An Application of Gaussian Processes for Analysis in Chemical Engineering

Brown Group

Author: Aaron Steven Yeardley

PhD Thesis

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Chemical & Biological Engineering

University of Sheffield

United Kingdom

2nd March 2023

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's guidance on the Use of Unfair means (https://www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously presented for an award at this, or any other, university. Part of this work is published in the following papers:

1. (Yeardley et al., 2020a) Yeardley, A. S., Bugryniec, P. J., Milton, R. A., & Brown, S. F. (2020). A study of the thermal runaway of lithium-ion batteries: A Gaussian Process based global sensitivity analysis. Journal of Power Sources, 456 (February), 228001. https://doi.org/10.1016/j.jpowsour.2-020.228001

2. (Yeardley et al., 2021) Yeardley, A. S., Bellinghausen, S., Milton, R. A., Litster, J. D., & Brown, S. F. (2021). Efficient global sensitivity-based model calibration of a high-shear wet granulation process. Chemical Engineering Science, 238, 116569. https://doi.org/10.1016/j.ces.2021.116569

3. (Yeardley et al., 2022b) Yeardley, A. S., Milton, R. A., Moghadam, P. Z., Cordiner, J., & Brown, S. F. (2022). Active subsets as a tool for structural characterisation and selection of metal-organic frameworks. Chemical Engineering Research and Design, 179, 424–434. https://doi.org/10.1016/-J.CHERD.2022.01.045

4. Yeardley, A. S., Roberts, D., Milton, R., & Brown, S. F. (2022). Robust Probabilistic Electricity Price Forecasting Using a Hybridisation of Gaussian Processes and Clustering [Manuscript submitted for publication]. Department of Chemical and Biological Engineering, University of Sheffield.

5. Yeardley, A. S., Ejeh, J. O., Allen, L., Cordiner, J., & Brown, S. F. (2022). Integrating Machine

Learning techniques into Optimal Maintenance Scheduling [Manuscript submitted for publication]. Department of Chemical and Biological Engineering, University of Sheffield.

6. (Dewulf et al., 2021) Dewulf, L., Chiacchia, M., Yeardley, A. S., Milton, R. A., Brown, S. F., & Patwardhan, S. V. (2021). Designing bioinspired green nanosilicas using statistical and machine learning approaches. Molecular Systems Design and Engineering, 6(4), 293–307. https://doi.org/10.1039/d0me00167h

7. (Yeardley et al., 2020b) Yeardley, A. S., Roberts, D., Milton, R., & Brown, S. F. (2020). An Efficient Hybridization of Gaussian Processes and Clustering for Electricity Price Forecasting. 30th European Symposium on Computer Aided Process Engineering. Elsevier B.V.

8. (Yeardley et al., 2022a) Yeardley, A.S., Ejeh, J.O., Allen, L., Brown, S.F., Cordiner, J., 2022a. Predictive Maintenance in the Digital Era, in: 32nd European Symposium on Computer Aided Process Engineering, Elsevier B.V

9. (Bugryniec et al., 2022) Bugryniec, P.J., Yeardley, A.S., Jain, A., Price, N., Vernuccio, S., Brown, S.F., 2022. Gaussian-Process based inference of electrolyte decomposition reaction networks in Li-ion battery failure, in: 32nd European Symposium on Computer Aided Process Engineering, Elsevier B.V.

10. (Biggins et al., 2022) Biggins, F.A.V., Ejeh, J.O., Roberts, R., Yeardley, A.S., Brown, S.F., 2022. Optimising a wind farm with energy storage considering curtailment and uncertainties, in: 32nd European Symposium on Computer Aided Process Engineering, Elsevier B.V.

Abstract

Industry 4.0 is transforming the chemical engineering industry. With it, machine learning (ML) is exploding, and a large variety of complex algorithms are being developed. One particularly popular ML algorithm is the Gaussian Process (GP), which is a full probabilistic, non-parametric, Bayesian model. As a blackbox function, the GP encapsulates an engineering system in a cheaper framework known as a surrogate model. GP surrogate models can be confidently used to investigate chemical engineering scenarios. The research conducted in this thesis explores the application of GPs to case studies in chemical engineering.

In many chemical engineering scenarios, it is critical to understand how input uncertainty impacts an important output. A sensitivity analysis does this by characterising the input-output relationship of a system. ML encapsulates a large system into a cheaper framework, enabling a Global Sensitivity Analysis (GSA) to be conducted. The GSA considers the model behaviour over the entire range of inputs and outputs. The Sobol' indices are recognised as the benchmark GSA method. To achieve a satisfactory precision level, the variance-based decomposition method requires a significant computational burden. Thus, one exciting application of GPs is to reduce the number of model evaluations required and efficiently calculate the Sobol' indices for large GSA studies.

The first three case studies used GPs to perform GSA's in chemical engineering. The first examined the effects of thermal runaway (TR) abuse on lithium-ion batteries. To calculate time-dependent Sobol' indices, this study created an accurate surrogate model by training individual GPs at each time step. This work used GPs to help develop a complex chemical engineering simulation model. The second GSA calibrated a high-shear wet granulation model using experimental data. This work developed a methodology, linking two GSA studies, to substantially reduce the experimental effort required for model-driven

design and scale-up of model processes. This enabled the creation of a targeted experimental design that reduced the experimental effort by 42%. The third case study developed a novel reduced order model (ROM) for predicting gaseous uptake of metal-organic framework (MOF) structures using GPs. Based on previous GSA research, the Active Subspaces were located using the Sobol' indices of each pore property for the MOF structures. The novel ROM was shown to be a viable tool to identify the top-performing MOF structures showing its potential to be a universal MOF exploration model.

The final two case studies applied GPs as a tool in novel techniques that combined ML algorithms. First, GPs are seldom used for mid-term electricity price forecasting because of their inaccuracy when extrapolating data. This research aimed to improve GP prediction accuracy by developing a GP-based clustering hybridisation method. The proposed hybridisation method outperformed other GP-based forecasting techniques, as demonstrated by the Diebold-Mariano hypothesis test. In the final case study, ML models were used to develop an effective maintenance strategy. The work compares ML algorithms for predictive maintenance and maintenance time estimation on a cyber-physical process plant to find the best for the maintenance workflow. The best algorithms for this case study were the Quadratic Discriminant Analysis model and the GP. The overall plant maintenance costs were found to be reduced by combining predictive maintenance with maintenance time estimation into a workflow. This research could help improve maintenance tasks in Industry 4.0.

This thesis focused on using GPs to enhance collaborative efforts and demonstrate the enormous impact that ML can have in both research and industry. By proposing several novel ideas and applications, it is shown that GPs can be an efficient and effective tool for the analysis of chemical engineering systems.

Keywords: Gaussian Process, Machine Learning, Global Sensitivity Analysis, Forecasting, Regression Model, Chemical Engineering

Acknowledgements

First of all, I would like to thank both my supervisors, Dr Solomon Brown and Professor Joan Cordiner, who have been invaluable throughout my PhD in both work and life. Sol your guidance from the very beginning has been incredible, and it is you who I must thank for my interest in machine learning, without this, I think I would still be looking for the path I want to go down. Joan, your compassion and advice will always be appreciated, you have opened the door to so many opportunities and no doubt you will continue to help me, always understanding the work-life balance when choosing a career. So thank you both, and I hope we stay in contact as friends.

I would also like to thank Dr Robert Milton, my python guru, who without I would probably still be writing "Hello World". Truly, **Rob Milton** (*what a guy!*).

I would like to thank everyone in The Brown Group, making us the best research group at the University of Sheffield. And finally, I would also like to thank all the lecturers I have worked for as a GTA during my PhD. Teaching, was genuinely an enjoyable experience that allowed me to switch off from my research and earn extra money. Without, I probably wouldn't have survived.

Contents

List of Figures

List of Tables

Glossary

Abbreviation	Description	Page
AI	artificial intelligence	1
ARD	automatic relevance determination	13
GP	Gaussian Process	iv, v, 3–14, 17–23, 25–28, 30–37
GSA	Global Sensitivity Analysis	iv, v, 2, 3, 10–15, 17–19, 21, 33–35
ML	machine learning	iv, v, 1–8, 14, 17, 21–25, 32, 33, 35–37
MOF	metal-organic framework	v, 19–21, 34, 35
ROM	reduced order model	v, 13, 14, 19–21, 35
TR	thermal runaway	iv, 16, 17, 33, 34

Chapter 1

Introduction

1.1 The Potential Machine Learning has for Chemical Engineering

A subfield of artificial intelligence (AI) is machine learning (ML), this is where useful information is extracted from data using sophisticated mathematical techniques. The popularity of AI for engineering has risen significantly due to improved computational power, increased data storage capabilities and new sensor technologies. All of this has driven a sudden and explosive rise in ML, causing the continuous development of a huge selection of complex algorithms. This thesis focuses on the potential applications of ML within the field of chemical engineering. The ML algorithms can be categorised into three main types (Murphy, 2012):

1. Supervised learning - uses training data to learn a relationship between inputs and outputs:
   (a) Classification problem - assign an input to a discrete category.
   (b) Regression problem - the output consists of continuous variables.
2. Unsupervised learning - find patterns in the given inputs where no training data is provided.
3. Reinforcement learning - find suitable actions to take in a given situation to maximize a reward.

The chemical engineering industry is undergoing an incredible transformation many are calling the fourth industrial revolution (also known as Industry 4.0). Modern industrial plants are in the ever-growing era of "big data" as industrial equipment generates more data than ever before (Yin and Kaynak, 2015).

In 2010 a total volume of 2 Zettabytes was produced annually. In comparison, it reached 64 Zettabytes in 2020 (See, 2021). Therefore, this vast quantity of information requires effective and efficient methods of analysis. Additionally, chemical engineering involves many complex systems that are typically governed by chemistry and physics controlling the underlying system. Thus, there is a need for analysis of large, complex and real data sets. Predicting outputs using ML is a promising technique that can be used to solve analytical problems within reliability, security, safety and risk assessment (Lepikhin et al., 2018). Murphy (2012) states that ML provides an automated method of data analysis, vital for today's era of big data.

In the opposite sense, chemical engineering can go from extremely large datasets in the industry to small datasets in research due to the costs associated with generating them. Many research problems in chemical engineering are solved using experiments and/or complex model simulations which can only be conducted a limited amount of times producing a few hundred data points. Such models include expensive aerospace models (Queipo et al., 2005), process flowsheet simulations (Palmer, 2013) and pharmaceutical process models (Jia et al., 2009). However, to capture the data needed to understand and analyse the key outputs, a very large number of model runs are required to apply data science techniques. For chemical engineering research, an ideal modelling approach would be to use a small amount of data to capture the underlying science in an interpretable model which can be used for further analysis. To mitigate this, direct interrogation of the underlying model (simulation or experiment) may be replaced by a surrogate model encapsulating the behaviour in a cheaper, simpler framework. Therefore, these issues are circumvented by the use of surrogate modelling techniques where the physical model is replaced by a mathematical approximation built from a few hundred data samples. Surrogate modelling techniques use ML algorithms such as the polynomial chaos expansion (Brown et al., 2013; Sudret, 2008), artificial neural networks (Li et al., 2016), and support vector regression (Shi et al., 2020).

Surrogate models can be used as powerful tools to implement data analysis techniques such as design, optimisation, and sensitivity analysis. Chemical engineers must understand how parameters affect the output of a system in order to accurately comprehend the behaviour. This is addressed by performing a sensitivity analysis, which quantifies the impact each input variable has on the desired output. A sensitivity analysis can be conducted locally or globally. A local sensitivity analysis investigates the effect of input variables one at a time over certain outputs of interest. In comparison, a Global Sensitivity Ana-

lysis (GSA) provides information over the whole range of input values. A GSA is critical for chemical engineers because it provides a deep understanding of the system's behaviour by analysing the impact each input has on the whole system. For GSA, the most popular method (Rohmer and Foerster, 2011; Al et al., 2019) is the Sobol' method (Sobol, 2001), which decomposes the variance of the model output to calculate Sobol' sensitivity indices. Frequently, the Sobol' indices are considered the benchmark technique for comparison to new GSA methods (Chastaing and Le Gratiet, 2015; Iooss and Lemaître, 2015). Calculating Sobol' indices requires a significant amount of time; almost 10,000 model assessments are necessary to achieve a 10% precision (Lamoureux et al., 2014). Surrogate models, on the other hand, give a more efficient method of estimating multidimensional integrals (Marrel et al., 2009).

This work focuses on a specific ML method, known as Gaussian Processes (GPs), to investigate the impact its application can have as a tool on "big data" in industry and as a surrogate model in research, providing efficient GSA insights to complex models.

1.2 Contribution to Science

1.2.1 Research Aims and Questions

This research aims to **maximise the applicability of GPs for a variety of domains in chemical engineering.** To do this, the following research questions are asked:

1. Can GPs be used as an efficient machine learning technique to aid data analysis of complex chemical engineering simulations?

2. Are GPs a suitable technique to link a complex computational model to experimental data using the design of experiments?

3. Given a large quantity of measurable data from experiments, can GPs create a blackbox function that can be used to provide knowledge that otherwise is unconfirmed?

4. Due to the difficulty in extrapolation when using GPs, could they be further developed to be used as a forecasting technique?

5. Can GPs be used in combination with other types of ML techniques to provide a novel workflow that has the potential to aid the process industry?

1.2.2 Contribution and Significance of the Thesis

The methods developed in this thesis focus on the application of a GP algorithm developed by Milton and Brown (2019). Collectively, the research prioritises the collaboration of data science and chemical engineering to show the importance of just one of many machine learning methods. Therefore, the works found in this thesis extend the knowledge in different areas found within chemical engineering while developing machine learning application techniques and effective use of data analysis. The contribution of each publication is discussed further in Chapter 3.

Overall, the research published in this thesis is significant because it provides novel insights into numerous disciplines within chemical engineering. The methodology developed in each chapter builds on GPs by supporting the application of them for data analysis, such as surrogate modelling, sensitivity analysis, design of experiments and forecasting. The main significance of this research is to show how important machine learning methods can be for chemical engineering during the Fourth Industrial Revolution (Industry 4.0).

1.2.3 Thesis Outline

The thesis is organised as follows.

Chapter 1 presents the Introduction. It begins by describing the background behind ML for chemical engineering. The chapter then concludes by discussing the contribution that the research undertaken during this PhD has had on science. Afterwards, a literature review is shown in Chapter 2, providing a summary of various ML methods and their use for data analysis in chemical engineering. The literature review uses previous research to show why GPs are an exciting ML method to apply to various chemical engineering case studies. Chapter 3 provides an introduction to each publication, presenting the publication information so that the contribution from each co-author is clearly stated. Following this, the contribution each paper has on the application of GPs is described in the Paper Contribution. All the publications presented in this thesis focus on the application of GPs. Therefore, Chapter 4 contains the full mathematical background, deriving the GP predictive equations that enable machine learning. Chapter 5 concludes the thesis with a summary of the research, evaluating the overarching importance of all the research undertaken before describing future work that this research has led to. Chapter 6 to Chapter 10 present each publication in full.

Chapter 2

Literature Review

In this chapter, a literature review will explore current research using machine learning (ML) for data analysis in chemical engineering. The literature will be reviewed to explain the reasoning behind choosing Gaussian Processes (GPs) as the focus of this thesis.

2.1 Surrogate Modelling Review

Chemical engineers can make important decisions using data analysis tools such as optimization, sensitivity analysis, design of experiments, and prediction. However, these techniques are becoming increasingly difficult to implement due to the computing burden associated with modelling real-world systems (Viana et al., 2021). To address such issues, surrogate modelling is emerging as a powerful tool to enable data analysis of complex systems. A surrogate model behaves as a blackbox function that approximates the input-output relationship of model systems using mathematical techniques. Different types of ML algorithms can be used as the mathematical approximation for input-output functions.

Choosing the best surrogate modelling technique is not a simple decision to make as there is no general agreement about which technique is superior to others. Further, creating an accurate model is not just dependent on choosing a surrogate model. Important considerations should also be made with respect to generating data for training, fitting the surrogate model, and evaluating the accuracy. All must be conducted before the surrogate model is used to conduct analysis techniques. Simpson et al. (2001) provided a survey with recommendations to help researchers with data collection, choosing a surrogate modelling technique, fitting the surrogate model, and validating the surrogate model.

Kajero et al. (2017) provided a summary of the different types of surrogate modelling techniques available. Previous literature has demonstrated an interest in comparing various techniques applied to the same engineering problem (Qian et al., 2006; Jin et al., 2003; Villa-Vialaneix et al., 2012). Clearly, there are many techniques available and there is no general consensus on the "best" surrogate modelling technique to use, making the choice of surrogate model difficult. Ultimately, the choice of surrogate modelling technique must depend on the application of the system being modelled (Jin et al., 2001). Other considerations to make when choosing a surrogate modelling technique include:

- Does the system being modelled follow a general pattern? A surrogate modelling technique that can capture that pattern should be considered. For example, a polynomial function could follow a general seasonal trend.
- Would confidence intervals improve the applicability of the surrogate models? Probabilistic surrogate models will provide the decision-maker with the surrogate model's uncertainty in predictions.
- Are predictions required away from the training data? Extrapolation is difficult for many surrogate modelling techniques. If so, consider the experimental design carefully to cover a large degree of input space before focusing on a surrogate modelling technique that was designed for forecasting purposes.
- Should a trade-off between efficiency and accuracy be made? Advanced ML algorithms are notoriously data-hungry as the long-term goal has been to develop accurate algorithms. It's not always possible to make such complex surrogate models given the time or data provided.

Table 2.1 summarises the popular surrogate modelling techniques available (Bhosekar and Ierapetritou, 2018; Viana et al., 2021). The table shows the advantages and disadvantages of each technique. Clearly, the choice depends on the scenario being modelled. For example, if the system follows a general pattern with limited noise, a polynomial response surface may be the best choice of surrogate modelling technique due to its efficiency.

In summary, the choice of ML algorithm for the surrogate model is always dependent on the scenario being modelled. That being said, this thesis will focus on the application of GPs because of their math-

Table 2.1: Surrogate Modelling Techniques

Technique	Advantages	Disadvantages
Polynomial Response Surfaces	• Efficient statistical method using the simplicity of polynomials. • Coefficients objectively determined using optimisation such as least square regression.	• Does not estimate the error in predictions. • Local analysis only. • Accurate models become unfeasible for complex systems that require high-order polynomials and more input variables.
Radial Basis Function	• ML algorithm that fine tunes parameters by reducing the Euclidian distance. • Different forms of the basis function are available e.g. linear, cubic, Gaussian, thinplate spline. • Weights of the basis functions and the polynomial coefficients can be determined using efficient optimisation methods.	• The sensitivity of the inputs is typically assumed identical. • Basis function has to be specified arbitrarily.
Support Vector Machines	• Very similar to radial basis functions but only involve a subset of the design site so that errors within a certain distance from the true values can be ignored. • Can directly control and reduce the surrogate model's sensitivity to noise.	• There are two specific vector parameters and the kernel function parameters that have to be optimised and they are mutually dependent, making parameter values difficult to interpret. • Require the storage of all the support vectors which grows significantly with the training data size. Thus, support vector machines have issues with memory requirements.
Gaussian Process	• Predicts a distribution describing the uncertainty in each prediction. • It is a non-parametric ML method. The hyperparameters in the kernel function are the only parameters that need to be optimised. Therefore, learning can be conducted by maximising the marginal likelihood. • A GP is versatile because many different kernels can be used to express a prior assumption of a model.	• Computational expensive to train the surrogate mode due to the need to invert the kernel. • GPs are victims to the curse of dimensionality, becoming inefficient when the input variables increase above a dozen.
Artificial Neural Network	• Most popular because they are highly flexible as they imitate the behaviour of biological brains. • Artificial neural networks can adapt to changes and do not need to be reprogrammed.	• Computationally expensive due to the requirement of parallel processing. • Artificial neural networks can be difficult to understand the functioning of the predictions. The decision process undertaken is not interpretable.

ematical tractability, resulting in a large body of research devoted to developing GPs. For example, the GP algorithm has recently been developed, creating Sparse GPs (Snelson and Ghahramani, 2005; Titsias, 2009), the GP Latent Variable Model (Lawrence, 2005), and Deep GPs (Damianou and Lawrence, 2013). Please note that this choice of research is not because the author believes GPs are the best surrogate modelling technique. Instead, this thesis aims to use the advantages of GPs to improve their applicability as a tool for a wide range of chemical engineering scenarios.

2.2 Gaussian Process Surrogate Modelling

A GP is an important stochastic process that corresponds to a class of normal distributions on a function. GP regression uses the stochastic process to make predictions based on previous data in a ML context.

For ML, GPs are popular due to three key advantages. First, they are a full probabilistic Bayesian model, which enables the best estimate of the desired function to be complemented by a probability distribution over likely functions; uncertainty in the estimate. Another appealing feature of GPs is the large range of covariance functions available making them flexible in nature. This means the GP can express many assumptions simply by choosing a covariance function that fits the underlying model. Finally, the last advantage is that it is a non-parametric ML method. This is when the algorithm does not make prior assumptions regarding the structure of the surrogate model and is free to learn any functional form from the training data. A parametric ML algorithm learns the coefficients of a simplified function of the underlying model. This means, non-parametric algorithms are more flexible but are at risk of overfitting. Together, these advantages make GPs a rare ML method that is non-parametric but still has the choice of many covariance functions. The GP is trained by integrating over a wide range of hypotheses so that the risk of overfitting is decreased compared to other non-parametric ML algorithms. In summary, GPs are simple to use and mathematically tractable (enable the use of data analysis techniques such as sensitivity analysis). These properties of the GP make it an ideal ML technique to use as surrogate models for chemical engineering applications.

The first use of GPs for surrogate modelling was proposed by Sacks et al. (1989) and Currin et al. (1988, 1991) who described how statistical inference can be used in computer modelling for estimating simulators. Each research group developed their concept using a different statistical framework and compared the results of their methods to each other using the same electronic circuit simulator example. The

methodology from Sacks et al. (1989) took a classic Frequentists stance while Currin et al. (1988) developed their emulator using a Bayesian framework. The work by both research groups was vital for the modern use of GPs as it set the basis for open statistical problems such as optimising the hyperparameter estimations and choosing appropriate design criteria. Both research groups agreed that the maximum likelihood estimation was the most efficient technique to optimise the hyperparameters. The question of which design criterion to use is dependent on the problem itself, for example, Sacks et al. (1989) used the average mean squared error while Currin et al. (1988, 1991) used the expected reduction in the entropy of the random vector at given input configurations to decide where predictions are to be made.

Since these ideas were first proposed, GPs have been widely used in various applications such as engineering structural reliability analysis (Su et al., 2017), prediction of tidal currents (Sarkar et al., 2018) and battery health forecasting (Richardson et al., 2017). The choices to be made when developing GP surrogate models include a crucial decision on the covariance function. Then, the optimisation method which estimates the hyperparameters in the covariance function has to be made. The choice of covariance function can be difficult due to the wide range of kernels and the flexibility of combining them within the covariance function. One of the toughest challenges for all researchers in this domain is to build an accurate surrogate model that can be used for engineering and science innovations. However, several methods are reported in the literature that aid in building a GP, including a method of automated kernel construction (Duvenaud et al., 2013). Another way to choose the most appropriate covariance function is to understand the modelling problem and the properties of various kernels. Therefore, the properties of the GP are influenced by the choice or kernel so that the surrogate model's behaviour is similar to the original modelling problem.

In addition, the optimisation method for the hyperparameters varies from the popular maximum likelihood estimation method to other techniques such as using a Markov Chain Monte Carlo algorithm (Andrieu et al., 2003). Once the models have been created, the method of validation needs to be chosen. Most authors use cross-validation (Queipo et al., 2005; Rohmer and Foerster, 2011; Gratiet et al., 2016) but authors have been seen to validate their emulator using separate validation points (Ba and Joseph, 2012; McDonnell et al., 2015). Finally, the diagnostic results that shall be used within the chosen method of validation need to be decided. Most diagnostics used to validate predictions compare just the mean predictions to the observed true predictions (such as root mean squared error or correlation coefficient)

but we can also consider the uncertainty in the model's predictions, from the GPs predicted STD.

An extensive list of examples that use GP surrogate models have been created by Al-Taweel (2018) showing the resulting choices that many authors have made when developing their GP surrogate model. The majority, 82%, of said GP examples presented in the table have used the squared exponential correlation function and the most popular method of hyperparameter estimation is the maximum likelihood estimation with 64% of the surrogates using it. The method of validating the emulator is much more varied between authors. Surprisingly, 20% of authors have not validated their model. Whereas, 33% use cross-validation methods and the remaining use separate data points for validation. The extensive review from Al-Taweel (2018) shows the choice of diagnostic method to be the most varied decision when using a GP surrogate model. Research studies validate their surrogate models by calculating various diagnostics from the root mean squared error (Montagna and Tokdar, 2016) to the predictivity coefficient (Marrel et al., 2015). Additionally, residual plots of predictions against true values help show the accuracy of the surrogate model (Sarkar et al., 2016). Consequently, Al-Taweel (2018) recommended that more than one diagnostic method should be used for validating surrogate models and at least one of the methods should consider the uncertainty in the predictions.

Overall, the difficulty of building GP surrogate models has been explored vigorously in prior studies emphasising the popularity of such techniques. As noted earlier, more work is necessary when validating GP surrogate models to ensure their applicability. In short, the literature strongly suggests that mathematics has been developed in detail highlighting GPs capability as surrogate models. However, additional studies are required that use robust validation techniques before focusing on applying GP surrogate models to aid chemical engineering domains.

2.3 Sensitivity Analysis using Gaussian Processes

Variance-based sensitivity analysis by calculation of Sobol' indices (Sobol, 2001) is one of the most widely used approaches for Global Sensitivity Analysis (GSA). However, the method requires a significantly large number of model evaluations, nearly $10,000$ are needed to reach 10% precision (Lamoureux et al., 2014). As previously mentioned, GPs provide an alternative method to estimate multidimensional integrals using Monte Carlo schemes (Oakley and O'Hagan, 2004; Jin et al., 2004; Marrel et al., 2009). The mathematical tractability of GPs enables sensitivity analysis by calculating the sensitivity indices

given a training set of model evaluations. Hence, using GPs will bypass a large number of model evaluations required by Monte Carlo methods. The derivation enabling the use of GP regression for analytical evaluation of variance-based sensitivity indices were first introduced by Jin et al. (2004) who applied the Sobol' index formula directly to the GP predictors. Previously, Oakley and O'Hagan (2004) used the global stochastic model of a GP, providing the calculations to produce random variables as new sensitivity measures. This has the advantage of the analysis of the sensitivity indices accuracy due to the distribution of the variables. This work was further improved by using the distributions to introduce confidence intervals for the Sobol' indices (Marrel et al., 2009). These methods were validated initially using toy mathematical functions showing very accurate sensitivity indices and satisfactory confidence intervals from the second method. However, when the approach was illustrated on real data to provide a sensitivity analysis on radionuclide groundwater transport, it was found that the confidence intervals were inaccurate for very low indices due to the overestimation of the lowest Sobol' indices. Overall, the work from the three research groups (Oakley and O'Hagan, 2004; Jin et al., 2004; Marrel et al., 2009) provided an excellent method of using GP surrogate models for GSA. Consequently, current research now has to decide whether the GSA method should calculate the indices using just the GPs prediction means or whether to use the GPs full distribution to produce confidence intervals on the sensitivity indices.

Using GP to compute the Sobol' sensitivity indices has been used in many disciplines. For example, Rohmer and Foerster (2011) used the technique to analyse large-scale numerical landslide models. Due to the limited knowledge of the slip surface in the La Frasse landslide, the simulator was approximated using small size training data for the GP. Therefore, Rohmer and Foerster (2011) inclined to use confidence intervals on the sensitivity measures to outline regions where the GP was unsure. The work concluded by providing ideas of future work by the design of experiments, whereby in the identified unsure regions, further simulator runs should be carried out allowing the GP model to learn further and reduce the confidence intervals.

The method has also been proven to aid in engineering system safety by providing an early validation of health indicators for detection and identification in design sites (Lamoureux et al., 2014). The work used a GP to emulate a pumping unit of an aircraft fuel system. Using Latin Hypercube sampling, the GP learnt from 400 model evaluations before the Sobol' indices for the 20 inputs were computed.

On this occasion, due to a large number of model evaluations and a large number of model inputs, the authors opted to calculate the Sobol' indices alone without any confidence intervals. Interestingly, this work computed both the first-order indices and the total indices, consequently finding very close results. Therefore, this indicates that the inputs have few or no correlations and so the most influential parameter for each health indicator was found.

However, issues may occur when calculating the Sobol' indices of input parameters which are correlated because the construction of such measures relies on the assumption that inputs are independent. This is because if inputs are dependent on one another then the amount of variance due to a given parameter may be influenced by its dependence on another input (Mara and Tarantola, 2012). Therefore, a lot of work has been conducted towards dealing with dependency in sensitivity analysis studies (Xu and Gertner, 2008; Li et al., 2010; Caniou and Sudret, 2010; Chastaing et al., 2015). However, these methods concentrate solely on the GSA method, choosing not to incorporate them with a GP surrogate model. Another issue with the previous studies is that the number of decomposition components exponentially grows with the model dimension (Chastaing and Le Gratiet, 2015). Although, other interesting methods have been invented to deal with correlated inputs which can be used with GPs. For example, Chastaing and Le Gratiet (2015) extended the work done by Durrande et al. (2013) to deal with dependent inputs. In this work, the GP is specified to a special class of ANOVA kernels (Berlinet, 2004), which is functionally decomposed as a sum of processes indexed by increasing dimension input variables. The work is similar to that of Caniou and Sudret (2010) except it considers the use of GPs instead of the polynomial chaos expansion. Another method (Li and Rabitz, 2012), decomposed the output variance by using a hierarchical orthogonality condition. The result of this reduces the prior formulas to them used for independent variables. Interestingly, Srivastava et al. (2017) investigated the Li and Rabitz (2012) framework, comparing it to the original variance-based method using GPs. The research used both methods in an attempt to understand which calibration parameters can be fixed without losing output variability for an aircraft model. Consisting of 100 calibration parameters which had both correlated inputs and independent inputs, the problem had over $4,000$ two-way interactions, ensuring the complexity was high enough to not be able to calculate all the two-way interaction indices. Therefore, the work incorporated all the complexities involved in real-world applications, inferring that although separate levels of interactions between inputs can be calculated with Sobol' indices, it is not always possible to separate the sensitivities

into structural and correlation elements. However, when the alternative method of approximations was used, the computed sensitivities had inaccuracies even though the relative ordering of variables was still correct. Therefore, it is important to understand that whichever method is used, care must be taken when inputs are correlated.

Reduced-order modelling has the potential to deal with correlated input variables as it reduces the dimensions d of a d-dimensional input $\mathbf{x}$ while preserving the behaviour of the output y. The result of this reduces the inputs to a set that is both mutually independent and highly relevant to the response. The curse of dimensionality is a significant driver for dimensionality reduction methods and so several approaches have been developed. For example, Constantine et al. (2014) calculates an Active Subspace through a derivative sensitivity measure. Similarly, Liu and Guillas (2017) reduced a GP surrogate model by gradient-based kernel reduction.

An exciting and novel method of dimensionality reduction (Milton and Brown, 2022) achieves the reduced order model (ROM) by an optimal rotation of the input basis. The ROM is achieved through the utilization of a GP surrogate model that facilitates GSA via Sobol' indices Sobol (2001). Simply, the method rotates the input basis so that the emulator only significantly depends on the most relevant rotated input dimensions. For a full derivation please refer to Milton and Brown (2022).

The overall goal is to replace the input matrix $\mathbf{X}$ of size $(d \times n)$ with a rotated input basis $\mathbf{U}$ of size $(d \times n)$. When looking at a sample datum $\mathbf{u}$, it can be defined to provide another input to the emulator upon rotation by a row orthogonal matrix $\boldsymbol{\Theta}$

$$\mathbf{x} =: \mathbf{u}^T \boldsymbol{\Theta} \tag{2.1}$$

Therefore, the GP can now be represented by the rotated input. Where the mean $\bar{f}(\mathbf{x})$, learnt from the original training data, can now be conditioned by calculating the variance of $\bar{f}(\mathbf{u}^T \boldsymbol{\Theta})$ due to the first d components of $\mathbf{u}$. Hence, GSA is proceeding in the same fashion now as it would have by calculating ordinary Sobol' indices. Currently, the analytic expressions for the variance can only be expressed using the automatic relevance determination (ARD) kernel Wipf and Nagarajan (2007). The algebra deriving these expressions would be vastly complicated by any other kernel.

The ROM is now completed by maximising the Sobol' index of each rotated input in turn by entirely

optimising $\boldsymbol{\Theta}$ row by row from top to bottom. The work is done using gradient descent, repeatedly rotating the input basis by $\boldsymbol{\Theta}$ until the Sobol' indices are maximised so much that $\boldsymbol{\Theta} \approx \mathbf{I}$.

This specific method of ROM is incredibly useful for the research conducted in this project due to its derivation from GPs and GSA. The method perfectly shows how the mathematical tractability of GPs enables powerful data analysis that will benefit a broad range of engineering topics. The desire to reduce the dimensions of inputs without impacting the output is twofold: to achieve accurate descriptions of a system at a much lower computational cost and to provide a means by which a process can be visualised. For example, an industrial plant may have an output that wants to be analysed but is difficult to obtain and many input variables can be varied. So analysing the change in these variables to the change in the output can be difficult. A ROM can be used to capture most of, if not all of, the system's fundamental dynamics with a much smaller amount of dimensions to consider. This could make a significant difference as simple as being able to plot the output with respect to just one or two input dimensions instead of having to consider a large degree of variables. This method also introduces far fewer complications involving the use of GSA with dependent inputs due to it reducing the input to independent sets. The rotation of the input space represents combinations of dependent inputs and so the analysis deals with the issue of dependent inputs by simply combining them, ensuring each rotated basis is independent of one another.

2.4 Literature Summary

The purpose of this chapter was to help the reader understand the current deployment of GPs. This is significant as it provides reasoning and motivation for the research conducted during this PhD. It helps explain the benefits of using GPs and highlights the challenges involved in using a GP. It is clear from the research reviewed that GPs are an extremely popular ML technique. We have also discussed the important decisions to be made when building a GP model. Most of the research found focused on developing surrogate models and used them for predictions and GSA. More research is required to use GPs as a tool prioritising the area they are aiming to fix instead of focusing on the GP itself. It is important to conduct more studies using the ML technique for the analysis of engineering data. Such techniques have been explored in research, for example, GSA. However, the previous studies explored have developed techniques for data analysis and so more work is necessary for focusing on the problem itself and then applying a technique to solve it. Thus, it has been found that GPs have a great potential

that offers chemical engineering a solution to provide data analysis (such as GSA) to complex underlying models.

Chapter 3

Introduction to Publications

3.1 Paper 1:

A study of the thermal runaway (TR) of lithium-ion batteries: A Gaussian Process based global sensitivity analysis

3.1.1 Publication Information

The first paper is published in the Journal of Power Sources (Yeardley et al., 2020a) which has provided confirmation that the published journal article "can be posted publicly by the awarding institution with DOI link back to the formal publications on ScienceDirect".

In this publication, I, the candidate, co-wrote the manuscript with Dr Peter J. Bugryniec. Together, we conceptualized the research through Dr Bugryniec's knowledge of lithium-ion battery modelling and my knowledge of sensitivity analysis. Together, myself and Dr Robert A. Milton developed the software used for the sensitivity analysis. For the manuscript itself, I, the candidate, completed the formal analysis and co-wrote the paper with supervision from Dr Solomon F. Brown.

3.1.2 Paper Contribution

The first paper published focused on aiding Dr Bugryniec in understanding a complex computational simulation that modelled the TR of lithium-ion batteries. The computation simulated the temperature of a lithium-ion cell per unit of time given thermo-characteristic input variables.

Thus, the goal here was to understand how the input variables impacted the cell temperature throughout the time simulated. A Global Sensitivity Analysis (GSA) calculates the impact of each input variable on the output (i.e. the cell temperature), but given the temperature is computationally calculated every second, a GSA conducted using the TR model would require too many model runs.

To solve this, an accurate Gaussian Process (GP) surrogate model encapsulated the computational model in a cheaper framework enabling the calculation of time-dependent Sobol' indices for the first time. During this study, a GP was successfully trained to encapsulate the underlying behaviour of a complex chemical engineering simulation. Individual GPs had to be trained to predict the temperature at each time step. Each GP was validated and concatenated to make a full surrogate model from GPs. Therefore, it was important to ensure that the learning of the machine learning (ML) hyperparameters produced accurate GP models that predicted a smooth temperature-time distribution. This, in itself, created novel research as often techniques such as time warping (Wang and Gasser, 1997; Maiy and Sudret, 2017) are used to transform each sample simulation to its own time scale. This ensures the sudden explosion from the TR is created at the same time for every sample. However, techniques such as this improved the accuracy of the predictions but negatively impacted the GSA because the impact of the variables that caused TR to occur earlier or later was not counted due to the time warping. All this was included in the publication and led to a methodology that used individual GPs as a collection to create an accurate surrogate model for a complex lithium-ion battery model simulation. This enabled data analysis of the simulation and the creation of time-dependent Sobol' indices.

In conclusion, this research provided an introduction into the application of a GP surrogate model before novel research was conducted to provide invaluable results that increased the understanding of the effect input variables have on the output of a complex model. The methodology in this research developed a technique to successfully carry out time-dependent GSA for the first time. This led to results required to further develop a complex simulation model. The research had a lasting impact because the lithium-ion battery research group later applied GPs again. This time to optimise the parameters of a reaction network model that is of greater

complexity to the original model (Bugryniec et al., 2022).

3.2 Paper 2:

Efficient global sensitivity-based model calibration of a high-shear wet granulation process

3.2.1 Publication Information

The second paper is published in Chemical Engineering Science (Yeardley et al., 2021) which have provided confirmation that the published journal article "can be posted publicly by the awarding institution with DOI link back to the formal publications on ScienceDirect".

In this publication, I, the candidate, co-wrote the manuscript with Dr Stefan Bellinghausen. Together, we conceptualized this research as Dr Bellinghausen's model for wet granulation Bellinghausen et al. (2022) was challenged with minimising the experimental effort for calibration. As such a novel method of model-driven design for scale up processes was required. In this research, the software developed by myself and Dr Robert A. Milton for the previously published paper was used to conduct a sensitivity analysis. Together, myself and Dr Bellinghausen formally analysed the GSA results to understand how it impacted the wet granulation model and linked the results to the experimental data. This manuscript was co-supervised by Dr James D. Litster and Dr Solomon F. Brown.

3.2.2 Paper Contribution

Within this paper, the collaboration focused on producing a more efficient design of experiments for a wet granulation simulation model. Previously, Stefan Bellinghausen created a high-shear wet granulation process which we used as a case study to help develop a novel methodology for model-driven design and scale up of the wet granulation process.

Similar to the previous research, a GP was used to create a surrogate model to interrogate the wet granulation simulation model so that the Sobol' indices could be calculated. The novelty in this work was from processing the results produced by the GSA which provided an insight into the critical process parameters. Here, the application of GPs enabled the proposal

of a model-driven design approach that consisted of an efficient experimental design and model calibration workflow. This was proven to improve the ability and efficiency of determining modelling parameter values. Results showed that only four out of twenty modelling parameters required estimated values to produce a generically applicable workflow. These findings made the subsequent model calibration possible considering the experimental data available. Therefore the research conducted in this manuscript successfully created a novel methodology using GPs and the design of experiments to optimise the model calibration workflow for a complex computational model with experimental data.

3.3 Paper 3:

Gaussian Process Identification of Active Subsets as a Tool for Structural Characterisation and Selection of Metal-Organic Frameworks

3.3.1 Publication Information

The third paper published in Chemical Engineering Research and Design (Yeardley et al., 2022b) has provided confirmation that the published journal article "can be posted publicly by the awarding institution with DOI link back to the formal publications on ScienceDirect".

In this publication, I, the candidate, wrote the manuscript with supervision from Dr Peyman Z. Moghadam, Professor Joan Cordiner and Dr Solomon F. Brown. Together, myself and Dr Robert A. Milton developed the software used for the initial sensitivity analysis. In this research, Dr Milton further developed the sensitivity analysis technique used in the previous research papers creating a novel reduced order model (ROM). I, the candidate, began tests on the ROM technique with Dr Moghadam's metal-organic framework (MOF) data (Moghadam et al., 2018), formally analysing the results from a GSA to understand where best to implement the ROM technique. The work, as shown in Chapter 8, describes a robust investigation into using GPs to identify the Active Subsets within the MOF database. Thus, I, the candidate, was the first person to test Dr Milton's ROM technique on real world data.

3.3.2 Paper Contribution

The third paper focused on using a novel ROM technique that has yet to be published. The technique locates Active Subspaces using GPs to calculate Sobol' indices. In this manuscript, the priority was to apply the ROM to predict the deliverable capacity of synthesised MOFs to help identify top-performing structures. MOFs are a unique class of polymers that have tailorable properties allowing an incredibly large number of potential structures. Therefore, the work conducted was able to apply a novel methodology to a very prominent field of study. By initially comparing the importance of pore properties on MOF structures deliverable capacities with existing research, the MOF database presented a case study that permitted an in-depth application of the ROM technique.

Ultimately, the novel methodology was proven to be invaluable in creating a technique that has the potential to build a universal MOF exploration model. A universal model that can predict a MOFs uptake of any gas would be extremely beneficial for scientists who design, synthesise and use MOFs. The work conducted in this research made a large initial contribution to this universal model by first providing a valuable understanding of the most important pore properties with respect to gas storage as the conditions change. Knowing this, we built a reversible model using Dr Milton's ROM code that could successfully predict the deliverable gas of a MOF storing oxygen when the data used for training were storing methane. The potential of such a model was shown by identifying the top-performing MOF structures for oxygen storage from a much larger database that does not have the deliverable capacity computationally measured. Results showed the ROM trained using oxygen data identified the same MOF structures as the ROM trained using methane data. Developing a MOF prediction technique that is successful on two similar storage gases is an exciting development in the goal of building a universal MOF exploration model.

This research has a large scope for further work as the overall goal for a universal model has only just started to be addressed by using two storage gases. However, the contributions from this research are significant as shown by both the implications it has to enhance MOF re-

search by accelerating the identification of better performing materials. While the GP research conducted in this thesis is further developed by using GPs for a novel ROM technique. In this research, GPs have created a blackbox function that has then been reduced to contain linear combinations of the most important input variables and trained to create a novel MOF exploration technique. I, the candidate, believe that this research has scope for a research grant in itself where a team of dedicated MOF scientists with an interest in ML could develop a universal model. With relevance to the research undertaken throughout this thesis, this manuscript has shown that GPs have the capability to model unknown science from a large quantity of experimental data to provide valuable knowledge. We have shown that the knowledge gained through GP based GSA's provided the spark to begin projects otherwise unachievable.

3.4 Paper 4:

Robust Probabilistic Electricity Price Forecasting Using a Hybridisation of Gaussian Processes and Clustering

3.4.1 Publication Information

The fourth paper has been submitted for consideration as a research paper in the International Journal of Forecasting. The research completed in this manuscript was initiated from a published conference paper (Yeardley et al., 2020b) where I, the candidate, presented the results for electricity price forecasting at a virtual international conference.

In this publication, I, the candidate, co-wrote the manuscript with Dr Diarmid Roberts. I conceptualized the research by combining electricity price clustering work that Dr Roberts had presented to our research group with my own GP studies. Dr Roberts had investigated how electricity prices can be clustered using ML classification techniques. I, the candidate, conceptualised the novel hybridisation methods by combining regression and classification to predict the electricity price. Together, myself and Dr Robert A. Milton developed the software used for training GPs, while Dr Roberts developed the software used for clustering. I, the candidate, then wrote the code that combined both the clustering and GPs together so that the results from the clustering methods could be used as input variables in the GP. For the

manuscript itself, I, the candidate, completed the formal analysis and co-wrote the paper with supervision from Dr Solomon F. Brown.

3.4.2 Paper Contribution

Electricity price forecasting is a very popular area of research due to the huge financial benefits successful forecasts can produce. However, the research is vast, ranging from continuous short-term predictions of the mean price to long-term forecasts into the future. The literature review, shown in Chapter 9, explains the gap in the literature when it comes to probabilistic electricity price forecasts. Specifically, previous research has struggled to provide statistical evidence that supports their new superior forecasting technique. Knowing a significant gap is available in the research, the research conducted in this paper focused on developing a GP technique that could provide a mid-term probabilistic forecast.

With respect to this thesis, the research conducted within this manuscript focused on developing the application of GPs when extrapolation is required in forecasting scenarios. This work is important because GPs famously requires a trend to be fitted into the model definition otherwise it will revert to the mean of the training data when the GP attempts to predict using input variables extrapolated away from the training data. For example, seasonal trends are included with a GP using a Seasonal AutoRegressive Integrated Moving Average for transmission system load forecasting (Nop and Qin, 2021). Often, literature has shown GPs to be the worst ML technique when compared to other forecasting methods (Qader et al., 2021). Therefore, the research conducted during this paper has developed a new method that combines clustering with GPs to produce a novel hybridisation technique, developed specifically for forecasting.

The work published in this manuscript focuses on developing GPs in a very popular field of study. However, I, the candidate, believe that this is the first paper to apply GPs to forecast the hourly resolution four weeks ahead. Additionally, the guidelines by (Nowotarski and Weron, 2018) have been followed to ensure the models were tested robustly using appropriate error metrics and the Diebold-Mariano hypothesis test. In the manuscript, the novel hybridisation method is tested against ordinary GP regression and a technique found in the literature which

clusters the data so that numerous GP regression models are used for each cluster (Mori and Nakano, 2015; Zhang et al., 2021). The results have shown that the novel hybridisation method does improve GPs' prediction accuracy and so by adding new input variables (which do not need to be extrapolated when making future predictions) the model is shown to be improved.

Due to the novelty of the technique developed in this manuscript, future work would consist of testing the method against the state of the art forecasting techniques. Within the manuscript, the results from a gradient boosting method are shown, but these are for a guideline to what is "good" and "bad" only. The research conducted in this thesis focuses on developing the application of GPs for chemical engineering. This hybridisation method successfully improves an ordinary GP regression model for forecasting. Further, this work has already made a clear contribution to science as the novel hybridisation technique has been used to generate a set of representative price scenarios, as an input for a stochastic scenario-based optimisation model (Biggins et al., 2022). This model optimises the scheduling of a hydrogen electrolyser and battery storage to maximise their owner's profits, under market uncertainty. The set of scenarios generated by the hybridisation technique encapsulates the underlying price uncertainty, by representing a range of realistic outcomes. Hence allowing the storage owner to make more informed decisions.

3.5 Paper 5:

Integrating Machine Learning techniques into Optimal Maintenance Scheduling

3.5.1 Publication Information

The final paper has been submitted for consideration as a research paper in Computers and Chemical Engineering.

In this publication, I, the candidate, wrote the manuscript with help editing and reviewing from Dr Jude O. Ejeh, Louis Allen and Dr Solomon F. Brown. The project was conceptualized by myself, Dr Ejeh and Professor Joan Cordiner. Together, Professor Cordiner and I focused our reading on how ML is used for process safety in the industry. Literature showed the popularity of using ML to predict faults in machinery and developing algorithms for predictive mainten-

ance. Together, we conceptualized this research by focusing on improving literature through the implementation of predictive maintenance powered by a combination of ML methods. The devised plan was to create a workflow that would employ predictive maintenance to save operational expenses in an industrial plant. Thus, Dr Ejeh contributed by writing a code that scheduled maintenance tasks based on various scenarios (inclusion of predictive maintenance or not). I, the candidate, Dr Ejeh and Professor Cordiner worked together to curate the plant maintenance data provided by Klein and Bergmann (2019) for a Fischertechnik factory model. The formal analysis was completed by myself and Dr Ejeh before Professor Cordiner ensured the results showed important findings for the industry. Finally, the project was co-supervised by Professor Cordiner and Dr Brown.

3.5.2 Paper Contribution

The research proposed a novel methodology that utilises machine learning to predict maintenance faults and maintenance repair time and uses this to underpin the scheduling of maintenance activities. The workflow developed can be used to plan maintenance and optimise the schedule based on the costs, labour availability and plant layout. Development of the methodology was completed by investigating various ML algorithms, testing them on a Fischertechnik model simulation. The final results showed the developed workflow has the potential to change the way maintenance tasks are approached as it reduced overall plant maintenance costs. However, the contributions to the science produced from the final manuscript are limited due to time constraints and data constraints. I, the candidate, believe the continuation of this work could be conducted by a team with a research grant.

The workflow developed during this research combines two popular research areas with a less explored branch related to machine maintenance, producing an effective maintenance strategy. Both predictive maintenance and maintenance scheduling are speciality areas that researchers have worked on in detail, using machine learning to continuously improve standard techniques. However, a large gap is available to develop models to accurately estimate the maintenance time required by machines. The gap in literature is identified in Chapter 10,

Section 10.2. Additionally, by combining the time required to fix a fault with predictive maintenance, more accurate machine maintenance scheduling can be achieved using an optimisation algorithm.

Most importantly, the final manuscript produced for this thesis has shown a much broader aspect to ML other than just GPs. The maintenance workflow was developed by investigating many machine learning algorithms instead of focusing on the deployment of GPs. Interestingly, the research found GP regression to be the most accurate method to estimate the maintenance times required. The combination of GP regression with the Quadratic Discriminant Analysis classification algorithm for predictive maintenance provided the most accurate results for the schedule optimisation. Conclusively, the research showed GPs and ML can have an important role in the future of the industry.

Chapter 4

Gaussian Processes

In this section we will provide an in-depth description of Gaussian Processes (GPs), introducing their origins and the mathematical derivation. A GP is an important stochastic process that corresponds to a class of normal distributions on a function as defined by (Williams and Rasmussen, 2006):

Definition 1 *A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

4.1 Origins

As a simple probability distribution, GPs have been studied for centuries, however, only since the 1940s have GPs been used to make predictions. A particular example is time series analysis where the work dates back to that from Kolmogorov (1941) and Wiener (1949).

It was not until the 1970s when GPs became widely accepted. In particular, at this time they were being used for geostatistics to make predictions in two or three dimensions. The method was developed by Krige (1952), a mining engineer from South Africa, where his research introduced mathematical statistics to the valuation of gold mines using a regression procedure. Theoretically, a weight was assigned to each sample assay and the combination of the available assays was used as the block grade estimator.

