems are presented to demonstrate the applications of CM to process manufacturing. Waste is an important consideration in manufacturing, with strong sustainability implications. The current focus has been on using CM for waste minimisation; however, process manufacturing offers waste as a resource (valorisation opportunities from diversifying co-products, reuse, recycle and energy recovery). Exploring CM's potential to characterise and evaluate alternative process routes for the valorisation of process manufacturing waste is considered for the first time. The specific limitations preventing CM adoption by process manufacturers are discussed. Finally, CM's place in the future of manufacturing is explored, including how

it will interact with, and complement other emerging manufacturing technologies to deliver a circular economy and personalised products.

Keywords

Cloud manufacturing, Circular economy, Process manufacturing, Resilient manufacturing, Waste valorization

Contents

Abstract 274

Keywords 275

1. Introduction 276

2. Cloud Manufacturing 277

2.1. Cloud manufacturing definition 277

2.2. Key characteristics of cloud manufacturing 279

2.3. Cloud manufacturing frameworks, architectures and implementation 282

2.4. Case studies of cloud manufacturing's application 284

3. Cloud manufacturing as a route to sustainable manufacturing 288

3.1. Sustainability drivers and barriers 288

3.2. Cloud manufacturing enabling collaboration for innovative designs 288

3.3. Greater automation from the adoption of cloud manufacturing systems 289

4. Application to process engineering manufacturing 289

4.1. Difference between discrete and process manufacturing 290

4.2. Application of collaborative design and greater automation to process manufacturing 291

4.3. Cloud manufacturing for waste valorisation 293

4.4. How cloud manufacturing increases process resilience 295

4.5. Examples of cloud manufacturing's application to process manufacturing 297

5. Limitations and future research directions 301

5.1. Data sharing among process manufacturers 301

5.2. Processing plant security 301

5.3. Characterisation of process streams 302

5.4. Barriers preventing SMEs adopting new technologies 302

6. Future of Manufacturing 302

7. Conclusion 303

Acknowledgements 306

Funding 306

References 306

1. Introduction

Today's manufacturing sector includes a broad range of industries, from food and drink, pharmaceutical, automobile and aerospace, to computer and electronic product manufacturing. Manufacturing is a vital sector in society, irrespective of being a high or low-income economy. In previous decades the manufacturing sectors of the USA, Japan, the UK, France and Italy have lost global market shares due to pressure from China. The exception being Germany, which has instead increased exports to China and Asia [1]. This is in part due to Germany's Government support of the application of manufacturing technology. Since the 2008 recession, policymakers in the West have focused on rebalancing the economy away from the service sector towards manufacturing [2]. Manufacturing in the USA and the UK account for 12% and 10% of the country's gross domestic product (GDP) respectively [3]. Whereas, in Germany the manufacturing sector accounts for 23% of GDP [3]. This led to the German economy making a faster recovery. It is not only Western economies that are facing new manufacturing challenges. The recent trend has been for manufacturers to move to low-income economies to benefit from cheaper production costs. However, China's manufacturing growth is slowing due the rapid increase in the cost of human resources, materials and energy [3]. Governments are facing the further challenge of developing policies which support manufacturing, while maintaining their commitment made in Paris to lower carbon emissions [4]. There are also additional drivers such as the EU Waste Framework, which promotes the waste hierarchy with financial incentives to minimise and/or reuse wastes [5]. The circular economy has emerged as an alternative economic model that supports growth, while minimising environmental impacts from production and consumption. The Waste and Resources Action Program charity offers a clear definition of the circular model: "A circular economy is an alternative to a traditional linear economy (make, use, dispose) in which we keep resources in use for as long as possible, extract the maximum value from them whilst in use, then recover and regenerate products and materials at the end of each service life." [6]. To create a sustainable manufacturing future, new manufacturing models need to be explored that support this circular approach.

Various advanced manufacturing models have been proposed to achieve this [7-10]. A concurrent concept across these models is the idea of an increase in collaboration, automation, and knowledge and data sharing throughout the supply chain. As a result, the manufacturing process benefits from an increase in flexibility, resource efficiency, and customisation. Recently a new manufacturing paradigm called Cloud Manufacturing (CM) has emerged, which serves to act as a platform to deliver this vision [11]. Underlying the CM model is the concept of sharing manufacturing capabilities and resources on a cloud platform capable of making intelligent decisions to provide the most sustainable and robust manufacturing route available. Limitations in largely computer processing power, data collection and analytics, and security solutions have hindered the deployment of previous advanced manufacturing models. Cloud manufacturing however, has the advantage of utilising new technologies such as cloud computing, big data, and the Industrial Internet of Things (IIoT, the industrial focus of the Internet of Things), enabling CM to overcome many of these previous limitations. Cloud manufacturing also has the potential to be a key component in the future manufacturing landscape. It complements emerging manufacturing processes, such as Additive Manufacturing (AM), to deliver consumer customised products via

sustainable processes. Where AM utilises 3D design data to build up a component in layers by depositing material, it is more commonly known as 3D printing [12].

Several projects have investigated research themes and applications of CM. In 2009, China launched the National High-Tech Research and Development Programme to promote research related to CM [13]. Two of the earliest and most cited papers on CM came from China. Tao et al. established and illustrated the basic ideas of CM including architecture, definition, characteristics and benefits [14]. Xu analysed how cloud computing is changing manufacturing and defined the difference between cloud computing alongside manufacturing and CM [15]. Under the European Seventh Framework Programme the project ManuCloud was established. The ManuCloud project objective was to investigate and develop a cloud-based infrastructure to provide better support for on-demand manufacturing supply chains in the photovoltaic, organic lighting and automotive sectors [16]. There are now companies emerging which offer cloud-based services for manufacturing. For example, Plex Systems has developed the Plex Manufacturing Cloud, which is a software as a service (SaaS) enterprise resource planner (ERP) that manages the manufacturing process [17]. The software is designed to provide managers and engineers with real time access to production data. However, the full implications of CM have yet to be fully realised in industry. Industry 4.0 was first coined in Germany and refers to the digitisation of the manufacturing sector driven by huge increases in data volumes, computational power and increasing manufacturing plant connectivity [18]. Cloud manufacturing has potential to work within Industry 4.0 to exploit an on-demand access to a shared collection of manufacturing resources to form temporary, reconfigurable supply chains with enhanced efficiency, reduced production costs, and optimal resource allocation.

This paper shall aim to explain how CM may offer a route to sustainable process manufacturing. To achieve this effectively, this paper will address five key questions:

1. What is CM, its key characteristics and what are the proposed frameworks and architectures for implementing CM?
2. How does CM offer a route to future sustainable manufacturing?
3. How does process manufacturing vary from discrete manufacturing?
4. What are the opportunities for CM to benefit process manufacturing?
5. What limitations specific to process manufacturing currently prevent the widespread adoption of CM, and what future research is necessary to overcome these limitations?

2. Cloud Manufacturing

2.1. Cloud manufacturing definition

Cloud manufacturing's first definition emerged in 2010 [11]. However, since that publication CM has generated considerable research interest and there now exists a variety of definitions. Unfortunately, there has yet to be an agreed common definition. Even the nomenclature and abbreviations can vary from author to author. A selection of some of the first CM definitions to emerge are detailed in Table 1, these definitions have since been commonly cited in recent publications on CM [19-25].

Table 1
Cloud manufacturing definitions arising from literature to date

Definition	Author/s
"Cloud-based design and manufacturing (CBDM) refers to a product development model that enables collective open innovation and rapid product development with minimum cost through social networking and crowd-sourcing platforms coupled with shared service pools of design, manufacturing resources and components."	[26]
"Cloud manufacturing (CMfg) is a computing and service-oriented manufacturing model developed from existing advanced manufacturing models (e.g., application service providers, agile manufacturing, networked manufacturing, manufacturing grids) and enterprise information technologies under the support of cloud computing, the Internet of Things (IoT), virtualization and service-oriented technologies, and advanced computing technologies."	[14]
"Cloud-based design and manufacturing (CBDM) refers to a service-oriented product development model in which service consumers are able to configure products or services as well as reconfigure manufacturing systems through Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), Hardware-as-a-Service (HaaS), and Software-as-a-Service (SaaS) in response to rapidly changing customer needs. CBDM is characterised by on demand self-service, ubiquitous access to networked data, rapid scalability, resource pooling, and virtualisation. The types of deployment models include private, public, and hybrid clouds."	[27]
"Cloud manufacturing (CM) is a customer-centric manufacturing model that exploits on-demand access to a shared collection of diversified and distributed manufacturing resources to form temporary, reconfigurable production lines with enhanced efficiency, reduced product lifecycle costs, and allow for optimal resource loading in response to variable-demand customer generated tasks."	[28]
"Cloud manufacturing (CM) is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable manufacturing resources (e.g., manufacturing software tools, manufacturing equipment, and manufacturing capabilities) that can be rapidly provisioned and released with minimal management effort or service provider interaction."	[15]
"A new service-oriented, high efficient and low consumptive, knowledge-based and intelligent networked agile manufacturing model and technology, allowing manufacturing resources and capabilities to be virtualized and transformed into on-demand services available to users through a product life cycle."	[29]

Each of these definitions prioritise different aspects of CM; for example, service oriented manufacturing, product development, resource allocation or customer centric manufacturing. The chosen aspect normally reflects the possible application of CM that was being demonstrated and was therefore highlighted in the authors' definition. Despite the differences in these definitions, all contain common elements such as, network manufacturing, ubiquitous access, multi-tenancy, virtualisation, big data, everything-as-a-service, scalability, and resource pooling. These can be used to

help define the fundamental principles and characteristics of CM, which in turn give thought to how it can be utilised to achieve a sustainable manufacturing route.

To draw out the main themes, these definitions have been entered into a word cloud generator [30], the result of which is presented in Fig. 1. The word cloud highlights the following concepts: service oriented, on-demand, network, intelligent, access, knowledge based, shared pool, rapid, scalability, advance and virtualised. This paper proposes a comprehensive CM definition combining the highlighted concepts from all these definitions and emphasising CM potential for sustainable manufacturing.

> "Cloud manufacturing is a service oriented manufacturing model that virtualises manufacturing resources and capabilities into on-demand services accessed through the manufacturing cloud. This transforms manufacturing supply lines to become temporary, and provide greater flexibility and scalability resulting in increased resilience and sustainability throughout the manufacturing process. Cloud manufacturing is a multi-tenant, intelligent, knowledge based platform that can provide sustainable solutions throughout the product and process life cycle via effective collaboration, data mining, and communication across a network of developers, manufacturers and consumers."

This definition, although comprehensive, is too long to easily bring to mind when asked what CM is. Therefore, a shortened definition is offered that captures the essence of CM.

> "The cloud manufacturing model is the concept of sharing manufacturing capabilities and resources on a cloud platform capable of making intelligent decisions to provide the most sustainable and robust manufacturing route available."



Fig. 1. Cloud manufacturing word cloud derived from [14, 15, 26-29]

2.2. Key characteristics of cloud manufacturing

Cloud manufacturing differentiates from other manufacturing paradigms because of several unique characteristics. Existing papers have explained in detail a selection of CM's characteristics or listed them without explanation [14, 28]. This paper complies all fundamental characteristics and explains their meaning, application within a CM

environment and the benefits they bring. This is necessary because defining the characteristics helps to provide a clearer understanding on how CM offers a sustainable manufacturing future.

Flexibility and scalability:

A major advantage of CM over traditional manufacturing systems is the scalability offered. By nature CM production lines are no longer fixed but temporary, allowing for the option of small or long production runs depending on the user and market demands [28, 31]. Users of manufacturing resources are matched to the nearest available and appropriate manufacturing resource provider depending on their need. These supply lines are highly reconfigurable, dynamic and agile giving CM the ability to adapt to unpredicted changes of circumstance. For the system to be agile when repurposing manufacturing resources across any number of sites, a high level of automation across the supply lines shall be required. Section 3.3 explains what is meant by automation in a manufacturing context and how CM will further automatise the manufacturing process.

Multi-tenancy:

An important characteristic of CM is its ability to connect tenants intelligently for mutual benefit without human intervention. Each tenant is given some degree of control of their manufacturing cloud's architecture, e.g. adaptability of the user interface, without changing the application code [32]. The mechanisms by how multiple tenants and manufacturing clouds will interact is expanded on in section 2.3.

An evaluation and reviewing system for users, like what is seen on Amazon, will be required to ensure quality of service provided. MFG.com is a real world example of a multi-tenant service that aims to quickly match manufacturers to suppliers [33]. It uses consumer generated ratings to ensure consumers are matched to suppliers that have demonstrated their products meets quality requirements, are safe and delivered on time.

Knowledge intensive, intelligent decision making tool:

Increasingly vast data sets are generated from data collected throughout the manufacturing process, but with varying degrees of utilisation. Currently the tendency is for the data to be compressed and archived for record keeping, and often only used for after the fact emergency analysis [34]. Cloud manufacturing acts as a mechanism to share this data within the manufacturing cloud so that useful manufacturing knowledge is extracted to support intelligent decision-making algorithms. Big data shall play a critical role in data collection and analysis [35]. As well as enhancing the process, data and manufacturing knowledge sharing across the manufacturing cloud is required for it to operate effectively. For example, four fundamental functions of CM that require knowledge support have been outlined as examples [14]:

1. Manufacturing resource and ability perception, connection, and virtualisation;
2. Cloud service description, match and search, aggregation, and composition;
3. Optimal allocation and scheduling;

4. Business workflow management.

However, the effectiveness of utilising data to improve manufacturing is dependent on the quality of the data collected. The data collected must be appropriate to the situation; therefore, calculated decisions must be taken to identify how the data is collected and the required granularity of that data.

Intelligent on-demand manufacturing:

Network manufacturing is a paradigm that has existed since 2000 [8]. It operates as a network of distributed manufacturing resources that are integrated on a joint manufacturing process. Where CM differs is in the method of communication between enterprises. Cloud manufacturing consists of a network of manufacturing resources that may operate automatically and/or independently via the decisions taken by the cloud platform. These manufacturing resources may be distributed within a single enterprise or shared across multiple enterprises.

On the cloud there will exist a 'pool' of shared manufacturing resources that is coordinated by the cloud platform enabling on-demand manufacturing. It must act intelligently, for example not overloading one manufacturing resource but utilising required resources in the most effective way to meet the user's requirements. Wu et al. illustrates this point effectively:

> "If manufacturing resource "A" is more heavily utilized than others in the network, yet the desired process can be performed by combining manufacturing resources "B" and "C", the CM environment will automatically realize and capitalize upon this alternative to avoid excessive loading of manufacturing resource "A". [28]

Manufacturing resource sharing and circulation are very important objectives in CM. It is a prerequisite of achieving on-demand and collaborative configuration of resources and services [20].

Manufacturing as a service:

A recurring theme from various CM definitions is the concept of manufacturing moving from production oriented manufacturing to service oriented manufacturing. Service oriented manufacturing integrates services throughout the product lifecycle, and allows producers a greater understanding of their product leading to a more efficient production process [36]. Cloud manufacturing achieves this by packaging manufacturing resources and capabilities as services, coordinated through the cloud on a pay-as-you-go scheme. Cloud manufacturing acts as centralised management platform for these services with a key goal of linking cloud consumers to cloud providers who can meet the product requirements of the consumer while maintaining cost, time and quality constraints. Cloud computing is currently operated as a service model and has an architecture split into three layers, Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [15, 32]. Extending this concept to CM Tao et al. explains how the CM environment is enabled by the creation of Manufacturing Cloud Services (MCSs) [14]. Manufacturing resources and capabilities are virtualised and categorised, before being sorted into an appropriate manufacturing cloud. These services are then sold through the cloud on a pay-as-you-go scheme, offering greater flexibility to manufacturers than traditional fixed contacts.

2.3. Cloud manufacturing frameworks, architectures and implementation

Since CM's inception in 2010 there can be seen numerous proposed architectures for its implementation. Two similar initial proposals focused on a layer hierarchal framework for a possible structure of CM's architecture [14, 15]. As research has progressed, this layer structure has been varied from three layers to twelve depending on the detail and depth the authors have undertaken [27, 29, 37, 38]. Common to these proposals is the issue of identifying and virtualising manufacturing resources and capabilities. These are encapsulated in the cloud as services to be managed though a centralised intelligent platform. To provide real time control of these manufacturing services they are connected by utilising IIoT technologies. This forms the bottom part of the layered framework. The middle section is then concerned with managing the cloud services. It is the backbone of the system and has a variety of duties to maintain and operate the cloud such as: scheduling, matching users with providers, quality control of services, remote monitoring of resources, fee calculation, evaluation, optimisation of process, etc. The higher layers of the framework are the application layers, which act as a portal through which cloud users interface with the system. Cloud users can communicate their individual needs and request various manufacturing cloud resources. A crucial feature of all the proposed frameworks is that knowledge is shared across the entire system and cloud security is constant and robust. A simplified, concise model of common features is presented in Fig. 2.

To understand how the CM architecture would be implemented, it is important to define the cloud users and their roles. There are commonly three defined roles, although the exact nomenclatural of each role varies. These are the cloud consumer, cloud provider and cloud operator [14, 15, 26, 29]. Cloud consumers and cloud providers have the obvious roles of one is utilising the cloud services for manufacturing purposes and one is providing these services. The cloud operator manages the use, performance, and delivery of cloud services as well as negotiating the relationship between cloud providers and cloud consumers. This role is sometimes split to involve a fourth cloud user known as a cloud carrier, which acts as the intermediary that provides connectivity and transport to enable exchange of services between consumers and providers [39]. Although it has been noted that one user may perform two or more of these roles [26], the implications of this on the possible manufacturing clouds and their business models has only recently been explored [23, 40]. Guo developed four possible models from different combinations of relationships between these users and the roles within the manufacturing cloud [23]. It was noted that the manufacturing cloud system design would be directly affected depending on the desired relationship between cloud users.

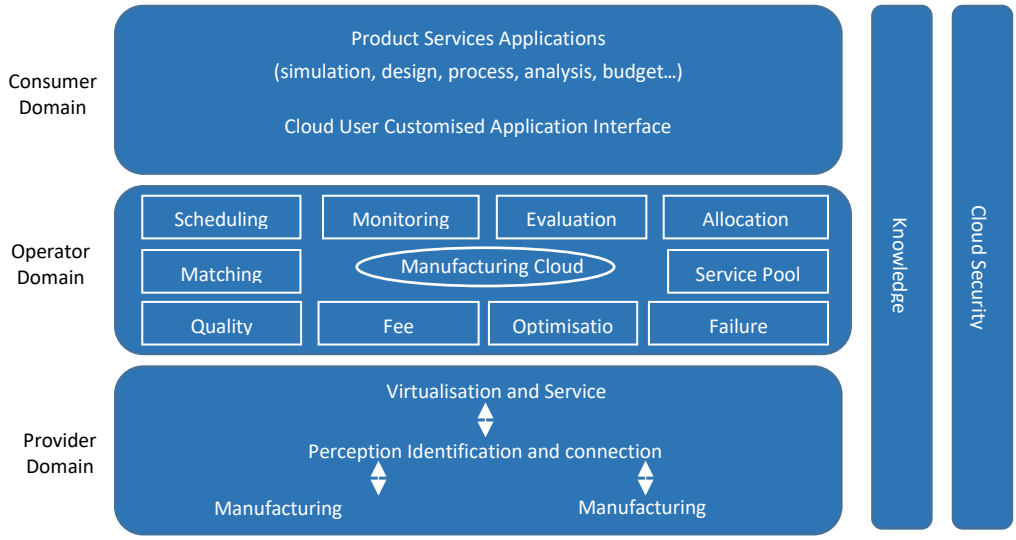


Fig. 2. Cloud manufacturing architecture adapted from [14, 15]

The current literature on CM's architecture broadly agrees on an overview for the framework and has instead moved focus onto how the system may be implemented. Part of that involves fully defining each layer and how it functions. Recent research has tried to achieved this by first defining an architecture, containing similar elements to those previously proposed, before breaking down each layer individually into flow charts to demonstrate how they would operate in practice [37].

The mechanism through which users shall interface depends on the manufacturing cloud deployment model. The manufacturing cloud has four potential deployment models: private, public, community and hybrid [15, 27, 41]. A private cloud would contain a single tenant's data and information and would provide services specific within that enterprise. By being private within an enterprise it allows security of data, with the exception of hacking. Multiple tenants who have a common interest may share certain knowledge or resources for mutual gain through the community cloud. On the public cloud manufacturing resources and capabilities and data is shared and managed by a third party and is the platform by which tenants from various backgrounds shall be able to interact with each other. Hybrid clouds consist of multiple external and internal clouds. An enterprise may have to utilise all four versions depending on the tasks.

Early research focused on single, isolated manufacturing clouds; however, for CM to be successful it will require the integration of multiple manufacturing clouds [42]. A hybrid manufacturing cloud allows enterprises to create multiple cloud models (public, private and community) depending on their business goals [43]. It provides a flexible and open environment that breaks down the barrier between organisations to allow for sharing of resources between enterprises, but it does not present a comprehensive method to integrate separate enterprises' manufacturing clouds. A hybrid integration framework for multiple manufacturing clouds was developed to enable users to find the most appropriate services across multiple manufacturing clouds, through a third party CM platform [21]. Similarly, a further system model has been considered where multiple 'cloud terminal subsystems' are connected to a 'cloud platform subsystem' [23]. The cloud terminal system is based on the enterprise information system. Industrial Internet of Things technologies collect information of enterpriser's

information elements and provide an interconnection data interface with the cloud platform. Furthermore, it is responsible for integrating the manufacturing resources and manufacturing capabilities of the enterprise into the cloud platform for remote access and control. It would act as an enterprise's individual manufacturing cloud. The cloud platform subsystem acts as a big data management centre responsible for the convergence, integration, distribution, monitoring and operation management of the various manufacturing cloud services provided by the connected enterprises. It also provides the human-computer interface. By splitting the CM system into various cloud terminals and a cloud platform, Guo helps to demonstrate how multiple enterprises may possibly operate within the CM architecture [23].

2.4. Case studies of cloud manufacturing's application

As the research into CM has expanded, two directions have emerged. Broadly speaking, academia have recently been focusing on the interpretation and analysis of CM, using case studies that demonstrate different applications of CM. A selection of the recent case studies from the literature is presented in Table 2. These have been chosen because they demonstrate CM application in various sectors, each highlighting different CM characteristics. How and why CM is used is briefly summarised, and the main positive outcomes from the CM approach are detailed. All the case studies demonstrate how CM may achieve the optimal solution, as defined by the author. However, no one case study reached an optimal solution, and further work has been suggested for each in order to reach a more complete solution.

The focus in industry has been on developing a CM architecture, as seen from examples from the industry leaders. Siemens has developed a cloud platform called 'MindSphere' that interlinks physical products and production facilities with digital data, enabling innovative solutions and getting products onto the market faster, better and more efficiently [44]. Similarly, Plex Manufacturing Cloud delivers production control, inventory management, supply chain management, and more [17]. IBM has the IBM Watson IoT, which mines the influx of data from an IIoT manufacturing process to make businesses more cognitive [45]. Industry has directed some application of CM, which is utilising manufacturing clouds to monitor and manage IIoT manufacturing systems within an enterprise. These examples deliver partially enabling technologies but are not true CM systems. Cloud manufacturing has further capabilities made possible from sharing of manufacturing resources and capabilities which enable innovative, sustainable solutions.

Table 2
Selected cloud manufacturing case studies from the literature

Case study summarised and sector, where given, highlighted	Cloud manufacturing features the study highlights	Outcomes from CM approach	Future work required suggested by this paper and *Further work suggested by original authors in italics*	Author and year
A proposed service optimisation configuration method for manufacturing assembly lines. Sector not specified, but model would apply to any sector with assembly lines.	**Manufacturing as a service**: Servitisation of manufacturing resources. **Intelligent on-demand manufacturing**: Optimal configuration method of services based on grey relational analysis.	CM was used for proactive discovery of manufacturing resources through servitisation. Method finds satisfying solutions with low cost and time consumption.	Method does not consider environmental impacts when determining optimal solution. *Real-time manufacturing information processing mechanism between bottom-level manufacturing machines and manufacturing cells. Batch volume of manufacturing tasks should be taken into account in the optimal selection of service composition*.	2016, [20]
A model to assess the sustainable manufacturing capability (SMC) of industrial robots (IR), used in discrete manufacturing.	**Decision making tool**: Sustainability of different manufacturing routes analysed.	Cloud platform was able to predict energy consumption of IR for assessing their capability to perform a manufacturing task.	Limited to predicting the energy consumption of a manufacturing process only to the energy consumed by the IRs. Could be expanded to include other energy outputs, such as transport of materials. *More factors should be considered to comprehensively describe the SMC of IRs, such as collaboration safety with human.*	2017, [46]
An adaptive artificial bee colony algorithm is used for decision making on the cloud	**Flexibility and scalability, and decision making tool**: Feasibility of different manufacturing routes analysed.	Model includes the geographical location or services when quantifying the quality of service.	A limitation is the long time required by complex models. Extra computing power can be accessed through CM. Further work should investigate balancing cost of increased computational power with time saved.	2015, [47]

service selection and composition.			*Integration of artificial neural networks for the definition of maximum cycle number*	
A cloud platform is established for the polymer industry for optimising the processes of blending polymer materials.	**Decision making tool**: Data mining for optimal solution. Computer-aided process design based on CM. **Flexibility and scalability:** Flexible CM system built that responds to customers' demands.	The knowledge-based intelligent searching algorithm finished in 10 minutes compared to the 5 hours it took by traditional methods.	Case study only explored how CM can be applied to the design phase. Further work should investigate how effectively the model would respond to different customer demands during production phase. N/A	2016, [22]
Public cloud manufacturing system for SMEs in manufacturing equipment industry in South China.	**Multi-tenancy and intelligent on-demand manufacturing**: Performs intelligent multidimensional matching of consumers to providers.	Enterprises can respond to market requirements more quickly and cost effectively than before. Crowdsourcing model demonstrates how CM can be used to develop innovative designs.	Exploring how enterprises' private clouds will integrate with public cloud. *To apply CM to more manufacturing applications.* *To investigate business models for CM.*	2015, [37]
A CM architecture is developed to optimise the production logistics of a paint making company.	**Manufacturing as a service:** Real-time monitoring of resources using IIoT.	The cloud coordinates an improved schedule resulting in logistics waiting costs and overtime working reductions of approximately 50%.	Resources monitored are all manually controlled. Therefore lacking real-time execution of management system. *Expand qualitative solution presented in paper to the quantitative stage to achieve more accurate application results.*	2016, [48]

	Decision making tool: Enterprise resource planning			
A CM system is constructed within a conglomerate which designs and manufactures multi-disciplinary complex products in the discrete manufacturing sector.	**Manufacturing as a service and multi-tenancy:** Sharing of manufacturing recourses and capabilities through a hybrid framework integrating multiple manufacturing clouds.	Cloud manufacturing links enterprises so institutions learn to employ core competences of other institutions. This allows them to focus on their own core competences.	Proposed system does not consider what information should be private between different institutions or intellectual property issues when designing new products. *Domain-specific demands and characteristics of business collaboration to design efficient collaborative algorithms for any industry. Service-oriented architecture based integration methods of manufacturing resources and capabilities.*	2016, [21]

3. Cloud manufacturing as a route to sustainable manufacturing

From understanding CM's characteristics and applications, the methods and opportunities for how CM may be used to enhance sustainability is explored. Three pillars of sustainability were formalised for future sustainable development during the 2002 World Summit, these are economic development, social development and environmental protection [49]. These pillars often come into conflict with one another, as seen during the 20th century where economic development for a better quality of life was favoured at the expense of the environment [50].

3.1. Sustainability drivers and barriers

Manufacturers have traditionally focused on just the economic aspect of sustainable development [50], whereas external drivers are now forcing companies to adopt environmental sustainable development strategies. Five key drivers have been identified pushing manufacturers towards overall sustainability which are: (1) legislation; (2) customer demands; (3) shortage of natural resources; (4) population increase; and (5) business competitiveness.

Despite these drivers, there are still numerous barriers in the way of sustainable manufacturing. These barriers can be sorted into four categories: (1) policy and market barriers; (2) financial and economic barriers; (3) technical and information barriers; and (4) managerial and organisational barriers [51]. From these barriers a lack of information and lack of allocated resources (time and human) are the most common given responses preventing companies adopting sustainable strategies [52]. This lack of information includes areas such as environmental impact and expertise knowledge. Lack of allocated resources will also include materials alongside time and human.

3.2. Cloud manufacturing enabling collaboration for innovative designs

Cloud manufacturing offers various solutions to the barriers preventing sustainable manufacturing. Two of the barriers highlighted by companies is a lack of expertise knowledge available within an enterprise and ease of access to quantitative environmental impact information of manufacturing processes [52]. A characteristic of CM is that knowledge and data is collected, stored and shared throughout the supply chain. Those sharing data can determine what data, how much data, and with whom this data is shared. During the initial stages of a partnership between cloud users limited data sharing may take place, but as the relationship progresses data sharing may increase with trust.

A new concept of 'Social Manufacturing' has recently emerged where manufacturers and individuals work together in the production of physical goods [53]. Cloud manufacturing complements this concept through its ability to crowdsource information and simulate innovative designs [37]. The CM platform can be used to match developers to relevant expertise, and ubiquitous access to design information [26, 27]. In the long term this should increase the amount of collaborative designs leading to innovative solutions to sustainable manufacturing problems [54]. Intellectual property barriers will arise here and deciding what information should be made available and what should remain private will be a challenge but overall, CM benefits from a share-to-gain philosophy [31].

3.3. Greater automation from the adoption of cloud manufacturing systems

Automation is required to enable some of CM's characteristics. Automation is defined as "any activity that is performed by a machine or an algorithm with no human involvement" [55]. Automation of the production stage of manufacturing is not new. For CM to be effective it will need to integrate with other emerging technologies such as smart factories and IIoT, which work to further increase automation in the production stage. Recently the automatisation of the planning and scheduling of manufacturing activities has gained more attention as an alternative route to furthering the sustainability of manufacturing processes [55]. Lean manufacturing, just in time manufacturing, supply chain management, and other manufacturing paradigms are previous attempts at this [56]. Cloud manufacturing stands above these paradigms because of the unprecedented level of data integration across the enterprise. This will allow for the identification of previously unseen optimisation routes and implementation from a CM centralised management centre, ensuring factory floor and business decisions are simultaneously made in relationship to one another.

One advantage from the automatisation of the supply chain is greater process agility. Market turbulence, volatility and unpredictability are some of the biggest challenges facing manufacturers and agile supply chains are becoming vital in order to respond quickly to these changes [57]. Cloud manufacturing supply chains are highly reconfigurable, because of the increased emphasis on modularity [54]. This is achieved by the packaging of manufacturing resources and manufacturing capabilities as services sold through the manufacturing cloud. The flexibility and scalability offered allows CM to respond to market changes quicker due to the shorter supply contracts and provides greater customisation of products increasing the economic sustainability of manufacturing processes. For CM to be able to manipulate manufacturing resources it will rely on IIoT technologies integrated throughout the supply chain [48]. The cloud platform intelligently schedules the manufacturing process, meaning that the providers of CM services no longer must actively search for contacts. These features enable CM to become proactive rather than reactive when responding to market volatility. How CM will determine the optimal configuration and proactive resource discovery is a major challenge preventing its deployment and is the subject of much of the present research [20, 58, 59]. As enterprises begin to rely on a variety of service providers for different stages of the product cycle, methods to quantify and guarantee a quality of service will need to be developed [47].

Furthermore, advanced planning and scheduling shall reduce manufacturing time, energy and material waste. As seen from section 2.4 an extensive amount of the current literature has focused on developing methods by which the optimal manufacturing route can be discovered and implemented by CM. As a route to process sustainability this has great potential for minimisation of all types of manufacturing waste. However, what is yet to be explored is how CM can be harnessed for waste valorisation opportunities.

4. Application to process engineering manufacturing

The literature up until now has focused on defining CM and constructing possible architectures for how CM may be applied to industry. In the literature, the limitations

and benefits of CM have been discussed in context to the overall manufacturing industry. However, what has yet to be touched upon is that within manufacturing two distinct branches exist “discrete manufacturing” and “process manufacturing” [60]. Although there are several similarities between the two, there also exists significant differences. The application and benefits of CM to discrete manufacturing stand out because of discrete manufacturing’s assembly lines and their high customisability. For example, the automobile industry already offers consumers the option to customise their cars during production. It is simpler to visualise a network of separate factories making unique components to be assembled at one location and being coordinated through a centralised system, or alternatively 3D industrial printers producing specialised products unique to each customer. These concepts are harder to visualise for process manufacturing, as the product produced is not an assembly of different components but rather a combination of raw materials or “ingredients”.

4.1. Difference between discrete and process manufacturing

The differences between discrete manufacturing and process manufacturing is summarised in Table 3. The fundamental cause of these differences is within the nature of the manufacturing process itself. In discrete manufacturing a specific quantity of material moves between different processing equipment as it is cut and assembled together [60]. Whereas in process manufacturing, raw or waste materials flow through the process undergoing thermal, chemical and/or biochemical conversion. The product(s) is then separated out from the process stream. There can be overlap within the two branches of manufacturing. For example, in the food industry a bar of chocolate is a distinct item or product but results from process manufacturing.

The methods by which CM enhances sustainability in manufacturing is outlined in section 3 and these are broadly applicable to both discrete and process manufacturing. There also exists opportunities for applying CM to process manufacturing scenarios. However, with the research on CM focusing on discrete manufacturing examples (section 2.4), these opportunities have yet to be fully explored. This paper shall focus on two aspects where process manufacturing differs from discrete manufacturing, including the impact on process manufacturing performance. The two aspects are:

1. Waste generation and management;
2. Feedstock variability.

Table 3
Differences between discrete and process manufacturing

	Process manufacturing	Discrete manufacturing
Manufacturing practice	Thermal and/or biochemical/chemical conversion of raw materials into final product and by-products, which typically are then extracted and purified.	Assembling parts and components into a final product.
Example industries	Chemicals, food and drink, paint, polymer, pharmaceutical, steel, fast moving consumer goods (FMCG).	Automobiles, computers and electronics, furniture, textile, construction.
Quality control	Product typically must meet a specified purity and yield. During start up or shut down product produced will not meet these specifications and only wasteful material is produced.	Quality is assured via inspection of a statistical determined amount of products for defects. Start up and shut down is instantaneous and makes no difference to product quality.
Feedstocks	Raw materials, often in liquid/gaseous form. Variation in feedstock characteristics often present.	Discrete units/bill of materials. Feedstock characteristics often constant.
Products	Product is typically produced as a bulk quantity item, such as chemicals, polymers, pharmaceuticals, food stuffs, rather than distinct items. Quantified by volume/mass of product.	Distinct items, such as automobiles, toys, tools etc. Quantified by number of items produced.
Waste produced	Liquid and solid waste streams containing potentially valuable material residue, co-products, recoverable heat or water.	Defects, unused parts, cuttings etc.

4.2. Application of collaborative design and greater automation to process manufacturing

Traditionally manufacturing has been a linear model (make, use, dispose). However, due to diminishing resources and legislative drivers, a circular approach is being

encouraged [6]. Process manufacturing has a unique opportunity for development of innovative designs because of its ability to manipulate the manufacturing route to increase the potential to valorise waste and by-product streams. One such innovative solution has been developed by Lindhurst Engineering ltd called H^2AD Micro AD. The H_2AD is a hybrid of anaerobic digestion (AD) and microbial fuel cell technology [61]. It reduces pollutants in wastewaters and generates energy that can be utilised by the manufacturer to directly reduce energy bills. Unlike traditional AD, H^2AD operates on the microscale meaning each batch takes only three days and can be retrofitted to existing small manufacturing plants making it suitable for Small and Medium Enterprises (SMEs). This technology may be modified to treat wastewater streams from a variety of industries. By using CM as platform to share historical H^2AD performance data, manufacturers may utilise this data when modifying the H^2AD unit for treating a previously untested industry's wastewater.

Both discrete and process manufacturing benefit from the application of CM for automated supply chains, as detailed in section 3.3. However, the application of automation for process manufacturing is different to discrete manufacturing. Discrete manufacturing has focused on automating manufacturing lines, by the introduction of autonomous machines used to assemble products instead of human labour [62]. Process manufacturing requires a high degree of automation, monitoring and control, and were among the first industries to be automated [63]. Traditional automation involves sophisticated process control which operates based on data collected from temperature, pressure, flow rate and level sensor measurements from an individual process plant. Cloud manufacturing could help to make more informed decisions relating to remote processing plant's operations, through the collection and analysis of data from multiple plant sources. This is facilitated by CM's characteristic of being a centralised repository for multi-site data [23, 64].

Automation in the process manufacturing industry is progressing by the development of advanced, commercial sensors and analytics [65-67]. These sensors can collect data not previously available, enabling manufacturers to identify alternative sustainable manufacturing routes. An example is the development of a multi-sensor system based upon ultraviolet optical fluorescence imaging and ultrasonic acoustic sensors for Clean-In-Place (CIP) optimisation within the food and drink industry [68]. Existing CIP processes are time intensive and waste large amounts of energy, water, and chemicals as they operate on a worst-case scenario. This system aims to take the data recorded and use a machine learning approach to determine the optimal time to stop cleaning, saving both time and materials [68]. The savings made in time are significant, as the food industry spends on average 20% of the day cleaning which could instead be utilised for production [69]. Cloud manufacturing enables the data collected to be analysed remotely requiring no additional expertise within the enterprise. This is particularly beneficial to SMEs who may not have the resources for data analytics within the enterprise. Caution is needed that some element of human integration with safety and quality control practices is retained, rather than a complete reliance on artificial models. For instance, with the CIP example switching between producing different food products on the same manufacturing plant may result in the production of pathogens the sensors are not able to detect.

4.3. Cloud manufacturing for waste valorisation

The waste hierarchy was first introduced into European waste policy by Directive 75/442/EEC, emphasising the importance of waste minimisation, and the protection of the environment and human health [70]. It has been updated in Directive 2008/98/EC to introduce a five-step waste hierarchy that states which waste management methods have priority based on environmental impact and extended beyond waste minimisation to reuse, recycling and recovery (Fig. 3) [5]. Preventing and reducing the amount of waste generated is always the first option in optimising the sustainability of a manufacturing process. This is an option available to both discrete and process manufacturing. However, regardless of how effective this stage is, the manufacturing environment will always produce some waste. The circular economy model treats this waste generated as a resource to feed back into the economy, rather than a waste to be disposed of. It takes a cyclical view where waste is constantly fed back into the economy, and there is minimal input of new material entering the system.

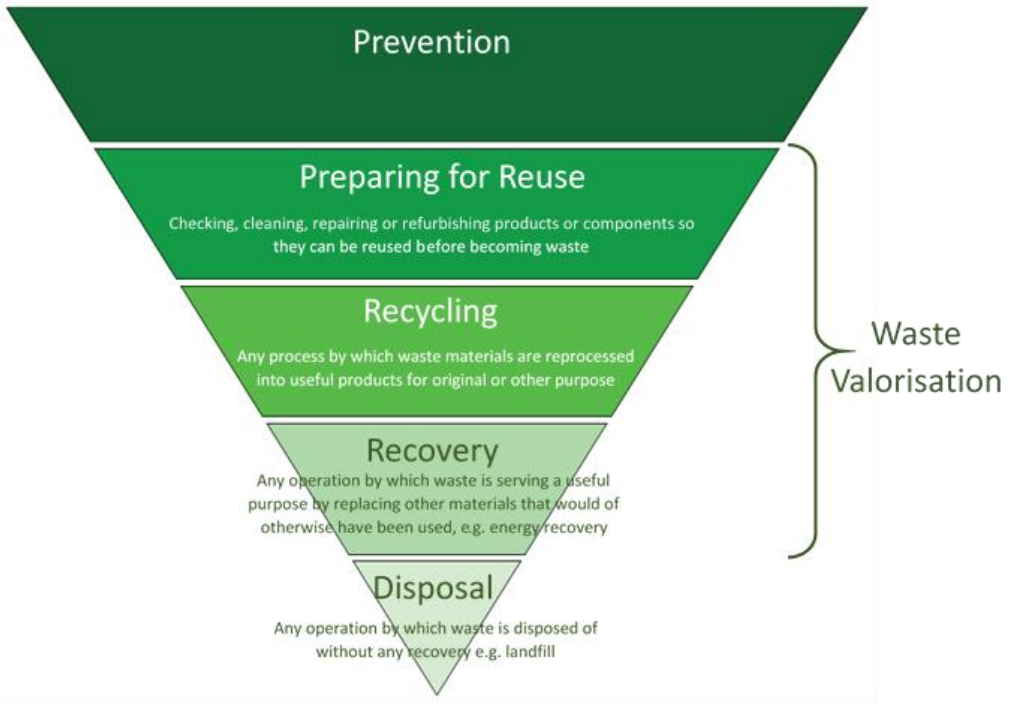


Fig. 3. Waste hierarchy adapted from Directive 2008/98/EC [5]

Discrete manufacturing produces mainly solid waste generated from defects, off cuttings, and unused parts during manufacturing [71]. Hence discrete manufacturing has limited waste valorisation opportunities, with instead a focus on waste minimisation and remanufacturing, which can be enhanced by CM [54]. Waste valorisation refers to industrial processing activities aimed at reusing, recycling, or recovering useful products or sources of energy from waste.

There are currently three waste valorisation strategies, geared towards: (1) production of fuel and/or energy to replace fossil fuels; (2) production of high-value chemicals from residues; and (3) production of other useful material(s) [72]. Process manufacturing generates a wide range of solid, liquid and gas waste streams that

contain significant quantities of valuable material, which may be recovered by the application of these strategies. Taking the food manufacturing industry as an example, its waste contains valuable compounds such as fats, proteins, bioethanol etc, which can be converted to soaps, rubber, fuels and more [73]. Furthermore, through biological, chemical and thermal processes, such as AD and pyrolysis, biogases and other gases can be produced as fuel or used for energy generation. Garcia-Garcia et al. effectively demonstrates how the waste hierarchy principles can be applied to selecting the most sustainable waste valorisation technology for treating food manufacturers' waste [73].

British Sugar's Wissington sugar production site is an excellent example of a manufacturing process that would not be economical sustainable without waste valorisation [74]. By reusing, recycling and recovering waste, British Sugar generates eleven co-products ranging from topsoil and animal feed to bioethanol and tomatoes [74]. This extensive waste valorisation has made British Sugar's Wissington factory one of the most efficient sugar factory in Europe and is an excellent example of the benefits of waste valorisation within process manufacturing [75].

Process manufacturing waste has a wide range of waste valorisation routes available [76-78]. As noted, a key characteristic of CM is the collection and sharing of data and expertise knowledge, which aids in the ability of CM to make intelligent, sustainable decisions. Cloud manufacturing can be used to evaluate the most sustainable valorisation waste method from this data. For example, in the production of olive-oil, solid waste pomace is produced. Traditionally pomace is a feedstock for production of pomace-olive oil and pomace wood, but there now exists several higher value products it can be converted to, including biomolecules, dyes and cosmetic products [79]. The flexible supply lines inherent within CM and real time analysis of market prices means that CM can be used to identify the most cost effective waste treatment method and co-orientate its implementation. This would be advantageous to SMEs who may not have their own waste treatment infrastructure on site but pay to have their waste taken away and treated by an external company.

As a multi-tenancy platform, CM can be used to match different cloud users together. Connecting and enhancing communication between local manufacturers will help to identify current inefficient waste treatment methods and identify local sustainable alternatives. As an example, CM may identify an alternative feed source for a process from another local manufacturer's waste, replacing a non-renewable source. This would be a preferred treatment option because it is higher up the EU waste hierarchy. However, it is important to recognise the importance of geographical location of manufacturing and waste treatment plants when constructing the manufacturing cloud [23]. The most sustainable waste treatment option may not be local, and the transportation of waste is both environmentally and economically costly. Therefore, there must be a balance between the best treatment option and feasibility to maximise overall sustainability. Cloud manufacturing also has the capability to identify suitable transportation routes and match suppliers and customers. It has the potential to utilise wasted transportation capacity, maximising the economic return for the transporter and decreasing CO_2 transport emissions from manufacturing. Cloud manufacturing could support a business model similar to the taxi service Uber where multiple consumers with a similar journey route share a single vehicle. This is an

expansion on the optimisation of supply chain scheduling CM offers, as discussed in section 3.3.

Real time monitoring of the composition of waste streams is required for CM to make intelligent decisions on how the waste should be utilised. For example, pyrolysis (the thermal decomposition of biomass in the absence of oxygen) is highly dependent on the moisture content of the feedstock [80]. At higher moisture contents, high levels of water are produced and at lower levels there is a risk that the process only produces dust instead of oil, char or gas [80]. Online measurement of the waste can feed back into the drying stage to ensure appropriate levels of drying takes place. Alternatively, in the design phase CM can be used to both access and analyse large volumes of stored process data to determine trends in moisture content of biomass waste. The conclusion may be reached to use the alternative biomass valorisation process hydrothermal carbonisation which operates in the presence of supercritical water. Hydrothermal carbonisation benefits from high moisture contents and may be judged to be the more sustainable route if a high moisture content trend is discovered [81]. Furthermore, if the waste stream is to be advertised through CM as a feedstock it will require its composition and properties to be defined to inform on the waste valorisation technology and where waste is considered a feedstock, considerations of the variability on the waste composition is required.

Process manufacturing has the potential for a wide range of waste valorisation methods. Cloud manufacturing can support this through three main methods:

1. It can be used as a decision tool for deciding the most sustainable waste valorisation routes, ideally the highest up the waste hierarchy to meet EU Directive;
2. It can identify and coordinate alternative waste valorisation routes if they prove to become more sustainable;
3. Cloud manufacturing can link and coordinate local supply chains to reuse waste streams as a feedstock, with minimal treatment.

4.4. How cloud manufacturing increases process resilience

Process manufacturing is reliant on process conditions that define how the feedstock is utilised and how equipment is operated in order to deliver a product at a predefined specification. The process conditions include but are not limited to process streams composition, temperature, pressure and flowrates. The nature of process manufacturing products having to meet a specific specification relies on a need for strong process resilience. There exist examples of process manufacturing goods being recalled because they did not meet specifications or face fines for breach of environmental limits [82]. This is also present in discrete manufacturing, though it is less common due to it being less susceptible to variations in manufacturing conditions. The current practise to ensure process resilience is to account for variability in the design stage. Section 4.4 will explore how CM enhances current methods, as well as offering new solutions.

A crude but effective method of ensuring process resilience is to design the manufacturing plant to operate within the design envelope. The design envelope of a system refers to its specified requirements that, if satisfied, would enable the system

to perform as intended in the operational environment [83]. This has the disadvantage of increased capital and operating costs. If the design envelope is underestimated the process plant shall fail, with a best case scenario resulting in off spec product and at worst a fatal incident [84]. Certain process conditions are outside the control of the engineers and technicians. Feedstock supply lines and composition is a critical process condition that is subject to external variables. As manufacturers move away from non-renewable sources towards waste and biomass feedstocks this will become an even greater challenge for future engineers [85]. This is again an example of how the large knowledge and data store available through CM would benefit the design stage of manufacturing. Advanced data analysis techniques available on the manufacturing cloud would be able to predict and model to a higher degree of accuracy the future trends of feedstock variations [86]. This allows for the design envelope to be reduced while still maintaining a safe operating range and increasing economic sustainability. During the plant operation, the reconfigurable supply lines inherent within CM will allow for easy identification and implementation of alternative local feedstock sources. This can help mitigate the risk of using waste and biomass sources as feedstocks in process manufacturing operations. Furthermore, advanced data analysis of historical manufacturing data can help designers gain a better understanding of the process and products. It is more effective to ensure quality in the products by building quality into the design rather than continued testing and adjustment of critical process parameters.

It has been highlighted that a current barrier in process manufacturing that prevents enterprise-wise optimisation is the lack of data integration and supply chain coordination across the enterprise and between enterprises [87]. Qin explains how data in process manufacturing is currently collected based on a hierarchy structure [34]. The bottom levels are the monitoring of equipment via sensors that happen virtually instantaneously. At the process level there is regular monitoring of process control data (temperature, pressure and flowrate), with product quality more often sampled irregularly and maybe off site. The top level is the customer feedback data. There is a significant time lag between the data at higher levels and its implementation to the manufacturing resources. Cloud manufacturing architecture is a data-centric system that ensures data is made available across all manufacturing and business levels [23]. Rather than having a hierarchy structure, it shall be circular with data from across the supply chain feeding back into the manufacturing process. This shall reduce the time lag and better inform on the process to find the optimal configuration.