In 1963, Matheron (1963) recognised Krige's contributions and used his name to describe

the spatial mineral evaluation process that formalised and generalised the GP regression procedure (Journel, 1977). Hence, the probabilistic process was first introduced to the field of geostatistics under the name of Kriging, before it was introduced in machine learning as GPs. Kriging has since found considerable applications in Geostatistics including an interesting Bayesian analysis (Kitanidis, 1986). An excellent overview of the origins of Kriging can be found here Cressie (1990).

The popularity of Bayesian analysis and, in particular, the use of GPs were increased when statisticians applied a Bayesian framework to overcome optimal design problems (O'Hagan and Kingman, 1978). The new multivariate regression model consisted of using a GP prior to aid curve fitting.

Williams and Rasmussen (1996) were the first to describe GP regression in a machine learning context in 1996. Before this, Neal (1995) had shown that the Central Limit Theorem applies to neural networks because the data converges to a GP prior as the contribution of hidden units increases. Motivated by this research into neural networks, Williams and Rasmussen (1996) specified GPs parametrically for regression problems. The research showed how to make predictions using a GP and how to estimate the hyperparameters. To summarise, the use of GPs for machine learning was proven to be a successful framework to build from, presenting examples of its performance with real-world problems Williams and Rasmussen (1996).

4.2 Mathematical Derivation

GPs are a popular machine learning technique used for regression. Here, we follow the training data in a Bayesian framework to derive the posterior of a GP that is used to make predictions on test data.

The mathematical derivation of a GP begins by specifying the initial training data, $\mathbf{D} = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_n}, y_n)\}$. The input data, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, consists of n amounts of input vectors of size d:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] = \begin{pmatrix} x_{1,1} & ... & x_{1,n} \\ ... & ... & ... \\ x_{d,1} & ... & x_{d,n} \end{pmatrix}$$

A GP is a stochastic process that takes deterministic input data and indexes them to a random variable, $f(\mathbf{x})$. The process maps the deterministic variables to observed multivariate Gaussian variables, $\mathbf{f} = [f_1, f_2, ..., f_n]$, which are jointly normal. In real life applications, it is assumed that only noisy outputs, $\mathbf{y}$, are observed. The noise, $\varepsilon(\mathbf{x}_i)$, is assumed to be independent with a Gaussian distribution of $\mathbf{0}$ mean and $\sigma_N^2\mathbf{I}$ variance.

$$\mathbf{y} = [y_1, y_2, ..., y_n]$$

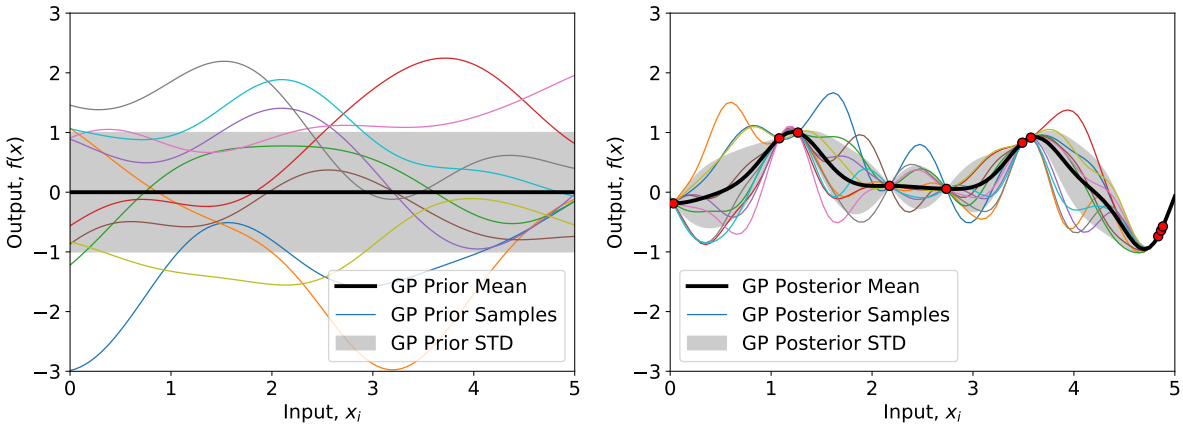
$$y_i = f(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i)$$

Therefore, the full training dataset, $\mathbf{D} \in \mathbb{R}^{n\times(d+1)}$, consists of input data, $\mathbf{X} \in \mathbb{R}^{n\times d}$, paired with noisy outputs, $\mathbf{y} \in \mathbb{R}^n$.

A GP regression model learns a function through probabilistic inference where it combines prior knowledge with the information provided by observations. Figure 4.1a shows how the prior information is expressed in a number of sample functions drawn at random from the prior distribution. Then the GP uses the training dataset, $\mathbf{D}$, to learn the functions that pass through the datapoints. Equation (4.1) shows how Bayesian Inference is used to combine the prior and the data leading to the posterior distribution as presented in Figure 4.1b after 10 training datapoints are observed.

$$\overbrace{(\mathbf{f}|\mathbf{X},\mathbf{y})}^{\text{Posterior}} = \frac{\overbrace{P(\mathbf{y}|\mathbf{f})}^{\text{Likelihood}}\overbrace{P(\mathbf{f}|\mathbf{X})}^{\text{Prior}}}{\underbrace{P(\mathbf{y}|\mathbf{X})}_{\text{Marginal Likelihood}}} \tag{4.1}$$

GP regression is a Bayesian approach which assumes a GP prior over functions. The prior



(a) (b)

Figure 4.1: (a) Ten samples drawn from the prior distribution. (b) The samples from the GP posterior once ten datapoints have been observed.

is usually assumed to have $\mathbf{0}$ mean but can also be linear in inputs.

$$P\left(\mathbf{f}|\mathbf{X}\right) = \mathsf{N}\left[\mathbf{f}|\mathbf{0}, \mathbf{K}_{ff}\right] = \frac{1}{\sqrt{\left(2\pi\right)^{n}\left|\mathbf{K}_{ff}\right|}} \exp\left(-\frac{1}{2}\mathbf{f}^{\intercal}{\mathbf{K}_{ff}}^{-1}\mathbf{f}\right)$$

The covariance matrix, $\mathbf{K}_{ff} \in \mathbb{R}^{n\times n}$, is evaluated between all pairs of $\mathbf{D}$. $\mathbf{K}_{ff}$ is dependent on the training data and some hyperparameters θ. In particular, the elements of $\mathbf{K}_{ff}$ are from the kernel function $k\left(\mathbf{x}_i, \mathbf{x}_j\right) : \mathbb{R}^{i+d} \times \mathbb{R}^{j+d} \rightarrow \mathbb{R}^{i} \times \mathbb{R}^{j}$, expressing the correlation between responses to input samples of sizes $(i \times d)$ and $(j \times d)$.

As previously mentioned, the observed data $\mathbf{y}$ is generated with Gaussian noise and so the likelihood is

$$P\left(\mathbf{y}|\mathbf{f}\right) = \mathsf{N}\left[\mathbf{0}, \sigma_N^2\mathbf{I}\right] = \frac{1}{\sqrt{\left(2\pi\right)^{n}\left|\sigma_N^2\mathbf{I}\right|}} \exp\left(-\frac{1}{2}\left(\sigma_N^2\mathbf{I}\right)^{-1}\right)$$

The marginal likelihood is defined by marginalising over the latent function, $\mathbf{f}$, from the product of the likelihood and the prior. The product of two Gaussian's is another Gaussian and so the marginal likelihood is available in the analytical form:

$$P\left(\mathbf{y}|\mathbf{X}\right) = \int P\left(\mathbf{y}|\mathbf{f}\right) P\left(\mathbf{f}\right)\ d\mathbf{f} = \mathsf{N}\left[\mathbf{y}, \mathbf{K}_{ff} + \sigma_N^2\mathbf{I}\right]$$
$$= \frac{1}{\sqrt{\left(2\pi\right)^n \left|\mathbf{K}_{ff} + \sigma_N^2\mathbf{I}\right|}} \exp\left(-\frac{1}{2}\mathbf{y}^\intercal\left(\mathbf{K}_{ff} + \sigma_N^2\mathbf{I}\right)^{-1}\mathbf{y}\right)$$

Throughout this work, we maximise the marginal likelihood to find the hyperparameter values and build a GP using the ROMCOMMA software library (Milton and Brown, 2019). The optimisation requires the training data to learn a number of hyperparameters dependent on the kernel function, $k\left(\mathbf{x}_i, \mathbf{x}_j\right)$, chosen for each case study.

The overall goal of GP regression is to predict the unobserved output, $\mathbf{f}_*$, from t amounts of test input points, $\mathbf{X}_*$. As previously discussed, the GP is trained using $\mathbf{D}$, consisting of noisy observed outputs, $\mathbf{y}$ and input data, $\mathbf{X}$.

Assuming the joint GP prior has zero-mean then the joint distribution of the noise free outputs (latent functions, $\mathbf{f}$ and $\mathbf{f}_*$) is

$$P\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_*\end{bmatrix}|\mathbf{X}, \mathbf{X}_*\right) = \mathsf{N}\left[\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_*\end{bmatrix}|\mathbf{0}, \begin{bmatrix}\mathbf{K}_{ff} & \mathbf{K}_{f*}\\ \mathbf{K}_{*f} & \mathbf{K}_{**}\end{bmatrix}\right]$$

Due to the additional data new covariance matrices are shown. $\mathbf{K}_{**} \in \mathbb{R}^{t\times t}$ is the covariance matrix between the test inputs, $\mathbf{X}_*$ and $\mathbf{K}_{f*} \in \mathbb{R}^{n\times t}$ and $\mathbf{K}_{*f} \in \mathbb{R}^{t\times n}$ are the cross covariance between the training and test inputs.

Given the GP is trained using observed noisy outputs, $\mathbf{y}$, we introduce the noise term, $\sigma_N^2\mathbf{I}$, to derive the joint prior distribution of the observed noisy outputs, $\mathbf{y}$, and the test latent outputs, $\mathbf{f}_*$, under the prior as

$$P\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_*\end{bmatrix}|\mathbf{X}, \mathbf{X}_*\right) = \mathsf{N}\left[\mathbf{0}, \begin{bmatrix}\mathbf{K}_{ff} + \sigma_N^2\mathbf{I} & \mathbf{K}_{f*}\\ \mathbf{K}_{*f} & \mathbf{K}_{**}\end{bmatrix}\right]$$

GP regression is a Bayesian method and so the joint prior distribution is restricted to the training data to derive the posterior distribution. This is done by conditioning the joint prior on

the observations using standard Gaussian Identities (Williams and Rasmussen, 2006):

$$P\left(\mathbf{f}_*|\mathbf{y},\mathbf{X},\mathbf{X}_*\right) = \mathsf{N}\left[\mathbf{K}_{*f}[\mathbf{K}_{ff}+\sigma_N^2\mathbf{I}]^{-1}\mathbf{y},\ \ \mathbf{K}_{**}-\mathbf{K}_{*f}[\mathbf{K}_{ff}+\sigma_N^2\mathbf{I}]^{-1}\mathbf{K}_{*f}\right]$$

Both the prior and the likelihood are Gaussian and so the posterior distribution over functions is a Gaussian with a joint distribution of observed values, $\mathbf{f}$ and predictions, $\mathbf{f}_*$. Therefore, the GP posterior can be used to make predictions, $\mathbf{f}_*$ at new inputs, $\mathbf{X}_*$. To make predictions, the posterior distribution is partitioned, using standard rules for conditioning Gaussians, to give Equation (4.2) that shows the predictive equations used for noisy observations:

$$\mathbf{f}_*(\mathbf{X}_*) \sim \mathsf{N}\left[\bar{\mathbf{f}},\boldsymbol{\Sigma}\right] \tag{4.2}$$

where

$$\bar{\mathbf{f}} := \mathbf{K}_{*f}[\mathbf{K}_{ff}+\sigma_N^2\mathbf{I}]^{-1}\mathbf{y} \tag{4.3}$$

$$\boldsymbol{\Sigma} := \mathbf{K}_{**}-\mathbf{K}_{*f}[\mathbf{K}_{ff}+\sigma_N^2\mathbf{I}]^{-1}\mathbf{K}_{*f} \tag{4.4}$$

The GP predictive distribution is a correlated prediction that gives the mean and marginal variance for each test point, making it a full GP in its own right. Implementation of Equation (4.2) uses Cholesky decomposition as it is faster and numerically more stable than directly inverting the matrix. Once the training data has been used to learn the hyperparameters, the predictions are also made using the ROMCOMMA software library (Milton and Brown, 2019).

In this Chapter, we have introduced the theory behind GPs, briefly discussing their origins before deriving the predictive equations used in GP regression. The publications presented in this Thesis have used the GP model as a powerful tool to overcome problems that arise in Chemical Engineering.

Chapter 5

Conclusion

5.1 Concluding Remarks

This thesis presents Gaussian Processes (GPs) as a machine learning (ML) technique that can be beneficial to chemical engineering. The research has focused on the application of GP regression and the analysis of results to aid in collaborative efforts, showcasing the huge impact ML can have in both research and industry. Throughout the work, the application of GPs was investigated as a surrogate modelling technique enabling analysis that previously had not been conducted. Overall, five case studies were used to develop our understanding of GPs. Each presenting novel research submitted as manuscripts to leading scientific journals.

By applying GPs to different subject areas within chemical engineering, the research has automatically had to focus on different methods of application and analysis to achieve results that are insightful and truly beneficial to each topic. To begin, the thesis introduced the reader to ML and explained the motivation behind using GPs, clearly defining the contribution it can make to chemical engineering. Then, Chapter 2 introduces each of the publications, the importance each one has had to its research field and the application of GPs. However, it is also important to understand how the published research collectively makes a positive contribution to chemical engineering.

Currently, the chemical engineering industry is transforming as data is becoming more and more accessible. Computational models used for engineering research and development are

becoming more and more complex. Therefore, data required for analysis is either large or sparse, depending on whether sensors are continuously measuring and recording vast quantities of information, or whether an experiment or computational model can only be conducted in a limited amount of time. Either way, data analysis techniques are now more important than ever for chemical engineering. With the data revolution, computer programmers and statisticians are vital for all disciplines, including chemical engineering, continuously developing complex ML algorithms that accurately learn patterns and make predictions. However, the collaboration between engineers and statisticians is often like a duck talking to a chicken.

I, the candidate, has shown that it is time for collaboration as the application of ML is shown to be hugely beneficial to chemical engineers. This thesis goes into great depth about how to use one of the most prominent ML approaches. The research focuses on using GPs to analyse complex engineering systems. It has shown, GPs are well suited to chemical engineering situations leading to many achievements throughout this research.

5.2 Limitations and Future Work

This thesis developed several methods that use GPs to provide insightful results in a variety of fields. The findings of this research have successfully answered the research questions proposed in Section 1.2.1 and ultimately, helped maximise the applicability of GPs for future work in chemical engineering. Regardless, there are still many avenues of ML to explore within chemical engineering. In particular, the research conducted has provided further insight into what work should be conducted to make ML and GPs more applicable for all chemical engineers.

Chapter 6 focused on using GPs to provide a Global Sensitivity Analysis (GSA) for a complex thermal runaway (TR) model. The benefits of using GP regression as a surrogate modelling technique are highlighted as they provide the key to calculating time-dependent Sobol' indices for the first time. However, the research conducted here came with significant difficulties and limitations. Creating an accurate surrogate model for the full time simulation was the first major challenge. This was caused by the requirement of independent GP regression models for each time-step of the computer simulation. Thus, a significant amount of computational power

was needed to train all of the GPs to create a surrogate model. In this work, it was managed by altering the time split throughout the computational simulation, during the beginning and end of the simulation when the model output was consistent, the time step was chosen to be large, then within the range of time that the model output changes (due to TR), the time-step was reduced. In future work, it would be beneficial to investigate the use of sparse GPs to allow reduced computational complexity so that more GP regression models can be used within the surrogate model. On top of this, future research should focus on making the GP regression models more accurate when the output value changes early or late in the model simulation. The work conducted used a limited range of data so that the model output never went 12 std's away from the mean value. However, model simulations showed this was a possibility if the TR event began significantly early in comparison to the others. Current literature uses techniques such as time warping (Wang and Gasser, 1997; Maiy and Sudret, 2017) to improve accuracy so that the significant event occurs at the same time. However, for a time-dependent GSA this would not help explain the importance of each input variable throughout the model simulation. Thus developing techniques to help GPs predict accurate outputs far away from the mean of the data is required.

The second published paper used GPs to develop an efficient experimental design using a wet granulation case study. In this work, the method successfully indicated any unidentified sub-processes using two GSA linking experimental data to model simulation data. For the given case study, the ordinary GP was shown to be accurate and efficient enough to provide calculations of Sobol' indices. Consequently, for the case study used, the GP and GSA tools do not require any further work to achieve successful results. However, as noted in section 7.6 it would be beneficial to further test the efficient experimental design methodology to other particulate processes. Ultimately, the method proposed enables reduced experimental effort in comparison to a conventional experimental design. So the goal is to optimise the modelling parameters against experimental data. The possibility of including parameter optimisation directly into the GP based GSA warrants further investigation.

Paper three investigates a novel method used to predict the deliverable capacity of metal-

organic framework (MOF) structures. In this sector, the overall goal of future work would be to develop a universal model. Additionally, the work opens the door for future work using ML algorithms for GSA on data with correlated input variables. Throughout the research conducted, the initial analysis of the calculated Sobol' indices was made difficult due to complex relationships between MOF pore properties. Due to the limited literature providing GSA techniques for correlated input variables, future research investigating the impact correlated inputs have on Sobol' indices is required. The novelty within this work was created by using a reduced order model (ROM) by rotating input variables to create primary dimensions. This in itself is believed to combat the impacts created from correlated inputs. However, due to the novelty of the method used, limitations in this research were created by not robustly investigating the novel properties created by calculating active subspaces using GPs beforehand. The publication of the innovative mathematics utilised to construct the technique should be a priority in future development.

Chapter 9 concentrates on applying GPs to the popular field of electricity price forecasting. In the manuscript, a novel hybridisation method is used for mid-term probabilistic electricity price forecasting with hourly resolution. The work utilised clustering techniques to enhance the forecasting capabilities of GP regression by creating an input variable that does not need extrapolation in future predictions (the cluster number in the test data will always have been seen in the training data). The work successfully showed the novel hybridisation method to improve ordinary GP regression and clustering to make numerous GPs. However, a major limitation of the hybridisation technique is that the GP regressions model must incorporate categorical input variables to use the cluster group as an input. To deal with this, the clustering techniques must classify the data into groups that have a rational ordering to it. This enables the relaxation of categorical variables so that they could be treated as continuous variables. Similarly, the day of the week was used as an input variable and relaxed as a continuous variable in the same way. Although this method of relaxing categorical variables is widely accepted, it does impact the hybridisation method as it forces the choice of the clustering method. Therefore, future work should focus on deriving a method to utilise categorical variables within GPs. For example,

Qian et al. (2008) developed a new covariance function that could incorporate both continuous and categorical factors. It would be of interest to investigate the use of this correlation function with this hybridisation method.

The final research project conducted for this thesis developed a novel maintenance workflow that can be used in industry to develop an optimum maintenance schedule. This was achieved by investigating three techniques and integrating them to quickly analyse process data and produce a framework that consists of three stages, predictive maintenance, maintenance time estimation and schedule optimisation. Each stage was robustly investigated on a manufacturing process simulation model successfully showing the potential that the novel maintenance workflow has for the way maintenance tasks are approached in Industry 4.0. The manufacturing simulation model provided the data required to model maintenance tasks in a cyber-physical model that is becoming more accessible during the fourth industrial revolution. The digitisation of the process industry is creating an abundance of interconnected data through the Internet of Things. In the manuscript, the research has shown that ML methods are ready for Industry 4.0 and that artificial intelligence will have significant benefits in the way that industry is run in the future. Interestingly, the work also found GPs to be the best regression algorithm to use in a maintenance time estimation model. The literature review (please see Section 10.2) has shown time estimation models are often used to predict the product manufacturing times. However, for maintenance time, traditional methods have still been used that lead to expensive overtime and rushed fixes. Therefore, investigating the use of ML algorithms for a maintenance time estimation model and discovering GP regression was the most accurate method concluded the research undertaken in this thesis by showing the huge potential GPs have in comparison to other ML algorithms. In future work, more research is required to apply and test the use of GPs for maintenance time estimation, investigating the tool in more detail and experimenting with the methods of applying GPs to improve results. Further, the novel maintenance workflow developed in this research can be improved by investigating the use of GP classification for predictive maintenance. Once again, this research has also shown limitations with using GPs and other ML algorithms as a large number of sensors provide many input variables. Some

of these sensors give categorical variables and some are correlated around the cyber-physical model. Future work to help commercialise the novel maintenance workflow should focus on using ML algorithms that successfully predict maintenance tasks and times, but also provides a detailed analysis into the cyber-physical model. This will enable process engineers the ability to understand the plant's process and understand which parts are the plant inherently affect the important outcomes, such as the yield. Conclusively, the research conducted for this manuscript has incredible potential, but lots of work has to be done to make it beneficial for the industry.

Overall, the research conducted for this thesis has provided exciting results and helped other researchers in different areas of chemical engineering. This has been done using the GP ML algorithm, presenting many ways to use and develop the technique for each subject area. Clearly, GPs are well suited to aid many disciplines within chemical engineering, particularly when data analysis of the system requires a reduced model that encapsulates the system behaviour in a cheaper, simpler framework. This thesis explains how to conduct efficient data analysis techniques using GPs in chemical engineering.

Chapter 6

A study of the thermal runaway of lithium-ion batteries: A Gaussian Process based global sensitivity analysis

6.1 Abstract

A particular safety issue with Lithium-ion (Li-ion) cells is thermal runaway (TR), which is the exothermic decomposition of cell components creating an uncontrollable temperature rise leading to fires and explosions. The modelling of TR is difficult due to the broad range of cell properties and potential conditions. Understanding the effect that thermo-physical and heat transfer characteristics have on the TR abuse model output is essential to develop more accurate and robust TR models. This study uses global sensitivity analysis (GSA) to investigate the effect of the cell parameters on the outcome of TR events. Using a Gaussian Process (GP) surrogate model to calculate the Sobol' indices, it is shown that the emissivity value is the dominant thermo-characteristic throughout the overall abuse scenario. Further analysis, investigating three key TR features shows the conductivity coefficient to be the most important with respect to the maximum temperature reached during TR. Results demonstrate that researchers can confidently estimate some thermo-characteristics but require accurate characterisation of

the emissivity and conductivity coefficient to ensure robust predictions. Given the importance of battery technology to aid in global de-carbonisation, these findings are key to increasing their safe design and operation.

6.1.1 Keywords

Gaussian Process; Thermal Runaway; Sobol' Indices; Global Sensitivity Analysis; Li-ion Cells

6.2 Introduction

Lithium-ion (Li-ion) cells are an increasingly popular electrochemical storage device (Steen et al., 2017; Kim et al., 2012; International Energy Agency, 2016) which play a pivotal role in applications such as electric vehicles and grid energy storage. Li-ion cells have been extensively studied in order to increase cell performance (Nitta et al., 2015; Ghadbeigi et al., 2015), reduce cell cost (Nitta et al., 2015), reduce environmental impact (Peters et al., 2017) and improve safety (Abada et al., 2016). However, a particular issue with Li-ion cells is thermal runaway (TR) which is the exothermic decomposition of cell components creating an uncontrollable rise in temperature leading to fires and explosions (Wang et al., 2012; Larsson et al., 2018; Feng et al., 2018). Hence, understanding the TR process, to prevent it occurring or reducing its severity, is essential for the development of safer batteries.

Current literature focuses on the development of models designed for carrying out studies to determine the effects of cell design (Kim et al., 2007; Lopez et al., 2015; Hu et al., 2017), environmental and abuse conditions (Chiu et al., 2014; Lopez et al., 2015) and pack design (Coleman et al., 2016; Xu et al., 2017; Duan et al., 2018) on TR behaviour or prevention. Recent studies have explored the implications of using Li-ion batteries in extreme environments, investigating how severe irradiation can cause TR (Shack et al., 2014; Ma et al., 2017; Wu et al., 2019). Yet little work has been carried out to understand the uncertainties within TR modelling and/or how variance in the physical parameters can affect the temperature during a TR event.

The need to understand TR events in Li-ion cells has motivated the development of physics-based models to provide assistance in the analysis of the mechanics allowing a more cost-effective and safer method rather than iterating cell abuse experimentally. As a result, research

has been undertaken with various approaches based on the Arrhenius formulation of cell decomposition reactions and heat generation (Hatchard et al., 2001; Kim et al., 2007; Lopez et al., 2015) to develop computational models of TR in cells with studies. For example, comparing cell chemistry (Peng and Jiang, 2016), utilising efficiency factors for the conversion of electrochemical energy to thermal energy (Coman et al., 2017), and the effect of nail penetration (Chiu et al., 2014). Previous work utilises an accelerated rate calorimeter (ARC) to determine the initial reaction kinetics of the solid electrolyte interface (SEI) and cathode reactions of cells (Richard and Dahn, 1999a; MacNeil et al., 2000). Mao et al. (2020) also used an ARC to investigate the self-heating reaction and TR criticality of Li-ion cells resulting in kinetic data that will aid in future modelling of Li-ion battery safety. Additionally, inverse modelling techniques have been used to estimate the parameters involved in the reaction kinetics for both cell components (Richard and Dahn, 1999b; MacNeil et al., 2000) and full cells (Ren et al., 2018; Liu et al., 2018). Liu et al. (2018) developed a one reaction model for an entire cell which takes into account the state of charge (SOC) by fitting experimental ARC data at different SOC's to determine the parameters of the Arrhenius equation and the heat of reaction as a third-order polynomial which is a function of SOC. Ren et al. (2018) determined parameter values through Kissinger and nonlinear fitting methods of direct scanning calorimetry (DSC) data. Fundamental thermal abuse experiments of ARC, DSC and oven exposure are used to validate such computational models. However, the TR models developed for Lithium iron phosphate (LFP) cells (Peng and Jiang, 2016) have not been validated well as inaccuracies are found when compared to new experimental work (Bugryniec et al., 2019). Therefore, the need to develop computational models of TR events are strained by the lack of understanding in the model parameters meaning an intensive study of the parameter sensitivities is essential for a thorough understanding of the TR of Li-ion cells.

A sensitivity analysis (SA) characterises the relationship between a system's inputs and outputs thus showing how the uncertainty of the outcome can be apportioned to the different sources of uncertainty in the input. This will provide important knowledge in relation to the safety and development of Li-ion cells given that:

1. Parameters are commonly estimated within TR models if they cannot be measured or have no measured values available in the literature,

2. There is an unknown certainty of the appropriateness of values in the literature for parameters such as composite materials i.e. electrodes, or for the overall cell properties. Especially when considering that exact cell compositions for literature sources or cells under study may be unknown. For example, Liu et al. (2018) and Drake et al. (2014) have both calculated the specific heat capacity of an 18650 LFP cell but resulted in different values. Various other literature have recorded a specific heat capacity ranging from $1100\,\mathrm{J\,kg^{-1}\,K^{-1}}$ to $1720\,\mathrm{J\,kg^{-1}\,K^{-1}}$ (Spinner et al., 2015; Wang et al., 2013; Chen et al., 2005a). Highlighting the significant variation, in just one parameter, of which the authors here have found that a 10% change can lead to significant variations in the TR simulation results.

3. Carrying out SA studies can help future developments, as the uncertainty in the model can be related to a real change in a physical cell.

This work will analyse the thermo-physical and heat transfer characteristics of a TR model to determine their influence on cell temperature under external heating of an LFP cell leading to a TR event. The focus is on an LFP cell as a case study due to it being considered the safest Li-ion chemistry (Jiang and Dahn, 2004; Bugryniec et al., 2018, 2019). SA has been implemented in many different research fields (for example Saltelli et al., 2005; Rohmer and Foerster, 2011; Al et al., 2019) as it is widely acknowledged as a good practice to better understand model behaviour. Specifically for battery research, sensitivity studies have been employed to advance the development and design of battery performance (Drews et al., 2003; Schmidt et al., 2010; Vazquez-Arenas et al., 2014; Zhang et al., 2014) without considering TR event occurring. Additionally, these sensitivity studies focused on minor parameter changes and how it affects the variation in the model output using a local SA study. In general, the most common class of SA method is the global sensitivity analysis (GSA) because it looks at the model behaviour over the whole range of inputs and outputs. Unlike other methods, such as local SA, GSA quantifies

the variation of the model response in the entire parameter domain fully exploring the input space. The first example, to the authors knowledge, to practise a GSA for battery development was Trembacki et al. (2016), where polynomial chaos expansion was used to calculate sensitivity index values for a thermal simulation of molten salt batteries. Recent literature by Lin et al. (2018) has applied a GSA to a large scale multiphysics Li-ion battery model. The SA was achieved by calculating the Sobol' indices using a polynomial chaos expansion approach to surrogate the 3D multiphysics model. Specifically, the safety of an LFP cell has been explored using GSA by Rajan et al. (2018) who considered the displacement, temperature and strain rate of a battery with respect to it's mechanical strength when subjected to external impacts that lead to TR. The research conducted a SA using an artificial neural network to emulate a finite element model and showed the temperature of a cell to have the most influence on the mechanical strength leading to catastrophic TR explosions. Whereas, the TR process of a Li-ion cell has only been analysed once with respect to sensitivity studies, when GPs were used as a surrogate model to optimise the reaction parameters in a TR abuse simulation (Milton et al., 2019). The work used the error of a heuristic fit (RMSE) between simulated and experimental results as an output to a GP, which was further reduced by an optimal rotation of the input basis. The input variables being optimised, were the parameters that control the reactions which govern the heat generation in the TR model.

The methodology presented in this work will use the Sobol' sensitivity indices, which are a variance-based decomposition method that is considered the benchmark for GSA methods (Rohmer and Foerster, 2011; Al et al., 2019). However, the calculation of Sobol' indices require a significant number of model evaluations to ensure convergence to a satisfactory precision level. Hence, the computational burden can be reduced by using meta-modelling techniques such as the polynomial chaos expansion (Brown et al., 2013; Sudret, 2008), artificial neural networks (Li et al., 2016), and Gaussian Processes (GPs) (Marrel et al., 2009). Here, based on our previous work, the surrogate model will be developed using GPs as they are a widely used tool for Bayesian nonlinear regression and provide an approach that predicts a distribution allowing uncertainties for each prediction. This novel approach will allow us to determine:

- which of the thermo-physical and heat transfer characteristics contribute most to the temperature output variability with respect to the time during a TR simulation,
- the number of interactions between the thermo-characteristics and with which inputs,
- which properties are insignificant and so do not need to be well characterised, helping to simplify TR models.

These goals are achieved by developing a TR model of an LFP cell which is emulated by a GP surrogate model enabling the calculation of Sobol' indices.

This work is organised as follows. Section 6.3 describes the methodology and the mathematics behind the three techniques before the results sections consist of the analysis of the three tools. First, the TR model is compared to experimental data, ensuring it accurately represents a real cell under TR allowing GPs to be created for a time-dependent SA on the temperature of a cell. The role of the thermo-characteristics on TR behaviour is then further investigated with analysis of the time at which self-heating and TR begins, as well as the the maximum temperature reached by the TR event. Once the GP's were produced, each one had to be validated ensuring inaccuracies were not carried over to the calculation of the Sobol' indices. The final results from the GSA presents the time-dependent Sobol' indices and the Sobol' indices from the three TR features. The report is then concluded in Section 6.5 detailing how the resulting sensitivity measures can guide the decision-making process when designing Li-ion cells and their models.

6.3 Methodology

This section presents the TR model, developed for an LFP cell, before it describes the formulation of the GP emulators and the calculation of the Sobol' indices.

6.3.1 TR Model

The TR abuse model for the LFP cell is constructed in the commercial finite element modelling software *COMSOL Muiltiphysics 5.2a* (COMSOL Multiphysics®V5.2a). The governing equations for heat transfer utilises one-dimensional forms of Fourier's law of heat conduction,

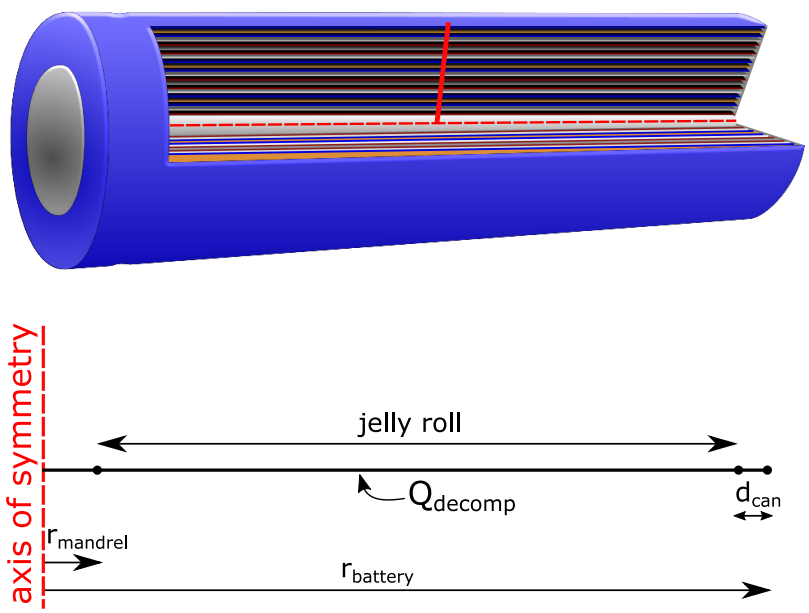


Figure 6.1: Top) Representation of a 18650 cell indicating (by the red line) the model simplification, bottom) schematic of model geometry.

Newton's law of cooling for convection and Stefan-Boltzmann equation for radiation (for example Kim et al., 2007). The model represents an 18650 LFP cell (1.5Ah) under oven exposure with free convection. The model assumes a 1D axi-symmetric geometry, consisting of three geometric domains; the mandrel, the jelly roll (the coil of electrodes, separator and current collector layers) and the cell casing, see Figure 6.1. The multilayer jelly roll is assumed to be constituted of a single homogeneous material to reduce computational complexity. This is to allow the use of measured properties of an entire cell and also to reduce model input parameters. Further, the model assumes all geometric domains take the average material values of a whole cell, as these are more accurately available, and it also reduces model input parameters.

Heat transfer in the model assumes a solid body throughout and considers conduction throughout the cell, while at the cells' surface free convection and radiation are considered.

The exothermic decomposition reactions driving TR are described by individual Arrhenius equations. Here the four common reactions are considered, these are the SEI decomposition, the negative electrode reaction (NE), positive material decomposition (PE) and electrolyte decomposition (E) (Spotnitz and Franklin, 2003), while the total decomposition heat is the summation

of these individual heat generation terms. This decomposition heat is assumed to be distributed over the entire jelly roll domain. The formulation of these governing reactions follows that outlined by Kim et al. (2007).

The rate of reaction, R_x (s^{-1}), for the SEI, positive and electrolyte reactions is given by:

$$R_x = A_x e^{\left(\frac{-E_{a,x}}{RT}\right)} C_x^{m_x} \left(1 - C_x\right)^{n_x} \tag{6.1}$$

where subscript x corresponds to one of the reactions sei, pe or e. A_x (s^{-1}) is the frequency factor, $E_{a,x}$ ($\mathrm{J\,mol}^{-1}$) is the activation energy, R and T the ideal gas constant ($8.3145\,\mathrm{J\,mol}^{-1}\,\mathrm{K}^{-1}$) and temperature (K) respectively, dimensionless quantities include: C_x the reaction species, while n_x and m_x are constants.

The negative electrode reaction is of a similar form:

$$R_{ne} = A_{ne} e^{\left(\frac{-E_{a,ne}}{RT}\right)} C_{ne}^{m_{ne}} \left(1 - C_{ne}\right)^{n_{ne}} e^{\left(\frac{-t_{sei}}{t_{sei,0}}\right)} \tag{6.2}$$

with the additional term, t_{sei}, a non-dimensional representation of the change in thickness of the SEI layer as it decomposes. $t_{sei,0}$ the initial thickness. The change in reaction species for each decomposition reaction and SEI layer thickness is given by:

$$\frac{\partial C_{sei}}{\partial t} = -R_{sei} \tag{6.3}$$

$$\frac{\partial C_{ne}}{\partial t} = -R_{ne} \tag{6.4}$$

$$\frac{\partial t_{sei}}{\partial t} = R_{ne} \tag{6.5}$$

$$\frac{\partial C_{pe}}{\partial t} = R_{pe} \tag{6.6}$$

$$\frac{\partial C_{e}}{\partial t} = -R_{e} \tag{6.7}$$

The volume-specific heat generation terms Q_z ($\mathrm{W\,m}^{-3}$) from each decomposition reaction

Table 6.1: Thermo-physical and heat transfer characteristics.

Parameter	Reference value	Range	STD
$\rho(\mathrm{kg\,m^{-3}})$	2418 [a]	2413 - 2426[a]	4.260[a]
$C_p(\mathrm{kJ\,kg^{-1}\,K^{-1}})$	1105 [a]	1092 - 1115[a]	8.680[a]
$h_{conv}(\mathrm{W\,m^{-2}\,K^{-1}})$	12.5 [b]	7.00 - 12.5[c]	1.00[d]
$\kappa_r(\mathrm{W\,m^{-1}\,K^{-1}})$	0.5 [d]	0.2 - 3[c]	0.1[d]
ε	0.8 [b]	0 - 1	0.1[d]

[a] Measured and calculated experimentally (Bugryniec et al., 2019)
[b] Reference value from Hatchard et al. (2001)
[c] Literature ranges Refs. (Guo et al., 2010; Hatchard et al., 2001; Kim et al., 2007; Hatchard et al., 2000; Dong et al., 2018; Coman et al., 2017; Liu et al., 2018; Zhao et al., 2014; Tanaka and Bessler, 2014)
[d] Estimated

are given by:

$$Q_z = H_z W_y R_z \tag{6.8}$$

where H_z ($\mathrm{J\,kg^{-1}}$) is the specific heat of reaction for reactions $z = sei, ne, pe, e$ and W_y ($\mathrm{kg\,m^{-3}}$) is the volume-specific content of reactive material, i.e. $y = c, p, e$ corresponding to carbon active material, positive active material and electrolyte respectively. W_c is used in the SEI and NE heat generation equations, W_p in the PE heat generation equation and W_e in the electrolyte heat generation equation.

Hence, the total heat generation from the decomposition of the cell is:

$$Q_{decomp} = Q_{sei} + Q_{ne} + Q_{pe} + Q_e \tag{6.9}$$

The thermo-physical parameters and heat transfer parameters are initially taken from experimental findings in the literature and are presented in Table 6.1. The parameters for the decomposition equations are estimated through comparison with experimental data and are presented in Table 6.2. Parameters describing the cell geometry, initial temperature and oven temperature and simulation time are presented in Table 6.3.

Due to the uncertainty in the experimental data, ranges for the thermo-physical and heat transfer characteristics (density, heat capacity, convection coefficient, radial thermal conduct-

Table 6.2: Kinetic parameters used in the decomposition equations.

Kinetic Parameter	**Carbon**	**SEI**	**$LiFePO_4$**	**Electrolyte**
Frequency Factor, $A(\mathrm{s}^{-1})$	2.50×10^{13}	1.67×10^{15}	2.00×10^{8}	5.14×10^{25}
Activation Energy, $E_a(\mathrm{J\,mol}^{-1})$	1.42×10^{5}	1.50×10^{5}	9.60×10^{4}	2.82×10^{5}
Reaction order, m	1.00	1.00	1.00	1.00
Reaction order, n	0.00	0.00	1.00	0.00
Heat, $H(\mathrm{J\,g}^{-1})$	1714	578.0	194.7	645.0
Specific weight, $W(\mathrm{kg\,m}^{-3})$	560	560	977	151
Species initial values, C_x, t_{sei}	0.75 ($t = 0.33$)	0.15	0.040	0.99

Table 6.3: Additional simulation parameters.

Parameter	**Value**
r_{bat} (mm)	18
r_{man} (mm)	2
d_{can} (mm)	0.3
T_{int} $(^{\circ}C)$	16.5
T_{oven} $(^{\circ}C)$	218
t_{length} (min)	90

ivity and emissivity value) were defined, while the abuse parameters were kept constant. The initial values (used in the parameter estimation of the model) of the varied parameters are taken to be the mean of the normal distribution. The standard deviation of the density and heat capacity is calculated from experimental findings, while the standard deviation of the convection coefficient, conductivity value and emissivity value are estimated arbitrarily by setting it to 1 relative to the order of magnitude. A random sampling of these five parameters is undertaken using Latin Hypercube Sampling (LHS) (Stein, 1987), normally distributed about the mean, and used to generate 753 sample points in the input space. Note that as emissivity has physical limits between 0 and 1, any parameter sets that have an emissivity value out of this range are removed, leading to a curtailment of the final distribution of emissivity values. Each of these 753 parameter sets is used (in conjunction with the original estimated abuse parameters) to carry out an oven simulation at $218^{\circ}C$ exposure. The resulting temperature data and initial input parameters sets are used to train a GP as described in the previous section.

6.3.2 GP Surrogate Model

In this study, standard Bayesian conditioning is used to take Gaussian priors and derive a predictive process. This allows the creation of a GP which takes a $(1 \times d)$ row vector of inputs $\mathbf{x}$

and returns a Gaussian random variable through calculations using the predictive equations

$$y(\mathbf{x}) \sim \mathsf{N}\left[\bar{f}(\mathbf{x}), \Sigma_y + \sigma_e^2\right] \tag{6.10}$$

where

$$\bar{f}(\mathbf{x}) := k(\mathbf{x}, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I})^{-1}\mathbf{y} = k(\mathbf{x}, \mathbf{X})\mathbf{K}^{-1}y \tag{6.11}$$

$$\Sigma_y := k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I})^{-1}k(\mathbf{X}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{X})\mathbf{K}^{-1}k(\mathbf{X}, \mathbf{x}) \tag{6.12}$$

whose mean $\bar{f}(\mathbf{x})$ and variance Σ_y is learnt from training data $\mathbf{y} = f(\mathbf{X}) + \mathbf{e}$. Standard Bayesian inference has been used to express the mean prediction in terms of the $(n \times 1)$ observed responses $\mathbf{y}$ to $(n \times d)$ training inputs $\mathbf{X}$. At the heart of this lies the kernel function $k\colon \mathbb{R}^{i+d} \times \mathbb{R}^{j+d} \to \mathbb{R}^i \times \mathbb{R}^j$, expressing the correlation between responses to input samples of sizes $(i \times d)$ and $(j \times d)$; for the purposes of this work the number of input dimensions is $d = 5$. This work exclusively uses the automatic relevance determination (ARD) kernel (Wipf and Nagarajan, 2007):

$$k(\mathbf{x}', \mathbf{x}) := \sigma_f^2 \exp\left(-\frac{(\mathbf{x} - \mathbf{x}')\Lambda^{-2}(\mathbf{x} - \mathbf{x}')^\intercal}{2}\right) \tag{6.13}$$

where Λ is a $(d \times d)$ diagonal positive definite lengthscale matrix. The learning from training data requires the optimisation $d + 2$ hyperparameters, constituting of Λ, σ_f, and σ_e, through the maximum marginal likelihood $p[\mathbf{y}|\mathbf{X}]$ using the ROMCOMMA software library (ROMCOMMA, 2019).

6.3.3 Sobol' Indices

In this work, the approach to the calculation of the Sobol' indices follows Jin et al. (2004) by substituting the true simulation model with the mean of the conditional GP resulting in semi-analytic Sobol' indices. This section will describe the calculation of Sobol' indices up to evaluating integrals, which will be evaluated using the GP surrogate model to allow more efficient computation.

Sobol' indices are calculated by considering a function $y = f(\mathbf{x})$, where $\mathbf{x} := [x_1, ..., x_d]$ is a d-dimensional row vector found in the input space, Ω, and y is the model output. Assuming that the inputs are mutually dependent and that $f(\mathbf{x}) \in L^2(\Omega)$ (Sobol, 1993, 2001). For a particular input x_i, it's first-order Sobol' index is defined by

$$S_{1,i} = \frac{Var\{E(y|x_i)\}}{Var\{y\}} = \frac{D_i}{D} \tag{6.14}$$

Then second-order Sobol' indices between input i and j shows the interactions between x_i and x_j:

$$S_{2,ij} = \frac{Var\{E(y|x_i x_j)\} - Var\{E(y|x_i)\} - Var\{E(y|x_j)\}}{Var\{y\}} = \frac{D_{ij}}{D} \tag{6.15}$$

To be able to express the whole effect of an input on the output, the total Sobol' index is (Saltelli and Homma, 1996)

$$S_{T,i} = S_{1,i} + \sum_{j \neq i}^{n} S_{2,ij} + \sum_{j \neq i, k \neq i, j<k}^{n} S_{3,ijk} + ... \tag{6.16}$$

Therefore, the first-order Sobol' indices measure the contribution to the variance solely attributable to x_i, in contrast, the total Sobol' index of i corresponds to its own contribution including interactions with the other inputs.

From here, the partial variances of y are determined through a decomposition method presented by Sobol (1993) which evaluates each term through multidimensional integrals. As previously mentioned in Section 6.2, here we use a GP surrogate model to calculate the Sobol' sensitivity indices by using Equation (6.11), the GP predictive mean, to analytically compute the integrals. Therefore, we have calculated the Sobol' indices of the predicted value so that $y = \bar{f}(\mathbf{x})$, in a similar way to that from Chen et al. (2005b).

6.4 Results

This section presents the results of the GSA of the LFP TR as described in the previous section. The case study considers a commercial 1500 mAh 18650 3.2V $LiFePO_4$ cell, and begins with

an analysis of the full order model results providing further understanding of the distribution of the data. Following this, the analysis is split into two sections, first the time-dependent temperature analysis and then the TR features analysis. Both sections include cross-validation of a GP surrogate model to assess it's predictive accuracy and reliability. A full SA is presented including both the first order and the total Sobol' indices in an attempt to understand the thermo-physical and heat transfer characteristics that govern TR. Finally, the GSA is efficiently analysed by comparing the Sobol' sensitivity indices for the inputs to further understand which properties govern the TR process. In particular, the research is focused on understanding the temperature-time profile during TR and so the maximum temperature reached (the TR severity) is detailed as much as the time to TR.

6.4.1 Full Order Model

Figure 6.2 shows the predicted cell surface temperature when exposing an LFP cell to an oven temperature of $218^{\circ}C$ when simulating using the estimated abuse parameters from Table 6.2 and the reference values for the thermo-physical and heat transfer characteristics shown in Table 6.1. The simulated fit is compared to an unpublished experimental dataset produced using an oven test that follows the methodology from Bugryniec et al. (2019). The simulation produces the general cell TR behaviour well, while specific details such as time to TR and TR severity are not accurately predicted. However, in the context of this work, utilising a TR model to produce a data set for SA and to produce more general comments for TR model development, the model is deemed appropriate for the remainder of this work.

Figure 6.3 shows the mean cell surface temperature of all 753 runs varying with time, alongside the original (reference) temperature profile for comparison. Also shown are the upper and lower bounds of the simulation results, i.e the mean plus and minus 2 standard deviations (STDs) respectively, while the value of the STD with time is also plotted. The difference between the mean and reference plots, as well as the magnitude and variation of the STD show how the predicted cell temperature is effected by simply changing the thermo-physical and heat transfer characteristics. Figure 6.3 highlights what is referred to as the "TR event", importantly

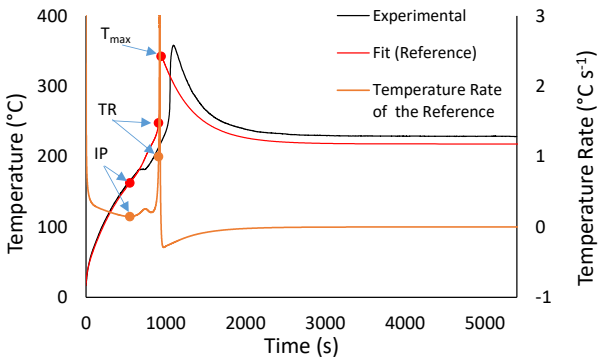


Figure 6.2: Full order model prediction of an LFP cell TR event due to oven exposure at $218^{\circ}C$ and compared to experimental results.

both before and after the TR event the mean results of the temperature profile are close to the corresponding reference values. During the TR event, the mean values differ from the reference plots, specifically the onset, rate and maximum temperature of TR. This difference is due to the TR event occurring at different times in each simulation run. Therefore, where the peak temperature occurs for the reference simulation, for other simulations at this point in time the temperature may be significantly lower as TR may not have started yet. This causes the overall mean temperature to have a lower maximum and shallower gradient during the TR event, rather than following a steep increase in temperature that is present in each individual simulation. Also, by looking at the STD, we can see that it increases significantly during the TR period, whilst in general the STD is greater for steeper temperature rates. This increase in STD causes a sudden drop in the lower bound of the data, even though for any of the 753 runs, a decrease in temperature similar to the dip in the lower bound is not possible.

6.4.2 Time-Dependent Temperature Analysis

GP Validation

The GP surrogate was used to predict the temperature at given times using the five thermo-physical and heat transfer characteristics as inputs to the GP. Therefore, $10s$ time steps were used up to the TR event and then, to capture the process accurately $2s$ time steps were used between $800s$ and $1200s$. Hence, the GP surrogate model consisted of 361 independent GPs, predicting the temperature at each time step. To be able to confidently calculate the Sobol'

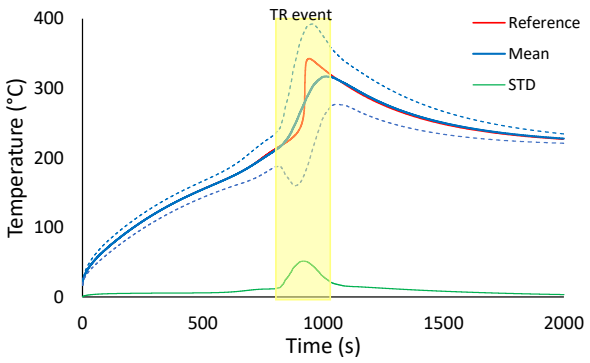


Figure 6.3: Mean results of all 753 oven simulations, each with randomly sampled thermo-physical characteristics, and compared to the reference simulation.

indices using a surrogate model the predictions produced by the GPs must be fully validated against the oven simulations to ensure inaccuracies are not inherently produced in the GSA. For this reason, the GP surrogate model was tested using the 5-fold cross-validation technique (Hastie, 2009) and the final results are as follows:

- Figure 6.4a presents the residuals, a comparison of the true values with the predicted mean values for each test prediction. A "high" predictive quality is indicated by the data following closely to the red $y = x$ trend line and a high coefficient of determination value of 0.969 (Marrel et al., 2008; Rohmer and Foerster, 2011).

- Figure 6.4b shows the test predictions as a function of time by considering the root mean squared error indicating the predictions have negligible errors up to and after the TR event. Between $850s$ and $1150s$, the RMSE increases to a maximum of 0.400.

- Figure 6.4c considers the predictive distribution by looking at the percentage of outliers at 2 STD's with respect to time. These outliers include any test prediction where it's true standardised value is outside of the predictions 95% uncertainty distribution. Again, before and after the TR event the number of outliers is negligible but during the TR event the percentage of outliers increase to be between 4% and 12%. Figure 6.4c also shows the percentage of predictions where the true test value is further than 4 predictive STD's away from the predicted mean.

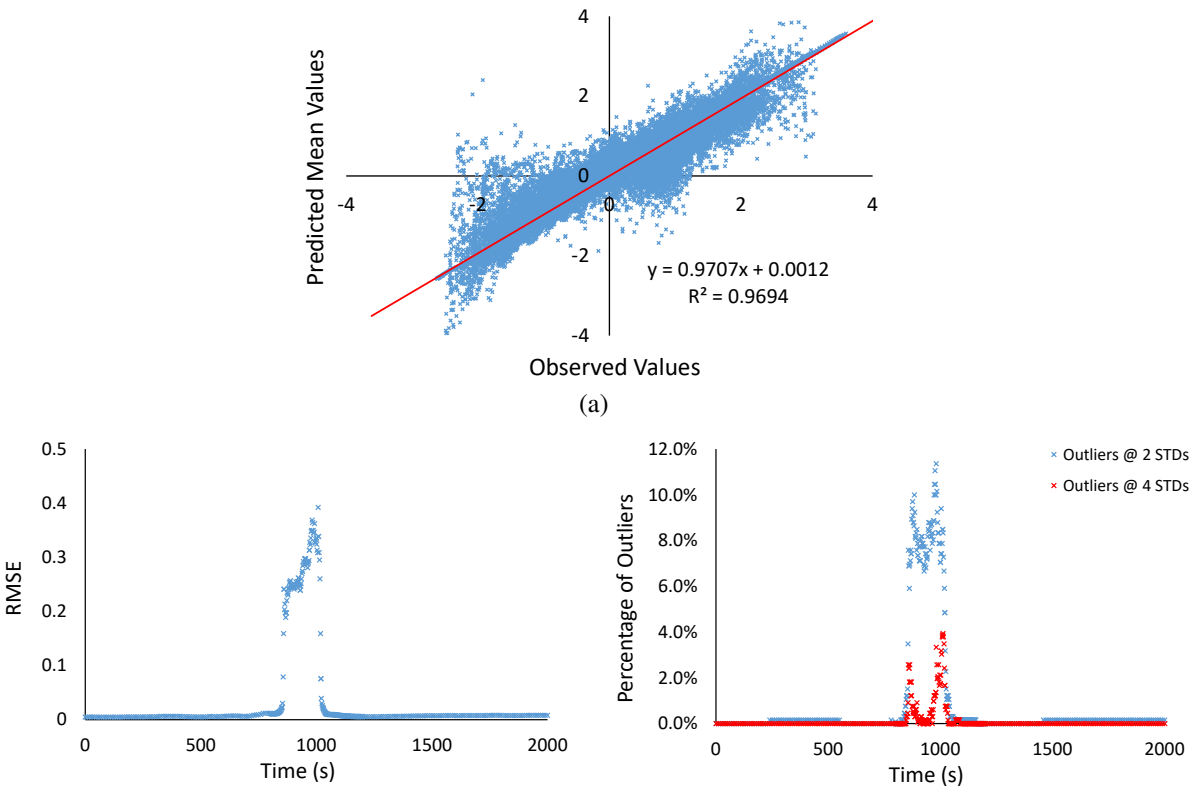


Figure 6.4: The diagnostics comparing the true standardised temperature to the standardised test predictions used to validate the time-dependent GPs using 5-fold cross-validation: a) the residuals b) the RMSE, c) the outliers.

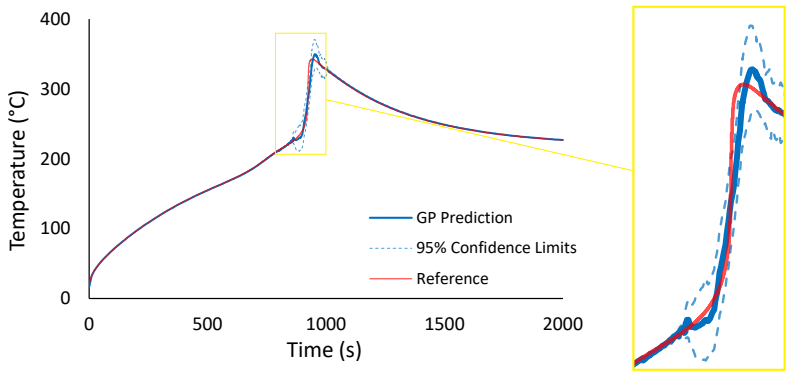


Figure 6.5: The predicted temperature-time profile from the GPs when the inputs are kept constant at the mean values.

Figure 6.5 shows the variation in temperature with time as predicted by the GP surrogate model using the mean input values. The temperature-time profile allows a direct comparison of the reference oven simulation to the predicted distribution which includes the mean predicted temperature (shown by the solid blue line) and the 95% confidence limits (shown by the dashed blue lines). At first, the accuracy of the GP is clear to see as the blue line closely follows the red reference line. Additionally, the uncertainty in the predictions is very small shown by the 95% confidence limits being barely distinguishable lines.

Overall, the test predictions have been analysed in detail by comparing the residuals, the diagnostics as a function of time and by fitting a GP to compare with the reference model. This has made it clear to see the predictions using the GP surrogate model are satisfactory and so we can confidently calculate the Sobol' indices.

Sobol' Indices

The Sobol' indices for each of the five inputs as a function of time, resulting from the GSA, are presented in Figure 6.6. Also shown in the figure is the mean temperature which allows us to study the dynamics of the system. Figure 6.6a shows the first-order Sobol' indices, including the interactions between all the inputs. Figure 6.6b shows the total Sobol' indices for each input. As can be seen in both figures, the surface temperature is first dominated by the conductivity coefficient. Hence, during the first instances of the cell being heated externally, it is the cells' ability to transfer heat to the bulk of the mass that has the most dominant effects on surface

temperature. In other words, a higher/ lower conductivity coefficient will mean that heat from the cells surface can be more/ less easily transferred through the cell to the bulk, leading to relatively lower/ higher surface temperatures, respectively.

From the beginning of the oven simulation to $500s$ the cells' temperature increases and so does the sensitivity measure of the emissivity and convection coefficient increase, with a corresponding decrease in the sensitivity measure of the conductivity coefficient. At $120s$ the sum of the sensitivity measures of the emissivity and convection coefficient equal that of the conductivity. Hence, at this temperature, we can say that the total heat exchange from the environment is the most important factor in cell temperature prediction. Beyond this time, the Sobol' indices of the emissivity coefficient continues to increase with further reduction in the Sobol' indices of the conductivity coefficient, while the Sobol' indices of the convection coefficient remain relatively constant. At $130s$ the emissivity becomes the most dominant factor in the overall uncertainty in cell temperature predictions, continuing up until the TR event reaching a maximum of 79.6% of the total output variance at $580s$.

Between $600s$ and $800s$, it can be seen from Figure 6.6a that the Sobol' indices of the conductivity coefficient increases. When comparing this to the temperature plot in the same graph, it can be seen that this corresponds to an increase in the slope of the temperature. The increase in the slope of the temperature is attributed to the increase in self-heating, as all else remains constant. Hence, the greater importance of the value of thermal conductivity shows the governing heat transfer to be the transfer of the decomposition heat from the jelly roll to the surface.

It can clearly be seen in Figure 6.6a that before $800s$, i.e. the start of TR, interactions between inputs are not involved in the process. Whereas, between $800s$ and $1100s$, we can see from Figure 6.6a that the interactions have a much larger effect on the output temperature. From Figure 6.6b we can see that these noticeable interactions must be between the heat capacity, conductivity coefficient, convection coefficient and emissivity, as the Sobol' indices of these inputs increase by a large amount from first-order Sobol' indices to total Sobol' indices. When the majority of the oven simulations have reached the peak of TR at around $1050s$ (as shown

by the mean temperature of the data) it can be seen that the total Sobol' indices of emissivity are large, almost reaching 1. However, at this same stage the first-order Sobol' indices of emissivity are below 0.3, showing that emissivity's interactions have a large effect on the output temperature when the TR event is reaching its finale. An increase from first order Sobol' indices to the total Sobol' indices for the conductivity coefficient and the convective coefficient show that the interactions must be between emissivity and these two thermal properties. Whereas, in comparison, the heat capacity and the density have a small increase in Sobol' values from first order to total. Hence, additional to the need to characterise emissivity well, as mentioned previously, it is shown that the convection coefficient and conductivity coefficient need also be characterised well to enable confident predictions of the TR event. Whereas, the results indicate that the density and heat capacity value could be estimated, knowing confidently that the values should have zero effect on the outcome of the oven simulation.

As shall be discussed in Section 6.5, calculating Sobol' indices for dependent inputs may create inaccuracies in the results as the method relies on independent variables Mara and Tarantola (2012). In spite of this, as the calculated Sobol' indices for both density and heat capacity are very small, we can assume that this is negligible.

Another important consideration is the high confidence in the experimental results which could cause the Sobol' indices for the density input variable to be small. This level of confidence in experiments has caused the relative standard deviation (RSTD) of density to be much smaller than the other input variables and so the variability of the density in the oven simulations is much smaller. This, along with the fact that the RSTD of density is less than a quarter to the RSTD of its dependent variable, the heat capacity, ensures that the small changes in density do not affect the output of the simulations. It should also be noted that the confidence of the heat capacity is relatively high. However, reported values in the literature between 18650 cells for the heat capacity vary greatly, and hence would increase the variance in the results if the C_p had a STD that encompassed the literature range of values.

Once TR has completed and the battery begins to cool, the variation in the Sobol' indices decreases as the cross effects become negligible. After this period the battery cools towards

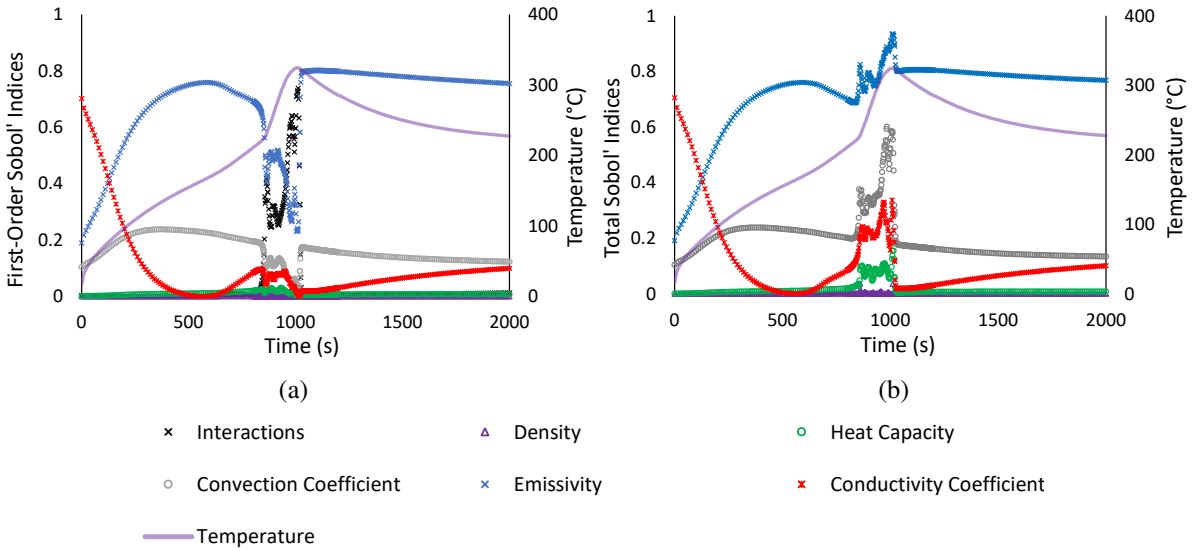


Figure 6.6: The Sobol' indices for each input as a function of time: a) shows the first-order Sobol' indices and the interactions, b) shows the total Sobol' indices.

the oven temperature and so the relevance of emissivity and the convection coefficient begin to slowly reduce linearly. Whereas, the conductivity coefficient starts to increase again, due to the dominant heat transfer from inside the cell to its surface and the environment, from almost zero relevance straight after the TR event to joining the convection coefficient's value of Sobol' indices. Nonetheless, it is clear that the most dominant thermo-physical and heat transfer characteristic is the emissivity, reducing from 0.81 at the time the TR event is over to 0.77 to the time the oven simulation has finished at $2000s$. In agreement with this result, the importance of radiation in TR of Li-ion cells has has also been shown experimentally (Hatchard et al., 2000; Chen et al., 2006).

6.4.3 TR Features Analysis

GP Validation

The time-dependent data showed emissivity to be the most dominant thermo-physical and heat transfer characteristic throughout the oven simulation, nevertheless, it is important to understand how this effects the key steps in a TR event. Consequently, after looking at the time-series data, the research shall now build on this knowledge, further understanding how the input variables have an impact on the three important TR features.

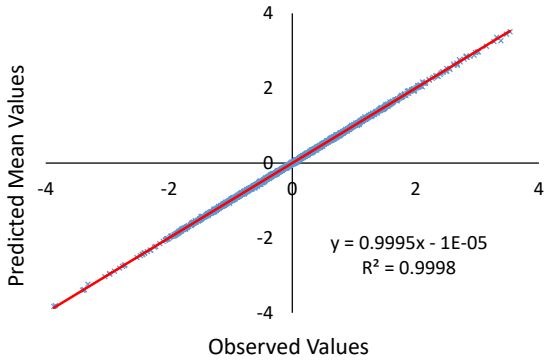


Figure 6.7: The residuals showing the observed standardised values against the predicted standardised values for all of the oven simulation test predictions.

The time-dependent data showed emissivity to be the most dominant thermo-physical and heat transfer characteristic throughout the oven simulation furthering our understanding of the TR modelling process. Additionally, a similar analysis was used to assist the time-series analysis, to consider how the input variables have an impact on three important TR features. We therefore perform a further GSA on the following features:

- Self-heating onset - defined as the point in time at which self-heating becomes dominant. Calculated as the time at which the rate of change in temperature goes from a negative slope to a positive slope, i.e the inflection point of temperature rate, marked "IP" on the temperature rate plot of Figure 6.2.

- TR onset time - the time at which the TR event begins. It is defined and calculated as the time at which the temperature rate goes above $1^{\circ}Cs^{-1}$ after the initial 500 seconds of heating, marked "TR" on the temperature rate plot of Figure 6.2

- Maximum temperature - marked as "T_{max}" on Figure 6.2.

Once again, a 5-fold cross validation of the GP outputs were performed just as before enabling a plot of the residuals shown in Figure 6.4a, comparing the observed values to the predicted mean values show the three GPs to have an excellent predictive quality, achieving a coefficient of determination value of 0.9998. Additionally, Table 6.4 shows the three outputs individually have low values close to zero for the outliers at 2 STD's and the RMSE.

Table 6.4: Resulting diagnostic values from the time prediction GPs

Output	**Outliers at 2 STD's**	**RMSE**
Inflection Point	0.266%	0.0225
TR Onset Time	0.133%	0.00763
Maximum Temperature	0.000%	0.00607

Sobol' Indices

The SA of the three important TR conditions were conducted to find which of the thermo-physical and heat transfer characteristics had the greatest effects on the time and severity of TR. The results are the total Sobol' indices calculated for each output can be seen in Figure 6.8 where the bars are split between the first-order and cross effects for the three different outputs. Figure 6.8a and Figure 6.8b clearly show the emissivity to be the most relevant thermo-physical and heat transfer characteristic for the important times dictating the TR event. Another interesting feature shown through the first two bar charts is that as the cell is first heated up to the inflection point, conductivity has little effect on the speed of the process. Although, for the TR onset time the total Sobol' indices for conductivity increases slightly. This infers that once self-heating of the battery becomes dominant (the inflection point) then conductivity has an increased effect on the TR. Once again, density and heat capacity have very little relevance on the output of the times during the TR process. Whereas, the TR severity has found the calculated Sobol' indices to show different relevance of importance for the input variables to all the previous results. The TR severity is described by the maximum temperature reached in the TR event and the total Sobol' indices calculated are shown in Figure 6.8c stating that the conductivity coefficient is the most dominant input, followed by the heat capacity and then the emissivity. In this case, both the density and the convection coefficient have no effect on the maximum temperature reached by the oven simulation. This shows the complexity involved in parametrising a TR model for Li-ion cells considering the whole process has previously shown to be dominated by the emissivity, it has now been shown that severity is dominated by a mixture of the three inputs. These results concur with the previous time-dependent Sobol' indices as the figures show very little red bars proving the interactions between inputs have small relevance leading up the TR events.

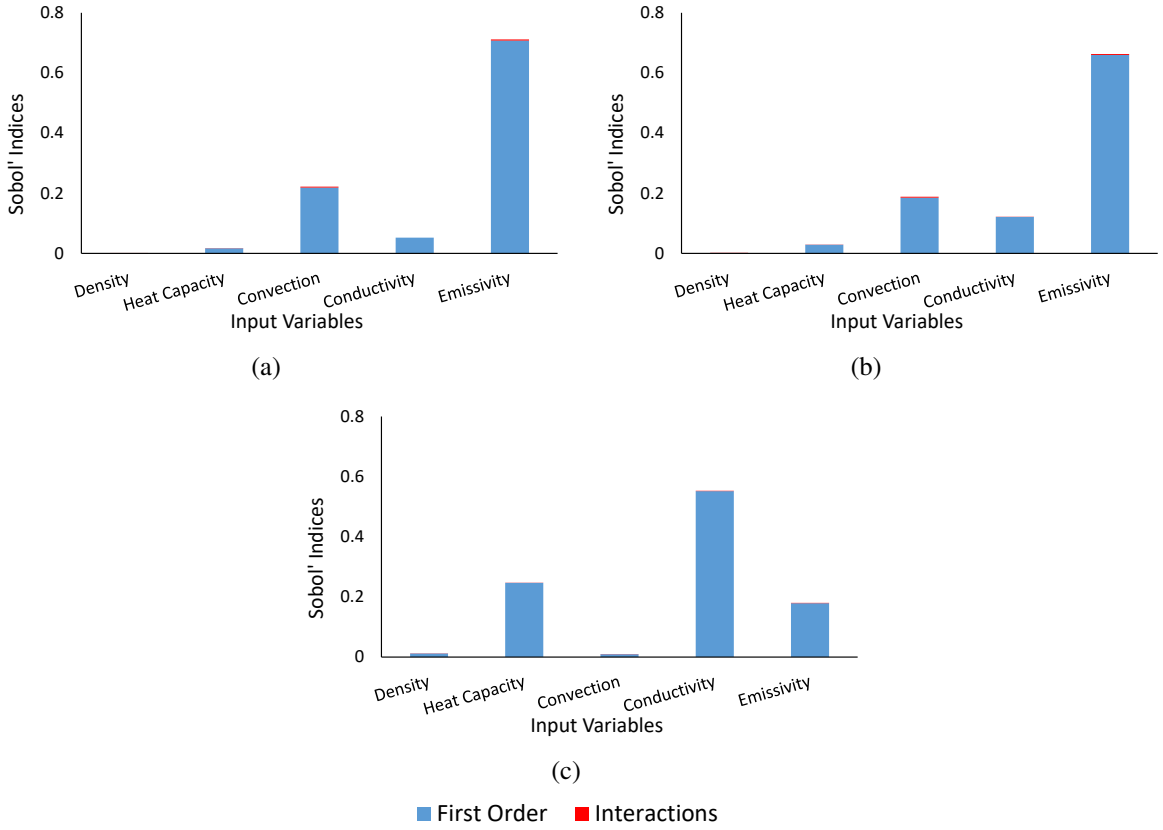


Figure 6.8: The total Sobol' indices for each input split to show the interactions and the first order Sobol' indices for each output: a) inflection point, b) TR onset time, and c) maximum temperature during TR.

To summarise, the SA outlines the importance of the values chosen for the development of TR models. It can be seen that estimating the emissivity value would not be adequate for a model as the sensitivity indices have shown that this dominates the whole TR process. Therefore, the SA has proven important in relation to Li-ion cell TR by providing vital information on the thermo-physical and heat transfer characteristics in respect to their effects on the temperature. From that, the research has also discovered that leading up to TR events the most dominant property is, again, emissivity as the inflection point and the TR onset time is considered. Although, when considering the severity of the TR event, analysing the maximum temperature reached by the oven simulations has found that the conductivity coefficient is the most dominant input. Additionally, heat capacity has zero impact on when TR occurs but it does have a say in the maximum temperature as the Sobol' indices for heat capacity increase with respect to the maximum temperature.

6.5 Conclusion

Modelling the thermal runaway (TR) of Li-ion cells is intrinsically as important as it is difficult due to the lack of understanding in the model parameters. A large number of parameters that govern the rise in temperature of a cell ensure the accuracy of simulation models difficult to obtain. Therefore, further understanding of some of the many properties was analysed in this research. Beginning with a GSA of the thermo-physical and heat transfer characteristics of a Li-ion cell provides guidelines into which parameters are the most important when designing the TR model. This was achieved by developing a TR abuse model for an LFP cell case study whereby all the parameters except the thermo-physical and heat transfer properties were kept constant to generate enough simulations for a variance-based GSA. With the use of a GP surrogate model, the Sobol' indices of the five inputs were calculated. Consequently, care was taken to validate the surrogate model ensuring inaccuracies were not carried onto the GSA.

The calculated Sobol' indices found that the most dominant parameter through the overall abuse scenario is clearly the emissivity value. This statement is true when considering the time-dependent Sobol' indices throughout the TR process as well as the Sobol' indices when

considering the two key time points as an output. However, when the severity of the TR event is considered, emissivity is no longer the dominant variable and now the conductivity coefficient has a much larger Sobol' indices. Hence, Li-ion safety will benefit by ensuring the radiation is controlled with care and when designing a TR model it is essential the emissivity is well characterised to avoid uncertainty in the model output. We have shown which parameters are key and at what point during TR, enabling other researchers to interpret their results better given their confidence in parameter values. The work carried out should ensure more accurate and robust TR models are developed by showing the important parameters to obtain. This further benefits the energy storage community as these uncertainties in the TR models are related to physical cells and so the relevant phenomena to consider for Li-ion safety are known.

As this model does not account of axial conductivity, which could lead to a greater temperature rate of the cell from increased heat transfer from the oven, reaction parameters may have to be altered to ensure accurate TR predictions. However, as the overall TR behaviour would be similar in any instance, one can assume the general dependences (of TR on parameters) would be similar. This is with the exception that the conductivity coefficient in the axial direction would have a greater influence on the model than the value in the radial direction. The GP SA techniques used here can be readily applied to higher dimensional models that include this extra heat transfer path but this is beyond the scope of this work.

For future work this study will be extended to the parameters governing the reaction kinetics and decomposition heat within the TR model, further developing the work begun by Milton et al. (2019) by understanding the effects each parameter has on the output. The GP model will be improved so that the uncertainty in predictions cover a wider range allowing a wider range of oven simulations to be used. Particularly, care will be taken in predictions for TR events that occur incredibly early/late so that the GP can successfully predict temperatures up to 10 STD's away from the mean at these early/late times. Another important aspect to consider is the range of input variables as this work gleaned the range of values from experimental work but an increase of 10% would reflect a larger variation of parameter values quoted in the literature. Furthermore, an unpublished analysis of the full order oven simulation has shown that a 10%

increase/decrease in heat capacity has a large effect on the resulting temperature-time graph. Whereas this research has shown that the range of heat capacity found from experimental work has very little significance on the resulting temperature profile. On a more practical level, it would be beneficial to consider the effects of all the model parameters together and apply the method for various other battery chemistries.

Chapter 7

Efficient Global Sensitivity-Based Model Calibration of a High-Shear Wet Granulation Process

7.1 Abstract

Model-driven design requires a well-calibrated model and therefore needs efficient workflows to achieve this. This efficiency can be achieved with the identification of the critical process parameters (CPPs) and the most impactful modelling parameters followed by a targeted experimental campaign to prioritise the calibration of these. To identify these parameters it is essential to perform a global sensitivity analysis (GSA).

Here, an efficient GSA is applied to a wet granulation case study with the Sobol' indices used to identify the CPPs and impactful modelling parameters. The population balance, mechanistic model that is used requires considerable computational effort for a GSA so a Gaussian Process surrogate is utilised to interrogate the underlying model. These key results reduce the input-space by 80% enabling the proposal of a targeted experimental design and model calibration workflow. This substantially improves the ability to deploy model-based design to determine the impactful parameter values, reducing the experimental effort by 42.1% compared

to a conventional experimental design.

7.1.1 Keywords

Gaussian Process; Sobol' Indices; Global Sensitivity Analysis; Model Calibration; Granulation; Experimental Design

7.2 Introduction

A common experimental design approach is the factorial design of experiments (DOE). In a factorial DOE, the experimental space is covered by identifying a high and a low level for every input factor (Montgomery and Runger, 2014). Every combination of input factor levels is tested experimentally and so, as a consequence, the experimental effort increases exponentially with the number of factors. In order to reduce the required experimental effort, the most critical process parameters (CPPs) should be identified *a priori* so only they are used as factors. However, the CPPs are commonly identified heuristically which can be unreliable if the process-specific experience is limited, e.g. due to recent process modifications or new formulations. Gaining experience through a rigorous experimental investigation of all process parameters requires a very high experimental effort. To identify the CPPs, a sensitivity analysis is proposed by using a predictive model that reduces the need for heuristics. However, models for particulate processes and product design are coupled with many parameters and degrees of freedom. Experimentally testing every combination of input variables, across 3 to 5 length scales makes designing particulate products a costly process in terms of time, money and materials. Consequently, researchers using experimental design methods overcame these issues using DOE and began developing more sophisticated methods such as the sequential approach. These methods are much more adaptive, as they offer a dynamic class of experimental design by incorporating system knowledge through progressive steps (Garud et al., 2017). However, these more sophisticated methods do not allow the opportunity to incorporate uncertainty using of a substantially reduced number of experiments. Therefore, this work focuses on the development of an application of a model-driven design workflow to identify the CPPs and thus reduce, and better target, the number of experiments to be performed using high shear wet granulation (HSWG) as a case study.

By applying a model-driven design workflow, the most important operating ranges are identified by evaluating model predictions to reduce the experimental effort further (Wang et al., 2019; Bellinghausen, 2020). The process model is based on a population balance modelling (PBM) framework. Mechanistic understanding of the rate processes is incorporated through appropriate kernels. An essential part of model-driven design is model calibration workflows, which typically involve a combination of designed characterisation tests, parameter estimation from lab-scale experiments and reasonable assumptions to determine all modelling parameters. It is important, therefore, to ensure that impactful modelling parameters are identified upfront, so that an efficient model calibration workflow can be applied (Bellinghausen, 2020). This is because the impactful parameters need to be determined more accurately, while a lower accuracy is acceptable for the remaining parameters. As such, by focussing on the impactful parameters, the experimental or computational effort to determine parameters is reduced advancing towards improved model predictions.

While there are a number of types of process models, in this study we will focus on PBM as it is the most frequently used method for particulate processes such as crystallisation (Costa et al., 2007; Sulttan and Rohani, 2019), polymerisation (Sood et al., 2016; Brunier et al., 2017), granulation (Meyer et al., 2015; Shirazian et al., 2019), milling (Kumar Akkisetty et al., 2010; Capece et al., 2011), and mixing (Sen et al., 2012; Boukouvala et al., 2012). PBM keeps track of particle properties over time using population balance equations as a process scaling approach. Often, models are validated by comparing simulation results to the same experimental results used for the parameter estimation study and so they are not critically assessed. For process design and scale up purposes it it essential that models are fully validated using data that has not been used to train the model (Chaudhury et al., 2014). Therefore, the need to develop efficient methods to reduce the cost of DOE for HSWG is constrained by the lack of understanding of both the modelling parameters and the experimental parameters. Hence, a model-driven design workflow that will identify and further the understanding of the CPPs is essential.

Critical modelling decisions in model-driven design need to be based on a good understanding of the process model. Additional insight is given by a sensitivity analysis that characterises

the relationship between the model's inputs and outputs. Thus, the uncertainty in the outcome can be apportioned to the different sources of uncertainty in the input. Sensitivity analysis has been implemented in many different research fields (for example Saltelli et al., 2005; Rohmer and Foerster, 2011; Al et al., 2019) as it is widely acknowledged as a good practise to better understand model behaviour. In particular, the use of a global sensitivity analysis (GSA) quantifies the variation of the model response in the entire parameter domain fully exploring the input space to identify the CPPs of interest. Sensitivity studies have been employed to advance particulate processes (Van Bockstal et al., 2018; Mortier et al., 2014) as it is an effective tool to rank and prioritise the process variables and the modelling parameters helping to focus experimental and model calibration efforts. Prioritising CPPs allow a reduction from a high dimensional model, hence further tools such as optimisation (Wang et al., 2017) can be applied to aid the development of particulate processing for design.

Focussing on granulation, sensitivity analysis studies have been applied to help determine the most important factors affecting the granulation output. For instance, Cryer and Scherer (2003) conducted a sensitivity analysis on fluid-bed granulation, beginning with a $1/2$ fraction factorial statistical DOE. The research determined the important factors affecting the granule size to be bed bowl charge, binder spray rate, air flow rate, and input air temperature. Then a PBM model sensitivity analysis was used to further the understanding available from the experimental DOE alone, suggesting 65% of the predicted variance is accounted for from the binder spray rate. Similarly, Metta et al. (2019) applied a GSA to further understand the complex interplay between process wide CPPs and critical quality attributes using a flowsheet model. Wet granulation was included in the continuous tablet manufacturing process and modelled using PBM in the model flowsheet. The GSA was achieved using the Morris method and the variance based method of Sobol' indices, both agreeing that the liquid feed rate to the granulator to be a CPP which affects the tablet properties. Interestingly, the work also showed significantly less samples were needed for the Morris method but it did not provide the detailed and quantitative information that Sobol' indices does. This research successfully identifies the CPPs throughout the continuous tablet manufacturing process but these studies were constrained by a computa-

tionally expensive flowsheet. The existing research has many problems in representing a high resolution for the twin screw wet granulation model, predicting only steady state outputs.

Therefore, this work will address the shortcomings from previous studies by concentrating on the HSWG case study as a single unit operation model. Although there are many sensitivity analysis techniques available (such as the Morris method), this work requires a detailed understanding of the importance of each parameter and their interactions. Consequently, this work will use a variance-based decomposition GSA method known as the Sobol' sensitivity indices which are considered the benchmark for GSA methods (Sumner et al., 2012; Kontoravdi et al., 2005; Xie et al., 2019). However, as noted by Metta et al. (2019), the calculations require significant amounts of data to ensure convergence of integrals to a satisfactory precision level. Due to the complexity of the underlying model, the traditional method of computation using Monte Carlo techniques (Kucherenko et al., 2009) are computationally impracticable for this research. Various surrogate modelling techniques reduce the computational burden by allowing the indirect interrogation of the model so that a significant reduction in the number of simulations can be used to evaluate the high-dimensional integrals. Examples of surrogate modelling techniques include polynomial chaos expansion (Brown et al., 2013; Sudret, 2008), artificial neural networks (Li et al., 2016), and Gaussian Processes (GPs) (Marrel et al., 2009; Yeardley et al., 2020a,b). Here, based on our previous work, the surrogate model will be developed using GPs as they are a widely used tool for Bayesian nonlinear regression and provide an approach that predicts a distribution allowing for uncertainties for each prediction. Furthermore, GP regression (also known as Kriging) has been proven for predictive modelling of pharmaceutical processes (Jia et al., 2009; Boukouvala et al., 2010). The novel approach will allow us to:

- identify of the most impactful modelling parameters for the HSWG process model,
- understand which of the operating parameters for the process are the CPPs,
- determine the parameters that will have a sufficient impact to influence the DOE independently and from interactions,
- investigate the more beneficial production conditions and a reduction in experimental ef-

fort.

These goals are achieved by developing a population balance based HSWG model that is directly emulated by a GP surrogate model enabling a reduction in the computational effort required to analytically calculate the Sobol' indices for a GSA.

7.3 Modelling Tools

To perform the GSA of the HSWG, we use PBM to model the HSWG, then GP regression acts as a surrogate model allowing the calculation of Sobol' indices. This section will describe the background behind each modelling tool. The nomenclature of the mathematical notation shown in the paper follows that of standard mathematics, where a bold lower case variable represents a vector and a bold upper case variable represents a matrix.

7.3.1 High-Shear Wet Granulation Process Model

This section summarises the process model for HSWG proposed by (Bellinghausen, 2020), as shown by the process schematic in Figure 7.1. The model is based on a 1-D, 1-compartment population balance modelling (PBM) framework (Ramkrishna and Mahoney, 2002) applying the lumped parameter approach (Hounslow et al., 2001). The particle size distribution is tracked by determining the particle number density n for every size bin i over time t:

$$\frac{\partial V n(v,t)}{\partial t} + \frac{\partial}{\partial v}\left[V n(v,t)\left(\dot{G}_{lay} + \dot{G}_{cons}\right)\right] = V\left[\dot{b}_{nuc}(v) + \dot{b}_{coal}(v) + \dot{b}_{br}(v) - \dot{d}_{coal}(v) - \dot{d}_{br}(v)\right] \quad (7.1)$$

where n is the volume-specific number density of particles, v is the particle volume, V is the control volume, $\dot{G}_{lay}$ and $\dot{G}_{cons}$ are the rate of change due to layering and consolidation

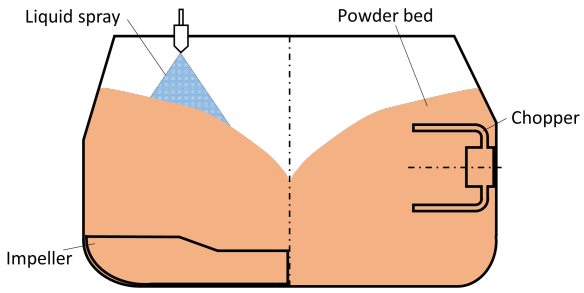


Figure 7.1: HSWG process schematic

respectively, $\dot{b}_{nuc}$, $\dot{b}_{coal}$ and $\dot{b}_{br}$ are the birth rates due to nucleation, coalescence and breakage, and $\dot{d}_{coal}$ and $\dot{d}_{br}$ are the death rates due to coalescence and breakage. Rate expressions for nucleation, consolidation, coalescence and breakage are incorporated in the model using inputs that determine the effect of each of these rate processes on the particle size. These rate expressions and underlying modelling assumptions are described in detail by (Bellinghausen, 2020). The model is implemented in gFormulate v1.5 (Process Systems Enterprise Ltd.).

7.3.2 Gaussian Process (GP) Regression

GP regression is used as a surrogate model for the PBM. This non-parametric machine learning technique enables direct interrogation of the PBM to be replaced by a reduced model encapsulating the systems behaviour in a cheaper, simpler framework.

The GP takes a $(1 \times d)$ row vector of inputs $\mathbf{x}$ and returns a Gaussian random variable through calculations using the predictive equations shown in Yeardley et al. (2020a). At the heart of this lies the kernel function $k\colon \mathbb{R}^{i+d} \times \mathbb{R}^{j+d} \to \mathbb{R}^{i} \times \mathbb{R}^{j}$, expressing the correlation between responses to input samples of sizes $(i \times d)$ and $(j \times d)$. This work exclusively uses the automatic relevance determination (ARD) kernel (Wipf and Nagarajan, 2007):

$$k(\mathbf{x}', \mathbf{x}) \coloneqq \sigma_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')\Lambda^{-2}(\mathbf{x}-\mathbf{x}')^\intercal}{2}\right) \tag{7.2}$$

where Λ is a $(d \times d)$ diagonal positive definite lengthscale matrix. The GP surrogate model is learnt from mapping the training inputs $\mathbf{X}$ to the observed responses $\mathbf{y}$, assuming the training data takes the form $\mathbf{y} = f(\mathbf{X}) + \mathbf{e}$ where $\mathbf{e}$ is an independent and identically distributed random error term. Regression uses the learned model to make predictions and so requires the optimisation of $d+2$ hyperparameters, constituting of Λ, σ_f, and σ_e, through the maximum marginal likelihood $p[\mathbf{y}|\mathbf{X}]$ using the ROMCOMMA software library (ROMCOMMA, 2019).

7.3.3 Sobol' Indices

This work implements the use of the variance based GSA using a GP surrogate model to calculate the Sobol' Indices. The calculation of the Sobol' indices follows Jin et al. (2004) by substituting the true simulation model with the mean of the conditional GP resulting in semi-

analytic Sobol' indices.

Sobol' indices are calculated as a variance-based GSA method, which describes how the variance of the output can be decomposed into terms that are dependent on the input factors (Iooss and Lemaître, 2015). Each input has two Sobol' indices calculated (Sobol, 1993, 2001). The first-order Sobol' index, S_i measures the contribution to the variance solely attributable to x_i. Whereas, the total Sobol' index, S_i^T expresses the whole effect of an input on the output, including interactions with all other inputs (Saltelli and Homma, 1996). Thus, the effect that interactions of an input variable has with other inputs on the variance on the output is calculated by the difference between S_i^T and S_i. Both indices are interpreted as the amount of variance ascribable by x_i and so the larger the value of the index, the more influential the input factor is.

The calculation of Sobol' indices are determined through a decomposition method presented by Sobol (1993), which evaluates each term through multidimensional integrals which requires a large sampling cost (Saltelli et al., 2008). Many alternative designs to calculate an estimation of Sobol' indices have been derived (Marrel et al., 2009; McKay et al., 2000). However, in this work we have calculated the Sobol' indices of the predicted value from a GP so that $y = \bar{f}(\mathbf{x})$, as shown by the mathematical details described in (Yeardley et al., 2020a).

7.4 Methodology

A suitable workflow was produced by applying a model-driven design (see Figure 7.2). GSA is the key to the model calibration and so both the modelling parameters and operating parameters undergo a GSA.

Figure 7.2 shows both types of parameters as separate pathways and so variations can be seen in both. Initially, modelling parameters may be known from physics or literature so if they do, those values can be used. If not a GSA is needed to discover the most impactful modelling parameters. Whereas, it is always important to identify the CPPs within the operating parameters straight away using a GSA. Figure 7.2 presents the GSA methodology as one box for both the modelling parameters and the operating parameters because they are both implemented with the same method. After both GSA's a criterion is applied to distinguish between parameters

with low or high impact dependent on the parameters Sobol' index. Determining which category each modelling and operating parameters falls into then helps determine the final steps for the model calibration. These final steps help determine the procedures necessary to efficiently calibrate the process model.

7.4.1 Parameter Identification and Sampling

For a GSA, representative model outputs need to be selected to assess the process performance. Particle size distribution and porosity are the most relevant product properties of wet granulation (Mittal, 2017). For this reason, the three properties of the particle size distribution, mass-median particle diameter $D50$, fines mass fraction $W\,(< 90\,\mu\mathrm{m})$ and coarse mass fraction $W\,(> 1\,\mathrm{mm})$, are selected as well as granule porosity ε.

The HSWG process model has been experimentally verified by Bellinghausen (2020) through an assessment of modelling assumptions and a comparison of model results to experimental trends. Individual wet granulation rate expressions used have been experimentally validated previously (Pohlman and Litster, 2015; Davis, 2016; Sayin, 2016; Bellinghausen et al., 2019). The results confirmed that the model is accurate across a wide range of conditions and process scales. In additional to the parameter estimation, model selection or validation is required to ensure the model is capable of predicting experimental results.

Therefore, initially a GSA is conducted as a screening tool using model runs with different combinations of modelling parameters. Then the model is run with different combinations of operating parameters for the second GSA. The parameter values for the process model simulations were produced using Quasi-random (Sobol') sampling with a normal distribution so that the input variables were based on reasonable uncertainty ranges. The GSA tool of gFormulate v1.5 is used for sampling and simulations. The modelling parameters and the rate process that they control are shown in Table 7.1. Additionally, their mean value and their standard deviations (STD) are listed with reasoning for each uncertainty range.

Table 7.2 shows the operating parameters mean and STD used for sampling the $10\,\mathrm{L}$ high-shear mixer filled with $2\,\mathrm{kg}$ of dry powder case study. Engineering rules for high-shear wet

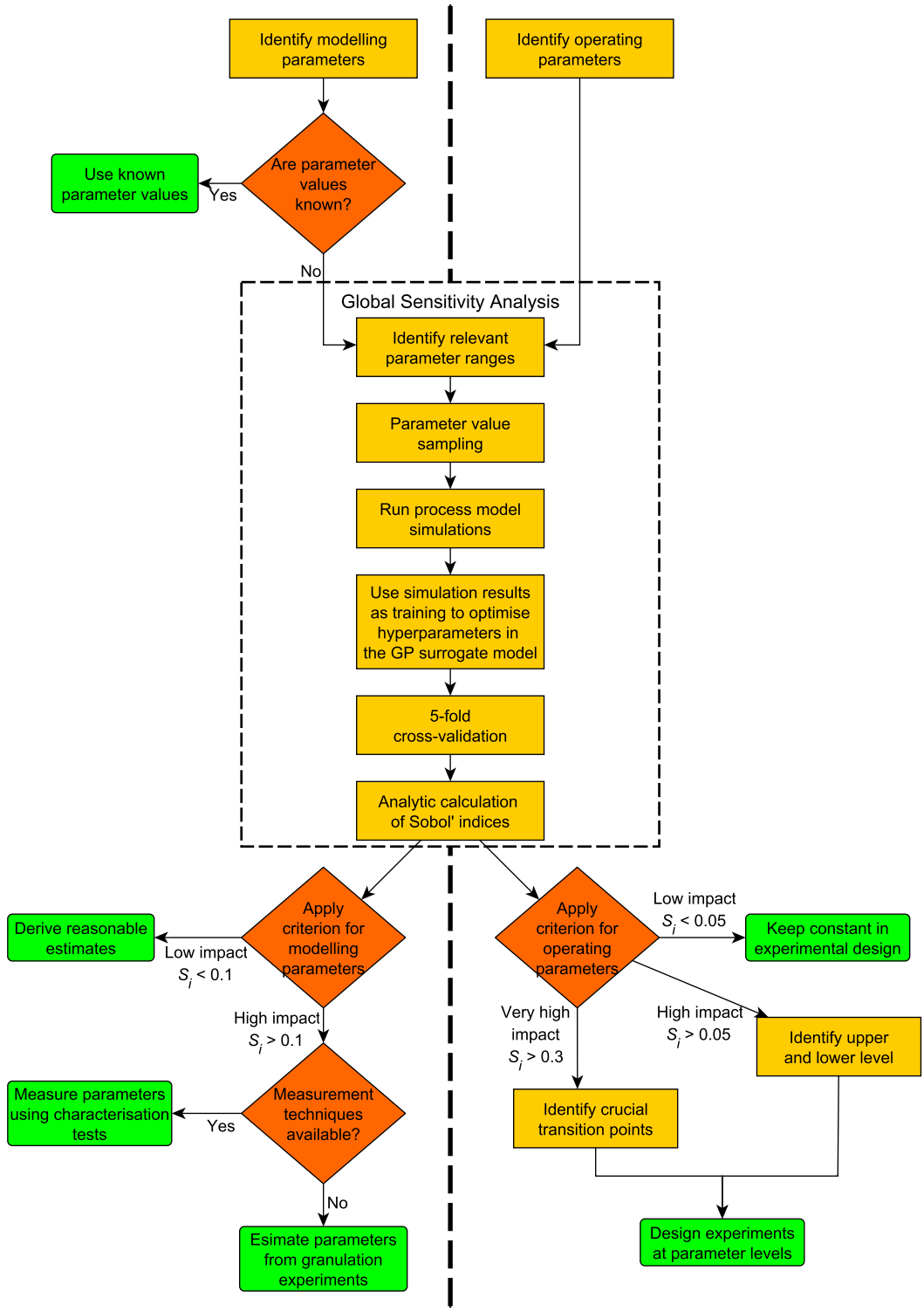


Figure 7.2: Model-driven design workflow using GSA with Sobol' indices criteria.

Table 7.1: The input parameters used for the Modelling Parameter GSA

Rate Process	Modelling Parameter	Mean	STD	Source
Nucleation	Nuclei-to-Drop Diameter Ratio K_d $[-]$	1.4	0.1	Hapgood et al. (2009)
Consolidation	Consolidation Coefficient k_{cons} $[-]$	1.5	0.25	Assumption
Consolidation/Coalescence	Dynamic Contact Angle θ [rad]	0.2	0.1	Pohlman and Litster (2015)
	Strength Parameter A $[-]$	7.0	0.5	Smith (2007)
	Strength Parameter B $[-]$	220	9	Smith (2007)
	Strength Parameter n $[-]$	0.59	0.01	Smith (2007)
	Surface Tension γ^{lv} $\left[\mathrm{mN\,m^{-1}}\right]$	73.9	0.1	Pallas and Harrison (1990)
Coalescence	Critical Pore Saturation S_{crit} $[-]$	0.16	0.01	Assumption
	Collision Coefficient $k_{I/II}$ $\left[\log_{10}\mu\mathrm{m}^{1.5}\right]$	13.5	0.5	Pohlman (2015)
	Elasticity Parameter k_E $[-]$	24.9	1.8	Pohlman and Litster (2015)
	Elasticity Parameter p_E $[-]$	0.17	0.01	Pohlman and Litster (2015)
	Elasticity Parameter q_E $[-]$	-6.9	0.6	Pohlman and Litster (2015)
	Elasticity Parameter r_E $[-]$	-1.50	0.015	Pohlman and Litster (2015)
	Height of Asperities h_a [µm]	1.51	0.2	Kastner et al. (2013)
	Poisson Ratio ν $[-]$	0.03	0.002	Pohlman and Litster (2015)
Breakage	Impact Energy $E_{m,kin}$ $\left[\mathrm{J\,kg^{-1}}\right]$	1000	50	Meier et al. (2008)
	Material Strength Parameter f_{Mat} $\left[\mathrm{kg\,m\,J^{-1}}\right]$	1.0	0.1	Meier et al. (2008)
	Breakage Coefficient k_{br} $[-]$	0.0035	0.001	Assumption
	Minimum Fragment Diameter $d_{j,min}$ [µm]	15	5	Vogel and Peukert (2005)
	Power Law Exponent q $[-]$	1.2	0.1	Assumption

Table 7.2: The input parameters used for the Operating Parameter GSA

Operating Parameter	Mean	STD	Design approach
Liquid to Solid Ratio L/S $\left[\mathrm{kg\,kg^{-1}}\right]$	0.17	0.015	Induction growth regime (Iveson et al., 2001)
Impeller Frequency n_{imp} $\left[\mathrm{min^{-1}}\right]$	321	32	Roping flow regime (Tran, 2015)
Kneading Time t_{kn} [s]	239	85	Induction time (Iveson et al., 2001)
Liquid Spray Rate $\dot{V}$ $\left[\mathrm{g\,min^{-1}}\right]$	71	9	Drop-controlled regime (Hapgood et al., 2003)

granulation were applied to derive the operating parameter ranges which are specific to the process scale investigated.

7.4.2 Application of GP Surrogate Modelling for the Calculation of Sobol' Indices

GP regression is used to encapsulate the HSWG process model in a much simpler framework enabling the semi-analytic calculation of Sobol' indices as shown by the modelling tools in Section 7.3. In this case study, the process model has four selected outputs and so the GP surrogate model captures each output as four independent GP regression predictive tools. In the first surrogate model twenty modelling parameters are used as a (1×20) row vector of inputs to predict the distribution of the four process outputs. Similarly, the four operating parameters are used in a second surrogate model. These predictions are made using a formula that is learnt from training data produced through sampling. GP regression involves optimising the

hyperparameters using the ROMCOMMA software library (ROMCOMMA, 2019). From there, the predictive equation is used to calculate Sobol' indices but first cross-validation ensures inaccuracies in the predictive capabilities of the GP surrogate models are not carried through to the GSA. Therefore, this research tested the GP surrogate model using 5-fold cross-validation, whereby, the training data is randomly split into five subsets. Then four of the five subsets are used as training data, while the remaining is used to test the GP surrogate models predictions. This procedure is repeated so that every single sampling point produced is used for testing.

7.4.3 Parameter Criterion

The threshold for each GSA criteria will be set by the average total Sobol' index value for the four outputs, $\hat{S}_i^T$. Firstly, the modelling parameters GSA acts a screening that uses a threshold of $\hat{S}_i^T > 0.1$. Any modelling parameters that fail the criteria are not impactful and can have reasonable values from literature used.

Similarly, the threshold for the operating parameters GSA is set so that CPPs have a value of $\hat{S}_i^T > 0.05$. Then these CPPs are used in the experimental design with various levels dependent on how much of an impact each has. The focus of the experimental design should be on operating parameters with a very decisive impact ($\hat{S}_i^T > 0.3$). This is achieved by introducing multiple levels for these parameters. The operating parameters with $\hat{S}_i^T < 0.05$ are kept constant in the experimental design. The different thresholds for the modelling parameters GSA and operating parameters GSA are due to the difference in the number of variables. There are a total of twenty modelling parameters in comparison to the four operating parameters and so the threshold for the modelling parameter GSA needs to be larger ensuring an efficient reduction in parameters.

7.4.4 Experimental Design Considerations

Based on the GSA results, an experimental design for model calibration and validation is proposed. Impactful modelling parameters that can be directly derived from physical properties need to be determined by characterisation tests. The remaining impactful modelling parameters need to be estimated from experimental data. To reduce the number of experiments and un-

derstand the impact of the operating conditions, all CPPs are varied in the experimental design but the very impactful parameters ($\hat{S}_i^T > 0.3$) are studied in more depth through smaller step changes. This more rigorous investigation of the very impactful CPPs is also needed to accurately estimate and validate the more impactful modelling parameters.