Cloud manufacturing will need to rely on big data technologies to achieve this. Big data is the collection and analysis of vast volumes of data throughout the manufacturing process to reveal trends, patterns and process optimisation routes [86]. Currently process manufacturing already collects a large amount of data from process control systems but it is not fully utilised, often only used for emergency after the event analysis [34]. Cloud manufacturing's reliance on advanced sensing technologies and widespread communication technologies will increase the amount and variety of data generated in manufacturing systems [88]. Data gathered from the manufacturing process could be used for

1. data mining and modelling of time series data for trend spotting and forecasting;

2. online use for real time operations optimisation and intelligent decision making;

3. offline troubleshooting as a valuable tool for continuous improvement.

Multivariate statistical approaches have been the recent favourite choice for data analysis, but new developments in machine learning and data mining offer opportunities to gain value from large, unclean data sets [34]. Machine learning advances in closed-loop optimisation may be used by CM to continuously optimise process control online. Houben and Lapkin summarise the recent advances in automatic optimisation of process conditions and explain how machine learning can build self-optimising algorithms [89]. By achieving tight process control the process plant will be able to operate closer to steady state process conditions despite variations in external conditions, lowering the energy and material consumption. Enterprises often lack effective analytic tools or computational power to get any real meaning out of this data. The service nature of CM allows enterprises to hire either extra computational processing power or outsource their data analysis to specialist data analysis companies on a pay-as-you-go business model.

4.5. Examples of cloud manufacturing's application to process manufacturing

To further reinforce how CM may be applied to process manufacturing, examples are presented in Table 4 for each of CM's unique characteristics applied to different industries.

Table 4
Examples of cloud manufacturing's application to process manufacturing industries.

Cloud manufacturing characteristic	Application to process manufacturing	Industrial examples
Flexibility and scalability	Continued monitoring of real time economic data to ensure the optimal manufacturing route.	Two adsorbents may be used for the recovery of CO_2 from flue gas by pressure swing adsorption, Zeolite and Activated Carbon [90]. During the lifetime of the plant these adsorbents may need to be replaced. Cloud manufacturing can be utilised to reassess which adsorbent is most sustainable, using the most recent economic and performance data.
Knowledge intensive	Utilise data across the supply chain to build an artificially intelligent (AI) algorithm to optimise process conditions.	The active component for all anti-malarial drugs is artemisinin and is harvested from a plant called *Artemisia annua* [91]. When the plant is harvested impacts the manufacturing plant's optimal process conditions to improve the artemisinin recovery [91]. Utilising IIoT technologies during growing and harvesting, CM system would integrate this data downstream to more accurately predict the required process conditions.
	More accurate pricing of feedstocks, wastes and products.	When a batch of potatoes arrives at a factory for the production of crisps, samples are taken to assess the quality of the potatoes and determine their value [92]. These tests can prove unreliable and new tests which utilise artificial neural networks are being developed [92]. By uploading these results to the manufacturing cloud, CM can be used to track various deliveries and coordinate accurate payments to each potato supplier.
Multi-tenancy	Co-ordination of multiple manufacturers' waste streams within geographical region.	Brewer Spent Gain (BSG) in a wet solid waste from the brewery process which has a wide variety of valorisations options [93]. Individually microbreweries do not produce enough BSG to make it feasible to valorise. However, CM could be used to coordinate the collection of BSG to enable one of these technologies to be successfully implemented by a third party.

	Sharing AI algorithms	An algorithm created to optimise the treatment of a wastewater by anaerobic digesters may be sold through CM to the benefit of the initial developer and other water treatment companies. The algorithm would be built upon valuable performance data that the developer would not want its competitor accessing. Cloud manufacturing acts as an impartial third party able to utilise performance data to retrain and optimise the algorithm, without disclosing eithers manufacturers' data.
On-demand	Customisation of products	The data generated by wearable 'lifestyle' technologies has the potential to revolutionise the method by which patients are diagnosed [94]. This data may also be used to determine an individual's required dosage. Cloud manufacturing would act as a platform sharing data between manufacturers, doctors and patients to enable personalised medicine.
	Responding to changes in customer's demands	Demand side management is the modification of consumer demand of energy, and is used to attempt to limit the peak demand on an electrical grid [95]. However, the energy sector is facing the challenge of trying to utilise unreliable renewable energy sources and an increase in customer demand [95]. One of the largest consumers of energy are process manufacturers. Cloud manufacturing could help limit the demand on the grid during peak times by increasing the communication between energy providers and manufacturers to schedule the most energy demanding processes when there is minimum demand on the electrical grid.
Manufacturing as a service	Outsourcing expertise	Affordable IIoT technologies are currently been developed aimed at automating the brewery process in microbreweries, generating new data sets [96]. Microbreweries have limited employees and do not have in-house data analyst to effectively make use of this data. Cloud manufacturing can be used to outsource their data analysis needs.
	Outsourcing parts of the supply chain	In the dairy industry whey is a by-product of cheese production, and has a wide variety of valorisation options [97]. Large scale cheese manufacturers normally valorise the

	whey on-site. Small independent cheese makers have limited capital to invest in these technologies to valorise on site, and instead the whey is treated as a waste. Through CM they can outsource the valorisation of the whey to either waste management companies or local large scale cheese manufacturers.

5. Limitations and future research directions

The literature to date has identified limitations that prevent widespread adoption of CM by manufacturers from the discrete manufacturing perspective. These limitations being identified as data sharing, cyber security, integration of technologies, modelling of material flow, and integrating existing systems within a CM system [42, 48, 64, 98]. However, these limitations are also relevant to process manufacturing, for instance data sharing among process manufacturers (Section 5.1). There are also limitations applicable only to process manufacturing, which are processing plant security and characterisation of process streams (Section 5.2 and 5.3 respectively). Potential limitations may also be industry specific within process manufacturing, when the cloud user is an SME (Section 5.4), though CM will also afford unique opportunities to SMEs [38].

5.1. Data sharing among process manufacturers

The current barriers obstructing free flow of data between enterprises have already been thoroughly discussed [98]. However, data sharing is important to elucidate on from a process manufacturing perspective. Large data sets are required to train and optimise the AI algorithms that CM will use [99]. As previously stated, the large process manufacturers were the first industries to be automated and thus have generated large quantities of data [63]. However, new companies and SMEs may not have access to enough data to build, train and validate the required mathematical models. Cloud manufacturing can overcome this by sharing either already built algorithms or manufacturing data from similar processes to act as a starting point in the construction of these algorithms. Manufacturers are naturally reluctant to share their data between competitors. A framework needs to be developed that will allow the manufacturing cloud to act as a third party to utilise another company's manufacturing data and/or models, without disclosing this information to the competing company. A pricing mechanism will also be required to assess and judge the value of the data used.

5.2. Processing plant security

Unlike discrete manufacturing, process manufacturing involves the thermal, chemical and or biochemical conversion of materials to products at temperatures and pressures which can be hazardous. Some industrial reactions are exothermic and there is the potential for runaway reactions if not controlled. This is a danger unique to process manufacturing industries and there are many reported cases where this has occurred and resulted in causalities [100]. Process plant safety has evolved in past 20-30 years, moving from corrective to preventive. The emphasis is on designing out hazards and where this is not possible, process control systems are used to monitor temperatures and pressures within a reactor ready to shut off if they reach dangerous levels. However, as with any use of online systems, the potential for hacking needs evaluating with risk mitigation. For process manufacturing this is a legitimate consideration and one not relevant to discrete manufacturing.

As manufacturing resources are increasingly controlled remotely there exist safety concerns protecting these control systems from malicious attacks. The "Aurora Project", performed by Idaho National Laboratory in 2007, first demonstrated the dangers from cyber criminals hacking into a process plant's control system to

manipulate and ultimately destroy physical systems [101]. Sophisticated cyber-attacks on control systems are becoming more frequent, resulting in the need for advanced cyber-security systems [102].

5.3. Characterisation of process streams

As highlighted in section 4.1, a fundamental difference between discrete and process manufacturing is the variability in process manufacturing's process streams. Be these feedstocks, wastes or by-product streams. For CM to be able to make intelligent decisions, it will require the characterisation of these process streams in real time, compared to the current practise of infrequent sampling and analysis off line [34]. This is because the AI algorithms that will be developed to support CM's decision-making ability will rely on real time data to ensure they are continuously learning. These AI algorithms must be capable of learning, so they can continually work to optimise all aspects of the manufacturing process and adapt to variations in process streams. Process manufacturing is currently limited from further automation by the development of commercially affordable sensors capable of characterising process streams in real time.

5.4. Barriers preventing SMEs adopting new technologies

Small and medium enterprises are considered the backbone of a country's economy, providing the majority of all new jobs and vital for stimulating economic growth [103]. Within process manufacturing certain industries have a strong SME portfolio, for instance the food and drink industry [104]. The wealth of SMEs are seen as a benefit because of the high value artisan products they produce, for example bakery products, cheeses and craft beers. There is strong growth within these industries as customers demand quality over quantity and are prepared to pay for this. The brewery industry in the UK is an example of this, observing a large increase in microbreweries despite a downward trend in beer consumption [105]. However, SMEs can struggle to adapt to external market forces and to make the next step in expanding their business due to a lack of expertise and resources. This is where CM shows great potential in enabling SMEs to overcome these barriers [38].

Small and medium enterprises can be slow in adopting new technologies, which are shown to improve revenue. They do not always have either the capital to invest in new technologies, or the expertise or guidance necessary to successfully implement them. The countries where SMEs have been most successful in adopting new technologies are countries which have coherent Government strategies in place to support them, for example in Germany (Industrie 4.0), China (Made in China 2025), and the USA (America Makes) [106]. This has resulted in higher levels of productivity (output per worker) in these countries over others, like the UK, which have lower levels of adoption [106]. If CM is to follow this example there needs to be a strong, coherent strategy in place to advice SMEs on how best to utilise CM for their needs.

6. Future of Manufacturing

Both government and academic literature agree that future manufacturing will evolve more sustainable manufacturing processes and business models radically different from those of today [56, 107]. A growing global population and the effects of climate changes shall cause volatility in supply chains, driving the development of

manufacturing processes that are more resilient and require less material, energy, water and land inputs. Resource insecurity shall become a greater threat in regions like the EU where rare earth minerals are depleted and there is a heavy reliance on imports. This will force economies to rebalance towards a circular economy, where no waste is created but materials are instead reused, remanufactured, cascaded, recycled and recovered [108]. The global community is beginning to take active steps towards closing the loop and becoming a circular economy [109-111]. Advances in waste valorisation and remanufacturing technologies will bring society closer towards achieving this, and CM offers a route to facility and enhance the circular economy.

Increased communication between manufacturers and customers, coupled with on-demand manufacturing, enables greater customisation of products to meet individual customer's needs. Customisation allows consumers the ability to create more personalised products and research by Loughborough University shows consumers are willing to pay an additional 10% for this service [112]. Customisation has great marketing potential allowing for customisation of products that reflect the event of the day. To enable effective customisation, manufacturing process must become flexible and agile to respond to customers' demands in a timely manner. To provide customised products manufacturers need more diverse resources and manufacturing capabilities, which CM can provide [113].

Additive manufacturing has real potential for customisation. At present its main application is printing geometrical complex components for high-value products in the aerospace, high-end automotive and bio-medical sectors [12]. This is because it is currently much faster to mass produce products using traditional production lines than AM. However, as the technology improves, and greater value is placed on personalised products, AM will shape the size and distribution of future factories. In one possible scenario there will be a mix of large super factories producing complex products, and small local factories producing personalised domestic products [107]. In 2015 Amazon filed a patent for lorries carrying 3D printers that will take orders online and produce the finished product by the time it reaches the customers door [114]. Within the process sectors, AM of food and pharmaceuticals products have recently made progress in overcoming the unique challenges these industries face in using 3D printing products [115, 116]. Additive manufacturing allows the product to be tailored for individual consumer's requirements. For example, there is research into the customisation of nutrition within 3D printed food products based on an individual dietary requirements for weight lost and further health benefits [117]. The application of CM is a feasible model to deploy large scale personalised AM [19, 118]. Cloud manufacturing supports increased communication between customers and manufacturers and may act as an interface for customers to upload their personalised requirements into the manufacturing process.

7. Conclusion

For the first time a detailed explanation of all of CM's key characteristics has been summarised. Recent case studies demonstrating the application of these characteristics in a CM system has been citied. The benefits and drawbacks of each study has been summarised, as well as the future work required to further develop these systems. Presented are examples of applying CM characteristics to process manufacturing industries. By understanding these characteristics and how they can be

applied to manufacturing, it enabled the exploration of how CM may be used to enhance sustainability. Four key methods of how CM increases sustainability have been identified: (1) collaborative design; (2) greater automation; (3) increase in process resilience; and (4) enhanced waste reduction, reuse and recovery (see Fig. 4).

As resource scarcity increases and climate change disrupts global supply chains, there will be an ever-increasing need for manufacturing processes that utilise waste valorisation strategies rather than relying on fresh material and energy inputs. As a model for implementing a circular economy approach in manufacturing, CM provides opportunities for both discrete and process manufacturing to improve sustainability. Cloud manufacturing research gained attention with the early models developed by Tao et al. and Xu, and the application of CM for sustainable process manufacturing complements both these models [14, 15]. Recent developments in technologies that are digitalising manufacturing have seen a rush of research into investigating the opportunities they offer. However, a framework for the infrastructure these different technologies shall operate within is currently lacking. Cloud manufacturing has the potential to act as an intelligent platform to interact remotely with the increasingly digitalised supply chain. This complete monitoring across the supply chain allows CM to identify, coordinate and implement opportunities for increased sustainability. This could not have been realised by the current manufacturing model where enterprises within a supply chain act with greater independence of each other. Before CM can be fully realised, examples of its operation in real world manufacturing models and the benefits it delivers will need to be demonstrated.

(1) **Collaborative design:** A critical barrier preventing sustainable manufacturing designs is the accessibility of environmental information and expertise knowledge within an enterprise. Cloud manufacturing operates on a multi-tenancy cloud platform that can crowdsource information and expertise to simulate the development of innovative sustainable designs. Process manufacturing has unique opportunities for new sustainable design from its ability to manipulate the manufacturing route to valorise waste and by-product streams.

(2) **Automation:** Cloud manufacturing can perform real time analyses of manufacturing and market data to intelligently schedule and plan the manufacturing process. This allows for CM to optimise the manufacturing process in real time and coordinate its implementation across the supply lines from a centralised management centre. This reduces the manufacturing costs through minimising energy consumption, material waste and downtime. Process manufacturing will become further automated by the implementation of advanced sensors, allowing CM to develop intelligent models. These models help to identity and implement more sustainable manufacturing solutions.

(3) **Waste valorisation:** Process manufacturing waste has the potential for valorisation as it can contain valuable material residue, co-products opportunities, residual heat, and energy recovery. It is not always obvious which is the most sustainable waste valorisation route, and CM can be used as an intelligent decision making tool. CM can monitor functions in the composition of waste streams and market prices to identify and coordinate alternative waste valorisation routes if they prove to become significantly more sustainable. Furthermore, CM can coordinate local supply chains to reuse waste streams as a feedstock. From these methods, CM aligns with the circular economy vision perceiving waste as a valuable resource to be reused, recovered and regenerated.

(4) **Process resilience:** Process manufacturing systems are particularly susceptible to variations in process conditions, especially feedstock composition. Currently over engineering is the best method to ensure variations do not impact product quality. Advanced data analytic techniques offer the potential to gain better process understanding to ensure quality in the design of the system. As advanced IIoT sensors are developed, this will provide CM with greater quantity and granularity of data meaning previously unidentifiably solutions can be implemented.

Fig. 4. Summary of how Cloud Manufacturing enables sustainable manufacturing

Acknowledgements

The author (RLG) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing. The author (OJF) would also like to acknowledge the University of Nottingham for his PhD scholarship. All authors would like to thank David Golightly of Nottingham University for his contributions and useful insights.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161].

References

[1] World Trade Organization. World Trade Statistical Review 2016, https://www.wto.org/english/res_e/statis_e/wts2016_e/wts16_toc_e.htm; 2016 [accessed 10/04/2017]
[2] HM Treasury. Budget 2011. The Stationery Office; London; 2011. p. 7-9.
[3] The World Bank. Manufacturing value added (% GDP), http://data.worldbank.org/indicator/NV.IND.MANF.ZS; 2017 [accessed 10/04/2017]
[4] UNFCCC Secretariat. Adoption of the Paris Agreement. In: Conference of the Parties. United Nations. New York. 2015. p. 22-4.
[5] Directive 2008/98/EC of the European Parliament and of the Council of 19 November 2008 on waste and repealing certain directives. Off J Eur Union L 312/3 2008.
[6] Waste & Resources Action Programme (WRAP). WRAP and the circular economy, http://www.wrap.org.uk/about-us/about/wrap-and-circular-economy; 2017 [accessed 13/04/2017]
[7] Mahesh M, Ong SK, Nee AYC, Fuh JYH, Zhang YF. Towards a generic distributed and collaborative digital manufacturing. Robot Comput-Integr Manuf 2007;23:267-75.
[8] Montreuil B, Frayret JM, D'Amours S. A strategic framework for networked manufacturing. Comput Ind 2000;42:299-317.
[9] Mehrabi MG, Ulsoy AG, Koren Y. Reconfigurable manufacturing systems: Key to future manufacturing. J Intell Manuf 2000;11:403-19.
[10] Da Silveira G, Borenstein D, Fogliatto FS. Mass customization: Literature review and research directions. Int J Prod Econ 2001;72:1-13.
[11] Li BH, Zhang L, Wang SL, Tao F, Cao JW, Jiang XD, Song X, Chai XD. Cloud manufacturing: a new service-oriented networked manufacturing model. Comput Integr Manuf 2010;16:1-7.
[12] Gao W, Zhang Y, Ramanujan D, Ramani K, Chen Y, Williams CB, Wang CCL, Shin YC, Zhang S, Zavattieri PD. The status, challenges, and future of additive manufacturing in engineering. Comput Aided Des 2015;69:65-89.
[13] Ministry of Science and Technology of the People's Republic of China. National High-tech R&D Program, http://www.most.gov.cn/eng/programmes1/200610/t20061009_36225.htm; 2016 [accessed 10/04/17]
[14] Tao F, Zhang L, Venkatesh VC, Luo Y, Cheng Y. Cloud manufacturing: a computing and service-oriented manufacturing model. In: Proc Inst Mech Eng Part B-J Eng Manuf. 2011. p.1969-76.

[15] Xu X. From cloud computing to cloud manufacturing. Robot Comput-Integr Manuf 2012;28:75-86.
[16] Manucloud-project.eu. Project Overview - The ManuCloud Project, http://www.manucloud-project.eu/; 2016 [accessed 05/12/16]
[17] Plex. Cloud ERP Software for Manufacturing, https://www.plex.com/; 2017 [accessed 10/04/2017]
[18] Qin J, Liu Y, Grosvenor R. A categorical framework of manufacturing for Industry 4.0 and beyond. Procedia CIRP 2016;52:173-8.
[19] Liu KH, Zhong PS, Zeng QL, Li DM, Li SS. Application modes of cloud manufacturing and program analysis. J Mech Sci Technol 2017;31:157-64.
[20] Zhang YF, Xi D, Li R, Sun SD. Task-driven manufacturing cloud service proactive discovery and optimal configuration method. Int J Adv Manuf Technol 2016;84:29-45.
[21] Yang C, Shen WM, Lin TY, Wang XB. A hybrid framework for integrating multiple manufacturing clouds. Int J Adv Manuf Technol 2016;86:895-911.
[22] Qiu XM, He G, Ji X. Cloud manufacturing model in polymer material industry. Int J Adv Manuf Technol 2016;84:239-48.
[23] Guo L. A system design method for cloud manufacturing application system. Int J Adv Manuf Technol 2016;84:275-89.
[24] Zhou J, Yao X. Hybrid teaching–learning-based optimization of correlation-aware service composition in cloud manufacturing. Int J Adv Manuf Technol 2017;91:3515-33.
[25] Jin H, Yao X, Chen Y. Correlation-aware QoS modeling and manufacturing cloud service composition. J Intell Manuf 2017;28:1947-60.
[26] Schaefer D, Thames JL, Wellman RD, Wu DZ, Yim S, Rosen DW. Distributed collaborative design and manufacture in the cloud - motivation, infrastructure, and education. In: 2012 ASEE annual conference. 2012.
[27] Wu DZ, Thames JL, Rosen DW, Schaefer D. Enhancing the product realization process with cloud-based design and manufacturing systems. J Comput Inf Sci Eng 2013;13:0410041-04100414.
[28] Wu D, Greer MJ, Rosen DW, Schaefer D. Cloud manufacturing: Strategic vision and state-of-the-art. J Manuf Syst 2013;32:564-79.
[29] Zhang L, Luo YL, Tao F, Li BH, Ren L, Zhang XS, Guo H, Cheng Y, Hu AR, Liu YK. Cloud manufacturing: a new manufacturing paradigm. Enterp Inf Syst 2012;8:167-87.
[30] WordArt. Word cloud generator, https://www.wordart.com; 2017 [accessed 10/04/2017]
[31] Wu DZ, Greer MJ, Rosen DW, Schaefer D, Asme. Cloud manufacturing: drivers, current status, and future trends. In: Proceedings of the ASME 8th international manufacturing science and engineering conference. 2013. p.V002T02A3.
[32] Wang P, Gao RX, Fan ZY. Cloud computing for cloud manufacturing: benefits and limitations. J Manuf Sci Eng-Trans ASME 2015;137:
[33] MFG.com. About MFG.com, https://www.mfg.com/about-mfgcom; 2016 [accessed 01/03/2017]
[34] Qin SJ. Process data analytics in the era of big data. AIChE J 2014;60:3092-100.
[35] Zhong RY, Lan SL, Xu C, Dai QY, Huang GQ. Visualization of RFID-enabled shopfloor logistics big data in cloud manufacturing. Int J Adv Manuf Technol 2016;84:5-16.
[36] Gao J, Yao YL, Zhu VCY, Sun LY, Lin L. Service-oriented manufacturing: a new product pattern and manufacturing paradigm. J Intell Manuf 2011;22:435-46.
[37] Ren L, Zhang L, Tao F, Zhao C, Chai XD, Zhao XP. Cloud manufacturing: from concept to practice. Enterp Inf Syst 2015;9:186-209.

[38] Huang BQ, Li CH, Yin C, Zhao XP. Cloud manufacturing service platform for small-and medium-sized enterprises. Int J Adv Manuf Technol 2013;65:1261-72.
[39] Wu DZ, Thames JL, Rosen DW, Schaefer D. Towards a cloud-based design and manufacturing paradigm: looking backward, looking forward. In: Proceedings of the ASME international design engineering technical conferences and computers and information in engineering conference. 2012. p.315-28.
[40] Ren L, Cui J, Wei YC, Laili YJ, Zhang L. Research on the impact of service provider cooperative relationship on cloud manufacturing platform. Int J Adv Manuf Technol 2016;86:2279-90.
[41] Wu DZ, Rosen DW, Wang LH, Schaefer D. Cloud-based manufacturing: old wine in new bottles? In: Variety management in manufacturing: proceedings of the 47th CIRP conference on manufacturing systems. 2014. p.94-9.
[42] He W, Xu LD. A state-of-the-art survey of cloud manufacturing. Int J Comput Integr Manuf 2015;28:239-50.
[43] Lu YQ, Xu X, Xu J. Development of a hybrid manufacturing cloud. J Manuf Syst 2014;33:551-66.
[44] Siemens. MindSphere – Siemens Cloud for Industry, http://w3.siemens.co.uk/services/uk/en/portfolio/plant-data-services/cloud-for-industry/pages/default.aspx#; 2017 [accessed 10/04/2017]
[45] IBM. Journey to Industry 4.0 and beyond with IBM Watson IoT for connected manufacturing, https://www.ibm.com/internet-of-things/iot-solutions/iot-manufacturing/; 2017 [accessed 10/01/2017]
[46] Zhao Y, Liu Q, Xu W, Wu X, Jiang X, Zhou Z, Pham DT. Dynamic and unified modelling of sustainable manufacturing capability for industrial robots in cloud manufacturing. Int J Adv Manuf Technol 2017;93:2753-71.
[47] Lartigau J, Xu XF, Nie LS, Zhan DC. Cloud manufacturing service composition based on QoS with geo-perspective transportation using an improved artificial bee colony optimisation algorithm. Int J Prod Res 2015;53:4380-404.
[48] Qu T, Lei SP, Wang ZZ, Nie DX, Chen X, Huang GQ. IoT-based real-time production logistics synchronization system under smart cloud manufacturing. Int J Adv Manuf Technol 2016;84:147-64.
[49] United Nations. Report of the world summit on sustainable development. Johannesburg. South Africa. New York. 2002. p. 14-5.
[50] Giret A, Trentesaux D, Prabhu V. Sustainability in manufacturing operations scheduling: A state of the art review. J Manuf Syst 2015;37:126-40.
[51] Shi H, Peng SZ, Liu Y, Zhong P. Barriers to the implementation of cleaner production in Chinese SMEs: government, industry and expert stakeholders' perspectives. J Clean Prod 2008;16:842-52.
[52] Bey N, Hauschild MZ, McAloone TC. Drivers and barriers for implementation of environmental strategies in manufacturing companies. CIRP Ann Manuf Technol 2013;62:43-6.
[53] Hamalainen M, Karjalainen J. Social manufacturing: When the maker movement meets interfirm production networks. Bus Horizons 2017;60:795-805.
[54] Bi ZM, Wang L. Manufacturing paradigm shift towards better sustainability. In: Li W, Mehnen J (editors.) Cloud manufacturing: distributed computing technologies for global and sustainable manufacturing, London: Springer London; 2013, p. 99-119.
[55] Monfared MAS, Yang JB. Design of integrated manufacturing planning, scheduling and control systems: a new framework for automation. Int J Adv Manuf Technol 2007;33:545-59.
[56] Esmaeilian B, Behdad S, Wang B. The evolution and future of manufacturing: A review. J Manuf Syst 2016;39:79-100.

[57] Jassbi J, Orio Gd, Barata D, Barata J. The impact of cloud manufacturing on supply chain agility. In: 2014 12th IEEE International conference on industrial informatics (INDIN). 2014. p.495-500.
[58] Yu C, Zhang W, Xu X, Ji Y, Yu S. Data mining based multi-level aggregate service planning for cloud manufacturing. J Intell Manuf 2015;1-11.
[59] Cao Y, Wang SL, Kang L, Li CS, Guo L. Study on machining service modes and resource selection strategies in cloud manufacturing. Int J Adv Manuf Technol 2015;81:597-613.
[60] Brandl D. Manufacturing control. In: Design patterns for flexible manufacturing, USA: ISA; 2007, p. 1-8.
[61] Micro H2AD. Welcome to H2AD, http://h2ad.org.uk/; 2017 [accessed 04/08/2017]
[62] Lastra JLM, Delamer IM. Ontologies for production automation. In: Dillon TS, Chang E, Meersman R, Sycara K (editors.) Advances in web semantics I: ontologies, web Services and applied wemantic web; 2008, p. 276-89.
[63] Canavas C. History of control and automation in chemical engineering. Chemie Ingenieur Technik 1995;67:709-17.
[64] Helo P, Suorsa M, Hao Y, Anussornnitisarn P. Toward a cloud-based manufacturing execution system for distributed manufacturing. Comput Ind 2014;65:646-56.
[65] Saguy IS, Singh RP, Johnson T, Fryer PJ, Sastry SK. Challenges facing food engineering. J Food Eng 2013;119:332-42.
[66] Yeh ML, Chang CT. An automaton-based approach to evaluate and improve online diagnosis schemes for multi-failure scenarios in batch chemical processes. Chem Eng Res Des 2011;89:2652-66.
[67] Kumar A, Baldea M, Edgar TF. On optimal sensing and actuation design for an industrial scale steam methane reformer furnace. AIChE J 2016;62:3225-37.
[68] Simeone A, Watson N, Sterritt I, Woolley E. A multi-sensor approach for fouling level assessment in clean-in-place processes. Procedia CIRP 2016;55:134-9.
[69] Jude B, Lemaire E. How to optimize clean-in-place (CIP) processes in food and beverage operations. Schneider Electric White Paper; 2013.
[70] Directive 75/442/EEC of the European Parliament and of the Council of 15 July 1975 on waste Off J Eur Union L 442. 1975.
[71] Lieder M, Rashid A. Towards circular economy implementation: a comprehensive review in context of manufacturing industry. J Clean Prod 2016;115:36-51.
[72] Arancon RAD, Lin CSK, Chan KM, Kwan TH, Luque R. Advances on waste valorization: new horizons for a more sustainable society. Energ Sci Eng 2013;1:53-71.
[73] Garcia-Garcia G, Woolley E, Rahimifard S. A framework for a more efficient approach to food waste management. Int J Food Eng 2015;1:65-72.
[74] British Sugar. How our factory operates, http://www.britishsugarlearningzone.com/how-our-factory-operates/#; [accessed 16/02/2017]
[75] Short SW, Bocken NMP, Barlow CY, Chertow MR. From Refining Sugar to Growing Tomatoes. J Ind Ecol 2014;18:603-18.
[76] Ferrasse JH, Seyssiecq I, Roche N. Thermal gasification: a feasible solution for sewage sludge valorisation? Chem Eng Technol 2003;26:941-5.
[77] Briens C, Piskorz J, Berruti F. Biomass valorization for fuel and chemicals production: a review. Int J Chem Reactor Eng 2008;6:

[78] Uçkun Kiran E, Trzcinski AP, Ng WJ, Liu Y. Bioconversion of food waste to energy: a review. Fuel 2014;134:389-99.
[79] Valta K, Aggeli E, Papadaskalopoulou C, Panaretou V, Sotiropoulos A, Malamis D, Moustakas K, Haralambous KJ. Adding value to olive oil production through waste and wastewater treatment and valorisation: the case of Greece. Waste Biomass Valorization 2015;6:913-25.
[80] Westerhof RJM, Kuipers NJM, Kersten SRA, van Swaaij WPM. Controlling the water content of biomass fast pyrolysis oil. Ind Eng Chem Res 2007;46:9238-47.
[81] Funke A, Ziegler F. Hydrothermal carbonization of biomass: a summary and discussion of chemical mechanisms for process engineering. Biofuels, Bioprod Biorefin 2010;4:160-77.
[82] Food Standards Agency. Food law prosecutions, https://www.food.gov.uk/enforcement/prosecutions; 2017 [accessed 10/04/2017]
[83] Towler G, Sinnott R. Chapter 1 - introduction to design. In: Chemical engineering design (second edition), Boston: Butterworth-Heinemann; 2013, p. 3-32.
[84] Crowl DA, Louvar JF, Chemical process safety: fundamentals with applications, 3 ed., Pearson Education, Boston, MA, United States, 2011.
[85] Kenney KL, Smith WA, Gresham GL, Westover T. Understanding biomass feedstock variability. Biofuels 2013;4:11-127.
[86] Waller MA, Fawcett SE. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J Bus Logist 2013;34:77-84.
[87] Grossmann I. Enterprise-wide optimization: A new frontier in process systems engineering. AIChE J 2005;51:1846-57.
[88] Bao Y, Ren L, Zhang L, Zhang X, Luo Y. Massive sensor data management framework in cloud manufacturing based on Hadoop. In: IEEE 10th International conference on industrial informatics. 2012. p.397-401.
[89] Houben C, Lapkin AA. Automatic discovery and optimization of chemical processes. Curr Opin Chem Eng 2015;9:1-7.
[90] Chue KT, Kim JN, Yoo YJ, Cho SH, Yang RT. Comparison of Activated Carbon and Zeolite 13X for CO2 Recovery from Flue Gas by Pressure Swing Adsorption. Ind Eng Chem Res 1995;34:591-8.
[91] Pilkington JL, Preston C, Gomes RL. Comparison of response surface methodology (RSM) and artificial neural networks (ANN) towards efficient extraction of artemisinin from Artemisia annua. Industrial Crops and Products 2014;58:15-24.
[92] Lopez-Juarez I, Rios-Cabrera R, Hsieh SJ, Howarth M. A hybrid non-invasive method for internal/external quality assessment of potatoes. European Food Research and Technology 2018;244:161-74.
[93] Xiros C, Christakopoulos P. Biotechnological Potential of Brewers Spent Grain and its Recent Applications. Waste Biomass Valorization 2012;3:213-32.
[94] Manufacturing vision for UK pharma. Medicines Manufacturing Industry Partnership; London; 2017.
[95] Hosseini Imani M, Niknejad P, Barzegaran MR. The impact of customers' participation level and various incentive values on implementing emergency demand response program in microgrid operation. International Journal of Electrical Power & Energy Systems 2018;96:114-25.
[96] Gonzalez Viejo C, Fuentes S, Li G, Collmann R, Condé B, Torrico D. Development of a robotic pourer constructed with ubiquitous materials, open hardware and sensors to assess beer foam quality using computer vision and pattern recognition algorithms: RoboBEER. Food Research International 2016;89:504-13.

[97] Smithers GW. Whey and whey proteins—From 'gutter-to-gold'. International Dairy Journal 2008;18:695-704.
[98] Buckholtz B, Ragai I, Wang LH. Cloud manufacturing: current trends and future implementations. J Manuf Sci Eng-Trans ASME 2015;137:
[99] Himmelblau DM. Applications of artificial neural networks in chemical engineering. Korean Journal of Chemical Engineering 2000;17:373-92.
[100] Mannan MS, Reyes-Valdes O, Jain P, Tamim N, Ahammad M. The evolution of process safety: current status and future direction. Annu Rev Chem Biomol 2016;7:135-62.
[101] Liu CC, Stefanov A, Hong J, Panciatici P. Intruders in the Grid. IEEE Power and Energy Magazine 2012;10:58-66.
[102] You Y, Cho I, Lee K. An advanced approach to security measurement system. J Supercomputing 2016;72:3443-54.
[103] Airaksinen A, Luomaranta H, Alajääskö P, Roodhuijzen A. Statistics on small and medium-sized enterprises, http://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_small_and_medium-sized_enterprises; 2018 [accessed 05/01/2018]
[104] Defra. Food statistics pocketbook 2016. In: Department for Environment FaRA. The Stationary Office London. 2017. p. 12.
[105] Ellis V, Bosworth G. Supporting rural entrepreneurship in the UK microbrewery sector. British Food Journal 2015;117:2724-38.
[106] Maier J. Made smarter. Review 2017. In: Department for Business EIS. The Stationery Office. London 2017.
[107] Foresight. The Future of Manufacturing: a new era of opportunity and challenge for the UK. In: The Government Office for Science. London. 2013.
[108] Gregson N, Crang M, Fuller S, Holmes H. Interrogating the circular economy: the moral economy of resource recovery in the EU. Econ Soc 2015;44:218-43.
[109] European Commission. Report from the Commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions on the implementation of the circular economy action plan. Brussels; 2017.
[110] Bowdish L. Trash to treasure: changing waste streams to profit streams. In: U.S. Chamber of Commerce Foundation. Washington 2016.
[111] Yong R. The circular economy in China. J Mater Cycles Waste 2007;9:121-9.
[112] Dickens P, Kelly M, Williams J. What are the significant trends shaping technology relevant to manufacturing? In: Government Office for Science. Foresight. London. 2013.
[113] Yang C, Lan S, Shen W, Huang GQ, Wang X, Lin T. Towards product customization and personalization in IoT-enabled cloud manufacturing. Cluster Comput 2017;20:1717-30.
[114] Apsley LK, Bodell CI, Danton JC, Hayden SR, Kapila S, Lessard E, Uhl RB, inventors; Amazon Technologies Inc, assignee. Providing services related to item delivery via 3D manufacturing on demand. United States patent US 20150052024. 2015 Feb 19.
[115] Prasad LK, Smyth H. 3D Printing technologies for drug delivery: a review. Drug Dev Ind Pharm 2016;42:1019-31.
[116] Izdebska J, Zolek-Tryznowska Z. 3D food printing - facts and future. Agro Food Ind Hi Tec 2016;27:33-7.
[117] Lipton JI, Cutler M, Nigl F, Cohen D, Lipson H. Additive manufacturing for the food industry. Trends Food Sci Tech 2015;43:114-23.

[118] Cerdas F, Juraschek M, Thiede S, Herrmann C. Life Cycle Assessment of 3D Printed Products in a Distributed Manufacturing System. J Ind Ecol 2017;21:S80-S93.

Vitae

Oliver Fisher earned a MEng in Chemical Engineering at the University of Nottingham (UK) in 2016. He is currently undertaking a PhD investigating intelligent resource use in process manufacturing at the University of Nottingham, and is a member of the Food, Water and Waste Research Group (FWW) in the Faculty of Engineering.

Dr Nicholas Watson has a MEng in Mechanical Engineering (University of Hull, 2005) and a PhD in Chemical Engineering (University of Leeds, 2010). From 2010 to 2014 he was a Post-Doctoral Research Assistant in the School of Food Science and Nutrition at the University of Leeds. In 2013 Dr Watson became a chartered engineer and a member of the Institute of Mechanical. In 2014 he joined the University of Nottingham as an assistant professor in Chemical Engineering and became an Associated Member of the Institute of Chemical Engineers.

Rachel Gomes is an Associate Professor in Chemical and Environmental Engineering at the University of Nottingham, following an Anne McLaren Fellowship. Her research interests lie in sustainable resource utilisation towards wastewater treatment and process manufacturing. Rachel was selected in 2013 as an University Engineering Future Research Leader, and in 2014 became 1 of 20 on the Engineering and Physical Sciences Research Council Early Career Forum in Manufacturing Research and in the Top 40 under 40 Power List for the Analytical Scientist magazine in recognition of developing novel analytics to inform on wastewater treatment and chemical manufacture.

APPENDIX B: PUBLICATION 2

Johnson Matthey Technol. Rev., 2020, **64**, (1), 93–99

JOHNSON MATTHEY
TECHNOLOGY REVIEW

www.technology.matthey.com

Intelligent Resource Use to Deliver Waste Valorisation and Process Resilience in Manufacturing Environments

Moving towards sustainable process manufacturing

Oliver J. Fisher, Nicholas J. Watson

Food, Water, Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

Josep E. Escrig

i2CAT Foundation, Calle Gran Capità, 2–4 Edifici Nexus, I2ª planta, 08034 Barcelona, Spain

Rachel L. Gomes*

Food, Water, Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

***Email:** rachel.gomes@nottingham.ac.uk

Circular economy (CE) thinking has emerged as a route to sustainable manufacture, with related cradleto-cradle implications requiring implementation from the design stage. The challenge lies in moving manufacturing environments away from the traditional linear economy paradigm, where materials, energy and water have often been designed to move out of the system and into receivership of waste management bodies after use. Recent applications of industrial digital technologies (IDTs: for example internet of things, data-driven modelling, cyberphysical systems, cloud manufacturing, cognitive computing) to manufacturing may be instrumental in transforming manufacturing from linear to circular. However, although IDTs and CE have been the focus of intensive research, there is currently limited research exploring the relationship between IDTs and the CE and how the former may drive the implementation of CE. This article aims to close the knowledge gap by exploring how an IDT (data-driven modelling) may facilitate and advance CE principles within process manufacturing systems, specifically waste valorisation and process resilience. These applications are then demonstrated through two

realworld manufacturing case studies: (a) minimising resource consumption of industrial cleaning processes and (b) transforming wastewater treatment plants (WWTPs) into manufacturing centres.

Introduction

Manufacturing drives the economies of all countries from low to high incomes; contributing 16% to global gross domestic product and 23% of global employment in 2017 (1). Manufacturing is split into two branches: discrete manufacturing (automobiles, computers and electronics, textiles) and process manufacturing (chemicals, food and drink, pharmaceutical, fast-moving consumer goods (FMCG)). The difference is that discrete manufacturing is the cutting and assembly of a bill of materials into a final distinct product (2), whereas process manufacturing is the thermal, chemical or biochemical conversion of resources into products, byproducts and waste streams (3). Both share the core value that their products and processes must be profitable to remain in business (4). However, there is mounting pressure for businesses to become not just economically sustainable but also environmentally sustainable.

A key goal of sustainable manufacturing is to conserve energy and natural resources (5). The methods of achieving sustainable manufacturing differ between discrete and process manufacturing. Discrete manufacturers are largely limited to minimising resource consumption and redesigning their products to use sustainable resources where possible (6). Process manufacturers are often able to keep the characteristics of their products the same but redesign their systems to use alternative sustainable feedstocks (7). Despite progress made by manufacturers, more needs to be done to meet demands of a growing population. In 2017, the European Union (EU) expanded its list of critical raw materials (CRM) to 27. These are defined as materials considered to be of high importance to the EU economy and of high risk to their supply (8). CE thinking has emerged as a means of increasing the resilience of manufacturing to disruptions in the supply of resources.

The UK-based Waste and Resources Action Programme (WRAP) charity defines the CE as “an alternative to a traditional linear economy (make, use, dispose) in which we keep resources in use for as long as possible, extract the maximum value from them whilst in use, then recover and regenerate products and materials at the end of each service life” (9). The application of CE thinking to process manufacturing requires deeper consideration than discrete manufacturing, where it is simple to visualise a discrete product being reused, repaired and recycled (such as fixing a kettle). This school of thought does not lend itself to some process manufacturing products; for example, it is not possible to reuse, repair

or recycle a pharmaceutical drug once consumed. However, by expanding the system to include waste produced during manufacturing (i.e. capture pharmaceuticals in waste or water streams and recover for use) and waste from the consumer (in this example human waste entering wastewater systems) it is possible to recapture the resource into the economy through waste recovery and valorisation. This also offers opportunity to increase profitability, in addition to reducing environmental pollution.

The application of emerging IDTs to manufacturing (sometimes referred to as Industry 4.0) to support the implementation of the CE has strong potential (10). However, a keyword search on the Web of Science database returns only 29 documents for both "circular economy" and "Industry 4.0", compared to 4138 and 4031 when the keywords are searched independently. This article aims to close the knowledge gap by exploring how IDTs may support CE principles within process manufacturing systems, specifically waste valorisation and process resilience.

Two process manufacturing case studies are evaluated within the field of CE. The first case study focuses on applications of CE to support waste valorisation within waste and wastewater management. For process manufacturing, 50–100% of the starting materials (and hence resources used in that system) are removed from the system via the wastewater. The wastewater and resources within then enter waste management bodies such as solid waste or WWTPs, or depending on the country and regulations (that may or may not be evident or enforced) may be released directly to the natural environment (11). Recovering water, energy and nutrients is an established application of CE to WWTPs (12) and readers are referred elsewhere (13) for a review of current innovations within WWTPs. However, the manufacturing environment from the CE perspective offers more inclusive considerations. Here innovative thinking coupled with new technologies enables process manufacturing plants to adapt and emerge, switching priorities from treating waste and water for reuse to manufacturing products from waste streams. The second case study demonstrates how IDTs may increase the resilience of FMCG industries and reduce resource consumption. Recent literature has demonstrated some innovative applications of IDTs to FMCG (14, 15). This case study aims to demonstrate how these technologies may be utilised within the context of CE.

Waste Valorisation Between Manufacturing Systems

There is a consensus among researchers that it is not possible to reach a fully CE, as the process of recycling will also always create waste and side products due to increasing entropy (16). Instead, the goal is

to identify resources from renewable sources and to minimise the spent resource leaving the system (17). Process manufacturing systems can take this further as their waste can contain valuable material available to use as is or as a building block for a related process manufacturing environment. This is known as the cradle-to-cradle concept, where waste is both recycled and utilised as a raw material (18). Applying this concept expands the traditional CE model from a singular circular system to a series of connected circular systems, where one system's waste is another's resource. This reimagining of the CE concept is shown in **Figure 1**.

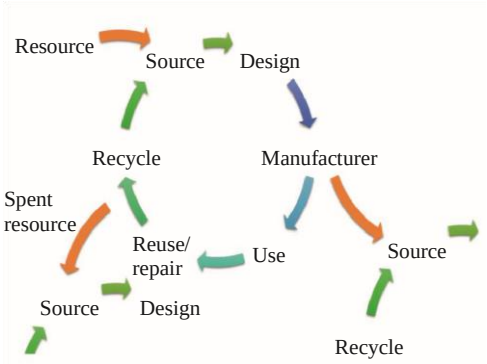


Fig. 1. Integrating the cradle-to-cradle concept within the CE model

The challenge of brokering relationships between manufacturers to form innovative collaborations that find ways to utilise waste from one as a raw material for another is known as industrial symbiosis (19–21). There are examples of support networks for industrial symbiosis, such as the National Industrial Symbiosis Programme, UK (22); Cleantech Östergötland, Sweden (23) and Kalundborg Symbiosis, Denmark (24). However, more needs to be done to make industrial symbiosis a wide-spread reality (25). There is an opportunity to utilise recent datadriven analysis and modelling advances to drive the implementation of industrial symbiosis (10). One such application is the development of data-driven models to describe manufacturing systems and enable intelligent decision-making on cloud platforms like Sharebox (a platform for physical resource sharing) (3, 26). Data-driven models can identify relationships between the system state variables (input and output) to inform on the system and make predictions on key process variables (for instance, product yield and quality) (27). By sharing and modelling a system's data, a data-driven model may evaluate the economic feasibility and environmental impact of utilising that waste as a resource. The model could further be used to evaluate multiple different systems to select the most sustainable valorisation route. For example, brewers' spent grain has multiple valorisation routes (animal and human food production, paper production, adsorbent material, enzyme production, energy source) (28). The different industrial processes to valorise the spent gains are impacted by the spent grains' characteristics (physical properties, microbial and chemical composition) (29). A data-driven model may evaluate which

valorisation route is the most sustainable dependant on the spent grains' characteristics.

Traditionally waste and WWTPs were the last stage of a product's lifecycle, either treating the waste before releasing it to the surrounding environment, or with some limited reuse in system (such as heat production). Nowadays, most WWTPs recover at least some of the energy trapped in the waste and water streams and the next generation of WWTPs are targeting energy neutrality and recovery of nutrients (11). A Severn Trent Water WWTP at Stoke Bardolph, UK, illustrates the innovative work already ongoing by process manufacturers to recapture phosphorus and energy back into the economy (30). P is on the EU's 27 CRM list and although fairly common is included because of its widespread use for crop fertilisers risking future supply. Globally, 20% of the manufactured P is contained within domestic wastewater (11) and if recovered could limit the depletion of phosphate rock that is not renewable (31). The Stoke Bardolph site uses two PHOSPAQ™ reactors (developed by Dutch company Paques BV) that recover approximately 736 tonne $year^{-1}$ of phosphorous in the liquor dewatered from the sludge from municipal sewage treatment (30). Also installed on the site is an ANAMMOX® reactor and a BIOPAQ® Upflow Anaerobic Sludge Blanket (UASB), which recover ammonia and generate biogas respectively. This is then used for the combined heat and power engines, contributing 7% to the site's energy use (30). With innovative thinking, coupled with new technologies, future WWTPs could take this further, so they are not only recovering resources but also manufacturing products onsite, **Figure 2**. The energy and water recovered from the wastewater could be used to support the manufacturing processes, helping to realise a truly CE by reducing the requirement of fresh resources, **Figure 2**. There is the potential to generate a wide range of products from wastewaters including: biofuels, biohydrogen, biopolymers, single-cell protein, fertiliser, cellulose and alginic acid as well as nutrients and metals recovery (11, 32). However, there is still some way to go before these circular thinking technologies can be deployed, as the technology readiness levels (TRL) for most are below TRL5 (11). TRLs measure the maturity of technology during the stages of its development between basic principles (TRL1) and commercialised operating product (TRL10) (33). One project aiming to achieve this is the NextGen initiative, funded by the EU (34). NextGen aims to demonstrate innovative technological, business and governance solutions for water in the CE, in ten high-profile, large-scale, demonstration cases across Europe.

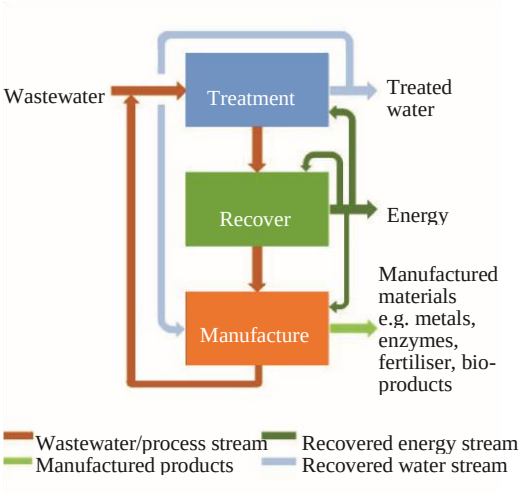


Fig. 2. Innovative WWTPs case study. Moving beyond the traditional concept of treating water so it is fit for purpose e.g. for release to the aqueous environment. Moving towards treatment plants of all sizes powered by the renewable bioenergy recovered from the waste or water and additional products where resources are recovered

Process Resilience Across Different Manufacturing Environments

Process manufacturers consume an extensive amount of resources during production processes. For example, the processing industries use 20% of the total global freshwater reserves during production (35). Towards the goal of sustainable resource use, the first steps are to investigate substituting raw materials for sustainable alternatives and to increase the efficiency of manufacturing processes to reduce the resources wasted during production. This falls under the field of 'process intensification', which is a chemical and process design approach that leads to substantially smaller, cleaner, safer and more energy-efficient processes (36). Process manufacturing systems are subject to variability throughout the system, including variations in feedstock characteristics and supply, unit operation process conditions and the waste produced by the system. This variability results in wasted resources, as manufacturers over-design their systems to limit the impact variability has on product quality and yield (37), and increases the complexity in process intensification attempts (38). As manufacturers move away from non-renewable sources towards renewable feedstocks this will become an even greater challenge for engineers (39). Previously, process manufacturing plants have been overdesigned in an attempt to capture the variability within the boundaries of the plant's system (40).