7.5 Results

7.5.1 Modelling Parameter Screening

GP Validation

The GP surrogate model was used to predict the outputs of the HSWG model using the twenty modelling parameters shown in Table 7.1. To be able to confidently calculate the Sobol' indices using a surrogate model, the predictions produced by the GPs must be fully validated against the true simulations to ensure inaccuracies are not inherently produced in the GSA. The GP surrogate model was tested using the 5-fold cross-validation technique (Hastie, 2009) and the resulting diagnostic values are presented in Table 7.3. The residuals of each prediction are plotted in Figure 7.3, comparing the true values with the predicted mean values showing a correlation that follows the red $y = x$ trend line. This is further exemplified by the coefficient of determination (r^2) and the root mean squared error (RMSE) values presented in Table 7.3 for each output. Both diagnostics measure a skill score, corresponding to the accuracy the predicted mean (Al-Taweel, 2018), but with different scales. Further, the predictive distribution is analysed to ensure the the GPs are not predicting with over confidence. This is shown by counting the outliers for any prediction where it's true standardised value is outside of the predictions 95% uncertainty distribution. Table 7.3 presents the outliers of all four outputs to be between 8.80% and 11.5% inferring the residuals are not Gaussian as the outliers are greater than the expected 5%. This shows the GPs are overconfident due some the predicted STD's being to small making the true validation values outside the predicted distribution. The prediction results here are in good agreement, and have shown the GP surrogate model is sufficiently accurate for the purposes of this work.

The results from the cross-validation for the modelling parameters are satisfactory but do

Table 7.3: Resulting diagnostic values from the modelling parameters prediction GPs

Output	**Correlation Coefficient,** r^2	**RMSE**	**Outliers at** 2 **STD's**
D50	0.661	0.622	11.5%
Fines Fraction	0.886	0.340	8.80%
Coarse Fraction	0.741	0.509	10.1%
Granule Porosity	0.919	0.285	9.90%

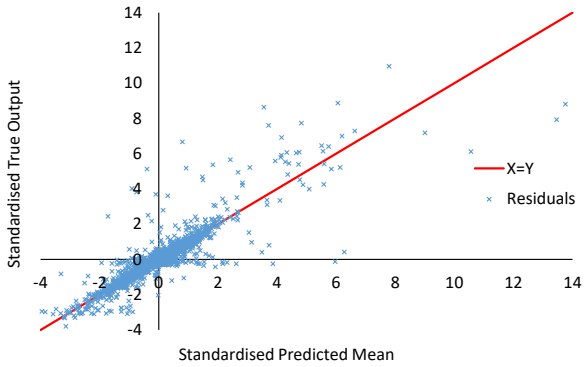


Figure 7.3: The residuals showing the observed standardised values against the predicted standardised values for all of the test predictions using cross-validation.

not show accuracy where the GPs predict exact test values. This is because of an issue with surrogate models known as the curse of dimensionality, where the surrogate requires an exponentially increasing number of output results throughout the input space as the number of parameters increases.

Sobol' Indices

The results are the total Sobol' indices for the modelling GSA calculated for each output are shown in in Figure 7.4 where each bar corresponds to the first-order Sobol' indices plus the remaining indices due to interactions with other input variables. The split in the bars are to illustrate the first-order value as the bottom and the interactions as the top. Instantly, it is clear to see in Figure 7.4 that interactions have a large effect on the four outputs. Additionally, each output has different input variables influencing it with great variability. For example, the collision coefficient produces over 80% of the total variance for D50 but only 1% of the variance for the granule porosity. Most importantly, the total Sobol' index values for only eight of the twenty modelling parameters are large enough to be shown clearly in Figure 7.4 and so the

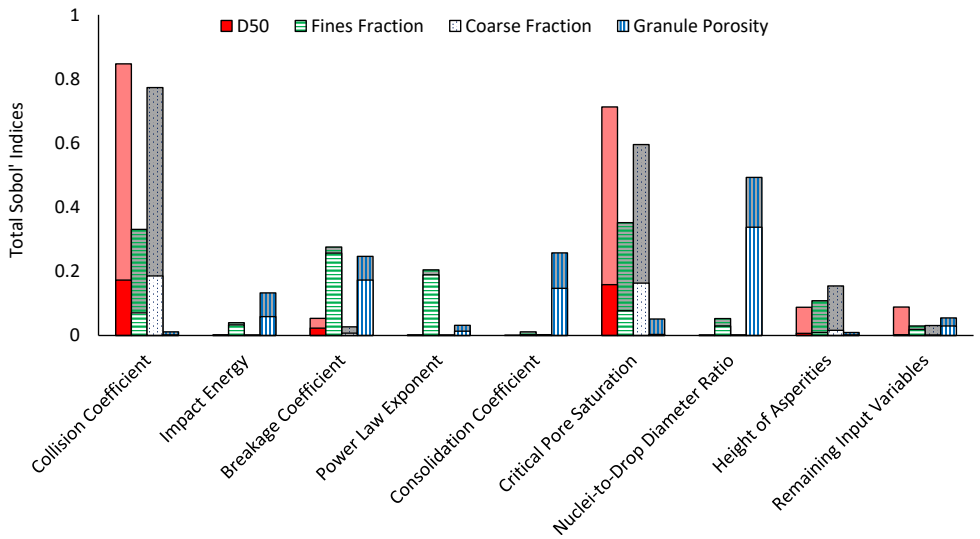


Figure 7.4: The total Sobol' indices for the modelling parameters with respect to each output of the HSWG model. Each bar is split to show the first-order Sobol' indices (bottom) and the interactions (top)

remaining are grouped into the "remaining input variables" category.

For each modelling parameter, further analysis calculating the average Sobol' index value per output ($\hat{S}_i^T$) shows which are the most impactful modelling parameters to be included in a parameter estimation. The criteria, set in Section 7.4, states the $\hat{S}_i^T > 0.1$ for i to be an impactful parameter. All the Sobol' index values shown in Figure 7.4 have there $\hat{S}_i^T$ presented in Table 7.4 clearly showing which output the modelling parameter affects the most by the colour of the shading.

Table 7.4 shows an $\hat{S}_i^T$ value greater than 0.1 for just four of the twenty modelling parameters. Therefore, they can be reduced so that in a characterisation test or a parameter estimation analysis the collision coefficient, breakage coefficient, critical pore saturation and nuclei-to-drop diameter ratio are estimated accurately. Whereas the remaining modelling parameters can be fixed using default values from literature recommended previously as the means used in the GSA sampling.

Interestingly, the two most impactful parameters, collision coefficient and critical pore saturation are the most impactful coalescence parameters followed by the height of asperities which has a $\hat{S}_i^T$ just below the threshold set. The coalescence rate is proportional to the collision coefficient and the critical pore saturation impacts the coalescence criterion. Thus, both impactful

Table 7.4: The impact each modelling parameter has on the four outputs where green is highly impactful and red is negligible. Summarised in the last column by the average Sobol' index value for each modelling parameter, $\hat{S}_i^T$

	D50 S_i^T	**Fines Fraction** S_i^T	**Coarse Fraction** S_i^T	**Granule Porosity** S_i^T	**Average** $\hat{S}_i^T$
Collision Coefficient					0.49
Impact Energy					0.04
Breakage Coefficient					0.15
Power Law Exponent					0.06
Consolidation Coefficient					0.07
Critical Pore Saturation					0.43
Nuclei-to-Drop Diameter Ratio					0.14
Height of Asperities					0.09
Remaining Input Variables					0.05

modelling parameters impact the three particle size properties the most. Whereas, the impact the collision coefficient and the critical pore saturation has on the granule porosity is negligible as shown in Table 7.4 by the red shading. The strong interaction between the two parameters shows that coalescence depends heavily on both parameters. Additionally, the high impact of both coalescence parameters provides evidence that coalescence is the most dominant rate process for particle size.

The breakage coefficient dominates the impact energy and the power law exponent as the most most impactful breakage parameter. Overall, the breakage coefficient has a high impact on the granule porosity but less of an impact on the D50 and coarse fraction showing that breakage has a negligible effect on coalescence. Yet, it has a high impact on fines fraction because the increase in breakage results in more fine particles.

The most impactful nucleation parameters is the nuclei-to-drop diameter ratio which determines the initial granule size and porosity. Due to the high impact of other rate processes like coalescence, the impact which nuclei-to-drop diameter has on size is negligible. Therefore, it mainly has a large impact on granule porosity.

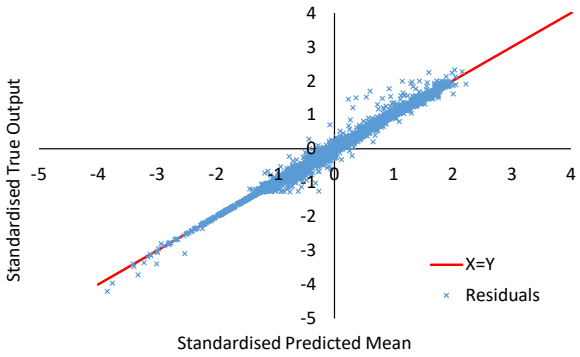


Figure 7.5: The residuals showing the observed standardised values against the predicted standardised values for all of the test predictions using cross-validation.

Table 7.5: Resulting diagnostic values from the operating parameters prediction GPs

Output	**Correlation Coefficient,** r^2	**RMSE**	**Outliers at** 2 **STD's**
D50	0.992	0.0869	6.00%
Fines Fraction	0.991	0.0959	6.05%
Coarse Fraction	0.996	0.0662	5.25%
Granule Porosity	0.999	0.0302	8.45%

7.5.2 Operating Parameters

GP Validation

A 5-fold cross-validation of the GPs used within the operating parameters GP surrogate model was performed. The GPs predicted each of the four HSWG model outputs using the four operating parameter inputs. A plot of the residuals shown in Figure 7.5 compares the observed values to the predicted mean values. The four GPs have a high predictive quality as the datapoints lie closely to the red $y = x$ trend line. Table 7.5 shows the three diagnostic skill scores showing accurate mean predictions through the r^2 and RMSE values. Further the predictions are neither overconfident nor underconfident as the predictions are shown to be normally distributed with the outliers now much closer to 5%. Overall, the high predictive accuracy shown from the GP surrogate models enables confidence in calculation of the Sobol' indices required in the GSA.

Sobol' Indices

The Sobol' indices for each of the four operating parameters were calculated with respect to the four model outputs. The results are the total Sobol' indices which are presented in Figure 7.6

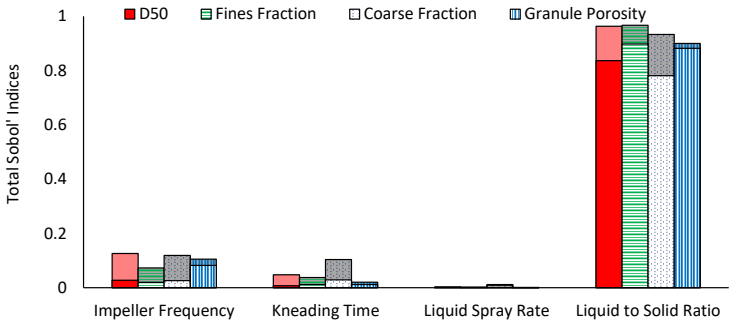


Figure 7.6: The Sobol' indices for the operating parameters with respect to each output.

Table 7.6: The impact each operating parameter has on the four outputs where green is highly impactful and red is negligible. Summarised in the last column by the average Sobol' index value for each modelling parameter, $\hat{S}_i^T$

	D50 S_i^T	**Fines Fraction** S_i^T	**Coarse Fraction** S_i^T	**Granule Porosity** S_i^T	**Average** $\hat{S}_i^T$
Impeller Frequency					0.11
Kneading Time					0.05
Liquid Spray Rate					0.004
Liquid to Solid Ratio					0.94

where each bar corresponds to the first-order Sobol' indices plus the remaining indices due to interactions with other input variables. Clearly, the liquid to solid ratio dominates all four model outputs, consistent with previous literature (Ameye et al., 2002). Both the impellar frequency and kneading time have a small influence. The liquid spray rate has negligible total Sobol' indices for all four outputs.

The average Sobol' index value per output ($\hat{S}_i^T$) for each operating parameter are shown in Table 7.6 clearly showing the CPP's and the impact each operating parameter has on each model output from the colour of the shading.

An $\hat{S}_i^T = 0.94$ shows that the liquid to solid ratio has a very high impact and so Figure 7.2 shows the CPP should have its crucial transition points identified. Clearly the liquid to solid ratio is dominant throughout the HSWG process and so it is vital that it should be accurately studied and be the focus of the experimental design. Therefore, it is recommended to have four levels in smaller step changes to identify potential critical levels of the liquid to solid ratio,

The value for both the impeller frequency and the kneading time is $0.05 < \hat{S}_i^T < 0.3$ and so the experimental design must identify the upper and lower level of these CPPs. Interest-

ingly, Figure 7.6 shows both the impeller frequency and kneading time to have 63% and 71%, respectively, of their $\hat{S}_i^T$ controlled by their interactions. These results are consistent with an experimental observation (Iveson and Litster, 1998) showing that at high liquid to solid ratio, the impeller frequency and kneading time have a higher impact because the high liquid to solid ratio aids in coalescence of particles, ensuring major impact from the particle size.

The liquid spray rate was only varied within the recommended drop-controlled regime. Within this narrow range, the liquid spray rate has been proven to have negligible impact on the output of the HSWG model. These findings are in agreement with data in the literature (Smrčka et al., 2015). Therefore, the liquid spray rate can be assumed constant because good experimental design recommendations are already available.

7.5.3 Experimental Design Proposal

The two GSA's have been used for the HSWG case study in an attempt to improve the experimental design. Here, we show the considerations that have to be made and how the analysis has reduced the work considerably. This section is in essence a demonstration of concept to show the practicality of this research.

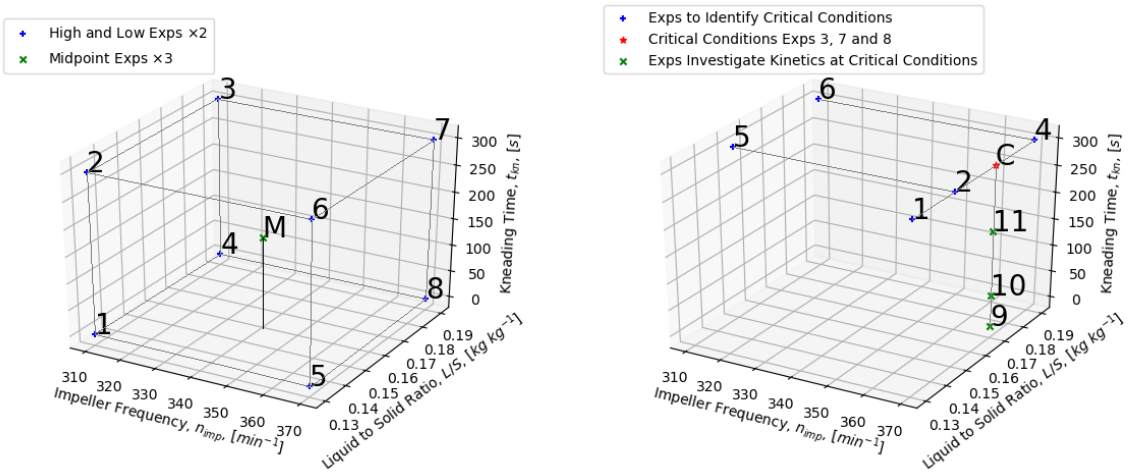
A conventional factorial DOE is not very useful for parameter estimation. A better experimental design is needed that allows the identification of critical conditions. In this way, modelling parameters that depend on critical operating conditions can be accurately estimated in an efficient way. Therefore, the results presented from both GSA's helps guide the experimental design that would lead to a well-calibrated HSWG model. Only four of the twenty modelling parameters have a significant impact and so only the values for the impactful modelling parameters have to be determined accurately. This is done through the use of characterisation tests to derive the physical properties and parameter estimations for empirical parameter using lab-scale experimental data. An experimental investigation of operating parameters with a negligible impact is not beneficial for model calibration. Therefore, the proposal of a reduced experimental design only based on the CPPs will significantly reduce the experimental effort and still reach the goal of a well-calibrated process model.

Table 7.7: Recommended experimental design for parameter estimation

Exp	$n_{imp}\ \left[\mathrm{min}^{-1}\right]$	$L/S\ \left[\mathrm{kg\,kg}^{-1}\right]$	t_{kn} [min]	Explanation
$1-4$	370 (high)	0.13, 0.15, 0.17, 0.19	5 (high)	Vary L/S to determine critical L/S
$5-6$	310 (low)	0.15, 0.19	5 (high)	and estimate the porosity parameters
$7-8$	370 (high)	0.17 (critical)	5 (high)	Triplicates to assess reproducibility
$9-11$	370 (high)	0.17 (critical)	0, 1, 3	Reduce t_{kn} to determine critical t_{kn}

Using these results an experimental design proposal based on the GSA results was developed for a $10\,\mathrm{L}$ granulator which is shown in Table 7.7. The ranges for impeller frequency and liquid to solid ratio depend on the equipment scale and formulation, respectively. In Exp $1-4$, the liquid to solid ratio of the critical granulation conditions is determined. The critical granulation conditions are defined as the lowest liquid to solid ratio at which the critical pore saturation is reached and rapid growth occurs (Iveson et al., 2001). Rapid growth is identified by a significant increase in particle size $D50$ and the formation of particles that are much greater than $1\,\mathrm{mm}$. The empirical minimum porosity correlation (Bellinghausen, 2020) is fitted using the results of Exp $1-6$. At critical conditions, triplicates are performed to assess the reproducibility (Exp $7-8$). In Exp $9-11$, the kneading time is reduced at critical conditions to determine the critical time and investigate the process kinetics in more detail at critical conditions. This is limited to the critical conditions because the kneading time is expected to have the highest impact at these operating conditions. In this scenario, it is recommended to use the results at critical conditions (Exp 3 and $7-11$) to estimate the three impactful modelling parameters: collision and breakage coefficient, and critical pore saturation. The remaining experiments are available to validate the estimates and assess the predictive power of the model.

Additionally, Figure 7.7 compares the proposed experimental design to a conventional experimental design that would have been conducted without the GSA results. The conventional experimental design is a 2^4 factorial DOE with three midpoint triplicates. The 3D scatter plot shown in Figure 7.7a shows eight high and low experiments with a midpoint experiment shown in the middle. It must be noted that there are three midpoint experiments and another eight high and low experiments for when the liquid spray rate value has changed. This conventional design therefore includes a total of 19 experiments. Whereas, the proposed experimental design shown



(a) Conventional (b) Proposed

Figure 7.7: The experimental designs for the HSWG case study.

in Figure 7.7b, consists of eight initial experiments to determine the critical liquid to solid ratio including three triplicates to asses reproducibility. Then an additional three experiments to determine the critical kneading time. Therefore, the proposed design of experiments reduces the conventional design from 19 experiments to 11 experiments. The workflow proposed reduces the experimental effort by 42.1%. Further, the proposed method has additional benefits as it will determine the critical pore saturation, an impactful parameter, whereas, the conventional method is not a very promising method with only two levels of each factor. Consequently, the proposed experimental design is a very promising approach to efficiently calibrate a HSWG process model.

7.6 Conclusion

Designing a new product via particulate processes involves testing every combination of input variables and so an experimental study for high shear wet granulation (HSWG) is often a costly process in time, money and materials. To help, an appropriate process model is used within a model-driven design workflow. Therefore, a model calibration workflow is proposed based on the identified critical process parameters (CPPs) and most impactful modelling parameters. Thus better targeted experimental design can be derived with the goals to increase the usability for model calibration and reduce the experimental effort.

To do this, an efficient global sensitivity analysis (GSA) is proposed to calculate the Sobol' indices for the HSWG model. In order to make the computational times practical, a Gaussian Process (GP) surrogate model is utilised to avoid the need of a very significant number of model simulations for the evaluation of high-dimensional integrals. The novel approach is applied using a suitable workflow so that both the modelling and operating parameters undergo individual GSA's, as both sets of parameters need to be considered for the experimental design.

Criteria for both the modelling parameters and the operating parameters were set so that the average total Sobol' index value for the four outputs, $\hat{S}_i^T$, were used to reduce experimental effort. The most impactful modelling parameters were determined using a threshold of $\hat{S}_i^T > 0.1$ reducing the modelling parameters from twenty to just four. The collision coefficient, breakage coefficient, critical pore saturation and nuclei-to-drop diameter ratio were recognised as the most impactful modelling parameters where the values need to be estimated in a characterisation test or through parameter estimation. The operating parameters used a threshold of $\hat{S}_i^T > 0.05$ to determine the CPPs. The calculated Sobol' indices found the liquid to solid ratio to dominate the HSWG process and so it should be the focus of the experimental design. Both the impeller frequency and kneading time were found to have an impact, primarily controlled by their interactions. Whereas, the liquid spray rate has negligible impact on the output of the process so it can be excluded from the experimental design by keeping the values constant from literature. Hence, we have recommended various levels for each operating parameter as factors in an experimental design in an attempt to estimate the most impactful modelling parameters that require parameter estimation for the model calibration of a HSWG process model. The results show the experimental design should be focused on the liquid to solid ratio and so four levels are recommended for this factor. Whereas, impeller frequency and kneading time can be analysed using just two levels. Hence we have shown that the HSWG process model can be efficiently calibrated through the use of a GP based GSA. Overall, the research has successfully completed the goals set out in Section 7.2 by identifying both the most impactful modelling parameters and the CPPs. From this, the impact of each parameter on a DOE has been analysed ensuring a proposal of a more efficient experimental design for the model calibration of a

HSWG process. The proposal substantially improves the ability to quickly deploy model-based design by determining the critical pore saturation value while reducing the experimental effort of 42.1% in comparison to a conventional experimental design.

During this research we have applied the methodology and proposed an efficient experimental design that can now be experimentally validated. Conducting this parameter estimation with experimental data will fully calibrate the PBM model to a $10\,\mathrm{L}$ high-shear mixer filled with $2\,\mathrm{kg}$ of dry powder case study. Consequently, we have proposed a much more efficient experimental design that will indicate any unidentified subprocesses. On a more practical level, it would be beneficial to extend the work carried out in this research to any particulate processes, further reducing the experimental costs of designing new products. Additionally, the GP based sensitivity analysis method will be developed further to enable optimisation of the parameter values.

Chapter 8

Active Subsets as a Tool for Structural Characterisation and Selection of Metal-Organic Frameworks

8.1 Abstract

To date over 80000 metal-organic framework (MOF) structures have been synthesised and only *ca.* 3% of these have had their adsorption capabilities measured for storing oxygen alone. As such, in order to aid the process of producing top performing MOFs for storing various gases, accurate methods to predict the deliverable capacity of MOFs that have their synthesis method already known is increasingly important. For this purpose, this paper develops a reduced order model (ROM) that can predict the deliverable capacity of synthesised MOFs irrespective to the storage gas across similar gases.

The ROM is constructed by identifying the Active Subspaces through a Sobol' index-based global sensitivity analysis (GSA). The resulting Gaussian Process (GP) regression model efficiently predicts the deliverable capacity given a MOFs pore properties with this reduced dimensional space.

This approach was applied to a practical MOF exploration example by training a ROM with

2745 MOFs storing methane at 30 bar. The ROM was robustly tested and analysed before using it to predict the deliverable capacity of 82221 synthesised MOFs storing oxygen at 30 bar. To ensure validity in the exploration example, the predictions produced from the methane trained ROM were compared to a separate ROM trained using the same MOFs but storing oxygen gas. The methane trained ROM was found to be in agreement with the oxygen trained ROM, and was shown to be a viable tool to identify the top performing MOF structures for oxygen storage.

8.1.1 Keywords

Metal-Organic Frameworks; Active Subspaces; Reduced Order Modelling; Gaussian Process; Sobol' Indices; Gas Adsorption

8.2 Introduction

Metal-organic frameworks (MOFs) (Carraro and Gross, 2014; Moghadam et al., 2017b) are a unique class of crystalline porous polymers that allow the design of pore structure and functionality due to their self-assembly synthesis process where metal building units are bridged by organic ligands. The nature of MOFs excites scientists due to their tailorable structural properties allowing a large number of MOFs to be potentially synthesised (Wilmer et al., 2012; Moghadam et al., 2017b). Consequently, the flexibility of MOFs allows them to be applied in a multitude of industrial settings such as gas storage (Mason et al., 2015; Thornton et al., 2017; Tian et al., 2018) and separation (Li et al., 2009; Bobbitt et al., 2017; Moghadam et al., 2017a).

To date *ca.* 88000 (Moghadam et al., 2017b) structures have been synthesised imposing difficulties in identifying top materials for each purpose. Such large datasets are often analysed through machine learning techniques and these methods have already shown accurate predictions for various gas uptake capacities of MOFs (Fernandez et al., 2013b; Ohno and Mukae, 2016; Fanourgakis et al., 2019). Examples of such techniques include polynomial chaos expansion (Sudret, 2008; Brown et al., 2013), artificial neural networks (Li et al., 2016), and Gaussian Processes (GPs) (Marrel et al., 2009; Yeardley et al., 2020a, 2021). These machine learning methods and high-throuput screening (Wilmer et al., 2012; Banerjee et al., 2016; Moghadam et al., 2019) aid in speeding up the process of finding new materials.

Previous studies have shown that machine learning methods enable the discovery and optimal design of MOFs (Moghadam et al., 2019). Fernandez et al. (2013b) compared three machine learning models to predict the storage capacity of MOFs storing methane. The research used 1.3×10^5 hypothetical MOFs at 1, 35 and 100 bar using the pore size, surface area, void fraction and framework density as input variables to multi-linear regression models, decision trees and nonlinear support vector machines. Soon after, Fernandez et al. (2014) furthered the research by introducing the atomic property weighted radial distribution function descriptor as an input variable in addition to the MOFs geometric features. This enables the models to capture the chemical features of a periodic material. The results suggested that more compact MOFs with interatomic ranges from 6 to 9 Å produce higher storage capacities for CO_2 at 0.15 and 1 bar. The first GP applied towards MOFs aided in predicting the methane uptake using the pore properties and the structural relationship as input variables (Ohno and Mukae, 2016). An automatic relevance determination (ARD) kernel calculated the lengthscales to discover the density, volumetric surface area and the void fraction to be the most dominant input variables. The GP surrogate model was then used to successfully identify new MOFs that could outperform them used in the training dataset for the methane uptake at 3.5 MPa. Further work using decision trees and support vector machines (Aghaji et al., 2016) found pore size, void fraction and surface area to be the most important factors when designing MOFs for CO_2 uptake and methane purification. Additionally, research has used the random forest technique to discover new MOFs for separation of xenon and krypton (Simon et al., 2015) and both CO_2 and N_2 uptake (Fernandez and Barnard, 2016). However, the biggest drawback with all the previous research is that only hypothetical MOF data were used to train the machine learning models and the synthesis of top performing materials is not always straight-forward. For databases containing MOF structures that are already synthesised Fanourgakis et al. (2019) used the Computation-ready, experimental (CoRE) MOFs (Chung et al., 2014), which have been experimentally synthesised with accurate calculations of their structural features. This research used a random forest algorithm to investigation to predict the methane adsorption in MOFs, discovering a smaller training set size were needed for convergence when using synthetic MOFs

compared to the hypothetical MOF databases previously used.

In this study, we have adopted a GP surrogate model that directly emulates a dataset storing experimentally synthesised MOFs with their pore properties computationally measured through simulations (Moghadam et al., 2018). The use of synthesised MOFs has a significant advantage over hypothetical MOFs because knowing the synthesis method can accelerate the production process once top performing MOFs have been identified (Chung et al., 2014; Thornton et al., 2017). Given the complexity of the data set, a versatile, non-parametric GP approach allows characterisation of the data, while also allowing the semi-analytic evaluation of the Sobol' indices that would otherwise be calculated at significant cost. As such we use the GP surrogate to predict the gravimetric deliverable capacity (GDC) and volumetric deliverable capacity (VDC) of MOFs given their pore properties. Initially, global sensitivity analyses (GSAs) are conducted at various pressures to understand the influence each pore property has on the deliverable capacity at different storage pressures. Then an efficient comparison between MOFs storing oxygen and methane enables the production of a ROM for property prediction.

Previous research studies investigating the design of MOFs identified the important variables without quantifying the impact of each variable. To the authors knowledge only Ohno and Mukae (2016) conducted a sensitivity analysis by calculating qualitative values for the most dominant input variables. Here, our contributions allow the further understanding of MOFs by developing a GP surrogate model that accurately predicts the deliverable capacity of MOFs. Additionally, the model investigates the effect each pore property has on the gaseous uptake of MOFs using GSAs.

Principal component analysis techniques have been widely applied in MOF research (Qiao et al., 2018, 2017; Fernandez et al., 2013a) to assess the interrelationships among MOF descriptors. A principal component analysis allows the dimensionality of multivariate data to be reduced providing a concise representation of data. Qiao et al. (2017) performed a principal component analysis combining four descriptors to make a two-dimensional model to help predict the performance of MOFs for thiol capture. This method of projecting four descriptors into a two-dimensional model was conducted again to aid in separating natural gas (Qiao et al., 2018). In

this research, we have developed a stronger technique based on Active Subspaces (Constantine et al., 2014; Constantine and Diaz, 2017) which utilises the GSA to build a reduced order model (ROM). Overall, we present the ROM to be used as a prediction technique which enables fast and efficient material design and discovery for gas storage applications due to the ability to predict irrespective to whether the MOF is storing oxygen or methane.

The novel approach presented here will allow us to determine: *i).* The most dominant pore properties and their interactions *ii).* The effect pressure has on the most influential pore properties in MOFS *iii).* How the pore properties influence the uptake of oxygen storage in comparison to methane storage and *iv).* A fast and efficient prediction technique for the gravimetric and volumetric deliverable capacity of synthesised MOFs irrespective to the storage gas across similar gases.

8.3 Mathematical Background

8.3.1 Gaussian Process Regression

GP regression is a non-parametric machine learning method used to create a blackbox function that predicts the deliverable capacity given a MOFs pore properties. Bayesian conditioning (Williams and Rasmussen, 2006) is used to learn the mean $\left(\bar{f}(\mathbf{x})\right)$ and variance $\left(\Sigma_f\right)$ using a MOF database which includes paired pore properties and measured deliverable capacities as training data $(\mathbf{y} = f(\mathbf{X}) + \mathbf{e})$. Once trained, the GP takes a $(1 \times d)$ row vector as inputs $\mathbf{x}$ to predict output as a Gaussian random variable using the predictive equations

$$y(\mathbf{x}) \sim \mathsf{N}\left[\bar{f}(\mathbf{x}), \Sigma_f + \sigma_e^2\right] \tag{8.1}$$

where

$$\bar{f}(\mathbf{x}) := k(\mathbf{x}, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I})^{-1}\mathbf{y} = k(\mathbf{x}, \mathbf{X})\mathbf{K}^{-1}y \tag{8.2}$$

$$\Sigma_f := k(\mathbf{x},\mathbf{x}) - k(\mathbf{x},\mathbf{X})(k(\mathbf{X},\mathbf{X}) + \sigma_e^2\mathbf{I})^{-1}k(\mathbf{X},\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - k(\mathbf{x},\mathbf{X})\mathbf{K}^{-1}k(\mathbf{X},\mathbf{x}) \quad (8.3)$$

At the heart of this lies the kernel function $k\colon \mathbb{R}^{i+d} \times \mathbb{R}^{j+d} \to \mathbb{R}^i \times \mathbb{R}^j$, expressing the correlation between responses to input samples of sizes $(i \times d)$ and $(j \times d)$. This work exclusively uses the automatic relevance determination (ARD) kernel (Wipf and Nagarajan, 2007):

$$k(\mathbf{x}',\mathbf{x}) := \sigma_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^\intercal \Lambda^{-2}(\mathbf{x}-\mathbf{x}')}{2}\right) \quad (8.4)$$

where Λ is a $(d \times d)$ *diagonal* positive definite lengthscale matrix. The learning from training data requires the optimisation $d+2$ hyperparameters, consisting of Λ, σ_f, and σ_e, through the maximum marginal likelihood $p[\mathbf{y}|\mathbf{X}]$ using the ROMCOMMA software library (Milton and Brown, 2019).

8.3.2 Global Sensitivity Analysis

This works implements a GSA using the variance based Sobol' indices technique (Sobol, 1993, 2001). The calculation of Sobol' indices requires the evaluation of complex multi-dimensional integrals. Here, we shall describe the calculation of Sobol' indices up to evaluating the integrals, which are calculated using the GP regression model resulting in semi-analytic Sobol' indices.

Sobol' indices are calculated by considering a function $y = f(\mathbf{x})$, where $\mathbf{x} := [x_1, ..., x_d]$ is a d-dimensional row vector found in the input space, Ω, and y is the model output. Assuming that the inputs are mutually dependent and that $f(\mathbf{x}) \in L^2(\Omega)$ (Sobol, 1993, 2001), then a particular input x_i has a first-order Sobol' index defined by

$$S_{1,i} = \frac{Var\left[E\left[\bar{f}(\mathbf{x})|x_i\right]\right]}{Var\left[\bar{f}(\mathbf{x})\right]} \quad (8.5)$$

To be able to express the whole effect of an input on the output, the total Sobol' index is

$$S_{T,i} = S_{1,i} + \sum_{j\neq i}^{n} S_{2,ij} + \sum_{j\neq i, k\neq i, j<k}^{n} S_{3,ijk} + ... \quad (8.6)$$

Therefore, the first-order Sobol' indices measure the contribution to the variance solely attributable to x_i, in contrast, the total Sobol' index of i corresponds to its own contribution including interactions with the other inputs (Saltelli and Homma, 1996; Saltelli et al., 2008).

From here, the partial variances of y are determined through a decomposition method presented by Sobol (1993) which evaluates each term through multi-dimensional integrals. Here, we compute the semi-analytic evaluation of the integrals using the GP regression model's predicted mean in Equation (8.2). Therefore, we have used GPs as a blackbox function providing an efficient sampling method for $y = f(\mathbf{x})$ in a similar way to that from Chen et al. (2005b).

8.3.3 Reduced Order Modelling

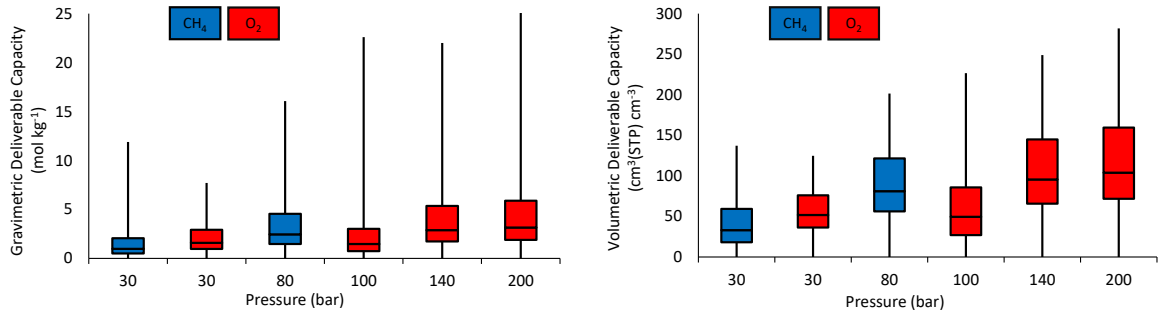
The ROM conducts an optimal dimension reduction by locating the Active Subspaces (Constantine et al., 2014; Constantine and Diaz, 2017). The ROM reduces the inputs to a set which is highly relevant to the response through the use of Sobol' indices. It is achieved by finding an orthogonal rotation matrix θ which combines the inputs into a low dimensional subspace. The methodology is achieved by rotating the input basis:

$$\mathbf{x} \coloneqq \mathbf{\Theta}^{\mathsf{T}}\mathbf{u} \tag{8.7}$$

Then by calculating the proportion of response variance ascribable to the first m basis directions of $\mathbf{u}$ through the Sobol' index, the optimum rotation can be determined by maximising this relevance.

$$S_m(\mathbf{\Theta}) \coloneqq \frac{Var\left[E\left[\bar{f}(\mathbf{x})|\mathbf{\Theta}\mathbf{x}\right]_m\right]}{Var\left[\bar{f}(\mathbf{x})\right]} \tag{8.8}$$

Where the dimension $m << d$. Therefore, S_m has to be optimised for $m = 1, ..., d$ in turn, to find the most relevant direction, then the second most relevant, and so on. Essentially, optimization is then a two-stage process: 1) a GP predictor is optimised, then GSA using the predictor rotates the input basis; 2) this rotated basis is then used in a new predictor. Ultimately this yields a predictor that only significantly depends on a dimensionally reduced basis.



(a) (b)

Figure 8.1: Box and whisker plots for the descriptive statistics of the a) gravimetric deliverable capacity and b) volumetric deliverable capacity for all 2745 MOF structures used in this research.

8.4 Method

In this section, we outline how the mathematical tools presented in Section 8.3 are applied to 2745 structures of synthesised MOFs. Understanding the MOF data is essential in the development of the methodology for creating an efficient predictive technique.

The MOF database used in this study is taken from Moghadam et al. (2018). The synthesised MOFs was first developed by Nazarian et al. (2016) before the deliverable capacities at various pressures were modelled using grand canonical Monte Carlo simulations." In this work, the deliverable capacity is defined as the difference between the amount of gas adsorbed at the storage pressure and the release pressure of 5 bar. For each storage gas and pressure, the descriptive statistics of the deliverable capacity for all 2745 MOF structures are shown in Figure 8.1 as box and whisker plots. The box in each plot bounds the first and the third quartile showing the median value as the middle line. The whiskers show the minimum and maximum deliverable capacity values. This research uses the deliverable capacity of each MOF as the important output value to analyse.

An ideal MOF structure would have a large deliverable capacity, which likely depends on several pore properties including the largest cavity diameter (LCD), volumetric surface area (A_V), gravimetric surface area (A_G), density (ρ), void fraction (VF), and heat of adsorption at 5 bar (H5) relative to the storage gas. The descriptive statistics showing the spread of values for each pore property throughout the 2745 MOF structures are shown in Table 8.1. Please note,

Table 8.1: The pore properties available for use as input variables for a GP.

Pore Property	LCD	A_V	A_G	VF	H5 of O_2	H5 of CH_4	ρ
Units	Å	$m^2\ m^{-3}$	$m^2\ g^{-1}$	-	kJ mol^{-1}	kJ mol^{-1}	g cm^{-3}
Mean	5.70	651	616	0.322	17.5	20.7	N/A
STD	2.09	727	828	0.183	3.63	5.50	N/A
Range	15.4	3096	5203	0.880	31.5	36.2	N/A

that although the density of each MOF was available, it was not used in calculations due to its dependency with A_G and A_V. Here we focus on using these pore properties and input variables to predict the deliverable capacity of MOFs. Further, these will be analysed using a GSA to understand the influence of each pore property on the deliverable capacity.

The deliverable capacities were available at different pressures; however, they were only measured at 30 bar for both oxygen and methane storage. Therefore, the GSA investigating the deliverable capacity at various pressures was conducted using MOFs storing oxygen. Subsequently, we investigated the effect which changing the storage gas has on the GSA and so kept the storage pressure constant at 30 bar.

GP regression is used to encapsulate the MOF data into a simple framework that can be used to make predictions and to enable the semi-analytic calculation of Sobol' indices. In this research, we are interested in both the gravimetric and volumetric deliverable capacity, and so for each scenario (storage pressure or storage gas), we have to produce two independent GP regression models. The models are regressed using standardised data and so the mean and standard deviations for the input-output data used for training were saved for standardising the test data. GP predictions work by taking a (1×5) row vector of pore property input values to predict the distribution of the output. To do this, the hyperparameters are optimised using the ROMCOMMA software library (Milton and Brown, 2019) where learning is achieved using the MOF database as training data. From there, the predictive equation is used for semi-analytic calculations of Sobol' indices once the GP models have been tested for inaccuracies.

The predicted deliverable capacity of MOFs are tested by comparing them to observed deliverable capacities and to ensure inaccuracies from the GP models are not inherently produced in the GSA. For this reason, 5-fold cross-validation (Hastie, 2009) is used in this research,

whereby, the MOF dataset is randomly split into five subsets. Then four of the five subsets are used as training data, while the remaining is used to test the GP model predictions. This procedure is repeated so that every single sampling point produced is used for testing. Once the GP models have been created, tested, and used for calculation of Sobol' indices, the process can be repeated with rotations and iterations as described in the previous section to produce a ROM.

8.5 Results

This Section presents the GSA results when comparing both the storage pressure and the storage gas. First, the pressure analysis investigates the impact each pore property has on oxygen storage at four various pressures. Then two GSAs are conducted, as described in the previous Section, to explore the impact that the storage gas has on the deliverable capacity of MOFs at 30 bar. For both studies, the GP models are assessed using cross-validation (Hastie, 2009) to ensure inaccuracies are both carried through to the semi-analytic calculation of Sobol' indices. Finally, the ROM's are built and we present the results to show the predictive capabilities even when the MOFs storage gas is unknown.

8.5.1 Sensitivity Analysis of Pressure

A GSA of the five pore properties with respect to the deliverable capacity at various pressures has been achieved using a GP regression model. Therefore, confidence in the resulting Sobol' indices depends entirely on the accuracy of the GP predictions. This accuracy was determined using 5-fold cross-validation so that each test point had not been used in training the GP model. The residuals for both the GDC and VDC are presented in Figure 8.2, comparing the true deliverable capacities to the predicted mean values for all 2745 MOFs storing oxygen at four pressures. The markers in Figure 8.2 lie close to the black $y = x$ trend line showing good test predictions. Additionally, the standardised diagnostic values are presented in Table 8.2, quantitatively showing the marks closeness to the trend line through high coefficient of determination values (r^2). The good predictions are given further evidence through low standardised root mean square error (RMSE) values and a predicted distribution shown to be normally distributed as close to 5% of the true standardised deliverable capacity values are outside of the

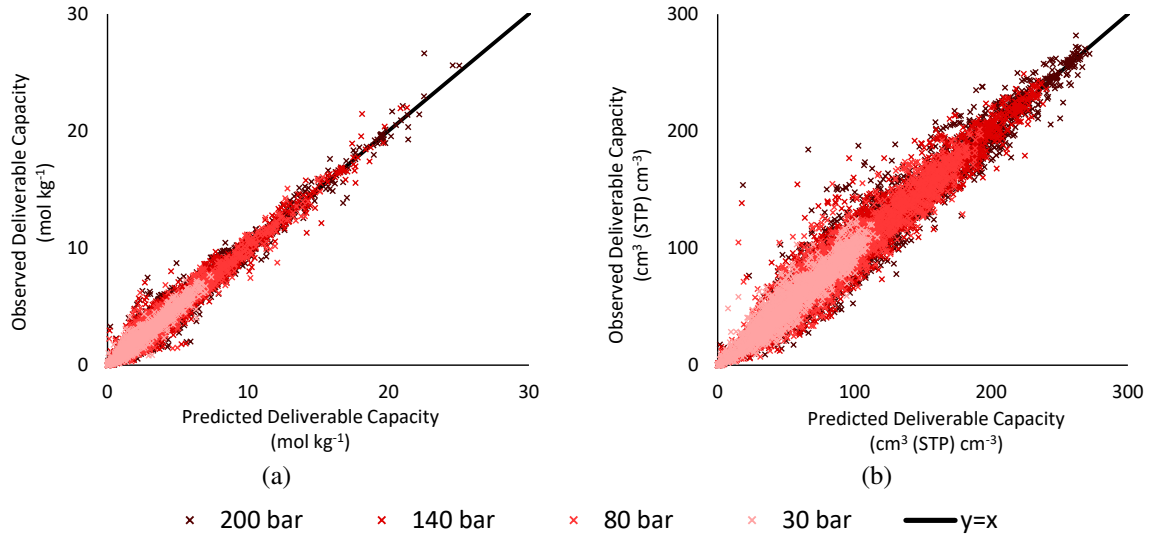


Figure 8.2: The residuals from testing the GPs predicting a) the gravimetric deliverable capacity and b) the volumetric deliverable capacity of MOFs storing oxygen at various pressures using 5-fold cross-validation.

Table 8.2: The diagnostic values of the GPs predicting the deliverable capacity of MOFs at various pressures.

Deliverable Capacity	Pressure (bar)	r^2	RMSE	Outliers	Max Error	MOF ref code
GDC	30	0.945	0.234	5.21%	2.53 mol kg^{-1}	NICZAA01
GDC	80	0.960	0.201	5.61%	3.86 mol kg^{-1}	NICZAA01
GDC	140	0.965	0.187	5.50%	4.39 mol kg^{-1}	NICZAA01
GDC	200	0.967	0.181	5.46%	4.64 mol kg^{-1}	NICZAA01
VDC	30	0.912	0.297	5.28%	45.4 cm^3(STP) cm^{-3}	NICZAA01
VDC	80	0.924	0.275	4.99%	89.9 cm^3(STP) cm^{-3}	LIHRIE
VDC	140	0.929	0.266	4.74%	121 cm^3(STP) cm^{-3}	LIHRIE
VDC	200	0.931	0.262	4.55%	136 cm^3(STP) cm^{-3}	LIHRIE

predictions uncertainty distribution. Overall, the high predictive accuracy shown from the GP models enables confidence in the calculation of Sobol' indices.

The first GSA of this research investigates the impact each pore property has on the deliverable capacity of synthesised MOFs at different pressures. The calculated Sobol' indices are presented in Figure 8.4 as total Sobol' indices split to show the first order values at the bottom plus the remaining indices due to interactions with other pore properties at the top. In Figure 8.3a we can see the impact each pore property has on the GDC, whereas the output for Figure 8.3b is the VDC. A direct comparison of the storage pressure can be seen for each pore property as the pressure changes from 30 bar on the left to 200 bar on the right as shown by the darkening colours highlighted in Figure 8.3's legend.

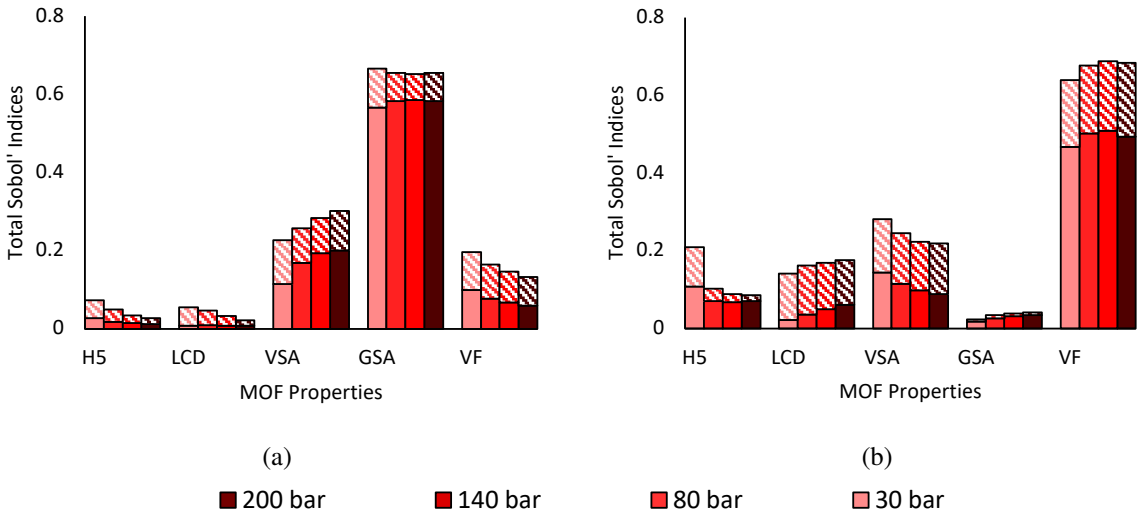


Figure 8.3: The bar charts showing the calculated total Sobol' indices for each pore property with respect to a) the gravimetric deliverable capacity and b) the volumetric deliverable capacity of MOFs storing oxygen at 30, 80, 140 and 200 bar. Each bar is split with the bottom solid fill corresponding to the first-order Sobol' indices and the top diagonal pattern being the cross effects.

A comparison between Figure 8.3a and Figure 8.3b shows the Sobol' index value for gravimetric surface area for the GDC to significantly increase by almost twenty times compared to what it was for the VDC. Therefore, gravimetric surface area is the most dominant pore property impacting the GDC but has negligible impact on the VDC. Whereas, the void fraction's Sobol' indices are four times larger for the VDC compared to the GDC. For either deliverable capacity and all four storage pressures the heat of adsorption at 5 bar and the largest cavity diameter both have first-order Sobol' indices of 0.1 or below and so these pore properties have a negligible impact on MOFs oxygen storage capabilities.

Interestingly, the increasing pressure does causes the impact of the heat of absorption at 5 bar to decrease. However, the importance of the remaining pore properties either increase for the GDC or decrease for the VDC. This may be due to the correlation between outputs or, alternatively, it could simply be a coincidence drawn from the uncertainty in the calculated Sobol' indices. Until further research, the minor changes between pressures are considered to be equivalent and so these findings demonstrate that the influence each pore property has on the MOFs oxygen storage capability in not impacted by the storage pressure.

Table 8.3: The diagnostic values of the GPs predicting the deliverable capacity of MOFs storing oxygen and methane at 30 bar.

Deliverable Capacity	Storage Gas	r^2	RMSE	Outliers	Max Error	MOF ref code
GDC	O_2	0.945	0.234	5.21%	-2.54 mol kg^{-1}	NICZAA01
GDC	CH_4	0.943	0.239	5.90%	-6.38 mol kg^{-1}	HIFTOG
VDC	O_2	0.912	0.297	5.28%	-45.9 cm^3(STP) cm^{-3}	NICZAA01
VDC	CH_4	0.897	0.321	4.92%	-61.1 cm^3(STP) cm^{-3}	LIHRIE

8.5.2 Sensitivity Analysis of Storage Gas

This Section presents the findings and contributions made when comparing the GSAs made for MOFs storing different gases. Once again, the GP regression model had to be fully validated to ensure trust in the GSAs. Following the 5-fold cross-validation each model had the standardaised error diagnostic values calculated and presented in Table 8.3. Interestingly, Table 8.3 show the maximum prediction error for the GP predicting the VDC of MOFs storing methane to be LIHRIE which is the same MOF causing the largest prediction errors from storing oxygen at 80, 140 and 200 bar as shown in Table 8.2. Whereas, a GP model for the GDC of methane at 30 bar results in a maximum error for a different MOF structure. Here, Table 8.2 shows HIFTOG has the largest error of -6.38 mol kg^{-1}. Anomalies such as these often occur in machine learning because the input variables are either outside of the range of the training data used. Alternatively, the input variables could be close to them from training data but have a different output value. Analysis of the MOF structures causing anomalies have shown that the input variables was within the range used for training. Therefore, the MOF is behaving differently to them with similar pore properties. The reasoning for the difference in behaviour of these MOF outliers is unclear at this point as there is no common characteristic of these MOFS. In future work, it would be interesting to conduct further chemical analysis to understand why these MOFs behave differently. Instead, this research focuses on a methodology that provides statistical results, validating the GPs showing predictions that enable accurate modelling of synthesised MOFs.