Recently there have been advances in the application of data-driven modelling to better understand and optimise process manufacturing systems (41, 42). By modelling historic data, it was possible to identify and understand the effect of the variables that strongly affect the system (42). The model can then be used to make predictions on how these variables can be manipulated to reach the model's goal, defined by the manufacturer. Currently, process manufacturers collect a large amount of data from process control systems that is not fully utilised and often only used for after-the-event analysis

(43). There is an opportunity to develop data-driven models from this data to investigate the cause of resource waste within the systems and ensure this waste is designed out of future systems. For example, in the biopharmaceutical industry yields can vary from 50% to 100% for no immediately discernible reason (42). Sadati et al. were able to develop a data-driven model from historic data that identified the control variables resulting in fluctuations in yield (42).

Data-driven models can also improve the development of affordable sensors to monitor systems that have previously been economically unfeasible to monitor. This is particularly true for the food and drink industry that is characterised by a large number of small and medium enterprises (SMEs) who face small profit margins and have limited resources to utilise new IDTs (44). For example, clean-in-place (CIP) systems (a method of cleaning the interior surfaces of pipes and process equipment without disassembly) use excessive amounts of water, chemicals and time. Cleaning is essential to ensure the equipment remains hygienic but inefficient in terms of resource use as systems are always designed to clean the materials which cause the worst fouling to the equipment (45). If manufacturers were able to model and predict the fouling behaviour of each product it would reduce the amount by which the CIP systems are over-designed, resulting in a saving of resources. An example of a sensor and data-driven model system used to reduce resource use during industrial equipment cleaning processes is presented in **Figure 3**. A data-driven model was developed to predict the point at which cleaning

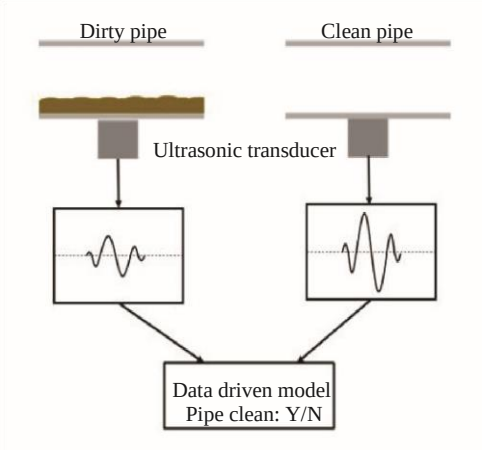


Fig. 3. CIP case study (45). Development of an ultrasonic sensor which uses machine learning to monitor and therefore optimise the cleaning of industrial processing equipment. This technology will enable the optimisation of the cleaning processes, reducing cleaning time and chemical, energy and water use

was achieved by modelling measurements from an affordable, non-intrusive ultrasonic sensor.

When aiming for sustainable manufacturing it is important to recognise that what is sustainable in one environment may not be sustainable for another. Different geographical regions and industries face different regulations and have access to different resources. The EU's rare earth minerals are depleted and there is a heavy reliance on imports, as reflected in the EU's list of CRM (8). Regions that still contain plentiful mineral

reserves may face different challenges such as water stress (46). Datadriven modelling for dynamic forecasting of a range of applications has already been tested in a variety of fields (47). In the future, it may be possible for a manufacturer to model and forecast the availability and variability of the supply of resources. This will enable intelligent decision making to identify and implement the most sustainable manufacturing route dependent on available resources. By recovering, reusing, recycling and valorising waste resources it will further increase the resilience of manufacturing systems from disruptions in the supply chain.

Final Remarks

During the 20th Century, our manufacturing systems were driven by economic incentives to expand the production of goods and services, with little to no regard to the environmental impacts this caused (48). In the past couple of decades there has been a recognition by the international community that this is unsustainable and manufacturers have been pressured to introduce sustainable strategies. However, the progress made has not been sufficient and the resources required to support the global population now exceed those available (49). This article has offered a brief overview of the applications of data-driven models (a key IDT) to support the CE principles of waste valorisation and process resilience. Data-driven modelling applications for analysis, decision making and forecasting were presented in the context of two process manufacturing case studies. A gap currently exists between the research fields of CE and Industry 4.0, as demonstrated by the limited number of joint publications between these fields. Therefore, there is great potential for research demonstrating the application of further IDTs (for example internet of things, cyber-physical systems, cloud manufacturing, cognitive computing) to drive the CE.

Acknowledgements

Rachel Gomes would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) (grant number EP/ K014161), Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing and EPSRC (grant number EP/M022234/1), Biotransforomics: Bioanalysis to Engineer Understanding in Wastewater Treatment. Nicholas Watson would like to acknowledge the support of Innovate UK (ref:132205). Oliver Fisher would also like to acknowledge the University of Nottingham Faculty of Engineering for his PhD scholarship.

References

1. ‘World Bank Open Data – Free and Open Access to Global Development Data’, The World Bank,

Washington, DC, USA: https://data.worldbank.org/ (Accessed on 11th July, 2019)

2. D. Brandl, 'Manufacturing Control', in "Design Patterns for Flexible Manufacturing", Ch. 1, Instrumentation Systems and Automation Society, Research Triangle Park, USA, 2007, pp. 1–6
3. O. Fisher, N. Watson, L. Porcu, D. Bacon, M. Rigley and R. L. Gomes, *J. Manuf. Syst.*, 2018, **47**, 53
4. J. A. Brierley, C. J. Cowton and C. Drury, *Int. J. Prod. Econ*., 2006, **100**, (2), 314
5. L. Gardner and J. Colwill, *Procedia CIRP*, 2016, **41**, 282
6. M. G. Yang, P. Hong and S. B. Modi, *Int. J. Prod. Econ*., 2011, **129**, (2), 251
7. F. Roschangar, Y. Zhou, D. J. C. Constable,
J. Colberg, D. P. Dickson, P. J. Dunn, M. D. Eastgate,
F. Gallou, J. D. Hayler, S. G. Koenig, M. E. Kopach,
D. K. Leahy, I. Mergelsberg, U. Scholz, A. G. Smith,
M. Henry, J. Mulder, J. Brandenburg, J. R. Dehli, D. R. Fandrick, K. R. Fandrick, F. Gnad-Badouin, G. Zerban, K. Groll, P. T. Anastas, R. A. Sheldon and C. H. Senanayake, *Green Chem*., 2018, **20**, (10), 2206
8. 'Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the 2017 List of Critical Raw Materials for the EU', COM/2017/0490 final, European Commission, Brussels, Belgium, 13th September, 2017
9. 'WRAP and the Circular Economy – What is a Circular Economy?', The Waste and Resources Action Programme (WRAP), Banbury, UK: http:// www.wrap.org.uk/about-us/about/wrap-andcircular-economy (Accessed on 15th November 2018)
10. M.-L. Tseng, R. R. Tan, A. S. F. Chiu, C.-F. Chien and T. C. Kuo, *Resour. Conserv. Recycl*., 2018, **131**, 146
11. D. Puyol, D. J. Batstone, T. Hülsen, S. Astals, M. Peces and J. O. Krömer, *Front. Microbiol.*, 2017, **7**, 2106
12. C. C. Flores, H. Bressers, C. Gutierrez and C. de Boer, *Manag. Res. Rev*., 2018, **41**, (5), 554
13. J. M. Lema and S. Suarez, "Innovative Wastewater Treatment and Resource Recovery Technologies – Impacts on Energy, Economy and Environment", IWA Publishing, London, UK, 2017
14. A. M. C. Adikari and T. P. Amalan, 'Distribution Cost Optimization Using Big Data Analytics, Machine Learning and Computer Simulation for FMCG Sector', International Research Conference on Smart Computing and Systems Engineering, Colombo, Sri Lanka, 28th March, 2019, IEEE, Piscataway, USA, pp. 63–69
15. F. Odważny, O. Szymańska and P. Cyplik, *LogForum*, 2018, **14**, (2), 257
16. N. Millar, E. McLaughlin and T. Börger, *Ecol. Econ.,* 2019, **158**, 11
17. J. Korhonen, A. Honkasalo and J. Seppälä, *Ecol. Econ.*, 2018, **143**, 37
18. W. McDonough and M. Braungart, "Cradle to Cradle: Remaking the Way We Make Things", North Point Press, a division of Farrar, Straus and Giroux, New York, USA, 2002
19. R. G. Charles, P. Douglas, J. A. Baker, M. J. Carnie,
J. O. Douglas, D. J. Penney and T. M. Watson, *J. Clean. Prod*., 2018, **202**, 1167
20. R. Álvarez and C. Ruiz-Puente, *Waste Biomass Valor*., 2017, **8**, (5), 1521
21. L. Fraccascia, V. Albino and C. A. Garavelli, *Int. J. Prod. Econ*., 2017, **183**, (A), 273
22. 'National Industrial Symbiosis Programme', International Synergies Ltd, Birmingham, UK: https://www.international-synergies.com/projects/national-industrial-symbiosisprogramme/ (Accessed on 27th November, 2019)
23. Cleantech Östergötland, Linköping, Sweden: https://cleantechostergotland.se/ (Accessed on 27th November, 2019)
24. 'Kalundborg Symbiosis' SymbiosisCenter, Kalundborg, Denmark: http://www.symbiosis.dk/ en/ (Accessed on 27th November, 2019)
25. R. Taddeo, A. Simboli, G. Ioppolo and A. Morgante, *Sustainability*, 2017, **9**, (2), 169
26. 'Project Overview', Sharebox – Secure Sharing, Barcelona, Spain: http://sharebox-project.eu/#overview (Accessed on 10th July 2019)
27. L. Angria, S. Y. D. Sari, M. Zarlis and Tulus, *IOP Conf. Ser.: Mater. Sci. Eng*., 2018, **300**, 012013
28. S. I. Mussatto, *J. Sci. Food Agric.*, 2014, **94**, (7), 1264
29. J. A. Robertson, K. J. A. I'Anson, J. Treimo, C. B. Faulds, T. F. Brocklehurst, V. G. H. Eijsink and K. W. Waldron, *LWT – Food Sci. Technol.*, 2010, **43**, (6), 890
30. N. Durose and T. Jeffcoat, "Stoke Bardolph STW Centrate Scheme – First UK Installation to Utilise the Phospaq, UASB^{+} and Anammox Technologies in a Single Process

Solution", Water Projects Ltd, Manchester, UK, 2014, 4 pp

31. K. Senthilkumar, A. Mollier, M. Delmas, S. Pellerin and T. Nesme, *Resour. Conserv. Recycl.*, 2014, **87**, 97
32. J. P. van der Hoek, H. de Fooij and A. Struker, *Resour. Conserv. Recycl.*, 2016, **113**, 53
33. J. Straub, *Aerosp. Sci. Technol*., 2015, **46**, 312
34. 'Towards a Next Generation of Water Systems and Services for the Circular Economy', NextGen, Project ID 776541, Community Research and Development Information Service (CORDIS), European Commission, Brussels, Belgium, 29th July, 2019
35. A. Howe, "Water Management in the Food and Drink Industry", IChemE, Rugby, UK, 2014, 6 pp
36. D. Reay, C. Ramshaw and A. Harvey, "Process Intensification – Engineering for Efficiency, Sustainability and Flexibility", Elsevier Ltd, Oxford, UK, 2008, 444 pp
37. R. Gani, *Comput. Chem. Eng.*, 2004, **28**, (12), 2441
38. M. Dogru and A. Erdem, *Energy Fuels*, 2019, **33**, (1), 340
39. K. L. Kenney, W. A. Smith, G. L. Gresham and T. L. Westover, *Biofuels*, 2013, **4**, (1), 111
40. G. Towler and R. Sinnott, 'Introduction to Design', in "Chemical Engineering Design", 2nd Edn., Ch. 1, Elsevier Ltd, Oxford, UK, 2013, pp. 3–32
41. A. Simeone, B. Deng, N. Watson and E. Woolley, *Sensors*, 2018, **18**, (11), 3742
42. N. Sadati, R. B. Chinnam and M. Z. Nezhad, *Expert Syst. Appl.*, 2018, **93**, 456
43. S. J. Qin, *AIChE J.*, 2014, **60**, (9), 3092
44. "Food Statistics Pocketbook 2016", Department for Enviroment Food and Rural Affairs, The Stationery
45. Office Limited, London, UK, 2017, 62 pp
46. J. Escrig, E. Woolley, S. Rangappa, A. Simeone and N. J. Watson, *Food Control*, 2019, **104**, 358
47. W. J. Cosgrove and D. P. Loucks, *Water Resour. Res.*, 2015, **51**, (6), 4823
48. S. Makridakis, E. Spiliotis and V. Assimakopoulos, *PLoS ONE*, 2018, **13**, (3), e0194889
49. A. Giret, D. Trentesaux and V. Prabhu, *J. Manuf. Syst.*, 2015, **37**, (1), 126
50. N. Supanchaiyamat and A. J. Hunt, *ChemSusChem*, 2019, **12**, (2), 397



Oliver Fisher earned a MEng in Chemical Engineering at the University of Nottingham, UK, in 2016. He recently completed a PhD investigating intelligent resource use in process manufacturing at the University of Nottingham and is a member of the Food, Water, Waste Research Group (FWW) in the Faculty of Engineering. As part of his work he won additional funding to undertake a placement at Lindhurst Engineering Ltd, UK, developing a datadriven model of Lindhurst's H^2AD bioprocess.



Nicholas Watson has a MEng in Mechanical Engineering (University of Hull, UK, 2005) and a PhD in Chemical Engineering (University of Leeds, UK, 2010). From 2010 to 2014 he was a postdoctoral research assistant in the School of Food Science and Nutrition at the University of Leeds. In 2013 Nicholas Watson became a chartered engineer and a member of the Institution of Mechanical Engineers. In 2014 he joined the University of Nottingham as an assistant professor in Chemical Engineering and became an associate member of the Institution of Chemical Engineers.



Josep Escrig is the director of the Distributed Artificial Intelligence area of the i2CAT private foundation. He holds an Industrial Engineer degree and MEng in Sustainability and Energy Efficiency from the University Jaume I, Spain; MEng in Mechanical Engineering from the Institut National des Sciences Appliquées de Toulouse (INSA Toulouse), France and PhD in Chemical and Environmental Engineering from the University of Nottingham. After his PhD he got a position as data manager of the Consortium on Transient and Complex Multiphase Flows and Flow Assurance (TMF consortium) and later he did postdoctoral research at the University of Nottingham on intelligent sensors applied to industry. Currently, Josep works in two Horizon 2020 projects in i2CAT and other private projects with companies.



Rachel Gomes is an associate professor in Chemical and Environmental Engineering at the University of Nottingham, following an Anne McLaren Fellowship. Her research interests lie in sustainable resource utilisation towards wastewater treatment and process manufacturing. Rachel was selected in 2013 as a University Engineering Future Research Leader and in 2014 became one of 20 on the Engineering and Physical Sciences Research Council Early Career Forum in Manufacturing Research and in the Top 40 under 40 Power List for The Analytical Scientist magazine in recognition of developing novel analytics to inform on wastewater treatment and chemical manufacture.

APPENDIX C: PUBLICATION 3

Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems

Oliver J FISHER[1], Nicholas J WATSON[1], Josep E ESCRIG[2], Rob WITT[3], Laura PORCU[4/5], Darren BACON[5], Martin RIGLEY[5], Rachel L GOMES[1]

[1] Food, Water, Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
[2] i2CAT Foundation, Calle Gran Capita, 2 -4 Edifici Nexus (Campus Nord Upc), 08034, Barcelona, Spain
[3] Totally Brewed, Unit 8-9 Wholesale District, Meadow Lane, Nottingham, NG2 3JJ, United Kingdom
[4] Energy Innovation & Collaboration, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, United Kingdom
[5] Lindhurst Engineering Ltd., Midland Road, Sutton in Ashfield, Nottinghamshire, NG17 5GS, United Kingdom

Corresponding author
Author: Rachel L Gomes
Tel.: +44 115 846 8883.
E-mail address: rachel.gomes@nottingham.ac.uk

Abstract

The increasing availability of data, due to the adoption of low-cost industrial internet of things technologies, coupled with increasing processing power from cloud computing, is fuelling increase use of data-driven models in manufacturing. Utilising case studies from the food and drink industry and waste management industry, the considerations and challenges faced when developing data-driven models for manufacturing systems are explored. Ensuring a high-quality set of model development data that accurately represents the manufacturing system is key to the successful development of a data-driven model. The cross-industry standard process for data mining (CRISP-DM) framework is used to provide a reference at to what stage process manufacturers will face unique considerations and challenges when developing a data-driven model. This paper then explores how data-driven models can be utilised to characterise process streams and support the implementation of the circular economy principals, process resilience and waste valorisation.

Highlights

- Data-driven models (DDMs) will become widespread across manufacturing
- Paramount to DDMs is the collection of an accurate set of model development data
- Process manufacturers face unique considerations and challenges in collecting data
- These points are presented in the context of the CRISP-DM framework

- This supports the development of DDMs to meet manufacturers' requirements

Keywords

Data-driven models, Process resilience, Waste valorisation, Mathematical modelling, Machine learning, Industry 4.0

1. Introduction

Fundamental to manufacturing is mathematical modelling (Hangos and Cameron, 2001), which utilises analogies to help understand the behaviour of a complex system (Cross and Moscardini, 1985). Models translate theories of how the world functions into the language of mathematics. Once built, a model can be used to aid in decision-making, develop scientific understanding, communicate knowledge and/or make predictions (Schichl, 2004). Manufacturers have utilised mathematical modelling for four primary applications: (1) planning and design; (2) monitoring and control; (3) process optimisation and (4) risk mitigation (Perry and Green, 2008). There are two branches of manufacturing, discrete manufacturing and process manufacturing. Discrete manufacturing consists of a bill-of-materials that moves between a set of manufacturing equipment, as it is cut and assembled together (Brandl, 2007). Whereas in process manufacturing, raw or waste materials flow through the manufacturing plant undergoing thermal, chemical and/or biochemical conversion (Fisher et al., 2018). This fundamental difference between the two causes numerous disparities between the manufacturing practices/processes (Brandl, 2007). Therefore, when constructing a model of a system, both discrete and process manufacturing have a unique set of considerations, challenges and opportunities. This paper shall focus on those faced in process manufacturing environments.

Within the field of mathematical modelling, there are two distinct branches: first principles modelling (often referred to as mechanistic models) and empirical modelling. First principle models build a series of equations by examining the workings of the system's individual parts (Schichl, 2004). First principle models rely on system understanding to compensate for lack of data. Because of this, they have a greater potential for extrapolation compared to empirical models (Mathews, 2004). Empirical models are mathematical equations derived from the analysis of data; therefore, requiring less knowledge of the system (Solomatine et al., 2008). Empirical models rely on the assumption that the data is of sufficient granularity and/or quantity to define the system. However, if the system is defined within the data, empirical models are valuable tools for characterising the system input-output relationships, especially when there is limited engineering-domain knowledge to characterise complex systems (Luo et al., 2016). For example, empirical models have proven efficiency at modelling novel configuration bioelectrochemical systems (BES), which otherwise would require detailed knowledge of complex interactions between physical, chemical and electrochemistry principles (Luo et al., 2016).

Empirical models are already well utilised in process manufacturing because of the volume of data manufacturers produce (Rasmuson et al., 2014). Recently the capabilities of empirical models have greatly expanded due to advances in the fields of computational intelligence and machine learning, these new approaches are

encompassed into the field of data-driven modelling (Solomatine et al., 2008). Computational intelligence are nature-inspired computational approaches to problem-solving; for example, algorithms that mimic the behaviour of animals (e.g. swarm intelligence) (Saka et al., 2013) or algorithms that replicate the behaviour of how humans solve problems (e.g. artificial neural network, ANN) (Kim, 2017). Machine learning focuses on the development of algorithms that can access data and use it to learn for themselves (e.g. support vector machines or random forest) and algorithms can belong within each fields (e.g. ANN) (Coley et al., 2018). Data-driven models (DDMs) are able to find relationships between the system state variables (input and output) without prior knowledge of the system (Angria et al., 2018); although, the incorporation of prior may enhance DDM predictive capabilities (Lauer and Bloch, 2008). Data-driven models derive the system's relationships by training an algorithm (e.g. linear regression, ANN, Gaussian process) on manufacturing data (Kim, 2017). Although data-driven algorithms have long existed (Ojha et al., 2017), their use in industry has been limited because of data constraints (e.g. lack of data, data not stored in useable format) and lack of computational processing power (Ge, 2017). There is an unprecedented rise in the volume of manufacturing data being generated due to the adoption of cyber-physical systems, smart factories and the industrial internet of things (IIoT) (Sadati et al., 2018). In 2015, it was reported that manufacturers globally generated more than 1000 Exabyte of data and by 2025 data generation in manufacturing will increase 20-fold (Yin and Kaynak, 2015). This means that the volume of data available to build DDMs has never been greater. Data-driven models were not always considered suitable for enterprise-wide modelling due to the computational cost in modelling vast volumes of data (Boukouvala et al., 2016). However, with the introduction of cloud computing, manufacturers now have affordable access to the processing power required to model large data sets (Ge, 2017). Subsequently, DDMs are becoming prevalent across industry for modelling and monitoring of plant-wide industrial processes (Ge, 2017).

There are well-known methodologies that provide a structured approach to developing a DDM. These include data mining and knowledge discovery in databases (KDD) (Fayyad and Stolorz, 1997), the cross-industry standard process for data mining (CRISP-DM) (Shearer, 2000), and sample, explore, modify, model and assess (SEMMA) (Shafique and Qaiser, 2014). Out of these the CRISP-DM is the most widely used methodology for developing DDMs and considered the *de facto* standard by industry (Mariscal et al., 2010). The CRISP-DM is composed of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Business understanding focuses on defining DDM objectives and requirements from a business objective. Data understanding is the collection and exploration of the data. Data preparation is the selection, cleaning and transformation of the data, in order to format the data from the next phase modelling. Modelling concerns the selection and application of various modelling techniques. Evaluation is the evaluation of obtained models and how to use their results. Finally, deployment focuses on utilising of the obtained results, knowledge and/or models to benefit the business. Developing a DDM is iterative as knowledge gained during the process may redefine the objectives and modelling approach, as shown in Fig. 1.

Utilising a structured methodology to develop a DDM helps the developer avoid common data modelling mistakes that may result in models built that exhibit poor

generalisation and overfitting problems. Unless properly addressed these are the two main sources of error in DDMs (Kim, 2017) and defined as:

- *Generalisation*: is the capability of a DDM to fit and make predictions of data that was not used during the development of the model (Kim, 2017). The challenge of making the model performance consistently between data used to develop the model and new input data is known as generalisation.
- *Overfitting*: is the generation of a model that corresponds too closely or exactly to the noise (error) within the dataset, which negatively impacts future predictions (Srivastava et al., 2014).

Process manufacturers can face some unique challenges when facing generalisation and overfitting problems, due to the difficulty in defining the process manufacturing system. A manufacturing system is defined as an input stream(s) which passes through a process changing their physical and/or chemical nature into an output stream(s), which may consist of multiple products, by-products and/or waste material. Because of characteristics specific to process manufacturing systems (e.g. feedstock and waste variability, non-linearity of processes, product specification) deciding on what data is required and knowing whether the system is accurately represented in the model's development data is a challenge. This problem is further compounded by a system where data availability is limited (e.g. batch production of multiple products, frequent changes to manufacturing practise, implementation of new technologies). This is because the system is likely to contain regions underrepresented by the data, which possibly restricts the model's capability to make predictions in that region. Models of said systems will have to undergo frequent retraining as more data becomes available. These considerations and challenges have often been neglected in previous process manufacturing DDM (Charte et al., 2017; Ning and You, 2018; Sadati et al., 2018).

The CRISP-DM approach has been successfully applied to process manufacturing scenarios where data is abundant and the challenge has been extracting useful information from the data (Arce et al., 2018; Atzmueller et al., 2017). There have been several attempts to update the CRISP-DM approach from a manufacturer perspective (Harding et al., 2006; Soroush Rohanizadeh and Moghadam, 2009), but these approaches have taken a general view of manufacturing as a whole. Process manufacturers whose systems are poorly defined will face challenges specific to process manufacturers from data availability and variability. Therefore, the aim of this paper is to present the considerations and challenges unique to process manufacturers. This knowledge may benefit future manufacturers who wish to develop models of their systems and avoid overfitting and generalisation problems. These points are presented in the context of the CRISP-DM framework to provide a structure to the modelling task, as shown in Fig. 1. This may be used as a guide by process manufacturers and modellers to facilitate the development of DDMs for their particular application.

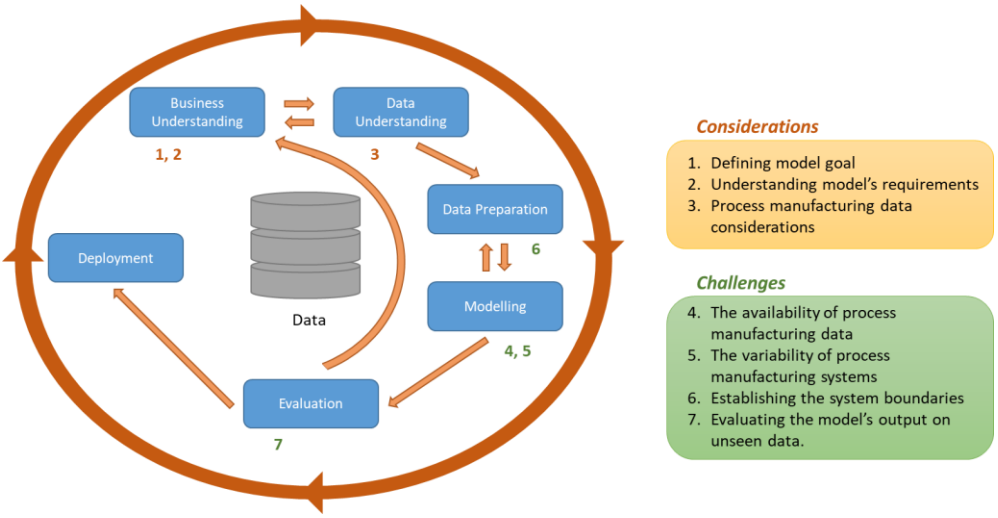


Fig. 6. Phrases of the CRISP-DM detailing updated to show the unique considerations and challenges faced when developing a data-driven model describing a process manufacturing system. Adapted from (Shearer, 2000).

Through two case studies, the considerations and challenges faced when modelling process manufacturing systems will be presented, before discussing what new opportunities arise from data-driven modelling. Each case study project was performed with a small and medium enterprise (SME). The aim of these projects was to increase the economic and environmental sustainability of these companies through the development of DDMs which utilised data collected from each manufacturing process. Both projects presented a variety of considerations and challenges that the authors and SMEs had to overcome (Table 1).

1.1. Case study 1: Waste/water management

SMEs are under pressure to reduce manufacturing costs and an effective waste/water management strategy is paramount. By 2050 there will be a 400% increase in demand for water by the manufacturing sector (OECD, 2012). Larger manufacturers have been utilising processes, like anaerobic digestion (AD), to valorise their waste to produce bioenergy and consider the resulting solid waste as fertiliser (Lin et al., 2013). The upfront capital costs of anaerobic digestion are a significant barrier for SMEs, which are generally more sensitive to additional financial costs (Rizos et al., 2015).

In the UK, Lindhurst Engineering Ltd. in partnership with the University of Nottingham has developed a technology called H^2AD Micro AD. The H^2AD bioprocess is capable of treating a variety of wastewaters, from sources including agriculture, brewing, soft drinks, foods, bio-manufacture residues, to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy. The H^2AD is a hybrid of anaerobic digestion and a microbial fuel cell, targeted at treating SME process manufacturer's wastewaters due to its modular, low-cost design.

Variations in the composition and characteristics of wastewater require management practices that can accommodate fluctuations in feedstock properties and process conditions, yet still deliver appropriate outputs without compensation in capital or operating costs. These wastewater variations, as well as potential environmental factors (e.g. seasonal), can decrease the efficiency of the bacteria to reduce pollutant load and generate biogas, which are the heart of the H^2AD bioprocess. This case study

aims to develop a DDM aimed at understanding how variations in the input wastewater stream and H^2AD process conditions inform on the process performance. The performance is defined as the percentage removal of wastewater pollutants and volume of biogas generated. The model is trained on data collected from an operating H^2AD plant currently treating wastes produced by a 300 cattle dairy farm. The farm waste contains cattle slurry, bedding waste, waste milk, footbath, parlour washing and rainfall. The model will then be used to:

(1) Analyse and predict wastewater variations effect on the bioprocess performance and respond to energy challenges.
(2) Predict the bioprocess's ability to handle new feedstocks.

1.2 Case study 2: Craft breweries

There are over 2,000 craft breweries within the UK, most utilising equipment and processing techniques that have not changed for +30 years (Simmonds, 2017). This makes it an often unpredictable and inefficient process, especially regarding water overconsumption (Edmonds, 2016). A critical stage of the brewing process is fermentation, where yeast is added to the wort (the liquid extracted from the mashing process downstream) to convert sugar to alcohol (ethanol). The fermentation process is complete once the beer has reached the desired alcohol content and flavour profile. For SME breweries, this is currently determined by removing samples from the vessel and manually measuring the specific gravity with a hydrometer. Although the fermentation duration should be identical for each batch of a particular beer, this is rarely the case due to seasonal variability in ingredient (malts, hops, water) properties and natural fluctuations in process temperature. As specific gravity measurements are only taken every 4-10 hours (or longer if no one is working overnight) this often leads to overfermentation, affecting product quality and resource utilisation.

Ultrasonic sensors can be used to monitor industrial processes such as equipment cleaning (Escrig et al., 2019) and multiphase flow (Al-Aufi et al., 2019). Previous research has shown that ultrasonic measurements can be used to monitor changes in ethanol volume concentration during beer fermentation processes (Krause et al., 2011; Resa et al., 2004). The authors of this paper are currently working on a project to develop a low-cost ultrasonic sensor designed for craft brewers. They are collaborating with an SME craft brewery in Nottingham, UK called Totally Brewed. This project records ultrasonic and temperature measurements during fermentation and uses supervised machine learning algorithms, such as artificial neural networks (ANNs), to predict the alcohol concentration from these measurements.

These case studies were chosen because they contain characteristics unique to process manufacturing systems (Table 1). Examples of the considerations and challenges resulting from these characteristics are referred to throughout this paper to support the topics discussed in each section.

Table 5. Summary of case studies (CSs)

	Waste/water management CS1	**Brewery CS2**
Manufacturing system	The H^2AD bioprocess is a circular economy technology,	Fermentation is a critical process in beer production

	which treats wastewater to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy.	where the sugar in the wort is converted to alcohol.
Project's aim	To develop a model to analyse and predict the effect of wastewater variability (e.g. suspended solid content) has on the bioprocess performance. Utilise model to make predictions on bioprocess's ability to treat new feedstocks.	To develop a model to predict alcohol concentration from temperature and ultrasonic measurements during a fermentation process.
Feedstock characteristics	Farm waste which has variability due to changes in farm practices, the welfare of cattle, rainfall, seasons and more.	Raw ingredients are cereal grain, hops, water and brewer's yeast. Each ingredient contains inherent variation which can affect flavour, colour, carbonation, alcohol content and other subtle changes in the beer.
Product specification	Wastewater pollutants are reduced so water is fit for reuse. And biogas generated contains a minimum of 70% by volume of methane.	Slap in the Face beer containing 4% alcohol concentration.
Process description	The H^2AD is a semi-batch bioprocess, meaning that X volume of liquid enters and leaves the system at predefined time intervals. When entering the system the feedstock is heated to 30°C. The process stream is recirculated through the system.	The main stages in the brewing process are: wort production, alcoholic fermentation and maturation. It is a batch process with the fermentation lasting between 3-6 days.
Model inputs	Water quality analysis parameters of feedstock (e.g. pH, dissolved oxygen, chemical oxygen demand, total suspended solids). H^2AD process conditions (e.g. temperature, system pressure).	Temperature, ultrasonic velocity and received ultrasonic signal amplitude, during fermentation.
Model outputs	Chemical oxygen demand (mg/L) and total suspended solids (mg/L) of the output stream.	Alcohol concentration during fermentation.

	Daily biogas production (L) and methane concentration of biogas (% vol).	
Volume of data	Samples were collected once a week for a 1 year period, resulting in 52 data points.	Four datasets were collected from four batches of the Slap in the Face beer. Each data date set was made up of between 830-835 data points.

2. Considerations for data-driven process manufacturing models

When developing a model, there are considerations, which must be made before the data collection and model building stages. Primarily, these considerations are

(1) Defining the model's goal;
(2) Understanding what is required from the model and,
(3) Data considerations (what data is already collected, what additional data points are required, how much data is needed, how the data is to be collected and how often).

Detailed discussions between the manufacturer and modeller must take place to address these considerations. The modeller (or team of modellers) may be an internal employee(s) or manufacturers may sub-contract data analysists and software specialists to model their processes. Digital companies, like Microsoft's Azure software, are now offering these services through the cloud (Microsoft, 2018). This service-orientated approach is often more affordable as they operate on a pay-as-you-go business model (Fisher et al., 2018). This means they are more readily available, especially for SMEs, who are understandably unlikely to have the required expertise within their current workforce or be able to afford the cost of developing, running and maintaining the models themselves.

2.1 Defining the model's goal

When constructing a mathematical model, the first stage is to identify and define the model's goal, as this will define the model's outputs (Cross and Moscardini, 1985). A model may have multiple goals; these may change over time and may come into conflict with each other. For example, in the waste/water management case study (CS1), the model's aim is to understand and predict how the system' variability affects the H^2AD bioprocess performance and make predictions how the H^2AD bioprocess will perform on new feedstocks. The performance of the process is defined as the percentage removal of wastewater pollutants and volume of biogas generated. The model's goal is to make predictions using the feedstock and H^2AD process data to determine the H^2AD process conditions that maximise the removal of key pollutants and volume of biogas generated. These two goals may come into conflict with one another; for example, maximising pollutant removal may adversely affect the rate of biogas production. Determining which goal to favour will be driven by current economic and environmental factors, customer demands and regulations. In the brewery case study (CS2), the model's aim is to predict when fermentation is complete and the beer has reached desired alcohol concentration and flavour profile. By doing

so it will minimise over-fermentation, which wastes resources and results in an inferior product. This defined the model's goal of developing a model that could predict the alcohol concentration in real-time from affordable ultrasonic and temperature measurements.

2.2 Understanding the requirements from the model

A model's goal will inform on what function the model is to perform and what degree of accuracy is required in the model outputs. A DDM may perform two functions:

(1) **Fit data** (by regression, classification and/or clustering) to identify relationships between input, output, process and material properties.
(2) **Predict** process performance and resource use.

Dependant on the goal, a DDM may only be required to perform function (1) or it may perform both functions (1) and (2). Data-driven models are able to model highly complex, nonlinear systems to identify these relationships (Pasini, 2015). Dependant on the manufacturer's requirement, DDMs can also be utilised for predictions (Almeida, 2002). However, the complexity of the model may increase, which will be dependant on the required accuracy of the model's predictions (Almeida, 2002). Table 2 gives examples of both functions (1) and (2).

Table 6. Examples of data-driven models requirements for the 2 case studies, waste/water management and brewery

	Identify relationships	**Make predictions**
Waste/water management CS1	To understand how variations in both the feedstock's characteristics and process conditions affect the H^2AD bioprocess. In addition, identify which key variables in both strongly affect performance.	(1) To predict how variations in feedstock will affect the bioprocess performance and optimise the process conditions to maximise performance. (2) To predict the H^2AD's performance when treating new feedstocks to determine if they are financially viable
Brewery CS2	Models data to identify the relationship between temperature and ultrasonic measurements during fermentation to alcohol. Furthermore, to understand how variations in these measurements affect alcohol generation.	(1) Utilises the model to make a prediction of fermentation endpoint for the current batch. (2) To evaluate feedstocks (hops, malt, water, yeast) influence the fermentation time and adjust process conditions to accommodate.

An important consideration for the manufacturer is assessing the required accuracy between the model's output and the true value. Is it important to the manufacturer whether the model is able to make predictions to 99.9% accuracy or is a model with reduced accuracy sufficient to complete the task? According to Dean Abbott, president of Abbott Analytics, it is more desirable to a manufacturer to build a model

that meets the requirements for the task at a cheaper cost than to invest more money in unnecessary accuracy (Garment, 2014). The accuracy of a DDMs is widely defined by the norm of the differences ("residuals") between estimates and observations (Duarte et al., 2004). The required accuracy will be determined by any one or a combination of the following factors:

(1) **Product specification:** when DDMs are utilised to predict a product's specification, the model's accuracy is influenced by the acceptable tolerance around the product specification. The tolerance may be determined by regulations, customer demands and/or economic feasibility.
(2) **Regulation(s):** manufacturers are subject to regulations regarding the accuracy to which they state the composition of their product, co-product and waste streams. For example, beers of an alcoholic strength not exceeding 5.5% volume that are sold within the European Union are subject to a 0.5% tolerance (European Union, 2011).
(3) **Economic:** when predicting an output that has a strong correlation to the economic feasibility of the process. Dependant on how precarious the process is between profit or loss making will influence how accurate the DDM is required to be. For example, when predicting the product yield of yield-driven processes. Yield-driven processes are productions processes whose economic feasibility are substantially influence by-product yield.
(4) **Safety:** when making predictions that influence the safety of the manufacturing system the DDM's accuracy is influenced by the severity and likelihood of the hazard. For example, when predicting hazardous gas dispersion under complex terrain conditions a high level prediction accuracy is required (Wang et al., 2018).

2.3 Process manufacturing data considerations

Data-driven models can be harnessed to support the development and implementation of new technologies and processes (Qiao et al., 2003). However, this requires the generation of a set of data, which can be costly and time-consuming (Sadati et al., 2018). Ensuring the data collected is representative of manufacturing the system is paramount to the performance of DDMs (Batista et al., 2004). To achieve this there are four questions to consider when generating a set of data:

(1) What data already exists, to what granularity (the scale and level of detail in the data) and how representative of the system is the data?
(2) What volume of data is required for model development?
(3) How was the data collected?
(4) What, if any, additional data will the model require and how will it be collected?

Manufacturing companies today are collecting vast troves of process data but typically use them only for monitoring purposes and after the event analysis, and not as a basis for improving process resilience (Sadati et al., 2018). As part of the data consideration stage, the modeller will need to decide on the optimal number of input and output variables the model will require. This is achieved through feature selection and

dimensionality reduction techniques, which improve the model's performance by identifying and removing noisy and/or irrelevant variables from the data (Sadati et al., 2018). Irrelevant variables are variables that have little to no impact on the model's output. Conversely, there exist process manufacturing systems that are not currently measuring/storing data to the required granularity to capture the system being modelled within the data. Therefore, a plan needs to be developed to decide which additional measurements and what volume of additional data is required, as well as how this data will be collected. For example, in CS1 the H^2AD bioprocess was already collecting 7 process performance variables but no data on the characteristics of the incoming or outgoing wastewater streams. Nineteen water quality variables were identified for initial collection, though with appreciation that not all of these would be truly relevant to modelling the output data. This made for a total of 26 input variables that was reduced to 9 by feature selection followed by principal component analysis. The volume of data is determined by three factors.

(1) The data needs to extend to, and preferably beyond, the system boundaries. Testing the model's capability to fit and predict beyond the system's boundaries is important to establish the confidence boundaries of the model.
(2) The duration period of data collection must be sufficient to capture any temporal variations observed in the system.
(3) There must be an equal distribution of the data within the system boundaries, to ensure the model is capable of fitting data and making predictions throughout the system.

If it is possible to collect data that achieves these three factors efficiently, then the volume of data required will be less. This can be done using Design of Experiments (DoE) techniques, such as central composite design and Box-Behnken design, to systematically collect data from across the system (Hamid et al., 2016). However, collecting data in industrial environments has numerous challenges (see Section 3) and a less intrusive method may be required, see (Section 3.3).

Data is collected either manually or automatically. Data collected manually is time-consuming and introduces the possibility of human error (Skoogh et al., 2012). Whereas, data collected automatically generally increases the volume of data recorded and removes human error (Skoogh et al., 2012) However there is a capital cost in purchasing and installing the required systems to collect the data automatically. Data-driven models often benefit from the automatic collection of data, as they can react in real-time to changes in the input variables. The DDM's response can feed into a larger online system creating a feedback loop. However, automatic data collection comes with challenges associated with handling large datasets. When handling a large dataset, the five V's are often employed to characterise the data (Addo-Tenkorang and Helo, 2016):

- **Volume:** refers to the size of the datasets.
- **Velocity:** is the speed to which data is generated, collected and analysed.
- **Variety:** is defined as the different types of data (e.g. time series data, image, audio, log files) incorporated into the datasets.
- **Value:** refers to the value extracted from the data.
- **Veracity:** the trustworthiness and quality of the data.

3. Challenges for data-driven process manufacturing models

Process manufacturers face unique challenges when collecting a set of data that is representative of the manufacturing system. The machine learning techniques utilised by DDMs have increased capabilities to fit data and make predictions when more data is provided (Kim, 2017). However, when ensuring there is sufficient data, and that it is representative of the system, there are a number of challenges in:

(1) The availability of process manufacturing data;
(2) The variability of process manufacturing systems;
(3) Establishing the system boundaries;
(4) Evaluating the model's output on unseen data.

3.1 Availability of manufacturing data

Data collection from real industrial processes faces a plethora of well-known challenges: sampling frequency, spatial representation of process environments, incomplete data, working conditions, sensor malfunction, communication exception or database shutdown, accuracy, etc. (Shang et al., 2014; Souza et al., 2016). This can result in poor quality data for model development that may require cleaning. Cleaning may be necessary to remove outliers within the data that are not representative of the system; however, the challenge is in knowing what data is and isn't representative of the system. Although DDMs developed from machine learning algorithms can overcome these challenges, large volumes of data are required (Qin, 2014). This is not always possible from process manufacturing systems, as they may be subject to additional challenges concerning the availability and variability of data. An established process manufacturing system may already have a wealth of historical data available for the initial training stage. However, this may not always be the case and the reasons for this may include:

(1) Modelling a new or adapted process, meaning no historical data representative of the current process exists;
(2) Variables (e.g. process conditions, external influences, new or changing feedstock) required for the model were not previously measured;
(3) Manufacturers may not have stored their data or not stored their data in a useable format;
(4) Manufacturers may produce a variety of products, meaning data on any one product is limited;
(5) Manufacturers may not keep potentially damaging data to the business if not required to.

Furthermore, industries that produce a variety of products from the same set of processing equipment (common in the pharmaceutical and food and drinks industries) will require the creation of a dataset for each of their products; as the process conditions, feedstocks and resources will be specific to that product. This data may not be immediately available, as these industries tend to produce a batch of one product then switch to another. This means the time between batches of the same product can be extensive delaying the development of the model. The CS2 faced these challenges, where craft breweries tend to ferment a range of different beers and a

model would need to be developed/adapted for each beer. The beer in question for this study is currently only brewed once a month so this means only 12 data sets could be recorded each year, which may not be sufficient for the machine learning models. Equally, for CS1 a new dataset will be required for optimising the H^2AD bioprocess's performance on new feedstocks (e.g. a different farm's waste, food waste).

The challenge of process manufacturing data availability will be less to manufacturers that utilise industrial control systems as part of their manufacturing systems. Systems like SCADA (supervisory control and data acquisition) have existed since the 1970's and are used to monitor and control a plant in industries such as wastewater treatment, waste management, energy, oil and gas refining and food production (Qin, 2014). Industries that make a consistent product and have extensively utilised industrial control systems (e.g. wastewater treatment or oil and gas refining) are well placed to recover value by developing a DDM to discovery knowledge within their historic data (Qin, 2014).

3.2 Variability in manufacturing systems

All manufacturing systems contain inherent variation (e.g. fluctuations in process temperatures, pressure, and flowrates, human operators, leakages) which can affect the performance of a DDM. This variability may be overcome by collecting sufficient quantity of relevant data as the variability will be captured within the model (Kay et al., 1999). However, this is not always possible for process manufacturers (see Section 3.1). Process manufacturers face the further challenge from the variations present in the resource flow, particularly when the resource flow is a waste stream. The feedstock physical and compositional variability can have a significant impact on the biochemical and thermochemical conversion to the final product (Williams et al., 2016). This variation can occur for a number of reasons, as summarised in Table 3.

Table 7. Causes of variation in process manufacturers' feedstock and some industry examples

Cause of variation	Industry example
1. Processes upstream	Water is an essential brewing ingredient. Inherent variability in the water treatment methods employed upstream has a direct impact on the water's characterises, and thus the final beer's characteristics (Simate, 2015).
2. Changes in operating practices upstream	Farming practices vary throughout the year. For example, the practices of spreading the farm waste stored in a slurry tank increases between the months of April and September, as per the EU waste framework Directive 91/676/EEC (European Union, 1991). This affects the slurry's characteristics (including total suspended solids, metal content and nutrient and organic load) and impacts processes utilising it as a feedstock, e.g. the H^2AD bioprocess.
3. Changes in supplier	Anaerobic digestion is an attractive option for the valorisation of food waste and other wastes with high organic load and calorific value. However, instability of anaerobic digesters is a common problem that can be exasperated by changes to the feedstock (Fisgativa et al., 2016). Food waste characteristics vary hugely dependant

	on its source, as well as being inherently variable (Fisgativa et al., 2016). Introducing a new feedstock to a digester can reduce biogas generation as the system takes time to adapt (Zhang et al., 2014).
4. Local Agronomic Conditions	Geographic location affects feedstock characteristics through variations in local agronomic conditions (Williams et al., 2016). The structure of corn harvested in the US has been shown to have a stronger correlation to geographic location than genetic variety (Templeton et al., 2009).
5. Cultivation and harvesting practises	There is extensive exploitation of plants for active ingredients for drug development (Ncube et al., 2012). The active ingredient derives their therapeutic effects from secondary metabolites, which is influenced by numerous natural factors (Ncube et al., 2012). The timing of harvesting and/or handling of the plant material also has an impact on plant quality (Ncube et al., 2012).
6. Seasonal	Sewage sludge generated by wastewater treatment plants is often used for agriculture as it recycles nutrients and organic matter to land. However, there is an environmental threat from the heavy metal content in the sludge (García-Delgado et al., 2007). The heavy metal content has been shown to vary by seasons and understanding this variability is important to minimise any negative environmental impact from using the sludge as a fertiliser (García-Delgado et al., 2007).
7. Storage and transportation	Potatoes degrade after they are harvested, at a rate determined by the storage: temperature, relative humidity, air circulation and gas composition (Eltawil et al., 2006). During transportation, the potatoes' quality is further reduced by bruising (Eltawil et al., 2006). Food manufacturers receive potatoes from multiple suppliers. Therefore, classification is required to determine if a potato quality is sufficient for further food processing (Lopez-Juarez et al., 2018).

During CS1, the H^2AD bioprocess was affected from variations in the feedstock supply, caused by changes to farm practice. There is an inherent variation present in the feedstock, as shown by the fluctuations in concentrations of four key water quality parameters in Fig. 2. The feedstock is stored in a slurry tank open to the environment, meaning during the winter months the total suspended solids (TSS) would decrease as the feedstock became diluted by increased rainfall.