The GSAs conducted for MOFs storing both oxygen and methane at 30 bar produced total Sobol' indices presented in Figure 8.4, a bar chart similar to Figure 8.3. The storage gas is

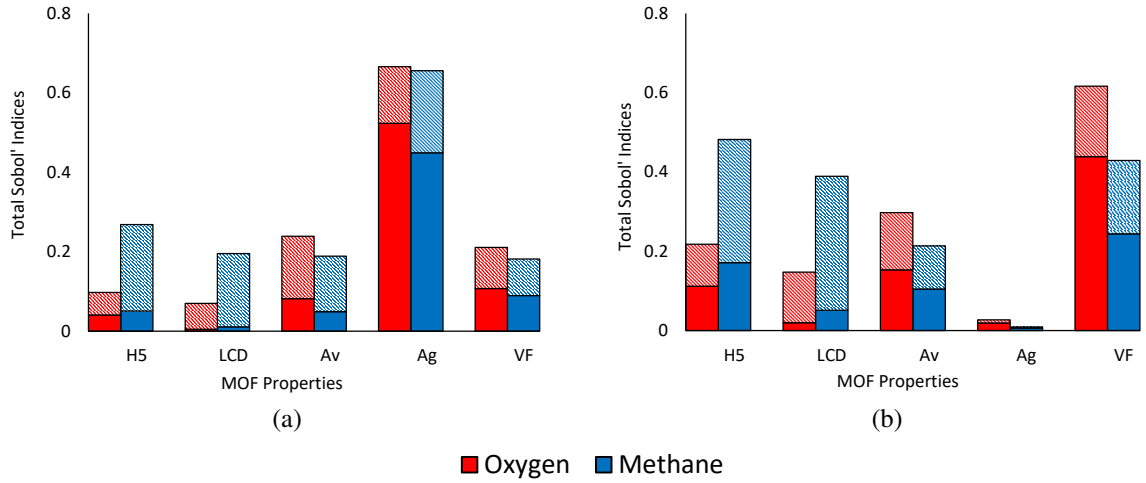


Figure 8.4: The bar charts showing the calculated total Sobol' indices for each property with respect to a) the gravimetric deliverable capacity and b) the volumetric deliverable capacity of MOFs storing oxygen (red) and methane (blue). Each bar is split with the bottom solid fill corresponding to the first-order Sobol' indices and the top diagonal pattern being the cross effects.

compared using the colour of each bar with red representing oxygen and blue representing methane. As expected, both this GSA and the previous GSA calculate identical values for the Sobol' indices of MOFs storing oxygen at 30 bar. Together, this supports the reliability of the results, showing the clear impact each pore property has on the capability of MOFs storing oxygen.

Comparing each GSA demonstrates many similarities between the effects each pore property has depending on the storage gas. Figure 8.4a shows the gravimetric surface area to dominate the GDC for both storage gases whereas the Sobol' index value for the heat of adsorption at 5 bar increases by more than double when storing methane compared to oxygen for both deliverable capacities. It is important to highlight the first-order Sobol' indices for the heat of adsorption at 5 bar only increases by 24% for GDC and 53% for VDC. This implies that the significant increase ($> 200\%$) in the total Sobol' index is associated with the heat of adsorptions interactions with other pore properties. This phenomenon is clear to see for the largest cavity diameter as the cross effects bar increases significantly from blue (oxygen storage) to red (methane storage) in Figure 8.4.

Overall, the GSAs investigating the storage gas successfully quantified the effect each pore

property has on both the deliverable capacities of MOFs storing either oxygen or methane. The results show that the interactions between the pore properties have a larger effect on the MOFs capabilities when storing methane compared to oxygen.

It can be seen that the relevance of the pore properties vary depending on whether the deliverable capacity is gravimetric or volumetric. For the GDC, the Sobol' index value for the gravimetric surface area is 66.6% when storing oxygen and 65.6% when storing methane. Whereas, the void fraction dominates the VDC with Sobol' indices of 61.7% when storing oxygen and 42.9% when storing methane. Determining these two as the most important pore properties are consistent with the qualitative findings in the literature (Ohno and Mukae, 2016).

Nonetheless, we believe these results build on previous MOF prediction studies by showing the possibility of predicting a MOFs deliverable capacity given limitations on the gas stored within the known dataset which is used to train machine learning models.

8.5.3 ROM Predictor Capabilities

The previous sections have performed GSAs investigating the impact the storage gas and storage pressure has on the effect of each pore property on MOFs deliverable capacity. This study has shown that the MOFs capabilities are largely unaffected by the storage capacity. Therefore, in this Section we have applied the novel GP based ROM technique to the MOF database, testing the methods capabilities by predicting the deliverable capacity of a MOF storing the opposite gas to that the training database stored. Once again, 5-fold cross-validation was used to ensure any MOF structures used in training the ROM is not used in testing the ROM. For example, a ROM is trained using four folds of MOF structures storing oxygen before it is tested by predicting the deliverable capacity of the remaining MOFs but storing methane.

Individually, four ROMs were trained and tested producing satisfactory prediction results. Figure 8.5 and Table 8.4 demonstrate these results in two residual plots and a table showing the standardised diagnostic values. The results are colour coordinated to show the predicted MOFs storage gas. For example, the blue residuals in Figure 8.5 show the predicted deliverable capacity of MOFs storing methane produced from a ROM trained using MOFs storing oxygen.

Table 8.4: The diagnostic values of the ROMs predicting the deliverable capacity of MOFs storing oxygen and methane at 30 bar.

Deliverable Capacity	Storage Gas	r^2	RMSE	Outliers	Max Error	MOF ref code
GDC	O_2	0.865	0.373	2.66%	3.83 mol kg^{-1}	XEBHOC
GDC	CH_4	0.826	0.546	4.74%	6.68 mol kg^{-1}	HIFTOG
VDC	O_2	0.829	0.417	1.35%	58.0 cm^3(STP) cm^{-3}	NICZAA01
VDC	CH_4	0.764	0.710	6.34%	65.9 cm^3(STP) cm^{-3}	RELLAW

Having a large correlation coefficient (r^2) and a low percentage of outliers outside two predicted standard deviations provide evidence for our novel ROMs to be used with confidence to predict the deliverable capacity of MOFs storing another gas to what was used for training.

Table 8.4 highlights the MOF with the maximum error for each ROM showing slight inconsistency with the worst prediction. This is in comparison to previous GP validation tests where Table 8.3 and Table 8.2 show NICZAA01 and LIHRIE to cause the worst predictions. Further, Figure 8.5 clearly shows lower deliverable capacities to be easier to predict as predictions begin to deviate away from the $y = x$ black trend line in both residual plots as the deliverable capacity increases. We believe this could be due to the general population of MOFs used for training being able to store more methane than oxygen on average. Further, the majority of these outliers begin to occur as the heat of absorption for MOFs storing oxygen become closer to the heat of absorption for MOFs storing methane. As the heat of absorption is the only input variable that changes when predicting the uptake of each gas, then it becomes understandable why the ROM prediction capabilities become less accurate when the MOFs have a higher deliverable capacity for methane than oxygen, but a similar heat of absorption. Therefore, as the deliverable capacities of the MOFs increase, the majority have a larger value when storing methane than oxygen, causing minor errors in the ROM when predicting an unknown gas. Hence, there are limitations to using the ROM to predict the deliverable capacity of MOFs at storage pressures greater than 30 bar. At this stage, we believe that as the pressure increases, being able to predict the deliverable capacity irrelevant to the storage gas would become significantly more difficult. Nonetheless, we believe developing this novel technique and testing it on synthesised MOFs storing just oxygen and methane builds on previous MOF prediction studies by being an initial step towards developing a universal model applicable to synthesised MOFs.

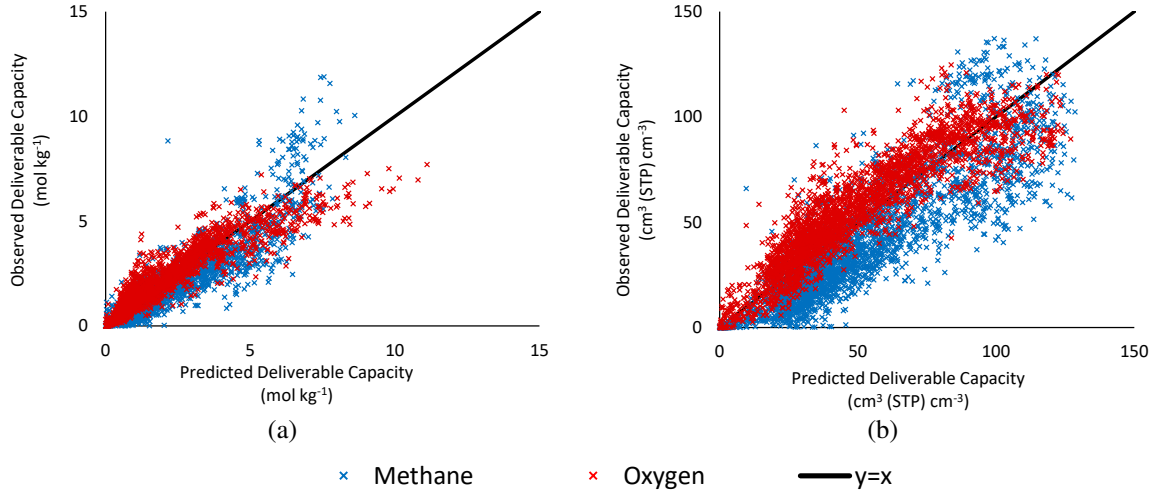


Figure 8.5: The residuals from testing the ROMs using 5-fold cross-validation to predict a) the gravimetric deliverable capacity and b) the volumetric deliverable capacity, showing the predictions of MOFs storing both oxygen (red) and methane (blue).

The ROM has been shown be able to make accurate predictions but more importantly it can also be reduced to a lower amount of dimensions, improving the efficiency. Figure 8.6 shows the cumulative Sobol' indices for each rotated dimension in the four ROMs. For similarity to previous figures, Figure 8.6 has the bars coloured depending on the function of the ROM. For example, a ROM trained using MOFs storing methane is used to predict the deliverable capacity of MOFs storing oxygen and therefore the bars for this ROM are coloured in red for oxygen. The cumulative Sobol' indices show how much of the variance is ascribable to i amount of rotated dimensions (U_i). Using Figure 8.6 we can see how many rotated dimensions each ROM can be reduced to. As expected, when all five dimensions are considered, 100% of the variance of the output is captured by the ROM. The ROM used for predicting the GDC of MOFs storing methane has 80.1% of the variance captured by the primary dimension, U_1. Therefore, one may be confident of making sufficient predictions for a given MOF using just one dimension. Whereas ROMs used for oxygen predictions should provide sufficient predictions using three dimensions. All of the ROMs capture over 90% of its output variance using four of the five dimensions and so all four models can be confidently reduced to four dimensions. This provides the potential mechanism for Active Subspaces which will provide fewer dimensions

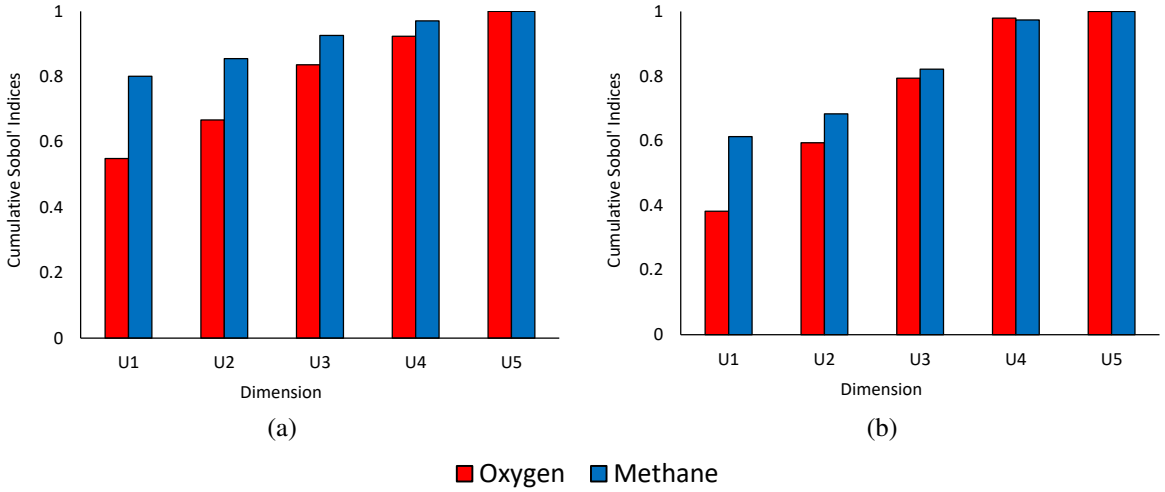


Figure 8.6: The bar charts showing the cumulative Sobol' indices for each dimension with respect to a) the gravimetric deliverable capacity and b) the volumetric deliverable capacity of MOFs storing oxygen (red) and methane (blue).

making MOF data far easier to understand and explain.

8.5.4 MOF Selection for Oxygen Storage

This section applies the ROM to a practical MOF exploration example. Oxygen storage using MOFs is rarely studied (Moghadam et al., 2018) but provides a significant promise for storage in healthcare, military and industrial applications (DeCoste et al., 2014). For oxygen storage, it is important to both increase storage per unit volume and reduce the storage pressure. Therefore, here we apply the ROM to help predict the top performing MOFs with respect to the deliverable capacity of oxygen at 30 bar.

This exploration study analysed the Cambridge Structural Database (Moghadam et al., 2017b) to predict the top performing MOFs for oxygen storage out of the 82221 synthesised MOFs. Predicting the performance of MOFs that have already been synthesised is more beneficial than predicting the performance of hypothetical MOFs because synthesis methods are not straightforward for newly proposed MOFs (Moghadam et al., 2018). After synthesis, measuring the adsorption of MOF structures has only been conducted on *ca.* 3% of the synthesised MOFs (Nazarian et al., 2016). Therefore predicting the deliverable capacity of synthesised structures is hugely important for producing promising MOF structures (Chung et al., 2014; Thornton

et al., 2017).

Previously, 2745 MOF structures (Nazarian et al., 2016) out of the 82221 MOFs within the Cambridge Structural Database were used to train the ROMS because the deliverable capacity of these synthesised MOFs were calculated using computational simulations (Moghadam et al., 2018). From this, two sets of data were used to train the ROMs, one using MOFs storing oxygen and the other used the same MOFs but storing methane. These models provided a belief that they can be used to make satisfactory predictions irrelevant to a MOFs storage gas across similar gases. Thus, in this section we predict the top performing MOFs for oxygen storage using a ROM trained from MOFs storing methane. The results are then compared to that from a ROM trained on the correct storage gas data.

Table 8.5 compares the top five MOF structures for oxygen storage predicted from our novel machine learning method. Here, the five promising structures were chosen because of their high predicted values for both gravimetric and volumetric deliverable capacity at 30 bar storage. However, due to limited volume in oxygen storage tanks, the VDC was prioritised over the GDC when needed. Each MOF had the deliverable capacity's mean value and the 95% confidence error predicted as shown in Table 8.5. Here, the ROM trained using 2745 MOFs storing methane, predicted the top MOF for oxygen storage at 30 bar to be VIKDID with a VDC of $118 \pm 15.0 \text{cm}^3\text{(STP) cm}^{-3}$. Additionally, the predicted GDC was high at $7.64 \pm 0.655 \text{mol kg}^{-1}$ making this a very promising MOF for oxygen storage. A comparison to MOF structures with known deliverable capacities from the training data found just two MOFs similar to the five predicted by the ROM. Table 8.5 shows these known MOFs to have similar pore properties to the five predicted MOFs giving confidence in the predictions. Hence, it is clear that the ROM is predicting top performing MOFs from structures with pore property values consistent to what have been found in previous research (Moghadam et al., 2018).

Additionally, Figure 8.7 compares the ten predicted top performing MOFs with the distribution from the training MOF database. The training MOF database is the 2745 MOF structures that Moghadam et al. (2018) previously screened and analysed (through grand canonical Monte Carlo simulations) as top performing MOF structure. Therefore, it is remarkable that new top

Table 8.5: The top five MOFs for oxygen storage identified by a ROM with comparison to two MOF structures with computationally calculated oxygen deliverable capacities.

MOF ref code	LCD (Å)	A_V ($m^2\ m^{-3}$)	A_G ($m^2\ g^{-1}$)	VF (-)	GDC (mol kg^{-1})	VDC (cm^3(STP) cm^{-3})
Predicted MOFs						
VIKDID	7.45	2280	3557	0.748	7.64 ± 0.655	118 ± 15.0
HEFDUT	9.94	2127	2661	0.732	6.57 ± 0.498	117 ± 13.2
NIMJUP	8.05	2289	2972	0.726	6.60 ± 0.488	115 ± 13.0
EJEKEK	7.62	2197	3155	0.703	7.22 ± 0.516	113 ± 13.4
VODSUC	7.70	2159	3176	0.687	7.39 ± 0.518	113 ± 13.4
Computationally Measured MOFs						
BEPROF	8.00	2120	3013	0.700	6.95	114
MATVEJ	8.53	2116	2956	0.690	6.27	101

performing MOFs from the Cambridge Structural Database consistently have a predicted deliverable capacity within the limits or above the synthesised MOFs used for training. This result enables the use MOFs with high storage capabilities that have already been synthesised. However, the close boundary to previously screened and simulated MOFs may actually hint towards the limit of synthesised MOFs to date. Clearly, with so many MOFs structures available to model, synthesise and then test, predicting top-performing MOFs for so many different uses will always be a challenge. The extent to whether it is possible to quantify all synthesised and non-synthesised MOFs is still unknown. Nonetheless, we believe that this new method provides a step towards quickly assessing capabilities of readily available MOFs for oxygen and methane storage.

Further, Figure 8.7 compares the predicted deliverable capacities from the ROM trained using MOFs storing methane and a ROM trained using MOFs storing oxygen. This comparison shows the reliability in using a methane trained ROM to predict the oxygen deliverable capacity due to the closeness between the two results. Figure 8.7 shows the oxygen trained predictions (red markers) to always be within the 95% error bars predicted by the methane trained ROM (in blue) for these top performing MOFs. Therefore, these results provide confidence that the novel machine learning method can be used to predict the deliverable capacity of MOFs irrelevant to the storage gas across similar gases. This is am important discovery due to the limitations involved in measuring the adsorption of so many MOF structures for so many gases.

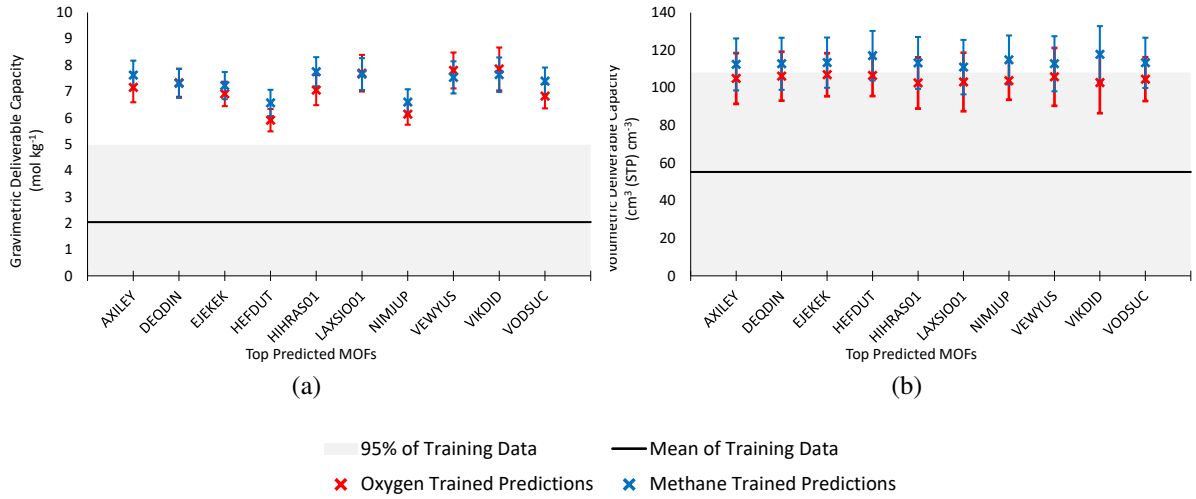


Figure 8.7: The predicted gravimetric (a) and volumetric (b) deliverable capacities at 30 bar storage for the ten most promising MOFs in comparison to the distribution of the training data. As predicted by a ROM trained using MOFs storing oxygen (red) and methane (blue)

8.6 Conclusion

With *ca.* 88000 structures of synthesised metal-organic frameworks (MOFs) available, identifying top materials for gaseous storage is difficult. Therefore, this paper investigates the pore properties affecting the gaseous deliverable capacity of synthesised MOFs. This paper has conducted two investigations using global sensitivity analyses (GSAs) before developing a reduced order model (ROM) prediction technique that results in faster more efficient material design and discovery for gas storage applications. A dataset of experimentally synthesised MOFs provides the information for learning producing a pathway to conduct GSAs discovering the most dominant pore properties to be accurately synthesised when designing MOFs. The GSAs was achieved by developing a Gaussian Process (GP) surrogate model enabling the semi-analytic calculation of Sobol' indices. Consequently, care was taken to validate the surrogate model ensuring inaccuracies were not carried onto the GSA.

The first GSA confirmed that the surface area and void fraction are the two most dominant pore properties while the heat of adsorption at 5 bar and the largest cavity diameter have a negligible effect on the deliverable capacity. Additionally, further GSAs conducted during this investigation has shown that the pore properties Sobol' indices are the same for different storage

pressures. Hence, pressure has little effect on changing the impact each pore property has on the deliverable capacity of MOFs. The second study of GSAs explored the effect the storage gas has on the influence of the pore properties by comparing oxygen storage to methane storage. Once again, the findings were consistent with the literature and the previous study showing the gravimetric surface area and void fraction to be the dominant pore properties. Further, it has shown that the pore properties interactions between themselves increase significantly when storing methane instead of oxygen. This aspect of the research showed the first-order Sobol' indices to be similar for MOFs storing either gas and so suggests that the impact of the pore properties is irrelevant to the storage gas.

This hypothesis was tested by training a ROM using MOF data storing one gas and then tested the ROM using MOFs storing the other gas. Importantly, our results provided evidence for good predictor capabilities irrelevant to the MOF storage gas shown by the standardised diagnostic values. Additionally, further analysis of the ROMs showed the capability to reduce the dimensions with confidence from five to four or three rotated dimensions. The ROM was applied to a practical MOF exploration example, to help identify the top performing MOFs for oxygen storage at 30 bar. Here, the methane trained ROM was found to be in agreement with the oxygen trained ROM, predicting VIKDID as the top performing MOF structure for oxygen storage with a VDC of 118 ± 15.0 cm^3(STP) cm^{-3} and a gravimetric deliverable capacity of 7.64 ± 0.655 mol kg^{-1}.

These findings provide a potential mechanism to improve on commonly used principal component analyses by using Active Subspaces instead of linear combinations. Future investigations are necessary to further understand the applicability of these kinds of conclusions to other storage gases. This would require improved understanding of correlated inputs and further additional data would be required that includes MOFs storing various other gases to further develop and confirm these initial findings. This work has only compared oxygen storage to methane storage which are similar in nature in the way that they interact with the MOF framework. The adsorption in mainly dominated by geometric properties and van der Waals forces, whereas the electrostatic forces are week. Therefore, the next stage would be to investigate whether the

ROM can be used to successfully make predictions for MOFs storing any gas or whether there is a select list of similar storage gases. For example, water and carbon dioxide are not considered similar to oxygen and methane as the polar and quadrupolar adsorbates have larger interactions from the surface chemistry. Hypothetically, numerous ROM's could be created dependent on the type of storage gas which could be classified based on these findings and future research. We believe that apart from looking at pore properties future research should also investigate the chemical and mechanical properties of MOFs.

Our novel contributions from this research can be summarised as follows:

- We have successfully predicted the GCD and VDC using an extensive database of synthesised MOF providing a basis for selecting appropriate structures for given applications.
- The predictions have been used to calculate a GSA to identify the most dominant pore properties and quantify their interactions depending on the application. Understanding how the storage gas or storage pressure impacts the behaviour of the pore properties enables researchers to consider the MOFs proposed usage before synthesising new materials.
- We presented a novel methodology for an Active Subset ROM which successfully predicts the gaseous uptake of MOFs for similar gas molecules.
- We have applied a ROM on a practical MOF exploration example, discovering the methane trained ROM to be in agreement with the oxygen trained ROM predicting VIKDID to be the top performing MOF structure for oxygen storage at 30 bar.

Overall, this study provides vital information regarding the importance of the pore properties for synthesised MOFs. It has provided an efficient method to predict the deliverable capacity of MOFs. Scientists wanting to predict and search for promising MOFs used for storing any gas can adopt the ROM.

Chapter 9

Robust Probabilistic Electricity Price Forecasting Using a Hybridisation of Gaussian Processes and Clustering

9.1 Abstract

As the deployment of intermittent renewable energy sources accelerates, consideration of electricity prices is becoming key. Therefore, to aid decision-making, accurate methods to forecast electricity prices are increasingly important. Here, we address mid-term probabilistic hourly price forecasting which is relatively overlooked in comparison to short-term point forecasting. As such, we introduce an efficient hybridisation method by combining Gaussian Processes with clustering to achieve accurate probabilistic electricity price forecasting four weeks ahead with an hourly resolution.

Given the volatility of electricity prices, a probabilistic method is required, therefore, the chosen accuracy measure is the continuous ranked probability score because it is sensitive to the entire predicted distribution. The proposed hybridisation method has its efficacy robustly tested by forecasting the price throughout 2019 in the UK.

It is, firstly, found that ten months of training data gives an optimal accuracy for forecasting

the electricity price. The little used, but essential, Diebold-Mariano test is applied to provide statistical evidence that the novel hybridisation method produces superior predictions when compared to other probabilistic methods. This probabilistic forecasting method provides the decision-maker with greater accuracy and reliability four weeks in advance so that the true hourly electricity price will be within the confidence limits predicted.

9.1.1 Keywords

Gaussian Process; Probabilistic Electricity Price Forecasting; Clustering; GP Hybridisation

9.2 Introduction

Increasingly, the rollout of intermittent renewable power sources with zero marginal cost of generation is resulting in a more dynamic electricity supply price curve. Coupled with variation in demand, this has the potential to increase the volatility of wholesale electricity prices (Weron, 2006; Steinert and Ziel, 2019). Within the existing electricity market structures, all of the proposed approaches to dealing with intermittency – storage, demand-side response and spatial interconnection - may be encouraged by exposing all customers, from domestic to industrial, to these variations. Hence accurate price forecasting will be of the utmost importance in the coming years (Weron, 2014). Additionally, probabilistic price forecasting is becoming increasingly valuable to stakeholders as a probability density function provides insight into the uncertainty in the predictions.

Forecasting electricity prices has been extensively researched in academia leading to a large number of papers published on this topic, for which there exist extensive reviews (Weron, 2014; Maritnez-Alvarez et al., 2015). Attempts to construct bottom-up models for electricity price based on predictions of demand and supply price curve at a given time run into difficulties due to the non-transparency of constraints in the system (Staffell and Green, 2016) and high levels of residual variance are associated with this approach (Pape et al., 2016). In particular, the electricity price forecast competition, EEM2016 EPF (Saraiva and Fidalgo, 2016) aided the interest in electricity price forecasting as competitors provided spot price forecasts for the daily market. Probabilistic forecasting methods are now gaining popularity for both load forecasting

(Rodrigues et al., 2014; Mordjaoui et al., 2017; Shepero et al., 2018) and price forecasting (Rafiei et al., 2017; Monteiro et al., 2018; Muniain and Ziel, 2020; Gao and Dowling, 2020).

Further, the volatility of electricity prices is due to both seasonal patterns and contributions from the supply-side. Hence, many factors influence electricity prices, such as economic growth, fuel price, generation, etc (Mandal et al., 2009). It has been argued that it is necessary to incorporate factors that reflect both the demand-side and supply-side (Keles et al., 2013) and it has been shown that including more variables will improve the prediction model by over 20% (Naumzik and Feuerriegel, 2020). Therefore, there is a role for statistical models that can capture various factors that cause sudden nonlinear variations and so improve the predictions of the electricity price.

Gaussian Processes (GPs) offer a computationally efficient machine learning method commonly used for probabilistic electricity price forecasting (Mori and Nakano, 2015; Kou et al., 2015; Gao and Dowling, 2020). The flexibility of GPs enables other mathematical techniques to be incorporated directly into the machine learning method (e.g. the calculation of Sobol' indices (Yeardley et al., 2020a) and experimental design (Yeardley et al., 2021)). Therefore, the predictive capabilities of GPs can be improved by combining them with clustering techniques that divide the data into smaller groups allowing data similarity to benefit the training of GPs.

Combining a classification method (such as clustering) with a regression technique (such as GPs) is a commonly adopted approach used in machine learning. One particular method that is often applied is Classification and Regression Trees (Breiman et al., 1984) which uses regression to create a predictive model and combines this with classification to explain how the output value can be predicted based on the input variables (Questier et al., 2005). However, the use of classification to improve the accuracy of regression is often applied using classification before regression (Khadem et al., 2020; Wang et al., 2014). Recently, cluster-boosted regression has been shown to improve predictive power in clinical decision-making (Rouzbahman et al., 2017; Kongburan et al., 2019) and to enhance solar forecasts (Najibi et al., 2021). Research has shown classification has a positive impact on the accuracy of electricity spot price forecasts, often leading to preferable outcomes (Fraunholz et al., 2021).

Specifically for GP electricity price forecasting, this combination was researched by Mori and Nakano (2015) who applied individual GPs to each cluster of inputs. This research uses the clustering of the inputs to be directly incorporated into the prediction of the output by conditioning a single GP producing a hybridisation method. The cluster number becomes an additional input variable that aids the GP learning process through data similarity.

This new hybridisation method is robustly compared to a GP without clustering and to the known clustering method by Mori and Nakano (2015). Additionally, a benchmark score enabling a guide for "good" predictions is calculated using a forecasting method based on gradient boosting. The comparison is achieved through an analysis of electricity price forecasting using real data from the UK (NordPool) providing true electricity prices for the analysis of each GP method's forecasting capabilities.

A review (Nowotarski and Weron, 2018; Ziel and Steinert, 2018) of the electricity price forecasting literature has identified mid-term (*ca.* one month ahead) and long-term (a year or longer) forecasting was rarely the focus of work. Ziel and Steinert (2018) found less than 1% of all of the published electricity price forecasting papers were related to mid-term or long-term probabilistic forecasting. Similarly, Nowotarski and Weron (2018) performed a bibliometric analysis showing 'neural network'-type methods to be the most popular forecasting technique. This review described the importance of 'maximizing sharpness subject to reliability' (Gneiting and Raftery, 2007). Here, reliability refers to maximising the number of observed prices covered by the predicted probability distribution and sharpness describes the width of the predicted probability distribution. However, the vast majority of the papers analysed in the bibliometric reviews showed forecast results to be using accuracy measures that should only be used for point forecasts (Yan and Chowdhury, 2013; Chakravarty et al., 2016). Deterministic error metrics (such as the mean absolute percentage error (de Myttenaere et al., 2016)) evaluate the performance of the forecasting method by comparing the mean of the prediction only and so neglect the reason we chose probabilistic forecasting to begin with. For example, Barta et al. (2015) developed an ensemble method using gradient boosting and benchmarked the experimental results using the popular ARMAX method (Yan and Chowdhury, 2013; Marín et al., 2018). But the two

forecasting methods were compared to the known prices using the root mean squared error and the mean absolute error, disregarding the importance of 'maximizing sharpness subject to reliability' (Gneiting and Raftery, 2007). Further, the mean absolute error tends towards infinity when the observed output nears zero which is a phenomenon that does occur in the UK electricity market especially as negative power pricing is becoming increasingly common (Hoover, 2006). Some authors have driven the further development of mid-term probabilistic electricity price forecasting by developing a nested combination of Monte Carlo simulation with spatial interpolation techniques and evaluating the new method using the pinball loss function (Bello et al., 2016). This error metric is similar to the continuous ranked probability score (CRPS) but conceptually less complex. Once again, the pinball loss function was used by Bello et al. (2017) to compare an innovative quantile regression technique with four other well-established electricity price techniques. To the authors' knowledge, these previous studies are the only mid-term probabilistic electricity price forecasting methods with an hourly resolution that evaluate the methods using a suitable error metric. However, neither of these papers provides statistical evidence that supports their new superior forecasting technique. To address this shortcoming, this paper identifies the importance of providing hypothesis tests to directly compare forecasting techniques.

This paper contributes directly towards scientific research and practice by:

- Developing a novel GP technique using hybridisation with clustering to perform mid-term electricity price forecasting with hourly resolution.
- Using relevant supply-side and demand-side factors to ensure the potential of the forecasting model (Naumzik and Feuerriegel, 2020).
- Comparing the results of each GP method using a robust hypothesis test, ensuring statistical evidence throughout the whole year.

The paper is structured as follows. Section 9.3 provides the mathematical background that has been used to develop the hybridisation technique for forecasting electricity prices. Afterwards, Section 9.4 describes the electricity price data before Section 9.3.4 describes the

mathematics required to evaluate probabilistic electricity forecasting methods. In Section 9.5.1 we analyse the amount of training data needed to increase the accuracy of the GP electricity price forecast. Section 9.5.2 provides a robust comparison between the developed hybridisation method and other electricity forecasting techniques and finally, conclusions are drawn in Section 9.6.

9.3 Methodology

9.3.1 Clustering

Clustering refers broadly to processes that can partition a set of vectors into subsets based on a similarity measure. In this work, each 5D vector consists of the hourly values of transmission level power demand and the generation of wind, solar, gas and nuclear power.

At a given point in time, the set of vectors forecast for the coming 4 weeks is combined with the set of vectors from the training period, each column in the array is standardised, and clustering is performed on the combined set. The length of the training period is varied, as described in Section 9.5.1. The cluster membership of each vector (an integer) is appended to it, and the set is then re-partitioned into training and test data to be used used as input for the GP.

In this work we apply two clustering techniques, k-means and hierarchical agglomeration, each with a heuristic applied in an unsupervised manner to judge an appropriate cutoff to avoid over-fitting.

K-means clustering is a popular method that is computationally efficient but requires the number of clusters, k, as an input. It is also non-deterministic as the initial positions of the cluster centroids are randomly assigned (George Seif, 2018). For this reason, a scan was performed whereby the process was repeated five times for each k value from 1 to 30. For each k value, the run that gave the lowest intra-cluster sum of squares (inertia) was retained. The inertia values were ordered in reverse (starting from k = 30) and the second derivative of the series taken. The value of k was taken as the index where the acceleration in inertia is greatest. This method is illustrated graphically in Goutte et al. (1999).

Hierarchical agglomerative clustering is a bottom-up deterministic approach that does not assume the number of clusters but requires a judgement on the minimum appropriate merge distance associated with genuine clusters. An acceleration plot is used, as described in Roberts and Brown (2020).

9.3.2 Gaussian Process (GP) Regression

A GP is an important stochastic process that corresponds to a class of normal distributions on a function as defined by (Williams and Rasmussen, 2006):

Definition 1 *A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

The posterior distribution is optimised to make predictions on test data for machine learning regression. This is done using standard Bayesian conditioning to take Gaussian priors and derive the predictive equations.

$$y(\mathbf{x}) \sim \mathsf{N}\left[\bar{f}(\mathbf{x}), \Sigma_y + \sigma_e^2\right] \tag{9.1}$$

where

$$\bar{f}(\mathbf{x}) := k(\mathbf{x}, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I})^{-1}\mathbf{y} = k(\mathbf{x}, \mathbf{X})\mathbf{K}^{-1}y \tag{9.2}$$

$$\Sigma_y := k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma_e^2 \mathbf{I})^{-1}k(\mathbf{X}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{X})\mathbf{K}^{-1}k(\mathbf{X}, \mathbf{x}) \tag{9.3}$$

whose mean $\bar{f}(\mathbf{x})$ and variance Σ_y is learnt from mapping the training inputs $\mathbf{X}$ to the observed responses $\mathbf{y}$, assuming the training data takes the form $\mathbf{y} = f(\mathbf{X}) + \mathbf{e}$ where $\mathbf{e}$ is an independent and identically distributed random error term.

At the heart of this lies the kernel function $k\colon \mathbb{R}^{i+d} \times \mathbb{R}^{j+d} \to \mathbb{R}^i \times \mathbb{R}^j$, expressing the correlation between responses to input samples of sizes $(i \times d)$ and $(j \times d)$. This work exclusively uses the automatic relevance determination (ARD) kernel (Wipf and Nagarajan, 2007):

$$k(\mathbf{x}', \mathbf{x}) \coloneqq \sigma_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')\Lambda^{-2}(\mathbf{x}-\mathbf{x}')^\intercal}{2}\right) \tag{9.4}$$

where Λ is a $(d \times d)$ *diagonal* positive definite lengthscale matrix. Regression uses the learned model to make predictions and so requires the optimisation of $d + 2$ hyperparameters, constituting of Λ, σ_f, and σ_e, through the maximum marginal likelihood $p[\mathbf{y}|\mathbf{X}]$ using the ROMCOMMA software library (Milton and Brown, 2019).

9.3.3 Hybridisation Method

On a given day, we wish to predict the hourly electricity price for 4 weeks ahead given the input variables. We apply a novel method, combining clustering techniques with a GP to create a hybrid clustering-GP approach.

Figure 9.1 and Figure 9.2 schematically present the methodology for predicting the electricity profile as the input data for both training and testing are put through the clustering method creating clusters of data that are similar. The cluster number then becomes an additional input variable used in the GP for both the training data and the test data. Consequently, our GP regression models must incorporate categorical (i.e. discrete) input variables because of the cluster number and the time-series data (such as the day of the week). Both clustering methods classify the electricity price input data into groups that are ordered depending on the similarity measure. Therefore, rational ordering enables the relaxation of categorical variables so that we could treat them as continuous variables. It is important to understand that this relaxation of discrete variables is a key difference to how GPs are ordinarily used.

9.3.4 Hypothesis Testing

In this section, we present a description of the hypothesis testing methodology used for evaluating the probabilistic forecasting methods being compared. The goal of probabilistic forecasting is to maximize sharpness subject to reliability. This means when evaluating forecasting methods, the choice of error metric is very important. In this work, we consider both the coefficient of determination (R^2) and the root mean squared error (RMSE) to evaluate the reliability of the point prediction. The percentage of outliers counts the number of observed values found

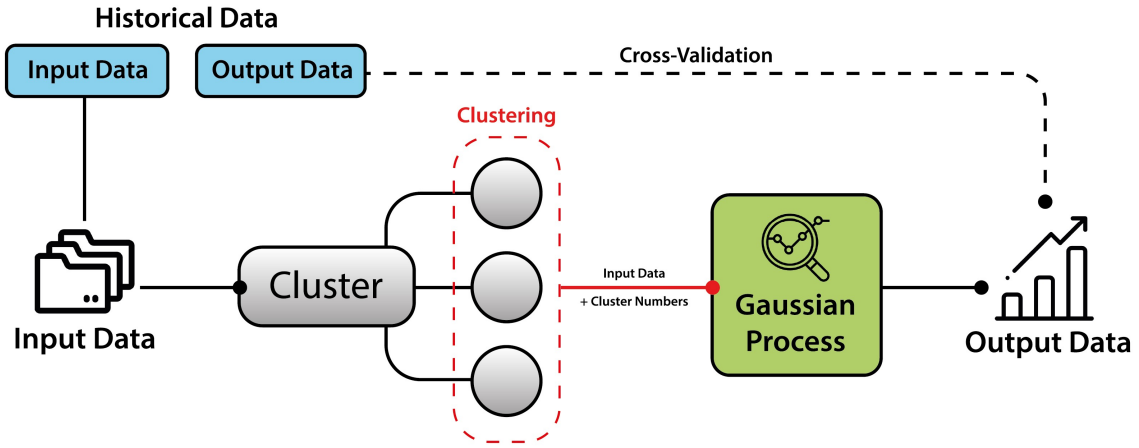


Figure 9.1: A flowchart showing the methodology for using historical data to train the hybridisation method of clustering and GPs.

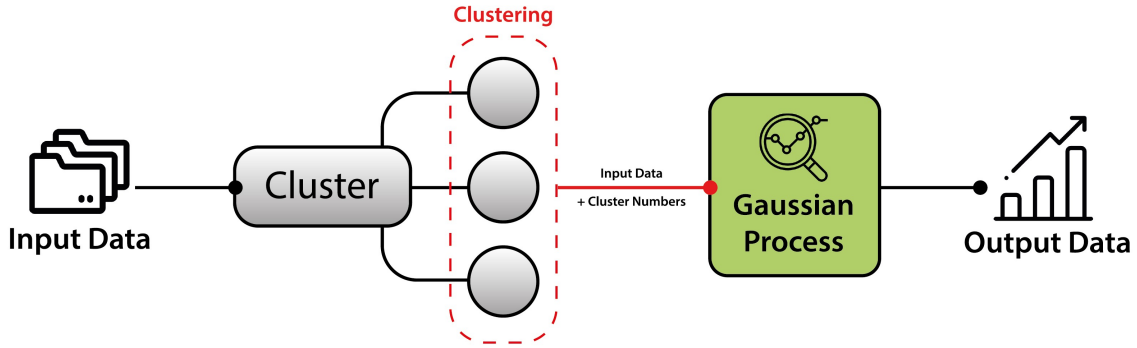


Figure 9.2: A flowchart showing the methodology for forecasting the price of electricity given input data using the hybridisation method of clustering and GPs.

outside of the predicted uncertainty distribution and so focuses on the reliability of the whole predicted distribution instead of just the point prediction.

However, as mentioned the sharpness of the distribution is also important and so this work uses the continuous ranked probability score (CRPS) (Matheson and Winkler, 1976; Gneiting and Raftery, 2007) as the main error metric used for hypothesis testing. Several advantages of the CRPS were presented by Hersbach (2000) including its ability to be decomposed into three main sections, showing the detailed behaviour of a forecast; reliability, uncertainty and resolution. For this research, the main advantage of the CRPS is that it is sensitive to the entire predicted distribution and for a deterministic forecast, the CRPS is equal to the mean absolute error (MAE). The combination of these two advantages means it further improves the MAE by directly incorporating the predictive continuous distribution for probability forecasts. This gives the CRPS a clear interpretation motivating its popularity for use as an accuracy metric for probability forecasts.

For a single observed electricity price, y_0, and the corresponding predicted price distribution, $\hat{F}(x)$, the CRPS is defined as:

$$CRPS = \int_{-\infty}^{\infty} \left(\hat{F}(x) - F_0(x) \right)^2 dx \tag{9.5}$$

where $F_0(x)$ is the Heaviside function:

$$F_0(x) = \begin{cases} 0, & \text{if } x - y_0 < 0, \\ 1, & \text{if } x - y_0 \geq 0 \end{cases} \tag{9.6}$$

The integral shown in Equation (9.5) can be calculated analytically for a Gaussian distribution (Gneiting and Raftery, 2007) enabling the CRPS to be implemented as a loss function. As such, we can formally evaluate the differences in two predictive performances using the Diebold-Mariano (DM) test (Diebold and Mariano, 1995). This test is used to statistically compare the forecasts of two predictive models. Here, the prediction results yield forecasting series for each method of equal size N. When comparing two of the electricity price forecasting

techniques we calculate two loss functions denoted by $CRPS_1$ and $CRPS_2$. Based on the following hypotheses we can test the differences between the two functions:

$$H_0 : E[CRPS_1] = E[CRPS_2] \tag{9.7}$$

versus

$$H_1 : E[CRPS_1] \neq E[CRPS_2] \tag{9.8}$$

In terms of a loss differential series

$$d_t = CRPS_1 - CRPS_2 \tag{9.9}$$

The DM test statistic is computed using the sample mean of d_t, $\hat{\mu}_{d_t}$, the standard deviation of d_t, $\hat{\sigma}_{d_t}$, and the number of predictions, N.

$$DM = \sqrt{N}\frac{\hat{\mu}_{d_t}}{\hat{\sigma}_{d_t}} \tag{9.10}$$

Diebold and Mariano (1995) show that when H_0 is true, the statistical test converges and so high values of DM are evidence against H_0. In this work, we use the DM test to compare the accuracy of two forecasts for each combination of forecasting methods by comparison of P-values corresponding to a significance level of $\alpha = 0.1$ (Ding, 2018). Due to the comparison of six prediction techniques, totalling 15 DM tests, the Bonferroni correction (Hommel, 1988) is required to account for the number of comparisons performed:

$$\alpha_{\text{critical}} = 1 - \left(1 - \frac{\alpha}{n}\right)^n \tag{9.11}$$

Hence, after applying a Bonferroni correction, a P-value greater than 0.0955 will fail to reject the H_0 showing the forecast of method 1 is statistically not significantly different to the forecast of method 2.

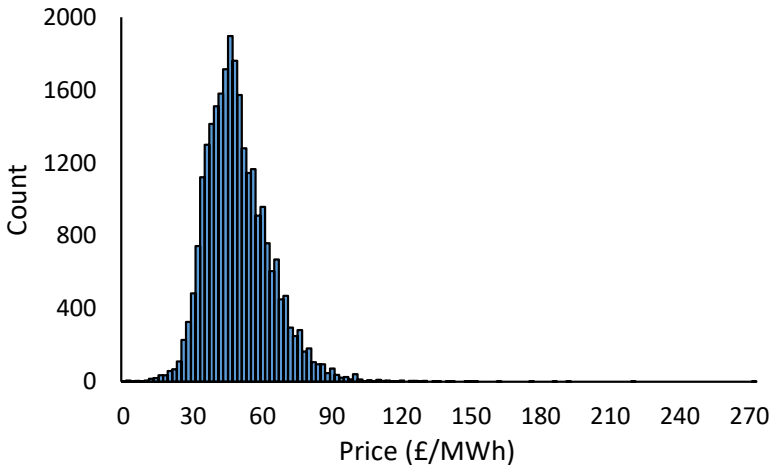


Figure 9.3: Histogram of the hourly electricity prices for the UK from January 1st 2017 to December 31st 2019. The bin sizes are in intervals of £2.00/MWh.

9.4 Case Study Data

The electricity price data used for this study was collected from NordPool providing hourly prices from the UK for 2017, 2018 and 2019. The distribution of electricity prices is shown in Figure 9.3 where the histogram indicates the high volatility of prices. In addition, the mean electricity price amounts to £48.56/MWh and the standard deviation is £14.14/MWh. As expected, the kurtosis is 6.34 indicating heavy tails caused by price spikes. Overall, the high volatility and extreme price spikes help to explain the difficulty in predicting accurate electricity prices.

Naumzik and Feuerriegel (2020) performed a sensitivity analysis to compare the use of potential variables known as external predictors (from the supply-side, demand-side, fuel-related and economic influences) in addition to time-dependent variables. The work conducted a DM test, providing statistical evidence that the accuracy of forecasting models including external predictors is superior to those not using external predictors. Specifically, the demand was ranked as the most important variable, followed by supply-side factors such as wind and solar generation. To enhance the model price forecasting capabilities, the electricity price data was combined with additional time-series data from ESO and BMReports. The extra variables used included the transmission system demand and generation data from wind, solar, gas and nuclear. Further, these variables can be forecast (BMReports) and so can be successfully used for four

weeks ahead electricity price forecasting. Together with five time variables (day of week, hour, date, month and year) and the previous price (electricity price four weeks previous) a total of eleven input variables were used to pair with the electricity price output.

9.4.1 Implementation of Hybridisation Method

In this work, the full set of electricity price data combined with the additional time-series data is used to implement the hybridisation methods for the first time. The novel hybridisation method is developed to improve GP's ability to forecast mid-term electricity prices. Therefore, four-week ahead predictions were made for each month in 2019 to analyse the novel method's ability in comparison to an ordinary GP.

We propose Algorithm 1 to show how to create a novel hybridisation method using the electricity price data. Please note, the total amount of data is not shown because a training data analysis will be performed to ensure the optimum amount of data is used for this specific study.

Algorithm 1

Require: $N \geq N_{test}$ ▷ The full data includes more datapoints than the test data.
Require: $M = 1$ ▷ The output is the electricity price.
Require: The number of input variables is equal to d
Ensure: Price predictions are made by a trained GP

1. Split entire data to create:
   - y_{test} of size N_{test} ▷ The observed prices for the forecasting period of testing.
   - X_{test} of size $N_{test} \times d$ ▷ The input variables the GP uses for predictions.
   - X_{train}
   - y_{train}
2. Cluster X_{train} and concatenate cluster number to each data point, making $d = d + 1$
3. Train GP to map $X_{train} \rightarrow y_{train}$
4. Custer X_{test} and concatenate cluster number to each data point, making $d = d + 1$
5. Forecast the price of electricity, $y_{predict}$, using X_{test} as input variables to the trained GP
6. Compare $y_{predict}$ with y_{train} by calculating an error metric for all predicted prices
7. Repeat to gather forecast results for an adequate amount of time (e.g. a full year)

9.5 Results

9.5.1 Training Data Analysis

It is well known that a limitation of GPs is that the training cost has $\mathcal{O}(N^3)$ complexity (Snelson, 2007). Thus, forecasting electricity prices can cause computational difficulties if large data sets are used for training the GPs. Further, literature has shown that forecasting performance is dependent on the calibration window (Pesaran and Timmermann, 2007). In this section, we focus on an experiment to further our understanding of the size of the calibration window and the amount of training data required to optimise a GPs electricity price forecasting capabilities.

Previous studies (Hubicka et al., 2019; Serafin et al., 2019; Marcjasz et al., 2018) have shown a significant accuracy gain can be achieved by investigating the calibration windows required for electricity price forecasting models. For example, (Serafin et al., 2019) has shown that by averaging a small number of both short-term and long-term calibration windows, then the best performing probabilistic forecast methods can be trained. However, the existing literature is not always in agreement with the optimum amount of training data, primarily because each study can be extended to investigate other datasets. One of the tough challenges for all researchers in electricity price forecasting is the volatility of electricity prices, ensuring not one price market is the same. Therefore, choosing an optimum amount of training data for every dataset is impossible. Instead, in this research, we focus on performing an experiment that will investigate the optimum calibration window for the electricity price data used in this case study.

To do this the 2019 electricity price data was used to test ordinary GPs trained using various amounts of previous price data as the independent variable. Altogether, nine ordinary GP forecasting techniques were trained using 2 to 18 months worth of previous data. Each of the ordinary GPs had the average time for training and testing each month of 2019 measured. Then the forecast prices were compared to the observed prices of 2019 and used to calculate the CRPS value for each ordinary GP forecasting technique.

Figure 9.4 presents a learning curve that shows the results from the time analysis experiment. All nine ordinary GP techniques have their CRPS value presented by the bar while the line

shows a decreasing trend for the time taken for each technique. As expected, as the number of months decreases, the amount of training data decreases and so does the amount of time taken to train and forecast each month's electricity price. However, even though a small amount of training data (such as two months) takes a negligible amount of time, it is not beneficial to the forecasting model as the CRPS value rises sharply with less data. The reasoning for this can be seen in Figure 9.5 where we compare the time-series predictions from three GP models using different training times as the calibration window. The forecasts made are probabilistic functions and so the predicted regions of 95% uncertainty are shown by shadings of yellow (two months calibration window), red (ten months), and blue (eighteen months). In both plots (January and July), the observed electricity price can be seen by the black markers.

Specifically, predictions made in January using a GP model trained on two months of data forecasts a constant price between £20/MWh and £106/MWh, with a mean prediction of £63/MWh. Clearly, the large uncertainty region shows the GP does not predict variations in electricity prices. Similarly, in July the yellow prediction area varies only slightly. Therefore, for this dataset, a calibration window of just two months of training shows results in a case of under-training.

On the other hand, an optimum amount of training data can be found for the calibration window. Figure 9.4 shows the CRPS value to decrease as the amount of training data decreases until just ten months of training data are used to achieve a score of just 0.3807 and more than a 70% reduction in time compared to when using eighteen months. This result provides evidence that an optimum amount of training data can be found for GPs. It is difficult to explain such results within the context of hourly electricity price forecasting nevertheless Figure 9.5 is used again to further understand the probabilistic forecasts from each GP. To compare the GP trained on eighteen months to that trained on ten months, Figure 9.5 shows a probability distribution in blue (eighteen months) and red (ten months). Each uncertainty area is close to the other as Figure 9.5 shows both the red and blue confidence intervals to surpass and shade over an equal amount of black markers. Showing, both GPs predict a price of electricity where the observed price is within the uncertainty region predicted an equal amount of times. However, the blue

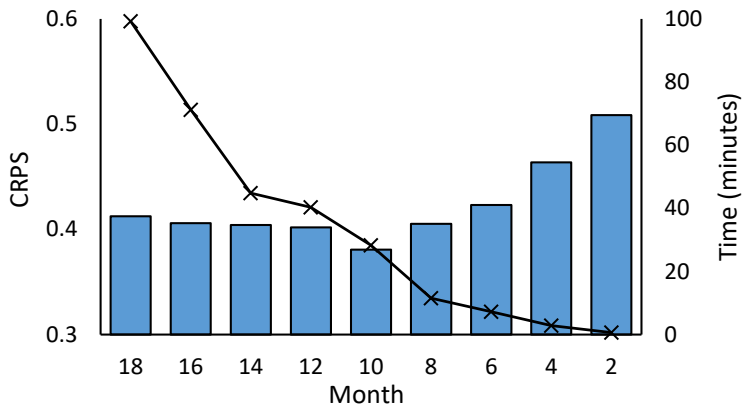


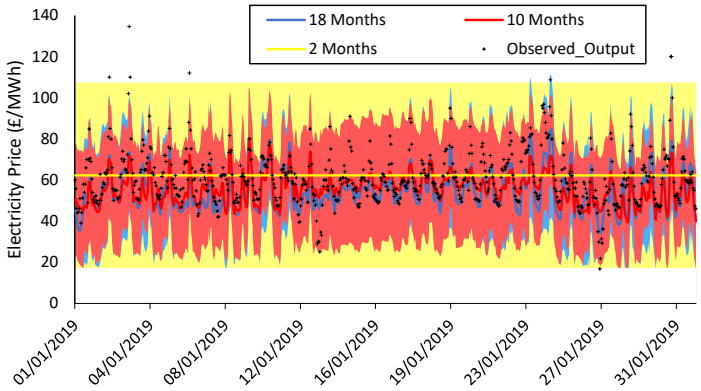
Figure 9.4: A learning curve measuring the CRPS and time taken for each GP using different amounts of training data.

region consistently reaches a higher electricity price than the red. Therefore, the GP trained on eighteen months of previous price data predicts a larger STD showing a larger uncertainty compared to the red distribution. This large uncertainty may be caused by over-training making the model more uncertain in its predictions. This result provides evidence that for this specific dataset the optimum amount of training data to use in the calibration window is ten months.

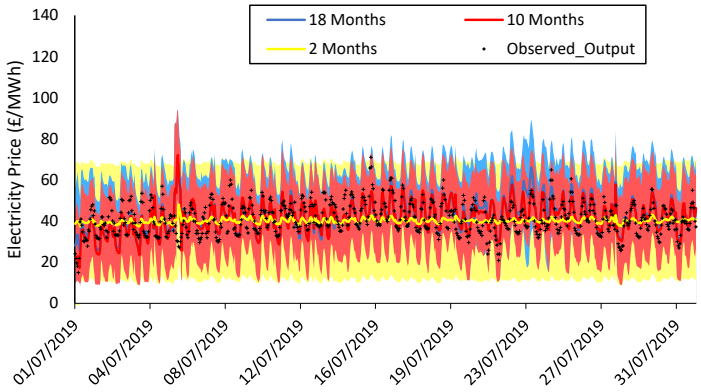
Overall, this simple computational experiment has been imperative to understand the optimum amount of data for this analysis. It has shown how GPs are sensitive to the amount of training data used. This is because too much data can cause over-training, creating uncertainty in forecasts. On the other hand, too little data forces predictions in areas not previously learnt by the GP. As discussed, we have obtained good results, achieving a low CRPS by forecasting electricity price data throughout 2019 with an ordinary GP using ten months of previous price data. We, therefore, use a calibration window of ten months for the remaining of this study.

9.5.2 Forecasting Results

This work has developed a novel hybridisation method to help improve GPs ability to forecast the price of electricity. Therefore, in this section, we directly compare the results of the proposed novel hybridisation methodology developed in Section 9.3 with those from an ordinary GP and a similar method that uses clustering to create individual GPs as developed by Mori and Nakano (2015). Both clustering forecast methodologies have been applied using hierarch-



(a)



(b)

Figure 9.5: A time series plot showing the predictions for GPs trained on 18 months, 10 months and 2 months of previous price data compared to the observed electricity prices in a) January and b) July.

ical clustering and K-means clustering to ensure the clustering technique does not impact the forecasting methodology. The focus of the experimentation is to understand the benefits of combining clustering with GPs for electricity price forecasting, hence the comparison to other GP methods. However, we have additionally applied a more conventional forecasting method based on gradient boosting (Zhang et al., 2018; Gaillard et al., 2016; Barta et al., 2015) so that we can see a benchmark score for an understanding of "good" forecast scores. Altogether, this requires the comparison of six electricity price forecasting techniques:

- An ordinary GP (GP)
- Novel hybridisation using K-means clustering (GP-Kc1)
- Novel hybridisation using hierarchical clustering (GP-Hc1)
- K-means clustering to make individual GPs (GP-Kc2)
- Hierarchical clustering to make individual GPs (GP-Hc2)
- Gradient boosting (GB)

A full year of monthly price predictions in 2019 was conducted for each method. This was conducted using ten months of historical data for training in a batch environment. Algorithm 1 was used with $N = 12432$ data points, corresponding to ten months of training data and 4 weeks of test data. The process is repeated twelve times in 2019 so that 4 weeks of every month has forecast results which are compared to the observed prices as described in Section 9.3.4.

Figure 9.6 shows the monthly performance of the six forecasting methods respective to the error metric being measured. The first two graphs, Figure 9.6a and Figure 9.6b present the coefficient of determination (R^2) and the root mean squared error (RMSE) respectively. An accurate forecasting method would present a high correlation and a low error thus these results would suggest that the gradient boosting method is consistently better than the GPs. However, these forecasts only take into account the mean point prediction and not the full probability. Thus, Figure 9.6c measures the percentage of true electricity prices outside of the predicted

confidence boundaries. In this case, the gradient boosting forecasting method consistently has more than 15% outliers, including the worst month when 83% of observed prices were found to be outliers in September. These high percentages could simply mean that the predicted confidence intervals from gradient boosting are far too small. The stark difference in the results between Figure 9.6a and Figure 9.6b compared to Figure 9.6c make the issue with measuring probability forecasts more apparent. It is therefore clear that we cannot measure the error from the mean point prediction alone due to the fact we are unable to analyse the confidence interval alone. As such, Figure 9.6d provides the CRPS which assesses the accuracy of a probabilistic forecast with respect to deterministic observations. Figure 9.6 reinforces our decision to conduct the comparisons made in this research using the CRPS as highlighted previously.

Wholesale electricity prices are increasingly volatile and throughout the change in seasons the prices are naturally difficult to predict. Figure 9.6 exemplifies this as the error-metric values for all six forecasting methods vary substantially through the months. For example, the low coefficient of determination in Figure 9.6a clearly shows the predictions in June are significantly worse than the other months. This particularly poor month of forecasts could be due to two extremely high prices, reaching £276.51/MWh and £219.32/MWh. Additionally, the average price for electricity during June is 1.14 standard deviations below the yearly average price. This appears to have caused predictions to be larger than the observed prices and the extremely high outliers increased the prediction error significantly. The deterministic error metrics provided in Figure 9.6 show poor point forecasts highlighting their unreliable nature. These findings support the notion that probabilistic electricity price forecasting is essential for policy-makers to consider the uncertainty in the predictions and make trusted decisions.

Figure 9.7 plots the average CRPS throughout a full year of testing for each of the six probabilistic electricity price forecasting methods suggesting the more conventional GB is the worst method due to its high CRPS of 0.5615. The comparisons show that the hybrid clustering-GP approach (GP-Hc1 and GP-Kc1) produces lower CRPS values than the other clustering variation developed by Mori and Nakano (2015) but a similar score to the ordinary GP. From these results, it is clear that this is a good choice for probabilistic electricity price forecasting

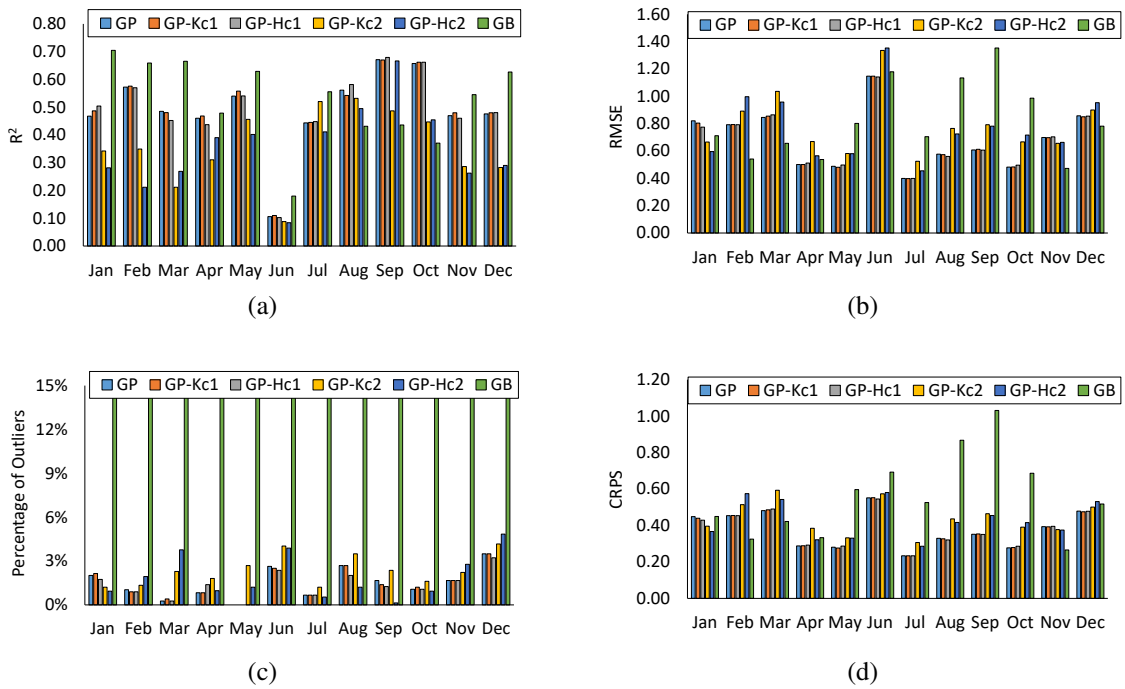


Figure 9.6: Bar charts comparing errors metrics for each of the six forecasting methods split for each of the twelve months of predictions.

due to the GP-Hc1 obtaining the lowest CRPS followed closely by GP-Kc1.

The results were statistically analysed using the DM test as described in Section 9.3.4 and presented in Table 9.1. Beginning with the GB column, consistently highlighted in red, the very low P-values (close to zero) reject H_0 showing significant differences between GB and the remaining methods. The hypothesis test yields statistical evidence supporting the advantages of GPs for predicting the UK electricity price throughout 2019 compared to the benchmark from the more conventional probabilistic electricity price forecasting method of gradient boosting. Additionally, the columns starting with GP-Kc2 and GP-Hc2 show a statistically significant difference from using clustering for individual GPs compared to an ordinary GP, GP-Kc1 and GP-Hc1 specific to the dataset used. We believe this justifies the use of our novel hybridisation method over clustering for individual GPs.

One concern about the findings from comparing the CRPS values was that the ordinary GP was very close to that of GP-Kc1 and GP-Hc1. Yet, results from the DM test demonstrate that this is not necessarily true. Table 9.1 compares the ordinary GP to the other forecasting

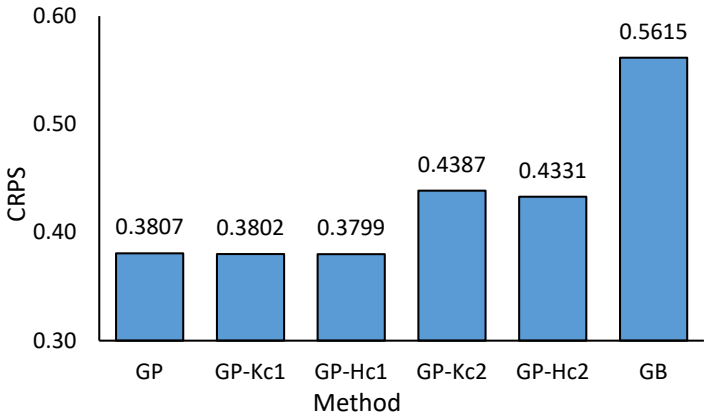


Figure 9.7: The CRPS for each of the forecasting methods.

method, by following along the GP row it is clear that there is statistical significance when using these methods for this dataset. When directly comparing the GP to GP-Kc1 and GP-Hc1, Table 9.1 shows orange filled cells indicating P-values close to the threshold but still below the 10% set previously. As with many hypothesis tests, this highlights shortcomings naturally arising from setting a statistical threshold. It must be noted that lowering it from 10% to 5% or 2.5% would find that for forecasting the price of electricity in the UK throughout 2019 the ordinary GP would have no statistical difference to our hybridisation method. However, from this standpoint, the novel hybridisation method does show an improvement in the probability of electricity price forecasts.

Table 9.1 highlights GP-Hc1 vs GP-Kc1 as green with a high P-value of 0.6534 and GP-Hc2 vs GP-Kc2 as green with a high P-value of 0.3275. Hence, we fail to reject the H_0 revealing that for both pre-clustering and hybridisation there is no statistical difference when using K-means clustering or hierarchical clustering to forecast the price of electricity in the UK throughout 2019. The results show that the clustering technique does not seem to impact the forecast results even though the way that said clustering technique is used impacts the results significantly. Therefore, the process by which clustering is used is more important than the actual clustering technique.

Table 9.1: The P-value from the DM test for the CRPS comparing each of the forecasting models. Cells filled in the green highlight when P-value > 0.100 so we fail to reject H_0 showing statistically the methods produce similar results.

Model	GP	GP-Kc1	GP-Hc1	GP-Kc2	GP-Hc2	GB
GP		0.0426	0.0739	<0.000	<0.000	<0.000
GP-Kc1			0.6534	<0.000	<0.000	<0.000
GP-Hc1				<0.000	<0.000	<0.000
GP-Kc2					0.3275	<0.000
GP-Hc2						<0.000
GB						

9.6 Conclusion

Accurate electricity price forecasting models are becoming increasingly important to aid in decision-making. In this work, we developed a novel GP-based hybridisation method to effectively forecast the price of electricity four weeks ahead with hourly resolution. The method utilises clustering techniques on the input variables to create an extra input variable for a GP regression model. This multistage hybridisation of GPs and clustering provides the GP with an extra input variable reducing the requirement for extrapolating the entire input space due to data similarity. The novel hybridisation technique builds on cluster-boosted regression techniques by combining the two into one regression technique instead of pre-classification creating numerous regressors.

The research rigorously tested the novel hybridisation method by comparing it to similar methods and a more conventional method by forecasting the electricity price for each of the twelve months in 2019 using real wholesale electricity price data for the UK from 2017 to 2019. To ensure the GPs were optimised to their maximum capability a training data analysis was conducted. The study found the optimum amount of training data to use for the GPs to be ten months. Following this, the comparison of forecast techniques was achieved by following the guidelines given by Nowotarski and Weron (2018). Thus, the results from each forecasting technique were analysed by calculating the CRPS before DM hypothesis tests were conducted to see whether two forecasting techniques achieve statistically different results.

The calculated CRPS values showed GP-Hc1 to predict the most accurate electricity prices in the UK throughout 2019, quickly followed by GP-Kc1 and the ordinary GP. The hypothesis

test provided statistical evidence that the novel hybridisation method is superior to an ordinary GP, a method that uses clustering to create individual GPs as developed by Mori and Nakano (2015) and a gradient boosting technique. However, when comparing GP-Hc1 vs GP-Kc1 high P-values from the DM test lead us to reject the null hypothesis providing statistical evidence that similar results were from two methods that only differ by a change in clustering technique. Alternatively, when comparing methods using the same clustering technique (e.g. GP-Hc1 vs GP-Hc2), a negligible P-value showed there is statistically a significant difference when the forecasting method itself changes (for example, when clustering is incorporated into one GP compared to when clustering creates independent GPs). From these results, an important insight is that the process by which clustering is used is much more important than the actual clustering technique for a multistage GP-based electricity price forecasting method.

In conclusion, this work has begun the development of a new GP-based hybridisation forecasting technique. The research used a hypothesis test to analyse a full year of predictions against similar forecasting techniques. These results showed the novel hybridisation technique has superior predictive capabilities. However, in reality, the electricity price market is extremely volatile and each year and country will have different price profiles to predict. This means the novel hybridisation technique requires continued analysis and hypothesis testing as time continues. Further, the technique needs to be tested on other electricity markets to ensure its predictive capabilities can be trusted. Successful results from this level of validation will allow a decision-maker to reliably estimate the degree of uncertainty in the electricity price efficiently, trusting that the true price is within the predicted distribution.

Future research should consider the potential effects of using supply-side and demand-side factors as input variables that need forecasting. Although the National Gird does provide forecast demand profiles, investigating how errors in these could affect the predicted electricity price would be beneficial for any decision-maker using this technique. In addition, this research aimed to directly contribute to probability electricity price forecasting by using a robust hypothesis test, ensuring statistical evidence throughout the whole year. While this was achieved, such a robust comparison to other methods provided implementation of the novel hybridisation

GP methodology impractical.

Chapter 10

Integrating Machine Learning techniques into Optimal Maintenance Scheduling

10.1 Abstract

Poor maintenance regimes often contribute to unplanned downtime, quality defects and accidents, thus it is crucial to apply an effective maintenance strategy to achieve an efficient and safe process. Industry 4.0 has brought about a proliferation of digital data and with it new opportunities to advance and improve the way maintenance activities are planned. Here, we propose a novel methodology that utilises machine learning to predict maintenance both faults and the repair time, and uses this to underpin a scheduling of maintenance activities. This can be used to plan maintenance, optimising the schedule for cost within the constraints of labour availability and plant layout. When applied to simulated data, using a simulated Fischertechnik (FT) model, this methodology reduces overall plant maintenance costs by reducing unplanned downtime and increasing maintenance efficiency. This work provides a promising first step towards improving the way maintenance tasks are approached in Industry 4.0.

10.1.1 Keywords

Predictive Maintenance; Optimisation; Maintenance Scheduling; Machine Learning; Time Estimation Model

10.2 Introduction

Machine maintenance is paramount to the process industry with regards to both safety and effectiveness. A poor maintenance system has a direct impact on costs, deadlines, quality, and accidents making it catastrophic to an organisation in terms of both operational performance and process safety. Unplanned downtime due to emergency repairs resulting from poor maintenance is estimated to have led to £13.1 billion in discrete manufacturing in 2016 (Thomas and Weiss, 2020). Given the important nature of maintenance, it must be conducted in parallel with a plant's normal operations to avoid compromising the plant's productivity levels (Kobbacy and Murthy, 2008).

Currently, industry is in a process of transformation towards Industry 4.0 where process automation and digitisation are becoming the norm (Gilchrist, 2016). One of the key factors that Industry 4.0 brings with it is the abundance of digital data, which can be used in the control and operation of a plant, improving production efficiency and managing process safety. Technologies using the Internet of Things provide the ability to measure and store large amounts of data from many sensors enriched by the control commands of actuators, transforming manufacturing environments into complex cyber-physical production systems (Gunes et al., 2014). The proliferation of data resulting from wide spread digitisation of manufacturing processes is opening up new opportunities to advance the way maintenance tasks are scheduled. These opportunities promise to improve process safety and reduced maintenance costs.

The most popular method that utilises machine learning and digital data is predictive maintenance (Zonta et al., 2020; Carvalho et al., 2019). Predictive maintenance utilises machine learning on real time sensor data to provide estimations of when maintenance is required on a machine (Yan et al., 2017). The most common predictive maintenance techniques use machine learning classification to predict a fault or failure occurring (Susto et al., 2015). The other techniques use machine learning regression to predict Remaining Useful Life of machines (Van Horenbeek and Pintelon, 2013) and forecast industrial aging processes (Bogojeski et al., 2021). A systematic literature review by Carvalho et al. (2019) showed that the most common machine

learning algorithms used are Random Forest, Neural Networks, Support Vector Machine and k-means clustering. Additionally, they found each machine learning method proposed was applied to a specific piece of equipment, for example turbines (Kumar et al., 2018), motors (dos Santos et al., 2017) and compressors (Prytz et al., 2015). For this reason, it becomes difficult to compare various machine learning algorithms as each study uses vastly different data for validation (Carvalho et al., 2019). Typically, predictive maintenance is employed on single machine systems. This approach lacks applicability to large industrial sites as it fails to consider the causal sequence of fault occurrence in process manufacturing. As such, further research that focuses on the application of predictive maintenance, integrating it within industrial sites to help develop maintenance workflow strategies instead of deriving novel machine learning algorithms is necessary.

Failure to properly consider time to complete maintenance tasks leads to prolonged downtime, increased technician time, and poorly executed jobs causing process incidents (Palmer, 2013). Estimating the time a maintenance task takes is a difficult task in large industrial settings, but it is essential to allow maintenance tasks to be accomplished more efficiently, leading to lower costs (Nyman and Levitt, 2006). Various methods are available to help estimate the time required including time study (Duffuaa and Raouf, 2015), predetermined motion time series (Alkan et al., 2016), or estimations based on past experience. However, these methods often lead to inaccurate errors leading to expensive overtime and rushed fixes.

Machine learning has been proven to be a valuable tool for time estimation models to aid the prediction of product manufacturing times (Liu and Jiang, 2005; Lingitz et al., 2018). Therefore, maintenance time estimation models using machine learning algorithms are possible. To the authors knowledge, Khalid et al. (2020) developed the only maintenance time estimation model. In their work, the historical work orders, functional locations and equipment related variables were used in machine learning algorithms to create better estimations of work hours for preventative maintenance tasks in an Oil Company. The work compared nine machine learning algorithms and found the Random Forest algorithm performed the best, decreasing the mean absolute error from 4.57 hours (when using estimation based on experience) to 3.83 hours.

The knock-on impact of a failure to consider maintenance time is a poor maintenance schedule. This is to the detriment of profitability (Vassiliadis and Pistikopoulos, 2001). Careful consideration of scheduling is required since performing more preventative maintenance will prevent serious failures but can cause unnecessary downtime and incur high maintenance costs. On the other hand, too little maintenance leads to corrective maintenance where tasks are performed due to failures occurring, thus, leading to process downtime and increased expenses. Traditional scheduling approaches rely on frequent periods of plant shutdown to perform maintenance tasks. Maintenance schedule optimisation is now a popular research topic due to its capabilities in increasing plant profits. It has been conducted for short-term scheduling of a multipurpose plant (Dedopoulos and Shah, 1995), long-term chemical plant turnarounds (Amaran et al., 2015), cleaning schedules in a furnace (Jain and Grossmann, 1998). These maintenance scheduling techniques are optimised to reduce costs focusing on when to schedule periodic maintenance. This, however, is not an optimal approach as it does not consider the relationship between maintenance and machine degradation. Thus, combining maintenance schedule optimisation with predictive maintenance offers opportunities to improve current practices. In literature, adaptive process scheduled have been developed to select acceptable process conditions based on predicted anomalies (Görür et al., 2021). Alternatively, condition-based maintenance scheduling enables dynamic maintenance scheduling based on the estimations from predictive maintenance (Mobley, 2002) to help create optimisation schedules (Jardine et al., 2006). Recently, preventative maintenance and optimisation were combined to create a schedule for a biomass boiler (Macek et al., 2017), a building heating ventilation and air conditioning system (Wu et al., 2021), and an ethylene cracking furnace system (Feng et al., 2021). Research studies have illuminated the combination of predictive maintenance and developed complex mathematical optimisation techniques, yet to the authors knowledge, no study to date has examined the optimisation of condition-based scheduling for a full industrial process.

Here, we present a novel methodology that can analyse the collection of machine sensor data to provide an optimum maintenance schedule through the combination of multiple machine learning techniques. We aim to provide a data-driven approach that automates the learning of

models for predictive maintenance and maintenance time estimation. The main objective of this work, therefore, is to build such a workflow that implements machine learning and optimisation in an approach that produces an optimum maintenance schedule. To do this, we first perform an investigation into the classification algorithms readily available for predictive maintenance. Predicting whether a fault has occurred in each machine can be used as the first tool in a promising application to develop robust maintenance scheduling in an industrial plant. We then seek to improve condition-based maintenance schedules by implementing a maintenance time estimation model that offers greater accuracy compared to physical observation. This work will build on the time estimation models previously created by Khalid et al. (2020) by estimating the maintenance time required to fix a predicted fault using live sensor readings as input variables. Finally, the workflow is completed by using the plant layout data, the predicted faults and the estimated maintenance time to optimise a full industrial industrial process.

The remainder of the paper is structured as follows. First, Section 10.3 presents the novel workflow created. Then a case study is described in Section 10.4 to which each stage of the workflow is tested on. Finally, Section 10.5 presents the analysis from each of the investigations and the results produced from the overall workflow.

10.3 Maintenance Policy Method

To create a robust maintenance schedule, we propose the workflow shown in Figure 10.1 that combines three techniques to enable accurate scheduling of maintenance tasks. Here, we outline the methodology behind the maintenance focusing on the algorithm that implements the policy.

Figure 10.1 shows historical maintenance records and sensor readings require feature engineering and separation to train and validate machine learning methods. This training process ensures the machine learning model can accurately predict whether a fault has occurred on each machine in the plant.

The maintenance time estimation then uses the sensor data of machines with predicted faults to estimate time required to fix the fault. In this scenario, it is assumed that the estimated time required to fix a fault is less than the time required to fix a failure that a fault could lead to.

Thus, by combining two machine learning methods the maintenance policy has used the sensor reading data to obtain the machines that are currently running with a fault and the time required to fix each fault before serious failures occur.

The results from both machine learning methods are passed on to the maintenance schedule optimisation. The maintenance schedule optimisation seeks to determine the most cost effective maintenance plan given the predicted faults for each machine and the time required to carry out such maintenance. At this stage, it is therefore essential to have a general understanding of plant procedures and the objectives required from the optimisation model. For example, knowledge of the availability of maintenance engineers and/or general layout of the plant are necessary to determine the number of simultaneous maintenance activities that may be carried out with the plant. Here, the mathematical model is explained in detail for the maintenance schedule optimisation with an objective to minimise the total cost comprising downtime costs, technician costs and the cost of part for replacement under corrective and/or preventative maintenance scenarios.

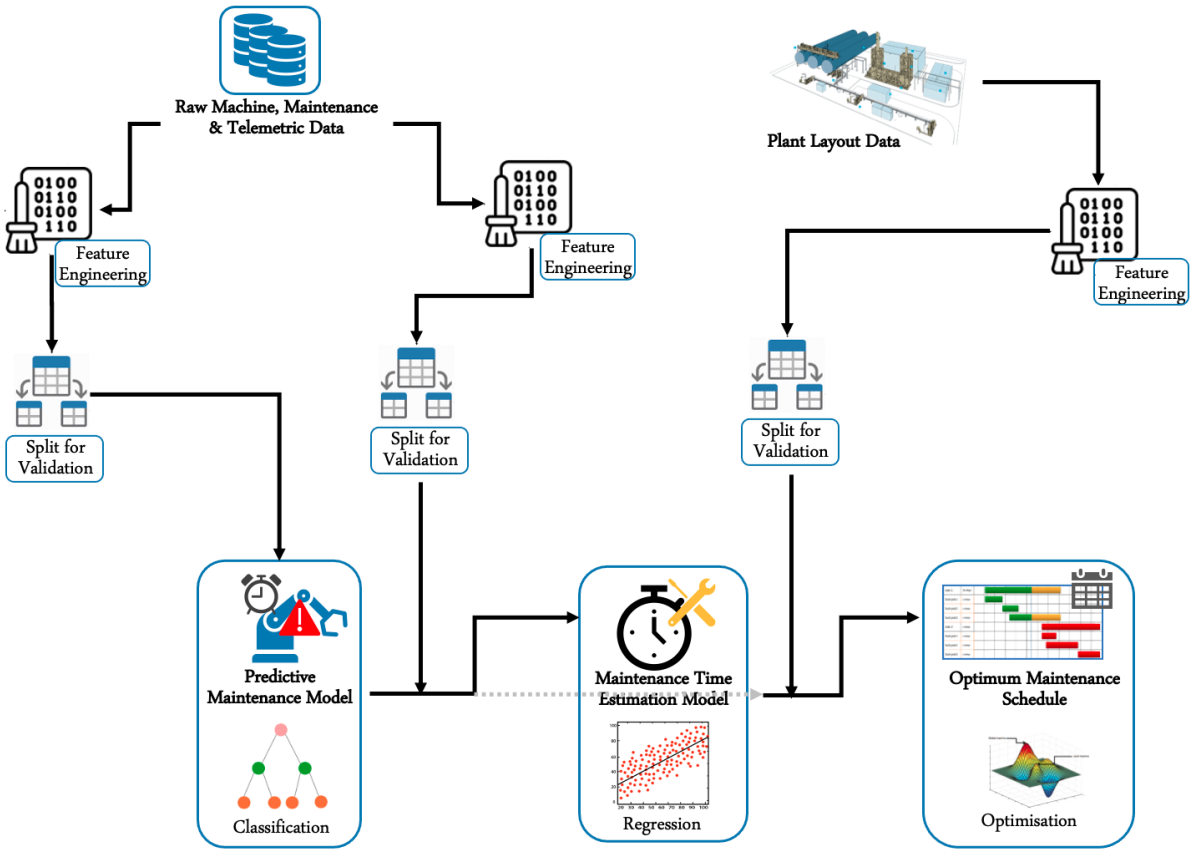


Figure 10.1: A flowchart of the ensemble of machine learning techniques used to produce an optimum maintenance schedule

10.4 Case Study

The following section describes the case study to which the methodology described in the previous section is applied.

10.4.1 Data

The application of the method outlines in Section 10.3 requires the availability of historical maintenance and complex sensor reading data related to an industrial plant. However, the availability of real-data from industry is extremely limited due to confidentiality issues. For this reason, data provided by a simulation, developed by Klein and Bergmann (2019), of a cyber-physical production system using a Fischertechnik (FT) factory model was used. The production plant consists of five workstations as described in an ontological knowledge base (Klein et al., 2019).

The large FT plant provides a realistic and challenging case study for detecting faults on 14 machines using 61 sensor readings indirectly related to each machine. The raw data is generated by multiple run-to-failure simulations where the sensor readings and the corresponding class are recorded. Altogether, the data includes $27,073$ data points recording 28 faults (29 classes in total).

The raw data provides sensor readings used as input variables for predictive maintenance and simulated faults can be used as test data to ensure accuracy. However, for the time estimation model, further information is required. Ideally, historical maintenance records where the time taken to fix faults and failures were recorded would provide outputs that can be used for training and testing. Given that the FT plant is a model simulation, this is not possible. Therefore, we have combined expert knowledge with that from the Ontology (Klein et al., 2019) to generate maintenance task times for each of the faults that occur and the failures that each fault could lead to. Additionally, Gaussian noise with zero mean and unit STD was added on to account for the variance between each maintenance task (for example, fixing a low wear fault may take 53 minutes previously but 62 minutes the next time). The average time for each maintenance task obtained in this manner is shown in Table 10.1.

The plant layout data (provided by Klein et al. (2019)) and the following assumptions were used to provide a basis for the maintenance scheduling:

- The plant operates at a sold out supply chain. When the plant is shutdown, it is losing profits.
- The plant is operating at a just in time manufacturing rate and overtime is not incurred.
- The plant is sequential so that a shutdown machine shutdowns the entire plant.
- Middle value products are packaged so that every hour of downtime is worth £10,000.
- A maintenance engineer can only work on one machine at a time.
- Multiple engineers are available.

- Each maintenance engineer is highly skilled in all departments and costs include planning time and overheads. This costs the plant £32.53 per hour of maintenance (Glassdoor Inc., 2021).
- Faults can be fixed without replacing parts.
- Failures are fixed by replacing parts that have a cost informed by an expert as shown in Table 10.1.

The objective of the maintenance scheduling is assumed to minimise cost under the assumptions mentioned. For the data given, we want to determine the maintenance schedule for each fault that occurs on a machine in the plant. This can be obtained using mathematical formulation described in Section 10.4.4.

Table 10.1: Summary of FT model data.

Fault No.	Machine No.	Machine	Plant Area	Fault	Fault Fix Time (hours)	Failure Fix Time (hours)	Cost of Part
1	1	Conveyor	txt15	Driveshaft Slippage	1	5	£400
2	2	Lightbarrier	txt15	Lightbarrier Mode 1	0.5	2	£100
3	2	Lightbarrier	txt15	Lightbarrier Mode 2	0.5	2	£200
4	2	Lightbarrier	txt15	Lightbarrier Mode 3	0.5	2	£300
5	3	M1	txt15	High Wear	1	7	£200
6	3	M1	txt15	Low Wear	1	7	£100
7	3	M1	txt15	Type 2 Wear	1	7	£200
8	4	Pneumatic lift	txt15	Leakage Mode 1	0.5	3	£100
9	4	Pneumatic lift	txt15	Leakage Mode 2	0.5	3	£300
10	4	Pneumatic lift	txt15	Leakage Mode 3	0.5	3	£500
11	5	Conveyor	txt16	Driveshaft Slippage	0.5	5	£400
12	5	Conveyor	txt16	Big Gear Tooth Broken	2	12	£500
13	5	Conveyor	txt16	Small Gear Tooth Broken	2	12	£200
14	6	Switch	txt16	Switch Mode 2	0.5	3	£200
15	7	Lightbarrier	txt16	Lightbarrier Mode 1	0.5	2	£100
16	8	M3	txt16	High Wear	0.5	7	£500
17	8	M3	txt16	Low Wear	0.5	7	£200
18	8	M3	txt16	Type 2 Wear	0.5	7	£300
19	9	Switch	txt17	Switch Mode 1	0.5	2	£100
20	9	Switch	txt17	Switch Mode 2	0.5	2	£200
21	10	Pneumatic Lift	txt17	Leakage Mode 1	0.5	3	£100
22	11	Transport Workpiece	txt17	Transport Workpiece Missing	0.5	9	£300
23	12	Pneumatic Lift	txt18	Leakage Mode 1	0.5	3	£100
24	12	Pneumatic Lift	txt18	Leakage Mode 2	0.5	3	£300
25	12	Pneumatic Lift	txt18	Leakage Mode 3	0.5	3	£200
26	13	Transport Workpiece	txt18	Transport Workpiece Missing	0.5	9	£200
27	14	Lightbarrier	txt19	Lightbarrier Mode 1	0.5	2	£100
28	14	Lightbarrier	txt19	Lightbarrier Mode 2	0.5	2	£200

10.4.2 Predictive Maintenance

Here, we compare five classification techniques; Decision Tree, Random Forest, Neural Network, AdaBoost and Quadratic Discriminant Analysis. The implementation used here was taken from the Python library, Scikit Learn (Pedregosa et al., 2011), the techniques are applied to the data to provide and insight into the promising tools readily available.

To ensure a robust comparison, the full data is split into training and test data based on complete simulations from start to finish so that each individual simulation leading to a fault are either only included in the test or training data set (Klein and Bergmann, 2019). Therefore, a fault that continued for multiple time steps was not found in both training and test data. A summary of the classification data is shown in Table 10.3, where the clear split between training data and test data can be seen for each of the 15 classes.

Table 10.3: A summary of the data used to train and test predictive maintenance techniques

Machine No.	**Train**	**Test**
No Fault	22,763	3,233
1	9	2
2	14	24
3	207	234
4	21	6
5	23	27
6	7	2
7	28	45
8	105	97
9	36	31
10	16	11
11	13	20
12	42	34
13	6	0
14	13	4
Total	23,303	3,770

10.4.3 Time Estimation Model

The second stage of the proposed workflow involves creating a time estimation model. Thus, here the study provides a comparison between Gaussian Processes, Neural Networks, Gradient Boosting Regression, Support Vector Regression and Random Forests to evaluate the most

promising regression technique to estimate the required time for the maintenance policy.

Irrespective of the technique being used, the models are tested in the same way. Here, the models use the simulation data that consists of the 61 sensor readings and the machine number as input variables, and the output is the time required to fix the fault in hours. The distribution of the maintenance time in hours is shown in Figure 10.2.

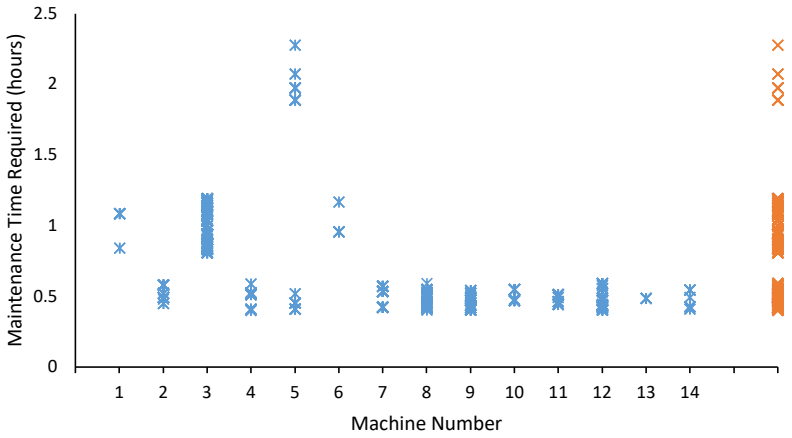


Figure 10.2: The maintenance time required to fix a fault for each machine in the FT model plant. At then end of the x-axis, the total spread for all of the machines is shown.

Here, cross-validation was chosen due to only $1,077$ data points classed as faults being used from the original FT model simulation. Therefore, this study split the data into 15 datasets so that each complete run to failure simulation is kept together in the same set, creating 14 historical maintenance datasets for training, and one remaining dataset as new work orders to be used as validation. This procedure is repeated to ensure all the data points are used for testing the regression techniques.

10.4.4 Maintenance Schedule Optimisation

The maintenance schedule optimisation utilises the predictions from the previous two methodologies to produce an optimal maintenance schedule. This work focuses on providing a maintenance schedule with the minimum cost for the case study. Therefore, in this section, the mathematical formulation for the maintenance schedule optimisation is described.

Mathematical Formulation

The objective of the maintenance schedule optimisation is to minimise the costs to the plant given system constraints resulting from plant procedures, plant layout data, and other operational considerations. Here, we present the mathematical optimisation model with a full nomenclature given in Section 10.7.

The problem is posed as follows:

Given:

- a set of machines (devices) in a plant;
- a set of possible faults per machine, the (predicted) time of occurrence and whether or not it causes a plant to be shutdown;
- estimated maintenance times required by each fault per machine before and after failure occurs;
- cost of parts and engineering personnel for each fault that occurs within a machine;
- downtime cost of the plant and machine;
- maximum number of available engineers for maintenance activities;

Determine:

- the maintenance schedule for each fault that occurs on a machine within the plant;

So as to:

- minimise the total cost over the time period of consideration. The cost comprises of the plant downtime cost, engineering personnel cost, and the cost of replacing machine parts during maintenance.

In addition to the assumptions previously stated, the following also apply:

- An engineer may only carry out a single maintenance activity for a specific fault at any given time;

- Only three machine states are considered: 'Running', 'Failed' or 'Under Maintenance';

- A maintenance activity may start on a machine for a predicted fault before the time of occurrence, further referred to as preventative maintenance (PM);

- Different times are allocated to preventative maintenance activities and maintenance activities carried out after a fault has occurred ie. a machine already in a failed state requires longer maintenance time (corrective maintenance, CM);

For any given time period, all machines must be in only one state (eq. (10.1)), and the total number of machines being simultaneously maintained must not surpass the available number of engineers (eq. (10.2)).

$$\sum_{s} S_{ist} = 1 \qquad \forall\, i, t \tag{10.1}$$

$$\sum_{i} S_{ist} \leq N^{p} \qquad \forall\, s \in \{M\} \tag{10.2}$$

In order to represent the states of each machine, and their transitions, over the time of consideration, eqs. (10.3) - (10.8) are introduced. The state of machines at the first time period ($t = 0$) are defined using eqs. (10.3) and (10.5), with a machine in a failed state if a fault is predicted to occur, else 'Running'. The machine remains in a 'Running' state at time t except a failure occurs or maintenance activity begins (eq. (10.4)).

$$S_{i,'R',0} \geq 1 - F_{i,0} \qquad \forall\, i \tag{10.3}$$

$$S_{i,'R',t} \geq S_{i,'R',t-1} - S_{i,'M',t} - S_{i,'F',t} \qquad \forall\, i, t > 0 \tag{10.4}$$

$$S_{i,'F',0} \geq F_{i,0} \qquad \forall\, i \tag{10.5}$$

A machine may only be in a 'Failed' state at any time period if a fault occurs (eq. (10.6)), and

it may only transition to the state 'Under Maintenance' (eq. (10.7)).

$$S_{i,'F',t} \leq S_{i,'F',t-1} + F_{it} \qquad \forall\, i, t > 0 \tag{10.6}$$

$$S_{i,'F',t} \geq S_{i,'F',t-1} - S_{i,'M',t} + F_{it} \qquad \forall\, i, t > 0 \tag{10.7}$$

Finally, the end of maintenance activities on a machine is tracked using the binary variable W^e_{it} in eq. (10.8).

$$S_{i,'M',t} \geq S_{i,'M',t-1} - W^e_{it} \qquad \forall\, i, t > 0 \tag{10.8}$$

The start and end times of maintenance activities are tracked using the binary variables W^s_{it} and W^e_{it} respectively, which are evaluated using eq. (10.9). Eqs. (10.10) and (10.11) ensure these variables may only take a value of 1 when the machine is under maintenance. Maintenance activities may not also start and end at the same time period (eq (10.12)).

$$W^s_{it} - W^e_{it} \geq S_{i,'M',t} - S_{i,'M',t-1} \qquad \forall\, i, t \tag{10.9}$$

$$W^s_{it} \leq S_{i,'M',t} \qquad \forall\, i, t \tag{10.10}$$

$$W^e_{it} \leq S_{i,'M',t-1} \qquad \forall\, i, t \tag{10.11}$$

$$W^s_{it} + W^e_{it} \leq 1 \qquad \forall\, i, t \tag{10.12}$$

Given that faults on each machines are being predicted, it becomes important to allow for preventative maintenance actions in the schedule, as opposed to the traditional corrective maintenance actions after a fault occurs. This feature is incorporated into the model using eq. (10.13).

$$W^s_{it} \leq \sum_{t'=0}^{t+\upsilon\rho^p} F_{it'} - \sum_{t'=0}^{t-1} W^s_{it'}\Big|_{t>0} \qquad \forall\, i, t \tag{10.13}$$

where υ denotes the number of time periods before fault occurrence maintenance activities are allowed to start for each fault.

As each machine can have a number of possible faults that can cause failure with different cost implications for maintenance, it is also important to predict each machine's state per fault that occurs and maintenance activity required. Eq. (10.14) evaluates the start and end times of a maintenance action (PM or CM) for each fault that can occur on a machine. The binary variable S^m_{ift} takes a value of unity if a machine is currently under maintenance at time t for fault f.

$$W^{s'}_{ift} - W^{e'}_{ift} \geq S^m_{ift} - S^m_{if,t-1} \qquad \forall\, (i,f) \in I^f, t \tag{10.14}$$

$$W^s_{it} = \sum_{f:(i,f)\in I^f} W^{s'}_{ift} \qquad \forall\, i,t \tag{10.15}$$

$$W^e_{it} = \sum_{f:(i,f)\in I^f} W^{e'}_{ift} \qquad \forall\, i,t \tag{10.16}$$

It is assumed that only one fault is corrected during a maintenance activity, hence the start time of a maintenance activity for a machine (evaluated using W^s_{it}) can only be mapped to one fault (eq. (10.15)). The same applies to the end times of maintenance activities (eq. (10.16)).
The binary variable S^m_{ift} may only take a value of unity if a corresponding fault is predicted to occur on a machine (eq. (10.17)), and only one fault is corrected (eq. (10.18))

$$S^m_{ift} \leq \sum_{t'=0}^{t+\upsilon\rho^p} F^f_{ift'} \cdot S_{i,'M',t} \qquad \forall\, (i,f) \in I^f, t \tag{10.17}$$

$$S_{i,'M',t} = \sum_{f:(i,f)\in I^f} S^m_{ift} \qquad \forall\, i,t \tag{10.18}$$

$$\tag{10.19}$$

A similar set of constraints to eqs. (10.17) and (10.18) are used to determine when a machine is in a failed state for a specific fault (eqs. (10.20) - (10.21)). In order to accurately calculate the downtime costs, eqs. (10.22) and (10.23) are introduced. Eq (10.22) determines if any machine under consideration in the plant is in a failed state for each time period using the binary variable $\overline{S}^f_t$. In eq (10.23) on the other hand, the binary variable S^f_t only takes a value of unity when a

machine with a fault f which causes plant shutdown is in a failed state.

$$S_{i,'F',t} = \sum_{f:(i,f)\in I^f} S^f_{ift} \qquad \forall\, i,t \tag{10.20}$$

$$S^f_{ift} \leq \sum_{t'=0}^{t+\upsilon\rho^p} F^f_{ift'} \cdot S_{i,'F',t} \qquad \forall\,(i,f)\in I^f, t \tag{10.21}$$

$$\overline{S}^f_t \geq S_{i,'F',t} \qquad \forall\, i,t \tag{10.22}$$

$$S^f_t \geq \mu_{if} S^f_{ift} \qquad \forall\,(i,f)\in I^f, t \tag{10.23}$$

In order to properly attribute maintenance duration's for corrective and preventative maintenance actions, eqs. (10.24) - (10.31) are introduced. The time difference between fault occurrence and the start of a maintenance activity is determined using eq. (10.24) when corrective ($\kappa^f_{if} > 0$) or preventative maintenance occurs ($\kappa^p_{if} > 0$) for each fault f on a machine i.

$$\kappa^p_{if} - \kappa^f_{if} = \sum_t t \cdot (F^f_{ift} - W^{s'}_{ift}) \qquad \forall\,(i,f)\in I^f \tag{10.24}$$

A binary variable, γ_{if}, which takes a value of 1 when PM actions are performed is then evaluated using eq. (10.25), and big '*M*' constraints introduced to ensure only one of κ^f_{if} and κ^p_{if} take non-zero values for each machine-fault pair.

$$\gamma_{if} \leq \kappa^p_{if} \qquad \forall\,(i,f)\in I^f \tag{10.25}$$

$$\kappa^p_{if} \leq \hat{M}\cdot\gamma_{if} \qquad \forall\,(i,f)\in I^f \tag{10.26}$$

$$\kappa^f_{if} \leq \hat{M}\cdot(1-\gamma_{if}) \qquad \forall\,(i,f)\in I^f \tag{10.27}$$

The number of contiguous times in which a machine with a particular fault is under maintenance is then enforced using eq. (10.28) depending on whether CM (M^f_{if}) or PM (M^p_{if}) actions are deemed optimal by the model.

$$S^m_{ift} = \gamma_{if}\sum_t \delta^p_{if\theta} W^{s'}_{if,t-\theta+1} + (1-\gamma_{if})\sum_t \delta^f_{if\theta} W^{s'}_{if,t-\theta+1} \qquad \forall\,(i,f)\in I^f, t \tag{10.28}$$

where

$$\delta^f_{ift} = \begin{cases} 1, & t < M^f_{if} \\ 0, & t \geq M^f_{if} \end{cases} \qquad \forall\ (i,f) \in I^f, t \tag{10.29}$$

$$\delta^p_{ift} = \begin{cases} 1, & t < M^p_{if} \\ 0, & t \geq M^p_{if} \end{cases} \qquad \forall\ (i,f) \in I^f, t \tag{10.30}$$

Finally, eq. (10.31) enforces the maintenance times (corrective or preventative) for each machine-fault pair only if the fault is predicted to occur.

$$\sum_t S^m_{ift} = F^f_{ift}\big(M^p_{if}\gamma_{if} + M^f_{if}(1-\gamma_{if})\big) \qquad \forall\ (i,f) \in I^f \tag{10.31}$$

The objective function is defined by eq. (10.32) and minimises the total cost accrued by the plant over the time of consideration. It comprises the sum of the plant downtime cost, machine downtime cost, personnel and parts cost for every maintenance activity per machine-fault pair.

$$min \sum_{it} \left(\hat{C}^d \cdot S^f_t + \hat{C}^f \cdot \overline{S}^f_t + \sum_{f:(i,f)\in I^f} \hat{C}^e_{if} \cdot W^{s'}_{ift}(1-\gamma_{if}) + \hat{C}^p \cdot S_{i,'M',t} \right) + N^p \tag{10.32}$$

subject to eqs. (10.1) - (10.31). This results in a mixed integer non-linear programming (MINLP) model owing to the bilinear terms in eq. (10.28) which can be solved using popular MINLP or mixed integer quadratically constrained programming (MIQCP) solvers.