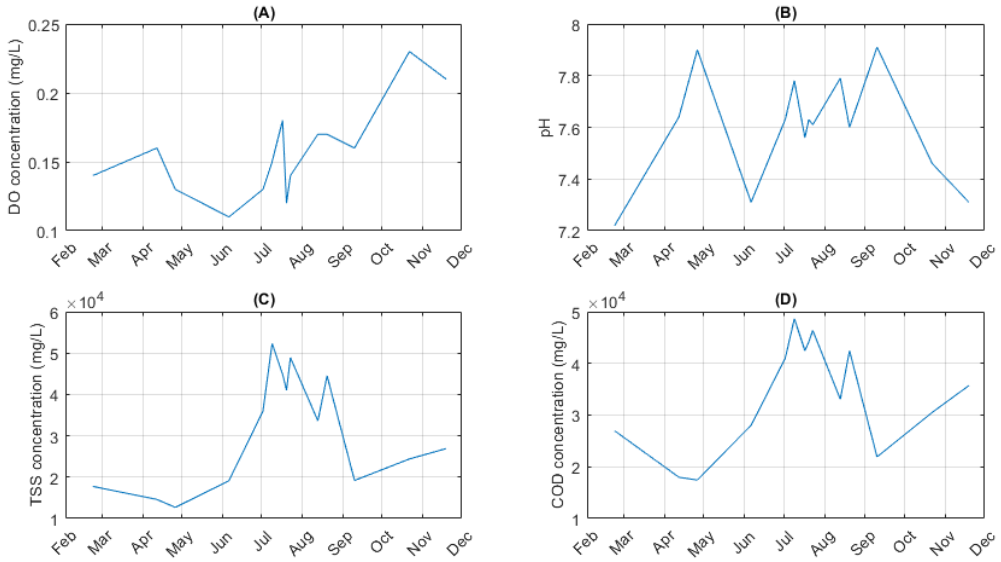


Fig. 7. Concentration of four water quality parameters of H^{2}AD feedstock during 2018: (A) dissolved oxygen (DO), (B) pH, (C) total suspended solids (TSS) and (D) chemical oxygen demand (COD).

There are further problems that arise from unexpected variability between batches of the same product. The brewery CS2 helps to illustrate the variability within supposedly identical processes. Fig. 3 shows the ultrasonic velocity, received ultrasonic signal amplitude and temperature as a function of fermentation time for four batches of a particular beer. Each of these measurements was recorded during different months in 2018, and are the input variables for a DDM predicting alcohol concentration during fermentation. The results indicate that the majority of variation within the signal is caused by temperature variations, which is typical of ultrasonic measurements. However, although the ultrasonic results show the same overall trends there are other variations caused by feedstock variability and marginally different brewing conditions (e.g. volume in the fermenter). These will affect the performance of the machine learning models and affect their accuracy. In Fig. 3, batch 4 shows a totally different temperature and speed of sound profile to the other batches. This was the result of a failure of the temperature control system during fermentation. Therefore, this batch was not representative of the system and is not suitable for training the model, further limiting the data available for model development.

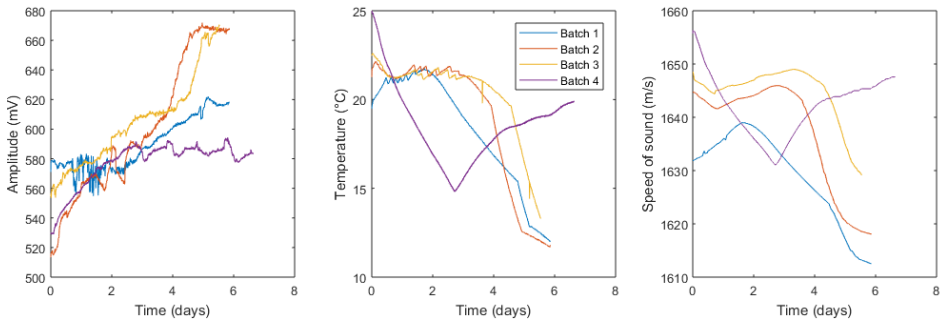


Fig. 8. Amplitude, temperature and speed of sound data during fermentation of four batches of Slap in the Face beer, collected from Totally Brewed in 2018.

3.3 Recognising the manufacturing system's boundaries

Predictions are a key feature of DDMs for process optimisation and intelligent decision-making. Data-driven models' predictive capabilities are generally strongest within the boundaries of the data used to develop them (Kim, 2017). If the data is only representative of a subsection of the manufacturing system, the model will fail at making accurate predictions of the entire system. Fig. 4 helps to visualise this problem. The model will be able to make accurate predictions for data inputted that is within the blue box but its prediction capability will likely decrease in accuracy the further away from the model boundaries. There are two methods to ensure the manufacturer can be confident the in model's results:

1. knowing the system boundaries and ensuring the model development data extends to these boundaries;
2. knowing the model boundaries and ensuring that only predictions made from this within the model development data range are acted upon.

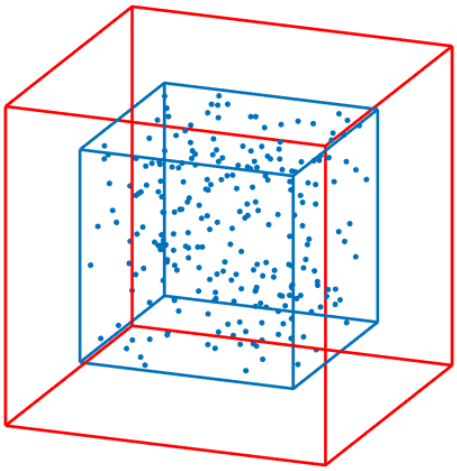


Fig. 9. Figure showing model development data that is only representative of one part of the whole manufacturing system. Where blue dots are model development data, the blue box represents model boundaries and the red box represents system boundaries.

Knowing the system boundaries of a process usually requires prior knowledge elicited from process industry practitioners (Shang et al., 2014). However, when modelling a new or changed system this knowledge may need reinforcement from trends identified within the model development data. Data visualisation plays an important role when understanding the space the system operates within (Lee and Ong, 1996). It is simple to visualise the boundaries' of a system of two or three dimensions, as demonstrated in Fig. 4. However, process manufacturing models typically contain data of higher dimensionality that will be impossible to visualise using conventional two-dimensional Euclidean space (Wang et al., 2004). There exist a number of approaches to plotting high dimensional data (Carr et al., 1987). Approaches like the scatter plot matrix can be utilised to plot the individual relationships between the variables alongside one another (Carr et al., 1987). The parallel coordinates system representation is particularly useful in visualising process manufacturing data in one plot (Wang et al., 2004). Wang et al. developed the Scan Circle algorithm that is able to identify regions of interest within the data that can be used to help describe the feasible region a DDM may operate successfully (Wang et al., 2002). Once the modeller is confident in the boundary conditions the model development data is selected from the historical data, ensuring that the model development data extends to these

boundaries. A scatter plot matrix was used to understand the model boundaries for the feedstock characteristics in CS1, a section is shown in Fig. 5. The diagram clearly shows that TSS is between 10,000 and 20,000 mg/L for the majority of the training data, implying that predictions for when TSS is greater than 20,000 are likely to be less accurate. It also indicates that there is a strong positive correlation between TSS and chemical oxygen demand (COD).

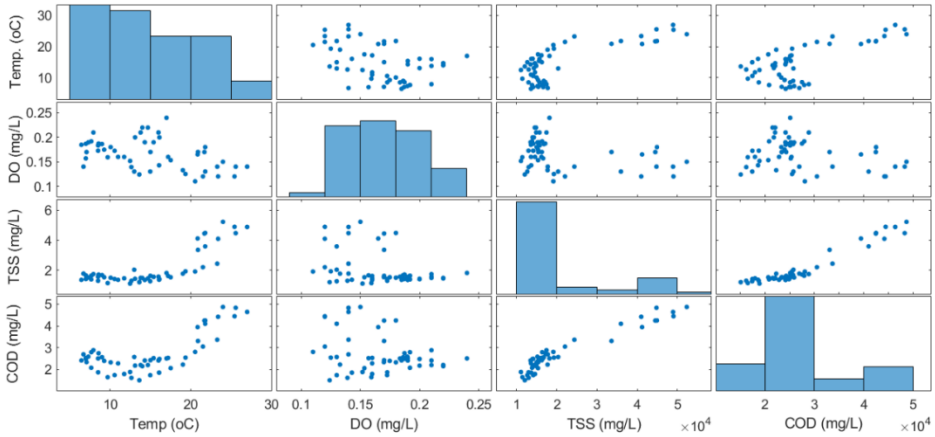


Fig. 10. Scatter plot matrix for four of case study 1's feedstock's characteristics [Temp.: temperature; DO: dissolve oxygen; TSS: total suspended solids; COD: chemical oxygen demand]

If no historical data exists, then data needs to be collected and it is likely that the system boundaries are yet to be defined. This can be overcome by collecting data over a suitable number of batches and/or timeframe of a continuous process, to be confident that the system is modelled within this data. This may be a time-intensive approach, as data may need to be collected for up to or beyond a year. For example, the process may be influenced by seasonal variations. An alternative method is to conduct a DoE, to determine a robust set of experiments that will define the system limits and generate data to extend to these boundaries (Hamid et al., 2016). By using DoE the data collection time and volume of data may be greatly reduced; however, the manufacturer may face additional costs from process disruptions necessary for experiment execution to achieve a robust dataset (Sadati et al., 2018). Dependant on the potential economic benefit of the model, it may be worth the additional cost or it may prove more feasible to conduct pilot-scale work to generate the data.

Prediction made by DDMs will always contain a degree of uncertainty. Uncertainty quantification is concerned with understanding the impact from the uncertainties inherent within the input data on the model's outputs (Iskandarani et al., 2016). There are a wide spectrum of techniques to consider when undertaking uncertainty quantification including: Bayesian methods, Latin hypercube sampling, polynomial chaos expansions, stochastic finite-element methods, Monte Carlo (Owhadi et al., 2013). Owhadi et al. proposed an optimal uncertainty quantification framework, which may be used as a guide for process manufacturers when accessing a DDM's predictive performance across the manufacturing system boundaries.

When implementing a new model into a system, careful monitoring is initially required to ensure the model is reflective of the system. Particularly if there is variability in the feedstock and other external factors. Once successfully installed, the model requires

monitoring to prevent a gradual degradation in performance. The decrease of the prediction quality is caused by the gradual and abrupt changes in the process (Kadlec and Gabrys, 2009). This can be avoided by retraining the model periodically as the availability and collection of data increases. Regular retraining often proves expensive and "online-learning" is an alternative approach to developing a DDM. Online learning is a different approach to machine learning where models continuously evolve as data becomes available sequentially in time (Chandrasekaran et al., 2012). This will be particularly effective for process manufacturers modelling data collected in real-time from control systems or IIoT.

3.4 Evaluating the model's output on unseen data

Model evaluation is necessary to determine how accurately the model reflects the system (Kim, 2017). Data is partitioned from the model development data to evaluate the model at different stages during its development. The model development data is partitioned into training, validation and testing datasets (Bishop, 2006). Training data is the data fitted to the model's algorithm, whilst validation data provides an evaluation of the model's fit to the training data and is used for tuning the algorithm's hyperparameters. A hyperparameter is an adjustable parameter that must be either manually or automatically tuned in order to obtain a model with optimal performance (Zeng and Luo, 2017). The testing data is used to evaluate of the model's fit on the model development data. While this may be sufficient when modelling systems whose boundaries are strongly defined, an additional evaluation will be required for systems whose boundaries are loosely defined. This will be likely for process manufacturers whose system may have challenges from data availability (Section 3.1) or from high levels of variability (Section 3.2). These systems will require additional evaluation using "unseen data", which is additional data not used during the development of a DDM. Unseen data is able to better evaluate the model's extrapolation capabilities beyond the model development data (Panerati et al., 2019). If the accuracy of the model does not meet the requirements defined by the manufacturer (see Section 2.2) the model must undergo redevelopment either through optimisation of the algorithm's architecture or the collection of additional data (Keviczky and Banyasz, 2015).

Process manufacturers that face data availability and variability challenges are more susceptible to overfitting and poor generalisation, as capturing these challenges increases the likelihood of that the model will exploit relationships within the data that do not describe the manufacturing system. Evaluating the model on unseen data is essential to avoid overfitting and ensure good generalisation, yet this stage has often been overlooked (Hamid et al., 2016). Unseen data has two sources:

1. **Partitioning model development data:** split data into model development data (training, validation and testing data) and unseen data.
2. **Experimental data:** an experiment is performed to generate unseen data that covers the range of the model boundaries. Design of experiments can be used to ensure the parameter range for the experiment is comprehensive.

Partitioning data is the most common approach (Liu and Cocea, 2017). The data partitioning can be performed randomly or by using a fixed method. However the data is partitioned, the unseen validation data must extend evenly to cover the boundaries of the model. This is to ensure the model is able to both fit and predict data through

the manufacturing system. The unseen data should also aim to go beyond the limits of the model. By using data beyond the model boundaries the modeller gains knowledge of the model's capability to fit and predict outside of the training data. An example of good and poor selection of unseen data is shown in Fig. 6.

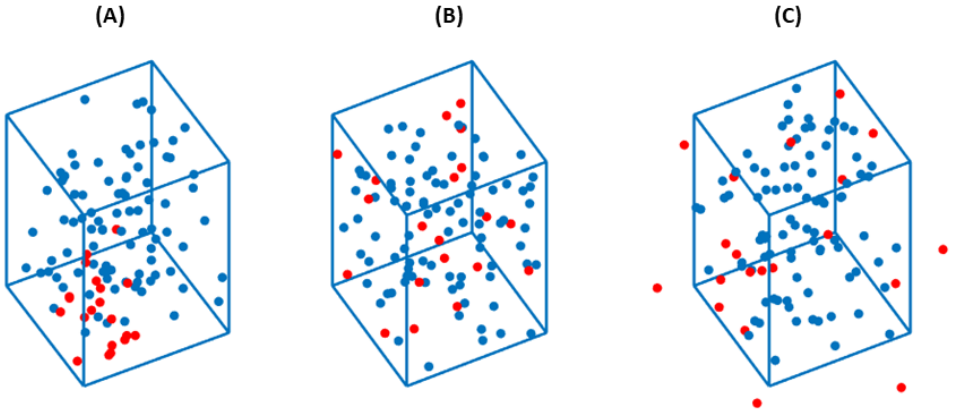


Fig. 11. (A) Example of unseen data only evaluating one region of the model, (B) example of unseen data evaluating the whole of the model, and (C) evaluating the whole of the model and the extrapolating capabilities of the model. Where blue dots are training data, red dots are unseen data and blue box is the model boundaries.

4. Opportunities arising from data-driven modelling of process manufacturing environments

In an age where manufacturing is under pressure to reduce cost and environmental impact, while maintaining product quality, the utilisation of data-driven modelling is a promising tool for process manufacturers. When the Aberdeen Group surveyed 223 global manufacturing organisations, 47% believed they needed to become more data-driven to remain competitive (Geiger, 2017). Four areas have been identified as an opportunity for process manufacturers to strengthen their manufacturing systems through the utilisation of DDMs:

(1) Utilising data-driven models to improve process manufacturing models;
(2) Enabling affordable characterisation of process streams;
(3) Ensuring greater process resilience;
(4) Evaluating waste valorisation potential.

4.1 Utilising data-driven models to improve process manufacturing models

The application of DDMs to support and improve traditional modelling tasks has been well established. Examples exist of their use for design and planning, control and monitoring, optimisation and safety (Boukouvala et al., 2016; Krenczyk, 2012; Pan and Hu, 2016; Wang et al., 2011). Furthermore, techniques are being designed to integrate data-driven and first principal models into a hybrid model structure (von Stosch et al., 2014). Within process manufacturing hybrid models have been developed for various application including chemical reactors (Azarpour et al., 2017), polymerization processes (Fiedler and Schuppert, 2008), crystallisation (Nicoletti et al., 2009), metallurgic processes (Hu et al., 2011) and distillation columns (Caballero, 2015).

Hybrid models are built by combining the predictions of a first principle model and a DDM into a single model. Hybrid models embrace the benefits of both techniques and overcome the disadvantages of both models (Azarpour et al., 2017). There are multiple methods by which the models may cooperate (von Stosch et al., 2014):

- **Proxy:** one model acts as a surrogate for the other;
- **Complement:** the solution is a combination of the two models;
- **Supplement:** one model provides a correction for the other model;
- **Embedment:** one model is embedded within the other model;
- **Integrate:** the output of one model serves as an input for the other model;
- **Inspiration:** the structure of one model is developed from knowledge provided by the other model.

Hybrid modelling is considered the state-of-the-art modelling techniques to model complex manufacturing systems (Barbosa and Azevedo, 2017). When compared to individual models, hybrid models tend to have a higher prediction accuracy, better calibration properties, enhanced extrapolation capabilities and better interpretability than DDMs (von Stosch et al., 2014). Process manufacturers that rely on existing first-principle models have an opportunity to improve the models results and prediction accuracy by developing and integrating a DDM into the existing model. However, manufacturers will face additional challenges in developing hybrid models. Developing hybrid models requires knowledge about different modelling techniques and flexibility from modellers to find a good fit between models (Barbosa and Azevedo, 2017). When deciding whether the development of a hybrid model is justified over a single model the manufacturer should considered if the model goal and requirements demands a hybrid modelling approach. If the model accuracy can be achieved through conventional single modelling techniques then the additional time and expertise demanded by a hybrid model is unjustified.

4.2 Enabling affordable characterisation of process streams

One of the main benefits of DDMs is their ability to enable real-time decisions by the collection of real-time manufacturing data (Chaturvedi et al., 1993). However, process manufacturing's feedstock, product and waste streams are not a set of specific discrete components. Instead, they are a multi-component and/or multi-phase matrix subject to inherent variation. Therefore, real-time online characterisation is expensive and sometimes unfeasible. This results in important process variables, like product quality, being measured infrequently offline (Yan et al., 2017). Because of the infrequency in which these measurements are performed, continuous direct monitoring of process streams' composition is not possible (Slišković et al., 2011).

Industrial processes are often described as *data rich but information poor.* There is often a large quantity of process data from conventional measurements (pressure, temperature, flow rate, etc.) but a lack of data on stream compositions. This is due to a lack of suitable and affordable technologies capable of performing these measurements non-invasively, online and in real-time. (Dong and Mcavoy, 1996). To overcome this, process manufacturers rely on soft sensors, which are able to model data collected from conventional measurements and used to predict key variables (Kleinert et al., 2011). As more novel IIoT technologies are introduced, the variety of

data available to train soft sensors shall expand. One such example is a multi-sensor system designed to optimise the Clean-In-Place (CIP) process, which utilises both visual and ultrasonic data (Simeone et al., 2018, 2016). By utilising DDMs, soft sensors are able to compute not only a wider range of data but also unclean data (Qin, 2014). This will allow for the development of cheaper sensors, as the requirement for high-quality data shall decrease, instead being replaced by volume of data (Qin, 2014). However, the data must always be representative of the system.

Being able to track the relevant composition and physical characteristics of a process stream in real-time, will become even more important to process manufacturers as regulations may change so that manufacturers have more responsibility to record and share data for compliance. Combating fraud in certain process manufacturing industries is an ongoing issue, with a particular focus on the food and drinks sector (Manning and Soon, 2016). The introduction of a greater number of affordable soft sensors would help build trust across the supply chain and with consumers. If this is to be successful, ensuring that these sensors are not susceptible to malicious interference or corrupted with false data is paramount.

4.3 Ensuring greater process resilience

As with all suppliers, process manufacturers are required to meet certain targets to be commercially successful and adhere to regulations. Paramount among these, from a manufacturing process viewpoint, is meeting minimum product specification (e.g. yield and quality) (Gani, 2004). Ensuring consistency for these targets is challenging as process manufacturers are subjected to a higher level of variability than other industries (Gani, 2004). For example, in the biopharmaceutical industry yields can vary from 50% to 100% for no immediately discernible reason (Sadati et al., 2018). In case study 2, the unexplained variation between batches, shown in Fig. 3 (Section 3.2), results in disparities in the final alcohol concentration in the product. By taking observational data (meaning data routinely collected by a manufacturing process) and developing a DDM, the variables that are having a significant effect on the industry yields can be identified and optimised (Sadati et al., 2018). The advantage of using data-driven techniques this way is that the DDM can intelligently identify significant variables responsible for the variability. This is achieved by using unsupervised machine learning techniques for exploratory data analysis to find hidden patterns or grouping in data (Bishop, 2006).

Although it is possible to model and limit the impact of some variability within a processing plant (Sadati et al., 2018), a substantial challenge for process manufacturers is limiting the impact of variables outside of an engineer's control. Paramount of these variables is the variability in the process's feedstock (see Section 3.2). This variability can have a significant impact on product quality and yield. For example, the production of anti-malarial drugs which is reliant on the active component artemisinin harvested from a plant called *Artemisia annua* (Pilkington et al., 2014). The plant's artemisinin concentration varies dependant on when it is harvested during its lifecycle and the storage conditions (Pilkington et al., 2014). By fitting historical harvesting, transportation and process data, a DDM would determine the optimal harvesting and storage point to maximise artemisinin recovery. The model could also be used to predict the feedstock's potential to produce artemisinin and whether to adjust the price for the supplier as not all feedstock is created equally.

Variation in the feedstock is not the only concern for a process manufacturer. The waste produced by process manufacturing is equally variable, in regards to the volume produced, the concentration of valuable components and pollutants (Parlikar et al., 2016). Governments and regulatory bodies around the world set maximum limits on what manufacturers may emit to the environment (European Union, 2010). Manufacturers may choose to treat their waste onsite, so it is within these limits, pay a third party to handle their waste or partially treat their waste to reduce cost. There is a cost involved with all options and this cost may fluctuate with waste characteristics. Being able to predict and accommodate these variations ahead of time may mean measures can be taken to reduce this cost. Process manufacturing waste is unique from discrete manufacturing in that there exists the potential for waste valorisation (Arancon et al., 2013). However, one of the barriers to waste valorisation is the detrimental effect variation in the waste has on the technologies in terms of performance. Case study 1 is an example of a waste valorisation technology that utilises a DDM to increase the resilience of the bioprocess. By collecting a set of training data over a 1-year period and fitting it to a model, it was possible to capture the variability of the feedstock due to seasons, changes in farm practices and inherent variation. The model is able to identify which key feedstock characteristics variations were detrimental to the bioprocess's performance. The bioprocess's process conditions were also varied throughout the year, in order to investigate their effect on performance. By combining and fitting this data to a model, it was possible to predict how to optimise the bioprocess conditions dependant on the incoming feedstock characteristics.

4.4 Evaluating waste valorisation potential

Increased competition for access to critical resources is a major concern for the manufacturing industry. In 2017, the EU expanded its list of critical raw materials (CRM) to 27, defined as materials considered to be of high importance to the EU economy and of high risk to their supply (European Commission, 2017). Greater focus has been placed on developing solutions towards implementing the circular economy into manufacturing (Lieder and Rashid, 2016). The Waste and Resources Action Program charity offers a clear definition of the circular model: "*A circular economy is an alternative to a traditional linear economy (make, use, dispose) in which we keep resources in use for as long as possible, extract the maximum value from them whilst in use, then recover and regenerate products and materials at the end of each service life.*" (Waste and Resources Action Program, 2018). The circular economy is being pursued by Governments, with Europe leading the way by implementing the 2018 Circular Economy Action Plan to create a set of measures to help manufacturers to adopt circular economy systems (Springer and Schmitt, 2018). Encouraging manufacturers to stop viewing waste as a *waste* but instead as *co-product* is an important step towards this goal.

Process manufacturing generates a wide range of solid, liquid and gas waste streams that have excellent potential for waste valorisation within a circular economy. Waste valorisation refers to industrial processing activities aimed at reusing, recycling, or recovering useful products or sources of energy from waste (Kabongo, 2013). There are currently three waste valorisation strategies, geared towards (1) production of fuel and/or energy to replace fossil fuels; (2) production and/or extraction of high-value

chemicals from residues; and (3) production of other useful material(s) (Arancon et al., 2013).

Process manufacturers have a variety of options when evaluating which waste valorisation route to take. Choosing which route to take depends on a number of factors: composition, volume and variability of the waste stream, capital and operational costs of onsite treatment versus valorisation by third party, pre-treatment costs, other waste being produced by nearby manufacturers, handling and transportation costs. Due to the large number of factors involved, it can be time and cost-intensive for a manufacturer to employ someone to evaluate the most sustainable route. There exists a number of Enterprise Resource Planning (ERP) programs, that use DDMs, capable of simulating the manufacturing process to intelligently determine the most sustainable waste valorisation route (The Access Group, 2018). These platforms are available via the cloud and systems like IBM Watson IoT are capable of integrating the manufacturer's existing ERP programs into their software (International Business Machines, 2016). However, while these platforms are capable of planning the logistics and costs, it has not been the focus to predict the performance of these different waste valorisation technologies at treating a waste stream. There is a great opportunity to have models incorporated into the process to take a holistic view, considering waste minimisation and/or valorisation alongside the key product objectives.

Conclusion and further work

The use of DDMs is on the rise across manufacturing primarily due to the following three points:

(1) There is an increasing amount of data generated in manufacturing as IIoT technologies become more widespread.
(2) Manufacturers have easier access to the computational power required to harness this data due to cloud computing enabling parallel processing.
(3) New machine learning techniques are enabling the utilisation of new sources of data, such as texts, image, audio, video, log files.

Before attempting to develop a DDM there are certain points the manufacturer must consider. The model's goal must be clearly defined and should aim to solve the manufacturing problem that has been identified. The manufacturer should also consider what is required from the model. Is it enough simply to fit the data, or will the model be required to make predictions? Being able to make predictions increases the model's value but also its complexity. The accuracy in which the model is required to perform these functions should be defined by the product specification, regulations, economic value and safety requirements.

Paramount to the success of DDMs is the generation/collection of a representative set of data used to develop the model. This data is split into model development data, used to train the model, and unseen validation data to evaluate the prediction capability of the model. Before generating/collecting this data the manufacturer should consider what data is already available and if it is representative of the system. Challenges arise when ensuring a representative data set is used for model development. Process manufacturers face unique challenges regarding this point

because of the limited availability and variability of industrial data. Even when data is available, knowing if it representative of the system and relevant to the model can be challenging. Furthermore, the boundaries of a process manufacturing system are hard to classify. The difficulty occurs in knowing when enough data has been collected to be representative of the whole system. Once a manufacturer is able to overcome the considerations and challenges involved when modelling a process manufacturing system, DDMs will allow the process manufacturers to:

(1) Improve process manufacturing models: By integrating DDMs with existing first-principle process manufacturing models into a hybrid model, which have been demonstrated to have greater prediction accuracy and extrapolation capabilities.

(2) Characterise process streams: Through the development of soft sensors that utilise recent advances in machine learning able to utilise cheaper, unclean data. This will allow process manufacturers to develop a deeper understanding of their processes, providing optimisation opportunities in a cost-effective manner.

(3) Enhance process resilience: By fitting data to the DDM, manufacturers are able to make accurate predictions on the effect, variability across the system, has on their process and take measures to optimise them.

(4) Evaluate waste valorisation routes: Existing DDMs may be used as a tool to evaluate new waste streams compatibility with an existing waste valorisation technology. Alternatively, use DDMs as a tool for manufacturers to see which technology is the most sustainable option for their waste.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing and [grant number EP/P001246/1], Network Plus: Industrial Systems in a Digital Age.

Acknowledgements

The author (RLG) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing. The author (NJW) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) Network Plus: Industrial Systems in a Digital Age [grant number EP/P001246/1], BREWNET: Intelligent Cloud Connected Sensors for Economic Small Scale Process Optimisation. The author (OJF) would also like to acknowledge the University of Nottingham Faculty of Engineering for his PhD scholarship.

References

Addo-Tenkorang, R., Helo, P.T., 2016. Big data applications in operations/supply-chain management: A literature review. Comput. Ind. Eng. 101, 528–543. https://doi.org/10.1016/J.CIE.2016.09.023

Al-Aufi, Y.A., Hewakandamby, B.N., Dimitrakis, G., Holmes, M., Hasan, A., Watson, N.J., 2019. Thin film thickness measurements in two phase annular flows using

ultrasonic pulse echo techniques. Flow Meas. Instrum. 66, 67–78. https://doi.org/10.1016/J.FLOWMEASINST.2019.02.008

Almeida, J.S., 2002. Predictive non-linear modeling of complex data by artificial neural networks. Curr. Opin. Biotechnol. 13, 72–76. https://doi.org/10.1016/S0958-1669(02)00288-4

Angria S, L., Sari, Y.D., Zarlis, M., Tulus, 2018. Data-driven Modelling for decision making under uncertainty, in: IOP Conference Series: Materials Science and Engineering. p. 12013.

Arancon, R.A.D., Lin, C.S.K., Chan, K.M., Kwan, T.H., Luque, R., 2013. Advances on waste valorization: new horizons for a more sustainable society. Energy Sci. Eng. 1, 53–71. https://doi.org/10.1002/ese3.9

Arce, D., Lima, F., Orellana Cordero, M.P., Ortega, J., Sellers, C., Ortega, P., 2018. Discovering behavioral patterns among air pollutants: A data mining approach. Enfoque UTE 9, 168–179. https://doi.org/10.29019/enfoqueute.v9n4.411

Atzmueller, M., Hayat, N., Schmidt, A., Kloepper, B., 2017. Explanation-aware feature selection using symbolic time series abstraction: Approaches and experiences in a petro-chemical production context, in: Proceedings - 2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017. Institute of Electrical and Electronics Engineers Inc., pp. 799–804. https://doi.org/10.1109/INDIN.2017.8104875

Azarpour, A., N.G. Borhani, T., R. Wan Alwi, S., A. Manan, Z., I. Abdul Mutalib, M., 2017. A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. Chem. Eng. Res. Des. 117, 149–167. https://doi.org/10.1016/j.cherd.2016.10.024

Barbosa, C., Azevedo, A., 2017. Hybrid Simulation for Complex Manufacturing Value-chain Environments. Procedia Manuf. 11, 1404–1412. https://doi.org/10.1016/j.promfg.2017.07.270

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explor. Newsl. 6, 20–29. https://doi.org/10.1145/1007730.1007735

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Boukouvala, F., Li, J., Xiao, X., Floudas, C.A., 2016. Data-driven modeling and global optimization of industrial-scale petrochemical planning operations, in: 2016 American Control Conference (ACC). pp. 3340–3345. https://doi.org/10.1109/ACC.2016.7525433

Brandl, D., 2007. 1. Manufacturing Control. Des. Patterns Flex. Manuf.

Caballero, J.A., 2015. Logic hybrid simulation-optimization algorithm for distillation design. Comput. Chem. Eng. 72, 284–299. https://doi.org/10.1016/j.compchemeng.2014.03.016

Carr, D.B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S., 1987. Scatterplot Matrix Techniques for Large N. J. Am. Stat. Assoc. 82, 424–436. https://doi.org/10.1080/01621459.1987.10478445

Chandrasekaran, M., Muralidhar, M., Krishna, C.M., Dixit, U.S., 2012. Online machining optimization with continuous learning, in: Computational Methods for Optimizing Manufacturing Technology: Models and Techniques. pp. 85–110.

Charte, F., Romero, I., Pï¿½rez-Godoy, M.D., Rivera, A.J., Castro, E., 2017. Comparative analysis of data mining and response surface methodology predictive models for enzymatic hydrolysis of pretreated olive tree biomass. Comput. Chem. Eng. https://doi.org/10.1016/j.compchemeng.2017.02.008

Chaturvedi, A.R., Hutchinson, G.K., Nazareth, D.L., 1993. Supporting complex real-

time decision making through machine learning. Decis. Support Syst. 10, 213–233. https://doi.org/10.1016/0167-9236(93)90039-6

Coley, C.W., Green, W.H., Jensen, K.F., 2018. Machine Learning in Computer-Aided Synthesis Planning. Acc. Chem. Res. 51, 1281–1289. https://doi.org/10.1021/acs.accounts.8b00087

Cross, M., Moscardini, A.O., 1985. Learning the Art of Mathematical Modelling, Ellis Horwood series in mathematics and its applications. Statistics and operational research. John Wiley & Sons, Inc., Chichester [Chichestershire] : New York, NY, USA.

Dong, D., Mcavoy, T.J., 1996. Nonlinear principal component analysis - Based on principal curves and neural networks. Comput. Chem. Eng. 20, 65–78. https://doi.org/10.1016/0098-1354(95)00003-K

Duarte, B., Saraiva, P.. M., C., P.C., 2004. Combined Mechanistic and Empirical Modelling. Int. J. Chem. React. Eng. https://doi.org/10.2202/1542-6580.1128

Edmonds, D., 2016. Before the Taps Run Dry: Incentivizing Water Sustainability in America's Craft Breweries. Georg. Washingt. J. Energy Environ. Law 7, 164–176.

Eltawil, M.A., Samuel, D.V.K., Singhal, O.P., 2006. Potato Storage Technology and Store Design Aspects. Agric. Eng. Int. CIGR J. 8, 1–18. https://doi.org/10.1093/hmg/ddw317

Escrig, J., Woolley, E., Rangappa, S., Simeone, A., Watson, N.J., Escrig, Escrig, J., Woolley, E., Rangappa, S., Simeone, A., Watson, N.J., 2019. Clean-in-place monitoring of different food fouling materials using ultrasonic measurements. Food Control 104, 358–366. https://doi.org/10.1016/J.FOODCONT.2019.05.013

European Commission, 2017. 2017 List of critical raw materials for the EU.

European Union, 2011. Regulation (EU) No 1169/2011 on the provision of food information to consumers. Off. J. Eur. Union 18–63. https://doi.org/10.1109/60.911397

European Union, 2010. Directive 2010/75/EU of the European Parliament and of the Council of 24 November 2010 on industrial emissions (integrated pollution prevention and control), OJ L 334.

European Union, 1991. Directive 91/676/CEE concerning the protection of waters against pollution caused by nitrates from agricultural sources. Off. J. Eur. Communities 1–8. https://doi.org/10.1017/CBO9781107415324.004

Fayyad, U., Stolorz, P., 1997. Data mining and KDD: Promise and challenges. Futur. Gener. Comput. Syst. 13, 99–115. https://doi.org/10.1016/s0167-739x(97)00015-0

Fiedler, B., Schuppert, A., 2008. Local identification of scalar hybrid models with tree structure. IMA J. Appl. Math. (Institute Math. Its Appl. 73, 449–476. https://doi.org/10.1093/imamat/hxn011

Fisgativa, H., Tremier, A., Dabert, P., 2016. Characterizing the variability of food waste quality: A need for efficient valorisation through anaerobic digestion. Waste Manag. 50, 264–274. https://doi.org/10.1016/J.WASMAN.2016.01.041

Fisher, O., Watson, N., Porcu, L., Bacon, D., Rigley, M., Gomes, R.L.R.L., 2018. Cloud manufacturing as a sustainable process manufacturing route. J. Manuf. Syst. 47. https://doi.org/10.1016/j.jmsy.2018.03.005

Gani, R., 2004. Chemical product design: challenges and opportunities. Comput. Chem. Eng. 28, 2441–2457. https://doi.org/10.1016/J.COMPCHEMENG.2004.08.010

García-Delgado, M., Rodríguez-Cruz, M.S., Lorenzo, L.F., Arienzo, M., Sánchez-Martín, M.J., 2007. Seasonal and time variability of heavy metal content and of its chemical forms in sewage sludges from different wastewater treatment plants.

Sci. Total Environ. 382, 82–92. https://doi.org/10.1016/J.SCITOTENV.2007.04.009

Garment, V., 2014. 3 Ways to Test the Accuracy of Your Predictive Models [WWW Document]. URL https://www.kdnuggets.com/2014/02/3-ways-to-test-accuracy-your-predictive-models.html (accessed 11.15.18).

Ge, Z., 2017. Review on data-driven modeling and monitoring for plant-wide industrial processes. Chemom. Intell. Lab. Syst. 171, 16–25. https://doi.org/10.1016/J.CHEMOLAB.2017.09.021

Geiger, D., 2017. Data-Driven Manufacturing – Monetizing the Analytical Edge [WWW Document]. URL https://www.aberdeen.com/opspro-essentials/data-driven-manufacturing-monetizing-analytical-edge/ (accessed 11.7.18).

Hamid, H.A., Jenidi, Y., Somerfield, C., Gomes, R.L., 2016. Predicting the capability of carboxylated cellulose nanowhiskers for the remediation of copper from water using response surface methodology (RSM) and artificial neural network (ANN) models. Ind. Crops Prod. 93, 108–120. https://doi.org/10.1016/J.INDCROP.2016.05.035

Hangos, K.M., Cameron, I.T. (Eds.), 2001. 1 - The Role of Models in Process Systems Engineering, in: Process Modelling and Model Analysis, Process Systems Engineering. Academic Press, pp. 3–18. https://doi.org/https://doi.org/10.1016/S1874-5970(01)80025-4

Harding, J.A., Shahbaz, M., Srinivas, Kusiak, A., 2006. Data mining in manufacturing: A review. J. Manuf. Sci. Eng. Trans. ASME. https://doi.org/10.1115/1.2194554

Hu, G., Mao, Z., He, D., Yang, F., 2011. Hybrid modeling for the prediction of leaching rate in leaching process based on negative correlation learning bagging ensemble algorithm. Comput. Chem. Eng. 35, 2611–2617. https://doi.org/10.1016/j.compchemeng.2011.02.012

International Business Machines, 2016. IBM Watson Internet of Things (IoT) [WWW Document]. URL https://www.ibm.com/internet-of-things (accessed 1.10.19).

Iskandarani, M., Wang, S., Srinivasan, A., Carlisle Thacker, W., Winokur, J., Knio, O.M., 2016. An overview of uncertainty quantification techniques with application to oceanic and oil-spill simulations. J. Geophys. Res. Ocean. 121, 2789–2808. https://doi.org/10.1002/2015JC011366

Kabongo, J.D., 2013. Waste Valorization, in: Idowu, S.O., Capaldi, N., Zu, L., Gupta, A. Das (Eds.), Encyclopedia of Corporate Social Responsibility. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 2701–2706. https://doi.org/10.1007/978-3-642-28036-8_680

Kadlec, P., Gabrys, B., 2009. Soft sensors: Where are we and what are the current and future challenges?, IFAC Proceedings Volumes (IFAC-PapersOnline). IFAC. https://doi.org/10.3182/20090921-3-TR-3005.00098

Kay, J.W., Titterington, D.M., Kay, S.L.S.J.W., 1999. Statistics and Neural Networks: Advances at the Interface, Royal Statistical Society Lecture Notes Series, 5. Oxford University Press.

Keviczky, L., Banyasz, C., 2015. 10. Process Identification, in: Two-Degree-of-Freedom Control Systems - The Youla Parameterization Approach. Elsevier, pp. 309–315.

Kim, S., 2017. MATLAB Deep Learning With Machine Learning, Neural Networks and Artificial Intelligence, 1st ed. Apress. https://doi.org/10.1007/978-1-4842-2845-6

Kleinert, T., Schladt, M., Muehlbeyer, S., Schocker, A., 2011. Combination of Process Analytical Technology with Soft Sensors for Online Process Data and Advanced Process Information. TM-TECHNISCHES Mess. 78, 589–602. https://doi.org/10.1524/teme.2011.0165

Krause, D., Schöck, T., Hussein, M.A., Becker, T., 2011. Ultrasonic characterization of aqueous solutions with varying sugar and ethanol content using multivariate regression methods. J. Chemom. 25, 216–223. https://doi.org/10.1002/cem.1384

Krenczyk, D., 2012. Data-driven modelling and simulation for integration of production planning and simulation systems. Sel. Eng. Probl. 119–122.

Lauer, F., Bloch, G., 2008. Incorporating prior knowledge in support vector machines for classification: A review. Neurocomputing 71, 1578–1594. https://doi.org/10.1016/j.neucom.2007.04.010

Lee, H.-Y., Ong, H.-L., 1996. Visualization support for data mining. IEEE Expert. Syst. their Appl. 11, 69–75. https://doi.org/10.1109/64.539019

Lieder, M., Rashid, A., 2016. Towards circular economy implementation: a comprehensive review in context of manufacturing industry. J. Clean. Prod. 115, 36–51. https://doi.org/10.1016/J.JCLEPRO.2015.12.042

Lin, C.S.K., Pfaltzgraff, L.A., Herrero-Davila, L., Mubofu, E.B., Abderrahim, S., Clark, J.H., Koutinas, A.A., Kopsahelis, N., Stamatelatou, K., Dickson, F., Thankappan, S., Mohamed, Z., Brocklesby, R., Luque, R., 2013. Food waste as a valuable resource for the production of chemicals, materials and fuels. Current situation and global perspective. Energy Environ. Sci. 6, 426–464. https://doi.org/10.1039/c2ee23440h

Liu, H., Cocea, M., 2017. Semi-random partitioning of data into training and test sets in granular computing context. Granul. Comput. 2, 357–386. https://doi.org/10.1007/s41066-017-0049-2

Lopez-Juarez, I., Rios-Cabrera, R., Hsieh, S.J., Howarth, M., 2018. A hybrid non-invasive method for internal/external quality assessment of potatoes. Eur. Food Res. Technol. 244, 161–174. https://doi.org/10.1007/s00217-017-2936-9

Luo, S., Sun, H., Ping, Q., Jin, R., He, Z., 2016. A Review of Modeling Bioelectrochemical Systems: Engineering and Statistical Aspects. Energies 9, 111. https://doi.org/10.3390/en9020111

Manning, L., Soon, J.M., 2016. Food Safety, Food Fraud, and Food Defense: A Fast Evolving Literature. J. Food Sci. 81, R823–R834. https://doi.org/10.1111/1750-3841.13256

Mariscal, G., Marban, O., Fernandez, C., 2010. A survey of data mining and knowledge discovery process models and methodologies. Knowl. Eng. Rev. 25, 137–166. https://doi.org/10.1017/S0269888910000032

Mathews, P., 2004. DOE Language and Concepts, in: Design of Experiments With MINITAB. ASQ Quality Press, Milwaukee, pp. 93–142.

Microsoft, 2018. Azure Machine Learning Studio [WWW Document]. URL https://azure.microsoft.com/en-gb/services/machine-learning-studio/ (accessed 10.24.18).

Ncube, B., Finnie, J.F., Van Staden, J., 2012. Quality from the field: The impact of environmental factors as quality determinants in medicinal plants. South African J. Bot. 82, 11–20. https://doi.org/10.1016/J.SAJB.2012.05.009

Nicoletti, M.C., Jain, L.C., Giordano, R.C., 2009. Computational intelligence techniques as tools for bioprocess modelling, optimization, supervision and control. Stud. Comput. Intell. https://doi.org/10.1007/978-3-642-01888-6_1

Ning, C., You, F., 2018. Data-driven stochastic robust optimization: General computational framework and algorithm leveraging machine learning for optimization under uncertainty in the big data era. Comput. Chem. Eng. https://doi.org/10.1016/j.compchemeng.2017.12.015

OECD, 2012. OECD Environmental Outlook to 2050, OECD Environmental Outlook.

OECD Publishing. https://doi.org/10.1787/9789264122246-en
Ojha, V.K., Abraham, A., Snášel, V., 2017. Metaheuristic design of feedforward neural networks: A review of two decades of research. Eng. Appl. Artif. Intell. 60, 97–116. https://doi.org/10.1016/j.engappai.2017.01.013
Owhadi, H., Scovel, C., Sullivant, T.J., McKerns, M., Ortiz, M., 2013. Optimal uncertainty quantification. SIAM Rev. 55, 271–345. https://doi.org/10.1137/10080782X
Pan, Y., Hu, M., 2016. A data-driven modeling approach for digital material additive manufacturing process planning, in: 2016 International Symposium on Flexible Automation (ISFA). pp. 223–228. https://doi.org/10.1109/ISFA.2016.7790165
Panerati, J., Schnellmann, M.A., Patience, C., Beltrame, G., Patience, G.S., 2019. Experimental methods in chemical engineering: Artificial neural networks–ANNs. Can. J. Chem. Eng. 97, 2372–2382. https://doi.org/10.1002/cjce.23507
Parlikar, U., Bundela, P.S., Baidya, R., Ghosh, S.K., 2016. Effect of Variation in the Chemical Constituents of Wastes on the Co-processing Performance of the Cement Kilns. Procedia Environ. Sci. 35, 506–512. https://doi.org/10.1016/J.PROENV.2016.07.035
Pasini, A., 2015. Artificial neural networks for small dataset analysis. J. Thorac. Dis. 7, 953–960. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61
Perry, R.H., Green, D.W., 2008. Perry's chemical engineers' handbook. McGraw-Hill, New York.
Pilkington, J.L., Preston, C., Gomes, R.L., 2014. Comparison of response surface methodology (RSM) and artificial neural networks (ANN) towards efficient extraction of artemisinin from Artemisia annua. Ind. Crops Prod. 58, 15–24. https://doi.org/10.1016/J.INDCROP.2014.03.016
Qiao, G., Riddick, F., McLean, C., 2003. New Manufacturing Modeling Methodology: Data Driven Design and Simulation System Based on XML, in: Proceedings of the 35th Conference on Winter Simulation: Driving Innovation, WSC '03. Winter Simulation Conference, pp. 1143–1148.
Qin, S.J., 2014. Process data analytics in the era of big data. AIChE J. 60, 3092–3100. https://doi.org/10.1002/aic.14523
Rasmuson, A., Andersson, B., Olsson, L., Andersson, R., Olisson, L., Andersson, R., 2014. Empirical model building, in: Mathematical Modeling in Chemical Engineering. Cambridge Univerisity Press, Cambridge, pp. 40–52. https://doi.org/10.1017/CBO9781107279124.005
Resa, P., Elvira, L., de Espinosa, F., 2004. Concentration control in alcoholic fermentation processes from ultrasonic velocity measurements. Food Res. Int. 37, 587–594. https://doi.org/10.1016/j.foodres.2003.12.012
Rizos, V., Behrens, A., Kafyeke, T., Hirschnitz-Garbers, M., Ioannou, A., 2015. The Circular Economy: Barriers and Opportunities for SMEs. CEPS Work. Doc. 1–25. https://doi.org/0006-8993(94)90176-7 [pii] ET - 1994/02/04
Sadati, N., Chinnam, R.B., Nezhad, M.Z., 2018. Observational data-driven modeling and optimization of manufacturing processes. Expert Syst. Appl. 93, 456–464. https://doi.org/10.1016/j.eswa.2017.10.028
Saka, M.P., Doğan, E., Aydogdu, I., 2013. Analysis of Swarm Intelligence–Based Algorithms for Constrained Optimization. Swarm Intell. Bio-Inspired Comput. 25–48. https://doi.org/10.1016/B978-0-12-405163-8.00002-8
Schichl, H., 2004. Models and the History of Modeling, in: Josef, K. (Ed.), Modeling Languages in Mathematical Optimization. Springer US, Boston, pp. 25–36. https://doi.org/10.1007/978-1-4613-0215-5
Shafique, U., Qaiser, H., 2014. A Comparative Study of Data Mining Process Models

(KDD, CRISP-DM and SEMMA). Int. J. Innov. Sci. Res. 12, 217–222.
Shang, C., Yang, F., Huang, D., Lyu, W., 2014. Data-driven soft sensor development based on deep learning technique. J. Process Control 24, 223–233. https://doi.org/10.1016/j.jprocont.2014.01.012
Shearer, C., 2000. The CRISP-DM model: the new blueprint for data mining. J. Data Warehous. 5, 13–22.
Simate, G.S., 2015. Water treatment and reuse in breweries. Brew. Microbiol. 425–456. https://doi.org/10.1016/B978-1-78242-331-7.00020-4
Simeone, A., Deng, B., Watson, N., Woolley, E., 2018. Enhanced Clean-In-Place Monitoring Using Ultraviolet Induced Fluorescence and Neural Networks. Sensors 18. https://doi.org/10.3390/s18113742
Simeone, A., Watson, N., Sterritt, I., Woolley, E., 2016. A Multi-sensor Approach for Fouling Level Assessment in Clean-in-place Processes. Procedia CIRP 55, 134–139. https://doi.org/10.1016/J.PROCIR.2016.07.023
Simmonds, J., 2017. Number of breweries in the UK breaks through the 2,000 barrier for the first time since the 1930s [WWW Document]. URL https://www.uhy-uk.com/news-events/news/number-of-breweries-in-the-uk-breaks-through-the-2000-barrier-for-the-first-time-since-the-1930s/ (accessed 12.19.18).
Skoogh, A., Perera, T., Johansson, B., 2012. Input data management in simulation – Industrial practices and future trends. Simul. Model. Pract. Theory 29, 181–192. https://doi.org/10.1016/J.SIMPAT.2012.07.009
Slišković, D., Grbić, R., Hocenski, Ž., 2011. Methods for Plant Data-Based Process Modeling in Soft-Sensor Development. Automatika 52, 306–318. https://doi.org/10.1080/00051144.2011.11828430
Solomatine, D., See, L M, Abrahart, R J, 2008. Data-Driven Modelling : Concepts , Approaches and Experiences, in: Abrahart, Robert J., See, Linda M., Solomatine, D.P. (Eds.), Practical Hydroinformatics. Water Science and Technology Library. Springer, Berlin, pp. 17–31. https://doi.org/https://doi.org/10.1007/978-3-540-79881-1
Soroush Rohanizadeh, S., Moghadam, M.B., 2009. A Proposed Data Mining Methodology and its Application to Industrial Procedures, Journal of Industrial Engineering.
Souza, F.A.A., Araújo, R., Mendes, J., 2016. Review of soft sensor methods for regression applications. Chemom. Intell. Lab. Syst. 152, 69–79. https://doi.org/10.1016/j.chemolab.2015.12.011
Springer, N.P., Schmitt, J., 2018. The price of byproducts: Distinguishing co-products from waste using the rectangular choice-of-technologies model. Resour. Conserv. Recycl. 138, 231–237. https://doi.org/10.1016/J.RESCONREC.2018.07.034
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Templeton, D.W., Sluiter, A.D., Hayward, T.K., Hames, B.R., Thomas, S.R., 2009. Assessing corn stover composition and sources of variability via NIRS. Cellulose 16, 621–639. https://doi.org/10.1007/s10570-009-9325-x
The Access Group, 2018. Access ERP systems for all industries [WWW Document]. URL https://www.theaccessgroup.com/supply-chain/industries/ (accessed 1.11.19).
von Stosch, M., Oliveira, R., Peres, J., Feyo de Azevedo, S., 2014. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Comput. Chem. Eng. https://doi.org/10.1016/j.compchemeng.2013.08.008

Wang, J., Chang, Q., Xiao, G., Wang, N., Li, S., 2011. Data driven production modeling and simulation of complex automobile general assembly plant. Comput. Ind. 62, 765–775. https://doi.org/10.1016/J.COMPIND.2011.05.004

Wang, K., Salhi, A., Fraga, E., 2002. Cluster identification using a Parallel Coordinate System for knowledge discovery and nonlinear optimization, in: 12th European Symposium on Computer Aided Process Engineering (ESCAPE-12).