10.5 Results

10.5.1 Predictive Maintenance

As described in Section 10.3, this research analysed five machine learning classification techniques and evaluated each using three popular classification error diagnostics. The standard measures considered are Precision, Recall and the F1 Score, each calculated as weighted averages based on the number of examples per class. An accuracy value was not chosen due to

Table 10.4: Resulting diagnostic values from the predictive maintenance

	Precision	Recall	F1 Score
Decision Tree	0.839	0.878	0.838
Random Forest	0.735	0.858	0.792
Neural Network	0.788	0.861	0.799
AdaBoost	0.736	0.858	0.793
Quadratic Discriminant Analysis	0.880	0.883	0.877

the imbalance between classes as shown in Table 10.3, where $3,233$ test points are in the class "No Fault" out of a total $3,770$. Therefore, if the classification algorithm predicted "No Fault" constantly would achieve an accuracy score of 85.76%. Each error metric counts the number of true positive predictions, but the precision score represents this as a ratio to all of the predicted positives, whereas, the recall is the ratio to all actual positives. To combine them both into one metric, the F1 scores weight both recall and precision equally.

Table 10.4 show the weighted averages obtained for each of the machine learning methods. As can be seen, the best results were obtained by the Quadratic Discriminant Analysis model due to it having the largest value for the precision, the recall and the F1 score. The results for all five machine learning models are satisfactory, but to get the best from the maintenance policy, trust in the predicted faults is of highest priority. Therefore, for the given case study, the Quadratic Discriminant Analysis was the chosen classification model to predict the faults passed on to the time estimation model.

10.5.2 Time Estimation Model

Once the predictive maintenance model has been trained, the next stage would be to ensure the time scheduled for maintenance is accurately fed into the optimisation model. Hence, validation of a time estimation model is of critical importance.

Once again, the case study was used to test five machine learning algorithms for regression, these being; Gaussian Processes, Neural Network, Gradient Boosting, Support Vector Regression, and Random Forest. Using these algorithms provides a variety of approaches that are widely available for use (Pedregosa et al., 2011; Milton and Brown, 2019) and so increases the chances of finding the optimum model.

Table 10.5: Resulting diagnostic values from the time prediction

	R squared	Standardised RMSE	RMSE (hours)
Gaussian Process	66.3%	0.600	0.202
Neural Network	57.9%	0.670	0.226
Gradient Boosting	56.9%	0.681	0.230
Support Vector Regression	51.8%	0.723	0.244
Random Forest	51.7%	0.720	0.243

As previously stated in Section 10.4.3, the machine learning models are tested using 15-fold cross-validation, ensuring every fault available as data were used for testing. Once again, three popular error diagnostics were chosen to evaluate each regression technique. For regression, these diagnostics were the coefficient of determination (R^2), the standardised root mean squared error (RMSE (-)), and the root mean squared error of time (RMSE (hours)).

Table 10.5 presents the values calculated for the validation of each method. Clearly, all five methods have an average R^2, above 50% but below 70%, providing an indication of the amount of variation in the observed maintenance time ascribable to the estimated maintenance time. The two RMSE diagnostics reveal the distance each prediction is away from the true value on average, but the RMSE (-) compares standardised values, whereas, the RMSE (hours) puts the error in the predictions into a more visual context for time estimation using units of hours. The average actual work is 0.696 hours, meaning an average offset of less than 0.35 hours gives a 50% time overlap to consider when setting work tasks. In comparison, previous work from Khalid et al. (2020) has shown using traditional methods the offset is 87%, but the research used machine learning algorithms to reduce this to 73%. Therefore, although predictions are initially seen to be satisfactory, the results from this time estimation investigation show the techniques applied provide an improved accuracy in predicting the maintenance time.

Importantly, Table 10.5 shows the Gaussian Process has the best performance as it has the largest R^2 value and lowest values for error. As such, the Gaussian Process is chosen to be used in this case study for a time-estimation model. Figure 10.3 clearly shows the performance of the Gaussian Process by plotting residuals of the predicted maintenance hours vs the actual maintenance hours. It can be clearly seen that the Gaussian Process performs exceptionally well, often predicting values close to the observed time.

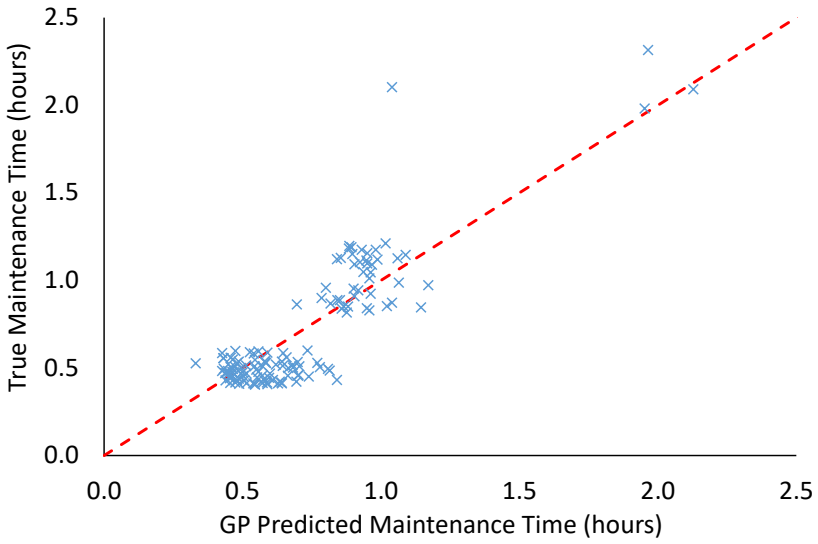


Figure 10.3: The residuals predicted by the Gaussian Process regression model.

10.5.3 Maintenance Schedule Optimisation

Following the preventative maintenance and time estimation model implementation, we use these models' outputs to obtain a cost optimal maintenance schedule for the case study presented in Section 10.4.1.

The data implemented was split into three cases based on calendar dates of fault occurrence. Figure 10.4 - Figure 10.6 shows the Gantt chart produced for each case. Each case corresponded to a day's worth of data showing the machines and their relative time periods of fault occurrence with each time unit corresponding to a five minute period. The maintenance scheduling model was solved using Gurobi 9.0.3 on 2 threads of an Intel i7-3615QM processor with 16GB RAM.

In order to demonstrate the benefits of our proposed maintenance policy two strategies are considered:

- Condition-based scheduling - using predictive maintenance model outputs and performs maintenance activities as soon as a fault occurs based on available resources;
- Our proposed workflow which extends the condition-based scheduling allowing for preventative maintenance actions. Hence, maintenance actions can be performed on machines before faults occur, which lead to failures in a preventative manner. This allows

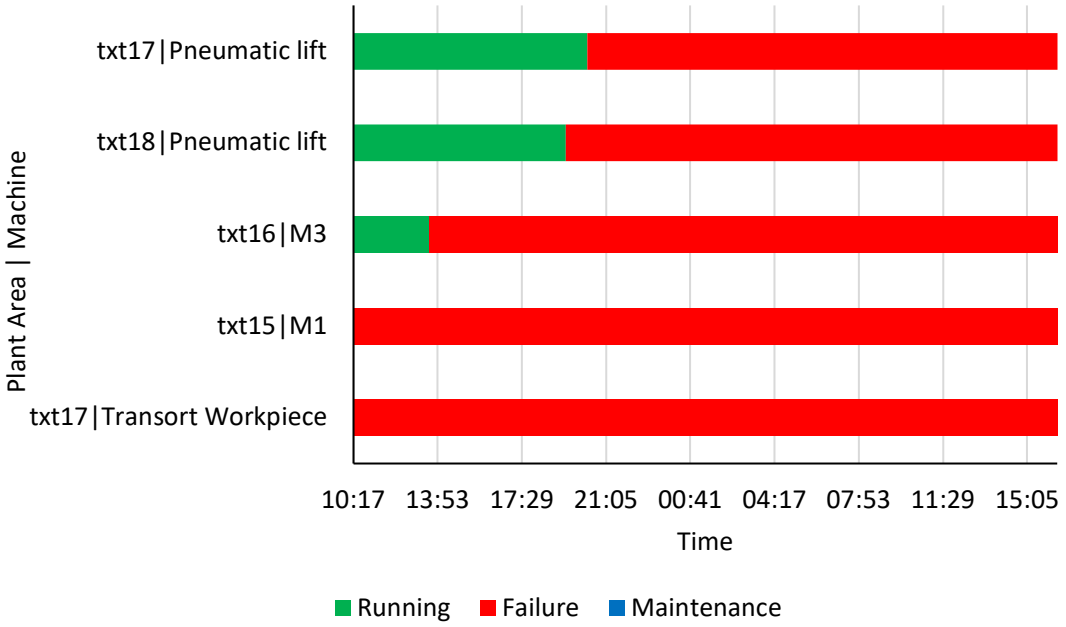


Figure 10.4: Case 1 Input Data

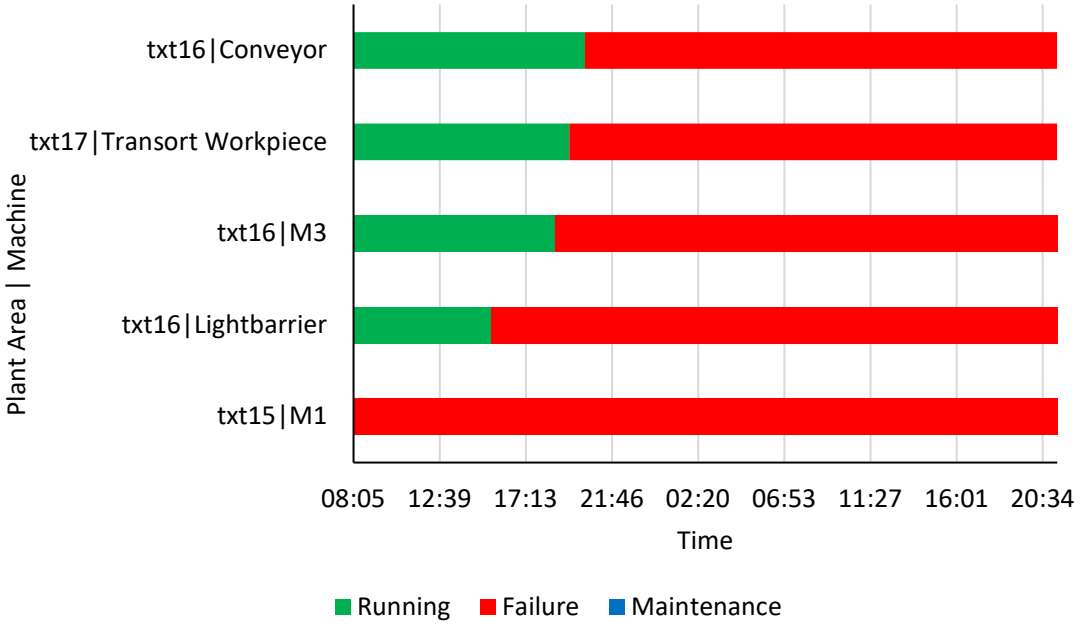


Figure 10.5: Case 2 Input Data

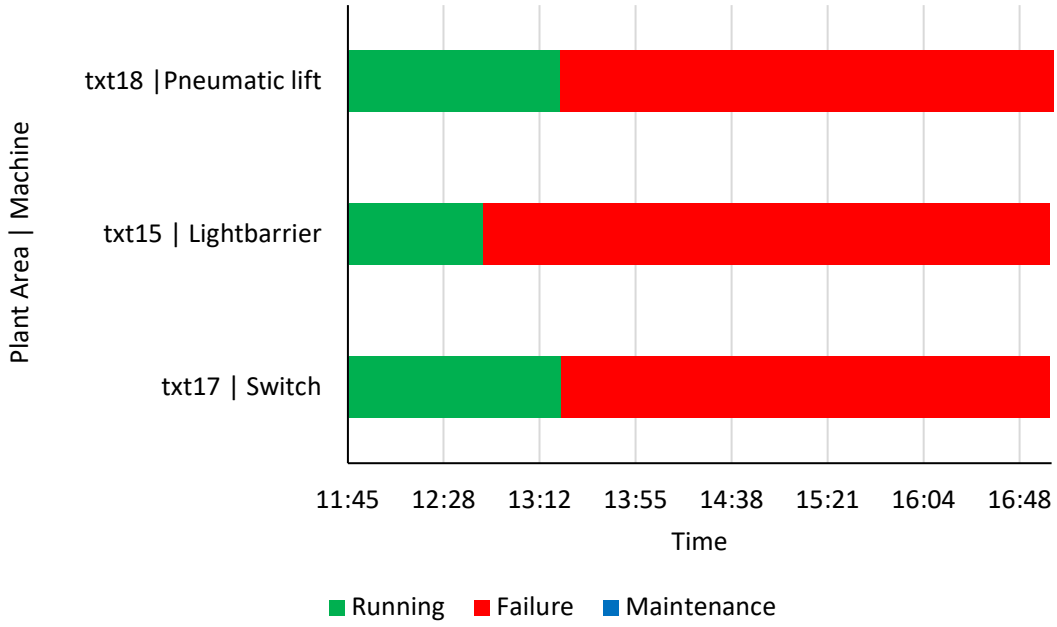


Figure 10.6: Case 3 Input Data

for a more flexible approach to maintenance and allows for downtime cost savings and a reduced maintenance time.

The Gantt chart in Figure 10.7 shows the results for Case 1 for the condition-based scheduling strategy. Using the preventative maintenance model predictions, and performing maintenance only at the point of, or after, a fault occurs, a total cost of £$8,196$ was obtained requiring three engineers during the period of consideration. A reduced cost of £$7,322$ however, is obtained using our proposed strategy with the same number of personnel (Figure 10.8). This is as preventative maintenance actions for faults detected take shorter times translating to reduced man-power costs, as well as a cost saving on parts replacement, overall reducing costs by 10.7%.

Similar sets of results are observed for Case 2 for the condition-based (Figure 10.9) and preventative maintenance scheduling (Figure 10.10) strategies. A cost reduction of up to 44% was obtained, which can be attributed not only to the difference in maintenance times for preventative and corrective maintenance actions, but also to the reduced number of engineering personnel required for the time period in question. As in Case 1, but also more obvious in the current case, our proposed maintenance strategy also leads to a shorter completion time despite the reduced

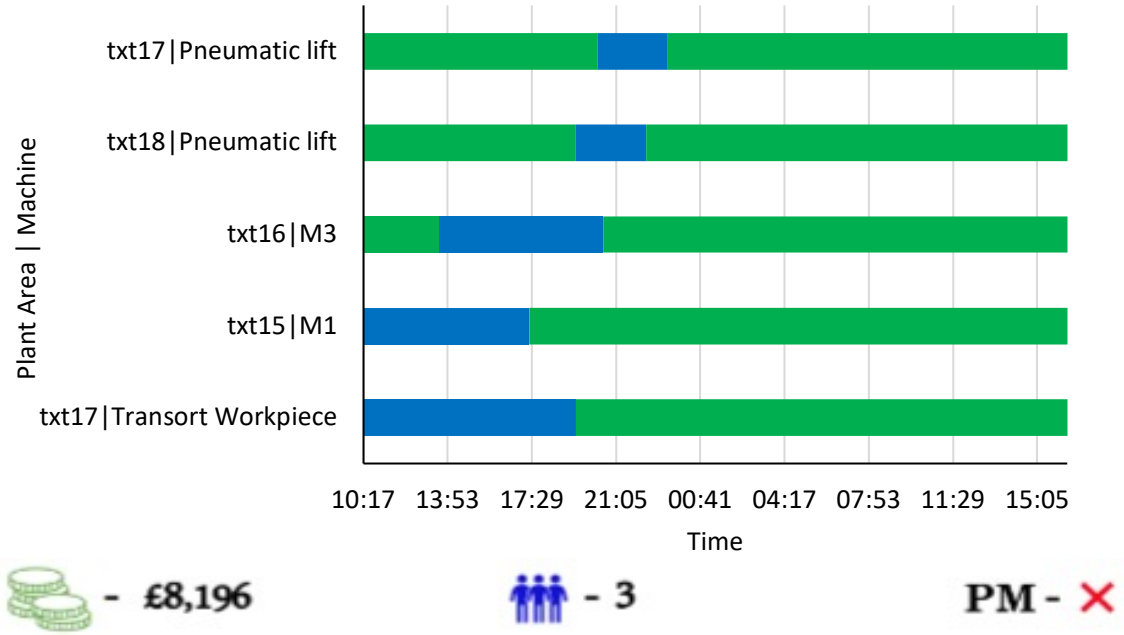


Figure 10.7: Case 1 Results - No Preventative Maintenance

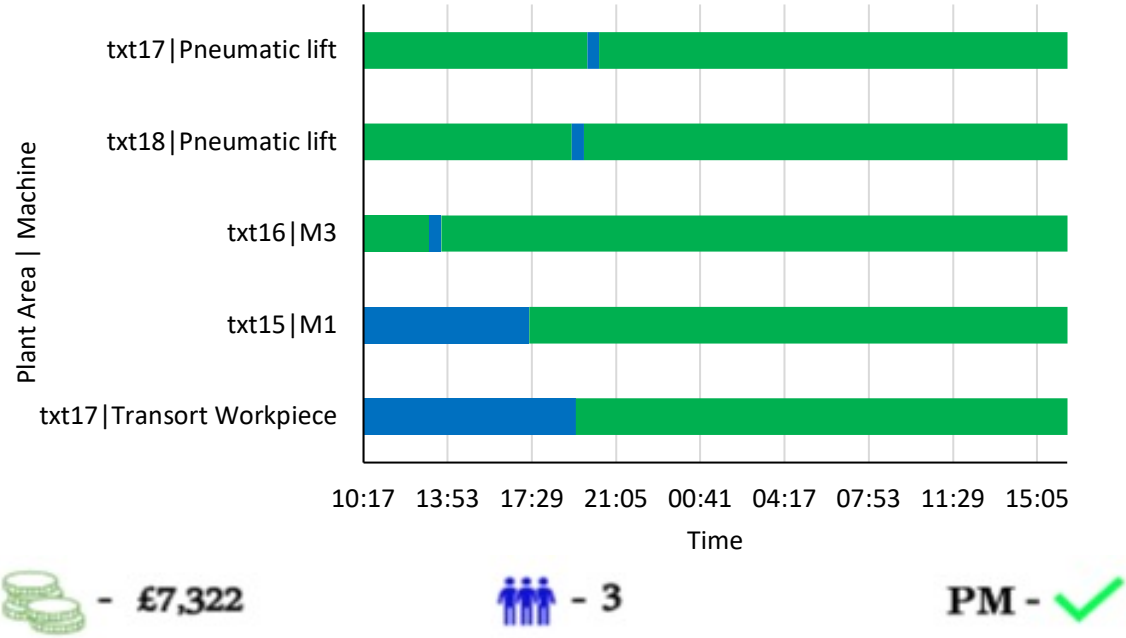


Figure 10.8: Case 1 Results - With Preventative Maintenance

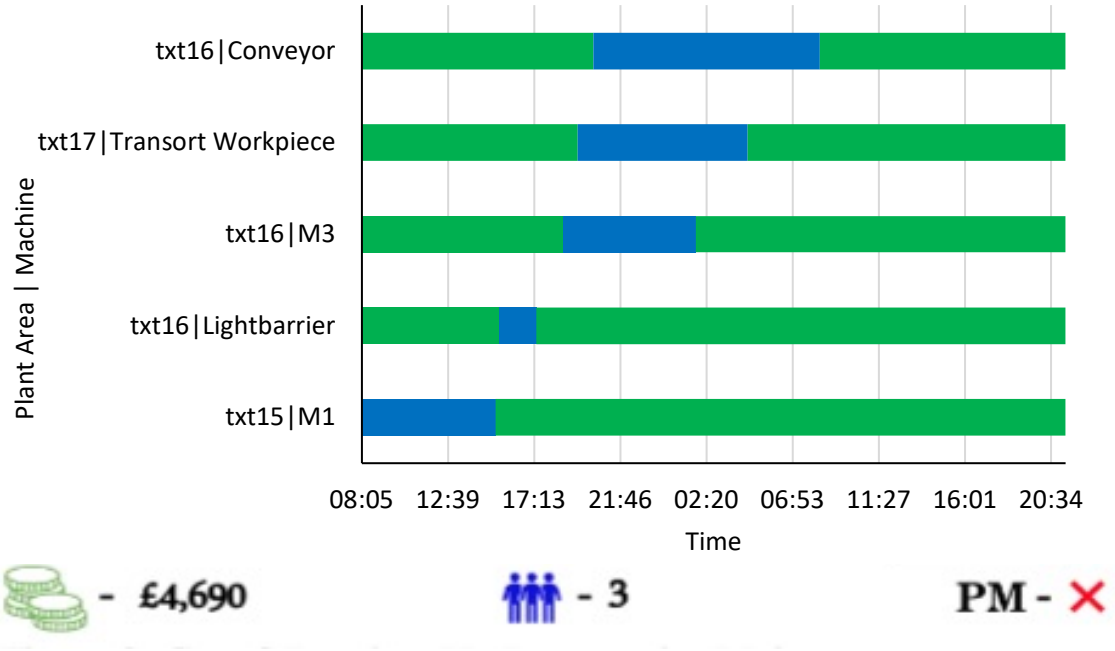


Figure 10.9: Case 2 Results - No Preventative Maintenance

number of personnel. This allows for additional production hours for the plant which can lead to higher revenues and/or flexibility of plant operation.

In Case 3 with a smaller number of machines subject to predicted faults, Figure 10.11 shows the cost optimal maintenance schedule when failures are corrected at the point of or after occurrence. A 25% reduction in cost is also observed with a reduced number of personnel when our preventative maintenance strategy is adopted instead (Figure 10.12). In each of these three cases, it becomes evident that obtaining optimal schedules for preventative maintenance tasks (in comparison to condition-based maintenance) leads not only to a general reduction in costs via reduced total maintenance times, but also to possible reduction in the number of personnel required.

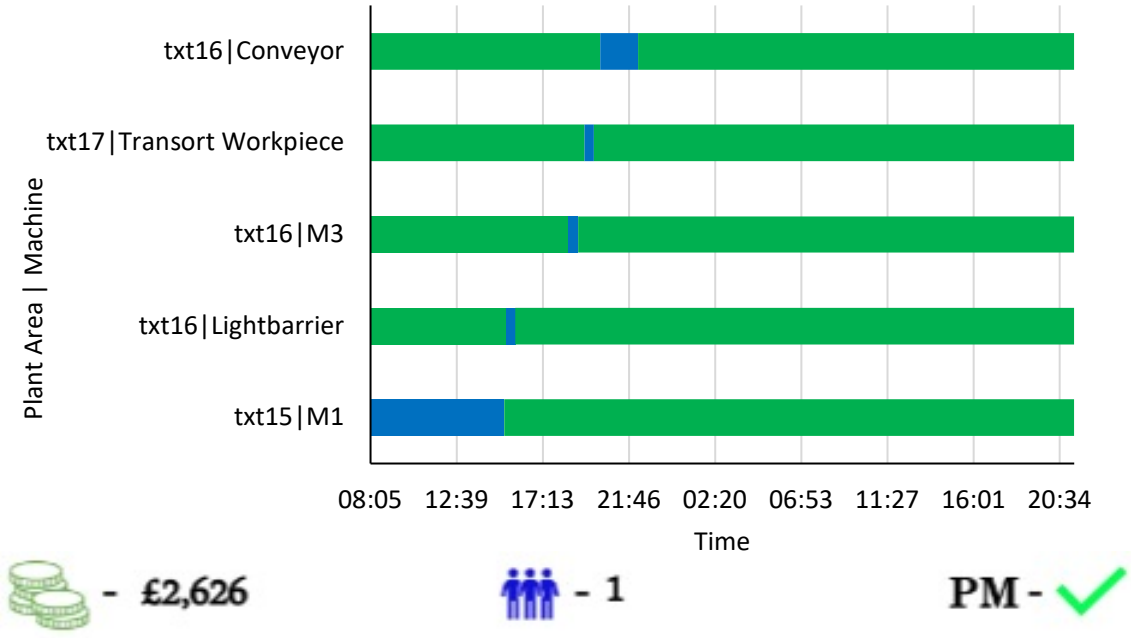


Figure 10.10: Case 2 Results - With Preventative Maintenance

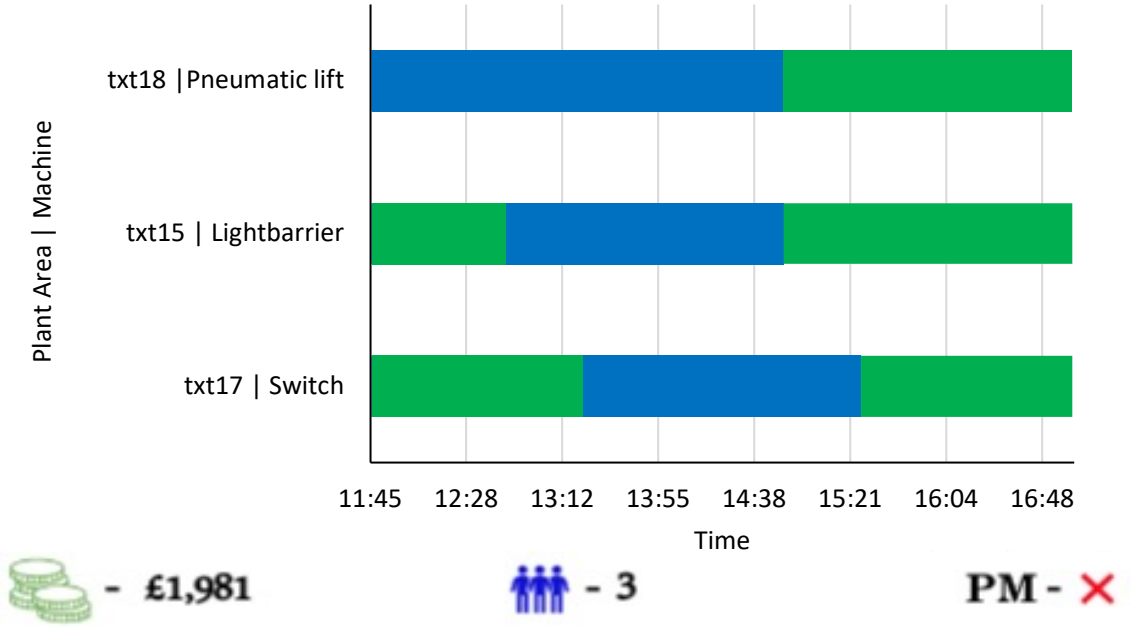


Figure 10.11: Case 3 Results - No Preventative Maintenance

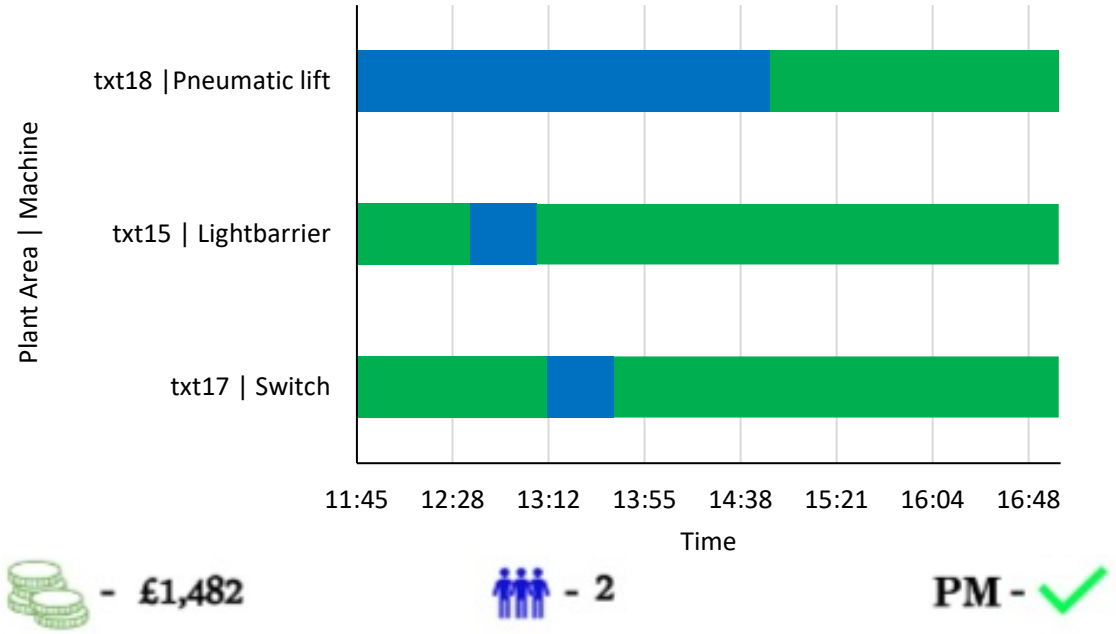


Figure 10.12: Case 3 Results - With Preventative Maintenance

10.6 Conclusion

A poor maintenance system can be catastrophic to an industrial plant's performance and safety. In this paper, we propose a novel methodology to be used in industry to create an optimum maintenance schedule. The methodology utilises the abundance of data made available due to Industry 4.0 by combining various machine learning methods to create an optimum maintenance schedule.

The proposed framework consists of three stages, predictive maintenance, maintenance time estimation and optimisation. In this research, the main objective of this work is to build such a methodology and investigate each stage as a proof of concept. Due to the lack of availability of real-data, we applied the proof of concept investigation to a Fischertechnik (FT) simulation model (Klein and Bergmann, 2019). Thus, each stage of the algorithm was investigated through batch training, whereas, in a real application, we would recommend this batch training approach before implementing an online version to ensure validation of the best machine learning model is chosen to fit the plant's data.

The algorithm describing the maintenance policy was robustly tested by investigating the

three steps of the framework. First, the popular method of predictive maintenance was analysed by comparing five readily available machine learning methods to ensure faults occurring in the FT plant can be identified. In this work, the demonstration of methodology, training and validation of machine learning methods, is vital to understanding how the maintenance policy begins its successful implementation. The predictive maintenance results showed the Quadratic Discriminant Analysis model to be superior of the five methods as it has the largest value for the precision, the recall and the F1 score. Therefore, the Quadratic Discriminant Analysis was chosen to identify faulty machines and share predictions to get the best from the maintenance policy.

The second stage of the algorithm requires accurate predictions to ensure the final proposed schedule can be followed without delays. Here, we addressed a gap in literature by investigation maintenance time estimation models. Once again, the method was demonstrated using five regression techniques that use historical data from the FT plant to map the sensor readings to the time it takes to fix a fault before failure occurs. Results found the Gaussian Process (Yeardley et al., 2020a, 2021) to be the best performing machine learning method, often predicting values close to the observed time. However, a significant anomaly did occur highlighting the importance in batch learning and cross-validation before implementing an online policy.

At the final stage, the FT model data was split into three cases based on calendar dates of fault occurrence. To demonstrate the benefits of the proposed maintenance policy, the maintenance schedules of the three cases were optimised using condition-based scheduling and our proposed strategy. The results of the optimisation provided high quality maintenance schedules and evidence that preventative maintenance obtains schedules with a lower cost, personnel requirement and/or overall maintenance times when compared to condition-based maintenance.

The workflow presented in this work could readily be applied to reduce maintenance costs; however, first, to assess its efficacy at a wider scale it would be beneficial to apply this data to other case studies based on real industrial plant data. Further, the workflow could be implemented in an online environment that allows data to flow automatically to update maintenance schedules.

10.7 Nomenclature

The abbreviations and symbols used are defined as follows:

Abbreviations/Acronyms

CM	Corrective Maintenance
PM	Preventative Maintenance
MINLP	Mixed Integer Non-linear Programming

Indices

i	devices/machines
t	time/period
s	device/machine state
f	fault description

Set

I	devices/machines
I^f	ordered pairs of device and possible faults
T	set of time/periods
S	set of possible device/machine states

Parameters

μ_{if}	0,1 parameter denoting if fault f on device i causes the plant to shutdown
ρ^p	0,1 parameter denoting if preventative maintenance is allowed
υ	number of times period before fault occurrence maintenance is allowed to start
$\hat{C}^p$	personnel/engineer cost per hour

$\hat{C}^e_{if}$	cost of device parts for replacement for fault f in device i
$\hat{C}^d$	plant downtime cost per hour
$\hat{C}^f$	device downtime cost per hour
F_{it}	0,1 values denoting time device i fails
F^f_{ift}	0,1 values denoting time fault f occurs on device i
$\hat{M}$	a 'big' number
M_i	time taken for maintenance on item i
M^f_{if}	time taken for maintenance on item i for fault f after failure occurs
M^p_{if}	time taken for preventative maintenance on item i for fault f on fault detection

Binary variables

γ_{if}	1 if preventative maintenance occurs on device i for fault f; 0 otherwise
S_{ist}	1 if device i is in state s at time t; 0 otherwise
S^f_t	1 if any device is in a failed state at time t and causes the plant to shutdown; 0 otherwise
$\overline{S}^f_t$	1 if any device is in a failed state at time t; 0 otherwise
S^f_{ift}	1 if device i is currently in a failed state for fault f at time t; 0 otherwise
S^m_{ift}	1 if device i is currently under maintenance for fault f at time t; 0 otherwise
W^e_{it}	1 if maintenance is completed on device i at time t; 0 otherwise
$W^{e'}_{ift}$	1 if maintenance is completed on device i for fault f at time t; 0 otherwise
W^s_{it}	1 if maintenance starts on device i at time t; 0 otherwise
$W^{s'}_{ift}$	1 if maintenance starts on device i for fault f at time t; 0 otherwise

Integer variables

N^p	number of technicians available for maintenance

Continuous variables

κ_{if}^p	difference in time periods between the time of fault occurrence and the start of maintenance for preventative maintenance actions
κ_{if}^f	difference in time periods between the time of fault occurrence and the start of maintenance for corrective maintenance actions

Bibliography

Abada, S., Marlair, G., Lecocq, A., Petit, M., Sauvant-Moynot, V., Huet, F., 2016. Safety focused modeling of lithium-ion batteries: A review. Journal of Power Sources 306, 178–192. doi:10.1016/j.jpowsour.2015.11.100.

Aghaji, M.Z., Fernandez, M., Boyd, P.G., Daff, T.D., Woo, T.K., 2016. Quantitative Structure–Property Relationship Models for Recognizing Metal Organic Frameworks (MOFs) with High CO2Working Capacity and CO2/CH4Selectivity for Methane Purification. European Journal of Inorganic Chemistry 2016, 4505–4511. doi:10.1002/ejic.201600365.

Al, R., Behera, C.R., Zubov, A., Gernaey, K.V., Sin, G., 2019. Meta-modeling based efficient global sensitivity analysis for wastewater treatment plants – An application to the BSM2 model. Computers & Chemical Engineering 127, 233–246. URL: https://doi.org/10.1016/j.compchemeng.2019.05.015, doi:10.1016/j.compchemeng.2019.05.015.

Al-Taweel, Y., 2018. Diagnostics and Simulation-Based Methods for Validating Gaussian Process Emulators. Phd thesis. University of Sheffield.

Alkan, B., Vera, D., Ahmad, M., Ahmad, B., Harrison, R., 2016. A Model for Complexity Assessment in Manual Assembly Operations Through Predetermined Motion Time Systems. Procedia CIRP 44, 429–434. URL: http://dx.doi.org/10.1016/j.procir.2016.02.111, doi:10.1016/j.procir.2016.02.111.

Amaran, S., Sahinidis, N., Sharda, B., Morrison, M., Bury, S., Miller, S., Wassick, J., 2015.

Long-term turnaround planning for integrated chemical sites. Computers and Chemical Engineering 72, 145–158. doi:10.1016/j.compchemeng.2014.08.003.

Ameye, D., Keleb, E., Vervaet, C., Remon, J.P., Adams, E., Massart, D.L., 2002. Scaling-up of a lactose wet granulation process in Mi-Pro high shear mixers. European Journal of Pharmaceutical Sciences 17, 247–251. doi:10.1016/S0928-0987(02)00218-X.

Andrieu, C., Freitas, N.D., Doucet, A., Jordan, M.I., 2003. An Introduction to MCMC for Machine Learning. Machine Learning 50, 5–43. URL: papers2://publication/uuid/A18A8ED8-98C6-4530-B94E-1C05CF3DE63C, doi:10.1023/A:1020281327116, arXiv:1109.4435v1.

Ba, S., Joseph, V.R., 2012. Composite Gaussian process models for emulating expensive functions. Annals of Applied Statistics 6, 1838–1860. doi:10.1214/12-AOAS570.

Banerjee, D., Simon, C.M., Plonka, A.M., Motkuri, R.K., Liu, J., Chen, X., Smit, B., Parise, J.B., Haranczyk, M., Thallapally, P.K., 2016. Metal-organic framework with optimally selective xenon adsorption and separation. Nature Communications 7, 1–7. doi:10.1038/ncomms11831.

Barta, G., Nagy, G.B.G., Kazi, S., Henk, T., 2015. GEFCOM 2014—probabilistic electricity price forecasting. Smart Innovation, Systems and Technologies 39, 67–76. doi:10.1007/978-3-319-19857-6_7, arXiv:1506.06972.

Bellinghausen, S., 2020. Modelling and scaling rules for high-shear wet granulation of pharmaceuticals. Ph.D. thesis. The University of Sheffield.

Bellinghausen, S., Gavi, E., Jerke, L., Barrasso, D., Salman, A.D., Litster, J.D., 2022. Model-driven design using population balance modelling for high-shear wet granulation. Powder Technology 396, 578–595. URL: https://doi.org/10.1016/j.powtec.2021.10.028, doi:10.1016/j.powtec.2021.10.028.

Bellinghausen, S., Gavi, E., Jerke, L., Ghosh, P.K., Salman, A.D., Litster, J.D., 2019. Nuclei size distribution modelling in wet granulation. Chemical Engineering Science: X 4, 100038.

Bello, A., Bunn, D.W., Reneses, J., Munoz, A., 2017. Medium-Term Probabilistic Forecasting of Electricity Prices: A Hybrid Approach. IEEE Transactions on Power Systems 32, 334–343. doi:10.1109/TPWRS.2016.2552983.

Bello, A., Reneses, J., Muñoz, A., Delgadillo, A., 2016. Probabilistic forecasting of hourly electricity prices in the medium-term using spatial interpolation techniques. International Journal of Forecasting 32, 966–980. URL: http://dx.doi.org/10.1016/j.ijforecast.2015.06.002, doi:10.1016/j.ijforecast.2015.06.002.

Berlinet, A., 2004. Reproducing kernel Hilbert spaces in probability and statistics. Kluwer Academic, Boston, [Mass.] ; London.

Bhosekar, A., Ierapetritou, M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. Computers and Chemical Engineering 108, 250–267. URL: https://doi.org/10.1016/j.compchemeng.2017.09.017, doi:10.1016/j.compchemeng.2017.09.017.

Biggins, F.A.V., Ejeh, J.O., Roberts, R., Yeardley, A.S., Brown, S.F., 2022. Optimising a wind farm with energy storage considering curtailment and uncertainties, in: 32nd European Symposium on Computer Aided Process Engineering, Elsevier B.V.

BMReports, . URL: https://www.bmreports.com/bmrs/?q=help/about-us.

Bobbitt, N.S., Mendonca, M.L., Howarth, A.J., Islamoglu, T., Hupp, J.T., Farha, O.K., Snurr, R.Q., 2017. Metal-organic frameworks for the removal of toxic industrial chemicals and chemical warfare agents. Chemical Society Reviews 46, 3357–3385. doi:10.1039/c7cs00108h.

Bogojeski, M., Sauer, S., Horn, F., Müller, K.R., 2021. Forecasting industrial aging processes with machine learning methods. Computers and Chemical Engineering 144,

107123. URL: https://doi.org/10.1016/j.compchemeng.2020.107123, doi:10.1016/j.compchemeng.2020.107123, arXiv:2002.01768.

Boukouvala, F., Dubey, A., Vanarase, A., Ramachandran, R., Muzzio, F.J., Ierapetritou, M., 2012. Computational approaches for studying the granular dynamics of continuous blending processes, 2–population balance and data-based methods. Macromolecular Materials and Engineering 297, 9–19.

Boukouvala, F., Muzzio, F.J., Ierapetritou, M.G., 2010. Predictive Modeling of Pharmaceutical Processes with Missing and Noisy Data. AIChE Journal 56, 2860–2872. doi:https://doi.org/10.1002/aic.12203, arXiv:0201037v1.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. Wadsworth, Belmont, Calif.

Brown, S., Beck, J., Mahgerefteh, H., Fraga, E.S., 2013. Global sensitivity analysis of the impact of impurities on CO2 pipeline failure. Reliability Engineering and System Safety 115, 43–54. URL: http://dx.doi.org/10.1016/j.ress.2013.02.006, doi:10.1016/j.ress.2013.02.006.

Brunier, B., Sheibat-Othman, N., Othman, S., Chevalier, Y., Bourgeat-Lami, E., 2017. Modelling particle growth under saturated and starved conditions in emulsion polymerization. The Canadian Journal of Chemical Engineering 95, 208–221.

Bugryniec, P.J., Davidson, J.N., Brown, S.F., 2018. Assessment of thermal runaway in commercial lithium iron phosphate cells due to overheating in an oven test. Energy Procedia 151, 74–78. URL: https://doi.org/10.1016/j.egypro.2018.09.030, doi:10.1016/j.egypro.2018.09.030.

Bugryniec, P.J., Davidson, J.N., Cumming, D.J., Brown, S.F., 2019. Pursuing safer batteries: Thermal abuse of LiFePO4 cells. Journal of Power Sources 414, 557–568. doi:10.1016/j.jpowsour.2019.01.013.

Bugryniec, P.J., Yeardley, A.S., Jain, A., Price, N., Vernuccio, S., Brown, S.F., 2022. Gaussian-Process based inference of electrolyte decomposition reaction networks in Li-ion battery failure, in: 32nd European Symposium on Computer Aided Process Engineering, Elsevier B.V.

Caniou, Y., Sudret, B., 2010. Distribution-based global sensitivity analysis using polynomial chaos expansions. Procedia - Social and Behavioral Sciences 2, 7625–7626. URL: http://dx.doi.org/10.1016/j.sbspro.2010.05.149, doi:10.1016/j.sbspro.2010.05.149.

Capece, M., Bilgili, E., Dave, R., 2011. Identification of the breakage rate and distribution parameters in a non-linear population balance model for batch milling. Powder Technology 208, 195–204.

Carraro, M., Gross, S., 2014. Hybrid materials based on the embedding of organically modified transition metal oxoclusters or polyoxometalates into polymers for functional applications: A review. Materials 7, 3956–3989. doi:10.3390/ma7053956.

Carvalho, T.P., Soares, F.A., Vita, R., Francisco, R.d.P., Basto, J.P., Alcalá, S.G., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Computers and Industrial Engineering 137, 106024. URL: https://doi.org/10.1016/j.cie.2019.106024, doi:10.1016/j.cie.2019.106024.

Chakravarty, S., Mohapatra, P., Dash, P.K., 2016. Evolutionary extreme learning machine for energy price forecasting. International Journal of Knowledge-Based and Intelligent Engineering Systems 20, 75–96. doi:10.3233/KES-160331.

Chastaing, G., Gamboa, F., Prieur, C., 2015. Generalized Sobol sensitivity indices for dependent variables: numerical methods. Journal of Statistical Computation and Simulation 85, 1306–1333. URL: https://doi.org/10.1080/00949655.2014.960415, doi:10.1080/00949655.2014.960415.

Chastaing, G., Le Gratiet, L., 2015. ANOVA decomposition of conditional Gaussian processes for sensitivity analysis with dependent inputs. Journal of Statistical Com-

putation and Simulation 85, 2164–2186. doi:10.1080/00949655.2014.925111, arXiv:arXiv:1310.3578v1.

Chaudhury, A., Barrasso, D., Pandey, P., Wu, H., Ramachandran, R., 2014. Population balance model development, validation, and prediction of CQAs of a high-shear wet granulation process: towards QbD in drug product pharmaceutical manufacturing. Journal of Pharmaceutical Innovation 9, 53–64.

Chen, S., Wan, C., Wang, Y., 2005a. Thermal analysis of lithium-ion batteries. Journal of Power Sources 140, 111 – 124. URL: http://www.sciencedirect.com/science/article/pii/S0378775304008596, doi:https://doi.org/10.1016/j.jpowsour.2004.05.064.

Chen, S.C., Wang, Y.Y., Wan, C.C., 2006. Thermal Analysis of Spirally Wound Lithium Batteries. Journal of The Electrochemical Society 153, A637. doi:10.1149/1.2168051.

Chen, W., Jin, R., Sudjianto, A., 2005b. Analytical Variance-Based Global Sensitivity Analysis in Simulation-Based Design Under Uncertainty. Journal of Mechanical Design 127, 875. doi:10.1115/1.1904642.

Chiu, K.C., Lin, C.H., Yeh, S.F., Lin, Y.H., Chen, K.C., 2014. An electrochemical modeling of lithium-ion battery nail penetration. Journal of Power Sources 251, 254–263. URL: http://dx.doi.org/10.1016/j.jpowsour.2013.11.069, doi:10.1016/j.jpowsour.2013.11.069.

Chung, Y.G., Camp, J., Haranczyk, M., Sikora, B.J., Bury, W., Krungleviciute, V., Yildirim, T., Farha, O.K., Sholl, D.S., Snurr, R.Q., 2014. Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. Chemistry of Materials 26, 6185–6192. doi:10.1021/cm502594j.

Coleman, B., Ostanek, J., Heinzel, J., 2016. Reducing cell-to-cell spacing for large-format lithium ion battery modules with aluminum or PCM heat sinks under failure conditions.

Applied Energy 180, 14–26. URL: http://dx.doi.org/10.1016/j.apenergy.2016.07.094, doi:10.1016/j.apenergy.2016.07.094.

Coman, P.T., Darcy, E.C., Veje, C.T., White, R.E., 2017. Modelling Li-Ion Cell Thermal Runaway Triggered by an Internal Short Circuit Device Using an Efficiency Factor and Arrhenius Formulations. Journal of The Electrochemical Society 164, A587–A593. doi:10.1149/2.0341704jes.

COMSOL Multiphysics®V5.2a, . COMSOL Multiphysics Reference Manual. www.comsol.com. COMSOL AB, Stockholm, Sweden.

Constantine, P.G., Diaz, P., 2017. Global sensitivity metrics from active subspaces. Reliability Engineering and System Safety 162, 1–13. URL: http://dx.doi.org/10.1016/j.ress.2017.01.013, doi:10.1016/j.ress.2017.01.013, arXiv:1510.04361.

Constantine, P.G., Dow, E., Wang, Q., 2014. Active subspace methods in theory and practice: Applications to kriging surfaces. SIAM Journal on Scientific Computing 36, A1500–A1524. doi:10.1137/130916138, arXiv:1304.2070.

Costa, C.B.B., Maciel, M.R.W., Maciel Filho, R., 2007. Considerations on the crystallization modeling: Population balance solution. Computers & Chemical Engineering 31, 206–218.

Cressie, N., 1990. The origins of kriging. Mathematical Geology 22, 239–252. doi:10.1007/BF00889887.

Cryer, S.A., Scherer, P.N., 2003. Observations and Process Parameter Sensitivities in Fluid-Bed Granulation. AIChE Journal 49, 2802–2809. doi:10.1002/aic.690491113.

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. Journal of the American Statistical Association 86, 953–963. doi:10.1080/01621459.1991.10475138.

Currin, C., Mitchell, T., Morris M., Ylvisaker D., 1988. A Bayesian Approach to the Design and Analysis of Computer Experiments. ORNL Oak Ridge National Laboratory (US) doi:10.2172/814584.

Damianou, A.C., Lawrence, N.D., 2013. Deep Gaussian Processes. Journal of Machine Learning Research 31, 207–215. URL: http://proceedings.mlr.press/v31/damianou13a.pdf.

Davis, N.J., 2016. Mechanical dispersion of semi-solid binders in high-shear granulation. Ph.D. thesis. Purdue University.

DeCoste, J.B., Weston, M.H., Fuller, P.E., Tovar, T.M., Peterson, G.W., LeVan, M.D., Farha, O.K., 2014. Metal-organic frameworks for oxygen storage. Angewandte Chemie - International Edition 53, 14092–14095. doi:10.1002/anie.201408464.

Dedopoulos, I., Shah, N., 1995. Optimal short-term scheduling of maintenance and production for multipurpose plants. Industrial and Engineering Chemistry Research 34, 192–201. doi:10.1021/ie00040a019.

Dewulf, L., Chiacchia, M., Yeardley, A.S., Milton, R.A., Brown, S.F., Patwardhan, S.V., 2021. Designing bioinspired green nanosilicas using statistical and machine learning approaches. Molecular Systems Design and Engineering 6, 293–307. doi:10.1039/d0me00167h.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–263. doi:10.1080/07350015.1995.10524599.

Ding, Y., 2018. A novel decompose-ensemble methodology with AIC-ANN approach for crude oil forecasting. Energy 154, 328–336. URL: https://doi.org/10.1016/j.energy.2018.04.133, doi:10.1016/j.energy.2018.04.133.

Dong, T., Peng, P., Jiang, F., 2018. Numerical modeling and analysis of the thermal behavior of NCM lithium-ion batteries subjected to very high C-rate discharge/charge operations. International Journal of Heat and Mass Transfer 117, 261–272. URL: https:

//doi.org/10.1016/j.ijheatmasstransfer.2017.10.024, doi:10.1016/j.ijheatmasstransfer.2017.10.024.