Wang, K., Salhi, A., Fraga, E.S., 2004. Process design optimisation using embedded hybrid visualisation and data analysis techniques within a genetic algorithm optimisation framework. Chem. Eng. Process. Process Intensif. 43, 657–669. https://doi.org/10.1016/J.CEP.2003.01.001

Wang, R., Chen, B., Qiu, S., Zhu, Z., Wang, Yiduo, Wang, Yiping, Qiu, X., 2018. Comparison of Machine Learning Models for Hazardous Gas Dispersion Prediction in Field Cases. Int. J. Environ. Res. Public Health 15. https://doi.org/10.3390/ijerph15071450

Waste and Resources Action Program, 2018. WRAP and the circular economy | WRAP UK [WWW Document]. URL http://www.wrap.org.uk/about-us/about/wrap-and-circular-economy; (accessed 11.15.18).

Williams, C.L., Westover, T.L., Emerson, R.M., Tumuluru, J.S., Li, C., 2016. Sources of Biomass Feedstock Variability and the Potential Impact on Biofuels Production. BioEnergy Res. 9, 1–14. https://doi.org/10.1007/s12155-015-9694-y

Yan, W., Tang, D., Lin, Y., 2017. A data-driven soft sensor modeling method based on deep learning and its application. IEEE Trans. Ind. Electron. 64, 4237–4245. https://doi.org/10.1109/TIE.2016.2622668

Yin, S., Kaynak, O., 2015. Big Data for Modern Industry: Challenges and Trends [Point of View]. Proc. IEEE 103, 143–146. https://doi.org/10.1109/JPROC.2015.2388958

Zeng, X., Luo, G., 2017. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Heal. Inf Sci Syst 5. https://doi.org/10.1007/s13755-017-0023-z

Zhang, C., Su, H., Baeyens, J., Tan, T., 2014. Reviewing the anaerobic digestion of food waste for biogas production. Renew. Sustain. Energy Rev. 38, 383–392. https://doi.org/10.1016/J.RSER.2014.05.038

APPENDIX D: PUBLICATION 4

A data-driven modelling approach for process manufacturing systems: An industrial bioprocess case study

Oliver FISHER[1], Nicholas J WATSON[1], Laura PORCU[2/3], Darren BACON[2], Martin RIGLEY[2], Rachel L GOMES[1]

[1] Food, Water, Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
[2] Lindhurst Engineering Ltd., Midland Road, Sutton in Ashfield, Nottinghamshire, NG17 5GS, United
Kingdom
[3] Energy Innovation & Collaboration, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, United Kingdom

Corresponding author
Author: Rachel L Gomes
Tel.: +44 115 846 8883.
E-mail address: rachel.gomes@nottingham.ac.uk

Abstract

A data-driven model (DDM) may be fitted to process manufacturing data to increase understanding about a system and may be further used for predictive analysis. Data-driven models are experiencing greater use across process manufacturing. This is due to the recent surge of data generated by industrial processes and increasing computational power available by cloud computing. There are frameworks in place to guide in the development of a DDM. However, these frameworks are not always suitable for modelling process manufacturing systems that face unique modelling challenges from data availability and variability. This paper expands the widely followed cross-industry standard process for data mining (CRISP-DM) framework. Two additional steps are added to the modelling stage which is to define the system boundaries and to evaluate the model on unseen data. Unseen data is data not used at any stage during model development. These additions to the CRIPS-DM framework are necessary if a process manufacturer plans to use the model to make predictions on future data as it will (1) evaluate the model's interpolation and extrapolation capabilities and (2) inform when to retrain the model.

The methodologies developed by this paper were demonstrated and evaluated by building a DDM of an industrial bioprocess. Four machine learning algorithms were evaluated trained and the algorithms' predictive capabilities were statistically evaluated on unseen data using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The random forest model had the highest accuracy on unseen data (R^2: 0.99; RMSE: 1.44; MAE: 2.81; MAPE: 6.04).

Highlights

- A data-driven models (DDMs) predictions can aid manufacturers decision making
- However, this relies on the DDM being representative of the system
- Process manufacturers face challenges from data availability and variability
- To aid process manufacturers a new modelling framework is proposed
- The framework is evaluated by modelling an industrial bioprocess

Keywords

Data-driven models, Industry 4.0, Process manufacturing, Artificial neural networks, random forest, support vector regression, Gaussian processes

1. Introduction

Manufacturers are reliant on mathematical models to describe the complex behaviour of their processes [1]. Models have been heavily utilised in process manufacturing to describe the thermal and bio/chemical processes at the heart of process manufacturing systems [2]. There are two fundamentally different categories of models: first-principal models and empirical, or data-driven models. First-principal modelling is the construction of a sequence of equations describing a system's components that, when assembled, describe the complete manufacturing system [3]. Data-driven modelling fits manufacturing data to algorithms to discover knowledge about a system and/or make predictions about the system [4]. Deciding which model to use is dependent on the level of knowledge known about the manufacturing system and quantity and granularity of data available. First principal models require a greater level of understanding of the system than data-driven models (DDMs) and are generally more time consuming to build [5]. However, the accuracy of a DDM's predictions relies on the modelling data being representative of the manufacturing system [6].

The application of data-driven modelling in manufacturing has progressed dramatically over the last two decades [7]. This rise can be attributed to two trends of

[A] *bbreviations:* NH_4^+, Ammonium; ANN, Artificial Neural Network; BES, Bioelectrochemical Systems; COD, Chemical Oxygen Demand; Cl^-, Chloride; R^2, Coefficient of Determination; Cond., Conductivity; Cu, Copper; DDM, Data-Drive Model; GP, Gaussian Process; HRT, Hydraulic Residence Time; MEA, Mean Average Error; MAPE, Mean Average Percentage Error; NCA; Neighbourhood Component Analysis; NO_3^-, Nitrate; NO_2^-, Nitrite; PCC, Pair-copula Construction; PCP, Parallel Coordinate Plot; PO_4^{3-}, Phosphate; PCA, Principal Component Analysis; PDF, Probability Density Function; RF, Random Forest; RMSE, Root Mean Squared Error; SO_4^{2-}, Sulphate; SVR, Support Vector Regression; SP, System Pressure; ST, System Temperature; TC, Total Carbon; TIC, Total Inorganic Carbon; TOC, tN; Total Nitrogen; Total Organic Carbon; TSS, Total Suspended Solids; WH, Water Hardness; WQA, Water Quality Analysis; Zn, Zinc;

the 21st Century; (1) cloud computing enabling a substantial increase in computer processing power and (2) increase in volume and variety of data being generated. Data generation by the manufacturing industry is overwhelming; in 2015 more than 1000 Exabytes of manufacturing data was produced worldwide and by 2025 this number is expected to increase 20-fold [8]. However, there are sectors of manufacturing where data generation and collection are inherently limited. For example, in the food and drink industry, the same manufacturing equipment tends to be used to produce a range of products. This means that time between batches of the same product can be extensive, reducing the data available to develop a model for the production of that specific product. Furthermore, the use of legacy equipment that do not feature as many sensors is limiting data collection.

Notwithstanding data limitation challenges, DDMs have great potential to deliver process efficiency. Previous work has demonstrated DDMs' potential to: increase product yields of pharmaceutical processes [9], reduce resource consumption in clean-in-place systems[10], enhance oil recovery from crude oil reservoirs [11], optimise the yield of a continuous flow reactor [12], maximise the harvest of active components of plants [13], develop soft-sensors for industrial processes [14], predicting corrosion in pipelines [15] and increasing efficiency of wastewater treatment processes [16].

The widely accepted framework for developing a DDM is the cross-industry standard process for data mining (CRISP-DM) [17]. The CRISP-DM is composed of six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The CRISP-DM is an iterative process and each phase may be revisited as more information is made available. Following the CRIPS-DM methodology aims to ensure a representative DDM is developed that has industry relevance and impact [18]. A stated advantage of the CRISP-DM is that it is applicable across various industries [17]. However, process manufacturers face some unique challenges from data availability and variability that the current CRISP-DM method fails to address, as it is has been designed to handle large-scale data mining problems [17]. These challenges may present problems in developing a DDM that is representative of the process manufacturing system.

There is no single definition of a manufacturing system. For this work, a manufacturing system is defined as input streams which pass through a process changing their nature into output streams, which may consist of multiple products and/or waste materials. The input and output streams and the manufacturing process can be monitored by measuring various parameters and these measurements can be used to develop a DDM of the system. Defining the boundaries of the system is challenging because of the variability present within manufacturing systems. When defining the system boundaries, a decision needs to be made on how much of the variability is required to be captured within the system. For example, Fig. 1 shows a manufacturing system whose boundaries have been defined to capture different levels of variability within the system. By extending the boundaries to capture more of the system's variability results in regions that are not as well defined by the data, which may negatively impact

on the predictive capabilities of the DDM. Therefore, how much variability is necessary to capture varies between manufacturing environments and the level of accuracy required of the DDM.

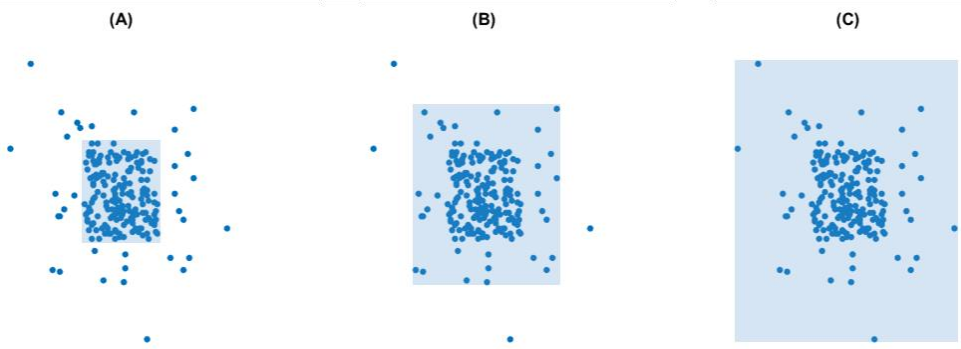


Fig. 12. Blue region represents the defined manufacturing system for: (A) Strongly represented system, weak variability is captured; (B) Moderately represented system, moderate variability captured; (C) All variability captured but system contains poorly represented regions.

For the model's data to be representative of the system, the data needs to not only extend to the system boundaries but also describe the space within. The majority of DDMs in the literature are developed for experimental or pilot-scale systems [19]; therefore, making it possible to control the system parameters to ensure data is evenly distributed across the whole system. This can be achieved by conducting a design of experiments [19]. However, when attempting to model from an operational manufacturing system it is not always possible to interrupt production and a non-intrusive method is required. An option often used is to train the model on historical manufacturing data [20]. Previous work often assumes that due to the large volume of historical data it is inherently representative of the current manufacturing system and little attempt is given to establish the boundaries of the system within the historical data [20]. Not only is this not possible for manufacturing systems where data is limited, but it also ignores the fact that manufacturing systems are continuously evolving. Manufacturing systems are generally described by high dimensional data increasing the complexities of extracting the key features of the data or visualising the system boundaries [21]. Exploratory data analysis may be performed to better understand the manufacturing system and to assist in defining the boundaries of the system.

There are several guides in the literature for developing DDMs of process manufacturing systems. Curteanu et al. studied three different modelling strategies to study the electrolysis process in wastewater treatment [22]. In this work, the authors demonstrated how a DDM's prediction capabilities may be improved by developing multiple DDMs of a system stacked together [22]. The separate work presented by Charte et al. [23] and Kashani et Shahhosseini [24] both explained the benefits of effective data partitioning to find a DDM best configurations. Kashani et Shahhosseini also demonstrated that applying dimensionality reduction techniques to datasets with a large number of input variables can improve the performance of DDMs [24]. What is

often overlooked in the literature is the importance of evaluating the DDM's prediction capabilities on unseen data [25]. Unseen data is data that has not been used during the development of the DDM. Traditionally testing data is partitioned from the model development data and used to evaluate the model's performance on future data and to avoid overfitting [26]. Overfitting is the generation of a model that corresponds too closely or exactly to the noise (error) within the dataset, which negatively impacts future predictions [27]. However, using only testing data to evaluate a model's prediction capability on future data is not sufficient for process manufacturing systems that face data availability and variability challenges. Unseen data is also required to evaluate what portioning of the manufacturing system does the model represent and the model's extrapolating capabilities beyond the model development data. Hamid et al. showed how additional experiments may be performed to generate unseen data to evaluate a model extrapolation capabilities beyond the model development data [19]. However, when modelling historical manufacturing data this may not be possible. A method is required to partition unseen data from the historical data that can be used to thoroughly evaluate the DDM's model predictive capabilities throughout and beyond the system.

This research proposes a methodology for modelling process manufacturing systems that face data availability constraints. To achieve this, it is proposed to add two new steps to the modelling phase of the CRISP-DM framework, as shown in Fig. 2. The first is to use exploratory data analysis to clearly define the system boundaries and the spread of data within the boundaries. The second is to use unseen data to evaluate how well the developed model captures the variability of the process manufacturing system. The proposed methodology is demonstrated by developing a DDM for an industrial case study that faces data availability and variability challenges. The case study aims to develop a DDM describing the novel H^2AD bioprocess developed by Lindhurst Engineering Ltd., which is described in Section 1.1.

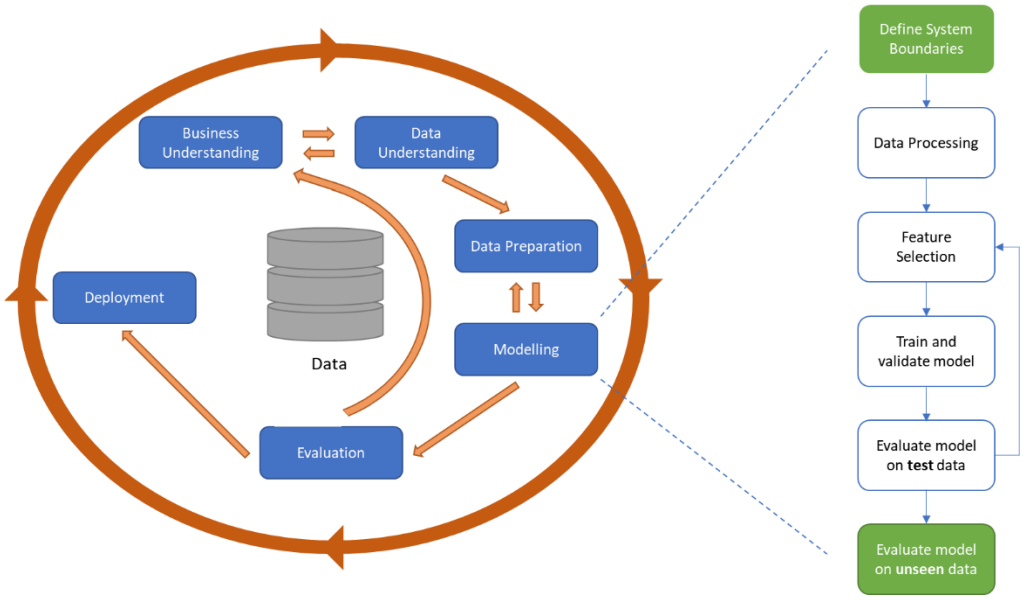


Fig. 13. The proposed CRISP-DM framework has been expanded to include two new steps (the green boxes) in the modelling stage.

1.1 H^2AD Case Study

Process manufacturers are under pressure to reduce manufacturing costs and an effective waste/water management strategy is paramount. Water demand in the manufacturing sector is expected to rise by 400% between 2012 and 2050 [28]. The manufacturing system studied in this work is a bioprocess developed by Lindhurst Engineering Ltd targeted at treating a SMEs manufacturer's wastewaters, due to its modular, low-cost design. At the heart of the bioprocess is the Micro H^2AD technology, which combines bioelectrochemical systems (BES) and anaerobic digestion technologies. Bioelectrochemical systems are unique systems capable of converting chemical energy into electrical energy (and vice-versa) by employing microbes as catalysts [29]. The H^2AD bioprocess is capable of treating a variety of wastewaters, from sources including agriculture, brewing, soft drinks, foods, bio-manufacture residues, to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy.

Bioelectrochemical systems are sensitive to variability within the environmental and operating conditions, which can negatively impact the efficiency of the system to remove pollutants and generate energy [30]. This is also true of the H^2AD bioprocess, whose performance varies with changes in the wastewater characteristics (temperature, pH, composition, etc.) and H^2AD operating conditions. The performance is defined as the percentage removal of wastewater pollutants and volume of biogas generated. Currently little is known on how and to what extent the characteristics of the wastewater have on the H^2AD bioprocess's performance. The development of a DDM will be used to analyse and predict the effect that varying wastewater characteristics have on the bioprocess performance. Furthermore, the model may be able to aid the commercialisation of the H^2AD product by predicting the bioprocess's ability to handle new wastewater feedstocks.

2. Material and methods

2.1 Manufacturing system

The manufacturing system is an operating H^2AD plant treating wastes produced by a high-performance dairy farm in the East Midlands of the UK. The farm waste contains cattle slurry, bedding waste, waste milk, footbath, parlour washing and rainfall. The farm waste is collected and stored in a 3 million litre slurry tank. The farm waste is fed into the H^2AD system that generates bioenergy for the farm and treats the wastewater fit for reuse on the farm. The steps outlined in Fig. 2 were followed to develop a DDM of the system.

2.2 Model data

The model was developed from labelled input and output data. The model input data originates from two sources, the H^2AD feedstock data and the H^2AD process data. To

characterise the H^2AD feedstock, water quality analysis (WQA) was performed on samples collected entering the H^2AD plant. A total of 18 WQA parameters were measured and are detailed in Appendix A. In addition to this, the H^2AD unit automatically collected data on four H^2AD process conditions via sensors connected by internet of things (IoT) technologies to a database hosted on a cloud platform. These were hydraulic retention time (HRT), pump speed, system temperature (ST) and system pressure (SP). The H^2AD outputs were characterised by the percentage removal of COD.

From previous work, there are 30 data-points available of WQA performed on the farm waste samples, taken from the slurry tank feeding the H^2AD system. In addition, there was one year's worth of historical H^2AD process data automatically collected by the unit. This informed the decision to generate an additional synthetic 22 WQA data points bringing the total to 52 data points, representing one year's worth of weekly sampling. When generating synthetic data the standard method is to randomly sample from a probability density function (PDF) fitted to existing data [31]. Pair-copula constructions (PCCs) are a popular tool for generating synthetic data from multivariate distributions, due to their simple structure and high flexibility [32]. Due to insufficient data relating input variables to the model's output variables, it was not possible to follow the same PDF method. The approach for generating synthetic output data followed that in Brissette et al., whereby trends identified from historical H^2AD data governed the development of non-linear equations to generate synthetic output data from the input data [33].

2.2 Defining system boundaries

Exploratory data analysis was performed to understand the manufacturing system represented in the data, using a combination of statistical and data visualisation techniques. The model development data was analysed to determine the system's features: central tendency, correlations, distribution, spread, modality, and extents. Of particular interest was defining the feedstock characteristics data and COD removal, as compared to the H2AD process data, these are subject to variability that is beyond the manufacturer's ability to control. The statistical measures employed were: mean, medium, standard deviation, Pearson's correlation coefficient, skewness, range, interquartile points, maximum, and minimum. Data visualisation plays an important role in detecting patterns, trends, and outliers in groups of data [21]. Visualising manufacturing systems' data is made challenging because of their high dimensional nature [21]. The data visualisation techniques capable of plotting high dimensional data utilised were: scatter plot matrix, box plots, heatmaps, and parallel coordinate plots. In order to plot variables of different magnitudes, the data was first normalised by Eq. 1, as described in Section 2.2.

2.3 Data processing

For this study data was integrated from two sources, WQA performed of H^2AD input and output streams samples and H^2AD process data automatically collected remotely by the unit. Data normalisation was undertaken to ensure the impact of variables of lower magnitudes are given the same weight by the model's algorithm as the variables of larger magnitudes. The COD and TSS data have magnitudes within the order of 10,000, while for DO is the magnitude is much lower and in the order of 0.1. To normalise the data the minimax function was applied, as this normalises the variables without any loss of information [20]. The equation to normalise the data is given below:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where x is the variable, x_{norm} is the normalised variable and x_{min} a*nd* x_{max} is the minimum and maximum values of the variable being normalised respectively.

2.4 Feature selection and dimensionality reduction

When attempting to develop a model with limited data points one risks the *curse of dimensionality*. The *curse of dimensionality* is the name given to the phenomena observed that as the number of input variables increases the volume of data required to avoid overfitting increases exponentially [20]. Two methods exist to avoid this problem; *feature selection* and *dimensionality reduction*. Feature selection reduces the number of variables in the dataset, keeping only the most relevant variables [34]. For this work, the machine learning technique neighbourhood component analysis (NCA) was utilised to identify the variables that had a minimum impact on the model output. The dimensions were then further reduced by the dimensionality reduction technique, principal component analysis (PCA). A PCA is a mathematical procedure that transforms potentially correlated data into an orthogonal system of linearly uncorrelated principal components (PC) [15].

2.5 Data partitioning

The data used to develop a DDM is then partitioned at several levels, as independent data is required to train, validate, test and evaluate the model, see Fig. 3. The data is first partitioned into 80% model development data and 20% unseen data from within and outside the system boundaries. To avoid sample representativeness issues the semi-random partitioning framework developed by Liu et Cocea 2017 was employed to partition the unseen data points within the system boundaries [35]. The model development data is further randomly partitioned into training, validation and testing data, following the widely applied ratio 70:15:15 respectively [36].

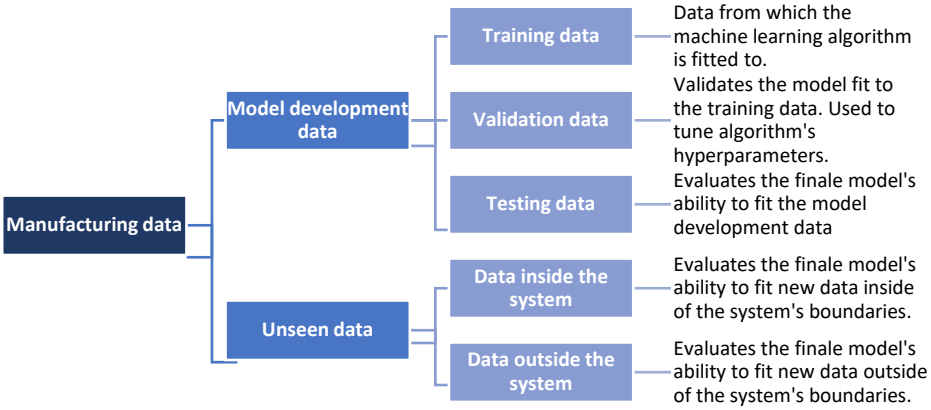


Fig. 14. Hierarchy of data partitioning during the development of a data-driven model.

2.6 Algorithm selection, training, validation and testing

For this work, four machine learning algorithms suitable for regression were investigated: Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural networks (ANN). Descriptions of each algorithm are provided in the supplementary material and for further in depth reading the readers are referred to The Elements of Statistical Learning by Hastie et al [37]. These algorithms were chosen as they have been shown to have stronger predictive powers over other algorithms like logistic regression [38,39].

When developing a DDM, each algorithm has model parameters and hyperparameters that require tuning to find the optimal configuration. Hyperparameters are parameters that are set before being trained on a model, compared to model parameters that are learnt during training [26]. For example, the number of neurons in an ANN's hidden layer is a hyperparameter; whereas, the weights between neurons is a model parameter. Hyperparameters are parameters that are set before being trained on a model, compared to model parameters that are learnt during training [26]. The training, validation and testing of the DDMs and hyperparameter optimisation were carried out by MATLAB R2017.b software developed by MathWorks.

2.7 Model evaluation

The prediction capabilities of the various models were statistically evaluated in terms of the coefficient of determination (R2), the root of the mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The R^2 is the square of the sample correlation coefficient between observed values and predicted values, and is a measure of the explained variance by the model [15]. The RMSE measures the mean square magnitude of the error [40], while MAE measures the absolute magnitude of the errors [41]. The MAPE is similar to MAE but is based on

percentage error, being that it is scaled independently and can be used to compare model performance across different data sets [41]. The statistically best model is the one that has an R^2 closest to 1 while minimising RMSE, MAE and MAPE. The R^2, RMSE, MAE and MAPE are defined as follows:

$$R^2 = \frac{(\sum_{i=1}^{n}(p_i - \bar{p}_i)(a_i - \mathrm{a}_i))^2}{\sum_{i=1}^{n}(y_a - \bar{y}_a)^2\left(y_p - \bar{y}_p\right)^2} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(a_i - p_i)^2}{n}} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{n}|a_i - p_i|}{n} \tag{4}$$

$$MAPE = 100 \times \frac{\sum_{i-1}^{n}\left|\frac{a_i - p_i}{a_i}\right|}{n} \tag{5}$$

Where n is the number of points, p_i is the predicted value, a_i is the actual observed value, and the symbol ¯ is the average of the related values.

3. Results and discussion

3.1 Generation of synthetic data

The Clayton copula method was used to produce synthetic input data [32]. The features of the new dataset were compared against the original to assess how successful the Clayton copula synthetic data captured the relationships in the original. The percentage change for mean, median and standard deviation for each variable from the original to the new dataset was calculated. A summary is provided by the mean percentage changes for the mean, median and standard deviation of the variables, which are 2.78%, 2.77% and 7.72% respectively. Therefore, the central point of the data has shifted by in a positive direction by 2.78% and that the spread of the data has increased by 7.72%. By comparing the Pearson correlation coefficients in the original dataset and synthetic dataset, shown in Fig. 4, it can be seen that the data generated widely captures the relationships in the original data. The complete input data is provided in the supplementary material.

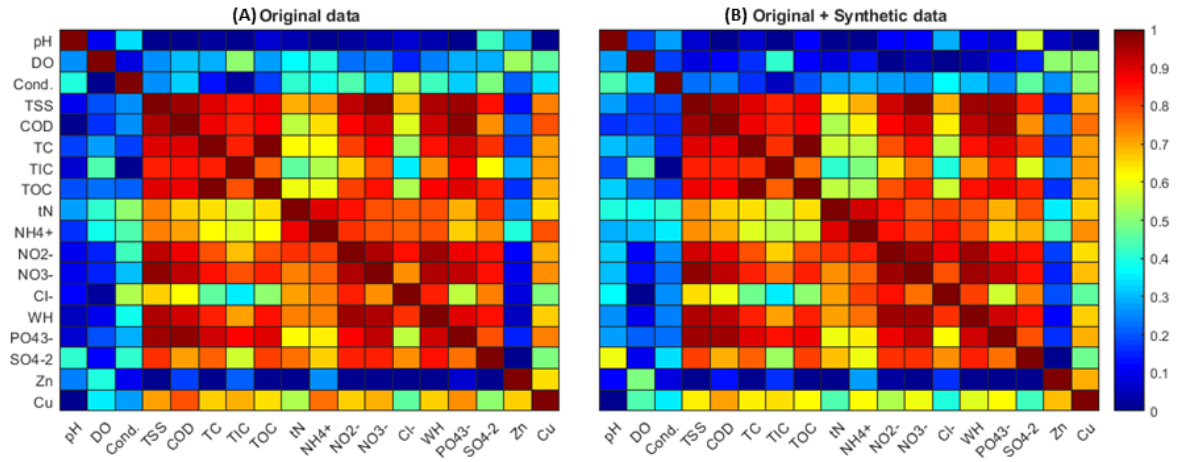


Fig. 15. Heatmaps displaying the scaled correlation coefficients between variables in (A) the original WQA data and (B) the original plus synthetic data.

The Pearson correlation coefficients were calculated from the historical data to explore the relationship between the model input data and the percentage of COD removed and are presented in Table 2. The correlations were then used to guide the development of synthetic output data, which is provided in the supplementary material.

Table 8. Correlation strength between model input variables and COD removal efficiency, where r is Pearson's correlation coefficient.

	Positive	**Negative**
Strong |0.7 < r < 1.0|	HRT, System temperature	DO, Cond., pH, WH
Moderate |0.35 < r <.7|	--	TIC, Cl-, TC, Zn, NO2-, NH4+, TOC, SO4-2, tN, NO3-, TSS, Cu, PO43-, Pump speed, system pressure
Weak |0 < r <.0.35|	--	COD

3.2 Data processing

An NCA for regression identified 7 input variables (TSS, COD, TC, TIC, NO_3^-, Cl^- and PO_4^{-3}) to remove, displayed in Fig. 5 (A), so that the H^2AD input variables are reduced from 22 to 15. The standard practice is to reduce the number of variables so that 95% of the variance within the dataset is captured by the PC [15]. By applying PCA to the H^2AD dataset, the input variables were further reduced from 15 to 9 PCs, displayed in Fig. 4(B).

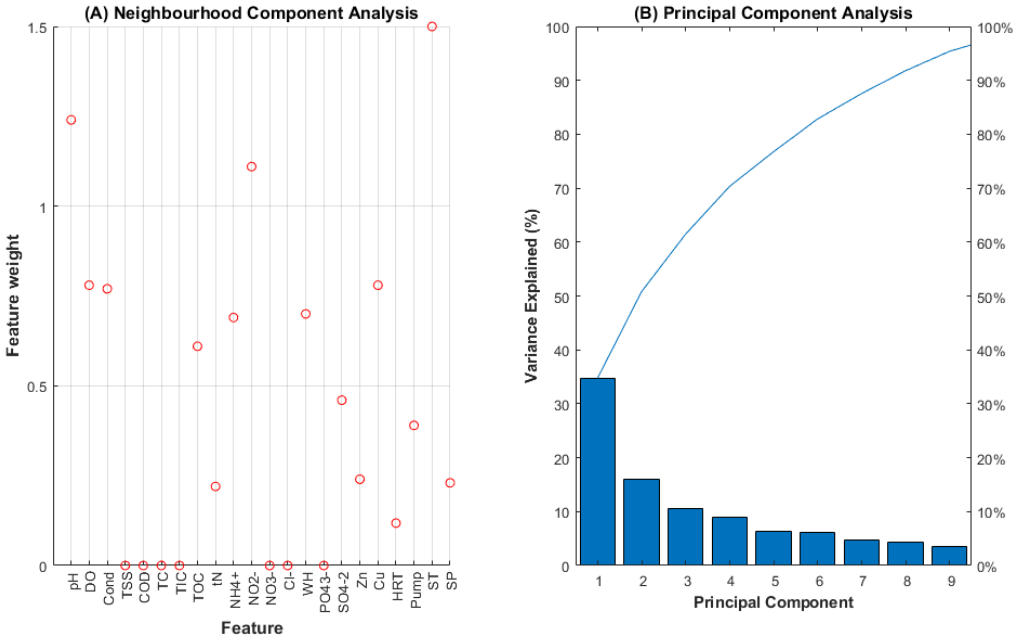


Fig. 16 Results from feature selection and dimensionality reduction techniques, (A) neighbourhood component analysis and (B) principal component analysis.

3.3 System boundaries

The central tendency can be used to describe the centre point of each variable within the system. Statistically, this is achieved with either the mean or median and has been calculated for each variable in Table 5. The variables TOC, tN and system pressure show greater than 10% difference between the mean and median. This can indicate the mean is being influenced by extreme scores. When visualising the data for TOC and tN in Figs. 7 and 8(B) there are noticeable clusters towards the lower values for both. On inspecting the histograms in Fig. 7, over half of the TOC values are clustered around the 0.2 value. This is similar but less extreme for tN, which shows the cluster is more spread out around the bottom values. The feedstock variables Nh_4^+, NO_2^-, WH and SO_4^- all have a cluster of measurements lowering the central tendencies. The pH, DO, conductivity, Zn and Cu, have similar values for mean and medium and the histograms and box plots, show a central tendency in the middle of the range values. Defining the mean for H^2AD process variables that are tightly controlled is less informative because these values operate at set values. For example, the system pressure has three values it operates at (Fig. 7), so a mean between these values does represent the value we expect for new data-points. Instead, the median is more informative in representing the centre of the system. Defining the central tendencies of the system informs what values we expect to observe in new data points collected. If the central tendencies of the system shift as new data is collected, there may have been a change to the system and the model needs retraining.

The relative standard deviation informs on the spread of the data around the mean. The pH and conductivity have smaller relative standard deviations than the other feedstock variables (2.24 and 5.33 compared to 18.1+), implying they are more closely

clustered around the mean and contain less variability. Whereas, the remaining feedstock variables have higher relative standard deviations, indicating that the data is more spread out. This shows these variables are subject to greater levels of variability. The scatter plot matrix visualised the distribution and spread of the data. The feedstock variables have data that is distributed across the system. Although there are variables that contain clusters (e.g. tN), there appear to be no poorly defined regions. Contrasting this the H^2AD process variables, pump speed and system pressure are only defined in 3-4 exact regions respectively. As these are control variables, this will only become an issue if the operator set the H^2AD process conditions outside of those the model was trained on. If the H^2AD system does move to operate at different conditions, new data will need to be collected to retrain the model.

The correlations within the dataset are shown in Fig. 8(A). Instantly recognisable in the heatmap, is that the TOC, tN, NH_4^+, NO_2^-, WH and SO_4^{-2} all show strong ($r > 0.8$) positive correlations to one another and Cu is moderately ($0.4 < r < 0.6$) correlated with this set as well. The strong correlation between tN, NH_4^+ and NO_2^- is expected, as they are all measuring nitrogen's different states within the feedstock. The pH has strong negative correlations with Cu and percentage of COD removal. Knowing the strong correlations within the dataset will help to identify new data-points that behave unexpectedly. This is another indicator that the system has evolved beyond the data that the model was developed on.

Table 9 Statistical analysis of the variables in the model development dataset

	Feature	Mean	Medium	Maximum	Minimum	Range	Relative Standard Deviation	Skewness
Feedstock characteristics	pH	7.54	7.52	7.90	7.28	0.623	2.24	0.391
	DO	0.182	0.190	0.280	0.110	0.170	22.9	-0.0864
	Cond.	11.7	11.6	12.8	10.3	2.54	5.33	-0.0784
	TOC	8020	6650	12200	5060	7140	27.2	0.750
	tN	1830	1560	3240	770	2470	33.0	0.612
	NH4+	944	880	1490	604	888	24.4	0.794
	NO2-	12.0	11.1	16.9	8.40	8.51	24.0	0.411
	WH	209	191	281	160	120	18.1	0.461
	SO32-	5460	5200	8170	3480	4700	25.8	0.313
	Zn	24.4	23.9	35.8	10.3	25.5	27.7	-0.310
	Cu	62.0	64.1	101	22	79	29.4	-0.620

H^2AD process conditions	HRT	3.71	3.48	6.09	2.25	3.84	25.8	1.20
	Pump speed	9.72	9.94	9.94	8.84	1.10	3.94	-1.76
	Proces Temp.	32.9	33.5	39.2	29.8	9.40	5.93	0.612
	System pressure	0.00333	0.00300	0.00500	0.00300	0.00200	16.0	1.32
H^2AD output	% COD removal	63.6	64.3	74.9	45.0	29.9	10.6	-0.731

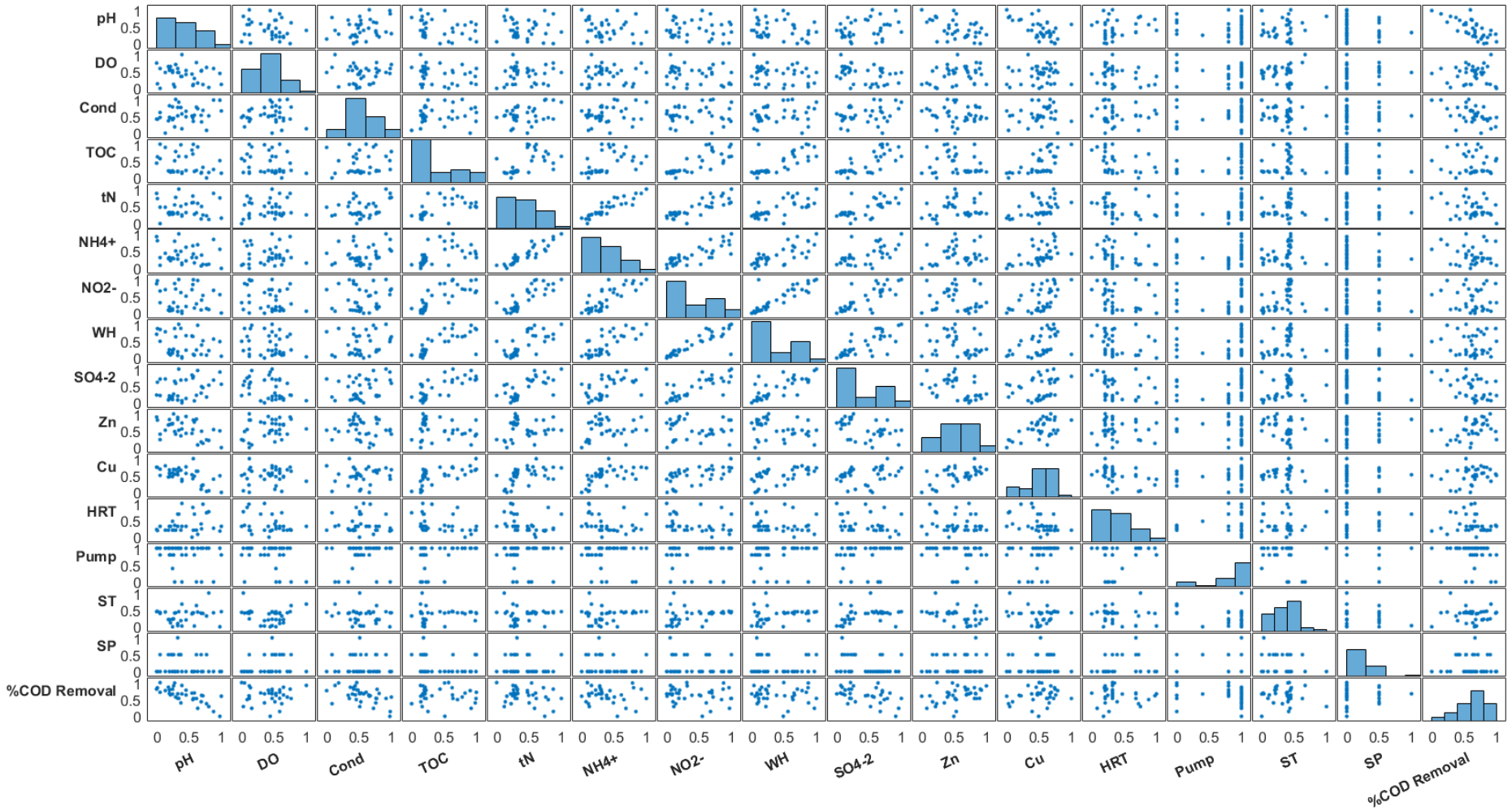


Fig. 17. Scatter plot matrix of variables in the model development dataset

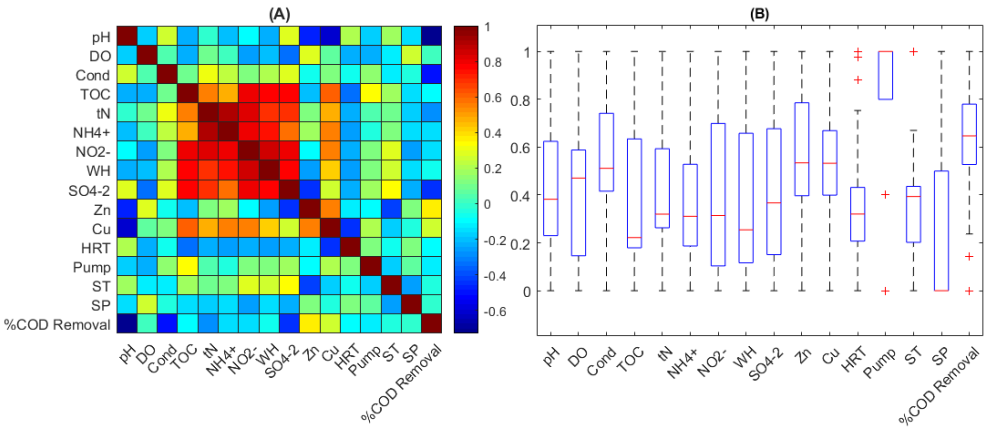


Fig. 18. (A) Heatmap displaying Pearson's correlation between the model development data and (B) boxplot displaying the distribution of the model development data

3.3.1 Determining if new data points are within the system

A useful tool for visualising all the systems variables in one figure is the parallel coordinate plot (PCP). The PCP is able to visualise high dimensional data by, setting each variable along as a point on the x-axis [21]. Data-points are normalised, so all values can be displayed on a single y-axis, and plotted as a series of lines, displaying a data-point normalised value for each variable [21]. By adapting the PCP, a traffic light classification was developed that can quickly inform manufacturers if variables of a new sample are within the boundaries of the model development data (green region), approaching the system boundaries (yellow region) or outside the system (red region), see Fig. 9. The red boundaries are defined as the maximum and minimum of the model development data and the yellow boundaries the 0.1 and 0.9 quantiles. In Fig. 9 two examples are given of a data-point that remained in the system, and a data-point that was well outside of the system for certain variables. The advantage of this traffic light classification is that it provides a simple visualisation that can be displayed on a handheld device and provide the operators with real-time information that the system is currently operating within the boundaries that the model was developed on. Therefore, informing them which model predictions can be trusted.

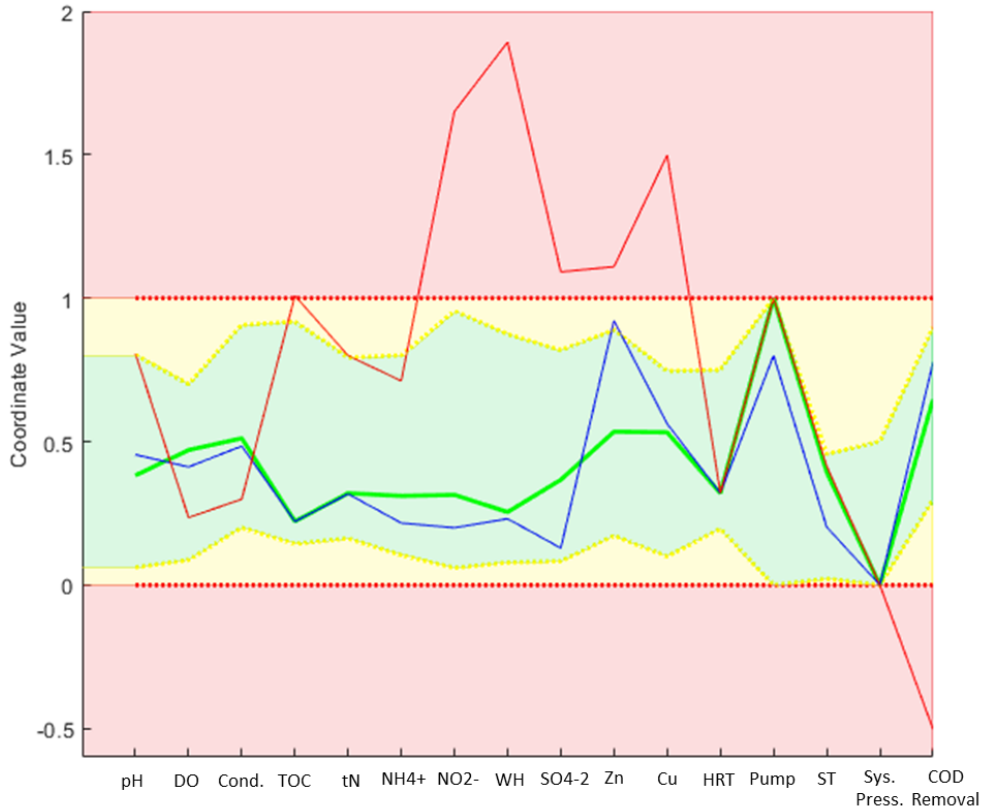


Fig. 19. Traffic light system informing manufacturers if new data points are within the manufacturing system (green region), approaching the boundaries of the system (yellow region) or outside the system (red system). Where, the green line is the median of each variable, the blue is a data point within the system and the red line is a data point outside of the system.

While the red boundaries will always be the maximum and minimum, the yellow boundaries may vary depending on the requirements of the model predictions and how accurate the model is at making predictions outside of the system boundaries. For example, when the RF model was tested on unseen data outside of the system (Section 3.3), it performed relatively well (MAE 4.96) and when predicting the percentage of COD removal, it meets the manufacturer's requirement. However, if this model was built for a predicting an output where a greater level of accuracy is required (e.g. predicting a safety critical variable), the yellow boundaries may have to be increased to give an earlier warning to the operators that the system is approaching the point outside the boundaries of the data the model was trained on.

3.4 Data-driven models results

With respect to predicting the H^2AD bioprocess to remove COD from wastewaters, several models were built to fit the manufacturing data, including the GP model, RF model, SVR model and ANN model. The model development data was used to determine the optimal hyperparameters and model parameters of the DDMs. The 45 model development data points were randomised and 85% (38 data points) were partitioned for the training and validation of the model and the remaining 15% of data (7 data points) were partitioned for testing the final model. A Bayesian optimisation

was performed to optimise the hyperparameters, using k-fold cross-validation to evaluate each configuration's ability to fit the validation data. For the k-fold cross-validation, a k value of 10 was chosen because of the limited number of data points [23]. The models' ability to fit new data was then evaluated using the testing data points. The optimised hyperparameter values for each model are displayed in Table 6.

Table 10. The hyperparameters values for the final machine learning models

Gaussian Process		**Random Forest**		**Support Vector Regression**		**Artificial neural network**	
H-parameter	Value	H-parameter	Value	H-parameter	Value	H-parameter	Value
Sigma	0.0203	Trees	301	Epsilon	0.88	Hidden layers	1
Basic function	Linear	Min no. of predictors	3	Box constraint	937.84	Neurons	6
Kernel function	Squared exp.	Min leaf size	1	Kernel function	Gaussian	Training function	Levenberg-Marquardt
Kernel scale	0.0562			Kernel scale	5.87	Learning rate	0.001

It has been common practice within process manufacturing to use regression plots, of actual values against a model's predictions when determining how well the model fits the data [15,19]. In Fig. 10 the regression plots for the developed DDMs indicates that all the models were able to achieve good agreement ($R^2 > 0.9$) between the actual and predicted values. From visual inspection, the GP, SVR and ANN models were slightly superior to the RF model and this is supported by the statistical analysis in Table 7. The GP model outperformed the other models in the MAE and MAPE error metrics. Both these scores present the magnitude of error around the predicted value and show that the average error of the GP model was the lowest; however, the GP model RMSE was higher The RMSE squares the error magnitude, giving a greater value to outliers. A higher RMSE score indicates that the GP model's predictions included some unexpected errors between the actual percentage of COD removal and the predicted. The full results, showing the model's predictions and residuals are provided in the supplementary material.

Regression plots are useful for a quick guide in determining how accurately a model's predictions match the actual data but are limited when analysing how a DDM performs throughout the system. This is because they are less informative in explaining how the model performs in different region of the system. Instead, residuals plots can be used to better understand where a model overestimates or underestimates the actual value and determine if there are fundamental flaws within the model. The residuals are calculated as the actual value minus the predicted percentage removal of COD, and the plot for each model are presented in Fig. 11. The RF model appears to be overestimating at values lower than approximately 60% COD removal and underestimating above this. Furthermore, it is possible to observe a linear relationship in the residual error. Similar, the residual plot for the SVR model shows that the model appears to consistently overestimate below 60% COD removal by the same amount.

Above 60% COD removal, the model appears to both overestimate and underestimate by consistently the same amount (with exceptions). Future testing on unseen data is required to see if these patterns are restricted only to the model development data. If not, it may be possible to define this relationship and using it to calibrate the models' predictions.

All four of the developed DDM are on a similar ($0.9 < R^2$ 0.99) performance level to other DDMs built to study a BES [42,43]. Lesnik and Liu built an ANN model predicting the COD removal by a microbial fuel cell and when tested on a validation data achieved MAPE 0.57% and 1.06%, similar to the ANN model built for this work which achieved MAPE 1.15% [42]. However, the authors did not evaluate the model on testing data or unseen data, reducing the confidence of the MAPE score. Garg et al. built and evaluated three DDMs of a BES; utilising GP, SVR and ANN machine learning algorithms [43]. After evaluating the DDMs on testing data, Garg et al. also observed the GP model to have the best overall performance [43]. What neither of these studies has done nor, to the best knowledge of the authors, other BES DDMs, is evaluate the DDMs' prediction capabilities on unseen data. Without evaluating on unseen data, it is harder to assess the prediction capabilities of a DDM on new data. This is vital if the model's predictions are to be trusted by a manufacturer.

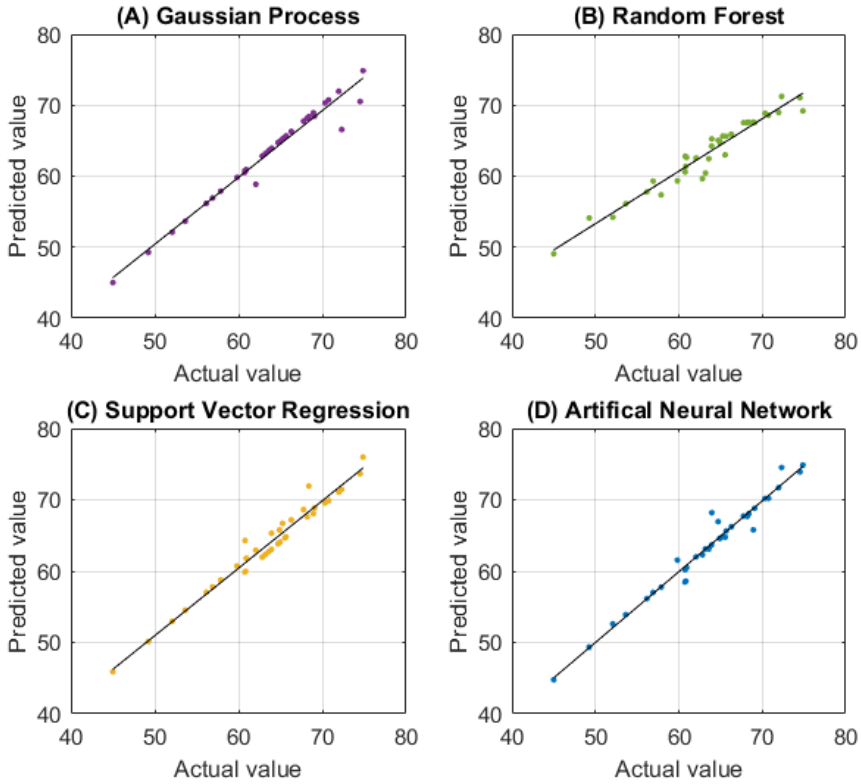


Fig. 20. The predicted percentage of COD removal plotted against the actual value, for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network

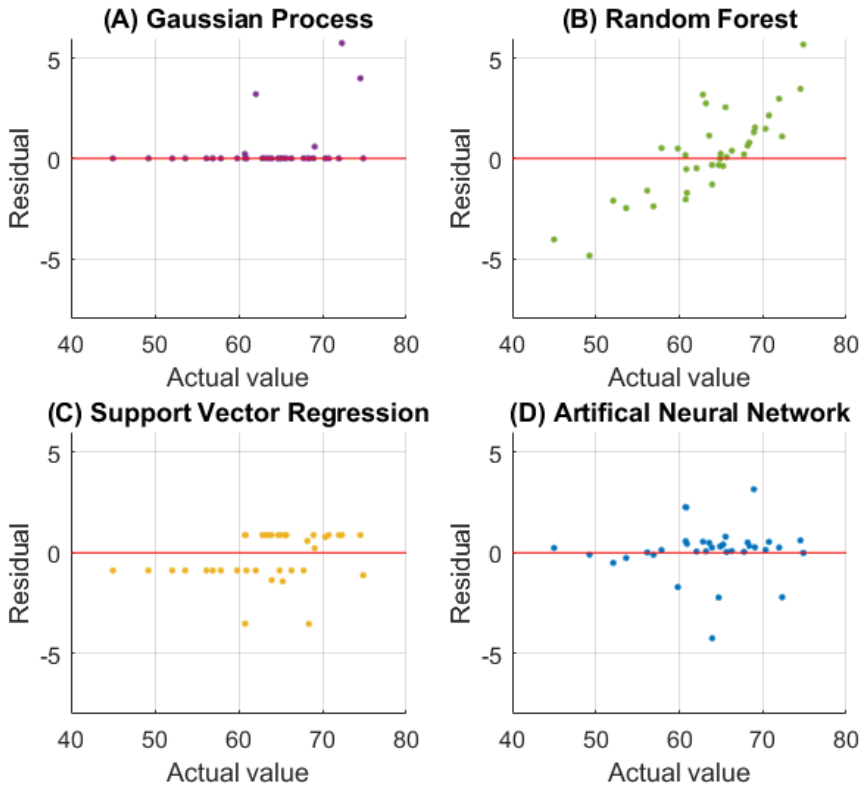


Fig. 21. Residuals plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network

Table 11: Statistical comparisons of the predictive capabilities of the Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural network (ANN) models

Data index	R^2				RMSE				MAE				MAPE			
	GP	**RF**	**SVR**	**ANN**	**GP**	**RF**	**SVR**	**ANN**	**GP**	**RF**	**SVR**	**ANN**	**GP**	**RF**	**SVR**	**ANN**
Training	0.966	0.945	0.969	0.967	1.202	1.226	1.166	1.262	0.383	1.595	1.031	0.737	0.549	2.604	1.638	1.150
Unseen	0.650	0.989	0.540	0.429	5.328	1.439	7.421	8.715	8.879	2.811	8.450	9.245	16.868	6.039	15.736	17.286
Inside	0.850	0.986	0.995	0.954	3.629	1.075	0.783	2.981	3.999	1.092	3.171	3.478	6.628	1.949	4.900	5.632
Outside	0.834	0.995	0.527	0.416	4.276	1.745	9.259	9.288	14.980	4.960	15.050	16.454	29.668	11.152	29.281	31.853

3.5 Evaluating the models' predictive capability on unseen data

Applying George Box's aphorism "all models are wrong, but some are useful" to manufacturing, helps to illustrate that while it is not possible to perfectly model the manufacturing system, a model may aid in decision-making, develop the scientific understanding about the system, communicate knowledge and/or make predictions [44]. However, the challenge is deciding which models are useful and which are not, to avoid giving the manufacturing false or misleading information. Part of the solution is to evaluate how well the model performs on new unseen data. Nine unseen data points were partitioned from the data before model development to evaluate the DDMs' predictive capabilities, see section 2.3.5. These unseen data points are further partitioned into points inside and outside of the system to give a comprehensive assessment of the models' predictive capabilities across and beyond the system.

The model abilities to predict the percentage of COD removal in wastewater were statistically evaluated using, R^2, RMSE, MAE and MAPE (Table 7). Although all models showed good quality predictions on the model development data (R2 > 0.95), the GP, SVR and ANN models' performance all decreased when evaluated on unseen data; whereas, the RF model's performance increased (0.99). The increase between model development data and unseen data of RF model's R^2 value is not reflected by the other error metrics. However, the RMSE, MAE and MAPE still show that the RF model is clearly superior to the other models when predicting on unseen data. In Fig. 12, the residual plot for the model's predictions on unseen data are displayed. The RF and SVR models are able to match the performance between model development data and unseen data within the system. However the SVR model performance significantly decreased outside of the system; whereas, for the RF model the decrease was not as extreme. This reflects the strong generalising capabilities RF algorithms are known for [45]. The RF model predictions on percentage COD removals were more accurate for high values of COD removal outside the system than they were for low values, being able to accurately predict one of the high COD values outside of the system.

Attempting to quantify a DDM's performance on unseen data is an active topic of interest in the field of data-driven modelling [46,47]. However, there are limited examples of using unseen data to evaluate process manufacturing models. One example is work by Hazren et al. to model the capability of carboxylated cellulose nanowhiskers for the remediation of copper from water [19]. Hazren et al. performed 14 additional experiments to generate unseen data both inside and outside the system boundaries [19]. When evaluating the ANN model on unseen, the authors recorded a decrease in performance (R^2: 0.99 → 0.94; RMSE: 1.66 → 6.29; MAPE: 1.15 → 7.98) [19]. The RF model developed for this work has a comparable performance (R^2: 0.99; RMSE: 1.44; MAPE: 6.04) on unseen data as Hazren et al.'s ANN model.

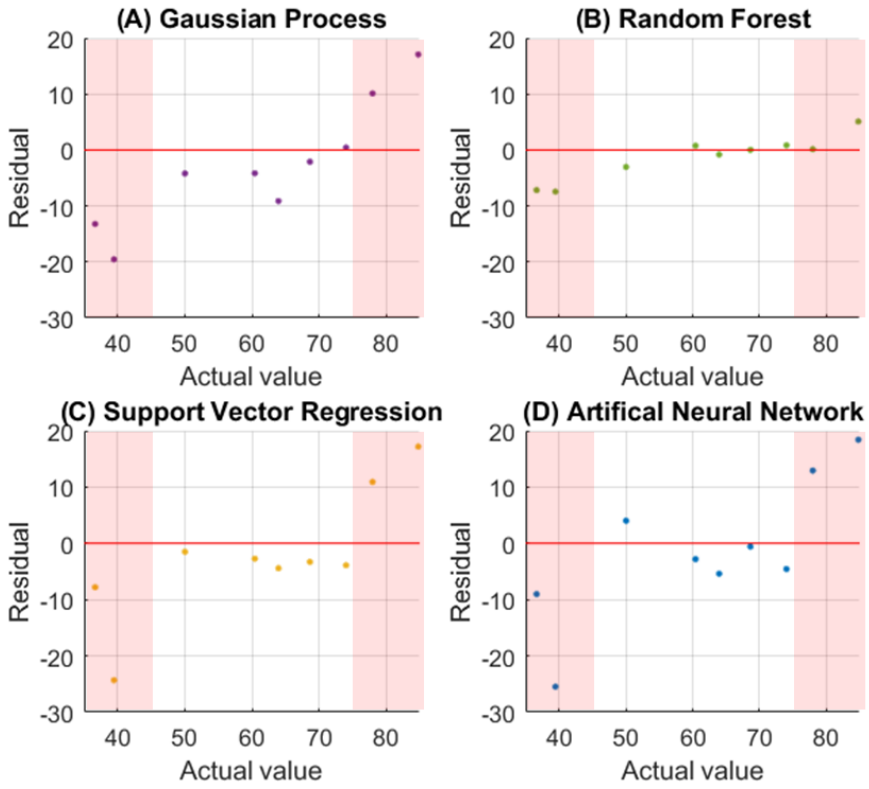


Fig. 22. Residuals of unseen predictions plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network. The red region is outside of the system.

The RF model was deemed to be the best overall model for predicting the percentage of COD removal by the H^2AD bioprocess, because of its consistent performance on model development data and unseen data. The RF model is an example of an ensemble model and this may be the reason the RF model performed better for predictions on unseen data. Ensemble models are built from integrating the predictions made by various models into one output [48]. Ensemble modelling has shown great potential for modelling bioprocesses as a way to account for uncertainty in the model structures and parameters [48]. Ensemble models, like stacked ANN, split the modelling space, by a means similar to cross-validation, so each model is a percentage of the overall space [49]. The 'random' in RF sets it apart from previous work, like stacked ANN, because it not only integrating models that have split the modelling space but has also split the number of input variables available for each model randomly [45]. When modelling systems that face data limitations, ensemble models can overcome the risk of overfitting and reduce the effect of the *curse of dimensionality*.

4. Conclusion and future recommendations

The development of DDM for manufacturing systems plays a significant role in understanding relationships within the system and making predictions. However, in order for a model to be trusted, the model's predictions capabilities must be thoroughly evaluated. This is of greater importance to manufacturing systems where

data is limited because models built from limited data are in danger of overfitting to the model development data. To avoid overfitting, the developed DDM's prediction capabilities should be evaluated on unseen data, taken both inside and outside of the system. Defining the boundaries of a manufacturing system and understanding the space inside the system is made more challenging because manufacturing systems are often highly dimensional. The purpose of this work was to propose an updated CRISP-DM framework that includes defining the manufacturing system boundaries and evaluating the model on unseen data within and outside these boundaries. These changes are aimed at increasing a manufacturers confidence that the model accurately represents the system being modelled. These changes will have two key benefits:

(1) **Determining the DDM's interpolation and extrapolation capabilities**: knowing the system boundaries within the data allows unseen data to be sampled from inside and outside the system. This can then be used to assess the model's interpolation and extrapolation capabilities.
(2) **Knowing when to retrain the model**: Manufacturing systems are continuously evolving and a DDM built from one set of data will likely deteriorate with time, as the system moves beyond the original defined system. By defining the system boundaries, manufacturers can utilise visualisation techniques, like PCP, to monitor when the system moves beyond the system boundaries. The model will then need retraining, incorporating the new data into the model development data. A traffic light protocol for easily classifying if new data collected is inside or outside the system was developed for this work.

In order to demonstrate and evaluate the proposed framework, a DDM was developed of an industrial bioprocess. This study compares the performance of four machine learning algorithms, GP, RF, SVR and ANN. The algorithms' capability to predict the percentage of COD removal from wastewater by the H^2AD bioprocess was evaluated. It was found that all the models performed well at fitting the model development data, with the GP model performing the best. However, based on their ability to fit both the model development data and the unseen data, the RF model showed the best overall performance (R^2: 0.99; RMSE: 1.44; MAE: 2.81; MAPE: 6.04).

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing and [grant number EP/P001246/1], Network Plus: Industrial Systems in a Digital Age.

Acknowledgements

The author (RLG) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing

– Towards Resilient and Scalable High Value Manufacturing. The author (OJF) would like to acknowledge the University of Nottingham, Faculty of Engineering for his PhD scholarship. The author (OJF) would also like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) Network Plus: Industrial Systems in a Digital Age [grant number EP/P001246/1], Connected Everything for funding his placement at Lindhurst Engineering Ltd.

References

[1] M. Cross, A.O. Moscardini, Learning the Art of Mathematical Modelling, John Wiley & Sons, Inc., Chichester [Chichestershire] : New York, NY, USA, 1985.

[2] D.M. Himmelblau, Accounts of Experiences in the Application of Artificial Neural Networks in Chemical Engineering, Ind. Eng. Chem. Res. 47 (2008) 5782–5796. doi:10.1021/ie800076s.

[3] H. Schichl, Models and the History of Modeling, in: K. Josef (Ed.), Model. Lang. Math. Optim., 1st ed., Springer US, Boston, 2004: pp. 25–36. doi:10.1007/978-1-4613-0215-5.

[4] D. Solomatine, L.M. See, R.J. Abrahart, Data-Driven Modelling : Concepts , Approaches and Experiences, in: R.J. Abrahart, L.M. See, D.P. Solomatine (Eds.), Pract. Hydroinformatics. Water Sci. Technol. Libr., Springer, Berlin, 2008: pp. 17–31. doi:https://doi.org/10.1007/978-3-540-79881-1.

[5] M. von Stosch, R. Oliveira, J. Peres, S. Feyo de Azevedo, Hybrid semi-parametric modeling in process systems engineering: Past, present and future, Comput. Chem. Eng. (2014). doi:10.1016/j.compchemeng.2013.08.008.

[6] C.W. Coley, W.H. Green, K.F. Jensen, Machine Learning in Computer-Aided Synthesis Planning, Acc. Chem. Res. 51 (2018) 1281–1289. doi:10.1021/acs.accounts.8b00087.

[7] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science (80-.). 349 (2015) 255 LP – 260. doi:10.1126/science.aaa8415.

[8] S. Yin, O. Kaynak, Big Data for Modern Industry: Challenges and Trends [Point of View], Proc. IEEE. 103 (2015) 143–146. doi:10.1109/JPROC.2015.2388958.

[9] N. Sadati, R.B. Chinnam, M.Z. Nezhad, Observational data-driven modeling and optimization of manufacturing processes, Expert Syst. Appl. 93 (2018) 456–464. doi:10.1016/j.eswa.2017.10.028.

[10] J. Escrig, E. Woolley, S. Rangappa, A. Simeone, N.J. Watson, J. Escrig, Escrig, E. Woolley, S. Rangappa, A. Simeone, N.J. Watson, Clean-in-place monitoring of different food fouling materials using ultrasonic measurements, Food Control. 104 (2019) 358–366. doi:10.1016/J.FOODCONT.2019.05.013.

[11] A. Hemmati-Sarapardeh, M.-H. Ghazanfari, S. Ayatollahi, M. Masihi, Accurate determination of the CO2-crude oil minimum miscibility pressure of pure and impure CO2 streams: A robust modelling approach, Can. J. Chem. Eng. 94

(2016) 253–261. doi:10.1002/cjce.22387.

[12] A.M. Schweidtmann, A.D. Clayton, N. Holmes, E. Bradford, R.A. Bourne, A.A. Lapkin, Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives, Chem. Eng. J. 352 (2018) 277–282. doi:10.1016/J.CEJ.2018.07.031.

[13] J.L. Pilkington, C. Preston, R.L. Gomes, Comparison of response surface methodology (RSM) and artificial neural networks (ANN) towards efficient extraction of artemisinin from Artemisia annua, Ind. Crops Prod. 58 (2014) 15–24. doi:10.1016/J.INDCROP.2014.03.016.

[14] D. Slišković, R. Grbić, Ž. Hocenski, Methods for Plant Data-Based Process Modeling in Soft-Sensor Development, Automatika. 52 (2011) 306–318. doi:10.1080/00051144.2011.11828430.

[15] M.H. Abbas, R. Norman, A. Charles, Neural network modelling of high pressure CO 2 corrosion in pipeline steels, Process Saf. Environ. Prot. 119 (2018) 36–45. doi:10.1016/j.psep.2018.07.006.

[16] D. Güçlü, Ş. Dursun, Artificial neural network modelling of a large-scale wastewater treatment plant operation, Bioprocess Biosyst. Eng. 33 (2010) 1051–1058. doi:10.1007/s00449-010-0430-x.

[17] C. Shearer, The CRISP-DM model: the new blueprint for data mining, J. Data Warehous. 5 (2000) 13–22. www.dw-institute.com/journal.htm (accessed December 13, 2019).

[18] U. Shafique, H. Qaiser, A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA), Int. J. Innov. Sci. Res. 12 (2014) 217–222. http://www.ijisr.issr-journals.org/ (accessed December 13, 2019).

[19] H.A. Hamid, Y. Jenidi, C. Somerfield, R.L. Gomes, Predicting the capability of carboxylated cellulose nanowhiskers for the remediation of copper from water using response surface methodology (RSM) and artificial neural network (ANN) models, Ind. Crops Prod. 93 (2016) 108–120. doi:10.1016/J.INDCROP.2016.05.035.

[20] S.M. Al-Fattah, H. A Al-Naim, H.A. Al-Naim, Artificial-Intelligence Technology Predicts Relative Permeability of Giant Carbonate Reservoirs, 2009. doi:10.2118/109018-PA.

[21] K. Wang, A. Salhi, E.S. Fraga, Process design optimisation using embedded hybrid visualisation and data analysis techniques within a genetic algorithm optimisation framework, Chem. Eng. Process. Process Intensif. 43 (2004) 657–669. doi:10.1016/J.CEP.2003.01.001.

[22] S. Curteanu, C.G. Piuleac, K. Godini, G. Azaryan, Modeling of electrolysis process in wastewater treatment using different types of neural networks, Chem. Eng. J. (2011). doi:10.1016/j.cej.2011.05.104.

[23] F. Charte, I. Romero, M.D. Pï¿½rez-Godoy, A.J. Rivera, E. Castro, Comparative analysis of data mining and response surface methodology predictive models for enzymatic hydrolysis of pretreated olive tree biomass, Comput. Chem. Eng. (2017). doi:10.1016/j.compchemeng.2017.02.008.

[24] M.N. Kashani, S. Shahhosseini, A methodology for modeling batch reactors using generalized dynamic neural networks, Chem. Eng. J. (2010). doi:10.1016/j.cej.2010.02.053.

[25] J. Panerati, M.A. Schnellmann, C. Patience, G. Beltrame, G.S. Patience, Experimental methods in chemical engineering: Artificial neural networks–ANNs, Can. J. Chem. Eng. 97 (2019) 2372–2382. doi:10.1002/cjce.23507.

[26] S. Kim, MATLAB Deep Learning With Machine Learning, Neural Networks and Artificial Intelligence, 1st ed., Apress, 2017. doi:10.1007/978-1-4842-2845-6.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.

[28] OECD, OECD Environmental Outlook to 2050, OECD Publishing, 2012. doi:10.1787/9789264122246-en.

[29] S. Bajracharya, M. Sharma, G. Mohanakrishna, X. Dominguez, D.P.B.T.B.B.T.B. Strik, P.M. Sarma, X. Dominguez Benneton, D.P.B.T.B.B.T.B. Strik, P.M. Sarma, D. Pant, An overview on emerging bioelectrochemical systems (BESs): Technology for sustainable electricity, waste remediation, resource recovery, chemical production and beyond, Renew. Energy. 98 (2016) 153–170. doi:10.1016/j.renene.2016.03.002.

[30] S. Li, G. Chen, S. Li, G. Chen, Factors Affecting the Effectiveness of Bioelectrochemical System Applications: Data Synthesis and Meta-Analysis, Batteries. 4 (2018) 34. doi:10.3390/batteries4030034.

[31] G. Albuquerque, T. Lowe, M. Magnor, Synthetic generation of high-dimensional datasets, IEEE Trans. Vis. Comput. Graph. 17 (2011) 2317–2324. doi:10.1109/TVCG.2011.237.

[32] I. Hobæk Haff, K. Aas, A. Frigessi, On the simplified pair-copula construction - Simply useful or too simplistic?, J. Multivar. Anal. 101 (2010) 1296–1310. doi:10.1016/j.jmva.2009.12.001.

[33] F.P. Brissette, M. Khalili, R. Leconte, Efficient stochastic generation of multi-site synthetic precipitation data, J. Hydrol. 345 (2007) 121–133. doi:10.1016/J.JHYDROL.2007.06.035.

[34] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, F. Herrera, A survey on data preprocessing for data stream mining: Current status and future directions, Neurocomputing. 239 (2017) 39–57. doi:10.1016/J.NEUCOM.2017.01.078.

[35] H. Liu, M. Cocea, Semi-random partitioning of data into training and test sets in granular computing context, Granul. Comput. 2 (2017) 357–386. doi:10.1007/s41066-017-0049-2.

[36] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

[37] T. Hastie, R. Tibshirani, J. Friedman, Elements Of Statistic Learning, Second, Springer New York, New York, 2009. doi:10.1007/978-0-387-84858-7.

[38] D. Gómez, A. Rojas, An Empirical Overview of the No Free Lunch Theorem and Its Effect on Real-World Machine Learning Classification, Neural Comput. 28 (2015) 216–228. doi:10.1162/NECO_a_00793.

[39] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205. doi:10.1016/J.RSER.2017.04.095.

[40] M. Dalmau, N. Atanasova, S. Gabarrón, I. Rodriguez-Roda, J. Comas, Comparison of a deterministic and a data driven model to describe MBR fouling, Chem. Eng. J. (2015). doi:10.1016/j.cej.2014.09.003.

[41] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688. doi:10.1016/j.ijforecast.2006.03.001.

[42] K. Larson Lesnik, H. Liu, Predicting Microbial Fuel Cell Biofilm Communities and Bioreactor Performance using Artificial Neural Networks, (2017). doi:10.1021/acs.est.7b01413.

[43] A. Garg, V. Vijayaraghavan, S.S. Mahapatra, K. Tai, C.H. Wong, Performance evaluation of microbial fuel cell by artificial intelligence methods, Expert Syst. Appl. 41 (2014) 1389–1399. doi:10.1016/J.ESWA.2013.08.038.

[44] G.E.P. Box, Science and Statistics, J. Am. Stat. Assoc. 71 (1976) 791–799. http://mkweb.bcgsc.ca/pointsofsignificance/img/Boxonmaths.pdf (accessed August 22, 2019).

[45] M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy Build. 147 (2017) 77–89. doi:10.1016/J.ENBUILD.2017.04.038.

[46] S. Redyuk, S. Schelter, T. Rukat, V. Markl, F. Biessmann, Learning to validate the predictions of black box machine learning models on unseen data, in: Proc. ACM SIGMOD Int. Conf. Manag. Data, Association for Computing Machinery, 2019. doi:10.1145/3328519.3329126.

[47] M. Zhang, K.H. Leung, Z. Ma, J. Wen, G. Avinash, A Generalized Approach to Determine Confident Samples for Deep Neural Networks on Unseen Data, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer, 2019: pp. 65–74. doi:10.1007/978-3-030-32689-0_7.

[48] Y. Liu, R. Gunawan, Bioprocess optimization under uncertainty using ensemble modeling, J. Biotechnol. 244 (2017) 34–44. doi:10.1016/J.JBIOTEC.2017.01.013.

[49] C.G. Piuleac, M.A. Rodrigo, P. Cañizares, S. Curteanu, C. Sáez, Ten steps modeling of electrolysis processes by using neural networks, Environ. Model. Softw. (2010). doi:10.1016/j.envsoft.2009.07.012.

[50] G. Tchobanoglous, D.H. Stensel, R. Tsuchihashi, F. Burton, M. Abu-Orf, G. Bowden, W. Pfrang, Wastewater engineering: treatment and resource recovery, 5th ed., New York : McGraw-Hill, New York, 2014.

[51] Q. Tao, J. Luo, J. Zhou, S. Zhou, G. Liu, R. Zhang, Effect of dissolved oxygen on

nitrogen and phosphorus removal and electricity production in microbial fuel cell, Bioresour. Technol. 164 (2014) 402–407. doi:10.1016/J.BIORTECH.2014.05.002.

[52] R.A. Rozendal, H.V.M. Hamelers, K. Rabaey, J. Keller, C.J.N. Buisman, Towards practical implementation of bioelectrochemical wastewater treatment, Trends Biotechnol. 26 (2008) 450–459. doi:10.1016/j.tibtech.2008.04.008.

[53] D. Pant, A. Singh, G. Van Bogaert, S. Irving Olsen, P. Singh Nigam, L. Diels, K. Vanbroekhoven, Bioelectrochemical systems (BES) for sustainable energy production and product recovery from organic wastes and industrial wastewaters, RSC Adv. 2 (2012) 1248–1263. doi:10.1039/C1RA00839K.

[54] B. Huang, H. Feng, M. Wang, N. Li, Y. Cong, D. Shen, The effect of C/N ratio on nitrogen removal in a bioelectrochemical system, Bioresour. Technol. 132 (2013) 91–98. doi:10.1016/J.BIORTECH.2012.12.192.

[55] H. Feng, B. Huang, Y. Zou, N. Li, M. Wang, J. Yin, Y. Cong, D. Shen, The effect of carbon sources on nitrogen removal performance in bioelectrochemical systems, Bioresour. Technol. 128 (2013) 565–570. doi:10.1016/j.biortech.2012.11.004.

[56] K. Rabaey, L. Angenent, U. Schroder, J. Keller, Bioelectrochemical systems : from extracellular electron transfer to biotechnological application, IWA Publishing, London, 2010.

[57] P.T. Kelly, Z. He, Bioresource Technology Nutrients removal and recovery in bioelectrochemical systems : A review, Bioresour. Technol. 153 (2014) 351–360. doi:10.1016/j.biortech.2013.12.046.

[58] M.A. Arugula, K.S. Brastad, S.D. Minteer, Z. He, Enzyme catalyzed electricity-driven water softening system, Enzyme Microb. Technol. 51 (2012) 396–401. doi:10.1016/J.ENZMICTEC.2012.08.009.

[59] X. Wang, S. Cheng, X. Zhang, X. Li, B.E. Logan, Impact of salinity on cathode catalyst performance in microbial fuel cells (MFCs), Int. J. Hydrogen Energy. 36 (2011) 13900–13906. doi:10.1016/J.IJHYDENE.2011.03.052.

[60] Y. V Nancharaiah, S.V. Mohan, P.N.L. Lens, Bioresource Technology Metals removal and recovery in bioelectrochemical systems : A review, Bioresour. Technol. 195 (2015) 102–114. doi:10.1016/j.biortech.2015.06.058.

[61] C. Wang, Y. Liu, S. Zheng, A. Jiang, Optimizing combustion of coal fired boilers for reducing NOx emission using Gaussian Process, Energy. 153 (2018) 149–158. doi:10.1016/J.ENERGY.2018.01.003.

[62] David S. Palmer, † Noel M. O'Boyle, and Robert C. Glen, J.B.O. Mitchell*, Random Forest Models To Predict Aqueous Solubility, (2006). doi:10.1021/CI060164K.

[63] B. Yildiz, J.I. Bilbao, A.B. Sproul, A review and analysis of regression and machine learning models on commercial building electricity load forecasting, Renew. Sustain. Energy Rev. 73 (2017) 1104–1122. doi:10.1016/j.rser.2017.02.023.

[64] M. Mesbah, E. Soroush, M. Rezakazemi, Development of a least squares support vector machine model for prediction of natural gas hydrate formation temperature, Chinese J. Chem. Eng. 25 (2017) 1238–1248. doi:10.1016/J.CJCHE.2016.09.007.

[65] H. Peng, X. Ling, Predicting thermal–hydraulic performances in compact heat exchangers by support vector regression, Int. J. Heat Mass Transf. 84 (2015) 203–213. doi:10.1016/J.IJHEATMASSTRANSFER.2015.01.017.

[66] Y. Rabbani, M. Shirvani, S.H. Hashemabadi, M. Keshavarz, Application of artificial neural networks and support vector regression modeling in prediction of magnetorheological fluid rheometery, Colloids Surfaces A Physicochem. Eng. Asp. 520 (2017) 268–278. doi:10.1016/J.COLSURFA.2017.01.081.

[67] D.M. Himmelblau, Applications of artificial neural networks in chemical engineering, Korean J. Chem. Eng. 17 (2000) 373–392. doi:10.1007/BF02706848.

[68] Scopus, document search, (n.d.). https://www.scopus.com/search/form.uri?display=basic (accessed August 22, 2019).

APPENDIX E: PUBLICATION 5

Multiple target modelling to enable sustainable process manufacturing: An industrial bioprocess case study

Oliver J FISHER[1], Nicholas J WATSON[1], Laura PORCU[2/3], Darren BACON[2], Martin RIGLEY[2], Rachel L GOMES[1]

[1] Food, Water, Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom
[2] Lindhurst Engineering Ltd., Midland Road, Sutton in Ashfield, Nottinghamshire, NG17 5GS, United
Kingdom
[3] Energy Innovation & Collaboration, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, United Kingdom

Corresponding author
Author: Rachel L Gomes
Tel.: +44 115 846 8883.
E-mail address: rachel.gomes@nottingham.ac.uk

Abstract

Process manufacturing sectors such as food and drink, pharmaceutical and waste processing are constantly striving to make their processes more sustainable from an environmental and economic perspective. A powerful tool to achieve this are data-driven models (DDMs). Data-driven models are developed using algorithms that learn hidden relationships within manufacturing data to make predictions. Previous attempts at modelling process manufacturing systems, have largely focused on building multiple independent single target models, each predicting one output. These models fail at capturing the relationships between the system's outputs. This paper develops multiple target (MT) models that are superior when describing a manufacturing system with multiple outputs.

For this study, the current leading MT models based on regression chains (RC) are studied. The manufacturing system modelled is a novel bioprocess that treats manufacturers' wastewaters to reduce pollutant load for reuse, while simultaneously generating biogas. This paper presents two novel RC models built using artificial neural networks (ANN) as a base model. The models were then compared against RC models using support vector regression (SVR) as base models and a standard multiple target ANN. Appraisal of the models through the coefficient of determination (R^2), normalised root mean square error (nRMSE) and mean absolute percentage error (MAPE) showed that the RC models built using ANN as a base model had superior predictive capabilities.

Highlights

- Sustainable manufacturing entails maximising economic and environmental objectives
- Within process manufacturing, these objectives are often conflicting
- Multiple target models can predict multiple process manufacturing system's outputs
- Two novel multi-target artificial neural network models are proposed
- Methods to utilise a multiple target model to increase sustainability is then explored

Keywords

Data-driven models, Machine learning, Industry 4.0, Process manufacturing, Multiple outputs, Multiple target regression

1. Introduction

Manufacturers are under pressure from environmental policy and public demands to develop and deliver sustainable manufacturing processes, whilst remaining economically feasible. Advanced manufacturing paradigms (e.g. intelligent/smart manufacturing, cyber-physical systems, cloud manufacturing) show potential to enable sustainable manufacturing by the intelligent allocation of manufacturing services and resources [1,2]. For a manufacturing system to be considered truly sustainable it needs to maximise environmental and social benefits, in addition to economic benefits [1]. This is particularly important for process manufacturing, which produces a variety of outputs with multiple, sometimes contradictory, sustainability implications. Process manufacturing is the thermal and/or bio/chemical conversion of resources to products, co-products, by-products and wastes streams [3]. Process manufacturing covers a range of industries, including chemicals, food and drink, paint, polymer, pharmaceutical, steel, fast-moving consumer goods (FMCG). All the process outputs (products, by-products and waste's composition and yields) need consideration to achieve sustainable manufacturing. This paper investigates how multiple target modelling of a process manufacturing system will provide manufacturers with the tools to evaluate the sustainability of a manufacturing system.

Modelling a manufacturing system can aid decision making to identify the most sustainable manufacturing route that may have been previously been unknowable [3]. There are two approaches when developing models: first-principal modelling and empirical/data-driven modelling. First-principal modelling is the development of a series of equations describing a system's individual parts that, when combined, describe the whole of the manufacturing system [4]. Data-driven modelling combines the fields of computational intelligence and machine learning, to build models by fitting algorithms to manufacturing data [5]. Computational intelligence refers to nature-inspired computation approaches to problem-solving; for example, algorithms that mimic the behaviour of animals (swarm intelligence) [6] or algorithms that replicate the behaviour of how humans solve problems (artificial neural network, ANN) [7]. Machine learning focuses on the development of algorithms that can access data and use it to learn for themselves [8]. There are three categories of machine

learning: supervised learning, unsupervised learning and reinforcement learning. Supervised learning learns from a set of known input and output pairs to map the relationships between the two [9]. Unsupervised learning identifies patterns, relationships and structures within a dataset without reference to known outputs [9]. Reinforcement learning is concerned with the problem of finding suitable actions to take to maximise a reward [9].

Data-driven models (DDMs) are able to model complex nonlinear relationships within multivariate data, without requiring a detailed understanding of the manufacturing system [10]. Deciding which to use is dependent on the level of knowledge known about the manufacturing system and if the manufacturing data is representative of the system. First principal models often require a greater level of understanding of the system than DDMs, which makes them more time consuming to build [11]. However, the accuracy of a DDM's predictions relies on the modelling data being representative of the manufacturing system [8]. When collecting a representative set of data, process manufacturers face challenges regarding the availability and variability of manufacturing data [8]. Compared to traditional first principle models, DDMs are more flexible and can be used in more complicated industrial processes [12]. Data-driven models use in process manufacturing has been on the rise since the 2000s [12], due to the on-going increase in manufacturing data generated by modern processes and increasing computing processing power from cloud computing.

Enhancing the sustainability of process manufacturing systems has gained growing attention over the last few decades [13–15]. Data-driven models may be used to discover knowledge about process manufacturing systems, making the process more understandable for efficient sustainability analyses [16]. The majority of published research utilising DDMs within process manufacturing systems are single target (ST) models. A ST model is trained to predict the value of a single, categorical or numerical, target variable of a manufacturing system [7]. Recent uses of ST models in process manufacturing show their capability to model: adsorption [17], clean-in-place systems [18], fluid mechanics [19], heat transfer [20], mass transport [21], reaction networks [22] and synthesise design [8]. These models may be utilised to discover the optimal sustainable configuration of the unit operation being modelled. For example, Halder et al. utilised a DDM model to determine the optimal process conditions to maximise the removal efficiency of the heavy metal chromium (VI) from industrial wastewaters [23]. Conventional metal removal techniques have some major disadvantages (high operational cost, incomplete removal and generation of toxic residuals) [24]. By utilising data-driven modelling techniques, Halder et al. were able to increase the chromium (VI) removal efficiency by superheated steam activated granular carbon (GAC) prepared indigenously from non-usable coconut shell, a sustainable alternative to conventional metal removal techniques [23]. However, industrial wastewaters often contain a variety of toxic heavy metals (e.g. cadmium, chromium, cobalt, copper, nickel and mercury) [24]. Although Halder et al. was able to use a ST model optimise the removal efficiency for one heavy metal, the adsorption process may be required to be optimised to remove multiple heavy metals. The development of a multiple target (MT) model would inform the manufacturer of the optimal process conditions to maximise the removal efficiency of multiple heavy metals.

Modelling a manufacturing system with the aim to increase sustainability, multiple outputs may be required for the model to identify both the most economical and environmental solution. There are examples of multiple independent ST models

predicting multiple outputs of a process manufacturing system. For example, Lesnik and Liu modelled the power density, chemical oxygen demand (COD) removal and Columbia efficiency of a microbial fuel cell treating wastewater [25]. The authors achieved this by developing three ST models, rather than a single model predicting all three outputs [25]. Yet, building multiple ST models of a system fails to capture the relationship between the multiple outputs and are computational expensive [26]. An alternative approach is to build a single multiple target model that captures the relationships between the output variables and are proven to use less computational power [27]. Multiple target models are capable of predicting multiple outputs from a set of input data [27]. The development of MT models has proven successful when applied to a wide range of practices: bioinformatics [28], chemometrics [29], ecology [30], gene function prediction [31], natural language processing [32], stock price forecasting [33]. There are a limited number of examples of MT models built for process manufacturing systems [34,35]. However, the potential of MT models has yet to be realised by the wider process manufacturing community.

There are two main approaches for using machine learning algorithms as a base method for MT learning: *problem transformation* and *algorithm adaptation* [27]. Problem transformation is the transformation of a MT model into multiple ST models, each solved separately using traditional DDM procedures [27]. Whereas, algorithm adaptation methods adapt machine learning algorithms to predict all outputs simultaneously in one DDM [27]. Algorithm adaption methods have been shown to perform better than problem transformation methods [26,31]. This is accredited to their ability to capture not only the relationships between a model's inputs and outputs but also between the different outputs [26]. A feature not possible in problem transformation models, because a single independent model is trained for each output [27]. This work compares the capability of both MT modelling approaches to make predictions for not only the model development data but also unseen data, something not yet tested in the literature. Unseen data is data that has not been used for the training, validation or testing of the model and is used to thoroughly evaluate the final model's predictive capabilities on new data [17]. This is particularly necessary when a

Once built, a MT model can then be used to evaluate different manufacturing routes to find the most sustainable solution. This sometimes proves challenging, as within process manufacturing the economic and environmental objectives are often competing [36]. There are two approaches when utilising a DDM model to ascertain the most sustainable manufacturing configuration, data exploration and self-optimisation. The first approach involves using the developed DDM to simulate output results from new synthetic input data. Variables outside of the process manufacturer's control are set at the one value for each data point, whilst the process conditions are assigned values to investigate each variables' effect on the system's outputs. For the second approach, the DDM model is used within a closed-loop optimisation system to automatically determine the optimal process conditions that maximise the model outputs [36]. This can be done by either fitting the DDM's predictions to a function that may be used to find a single optimal point of the system or by establishing the Pareto front for each process condition [36]. A Pareto front is a state from which it is impossible to optimise the process conditions further to increase a performance objective without making at least one performance objective worse off [37]. For example, Schweidtmann et al. were able to utilise a MT model for the self-optimisation of two chemical reactions performed in continuous flow, which determined the Pareto fronts of the four reactions conditions [36]. The use of MT models for multiple

objective optimisation shows great potential for finding a solution that maximises both economic and environmental objectives.

The aim of this research is to investigate the best techniques to develop a MT model of a process manufacturing system and demonstrate how that MT model may be utilised to increase sustainability within the system. The main contributions presented in this paper include:

- Evaluating the performance problem transformation and leading algorithm adaptation models when modelling a manufacturing system. Including evaluating their abilities to make predictions on unseen data, both inside and outside of the manufacturing system.
- Proposing a novel MT model by adapting state-of-the-art algorithm adaptation models to use ANN as a base model for the first time. The new model is then evaluated against existing MT models techniques.
- Utilising the final MT model to investigate the most sustainable manufacturing route of a novel industrial bioprocess for different manufacturing environments.

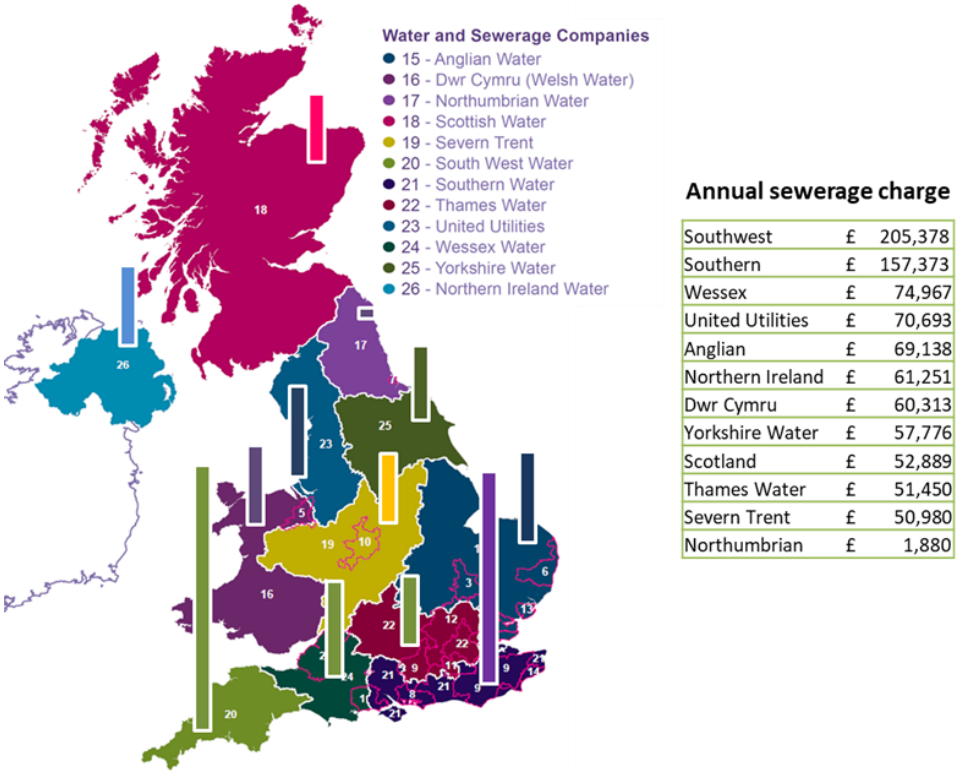
1.1 H^2AD Case Study

Manufacturers are under pressure to increase sustainable throughout their systems and an intelligent resource use strategy is paramount to achieving this. Two key resources that process manufacturers consume are water and energy. Water demand in the manufacturing sector is expected to rise by 400% between 2012 and 2050 [38], while energy consumption is projected to increase by 1.2%/year until 2040 [39]. Process manufacturers are in an advantageous position to reduce demand for water and energy because their systems offer an opportunity to recover both from waste streams via waste valorisation technologies [3]. Waste valorisation refers to industrial processing activities aimed at reusing, recycling, or recovering useful products or sources of energy from waste [40]. In the UK, Lindhurst Engineering Ltd. in partnership with the University of Nottingham has developed a technology called H^2AD Micro AD. The H^2AD is a process manufacturing bioprocess is capable of treating a variety of wastewaters, from sources including agriculture, brewing, soft drinks, foods and bio-manufacture residues, to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy. The H^2AD is a hybrid of an anaerobic digester and a microbial fuel cell, targeted at treating the wastewaters from SME manufacturing processes due to its modular, low-cost design.

The H^2AD system was selected to demonstrate the methodology developed in this paper because the bioprocess has multiple objectives to optimise (water pollutant remediation and biogas generation). These two objectives come into conflict with one another when deciding on which process conditions to use when operating the H^2AD. For example, increasing the hydraulic residence time (HRT) of the bioprocess increases the time available for the system to reduce pollutants in the wastewater. However, this reduces the rate of addition of wastewater to the system, reducing the fuel available for biogas generation. This is a simplistic example, the actual relation

between HRT and the objectives are more complex (due to the fact the bioprocess is a complex non-linear living system), hence the need for a DDM of the system.

The H^2AD system has two objectives: improve the wastewater quality for reuse and energy generation from waste streams. This was represented in the model by 3 outputs: percentage of COD removal (%COD removal), percentage of total suspended solids (TSS) removal (%TSS removal) and daily volume of methane produced by the H^2AD bioprocess (CH_4 generation). The pollutants COD and TSS were chosen because UK wastewater management companies use COD and TSS concentrations in the wastewater, as part of the Mogden formula, when calculating wastewater disposal cost [41]. That rate wastewater companies charge for wastewater disposal differs between companies, see Fig. 1. This will have implications when determining whether it is more sustainable to maximise pollutant removal or energy generation. The energy generation potential of the H^2AD bioprocess is represented by the volume of CH_4 produced. The H^2AD bioprocess produces biogas, made up of CH_4, carbon dioxide, oxygen and hydrogen sulphide components. Of these, only CH_4 has significant calorific value and can be used to estimate the energy potential from the biogas produced [42].



Annual sewerage charge

Southwest	£	205,378
Southern	£	157,373
Wessex	£	74,967
United Utilities	£	70,693
Anglian	£	69,138
Northern Ireland	£	61,251
Dwr Cymru	£	60,313
Yorkshire Water	£	57,776
Scotland	£	52,889
Thames Water	£	51,450
Severn Trent	£	50,980
Northumbrian	£	1,880

Fig. 23. Annual wastewater charge for different UK water and sewerage companies, based on 4m3/d of wastewater with a chemical oxygen demand of 85,000 mg/L and total suspended solids of 984 mg/L using WRAP Mogden calculator [41].

2. Overview of multiple target modelling techniques

There are serval approaches to use when building MT models, including linear target combinations for MT regression [43], multi-objective random forests [44], multiple output ANN [34], boosted-neural network ensemble [45] and multi-task deep learning [46]. Recently, Spyromitros-Xioufis et al. explored the possibility of stacking or chaining regression models together, with Regression Chains (RCs) proving the most effective [47]. In building a RC model a random chain, or sequence, of the output

variables is selected and for each output in the chain, ST models are built sequentially by using the output of the previous model as input for the next. The RC is an example of an algorithm adaptation multiple target modelling technique because a RC uses selected model outputs to predict the additional model outputs (Fig. 2-B). Whereas, a problem transformation creates multiple independent models for each output and fails to capture the relationships between model outputs (Fig. 2-A) [27].

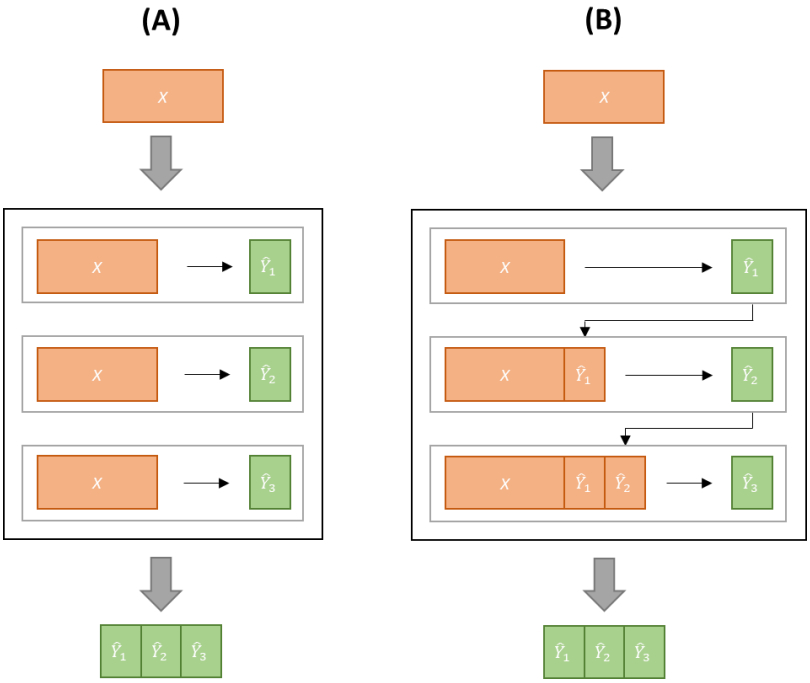


Fig. 24. Two examples of multiple target modelling: (A) problem transformation and (B) algorithm adaptation. Where the model input data (X) is represented by the orange box and the model output prediction ($\hat{Y}_i$) by the green box.

Despite the advantages previously discussed of using an algorithm adaption model over a problem transformation model, there are two main problems with using a RC model:

1. The relationships between the output variables at the start of the chain and the end of the chain are not exploited [26].
2. Prediction error at the start of the chain will propagate through the rest of the chain [47].

To overcome these problems, an Ensemble of Regressor Chains (ERC) was proposed [47]. The ECR works by creating multiple RCs for every possible permutation of the output sequence order. For example, the RC displayed in Fig. 2 (A) is ordered Y1-Y2-Y3, an ERC model would include multiple RC for orders Y1-Y3-Y2, Y2-Y1-Y3, Y2-Y3-Y1, Y3-Y1-Y2 and Y3-Y2-Y1. The final prediction values are obtained by taking the mean of the predicted values. Because of the computational time for building models with outputs that exceed 10 permutations, the sequence order is randomised, and 10 RCs are selected for constructing the ERC. As the number of output variables increases, the number of possible chains increases by a factorial function. Therefore, there is no guarantee that the 10 random chains generated will truly capture the relationships among the outputs. Additionally, building an ERC becomes requires exponentially greater computational power as the number of outputs increases. Melki et al.

proposed an adaption to the ERC, which instead used a single chain based on the maximisation of the correlations among the output variables [26]. Melki et al. referred to their techniques as SVR correlation chain (SVRCC), as it was built using a SVR as a base model. For this work, it is referred to as the maximum correlation chain (MCC) for when applying the technique to other base models. To construct a MCC the first step is to calculate the outputs' correlation coefficient matrix, which will describe the linear relationship among the output variables. This can then be sorted into descending order, creating the MCC. The MCC then follows the same training procedure as a standard RC.

Problem transformation and algorithm adaptation MT models use machine learning algorithms as a base model for each model developed in the stack or chain. Previous examples of ERC and MCC models have used random forest (RF) and support vector regression (SVR) based models [26,47]. Random forests are an ensemble technique that uses multiple decision trees and a statistical technique called 'bagging' [48]. Rather than just averaging predictions from multiple trees, an RF instead randomly samples from the training data for each tree and randomly subsets the input variables at splitting nodes until the minimum node size is reached [48]. A SVR solves a non-linear problem by transforming the non-linearity between features and target using linear mapping (kernels) [49]. They are highly effective machine learning algorithms, even when presented with small quantities of data [49]. Artificial neural networks (ANNs) are a machine learning algorithm that has yet to be utilised as a base model for ECR or MCC. [49]. Artificial neural networks are mathematical models that have been inspired by the connections made between neurons in a biological brain [7]. The neural network is composed of nodes (which represent the neurons) and the connection between these nodes are given weight values to mimic how the brain alters the association between neurons [7]. An ANN consists of an input layer, a single or multiple hidden layers and an output layer. Artificial neural networks have excellent capabilities when modelling non-linear systems and are the most extensively utilised DDM in process manufacturing [50].

The final MT modelling technique evaluated as part of this work is a multiple target artificial neural network (MT-ANN). An ANN model is highly suited to MT modelling as its architecture can be quickly adapted for multiple outputs [47], as shown in Fig. 3. Because of this, MT-ANNs are among some of the earliest examples of MT modelling [51,52]. An example is presented by Curteanu et al. modelling the electrolysis process in wastewater treatment, who found that MT-ANN outperformed multiple individual ANN models [34]. There has yet to be a comparison of the abilities of a MT-ANN to with either the ERC or MCC modelling techniques.

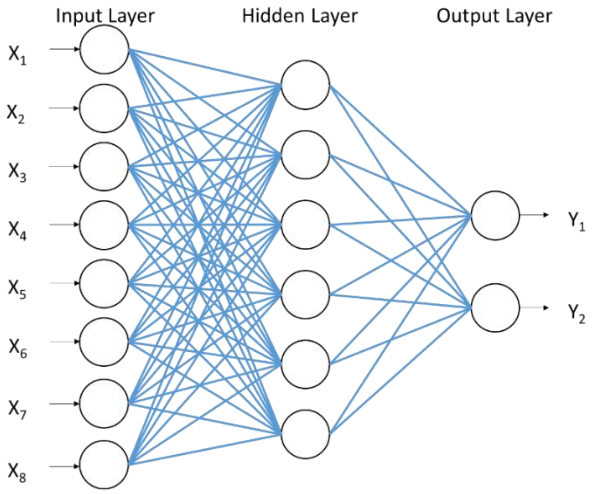


Fig. 25: Architecture of a Multiple Target Artificial Neural Network (MT-ANN), with 8 inputs, 2 outputs and 1 hidden layer containing 6 neurons

3. Material and methods

3.1 Modelling data

The manufacturing system being modelled is the H^2AD bioprocess described in Section 1.1. The model is developed from labelled input and output data. The model input data originates from two sources, the H^2AD feedstock data and the H^2AD process data. To characteristic the H^2AD feedstock, water quality analysis (WQA) was performed on samples collected entering the H^2AD plant. A total of 18 WQA parameters were measured and are detailed in Appendix A. In addition to this, the H^2AD unit automatically collected data on four H^2AD process conditions via sensors connected by internet of things (IoT) technologies to a database hosted on a cloud platform. These were hydraulic retention time (HRT), pump speed, system temperature (ST) and system pressure (SP). The H^2AD outputs were characterised by three variables, %COD removal, %TSS removal and CH_4 generation. The H^2AD bioprocess model is shown in Fig. 4.

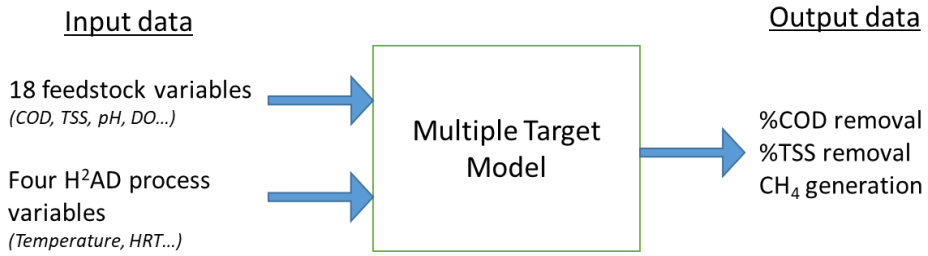


Fig. 26. The H2AD bioprocess model [chemical oxygen demand: COD; total suspended solids: TSS; dissolved oxygen: DO; hydraulic retention time: HRT; percentage of COD removal: %COD removal, percentage of TSS removal: %TSS removal, daily volume of methane generated: CH_4 generation].

From previous work, there are 30 data-points available of WQA performed on the farm waste samples, taken from the slurry tank feeding the H^2AD system. In addition, there is one year's worth of historical H^2AD process data automatically collected by the unit. This informed the decision to generate an additional synthetic 22 WQA data points bringing the total to 52 data points, representing one year's worth of weekly sampling. When generating synthetic data the standard method is to randomly sample from a

probability density function (PDF) fitted to existing data [53]. Pair-copula constructions (PCCs) are a popular tool for modelling multivariate data, due to their simple structure and high flexibility [54].

Due to insufficient data relating input variables to the model's output variables, it was not possible to follow the same PDF method. The approach for generating synthetic output data followed that in Brissette et al., whereby trends identified from historical H^2AD data governed the development of non-linear equations to generate synthetic output data from the input data [55]

3.2 Data preparation, pre-processing and partitioning

For this study, the dataset was normalised as part of the data preparation stage. This was to ensure variables of higher magnitudes are not given greater weight by the DDM. The COD and TSS data have magnitudes within the order of 10,000, while for DO is the magnitude is much lower and in the order of 0.1. To normalise the data the minimax function was applied, as this normalises the variables without any loss of information [56]. The equation to normalise the data is given below:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where x is the variable, x_{norm} is the normalised variable and x_{min} *and* x_{max} is the minimum and maximum values of the variable being normalised respectively.

This work utilised a neighbourhood component analysis (NCA) feature selection method and principal component analysis (PCA) to reduce the number of dimensions in the data. The NCA is a machine learning technique that learns the feature weights of the input data on the output data [57], weights close to zero are classified irrelevant. Principal component analysis is a mathematical procedure that transforms potentially correlated data into an orthogonal system of linearly uncorrelated principal components (PCs) [58]. All the PCs are orthogonal to each other and ordered so that the maximum variance is captured in PC1 and the variance decreases as one moves down the PCs. The standard practice to reduce the number of variables so that 95% of the variance within the dataset is captured by the PCs was followed [58].

The data used to develop a DDM is then partitioned at several levels, as independent data is required to train, validate, test and evaluate the model. The data is first partitioned into model development data and unseen data. The model development data is further partitioned into training, validation and testing data [9]. Unseen data-points are kept apart until the models are completed to evaluate the prediction capabilities of the model on data not used during the model's development. A systematic approach developed by Fisher et al. was used to partition the manufacturing data into 70% model development data and 30% unseen data [ref]. Ten points were taken from within the model development boundaries and six points outside the model boundaries.

The model development data is then portioned into training, validation and testing data. When developing a DDM, each algorithm has model parameters and hyperparameters that require tuning to find the optimal configuration. Hyperparameters are parameters that are set before being trained on a model,

compared to model parameters that are learnt during training [7]. For example, the number of neurons in an ANN's hidden layer is a hyperparameter; whereas, the weights between neurons is a model parameter. Validation data is used to tune the hyperparameters and training data is used to tune the model parameters. Testing data is utilised to evaluate the final model's ability to fit new data. The model development data points were randomised and 85% were partitioned for the training and validation of the model and the remaining 15% of data were partitioned for testing the final model. A Bayesian optimisation was performed to optimise the hyperparameters, using k-fold cross-validation to evaluate each configuration's ability to fit the validation data. For the k-fold cross-validation, a k value of 10 was chosen because of the limited (<100) number of data points [59]. The model ability to fit new data was then evaluated using the testing data points.

3.3 Multiple target models

With respect to modelling the H^2AD bioprocess's capability to remove COD and TSS from wastewaters and generate CH4 gas, 8 MT DDMs were built to fit the manufacturing data. Three of the models were problem transformation models and five were algorithm adaption models. The machine learning algorithms RF, SVR and ANN were used as base models. Five algorithm adaption MT models were built to describe the H^2AD system. Two models were built following ERC techniques using SVR and ANN as base models, these are referred to as ERC-SVR and ERC-ANN respectively. The ECR models were built following the procedure outlined by Spyromitros-Xioufis et al [47]. Two models were built following Melki et al. MCC procedure using SVR and ANN as base models, these are referred to as MCC-SVR and MCC-ANN respectively [26]. The final MT model built was a MT-ANN. All the base models built were tuned by Bayesian optimisation and cross-validation procedures.

3.4 Model evaluation

The prediction capabilities of the various models were statistically evaluated in terms of the coefficient of determination (R^2), the normalised root of mean squared error (nRMSE) and the mean absolute percentage error (MAPE). The R^2 is the square of the sample correlation coefficient between observed values and predicted values, and is a measure of the explained variance of the model [58]. The nRMSE measures the normalised mean square magnitude of the error [60], while MAPE measures the absolute percentage of the errors [61]. The statistically best model is the one that has an R^2 closest to 1 while minimising nRMSE and MAPE. The R^2, nRMSE and MAPE are defined as follows:

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}_i\right)(Y_i - \bar{Y}_i)\right)^2}{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}_i\right)^2 (Y_i - \bar{Y}_i)^2} \tag{2}$$

$$nRMSE = \frac{1}{Y_{max} - Y_{min}}\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}} \tag{3}$$

$$MAPE = 100 \times \frac{\sum_{i-1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|}{n} \quad (4)$$

Where n is the number of data points, $\hat{Y}$ is the predicted value, Y_i is the actual observed value, and the symbol ¯ is the average of the related values.

4. Results and discussion

4.1 Results from data-driven models

4.1.1 Problem transformation multiple target models

The RF, SVR and ANN models' capabilities to predict the %COD removal, %TSS removal and CH_4 generation were statistically evaluated using R^2, nRMSE and MAPE; the results are displayed in Table 1. The best model is one which has the highest R^2 score and lowest nRMSE and MAPE score. The RF model is inferior at predicting for each H^2AD output (Table 1) and is significantly inferior at predicting the %COD removal (R^2 of 0.63, nRMSE of 0.06 and MAPE of 8.56). The SVR and ANN show similar capabilities when predicting the %COD removal, though the ANN model has a slightly worst nRMSE score (nRMSE of 0.05-ANN compared to 0.04-SVR), suggesting that the ANN model is subject to a greater number of outliers within the predictions. This trend between the SVR and ANN model is repeated when predicting CH_4 generation (Table 1). However, when predicting %TSS removal the SVR model performs consistently worse than the ANN model (R^2 of 0.90, nRMSE of 0.08 and MAPE of 4.14, compared to R^2 0.96, nRMSE of 0.05 and MAPE of 3.22).

Table 12. Comparison of the random forest (RF), support vector regression (SVR) and artificial neural network (ANN) models ability to fit the model development data.

Model	Output	R2	nRMSE	MAPE
RF	COD	0.63	0.06	8.56
	TSS	0.90	0.07	5.54
	CH4	0.95	0.07	2.53
SVR	COD	0.97	0.04	2.26
	TSS	0.90	0.08	4.14
	CH4	0.96	0.05	3.00
ANN	COD	0.96	0.05	1.80
	TSS	0.96	0.05	3.22
	CH4	0.96	0.06	2.36

A ranking system was employed to determine the best overall model at predicting the H^2AD bioprocess outputs. The average statistical error metrics R^2, nRMSE and MAPE for each model were determined. Then for each average statistical error metrics, the models were ordered from best to worst and scored (top of the list scored one and the bottom scored three). The model's rank was then determined by summing together each model's scores for each error metrics and ordering the model's from lowest to highest. The results from averaging each model's R^2, nRMSE and MAPE and model ranking are displayed in Fig. 5. The average R2, nRMSe and MAPE scores

indicate that the ANN model is superior at predicting the H^{2}AD bioprocess outputs compared to the RF and SVR model (Fig. 5).

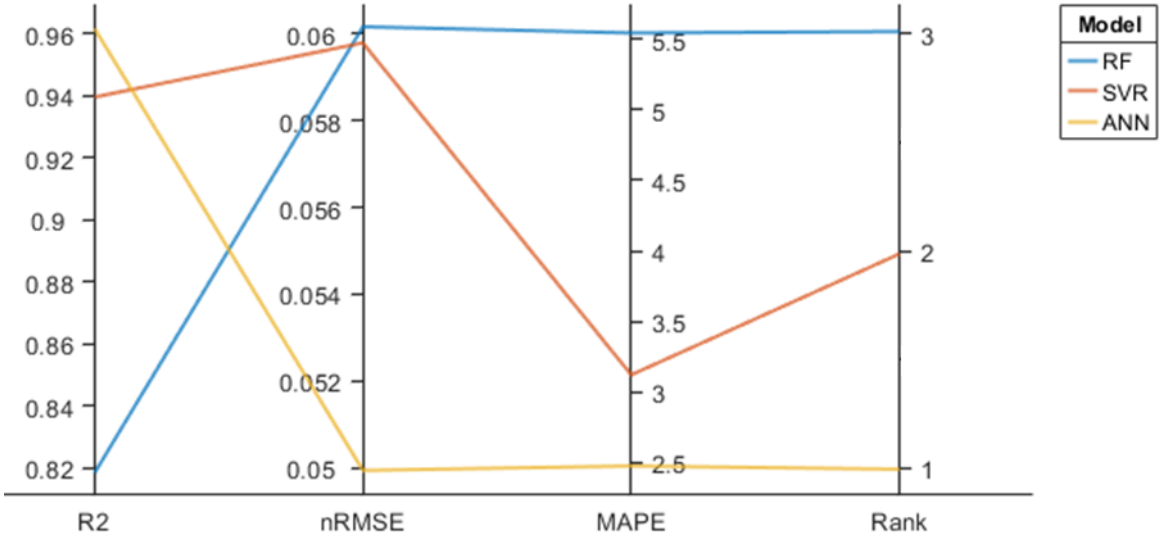


Fig. 27. Evaluation metrics of the problem transformation modes. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH_4 generation output. [Coefficient of determination, R^2; normalised root of mean squared error, nRMSE; mean absolute percentage error, MAPE; random forest, RF; support vector regression, SVR; artificial neural network, ANN].

Problem transformation models have been shown to be less accurate when predicting multiple outputs, as they are unable to capture the relationship between model outputs [27]. However, they can be effective at determining the suitability of machine learning algorithms as a base model for algorithm adaption models [26]. After statistical analysis of the predictive capabilities of the RF, SVR and ANN models, the decision was taken to use only SVR and ANN algorithms as base models for the algorithm adaption MT models.

4.1.2 Algorithm adaption multiple target models

The ERC-SVR, ERC-ANN, MCC-SVR, MCC-ANN and MT-ANN models' capabilities to predict the %COD removal, %TSS removal and CH_4 generation were also statistically evaluated using R^2, nRMSE and MAPE; the results are displayed in Table 2. The ranking system described in Section 3.1.1 was again used to determine the best model at predicting the H^{2}AD outputs. The results, displayed in Fig. 6, indicate that ERC-ANN is the superior model. The MCC-ANN and ERC-ANN have a similar MAPE score, indicating that on average they have similar prediction errors; however, the MCC-ANN has a higher nRMSE error. That is the result of the model containing a greater number of outlier predictions. The ERC-ANN is an example of an ensemble model, which are able to mitigate the influence of outlier predictions. This is because an ensemble model outputs are an average of multiple predictions; an outlier may raise the value of the prediction, but by averaging the prediction an ensemble model is capable of reducing the outliers' impact on model performance [62].

Table 13. Comparison of the ensemble of regressor chains (ERC) and maximum correlation chain (MCC) models using support vector regression (SVR) and artificial neural network (ANN) based models with a multiple target ANN (MT-ANN) ability to fit the model development data. [Coefficient of determination, R^2; normalised root of mean squared error, nRMSE; mean absolute percentage error, MAPE].

Model	Output	R2	nRMSE	MAPE
ERC-SVR	COD	0.92	0.04	3.92
	TSS	0.92	0.07	3.97
	CH4	0.97	0.03	2.46
ERC-ANN	COD	0.98	0.02	1.53
	TSS	0.99	0.03	2.20
	CH4	0.99	0.02	1.50
MCC-SVR	COD	0.91	0.05	3.13
	TSS	0.91	0.08	4.43
	CH4	0.92	0.06	2.85
MCC-ANN	COD	0.94	0.05	1.83
	TSS	0.96	0.05	1.79
	CH4	0.95	0.04	1.89
MT-ANN	COD	0.92	0.05	3.03
	TSS	0.91	0.08	3.93
	CH4	0.94	0.05	3.19

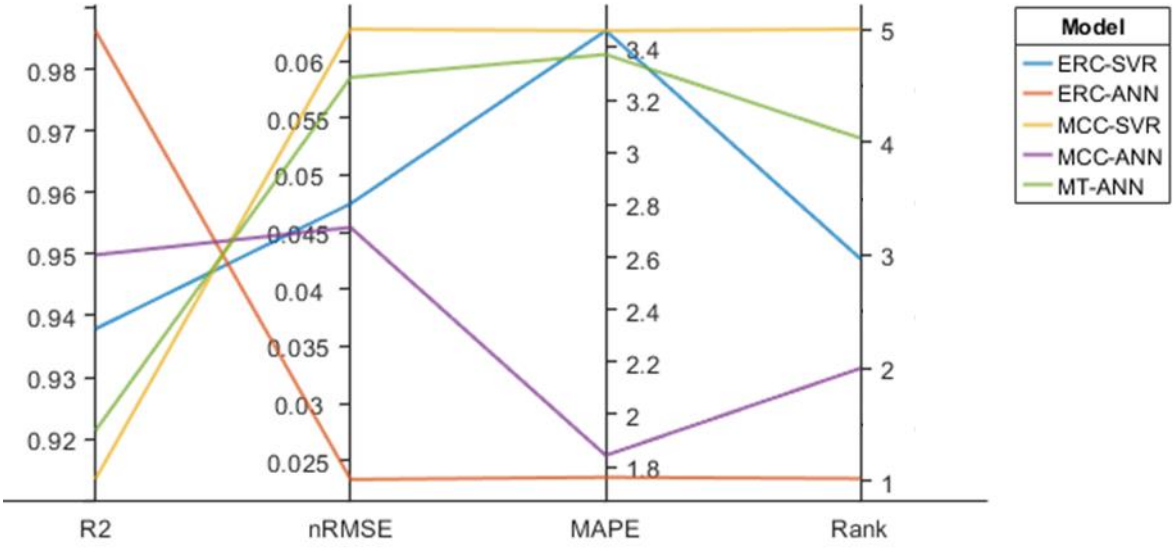


Fig. 28. Evaluation metrics of algorithm adaption modes. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 output. [Coefficient of determination, R^2; normalised root of mean squared error, nRMSE; mean absolute percentage error, MAPE; support vector regression, SVR; artificial neural network, ANN; ensemble of regressor chains, ECR; maximum correlation chain, MCC].

Comparing the average statistical error scores of the SVR and ANN based ST models and the MT models show little difference in performance (average for ST models: R^2 of 0.95, nRMSE of 0.06 and MAPE of 2.80; average for MT models: R^2 of 0.94, nRMSE of 0.05 and MAPE of 2.78). These results are inconsistent to previous work that has shown algorithm adaption models have been shown to definitely outperform problem

transformation models [26,31]. However, when comparing the average statistical error scores of the best performing ST model (ANN) with the best performing MT model (ECR-ANN), the algorithm adaption model is clearly superior reflecting the findings in the previous work (average ANN: R^2 of 0.95, nRMSE of 0.06 and MAPE of 2.80; average ECR-ANN: R^2 of 0.94, nRMSE of 0.05 and MAPE of 2.78).

The RC models built using an ANN as a based model were ranked first and second, while the RC models using a SVR as a base model were ranked third and fifth. Similarly, the ERC modelling ranked higher than the MCC model for when built with the same base model (ECR-ANN ranked first > MCC-ANN ranked second, ECR-SVR ranked third > MCC-SVR ranked fifth). This again suggests that the ANN is a superior base model and that the ERC MT modelling technique was superior to the MCC technique. Of particular interest was comparing the ANN RC models to the MT-ANN model, as these techniques have never been compared before. The MT-ANN model was outperformed by both the ECR-ANN and MCC-ANN models, shown in Fig. 6. The reason for this may be that the models were constructed on only 36 data-points. Multiple target ANN models are known to be less capable when modelling systems where data is limited [63]. This is because, as an ANN architecture increases its requirement for data points increases to avoid overfitting [56]. The RC ANN models built have smaller architectures (only one neuron in the outer layer), which may explain their better performance. Before drawing any final conclusions on the MT models, it is important to evaluate the models' performance when making predictions for unseen data. Testing the models on unseen data is utilised to evaluate then models' extrapolation capabilities beyond the data they were developed from [63].

4.1.3 Comparison of the models' performance on unseen data

A common error when developing a DDM, is to put too much emphasis on optimising the model's capability to fit the model development data [63]. The best DDMs are able to extrapolate beyond the data they were developed from, to make accurate predictions on unseen data. This is important if a process manufacturer wants to utilise the model for predictions on future data and to assess how much of the process manufacturing system's variability is captured in the model. The 16 unseen data points were partitioned from the manufacturing data to represent the parameter space both inside and outside the system. The prediction capabilities of the MT models for unseen data were statistical measured in terms of R^2, nRMSE and MAPE. Table 3 shows the statistical comparison of all models for each of the H^2AD bioprocess's outputs.

Table 14. Comparison of the ensemble of regressor chains (ERC) and maximum correlation chain (MCC) models using support vector regression (SVR) and artificial neural network (ANN) based models with a multiple target ANN (MT-ANN) predictive capabilities on unseen data. [Coefficient of determination, R^2; normalised root of mean squared error, nRMSE; mean absolute percentage error, MAPE].

Model	Output	R2		nRMSE		MAPE	
		Inside	Outside	Inside	Outside	Inside	Outside
ERC-SVR	COD	0.47	0.50	0.06	0.11	9.21	22.75
	TSS	0.75	0.88	0.12	0.11	6.26	12.72
	CH4	0.63	0.15	0.10	0.17	7.34	18.27
ERC-ANN	COD	0.33	0.82	0.17	0.10	10.17	16.23
	TSS	0.80	0.91	0.09	0.09	5.02	10.78

	CH4	0.75	0.36	0.08	0.14	6.58	16.10
MCC-SVR	COD	0.35	0.55	0.08	0.12	8.72	22.31
	TSS	0.75	0.90	0.12	0.11	7.34	11.71
	CH4	0.43	0.11	0.13	0.19	11.53	18.28
MCC-ANN	COD	0.62	0.55	0.13	0.16	8.23	21.40
	TSS	0.63	0.93	0.11	0.11	7.65	7.02
	CH4	0.82	0.49	0.06	0.14	7.25	14.43
M-ANN	COD	0.00	0.85	0.16	0.14	16.01	11.51
	TSS	0.40	0.50	0.21	0.33	10.89	17.69
	CH4	0.46	0.41	0.13	0.18	7.97	17.53

When evaluating the models' abilities to make predictions on unseen data within the system, it is of interest to note that the ECR is no longer clearly superior to the MCC technique; for example, both ERC-SVR and MCC-SVR have a R^2 of 0.75 for inside TSS predictions. The ERC-ANN model, is no longer a clearly superior model, due to its relatively poor performance when predicting the %COD removal (R^2 of 0.33, nRMSE of 0.17 and MAPE of 10.17). The ERC-SVR, ERC-ANN and MCC-ANN average performance are all relatively equal when making predictions for unseen data inside the system, each performing slightly better for a different statistical metric (ERC-SVR, nRMSE of 0.9; ERC-ANN, MAPE of 7.5; MCC-ANN, R^2 of 0.64). Although, the MT-ANN model remains significantly worse than all RC models (average R^2 of 0.29), emphasising the point raised earlier about the tendencies or larger ANN architectures to overfit. This highlights the advantages of the algorithm adaptions techniques developed by Spyromitros-Xioufis et al. and Melki et al. when building a MT model [26,47].

The ranking system described in Section 4.1.1 was again used to systematically determine which model was superior when making predictions on unseen both inside and outside of the system. The results in Fig. 7 show that the ECR-ANN model was the superior model. As the ECR-ANN has proved to be the statistically best model both when making predictions for the model development data and new unseen data, it was determined to be the best model to describe the three outputs of the H^2AD bioprocess. The ECR-ANN model was then used to increase the sustainability of the H^2AD bioprocess.

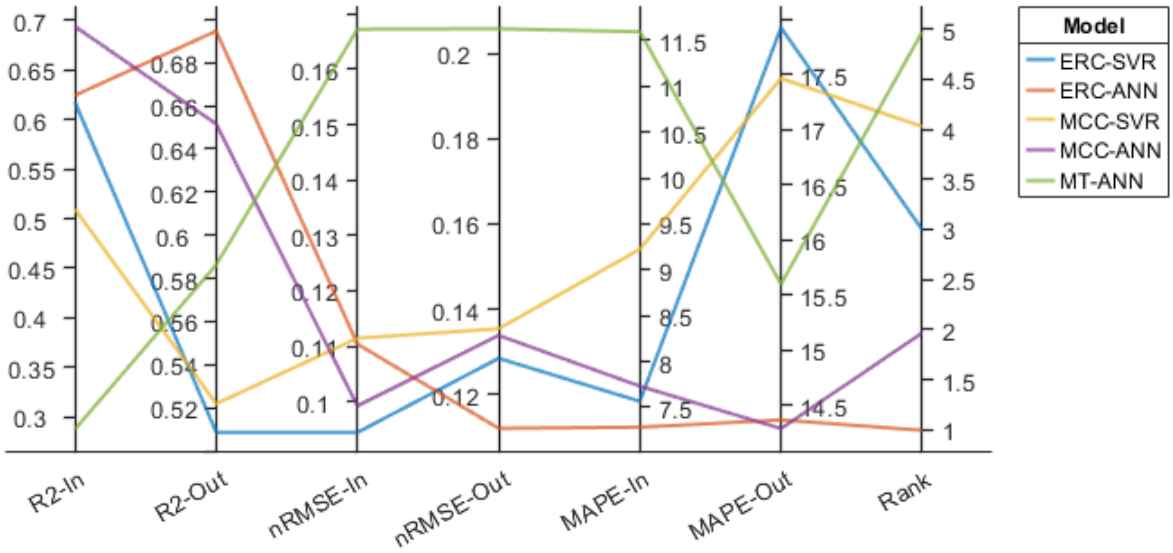


Fig. 29. Statistical analysis of algorithm adaption MT modes performance on unseen data. Where the black line is the average performance of all three outputs, the blue line is %COD removal output, the red line is %TSS removal output and the yellow line is CH4 output. [Coefficient of determination, R^2; normalised root of mean squared error, nRMSE; mean absolute percentage error, MAPE; support vector regression, SVR; artificial neural network, ANN; ensemble of regressor chains, ECR; maximum correlation chain, MCC].

4.2 Utilising the final model to optimise H2AD bioprocess sustainability

The H^2AD bioprocess can be manipulated to promote one output over the other by changing the set point of the H^2AD process conditions (HRT, pump speed, ST and SP). To better understand the effect of the H^2AD process conditions on the bioprocess, the ECR-ANN model was employed to generate three-dimensional surface plots. Fig. 8 and Figs A.1-5, show the effect of the four investigated H^2AD process conditions on the H^2AD bioprocess outputs (COD removal, TSS removal and CH_4 generation). The plots were produced by using the ECR-ANN to predict the effect of two H^2AD process conditions on the H^2AD bioprocess outputs, while holding wastewater quality parameters and the remaining two process conditions constant at intermediate values. The knowledge gained from the surface plots can inform the manufacturer how to optimise the bioprocess, dependant on the manufacturing environment's requirements, to find the most sustainable solution.

4.2.1 Understanding the relationship between process conditions and pollutant remediation and energy generation

Fig. 8 and Fig. A.1 shows the effect of HRT, pump speed, ST and SP on COD removal by the H^2AD bioprocess. The contour plot, in Fig. 8 (A), shows that at lower pump speeds the removal of COD is driven by the HRT. However, when the pump speed is increased beyond 9.2 L/min the effect of pump speed on COD removal increases. The contour plots in Fig. A.1 (A) and (B) also show that at pump speeds of less than 9.2 L/min the pump speed has less effect on the COD removal than the ST and SP respectively. If this was real data, rather than synthetic, it may be possible to hypothesise justifications for this observed phenomenon. For example, that pump speeds above 9.2 L/min cross a critical recirculation rate, within the reactor, that enables sufficient mixing of the fluid to reduce dead space within the reactor, increasing the reactor's efficiency. The

effect of recirculation rates on the volume of dead space within reactors is a well-studied field [64]. The advantage of the ECR-ANN based data analysis is that it is able to identify the critical recirculation rate and that is able to determine benefits of increasing recirculation rate in comparison to the other H^2AD process conditions. This can help to inform manufacturers which process conditions have a more significant effect on a system and therefore which ones to prioritise.

Figs. 8 (A) and (C) shows that, within the ranges investigated, HRT is the driving factor on COD removal compared to both over pump speed and SP, and that HRT rates should be prioritised if aiming to control the COD removal rates. Fig. 8 (B) indicate that HRT and ST have a similar effect and that both should be given consideration when attempting to manipulate COD removal rates. Generally, ECR-ANN based data analyses indicate that shorter HRT, higher pump speeds, higher set temperatures and lower system pressure increased the percentage removal of COD by the H^2AD bioprocess from the wastewater. By repeating this analysis for other H^2AD outputs (TSS removal and CH_4 generation) it was possible to identify the value of each process condition that will maximise each output.

Table 4 displays the values of each process condition that will maximise each output. By comparing these values it becomes apparent that outputs COD removal and CH_4 generation have almost the same configuration of process condition to maximise each output. The only difference is that a high ST increases the percentage of COD removed, but decreases the volume of CH_4 generated. Fig. 8 (B) and Fig. A.4 (B) indicate that ST has a stronger effect on COD removal rates than it does on CH_4 generation. Therefore, a manufacturer aiming to maximise both should consider setting a higher ST as this will have less of a negative effect on CH_4 than a low ST will have on percentage of COD removal. Manufacturers aiming to maximise all three H^2AD outputs are likely to face difficulty arising from conflicting optimal configuration of the HRT set point between TSS removal and the other two outputs. The ECR-ANN based data analyses indicate that HRT is a driving force for controlling all three of the H^2AD outputs. The TSS removal output favours a longer HRT, while the other two outputs favour a shorter HRT. Selecting which HRT to operate will depend on the manufacturer's goal, the incoming wastewater COD and TSS composition and the manufacturing environment operating within, as illustrated in Section 4.2.2.

Table 15 Optimal process conditions values for each of the H^2AD bioprocess outputs. [Hydraulic retention time, HRT; set temperature, ST; system pressure, SP; percentage chemical oxygen demand removal, %COD removal; percentage total suspend solid, %TSS removal; volume of methane generated, CH_4 generation].

Process condition	**%COD removal**	**%TSS removal**	**CH_4 generation**
HRT (days)	2.95	6.10	2.95
Pump speed (l/min)	9.94	9.94	9.94
ST (^oC)	39.20	39.20	29.6
SP (bar)	0.003	0.005	0.003

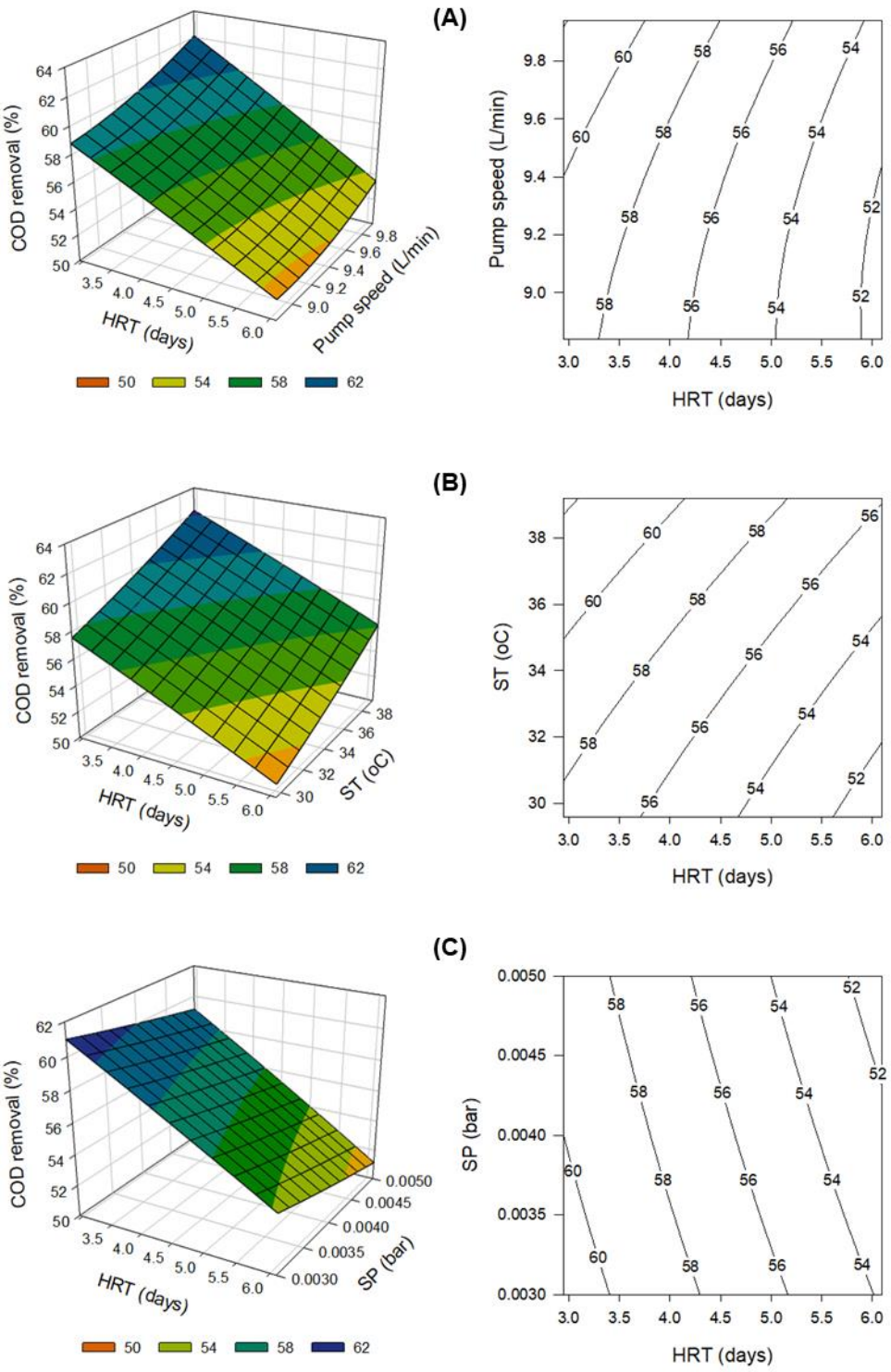


Fig. 30. Surface plots (left) and corresponding contour plots (right) showing the effects of the H²AD process conditions on the chemical oxygen demand (COD) removal as predicted by the ECR-ANN model with set temperature (ST) and system pressure (SP) held constant (A), with pump speed and SP held constant (B) and with pump speed and ST held constant (C). [Hydraulic retention time, HRT].

4.2.2 Determining the most sustainable configuration of the H²AD system

After analysing the impacts of the process conditions on the H²AD bioprocess, it became apparent when attempting to optimise for %COD removal and CH_4

generation, %TSS removal will decrease. Finding the pre-eminent solution will depend on maximising the economic and environmental benefits when treating a manufacturer's wastewater stream.

The environmental objective of the system is to minimise the pollutants leaving the system and maximise the energy recovered from the waste. This closely matches the economic objective of the H^2AD system to (A) reduce waste disposal costs associated with manufacturing process by reducing the pollutant load in wastewater and (B) reduce the cost of energy by recovering energy from the waste stream. However, there are also economic costs associated with operating the H^2AD. These are largely from the required energy to heat the incoming wastewater to the required system temperature and maintaining system temperature. These may come into conflict with each other, as operating the H^2AD system on higher temperatures was found to increase both pollutant removal and energy generation. The extra cost associated with at operating at higher temperatures may be greater than the economic benefits from the savings generated from the reduced pollutant load and onsite energy generation.

The most sustainable solution will differ between the manufacturing environments and the wastewater characteristics. When the H^2AD effluent is being emitted to a waste management body that charges for pollutant load, is the driving economic force of the system, the goal would be to maximise pollutant removal regardless of bioenergy generation. As the results from Section 4.2.1 indicate that %COD removal and %TSS removal have different optimal conditions, the process conditions configuration would depend on the characteristics of the wastewater being treated by the H^2AD system. However, if pollutant load charges in the effluent are lower the optimal economic solution may be to maximize energy recovery from the waste stream.

The ECR-ANN model was used to determine the economic benefit of optimising the H^2AD configuration dependant on wastewater composition and further treatment costs. Four different wastewater composition scenarios were proposed using different high and low values of COD and TSS concentrations, the values are presented in Table 5. The low and high values were taken as the highest and lowest values observed in the model development data from Section 3.2. The remaining wastewater characteristics were held constant and an intermediate value. The two H^2AD configurations were (1) optimising for COD% removal and CH_4 generation (configuration 1) and (2) optimising for TSS% removal. The H^2AD process conditions were taken from for each configuration were taken from results of the surface plot modelling in Table 4. The model results for each scenario and H^2AD configuration is presented in Table 5.

Table 16 Chemical oxygen demand (COD) and total dissolved solids (TSS) concentration in the H^2AD input and output streams for each wastewater composition scenario for both H^2AD set ups.

			Configuration 1		**Configuration 2**	
Scenario	**Pollutant**	**Pollutant concentration (mg/L)**	**Pollutant conc. (mg/L)**	**Methane generated (L/day)**	**Pollutant conc. (mg/L)**	**Methane generated (L/day)**
1	COD	40,000	11900	5270	20500	4040
	TSS	30,000	24100		21000	
2	COD	40,000	12300	5250	20500	4130

	TSS	13,000	10600		9060	
3	COD	17,000	5070	5240	8740	4010
	TSS	30,000	24100		21000	
4	COD	17,000	5070	5120	8740	4000
	TSS	13,000	10600		9060	

Economic analysis was conducted to identify which H^2AD configuration was the most sustainable for each scenario dependant on if the wastewater was being sent to a wastewater treatment company that charge a high amount (United Utilities) or a low amount (Wessex Water). The disposal charges were calculated for each scenario using the Mogden formula, assuming $4m^3$/day of wastewater was being treated [41]. The energy value from the methane gas was calculated by using equation 5.7 from Bauer *et al.* [42] and then converted to kilowatt hours (KWH). A biogas generator is not capable of converting all this energy to electricity, as approximately 65% of the energy is lost as heat and other mechanical losses [42]; therefore, the predicted energy generated by the H^2AD was reduced by 65% to account for inefficiencies. The average UK price of £0.124 per KWH at time of writing was used to determine the economic benefit from the biogas [65].

$$1\ m^3\ of\ CH_4 = 39.79\ MJ \tag{5}$$

The economic analysis of the ECR-ANN model results are presented in Table 6. The analysis indicated that for manufacturing environments that have a low disposal charge for H^2AD effluent, the optimal H^2AD configuration is the removal of COD and generation of CH_4. This is true for all the wastewater composition scenarios, though the largest economic benefit (£20.00 per day) from optimising the H^2AD configuration was for high COD and low TSS concentrations in the wastewater. Optimising the H^2AD configuration would save the process manufacturer £7000.00 per annum in this scenario. When operating in a manufacturing environment with a high disposal charge, the H^2AD configuration that favoured TSS removal produced the largest economic saving. For this manufacturing
environment, the biggest economic benefit was for low COD and high TSS concentration wastewater scenario. By using a H^2AD configuration to optimise TSS removal would benefit the process manufacturer by £9125.00 per annum.

Table 17 Economic analysis of the wastewater disposal costs before and post H2AD treatment and the economic benefit from the methane generated.

		Configuration 1					Configuration 2			
Wastewater company	**Scenario**	**Disposal charge with no treatment** £/day	**Disposal charge post H2AD** £/day	**Saving** £/day	**Value from methane** £/day	**Total economic benefit** £/day	**Disposal charge post H2AD** £/day	**Saving** £/day	**Value from methane** £/day	**Total economic benefit** £/day
	1	620	390	230	7	**240**	390	230	6	**240**
United	2	380	210	170	7	**180**	230	150	6	160
Utilities	3	500	360	140	7	150	340	160	6	**170**
	4	270	170	100	7	100	170	100	5	**110**
	1	170	90	80	7	**90**	110	60	6	70
Wessex	2	130	60	70	7	**80**	70	60	6	70
Water	3	120	80	40	7	**50**	80	40	6	**50**
	4	80	40	40	7	**50**	50	30	5	40

5. Conclusion

When searching for the most sustainable manufacturing solution, multiple economic and environmental objectives must be considered. The objectives often come into conflict with one another. By developing a MT model of the manufacturing system, one would be able to investigate the impact of various manufacturing factors on the objectives and understand how the objectives relate to one another. This will enable manufacturers to determine the most sustainable configuration of their process manufacturing systems. Despite recent advances in MT modelling, there are limited examples of MT models being developed for process manufacturing systems. The MT models that have been developed frequently use a problem transformation approach, where multiple independent ST models are developed for each output. This fails to capture the relationship between the model's outputs, resulting in models with inferior predictive capabilities.

In this study, two state-of-the-art MT modelling approaches were applied to a manufacturing system for the first time. These were RC based techniques called ERC and MCC. The RC technique treats the outputs of the system as inputs for subsequent outputs, thus capturing the relationship between model outputs. Previous work has used only RF and SVR algorithms as base models for RC models. This study was the first example of an ANN base model being used for either an ERC or a MCC MT model. Previous work has shown the versatility of ANNs when modelling process manufacturing systems [63]. This was reflected in this study, as the ANN was able to consistently outperform the ERC and MCC built from SVR base models. After statistical analysis of the model predictions on unseen data and after comparing the models' predictions distributions with that of the real data, the ECR-ANN model was found to be the superior model.

Utilising the ECR-ANN model to develop surface plots of %COD removal, %TSS removal and CH_4 generation, it was found that the %COD and CH_4 generation favoured closely similar configurations of the H^2AD system's process conditions. However, the %TSS removal was found to have the opposite relationship with all process conditions apart from the system temperature. This demonstrated the complex task when identifying the most sustainable solution, balancing both economic and environmental considerations. What is sustainable for one manufacturing environment may not be for another, as different environments have different economic and environmental objectives. Predictions form the ECR-ANN model were able to achieve economic benefits up of £9125 per annum.

Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing and [grant number EP/P001246/1], Network Plus: Industrial Systems in a Digital Age.

Acknowledgements

The author (RLG) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing

– Towards Resilient and Scalable High Value Manufacturing. The author (OJF) would like to acknowledge the University of Nottingham, Faculty of Engineering for his PhD scholarship. The author (OJF) would also like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) Network Plus: Industrial Systems in a Digital Age [grant number EP/P001246/1], Connected Everything for funding his placement at Lindhurst Engineering Ltd.

References

[1] T. Stock, G. Seliger, Opportunities of Sustainable Manufacturing in Industry 4 . 0, Procedia CIRP. 40 (2016) 536–541. doi:10.1016/j.procir.2016.01.129.

[2] R.Y. Zhong, X. Xu, E. Klotz, S.T. Newman, Intelligent Manufacturing in the Context of Industry 4.0: A Review, Engineering. 3 (2017) 616–630. doi:10.1016/J.ENG.2017.05.015.

[3] O. Fisher, N. Watson, L. Porcu, D. Bacon, M. Rigley, R.L.R.L. Gomes, Cloud manufacturing as a sustainable process manufacturing route, J. Manuf. Syst. 47 (2018). doi:10.1016/j.jmsy.2018.03.005.

[4] H. Schichl, Models and the History of Modeling, in: K. Josef (Ed.), Model. Lang. Math. Optim., 1st ed., Springer US, Boston, 2004: pp. 25–36. doi:10.1007/978-1-4613-0215-5.

[5] D. Solomatine, L.M. See, R.J. Abrahart, Data-Driven Modelling : Concepts , Approaches and Experiences, in: R.J. Abrahart, L.M. See, D.P. Solomatine (Eds.), Pract. Hydroinformatics. Water Sci. Technol. Libr., Springer, Berlin, 2008: pp. 17–31. doi:https://doi.org/10.1007/978-3-540-79881-1.

[6] M.P. Saka, E. Doğan, I. Aydogdu, Analysis of Swarm Intelligence–Based Algorithms for Constrained Optimization, Swarm Intell. Bio-Inspired Comput. (2013) 25–48. doi:10.1016/B978-0-12-405163-8.00002-8.

[7] S. Kim, MATLAB Deep Learning With Machine Learning, Neural Networks and Artificial Intelligence, 1st ed., Apress, 2017. doi:10.1007/978-1-4842-2845-6.

[8] C.W. Coley, W.H. Green, K.F. Jensen, Machine Learning in Computer-Aided Synthesis Planning, Acc. Chem. Res. 51 (2018) 1281–1289. doi:10.1021/acs.accounts.8b00087.

[9] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

[10] J. Wang, Y. Ma, L. Zhang, R.X. Gao, Deep learning for smart manufacturing: Methods and applications, J. Manuf. Syst. 48 (2018) 144–156. doi:10.1016/J.JMSY.2018.01.003.

[11] M. von Stosch, R. Oliveira, J. Peres, S. Feyo de Azevedo, Hybrid semi-parametric modeling in process systems engineering: Past, present and future, Comput. Chem. Eng. (2014). doi:10.1016/j.compchemeng.2013.08.008.

[12] Z. Ge, Review on data-driven modeling and monitoring for plant-wide industrial processes, Chemom. Intell. Lab. Syst. 171 (2017) 16–25.

doi:10.1016/J.CHEMOLAB.2017.09.021.

[13] W. Te Liew, A. Adhitya, R. Srinivasan, Sustainability trends in the process industries: A text mining-based analysis, Comput. Ind. 65 (2014) 393–400. doi:10.1016/j.compind.2014.01.004.

[14] G. Al-Sharrah, A. Elkamel, A. Almanssoor, Sustainability indicators for decision-making and optimisation in the process industry: The case of the petrochemical industry, Chem. Eng. Sci. 65 (2010) 1452–1461. doi:10.1016/j.ces.2009.10.015.

[15] B.R. Bakshi, J. Fiksel, The quest for sustainability: Challenges for process systems engineering, AIChE J. 49 (2003) 1350–1358. doi:10.1002/aic.690490602.

[16] Z. Ge, Z. Song, S.X. Ding, B. Huang, Data Mining and Analytics in the Process Industry: The Role of Machine Learning, IEEE Access. 5 (2017) 20590–20616. doi:10.1109/ACCESS.2017.2756872.

[17] H.A. Hamid, Y. Jenidi, C. Somerfield, R.L. Gomes, Predicting the capability of carboxylated cellulose nanowhiskers for the remediation of copper from water using response surface methodology (RSM) and artificial neural network (ANN) models, Ind. Crops Prod. 93 (2016) 108–120. doi:10.1016/J.INDCROP.2016.05.035.

[18] J. Escrig, E. Woolley, S. Rangappa, A. Simeone, N.J. Watson, J. Escrig, Escrig, E. Woolley, S. Rangappa, A. Simeone, N.J. Watson, Clean-in-place monitoring of different food fouling materials using ultrasonic measurements, Food Control. 104 (2019) 358–366. doi:10.1016/J.FOODCONT.2019.05.013.

[19] J. Ling, R. Jones, J. Templeton, Machine learning strategies for systems with invariance properties, J. Comput. Phys. 318 (2016) 22–35. doi:10.1016/J.JCP.2016.05.003.

[20] A. Baghban, M. Kahani, M.A. Nazari, M.H. Ahmadi, W.-M. Yan, Sensitivity analysis and application of machine learning methods to predict the heat transfer performance of CNT/water nanofluid flows through coils, Int. J. Heat Mass Transf. 128 (2019) 825–835. doi:10.1016/J.IJHEATMASSTRANSFER.2018.09.041.

[21] S. Kolouri, S.R. Park, M. Thorpe, D. Slepcev, G.K. Rohde, Optimal Mass Transport: Signal processing and machine-learning applications, IEEE Signal Process. Mag. 34 (2017) 43–59. doi:10.1109/MSP.2017.2695801.

[22] Z.W. Ulissi, A.J. Medford, T. Bligaard, J.K. Nørskov, To address surface reaction network complexity using scaling relations machine learning and DFT calculations, Nat. Commun. 8 (2017) 14621. doi:10.1038/ncomms14621.

[23] G. Halder, S. Dhawane, P.K. Barai, A. Das, Optimizing chromium (VI) adsorption onto superheated steam activated granular carbon through response surface methodology and artificial neural network, Environ. Prog. Sustain. Energy. 34 (2015) 638–647. doi:10.1002/ep.12028.

[24] S.A. Shahin, M. Mossad, M. Fouad, Evaluation of copper removal efficiency using water treatment sludge, Water Sci. Eng. 12 (2019) 37–44. doi:10.1016/j.wse.2019.04.001.

[25] K. Larson Lesnik, H. Liu, Predicting Microbial Fuel Cell Biofilm Communities and Bioreactor Performance using Artificial Neural Networks, (2017). doi:10.1021/acs.est.7b01413.

[26] G. Melki, A. Cano, V. Kecman, S. Ventura, Multi-target support vector regression via correlation regressor chains, Inf. Sci. (Ny). 415–416 (2017) 53–69. doi:10.1016/J.INS.2017.06.017.

[27] H. Borchani, G. Varando, C. Bielza, P. Larrañaga, A survey on multi-output regression, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 5 (2015) 216–233. doi:10.1002/widm.1157.

[28] Q. Liu, Q. Xu, V.W. Zheng, H. Xue, Z. Cao, Q. Yang, Multi-task learning for cross-platform siRNA efficacy prediction: an in-silico study, BMC Bioinformatics. 11 (2010) 181. doi:10.1186/1471-2105-11-181.

[29] A.J. Burnham, J.F. MacGregor, R. Viveros, Latent variable multivariate regression modeling, Chemom. Intell. Lab. Syst. 48 (1999) 167–180. doi:10.1016/S0169-7439(99)00018-0.

[30] D. Kocev, S. Džeroski, M.D. White, G.R. Newell, P. Griffioen, Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition, Ecol. Modell. 220 (2009) 1159–1168. doi:10.1016/J.ECOLMODEL.2009.01.037.

[31] D. Kocev, M. Ceci, Ensembles of Extremely Randomized Trees for Multi-target Regression, in: Springer, Cham, 2015: pp. 86–100. doi:10.1007/978-3-319-24282-8_9.

[32] M. Jeong, G.G. Lee, Multi-domain spoken language understanding with transfer learning, Speech Commun. 51 (2009) 412–424. doi:10.1016/j.specom.2009.01.001.

[33] T. Xiong, Y. Bao, Z. Hu, Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting, Knowledge-Based Syst. 55 (2014) 87–100. doi:10.1016/J.KNOSYS.2013.10.012.

[34] S. Curteanu, C.G. Piuleac, K. Godini, G. Azaryan, Modeling of electrolysis process in wastewater treatment using different types of neural networks, Chem. Eng. J. (2011). doi:10.1016/j.cej.2011.05.104.

[35] Y. Liu, Y. Pan, D. Huang, Q. Wang, Fault prognosis of filamentous sludge bulking using an enhanced multi-output gaussian processes regression, Control Eng. Pract. 62 (2017) 46–54. doi:10.1016/J.CONENGPRAC.2017.02.003.

[36] A.M. Schweidtmann, A.D. Clayton, N. Holmes, E. Bradford, R.A. Bourne, A.A. Lapkin, Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives, Chem. Eng. J. 352 (2018) 277–282. doi:10.1016/J.CEJ.2018.07.031.

[37] Z. Wang, G.P. Rangaiah, Application and Analysis of Methods for Selecting an Optimal Solution from the Pareto-Optimal Front obtained by Multiobjective Optimization, Ind. Eng. Chem. Res. 56 (2017) 560–574. doi:10.1021/acs.iecr.6b03453.

[38] OECD, OECD Environmental Outlook to 2050, OECD Publishing, 2012. doi:10.1787/9789264122246-en.

[39] U.S. Energy Information Administration, International Energy Outlook 2016, Washington, 2016. www.eia.gov (accessed September 2, 2019).

[40] J.D. Kabongo, Waste Valorization, in: S.O. Idowu, N. Capaldi, L. Zu, A. Das Gupta (Eds.), Encycl. Corp. Soc. Responsib., Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 2701–2706. doi:10.1007/978-3-642-28036-8_680.

[41] Tools: Water Efficiency in the Food and Drink industry | WRAP UK, (n.d.). http://www.wrap.org.uk/node/31479 (accessed August 13, 2019).

[42] A. Bauer, C. Leonhartsberger, P. Bösch, B. Amon, A. Friedl, T. Amon, Analysis of methane yields from energy crops and agricultural by-products and estimation of energy potential from sustainable crop rotation systems in EU-27, Clean Technol. Environ. Policy. 12 (2010) 153–161. doi:10.1007/s10098-009-0236-1.

[43] G. Tsoumakas, E. Spyromitros-Xioufis, A. Vrekou, I. Vlahavas, Multi-target Regression via Random Linear Target Combinations, in: Springer, Berlin, Heidelberg, 2014: pp. 225–240. doi:10.1007/978-3-662-44845-8_15.

[44] D. Kocev, C. Vens, J. Struyf, S. Džeroski, Ensembles of Multi-Objective Decision Trees, in: Mach. Learn. ECML 2007, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007: pp. 624–631. doi:10.1007/978-3-540-74958-5_61.

[45] E. Hadavandi, J. Shahrabi, S. Shamshirband, A novel Boosted-neural network ensemble for modeling multi-target regression problems, Eng. Appl. Artif. Intell. 45 (2015) 204–219. doi:10.1016/J.ENGAPPAI.2015.06.022.

[46] R. Ranjan, V.M. Patel, R. Chellappa, HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 121–135. doi:10.1109/TPAMI.2017.2781233.

[47] E. Spyromitros-Xioufis, · Grigorios Tsoumakas, W. Groves, I. Vlahavas, G. Tsoumakas, Multi-target regression via input space expansion: treating targets as inputs, Mach. Learn. 104 (2016) 55–98. doi:10.1007/s10994-016-5546-z.

[48] M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy Build. 147 (2017) 77–89. doi:10.1016/J.ENBUILD.2017.04.038.

[49] B. Yildiz, J.I. Bilbao, A.B. Sproul, A review and analysis of regression and machine learning models on commercial building electricity load forecasting, Renew. Sustain. Energy Rev. 73 (2017) 1104–1122. doi:10.1016/j.rser.2017.02.023.

[50] D.M. Himmelblau, Applications of artificial neural networks in chemical engineering, Korean J. Chem. Eng. 17 (2000) 373–392. doi:10.1007/BF02706848.

[51] J. Baxter, Jonathan, Learning internal representations, in: Proc. Eighth Annu. Conf. Comput. Learn. Theory - COLT '95, ACM Press, New York, New York, USA,

1995: pp. 311–320. doi:10.1145/225298.225336.

[52] R. Caruana, Learning Many Related Tasks at the Same Time with Backpropagation, in: G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Adv. Neural Inf. Process. Syst. 7, MIT Press, 1995: pp. 657–664. http://papers.nips.cc/paper/959-learning-many-related-tasks-at-the-same-time-with-backpropagation.pdf.

[53] G. Albuquerque, T. Lowe, M. Magnor, Synthetic generation of high-dimensional datasets, IEEE Trans. Vis. Comput. Graph. 17 (2011) 2317–2324. doi:10.1109/TVCG.2011.237.

[54] I. Hobæk Haff, K. Aas, A. Frigessi, On the simplified pair-copula construction - Simply useful or too simplistic?, J. Multivar. Anal. 101 (2010) 1296–1310. doi:10.1016/j.jmva.2009.12.001.

[55] F.P. Brissette, M. Khalili, R. Leconte, Efficient stochastic generation of multi-site synthetic precipitation data, J. Hydrol. 345 (2007) 121–133. doi:10.1016/J.JHYDROL.2007.06.035.

[56] S.M. Al-Fattah, H. A Al-Naim, H.A. Al-Naim, Artificial-Intelligence Technology Predicts Relative Permeability of Giant Carbonate Reservoirs, 2009. doi:10.2118/109018-PA.

[57] W. Yang, K. Wang, W. Zuo, Neighborhood Component Feature Selection for High-Dimensional Data Image enhancement and restoration View project Ischemia/Reperfusion-induced arrhythmias View project Neighborhood Component Feature Selection for High-Dimensional Data, J. Comput. 7 (2012) 161–168. doi:10.4304/jcp.7.1.161-168.

[58] M.H. Abbas, R. Norman, A. Charles, Neural network modelling of high pressure CO 2 corrosion in pipeline steels, Process Saf. Environ. Prot. 119 (2018) 36–45. doi:10.1016/j.psep.2018.07.006.

[59] F. Charte, I. Romero, M.D. Pï¿½rez-Godoy, A.J. Rivera, E. Castro, Comparative analysis of data mining and response surface methodology predictive models for enzymatic hydrolysis of pretreated olive tree biomass, Comput. Chem. Eng. (2017). doi:10.1016/j.compchemeng.2017.02.008.

[60] M. Dalmau, N. Atanasova, S. Gabarrón, I. Rodriguez-Roda, J. Comas, Comparison of a deterministic and a data driven model to describe MBR fouling, Chem. Eng. J. (2015). doi:10.1016/j.cej.2014.09.003.

[61] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688. doi:10.1016/j.ijforecast.2006.03.001.

[62] J. Levatić, M. Ceci, D. Kocev, S. Džeroski, Self-training for multi-target regression with tree ensembles, Knowledge-Based Syst. 123 (2017) 41–60. doi:10.1016/J.KNOSYS.2017.02.014.

[63] J. Panerati, M.A. Schnellmann, C. Patience, G. Beltrame, G.S. Patience, Experimental methods in chemical engineering: Artificial neural networks–ANNs, Can. J. Chem. Eng. 97 (2019) 2372–2382. doi:10.1002/cjce.23507.

[64] H.S. Fogler, Essentials of chemical reaction engineering, 2nd ed., Prentice Hall,

2017.

[65] UK Power, Compare Energy Prices Per kWh | Gas & Electric Per Unit | UKPower, (n.d.). https://www.ukpower.co.uk/home_energy/tariffs-per-unit-kwh (accessed September 30, 2019).

APPENDIX F: H^2AD MODELLING DATA

Table 1 Water quality data of H^2|AD input stream used in chapter 5

pH	DO	Cond.	TSS	COD	TC	TIC	TOC	tN	NH4+	NO2-	NO3-	Cl-	WH	PO43-	SO4-2	Zn	Cu
7.52	0.28	10.63	14800	20420	7940	1900	6040	1392	738	8.90	57.50	1260	165	216	3775	24.40	48.10
7.48	0.17	12.32	16300	25120	7900	1270	6630	1424	858	10.90	71.50	1370	177	292	4530	29.00	74.70
7.66	0.24	12.26	18200	25360	7800	1160	6650	1540	880	10.75	71.00	1410	189	279	4520	31.80	71.50
7.49	0.16	12.38	16200	24860	7870	1230	6640	1542	856	9.95	68.50	1335	180	268	4330	32.70	72.50
7.40	0.20	12.36	14000	25240	7810	1270	6540	1578	882	10.00	66.50	1430	190	281	4450	31.50	74.70
7.43	0.19	12.28	15600	24180	7800	1250	6560	1566	864	9.50	63.50	1345	186	261	4360	28.50	67.40
7.48	0.21	12.24	15700	24300	7830	1350	6480	1476	950	10.00	68.50	1535	183	270	4425	33.10	70.70
7.39	0.22	11.23	14550	22080	7710	1170	6540	1520	854	9.20	62.00	1280	179	239	3870	27.30	58.00
7.48	0.19	11.60	13500	21460	7600	1240	6360	1546	842	8.85	59.50	1305	170	222	3865	31.60	63.30
7.43	0.20	11.31	14300	21800	7770	1270	6490	1438	868	9.15	61.50	1365	177	231	3930	28.20	59.00
7.41	0.22	11.43	14850	21540	7590	1300	6290	1468	800	9.10	62.00	1305	171	241	4065	27.90	59.60
7.40	0.21	11.49	14900	22060	7710	1260	6450	1410	872	9.00	62.00	1300	175	243	4085	27.70	66.70
7.44	0.13	11.42	20350	25760	7920	1600	6320	1480	852	11.10	80.00	1490	192	320	4690	35.80	75.80
7.52	0.17	11.75	11100	18900	7340	1050	6290	1304	700	8.40	51.50	1345	160	197	3475	28.90	52.50
7.64	0.16	11.65	13600	22860	7590	1160	6430	1428	744	9.70	61.50	1280	170	227	3745	29.60	60.00
7.56	0.18	11.53	15000	23320	7800	1160	6640	1554	796	10.10	63.50	1375	188	238	4080	33.80	66.40
7.22	0.14	12.45	17760	26960	4560	1140	3540	1421	770	12.21	83.49	1991	223	299	3540	11.24	44.60

7.64	0.16	12.31	14600	17980	7440	904	6540	1624	632	10.00	69.00	1645	184	247	6540	12.28	36.00
7.90	0.13	11.97	12680	17440	6900	828	6060	1158	604	8.45	51.00	1185	167	179	6060	10.26	22.00
7.31	0.11	11.48	19150	28080	10400	1256	9140	770	682	11.05	82.50	1800	220	288	4480	16.30	61.60
7.91	0.16	12.84	19200	21920	8160	954	7240	2440	1116	12.80	91.60	2290	214	258	6100	15.68	58.80
7.49	0.19	12.63	25424	29097	10880	1258	9622	3238	1492	16.91	119.25	3018	281	335	8171	23.41	79.27
7.46	0.23	12.84	24400	30520	12840	1800	10640	2644	1052	14.50	109.50	1655	236	457	6180	20.40	75.00
7.31	0.21	11.85	26900	35720	14120	1928	12200	1944	880	14.00	98.50	1708	262	468	6780	20.40	76.40
7.51	0.17	12.23	28000	38160	14120	2060	12080	2008	1040	13.60	113.00	1780	266	529	6640	19.18	73.80
7.36	0.19	10.30	29200	37880	13680	2156	11520	1940	992	14.35	130.50	1500	254	504	6790	23.40	84.20
7.56	0.20	11.95	27100	31840	11640	2068	9560	2156	1148	12.70	125.00	1676	240	409	6670	21.00	81.80
7.63	0.13	11.19	35950	40980	14460	2300	12160	2280	974	16.85	142.00	2390	173	300	7150	31.10	101.00
7.37	0.14	12.32	13907	21040	5367	866	4442	1366	943	10.72	78.89	2203	195	217	4091	24.54	53.33
7.71	0.20	11.64	15152	20993	7459	1294	6071	1098	664	9.22	66.89	1270	173	274	4829	14.53	29.08
7.28	0.24	11.36	24856	34959	10498	1969	8186	2929	1422	15.82	112.44	2210	245	475	4552	33.39	80.94
7.63	0.12	11.46	20007	26313	11498	1287	10115	2737	1184	15.62	100.57	2881	237	345	6925	23.92	54.15
7.76	0.22	10.51	13834	19202	6589	1434	5060	1321	696	9.58	62.44	1346	171	215	4280	16.14	24.19
7.62	0.18	11.59	24683	31367	12570	1545	10953	2254	882	14.35	109.20	1881	239	427	7339	20.99	63.30
7.77	0.14	12.56	23104	29650	9855	1403	8455	2882	1363	15.42	112.91	2828	242	382	7158	25.93	66.93
7.89	0.19	12.78	21891	27906	10124	1337	8724	2602	1094	13.08	99.41	1620	219	418	7755	22.95	60.92
7.78	0.11	11.60	15114	20419	7004	998	6217	1059	709	9.34	65.18	1790	187	250	5929	15.70	25.69
7.63	0.11	10.59	24790	33302	12203	1531	10740	1824	872	15.08	109.82	2329	255	433	6933	17.19	48.52
7.72	0.13	12.01	23192	30616	8712	1475	7291	1380	707	12.88	95.79	1624	222	412	5567	13.49	39.34
7.67	0.14	12.14	29001	37994	13507	2023	11624	2215	1212	16.66	129.60	2534	277	525	7972	12.28	71.49
7.28	0.13	11.35	27468	36401	11102	1711	9527	2557	1317	16.60	119.85	2910	265	465	6337	31.06	80.39
7.45	0.12	10.77	26516	34925	10115	1533	8691	2103	1302	15.88	121.73	2779	266	408	6474	22.40	64.81

7.82	0.21	12.80	19666	23077	8289	1297	6857	2932	1348	14.16	96.67	2844	228	299	6637	23.90	52.98
7.55	0.12	10.35	12279	18299	4957	899	3990	879	734	8.66	56.16	1573	167	184	3755	30.19	29.88
7.71	0.20	11.15	16318	23697	7470	1301	6220	1993	946	11.71	78.37	2117	207	273	5733	22.60	37.18
7.25	0.41	10.55	29300	42760	13840	2112	11760	2004	952	13.75	140.50	1625	276	444	6820	25.60	89.20
7.78	0.15	11.06	52300	48640	15800	3540	12260	2744	1236	22.45	191.50	2240	388	691	8600	38.60	140.40
7.56	0.18	12.67	44900	42480	16600	2460	14160	2372	1112	20.75	154.00	2400	308	611	8780	30.40	115.80
7.63	0.12	13.22	41000	44160	18600	2644	15920	2452	1180	18.05	169.40	2570	292	660	8770	31.20	123.60
7.61	0.14	12.94	48900	46360	17400	3068	14320	2708	1216	20.15	192.40	2715	348	744	9800	35.80	132.60
7.60	0.17	12.40	44500	42440	16720	2788	13920	2788	1236	18.80	162.80	2320	302	670	9000	32.80	125.00
7.79	0.17	14.52	33600	33080	15400	3088	12320	2420	1212	15.74	134.80	2180	290	486	7570	30.40	96.00

Table 2 H2AD process data used in chapter 5

HRT	Pump speed	Process set temp.	System pressure
3.48	8.84	36.10	0.003
3.48	8.84	36.30	0.003
3.05	8.84	35.70	0.004
3.48	8.84	37.00	0.003
3.48	9.94	31.80	0.003
3.05	9.94	31.90	0.003
3.48	9.94	30.70	0.003
3.05	9.94	31.80	0.004
4.87	9.94	30.10	0.005
4.06	9.28	29.80	0.004
3.48	9.72	30.90	0.004
3.48	9.72	31.70	0.003
3.05	9.72	32.00	0.004
6.09	9.72	30.00	0.003
4.87	9.72	30.00	0.003
3.48	9.72	31.70	0.003
3.48	9.72	29.60	0.004
3.05	9.72	33.90	0.003
3.48	9.72	34.00	0.003
3.48	9.72	33.80	0.003
3.48	9.94	33.80	0.004
4.87	9.94	34.00	0.004
3.05	9.94	33.70	0.003
3.48	9.94	33.50	0.004
3.05	9.94	31.50	0.004
3.05	9.94	33.90	0.003
3.48	9.94	34.00	0.003
3.05	9.94	33.70	0.003
2.68	9.94	33.80	0.003
3.76	9.94	33.60	0.004
3.13	9.94	34.00	0.003
5.63	9.94	33.70	0.003
5.03	9.94	32.80	0.004
2.25	9.94	33.40	0.003
3.60	9.94	33.30	0.003
2.95	9.94	29.80	0.003
5.14	9.94	39.20	0.003
3.27	9.94	33.60	0.003
6.00	9.94	33.50	0.003
2.57	9.94	33.90	0.003

3.00	9.94	34.10	0.003
3.60	8.84	33.60	0.003
3.27	8.84	33.70	0.003
3.20	8.84	33.40	0.003
3.20	8.84	33.50	0.003
4.06	9.94	33.90	0.003
3.48	9.94	33.70	0.003
3.48	9.94	33.90	0.004
3.05	9.94	33.80	0.004
3.48	9.94	33.70	0.004
3.48	9.94	33.80	0.004
5.03	9.94	33.70	0.004

Table 3 Chapter 5 part (A) data-driven modelling results

Actual value	GP	Residual	RF	Residual	SVR	Residual	ANN	Residual
			Model development data					
72.319	66.568	5.751	71.228	1.092	71.442	0.878	74.529	-2.210
60.747	60.525	0.222	60.587	0.160	59.869	0.878	58.478	2.270
65.264	65.265	-0.001	65.643	-0.380	66.686	-1.422	64.850	0.414
68.933	68.932	0.001	67.630	1.303	68.055	0.878	65.781	3.152
60.927	60.928	-0.002	62.643	-1.716	61.805	-0.878	60.481	0.445
70.743	70.742	0.001	68.601	2.142	69.865	0.878	70.205	0.538
64.952	64.952	0.000	64.722	0.230	64.075	0.878	64.642	0.310
64.717	64.719	-0.001	65.052	-0.335	63.840	0.877	66.940	-2.223
63.946	63.947	-0.001	65.247	-1.301	65.307	-1.362	68.190	-4.245
68.387	68.387	0.000	67.599	0.787	71.918	-3.532	68.046	0.341
71.959	71.959	0.001	68.982	2.978	71.082	0.877	71.701	0.258
64.888	64.889	0.000	64.901	-0.013	65.766	-0.878	64.574	0.315
65.668	65.668	0.001	65.607	0.062	64.790	0.878	65.631	0.037
68.194	68.193	0.001	67.558	0.635	67.606	0.588	67.679	0.515
62.831	62.828	0.003	59.650	3.182	61.953	0.878	62.282	0.549
74.867	74.867	0.000	69.183	5.684	75.988	-1.121	74.876	-0.009

60.839	60.837	0.002	61.376	-0.537	59.962	0.878	58.586	2.253
62.061	58.854	3.207	62.552	-0.491	62.939	-0.878	61.999	0.062
66.302	66.302	0.001	65.918	0.384	67.180	-0.878	66.217	0.085
56.913	56.915	-0.002	59.304	-2.391	57.790	-0.877	57.011	-0.098
67.756	67.757	-0.001	67.552	0.204	68.634	-0.878	67.715	0.041
63.918	63.916	0.001	64.240	-0.323	63.039	0.878	63.663	0.254
60.773	60.773	0.000	62.820	-2.047	64.294	-3.521	60.201	0.571
57.884	57.884	0.000	57.369	0.514	58.761	-0.878	57.750	0.134
69.074	68.488	0.585	67.529	1.545	68.849	0.225	68.803	0.270
59.829	59.828	0.000	59.338	0.490	60.706	-0.878	61.536	-1.708
56.163	56.164	-0.001	57.776	-1.613	57.040	-0.877	56.146	0.017
65.548	65.545	0.002	62.995	2.553	64.670	0.877	64.752	0.796
45.006	45.008	-0.002	49.052	-4.046	45.884	-0.878	44.765	0.241
53.638	53.641	-0.003	56.122	-2.484	54.516	-0.878	53.899	-0.261
63.194	63.192	0.001	60.443	2.751	62.316	0.878	63.120	0.074
52.094	52.097	-0.003	54.204	-2.110	52.972	-0.878	52.596	-0.502
70.331	70.332	0.000	68.856	1.476	69.549	0.783	70.186	0.145
74.520	70.523	3.998	71.050	3.470	73.643	0.877	73.908	0.612
49.254	49.255	-0.001	54.106	-4.851	50.133	-0.878	49.350	-0.095
63.593	63.592	0.001	62.455	1.139	62.716	0.878	63.107	0.486

				Unseen data: Inside				
50.038	54.228	-4.190	53.072	-3.033	51.534	-1.496	46.015	4.023
60.440	64.592	-4.152	59.677	0.763	63.164	-2.724	63.261	-2.820
63.974	73.102	-9.129	64.787	-0.814	68.400	-4.427	69.373	-5.400
68.657	70.749	-2.092	68.649	0.008	71.955	-3.298	69.239	-0.582
74.042	73.612	0.430	73.198	0.844	77.950	-3.908	78.606	-4.563
36.648	49.886	-13.238	43.818	-7.170	44.448	-7.800	45.665	-9.017
				Unseen data: Outside				
39.467	59.004	-19.538	46.915	-7.449	63.773	-24.307	64.934	-25.467
77.989	67.902	10.087	77.850	0.139	67.082	10.907	65.067	12.922
84.800	67.743	17.057	79.719	5.081	67.613	17.187	66.389	18.411
50.038	54.228	-4.190	53.072	-3.033	51.534	-1.496	46.015	4.023

APPENDIX G: COAL AUTHORITY DATA

Table 1. Reactor 1 (no packing material) manufacturing data provided by the Coal Authority. Yellow highlighted cells indicate sensor malfunction.

Date	Week of Trial	Input						Output	
		Total Fe (mg/l)	TSS (mg/l)	pH	DO (mg/l)	Conductivity (uS/cm)	Flowrate (l/s)	Total Fe (mg/l)	Fe% Removed (%)
25/05/2018	0	7.5	25	7.3	-	2210	1.005	10	-33.33333
29/05/2018	1	5.28	43	7.2	-	2380	1.005	6.16	-16.66667
01/06/2018	1	10.6	24	7.4	-	2360	1.089	10.7	-0.943396
05/06/2018	2	12.6	28	7.4	-	1990	1.054	15.1	-19.84127
08/06/2018	2	12	31	7.6	-	2580	1.010	13.9	-15.83333
12/06/2018	3	11.3	32	7.4	-	2520	1.028	12.9	-14.15929
15/06/2018	3	11.6	23	7.5	-	2600	1.028	11.6	0
19/06/2018	4	10.2	30	7.4	-	2560	1.028	12.6	-23.52941
22/06/2018	4	8.61	30	7.4	-	2560	1.028	10.3	-19.62834
26/06/2018	5	12	29	7.2	-	2530	1.028	10.5	12.5
29/06/2018	5	9.46	64	7.5	-	2520	1.028	10.4	-9.936575
03/07/2016	6	11	26	7.4	-	2550	1.028	10.7	2.7272727
06/07/2018	6	7.81	22	7.5	-	2560	1.028	6.87	12.035851
10/07/2018	7	12.7	31	7.3	-	2540	1.028	9.78	22.992126
13/07/2018	7	9.03	31	7.5	-	2480	1.028	9.27	-2.657807
17/07/2018	8	11.4	58	7.5	-	2520	1.028	11.3	0.877193
20/07/2018	8	12.6	32	7.5	-	2490	1.028	12.5	0.7936508

24/07/2018	9	12.4	32	7.5	-	2380	1.028	12.8	-3.225806
27/07/2018	9	12.7	23	8.3	-	2310	1.028	12.9	-1.574803
31/07/2018	10	12.3	39	7.9	-	2380	1.028	12.4	-0.813008
03/08/2018	10	12.5	37	7.6	-	2470	1.028	13	-4
07/08/2018	11	12	26	7.6	-	2570	1.028	12.4	-3.333333
10/08/2018	11	11.9	28	7.5	1.028	2440	1.028	12.2	-2.521008
14/08/2018	12	11.7	32	7.4	0.6	2580	1.028	12	-2.564103
17/08/2018	12	12.5	31	8	0.6	2560	1.028	12.1	3.2
21/08/2018	13	12.6	24	7.7	0.6	2570	0.600	12.2	3.1746032
24/08/2018	13	12	37	7.6	0.6	2560	0.600	12.9	-7.5
28/08/2018	14	12.5	28	7.7	0.6	2540	0.600	14.8	-18.4
31/08/2018	14	10.9	35	7.5	0.6	2530	0.600	12.5	-14.6789
04/09/2018	15	12.1	27	7.5	0.6	2540	0.600	11.5	4.9586777
07/09/2018	15	11.2	39	8.4	0.6	2600	0.600	11.9	-6.25
11/09/2018	16	13	29	7.7	0.5	2560	0.600	13.6	-4.615385
14/09/2018	16	11.7	31	8.2	0.5	2580	0.600	11.2	4.2735043
18/09/2018	17	13	27	7.4	0.5	2510	0.500	13.1	-0.769231
21/09/2018	17	12.3	31	7.6	0.5	2030	0.500	11.8	4.0650407
25/09/2018	18	11.8	20	7.5	0.45	2620	0.500	11.2	5.0847458
28/09/2018	18	12.4	30	7.6	0.45	2590	0.500	12	3.2258065
02/10/2018	19	11.5	26	7.3	0.4	2620	0.450	11.5	0
05/10/2018	19	12.3	32	7.4	0.4	2530	0.450	12.1	1.6260163
09/10/2018	20	12.1	26	7.7	0.4	2530	0.400	12.9	-6.61157
12/10/2018	20	10.4	27	8.2	0.4	2560	0.400	11.5	-10.57692
16/10/2018	21	12.2	25	7.6	0.4	2550	0.400	11.8	3.2786885

19/10/2018	21	11.1	31	7.4	0.4	2560	0.400	11.4	-2.702703
23/10/2018	22	NO SAMPLES TAKEN - DESLUDGE DAY							
26/10/2018	22	-	26	7.4	0.4	2500	0.400	-	-
30/10/2018	23	12.3	35	7.4	0.4	2490	0.400	11.4	7.3170732
02/11/2018	23	12.7	29	7.4	0.4	2570	0.400	12	5.511811
06/11/2018	24	11.4	28	7.4	0.4	2610	0.400	10.6	7.0175439
09/11/2018	24	12.2	36	7.7	0.4	2410	0.400	12.6	-3.278689
13/11/2018	25	11.9	47	7.8	0.4	2490	0.400	11.6	2.5210084
16/11/2018	25	12.5	31	7.5	0.4	2170	0.400	11.8	5.6
20/11/2018	26	12.1	23	7.5	0.4	2130	0.400	11.3	6.6115702
23/11/2018	26	12.3	23	7.6	0.9	2460	0.400	11.6	5.6910569
27/11/2018	27	12.4	30	7.5	0.9	2300	0.400	12.4	0
30/11/2018	27	11.9	29	8.5	0.9	2330	0.900	11.4	4.2016807
04/12/2018	28	11.4	28	7.4	0.6	2200	0.900	11.3	0.877193
07/12/2018	28	13.4	29	7.6	0.9	2600	0.900	13	2.9850746
11/12/2018	29	12.3	25	8.3	1	2540	0.600	12.5	-1.626016
14/12/2018	29	12.8	35	7.8	0.9	2450	0.900	12.6	1.5625
18/12/2018	30	14.9	27	7.6	-	2580	1.000	75.9	-409.396
21/12/2018	30	14.5	26	8	-	2710	0.900	13.9	4.137931
25/12/2018	-	13.6	-	-	-	-	-	12.9	5.1470588
28/12/2018	-	-	-	-	-	-	-	-	-
02/01/2019	32	-	31	7.6	0.8	2710	0.800	-	-
04/01/2019	32	12.2	27	7.6	0.85	2730	0.800	11.7	4.0983607
08/01/2019	33	12.6	22	7.5	0.8	2600	0.800	12.4	1.5873016
11/01/2019	33	12.2	27	7.7	0.8	2750	0.850	11.9	2.4590164

15/01/2019	34	11.7	47	7.6	0.8	2700	0.800	11.9	-1.709402
17/01/2019	34	11.5	39	7.6	0.75	2690	0.800	11.7	-1.73913
22/01/2019	35	11.8	23	7.5	0.7	2660	0.800	12	-1.694915
25/01/2019	35	11	24	7.5	0.5	2670	0.750	10.6	3.6363636
29/01/2019	36	11.4	29	8.1	0.7	2680	0.700	11.2	1.754386
04/02/2019	37	10.9	44	7.5	0.7	2680	0.500	10.8	0.9174312
11/02/2019	38	11.9	24	7.9	0.6125	2710	0.700	11	7.5630252
19/02/2019	39	9.01	25	7.5	-	2700	0.700	9.88	-9.655938
26/02/2019	40	10.5	41	7.8	-	2710	0.613	10.3	1.9047619

Table 2 Reactor 2 (down-flow) manufacturing data provided by the Coal Authority. Yellow highlighted cells indicate sensor malfunction.

Date	**Week of Trial**	**Input**						**Output**	
		Total Fe (mg/l)	TSS (mg/l)	pH	DO (mg/l)	Conductivity (uS/cm)	Flowrate (l/s)	Total Fe (mg/l)	Fe% Removed (%)
25/05/2018	0	7.97	25	7.3	-	2180	1.062	10.2	-27.9799
29/05/2018	1	6.99	40	7.2	-	2350	1.062	2.68	61.65951
01/06/2018	1	11.9	24	7.3	-	2350	0.931	11.7	1.680672
05/06/2018	2	11.4	26	7.5	-	2020	0.996	14.2	-24.5614
08/06/2018	2	13.5	35	7.5	-	2590	1.004	13	3.703704
12/06/2018	3	12.97	32	7.4	-	2530	1.068	12.8	1.310717
15/06/2018	3	10.3	25	7.5	-	2600	1.068	10.5	-1.94175
19/06/2018	4	11.8	32	7.5	-	2550	1.068	13.4	-13.5593
22/06/2018	4	9	25	7.4	-	2570	1.068	11.8	-31.1111
26/06/2018	5	11	31	7.3	-	2540	1.068	9.31	15.36364

29/06/2018	5	10.9	28	7.5	-	2390	1.068	8.68	20.36697
03/07/2016	6	9.65	28	7.4	-	2560	1.068	7.86	18.54922
06/07/2018	6	8.39	28	7.5	-	2560	1.068	6.93	17.40167
10/07/2018	7	11.2	32	7.4	-	2550	1.068	7.7	31.25
13/07/2018	7	14	38	7.5	-	2440	1.068	9.01	35.64286
17/07/2018	8	11.2	31	7.5	-	2530	1.068	8.59	23.30357
20/07/2018	8	12.7	32	7.5	-	2500	1.068	9.14	28.0315
24/07/2018	9	12.9	30	7.5	-	2300	1.068	9.78	24.18605
27/07/2018	9	12.2	23	8.3	-	2310	1.068	8.64	29.18033
31/07/2018	10	12.3	39	7.9	-	2380	1.068	18	-46.3415
03/08/2018	10	12.5	37	7.6	-	2470	1.068	10.9	12.8
07/08/2018	11	12	26	7.6	-	2570	1.068	8.99	25.08333
10/08/2018	11	11.9	28	7.5	1.028	2440	1.068	8.98	24.53782
14/08/2018	12	11.7	32	7.4	0.6	2580	1.068	9.32	20.34188
17/08/2018	12	12.5	31	8	0.6	2560	1.068	8.69	30.48
21/08/2018	13	12.6	24	7.7	0.6	2570	0.620	8.85	29.7619
24/08/2018	13	12	37	7.6	0.6	2560	0.620	9.94	17.16667
28/08/2018	14	12.5	28	7.7	0.6	2540	0.600	9.38	24.96
31/08/2018	14	10.9	35	7.5	0.6	2530	0.620	10.1	7.33945
04/09/2018	15	12.1	27	7.5	0.6	2540	0.600	8.77	27.52066
07/09/2018	15	11.2	39	8.4	0.6	2600	0.600	10	10.71429
11/09/2018	16	13	29	7.7	0.5	2560	0.600	10.4	20
14/09/2018	16	11.7	31	8.2	0.5	2580	0.600	9.12	22.05128
18/09/2018	17	13	27	7.4	0.5	2510	0.500	10.7	17.69231
21/09/2018	17	12.3	31	7.6	0.5	2030	0.500	8.25	32.92683
25/09/2018	18	11.8	20	7.5	0.45	2620	0.500	9.99	15.33898
28/09/2018	18	12.4	30	7.6	0.45	2590	0.500	9.57	22.82258
02/10/2018	19	11.5	26	7.3	0.4	2620	0.450	10.6	7.826087
05/10/2018	19	12.3	32	7.4	0.4	2530	0.450	9.57	22.19512
09/10/2018	20	12.1	26	7.7	0.4	2530	0.400	10.9	9.917355
12/10/2018	20	10.4	27	8.2	0.4	2560	0.400	9.9	4.807692
16/10/2018	21	12.2	25	7.6	0.4	2550	0.400	9.36	23.27869

19/10/2018	21	11.1	31	7.4	0.4	2560	0.400	9.53	14.14414
23/10/2018	22	NO SAMPLES TAKEN - DESLUDGE DAY							
26/10/2018	22	-	26	7.4	0.4	2500	0.400	-	-
30/10/2018	23	12.3	35	7.4	0.4	2490	0.400	8.65	29.6748
02/11/2018	23	12.7	29	7.4	0.4	2570	0.400	10.6	16.53543
06/11/2018	24	11.4	28	7.4	0.4	2610	0.400	8.13	28.68421
09/11/2018	24	12.2	36	7.7	0.4	2410	0.400	9.7	20.4918
13/11/2018	25	11.9	47	7.8	0.4	2490	0.400	8.47	28.82353
16/11/2018	25	12.5	31	7.5	0.4	2170	0.400	8.67	30.64
20/11/2018	26	12.1	23	7.5	0.4	2130	0.400	8.5	29.75207
23/11/2018	26	12.3	23	7.6	0.9	2460	0.400	10.7	13.00813
27/11/2018	27	12.4	30	7.5	0.9	2300	0.400	8.77	29.27419
30/11/2018	27	11.9	29	8.5	0.9	2330	0.900	12.6	-5.88235
04/12/2018	28	11.4	28	7.4	0.6	2200	0.900	10.8	5.263158
07/12/2018	28	13.4	29	7.6	0.9	2600	0.900	12.5	6.716418
11/12/2018	29	12.3	25	8.3	1	2540	0.800	11	10.56911
14/12/2018	29	12.8	35	7.8	0.9	2450	0.900	9.64	24.6875
18/12/2018	30	14.9	27	7.6	-	2580	1.000	79.6	-434.228
21/12/2018	30	14.5	26	8	-	2710	0.900	11.9	17.93103
25/12/2018	-	13.6	-	-	-	-	-	10.4	23.52941
28/12/2018	-	-	-	-	-	-	-	-	-
02/01/2019	32	-	31	7.6	0.8	2710	0.800	-	-
04/01/2019	32	12.2	27	7.6	0.85	2730	0.800	9.97	18.27869
08/01/2019	33	12.6	22	7.5	0.8	2600	0.800	8.97	28.80952
11/01/2019	33	12.2	27	7.7	0.8	2750	0.850	12.8	-4.91803
15/01/2019	34	11.7	47	7.6	0.8	2700	0.800	11.2	4.273504
17/01/2019	34	11.5	39	7.6	0.75	2690	0.800	10.3	10.43478
22/01/2019	35	11.8	23	7.5	0.7	2660	0.800	10.8	8.474576
25/01/2019	35	11	24	7.5	0.5	2670	0.750	9.82	10.72727
29/01/2019	36	11.4	29	8.1	0.7	2680	0.700	9.56	16.14035
04/02/2019	37	10.9	44	7.5	0.7	2680	0.530	9.89	9.266055
11/02/2019	38	11.9	24	7.9	0.6125	2710	0.600	9.74	18.15126

19/02/2019	39	9.01	25	7.5	-	2700	0.550	9	0.110988
26/02/2019	40	10.5	41	7.8	-	2710	0.600	8.9	15.2381

Table 3 Reactor 3 (up-flow) manufacturing data provided by the Coal Authority. Yellow highlighted cells indicate sensor malfunction.

Date	**Week of Trial**	**Input**						**Output**	
		Total Fe (mg/l)	TSS (mg/l)	pH	DO (mg/l)	Conductivity (uS/cm)	Flowrate (l/s)	Total Fe (mg/l)	Fe% Removed (%)
25/05/2018	0	9.7	26	7.3	-	2160	1.090	7.53	22.37113
29/05/2018	1	5.8	47	7.2	-	2300	1.090	3.61	37.75862
01/06/2018	1	10.5	22	7.3	-	2340	1.042	6.95	33.80952
05/06/2018	2	13.1	27	7.4	-	2030	1.008	8	38.9313
08/06/2018	2	12.7	23	7.5	-	2600	0.982	4.34	65.82677
12/06/2018	3	12.57	34	7.4	-	2520	1.002	8.53	32.14002
15/06/2018	3	10.2	21	7.5	-	2590	1.002	6.52	36.07843
19/06/2018	4	12	31	7.3	-	2560	1.002	8.69	27.58333
22/06/2018	4	10.7	26	7.4	-	2560	1.002	7.17	32.99065
26/06/2018	5	12.2	29	7.2	-	2550	1.002	6.2	49.18033
29/06/2018	5	10.1	27	7.5	-	2510	1.002	7.69	23.86139

03/07/2016	6	10.1	27	7.4	-	2560	1.002	3.76	62.77228
06/07/2018	6	10.1	26	7.4	-	2560	1.002	4.42	56.23762
10/07/2018	7	11	33	7.3	-	2530	1.002	5.91	46.27273
13/07/2018	7	14	38	7.4	-	2450	1.002	7.98	43
17/07/2018	8	11.1	28	7.4	-	2520	1.002	9.66	12.97297
20/07/2018	8	11.9	31	7.4	-	2470	1.002	10.2	14.28571
24/07/2018	9	12	33	7.4	-	2280	1.002	10.1	15.83333
27/07/2018	9	12.7	23	8.3	-	2310	1.002	9.76	23.14961
31/07/2018	10	12.3	39	7.9	-	2380	1.002	9.35	23.98374
03/08/2018	10	12.5	37	7.6	-	2470	1.002	9.55	23.6
07/08/2018	11	12	26	7.6	-	2570	1.002	9.45	21.25
10/08/2018	11	11.9	28	7.5	1.028	2440	1.002	8.54	28.23529
14/08/2018	12	11.7	32	7.4	0.6	2580	1.002	8.56	26.83761
17/08/2018	12	12.5	31	8	0.6	2560	1.002	9.5	24
21/08/2018	13	12.6	24	7.7	0.6	2570	0.600	9.04	28.25397
24/08/2018	13	12	37	7.6	0.6	2560	0.600	9.55	20.41667
28/08/2018	14	12.5	28	7.7	0.6	2540	0.600	9.17	26.64
31/08/2018	14	10.9	35	7.5	0.6	2530	0.600	9.95	8.715596
04/09/2018	15	12.1	27	7.5	0.6	2540	0.600	9.11	24.71074
07/09/2018	15	11.2	39	8.4	0.6	2600	0.600	8.77	21.69643
11/09/2018	16	13	29	7.7	0.5	2560	0.600	9.99	23.15385
14/09/2018	16	11.7	31	8.2	0.5	2580	0.600	8.01	31.53846
18/09/2018	17	13	27	7.4	0.5	2510	0.500	10.7	17.69231
21/09/2018	17	12.3	31	7.6	0.5	2030	0.500	8.86	27.96748
25/09/2018	18	11.8	20	7.5	0.45	2620	0.500	8.67	26.52542

28/09/2018	18	12.4	30	7.6	0.45	2590	0.500	8.69	29.91935
02/10/2018	19	11.5	26	7.3	0.4	2620	0.450	8.15	29.13043
05/10/2018	19	12.3	32	7.4	0.4	2530	0.450	8.39	31.78862
09/10/2018	20	12.1	26	7.7	0.4	2530	0.400	9.88	18.34711
12/10/2018	20	10.4	27	8.2	0.4	2560	0.400	8.24	20.76923
16/10/2018	21	12.2	25	7.6	0.4	2550	0.400	8.7	28.68852
19/10/2018	21	11.1	31	7.4	0.4	2560	0.400	7.97	28.1982
23/10/2018	22	NO SAMPLES TAKEN - DESLUDGE DAY							
26/10/2018	22	-	26	7.4	0.4	2500	0.400	8.96	-
30/10/2018	23	12.3	35	7.4	0.4	2490	0.400	9.98	18.86179
02/11/2018	23	12.7	29	7.4	0.4	2570	0.400	6.97	45.11811
06/11/2018	24	11.4	28	7.4	0.4	2610	0.400	7.3	35.96491
09/11/2018	24	12.2	36	7.7	0.4	2410	0.400	7.49	38.60656
13/11/2018	25	11.9	47	7.8	0.4	2490	0.400	5.85	50.84034
16/11/2018	25	12.5	31	7.5	0.4	2170	0.400	5.14	58.88
20/11/2018	26	12.1	23	7.5	0.4	2130	0.400	4.96	59.00826
23/11/2018	26	12.3	23	7.6	0.9	2460	0.400	4.66	62.11382
27/11/2018	27	12.4	30	7.5	0.9	2300	0.400	4.95	60.08065
30/11/2018	27	11.9	29	8.5	0.9	2330	0.900	6.99	41.2605
04/12/2018	28	11.4	28	7.4	0.6	2200	0.900	9.07	20.4386
07/12/2018	28	13.4	29	7.6	0.9	2600	0.900	7.66	42.83582
11/12/2018	29	12.3	25	8.3	1	2540	0.800	10.2	17.07317
14/12/2018	29	12.8	35	7.8	0.9	2450	0.900	9.76	23.75
18/12/2018	30	14.9	27	7.6	-	2580	1.000	39.8	-167.114

21/12/2018	30	14.5	26	8	-	2710	0.900	11.6	20
25/12/2018	-	13.6	-	-	0.8	-	-	10.8	20.58824
28/12/2018	-	-	-	-	0.8	-	-	-	-
02/01/2019	32	-	31	7.6	0.8	2710	0.800	-	-
04/01/2019	32	12.2	27	7.6	0.85	2730	1.000	8.54	30
08/01/2019	33	12.6	22	7.5	0.8	2600	1.000	8.88	29.52381
11/01/2019	33	12.2	27	7.7	0.8	2750	1.000	8.82	27.70492
15/01/2019	34	11.7	47	7.6	0.8	2700	0.800	8.63	26.23932
17/01/2019	34	11.5	39	7.6	0.75	2690	0.800	8.78	23.65217
22/01/2019	35	11.8	23	7.5	0.7	2660	0.800	8.62	26.94915
25/01/2019	35	11	24	7.5	0.5	2670	0.730	8.35	24.09091
29/01/2019	36	11.4	29	8.1	0.7	2680	0.700	9.38	17.7193
04/02/2019	37	10.9	44	7.5	0.7	2680	1.000	9.15	16.05505
11/02/2019	38	11.9	24	7.9	0.6125	2710	0.750	9.72	18.31933
19/02/2019	39	9.01	25	7.5	-	2700	0.570	7.78	13.6515
02/2019	40	10.5	41	7.8	-	2710	0.450	6.02	42.66667