Drake, S., Wetz, D., Ostanek, J., Miller, S., Heinzel, J., Jain, A., 2014. Measurement of anisotropic thermophysical properties of cylindrical li-ion cells. Journal of Power Sources 252, 298 – 304. URL: http://www.sciencedirect.com/science/article/pii/S0378775313019502, doi:https://doi.org/10.1016/j.jpowsour.2013.11.107.

Drews, T.O., Braatz, R.D., Alkire, R.C., 2003. Parameter Sensitivity Analysis of Monte Carlo Simulations of Copper Electrodeposition with Multiple Additives. Journal of The Electrochemical Society 150, C807. doi:10.1149/1.1617305.

Duan, X., Jiang, W., Zou, Y., Lei, W., Ma, Z., 2018. A coupled electrochemical–thermal–mechanical model for spiral-wound Li-ion batteries. Journal of Materials Science 53, 10987–11001. URL: https://doi.org/10.1007/s10853-018-2365-6, doi:10.1007/s10853-018-2365-6.

Duffuaa, S., Raouf, A., 2015. Planning and control of maintenance systems: Modelling and analysis. Springer International Publishing. doi:10.1007/978-3-319-19803-3.

Durrande, N., Ginsbourger, D., Roustant, O., Carraro, L., 2013. ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. Journal of Multivariate Analysis 115, 57–67. URL: http://dx.doi.org/10.1016/j.jmva.2012.08.016, doi:10.1016/j.jmva.2012.08.016.

Duvenaud, D., Lloyd, J.R., Grosse, R., Tenenbaum, J.B., Ghahramani, Z., 2013. Structure discovery in nonparametric regression through compositional kernel search. 30th International Conference on Machine Learning, ICML 2013 28, 2203–2211. arXiv:arXiv:1302.4922v4.

ESO, N.G., . URL: https://www.nationalgrideso.com/balancing-data/data-finder-and-explorer.

Fanourgakis, G.S., Gkagkas, K., Tylianakis, E., Klontzas, E., Froudakis, G., 2019. A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials. Journal of Physical Chemistry A 123, 6080–6087. doi:10.1021/acs.jpca.9b03290.

Feng, W., Feng, Y., Zhang, Q., 2021. Multistage distributionally robust optimization for integrated production and maintenance scheduling. AIChE Journal 67, 1–19. doi:10.1002/aic.17329.

Feng, X., Ouyang, M., Liu, X., Lu, L., Xia, Y., He, X., 2018. Thermal runaway mechanism of lithium ion battery for electric vehicles: A review. Energy Storage Materials 10, 246 – 267. URL: http://www.sciencedirect.com/science/article/pii/S2405829716303464, doi:https://doi.org/10.1016/j.ensm.2017.05.013.

Fernandez, M., Barnard, A.S., 2016. Geometrical Properties Can Predict CO2 and N2 Adsorption Performance of Metal-Organic Frameworks (MOFs) at Low Pressure. ACS Combinatorial Science 18, 243–252. doi:10.1021/acscombsci.5b00188.

Fernandez, M., Boyd, P.G., Daff, T.D., Aghaji, M.Z., Woo, T.K., 2014. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO2 capture. Journal of Physical Chemistry Letters 5, 3056–3060. doi:10.1021/jz501331m.

Fernandez, M., Trefiak, N.R., Woo, T.K., 2013a. Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity. Journal of Physical Chemistry C 117, 14095–14105. doi:10.1021/jp404287t.

Fernandez, M., Woo, T.K., Wilmer, C.E., Snurr, R.Q., 2013b. Large-scale quantitative structure-property relationship (QSPR) analysis of methane storage in metal-organic frameworks. Journal of Physical Chemistry C 117, 7681–7689. doi:10.1021/jp4006422.

Fraunholz, C., Kraft, E., Keles, D., Fichtner, W., 2021. Advanced price forecasting in agent-

based electricity market simulation. Applied Energy 290. doi:10.1016/j.apenergy.2021.116688.

Gaillard, P., Goude, Y., Nedellec, R., 2016. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. International Journal of Forecasting 32, 1038–1050. URL: http://dx.doi.org/10.1016/j.ijforecast.2015.12.001, doi:10.1016/j.ijforecast.2015.12.001.

Gao, X., Dowling, A.W., 2020. Making Money in Energy Markets: Probabilistic Forecasting and Stochastic Programming Paradigms. Proceedings of the American Control Conference 2020-July, 168–173. doi:10.23919/ACC45564.2020.9147380.

Garud, S.S., Karimi, I.A., Kraft, M., 2017. Design of computer experiments: A review. Computers and Chemical Engineering 106, 71–95. doi:10.1016/j.compchemeng.2017.05.010.

George Seif, 2018. The 5 Clustering Algorithms Data Scientists Need to Know. URL: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

Ghadbeigi, L., Harada, J.K., Lettiere, B.R., Sparks, T.D., 2015. Performance and resource considerations of li-ion battery electrode materials. Energy Environ. Sci. 8, 1640–1650. URL: http://dx.doi.org/10.1039/C5EE00685F, doi:10.1039/C5EE00685F.

Gilchrist, A., 2016. Industry 4.0. The Industrial Internet of Things. Apress. doi:10.1007/978-1-4842-2047-4, arXiv:arXiv:1011.1669v3.

Glassdoor Inc., 2021. Chemical plant operator salaries. https://www.glassdoor.co.uk/Salaries/chemical-plant-operator-salary-SRCH_KO0,23.htm.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378. doi:10.1198/016214506000001437.

Görür, O.C., Yu, X., Sivrikaya, F., 2021. Integrating predictive maintenance in adaptive process scheduling for a safe and efficient industrial process. Applied Sciences (Switzerland) 11. doi:10.3390/app11115042.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F.A., Hansen, L.K., 1999. On clustering fMRI time series. NeuroImage 9, 298–310. doi:10.1006/nimg.1998.0391.

Gratiet, L.L., Marelli, S., Sudret, B., 2016. Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes. Handbook of Uncertainty Quantification - Part III: Sensitivity analysis doi:10.1007/978-3-319-11259-6_38-1.

Gunes, V., Peter, S., Givargis, T., Vahid, F., 2014. A survey on concepts, applications, and challenges in cyber-physical systems. KSII Transactions on Internet and Information Systems 8, 4242–4268. doi:10.3837/tiis.2014.12.001.

Guo, G., Long, B., Cheng, B., Zhou, S., Xu, P., Cao, B., 2010. Three-dimensional thermal finite element modelling of lithium-ion battery in thermal abuse application. Journal of Power Sources 195, 2393–2398. doi:10.1016/j.jpowsour.2009.10.090.

Hapgood, K.P., Litster, J.D., Smith, R., 2003. Nucleation regime map for liquid bound granules. AIChE Journal 49, 350–361.

Hapgood, K.P., Tan, M.X.L., Chow, D.W.Y., 2009. A method to predict nuclei size distributions for use in models of wet granulation. Advanced Powder Technology 20, 293–297.

Hastie, T., 2009. The elements of statistical learning : data mining, inference, and prediction. Springer series in statistics. 2nd ed. ed., Springer, New York.

Hatchard, T.D., MacNeil, D.D., Basu, A., Dahn, J.R., 2001. Thermal Model of Cylindrical and Prismatic Lithium-Ion Cells. Journal of The Electrochemical Society 148, A755–A761. doi:10.1149/1.1377592.

Hatchard, T.D., Macneil, D.D., Stevens, D.A., Christensen, L., Dahn, J.R., 2000. Importance of

Heat Transfer by Radiation in Li-Ion Batteries during Thermal Abuse. Electrochemical and Solid-Stae Letters 3, 305–308.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and forecasting 15, 559–570.

Hommel, G., 1988. A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika 75, 383–386. doi:10.1093/biomet/75.2.383.

Hoover, J., 2006. Measuring forecast accuracy: Omissions in today's forecasting engines and demand-planning software. Foresight: The International Journal of Applied Forecasting , 32–35URL: https://EconPapers.repec.org/RePEc:for:ijafaa:y:2006:i:4:p:32-35.

Hounslow, M.J., Pearson, J.M.K., Instone, T., 2001. Tracer studies of high-shear granulation: II. population balance modeling. AIChE Journal 47, 1984–1999.

Hu, B., Ma, Z., Lei, W., Zou, Y., Lu, C., 2017. A chemo-mechanical model coupled with thermal effect on the hollow core–shell electrodes in lithium-ion batteries. Theoretical and Applied Mechanics Letters 7, 199–206. URL: http://dx.doi.org/10.1016/j.taml.2017.09.001, doi:10.1016/j.taml.2017.09.001.

Hubicka, K., Marcjasz, G., Weron, R., 2019. A Note on Averaging Day-Ahead Electricity Price Forecasts Across Calibration Windows. IEEE Transactions on Sustainable Energy 10, 321–323. doi:10.1109/TSTE.2018.2869557.

International Energy Agency, 2016. Global EV Outlook 2016: Beyond one million electric cars. Technical Report. IEA. Paris. URL: Availableat:https://www.iea.org/publications/freepublications/publication/Global_EV_Outlook_2016.pdf. [accessed 2 August 2016].

Iooss, B., Lemaître, P., 2015. A review on global sensitivity analysis methods. Op-

erations Research/ Computer Science Interfaces Series 59, 101–122. doi:10.1007/978-1-4899-7547-8_5, arXiv:1404.2405.

Iveson, S.M., Litster, J.D., 1998. Fundamental studies of granule consolidation Part 2: Quantifying the effects of particle and binder properties. Powder Technology 99, 243–250.

Iveson, S.M., Wauters, P.A.L., Forrest, S., Litster, J.D., Meesters, G.M.H., Scarlett, B., 2001. Growth regime map for liquid-bound granules: further development and experimental validation. Powder Technology 117, 83–97.

Jain, V., Grossmann, I., 1998. Cyclic scheduling of continuous parallel-process units with decaying performance. AIChE Journal 44, 1623–1636. doi:10.1002/aic.690440714.

Jardine, A.K., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing 20, 1483–1510. doi:10.1016/j.ymssp.2005.09.012.

Jia, Z., Davis, E., Muzzio, F.J., Ierapetritou, M.G., 2009. Predictive modeling for pharmaceutical processes using kriging and response surface. Journal of Pharmaceutical Innovation 4, 174–186. doi:10.1007/s12247-009-9070-6.

Jiang, J., Dahn, J.R., 2004. ARC studies of the thermal stability of three different cathode materials: $LiCoO_2$; $Li[Ni_{0.1}Co_{0.8}Mn_{0.1}]O_2$; and $LiFePO_4$, in $LiPF_6$ and LiBoB EC/DEC electrolytes. Electrochemistry Communications 6, 39–43. doi:10.1016/j.elecom.2003.10.011.

Jin, R., Chen, W., Simpson, T.W., 2001. Comparative studies of metamodelling techniques under multiple modelling criteria. Structural and Multidisciplinary Optimization 23, 1–13. doi:10.1007/s00158-001-0160-4.

Jin, R., Chen, W., Sudjianto, A., 2004. Analytical metamodel-based global sensitivity analysis and uncertainty propagation for robust design. SAE Transactions Journal of Materials & Manufacturing doi:10.4271/2004-01-0429.

Jin, R., Du, X., Chen, W., 2003. The use of metamodeling techniques for optimization under uncertainty. Structural and Multidisciplinary Optimization 25, 99–116. doi:10.1007/s00158-002-0277-0.

Journel, A.G., 1977. Kriging in terms of projections. Mathematical Geology 9, 563–586. doi:10.1007/BF02067214.

Kajero, O.T., Chen, T., Yao, Y., Chuang, Y.C., Wong, D.S.H., 2017. Meta-modelling in chemical process system engineering. Journal of the Taiwan Institute of Chemical Engineers 73, 135–145. doi:10.1016/j.jtice.2016.10.042.

Kastner, C.A., Brownbridge, G.P.E., Mosbach, S., Kraft, M., 2013. Impact of powder characteristics on a particle granulation model. Chemical Engineering Science 97, 282–295.

Keles, D., Genoese, M., Möst, D., Ortlieb, S., Fichtner, W., 2013. A combined modeling approach for wind power feed-in and electricity spot prices. Energy Policy 59, 213–225. URL: http://dx.doi.org/10.1016/j.enpol.2013.03.028, doi:10.1016/j.enpol.2013.03.028.

Khadem, H., Eissa, M.R., Nemat, H., Alrezj, O., Benaissa, M., 2020. Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy. Talanta 211, 120740. URL: https://doi.org/10.1016/j.talanta.2020.120740, doi:10.1016/j.talanta.2020.120740.

Khalid, W., Albrechtsen, S.H., Sigsgaard, K.V., Mortensen, N.H., Hansen, K.B., Soleymani, I., 2020. Predicting maintenance work hours in maintenance planning. Journal of Quality in Maintenance Engineering 27, 366–384. doi:10.1108/JQME-06-2019-0058.

Kim, G.H., Pesaran, A., Spotnitz, R., 2007. A three-dimensional thermal abuse model for lithium-ion cells. Journal of Power Sources 170, 476–489. doi:10.1016/j.jpowsour.2007.04.018.

Kim, T.H., Park, J.S., Chang, S.K., Choi, S., Ryu, J.H., Song, H.K., 2012. The Current Move

of Lithium Ion Batteries Towards the Next Phase. Advanced Energy Materials 2, 860–872. doi:10.1002/aenm.201200028.

Kitanidis, P.K., 1986. Parameter Uncertainty in Estimation of Spatial Functions - Bayesian-Analysis. Water Resources Research 22, 499–507.

Klein, P., Bergmann, R., 2019. Generation of complex data for AI-based predictive maintenance research with a physical factory model. ICINCO 2019 - Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics 1, 40–50. doi:10.5220/0007830700400050.

Klein, P., Malburg, L., Bergmann, R., 2019. FTOnto: A domain ontology for a fischertechnik simulation production factory by reusing existing ontologies. CEUR Workshop Proceedings 2454.

Kobbacy, K.A.H., Murthy, D.N.P., 2008. Complex system maintenance handbook. Springer Science and Business Media, Berlin Heidelberg, Germany.

Kolmogorov, A.N., 1941. Interpolation und Extrapolation von stationaren zufaligen Folgen. Izv. Akad. Nauk SSSR Ser. Mat. 5, 3–14.

Kongburan, W., Chignell, M., Charoenkitkarn, N., Chan, J.H., 2019. Enhancing Predictive Power of Cluster-Boosted Regression with Text-Based Indexing. IEEE Access 7, 43394–43405. doi:10.1109/ACCESS.2019.2908032.

Kontoravdi, C., Mantalaris, A., Asprey, S., Pistikopoulos, E., 2005. Application of the sobol global sensitivity analysis method to a dynamic model of mab-producing mammalian cell cultures. Proceedings of the IASTED International Conference on Modelling, Identification, and Control, MIC , 361–366.

Kou, P., Liang, D., Gao, L., Lou, J., 2015. Probabilistic electricity price forecasting with variational heteroscedastic Gaussian process and active learning. Energy Conversion and Man-

agement 89, 298–308. URL: http://dx.doi.org/10.1016/j.enconman.2014.10.003, doi:10.1016/j.enconman.2014.10.003.

Krige, D.G., 1952. A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa , 201–215doi:10.2307/3006914.

Kucherenko, S., Rodriguez-Fernandez, M., Pantelides, C., Shah, N., 2009. Monte Carlo evaluation of derivative-based global sensitivity measures. Reliability Engineering and System Safety 94, 1135–1148. doi:10.1016/j.ress.2008.05.006.

Kumar, A., Shankar, R., Thakur, L., 2018. A big data driven sustainable manufacturing framework for condition-based maintenance prediction. Journal of Computational Science 27, 428–439. doi:10.1016/j.jocs.2017.06.006.

Kumar Akkisetty, P., Lee, U., Reklaitis, G.V., Venkatasubramanian, V., 2010. Population balance model-based hybrid neural network for a pharmaceutical milling process. Journal of Pharmaceutical Innovation 5, 161–168.

Lamoureux, B., Mechbal, N., Massé, J.R., 2014. A combined sensitivity analysis and kriging surrogate modeling for early validation of health indicators. Reliability Engineering and System Safety 130, 12–26. URL: http://dx.doi.org/10.1016/j.ress.2014.03.007, doi:10.1016/j.ress.2014.03.007.

Larsson, F., Bertilsson, S., Furlani, M., Albinsson, I., Mellander, B.E., 2018. Gas explosions and thermal runaways during external heating abuse of commercial lithium-ion graphite-LiCoO2 cells at different levels of ageing. Journal of Power Sources 373, 220 – 231. URL: http://www.sciencedirect.com/science/article/pii/S0378775317314398, doi:https://doi.org/10.1016/j.jpowsour.2017.10.085.

Lawrence, N., 2005. Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. Journal of Machine Learning Research 6, 1783–1816.

Lepikhin, A., Moskvichev, V., Machutov, N., 2018. Probabilistic modelling in solving analytical problems of system engineering, in: Kostogryzov, A. (Ed.), Probabilistic Modeling in System Engineering. IntechOpen, Rijeka. chapter 1. URL: https://doi.org/10.5772/intechopen.75686, doi:10.5772/intechopen.75686.

Li, G., Rabitz, H., 2012. General formulation of HDMR component functions with independent and correlated variables. Journal of Mathematical Chemistry 50, 99–130. doi:10.1007/s10910-011-9898-0.

Li, G., Rabitz, H., Yelvington, P.E., Oluwole, O.O., Bacon, F., Kolb, C.E., Schoendorf, J., 2010. Global sensitivity analysis for systems with independent and/or correlated inputs. The journal of physical chemistry. A 114, 6022–6032. doi:10.1021/jp9096919.

Li, J.R., Kuppler, R.J., Zhou, H.C., 2009. Selective gas adsorption and separation in metal-organic frameworks. Chemical Society Reviews 38, 1477–1504. doi:10.1039/b802426j.

Li, S., Yang, B., Qi, F., 2016. Accelerate global sensitivity analysis using artificial neural network algorithm: Case studies for combustion kinetic model. Combustion and Flame 168, 53–64. doi:10.1016/j.combustflame.2016.03.028.

Lin, N., Xie, X., Schenkendorf, R., Krewer, U., 2018. Efficient Global Sensitivity Analysis of 3D Multiphysics Model for Li-Ion Batteries. Journal of Electrochemical Society 165, A1169–A1183. doi:10.1149/2.1301805jes.

Lingitz, L., Gallina, V., Ansari, F., Gyulai, D., Pfeiffer, A., Sihn, W., 2018. Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer. Procedia CIRP 72, 1051–1056. doi:10.1016/j.procir.2018.03.148.

Liu, B., Jiang, Z.h., 2005. The man-hour estimation models and its comparison of interim products assembly for shipbuilding. International Journal of Operations Research 2, 14–19.

Liu, X., Guillas, S., 2017. Dimension Reduction for Gaussian Process Emulation: An Applica-

tion to the Influence of Bathymetry on Tsunami Heights. SIAM-ASA Journal on Uncertainty Quantification 5, 787–812. doi:10.1137/16M1090648.

Liu, X., Wu, Z., Stoliarov, S.I., Denlinger, M., Masias, A., Snyder, K., 2018. A Thermo-Kinetic Model of Thermally-Induced Failure of a Lithium Ion Battery: Development, Validation and Application. Journal of Electroanalytical Society 165, A2909–A2918. doi:10.1149/2.0111813jes.

Lopez, C.F., Jeevarajan, J.A., Mukherjee, P.P., 2015. Characterization of Lithium-Ion Battery Thermal Abuse Behavior Using Experimental and Computational Analysis. Journal of The Electrochemical Society 162, A2163–A2173. doi:10.1149/2.0751510jes.

Ma, Z., Wu, H., Wang, Y., Pan, Y., Lu, C., 2017. An electrochemical-irradiated plasticity model for metallic electrodes in lithium-ion batteries. International Journal of Plasticity doi:10.1016/j.ijplas.2016.10.009.

Macek, K., Endel, P., Cauchi, N., Abate, A., 2017. Long-term predictive maintenance: A study of optimal cleaning of biomass boilers. Energy and Buildings 150, 111–117. doi:10.1016/j.enbuild.2017.05.055.

MacNeil, D.D., Christensen, L., Landucci, J., Paulsen, J.M., Dahn, J.R., 2000. An Autocatalytic Mechanism for the Reaction of Li_xCoO_2 in Electrolyte at Elevated Temperature. Journal of The Electrochemical Society 147, 970–979. doi:10.1149/1.1393299.

Maiy, C., Sudret, B., 2017. Surrogate models for oscillatory systems using sparse polynomial chaos expansions and stochastic time warping. SIAM-ASA Journal on Uncertainty Quantification 5, 540–571. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85052092906&doi=10.1137%2f16M1083621&partnerID=40&md5=24d5d301482149149d4177a911f0eff2, doi:10.1137/16M1083621. cited By 25.

Mandal, P., Srivastava, A.K., Negnevitsky, M., Park, J.W., 2009. Sensitivity analysis of neural

network parameters to improve the performance of electricity price forecasting. International journal of energy research 33, 38–51. doi:10.1002/er.1469.

Mao, B., Huang, P., Chen, H., Wang, Q., Sun, J., 2020. Self-heating reaction and thermal runaway criticality of the lithium ion battery. International Journal of Heat and Mass Transfer doi:10.1016/j.ijheatmasstransfer.2019.119178.

Mara, T.A., Tarantola, S., 2012. Variance-based sensitivity indices for models with dependent inputs. Reliability Engineering and System Safety 107, 115–121. URL: http://dx.doi.org/10.1016/j.ress.2011.08.008, doi:10.1016/j.ress.2011.08.008.

Marcjasz, G., Serafin, T., Weron, R., 2018. Selection of calibration windows for day-ahead electricity price forecasting. Energies 11. URL: https://www.mdpi.com/1996-1073/11/9/2364, doi:10.3390/en11092364.

Marín, J.B., Orozco, E.T., Velilla, E., 2018. Forecasting electricity price in Colombia: A comparison between neural network, ARMA process and hybrid models. International Journal of Energy Economics and Policy 8, 97–106.

Maritnez-Alvarez, F., Troncoso, A., Asencio-Cortes, G., Riquelme, J.C., 2015. A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting. Energies 8, 13162–13193. doi:10.3390/en81112361.

Marrel, A., Iooss, B., Laurent, B., Roustant, O., 2009. Calculations of Sobol indices for the Gaussian process metamodel. Reliability Engineering and System Safety 94, 742–751. doi:10.1016/j.ress.2008.07.008.

Marrel, A., Iooss, B., Van Dorpe, F., Volkova, E., 2008. An efficient methodology for modeling complex computer codes with Gaussian processes. Computational Statistics and Data Analysis 52, 4731–4744. doi:10.1016/j.csda.2008.03.026, arXiv:0802.1099.

Marrel, A., Marie, N., De Lozzo, M., 2015. Advanced surrogate model and sensitivity analysis methods for sodium fast reactor accident assessment. Reliability Engineering and Sys-

tem Safety 138, 232–241. URL: http://dx.doi.org/10.1016/j.ress.2015.01.019, doi:10.1016/j.ress.2015.01.019.

Mason, J.A., Oktawiec, J., Taylor, M.K., Hudson, M.R., Rodriguez, J., Bachman, J.E., Gonzalez, M.I., Cervellino, A., Guagliardi, A., Brown, C.M., Llewellyn, P.L., Masciocchi, N., Long, J.R., 2015. Methane storage in flexible metal-organic frameworks with intrinsic thermal management. Nature 527, 357–361. URL: http://dx.doi.org/10.1038/nature15732, doi:10.1038/nature15732.

Matheron, G., 1963. Principles of Geostatistics. Economic Geology 58, 1246–1266. doi:10.2113/gsecongeo.58.8.1246.

Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions. Management science 22, 1087–1096.

McDonnell, J.D., Schunck, N., Higdon, D., Sarich, J., Wild, S.M., Nazarewicz, W., 2015. Uncertainty quantification for nuclear density functional theory and information content of new measurements. Physical Review Letters 114, 1–6. doi:10.1103/PhysRevLett.114.122501.

McKay, M.D., Beckman, R.J., Conover, W.J., 2000. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 42, 55–61. doi:10.1080/00401706.2000.10485979.

Meier, M., John, E., Wieckhusen, D., Wirth, W., Peukert, W., 2008. Characterization of the grinding behaviour in a single particle impact device: studies on pharmaceutical powders. European Journal of Pharmaceutical Sciences 34, 45–55.

Metta, N., Ghijs, M., Schäfer, E., Kumar, A., Cappuyns, P., Van Assche, I., Singh, R., Ramachandran, R., De Beer, T., Ierapetritou, M., Nopens, I., 2019. Dynamic flowsheet model development and sensitivity analysis of a continuous pharmaceutical tablet manufacturing process using the wet granulation route. Processes 7, 1–35. doi:10.3390/pr7040234.

Meyer, K., Bück, A., Tsotsas, E., 2015. Dynamic multi-zone population balance model of particle formulation in fluidized beds. Procedia engineering 102, 1456–1465.

Milton, R., Bugryniec, P., Brown, S., 2019. Parameter estimation for thermal runaway of li-ion cells: a gaussian process approach, in: Kiss, A.A., Zondervan, E., Lakerveld, R., Özkan, L. (Eds.), 29th European Symposium on Computer Aided Process Engineering. Elsevier. volume 46 of *Computer Aided Chemical Engineering*, pp. 775 – 780. URL: http://www.sciencedirect.com/science/article/pii/B9780128186343501302, doi:https://doi.org/10.1016/B978-0-12-818634-3.50130-2.

Milton, R.A., Brown, S.F., 2019. ROMCOMMA. URL: https://github.com/C-O-M-M-A/rom-comma.

Milton, R.A., Brown, S.F., 2022. Minimum reduced order modelling. In Preparation .

Mittal, B., 2017. Chapter 4 - pharmaceutical unit operations, in: Mittal, B. (Ed.), How to Develop Robust Solid Oral Dosage Forms from Conception to Post-Approval. Academic Press, pp. 69 – 95. URL: http://www.sciencedirect.com/science/article/pii/B9780128047316000042, doi:https://doi.org/10.1016/B978-0-12-804731-6.00004-2.

Mobley, R.K., 2002. 18 - world-class maintenance, in: Mobley, R.K. (Ed.), An Introduction to Predictive Maintenance (Second Edition). second edition ed.. Butterworth-Heinemann, Burlington. Plant Engineering, pp. 394–433. URL: https://www.sciencedirect.com/science/article/pii/B978075067531450018X, doi:https://doi.org/10.1016/B978-075067531-4/50018-X.

Moghadam, P.Z., Islamoglu, T., Goswami, S., Exley, J., Fantham, M., Kaminski, C.F., Snurr, R.Q., Farha, O.K., Fairen-Jimenez, D., 2018. Computer-aided discovery of a metal-organic framework with superior oxygen uptake. Nature Communications 9, 1–8. URL: http://dx.doi.org/10.1038/s41467-018-03892-8, doi:10.1038/s41467-018-03892-8.

Moghadam, P.Z., Ivy, J.F., Arvapally, R.K., Dos Santos, A.M., Pearson, J.C., Zhang, L., Tylianakis, E., Ghosh, P., Oswald, I.W., Kaipa, U., Wang, X., Wilson, A.K., Snurr, R.Q., Omary, M.A., 2017a. Adsorption and molecular siting of CO2, water, and other gases in the superhydrophobic, flexible pores of FMOF-1 from experiment and simulation. Chemical Science 8, 3989–4000. doi:10.1039/c7sc00278e.

Moghadam, P.Z., Li, A., Wiggin, S.B., Tao, A., Maloney, A.G., Wood, P.A., Ward, S.C., Fairen-Jimenez, D., 2017b. Development of a Cambridge Structural Database Subset: A Collection of Metal-Organic Frameworks for Past, Present, and Future. Chemistry of Materials 29, 2618–2625. doi:10.1021/acs.chemmater.7b00441.

Moghadam, P.Z., Rogge, S.M., Li, A., Chow, C.M., Wieme, J., Moharrami, N., Aragones-Anglada, M., Conduit, G., Gomez-Gualdron, D.A., Van Speybroeck, V., Fairen-Jimenez, D., 2019. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. Matter 1, 219–234. URL: https://www.sciencedirect.com/science/article/pii/S2590238519300062, doi:10.1016/J.MATT.2019.03.002.

Montagna, S., Tokdar, S.T., 2016. Computer emulation with nonstationary Gaussian processes. SIAM-ASA Journal on Uncertainty Quantification 4, 26–47. doi:10.1137/141001512, arXiv:1308.4756.

Monteiro, C., Ramirez-Rosado, I.J., Fernandez-Jimenez, L.A., 2018. Probabilistic electricity price forecasting models by aggregation of competitive predictors. Energies 11. doi:10.3390/en11051074.

Montgomery, D.C., Runger, G.C., 2014. Applied statistics and probability for engineers. John Wiley and Sons.

Mordjaoui, M., Haddad, S., Medoued, A., Laouafi, A., 2017. Electric load forecasting by using dynamic neural network. International Journal of Hydrogen Energy 42, 17655–17663. URL:

http://dx.doi.org/10.1016/j.ijhydene.2017.03.101, doi:10.1016/j.ijhydene.2017.03.101.

Mori, H., Nakano, K., 2015. Development of advanced Gaussian Process for LMP forecasting. 2015 18th International Conference on Intelligent System Application to Power Systems, ISAP 2015 , 1–6doi:10.1109/ISAP.2015.7325553.

Mortier, S.T.F.C., Gernaey, K.V., Beer, T.D., Nopens, I., 2014. Global Sensitivity Analysis Applied to Drying Models for One or a Population of Granules. AIChE Journal 60, 1700–1717. doi:https://doi.org/10.1002/aic.14383, arXiv:0201037v1.

Muniain, P., Ziel, F., 2020. Probabilistic forecasting in day-ahead electricity markets: Simulating peak and off-peak prices. International Journal of Forecasting 36, 1193–1210. URL: https://doi.org/10.1016/j.ijforecast.2019.11.006, doi:10.1016/j.ijforecast.2019.11.006.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.

de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F., 2016. Mean absolute percentage error for regression models. Neurocomputing (Amsterdam) 192, 38–48.

Najibi, F., Apostolopoulou, D., Alonso, E., 2021. Enhanced performance Gaussian process regression for probabilistic short-term solar output forecast. International Journal of Electrical Power and Energy Systems 130, 106916. URL: https://doi.org/10.1016/j.ijepes.2021.106916, doi:10.1016/j.ijepes.2021.106916.

Naumzik, C., Feuerriegel, S., 2020. Forecasting electricity prices with machine learning: predictor sensitivity. International Journal of Energy Sector Management doi:10.1108/IJESM-01-2020-0001, arXiv:2005.08005.

Nazarian, D., Camp, J.S., Sholl, D.S., 2016. A Comprehensive Set of High-Quality Point Charges for Simulations of Metal-Organic Frameworks. Chemistry of Materials 28, 785–793. doi:10.1021/acs.chemmater.5b03836.

Neal, R.M., 1995. Bayesian Learning for Neural Networks. Phd thesis. University of Toronto. URL: http://link.springer.com/10.1007/978-1-4612-0745-0.

Nitta, N., Wu, F., Lee, J.T., Yushin, G., 2015. Li-ion battery materials: present and future. Materials Today 18, 252 – 264. URL: http://www.sciencedirect.com/science/article/pii/S1369702114004118, doi:https://doi.org/10.1016/j.mattod.2014.10.040.

Nop, P., Qin, Z., 2021. Cambodia mid-term transmission system load forecasting with the combination of seasonal arima and gaussian process regression, in: 2021 3rd Asia Energy and Electrical Engineering Symposium (AEEES), pp. 700–707. doi:10.1109/AEEES51875.2021.9403196.

NordPool, . Historical Market Data. URL: https://www.nordpoolgroup.com/historical-market-data/.

Nowotarski, J., Weron, R., 2018. Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renewable and Sustainable Energy Reviews 81, 1548–1568. doi:10.1016/j.rser.2017.05.234.

Nyman, D., Levitt, J., 2006. Maintenance Planning, Scheduling and Coordination. 2nd edition ed., Industrial Press Inc.

Oakley, J.E., O'Hagan, A., 2004. Probabilistic sensitivity analysis of complex models : a Bayesian approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66, 751–769. doi:1369âĂŞ7412/04/66751.

O'Hagan, A., Kingman, J.F.C., 1978. Curve Fitting and Optimal Design for Prediction. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 40, 1–42.

Ohno, H., Mukae, Y., 2016. Machine Learning Approach for Prediction and Search: Application to Methane Storage in a Metal-Organic Framework. Journal of Physical Chemistry C 120, 23963–23968. doi:10.1021/acs.jpcc.6b07618.

Pallas, N.R., Harrison, Y., 1990. An automated drop shape apparatus and the surface tension of pure water. Colloids and Surfaces 43, 169–194.

Palmer, R.D., 2013. Maintenance planning and scheduling handbook. McGraw-Hill Education.

Pape, C., Hagemann, S., Weber, C., 2016. Are fundamentals enough? Explaining price variations in the German day-ahead and intraday power market. Energy Economics 54, 376–387. URL: http://dx.doi.org/10.1016/j.eneco.2015.12.013, doi:10.1016/j.eneco.2015.12.013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peng, P., Jiang, F., 2016. Thermal safety of lithium-ion batteries with various cathode materials: A numerical study. International Journal of Heat and Mass Transfer 103, 1008–1016. doi:10.1016/j.ijheatmasstransfer.2016.07.088.

Pesaran, M.H., Timmermann, A., 2007. Selection of estimation window in the presence of breaks. Journal of Econometrics 137, 134–161. URL: https://www.sciencedirect.com/science/article/pii/S0304407606000418, doi:https://doi.org/10.1016/j.jeconom.2006.03.010.

Peters, J.F., Baumann, M., Zimmermann, B., Braun, J., Weil, M., 2017. The environmental impact of li-ion batteries and the role of key parameters – a review. Renewable and Sustainable Energy Reviews 67, 491 – 506. URL: http://www.sciencedirect.com/science/article/pii/S1364032116304713, doi:https://doi.org/10.1016/j.rser.2016.08.039.

Pohlman, D.A., 2015. Multi-scale modeling of high-shear granulation. Ph.D. thesis. Purdue University.

Pohlman, D.A., Litster, J.D., 2015. Coalescence model for induction growth behavior in high shear granulation. Powder Technology 270, 435–444.

Prytz, R., Nowaczyk, S., Rögnvaldsson, T., Byttner, S., 2015. Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. Engineering Applications of Artificial Intelligence 41, 139–150. doi:10.1016/j.engappai.2015.02.009.

Qader, M.R., Khan, S., Kamal, M., Usman, M., Haseeb, M., 2021. Forecasting carbon emissions due to electricity power generation in Bahrain. Environmental Science and Pollution Research URL: https://doi.org/10.1007/s11356-021-16960-2, doi:10.1007/s11356-021-16960-2.

Qian, P.Z., Wu, H., Wu, C.F., 2008. Gaussian process models for computer experiments with qualitative and quantitative factors. Technometrics 50, 383–396. doi:10.1198/004017008000000262.

Qian, Z., Seepersad, C.C., Joseph, V.R., Allen, J.K., Wu, C.F., 2006. Building surrogate models based on detailed and approximate simulations. Journal of Mechanical Design, Transactions of the ASME 128, 668–677. doi:10.1115/1.2179459.

Qiao, Z., Xu, Q., Cheetham, A.K., Jiang, J., 2017. High-Throughput Computational Screening of Metal-Organic Frameworks for Thiol Capture. Journal of Physical Chemistry C 121, 22208–22215. doi:10.1021/acs.jpcc.7b07758.

Qiao, Z., Xu, Q., Jiang, J., 2018. High-throughput computational screening of metal-organic framework membranes for upgrading of natural gas. Journal of Membrane Science 551, 47–54. URL: https://doi.org/10.1016/j.memsci.2018.01.020, doi:10.1016/j.memsci.2018.01.020.

Queipo, N.V., Haftka, R.T., Shyy, W., Goel, T., Vaidyanathan, R., Kevin Tucker, P., 2005. Surrogate-based analysis and optimization. Progress in Aerospace Sciences 41, 1–28. doi:10.1016/j.paerosci.2005.02.001, arXiv:arXiv:1011.1669v3.

Questier, F., Put, R., Coomans, D., Walczak, B., Heyden, Y.V., 2005. The use of CART and multivariate regression trees for supervised and unsupervised feature selection. Chemometrics and Intelligent Laboratory Systems 76, 45–54. doi:10.1016/j.chemolab.2004.09.003.

Rafiei, M., Niknam, T., Khooban, M.H., 2017. Probabilistic electricity price forecasting by improved clonal selection algorithm and wavelet preprocessing. Neural Computing and Applications 28, 3889–3901. doi:10.1007/s00521-016-2279-7.

Rajan, A., Vijayaraghavan, V., Ooi, M.P.L., Garg, A., Kuang, Y.C., 2018. A simulation-based probabilistic framework for lithium-ion battery modelling. Measurement: Journal of the International Measurement Confederation 115, 87–94. URL: http://dx.doi.org/10.1016/j.measurement.2017.10.033, doi:10.1016/j.measurement.2017.10.033.

Ramkrishna, D., Mahoney, A.W., 2002. Population balance modeling. Promise for the future. Chemical Engineering Science 57, 595–606.

Ren, D., Liu, X., Feng, X., Lu, L., Ouyang, M., Li, J., He, X., 2018. Model-based thermal runaway prediction of lithium-ion batteries from kinetics analysis of cell components. Applied Energy 228, 633–644. URL: https://doi.org/10.1016/j.apenergy.2018.06.126, doi:10.1016/j.apenergy.2018.06.126.

Richard, M.N., Dahn, J.R., 1999a. Accelerating Rate Calorimetry Study on the Thermal Stability of Lithium Intercalated Graphite in Electrolyte I. Experimental. Journal of The Electrochemical Society 146, 2068–2077. doi:10.1149/1.1391894.

Richard, M.N., Dahn, J.R., 1999b. Accelerating Rate Calorimetry Study on the Thermal Stability of Lithium Intercalated Graphite in Electrolyte. II. Modeling the Results and Predicting Differential Scanning Calorimeter Curves. Journal of The Electrochemical Society 146, 2078–2084. doi:10.1149/1.1391894.

Richardson, R.R., Osborne, M.A., Howey, D.A., 2017. Gaussian process regression for forecasting battery state of health. Journal of Power Sources 357, 209–219. URL: http://dx.doi.org/10.1016/j.jpowsour.2017.05.004, doi:10.1016/j.jpowsour.2017.05.004.

Roberts, D., Brown, S., 2020. Identifying calendar-correlated day-ahead price profile clusters for enhanced energy storage scheduling. Energy Reports 6, 35–42. doi:10.1016/j.egyr.2020.02.025.

Rodrigues, F., Cardeira, C., Calado, J.M., 2014. The daily and hourly energy consumption and load forecasting using artificial neural network method: A case study using a set of 93 households in Portugal. Energy Procedia 62, 220–229. URL: http://dx.doi.org/10.1016/j.egypro.2014.12.383, doi:10.1016/j.egypro.2014.12.383.

Rohmer, J., Foerster, E., 2011. Global sensitivity analysis of large-scale numerical landslide models based on Gaussian-Process meta-modeling. Computers and Geosciences 37, 917–927. URL: http://dx.doi.org/10.1016/j.cageo.2011.02.020, doi:10.1016/j.cageo.2011.02.020.

ROMCOMMA, 2019. https://github.com/C-O-M-M-A/rom-comma. [Online]. [Accessed on 25 November 2019].

Rouzbahman, M., Jovicic, A., Chignell, M., 2017. Can Cluster-Boosted Regression Improve Prediction of Death and Length of Stay in the ICU? IEEE Journal of Biomedical and Health Informatics 21, 851–858. doi:10.1109/JBHI.2016.2525731.

Sacks, J., Welch, W., Mitchell, T., Wynn, H., 1989. Design and analysis of computer experiments. Journal of Statistical Science 4, 409–435. doi:10.1214/ss/1177012413.

Saltelli, A., Homma, T., 1996. Importance measures in global sensitivity analysis of model output. Reliab. Eng. Sys. Safety 52, 1–17.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis. The primer ed., John Wiley & Sons Ltd.

Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2005. Sensitivity Analysis for Chemical Models. Chemical Reviews 105, 2811–2828. URL: https://pubs.acs.org/doi/10.1021/cr040659d, doi:10.1021/cr040659d.

dos Santos, T., Ferreira, F.J.T.E., Pires, J.M., Damásio, C., 2017. Stator winding short-circuit fault diagnosis in induction motors using random forest, in: 2017 IEEE International Electric Machines and Drives Conference, (IEMDC), pp. 1–8. doi:10.1109/IEMDC.2017.8002350.

Saraiva, J.T., Fidalgo, J.N., 2016. Eem2016 price forecast competition. URL: http://www.eem2016.com/price-forecast-competition/COMPLATT{_}MKT{_}V4.pdf. [Accessed: 11-Nov-20].

Sarkar, D., Contal, E., Vayatis, N., Dias, F., 2016. Prediction and optimization of wave energy converter arrays using a machine learning approach. Renewable Energy 97, 504–517. URL: http://dx.doi.org/10.1016/j.renene.2016.05.083, doi:10.1016/j.renene.2016.05.083.

Sarkar, D., Osborne, M.A., Adcock, T.A., 2018. Prediction of tidal currents using Bayesian machine learning. Ocean Engineering 158, 221–231. URL: https://doi.org/10.1016/j.oceaneng.2018.03.007, doi:10.1016/j.oceaneng.2018.03.007.

Sayin, R., 2016. Mechanistic studies of twin screw granulation. Ph.D. thesis. Purdue University.

Schmidt, A.P., Bitzer, M., Imre, Á.W., Guzzella, L., 2010. Experiment-driven electrochemical modeling and systematic parameterization for a lithium-ion battery cell. Journal of Power Sources 195, 5071–5080. doi:10.1016/j.jpowsour.2010.02.029.

See, A.v., 2021. Total data volume worldwide 2010-2025. URL: https://www.statista.com/statistics/871513/worldwide-data-created/.

Sen, M., Singh, R., Vanarase, A., John, J., Ramachandran, R., 2012. Multi-dimensional population balance modeling and experimental validation of continuous powder mixing processes. Chemical Engineering Science 80, 349–360.

Serafin, T., Uniejewski, B., Weron, R., 2019. Averaging Predictive Distributions Across Calibration Windows for Day-Ahead Electricity Price Forecasting. Energies 12, 1–12. doi:10.3390/en12132561.

Shack, P., Iannello, C., Rickman, S., Button, R., 2014. NASA Perspective and Modeling of Thermal Runaway Propagation Mitigation in Aerospace Batteries.

Shepero, M., van der Meer, D., Munkhammar, J., Widén, J., 2018. Residential probabilistic load forecasting: A method using Gaussian process designed for electric load data. Applied Energy 218, 159–172. URL: https://doi.org/10.1016/j.apenergy.2018.02.165, doi:10.1016/j.apenergy.2018.02.165.

Shi, M., Lv, L., Sun, W., Song, X., 2020. A multi-fidelity surrogate model based on support vector regression. Structural and Multidisciplinary Optimization 61, 2363–2375. URL: http://dx.doi.org/10.1007/s00158-020-02522-6, doi:10.1007/s00158-020-02522-6.

Shirazian, S., Ismail, H.Y., Singh, M., Shaikh, R., Croker, D.M., Walker, G.M., 2019. Multi-dimensional population balance modelling of pharmaceutical formulations for continuous twin-screw wet granulation: Determination of liquid distribution. International Journal of Pharmaceutics 566. doi:10.1016/j.ijpharm.2019.06.001.

Simon, C.M., Mercado, R., Schnell, S.K., Smit, B., Haranczyk, M., 2015. What Are the Best Materials to Separate a Xenon/Krypton Mixture? Chemistry of Materials 27, 4459–4475. doi:10.1021/acs.chemmater.5b01475.

Simpson, T.W., Peplinski, J.D., Koch, P.N., Allen, J.K., 2001. Metamodels for computer-based engineering design: Survey and recommendations. Engineering with Computers 17, 129–150. doi:10.1007/PL00007198.

Smith, R.M., 2007. Wet Granule Breakage in High Shear Mixer Graunlators. Ph.D. thesis. The University of Queensland.

Smrčka, D., Dohnal, J., Štěpánek, F., 2015. Effect of process scale-up on the dissolution of granules with a high content of active pharmaceutical ingredient. Powder Technology 285, 88–95.

Snelson, E., 2007. Flexible and efficient Gaussian process models for machine learning. Phd thesis. University College London. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.4041{&}rep=rep1{&}type=pdf{%}5Cnhttp://portal.acm.org/citation.cfm?id=1117456.

Snelson, E., Ghahramani, Z., 2005. Sparse Gaussian Processes using Pseudo-inputs. Advances in Neural Information Processing Systems , 1257–1264.

Sobol, I.M., 1993. Sensitivity analysis for nonlinear mathematical models. Mathematical Modelling Computational Experiments 1, 407–414. URL: http://max2.ese.u-psud.fr/epc/conservation/MODE/SobolOriginalPaper.pdf{%}0Ahttp://www.mathnet.ru/eng/mm2320, doi:10.18287/0134-2452-2015-39-4-459-461., arXiv:arXiv:1305.4373v1.

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models. Review. Mathematics and Computers in Simulation , 271–280.

Sood, A., Awasthi, A., Bharti, R., 2016. A population balance model for butyl acrylate emulsion polymerization. Indian Chemical Engineer 58, 40–60.

Spinner, N.S., Mazurick, R., Brandon, A., Rose-pehrsson, S.L., Tuttle, S.G., 2015. Analytical, Numerical and Experimental Determination of Thermophysical Properties of Commercial 18650 LiCoO 2 Lithium-Ion Battery. Journal of The Electrochemical Society 162, A2789–A2795. doi:10.1149/2.0871514jes.

Spotnitz, R., Franklin, J., 2003. Abuse behavior of high-power, lithium-ion cells. Journal of Power Sources 113, 81–100.

Srivastava, A., Subramaniyan, A.K., Wang, L., 2017. Analytical global sensitivity analysis with Gaussian processes. Artificial Intelligence for Engineering Design, Analysis and Manufacturing: AIEDAM 31, 235–250. doi:10.1017/S0890060417000142.

Staffell, I., Green, R., 2016. Is There Still Merit in the Merit Order Stack? The Impact of Dynamic Constraints on Optimal Plant Mix. IEEE Transactions on Power Systems 31, 43–53. doi:10.1109/TPWRS.2015.2407613.

Steen, M., Lebedeva, N., Di Persio, F., Boon-Brett, L., 2017. EU Competitiveness in Advanced Li-ion Batteries for E-Mobility and Stationary Storage Applications – Opportunities and Actions. Technical Report. European Commission's Joint Research Centre. URL: http://publications.jrc.ec.europa.eu/repository/bitstream/JRC108043/kjna28837enn.pdf. [accessed 1 September 2018].

Stein, M., 1987. Large sample properties of simulations using Latin hypercube sampling. Technometrics 29, 143–151. URL: http://www.jstor.org/stable/10.2307/1269769.

Steinert, R., Ziel, F., 2019. Short- to mid-term day-ahead electricity price forecasting using futures. Energy Journal 40, 105–127. doi:10.5547/01956574.40.1.rste, arXiv:1801.10583.

Su, G., Peng, L., Hu, L., 2017. A Gaussian process-based dynamic surrogate model for complex engineering structural reliability analysis. Structural Safety 68, 97–109. URL: http://dx.doi.org/10.1016/j.strusafe.2017.06.003, doi:10.1016/j.strusafe.2017.06.003.

Sudret, B., 2008. Global sensitivity analysis using polynomial chaos expansions. Reliability Engineering and System Safety 93, 964–979. doi:10.1016/j.ress.2007.04.002.

Sulttan, S., Rohani, S., 2019. Coupling of CFD and population balance modelling for a continuously seeded helical tubular crystallizer. Journal of Crystal Growth 505, 19–25.

Sumner, T., Shephard, E., Bogle, I.D.L., 2012. A methodology for global-sensitivity analysis of time-dependent outputs in systems biology modelling. Journal of The Royal Society Interface 9, 2156–2166. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2011.0891, doi:10.1098/rsif.2011.0891, arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2011.0891.

Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A., 2015. Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics 11, 812–820. doi:10.1109/TII.2014.2349359.

Tanaka, N., Bessler, W.G., 2014. Numerical investigation of kinetic mechanism for runaway thermo-electrochemistry in lithium-ion cells. Solid State Ionics 262, 70–73. URL: http://dx.doi.org/10.1016/j.ssi.2013.10.009, doi:10.1016/j.ssi.2013.10.009.

Thomas, D.S., Weiss, B.A., 2020. Economics of manufacturing machinery maintenance: A Survey and Analysis of U.S. Costs and Benefits Douglas. Technical Report. National Institute of Standards and Technology. Gaithersburg, MD. URL: https://nvlpubs.nist.gov/nistpubs/ams/NIST.AMS.100-34.pdf, doi:10.6028/NIST.AMS.100-34.

Thornton, A.W., Simon, C.M., Kim, J., Kwon, O., Deeg, K.S., Konstas, K., Pas, S.J., Hill, M.R., Winkler, D.A., Haranczyk, M., Smit, B., 2017. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. Chemistry of Materials 29, 2844–2854. doi:10.1021/acs.chemmater.6b04933.

Tian, T., Zeng, Z., Vulpe, D., Casco, M.E., Divitini, G., Midgley, P.A., Silvestre-Albero, J., Tan, J.C., Moghadam, P.Z., Fairen-Jimenez, D., 2018. A sol-gel monolithic metal-organic framework with enhanced methane uptake. Nature Materials 17, 174–179. doi:10.1038/NMAT5050.

Titsias, M.K., 2009. Variational learning of inducing variables in sparse Gaussian processes. Proceedings of the 12th International Conference on Artificial Intelligence and Statistics , 567–574.

Tran, A.L.H., 2015. Powder flow in bertical high shear mixer granulators. Ph.D. thesis. The University of Queensland.

Trembacki, B., Harris, S.R., Piekos, E.S., Roberts, S.A., 2016. Uncertainty Quantification, Verification, and Validation of a Thermal Simulation Tool for Molten Salt Batteries. 47th Power Sources Conference URL: https://www.osti.gov/servlets/purl/1365182.

Van Bockstal, P.J., Mortier, S.T.F., Corver, J., Nopens, I., Gernaey, K.V., De Beer, T., 2018. Global Sensitivity Analysis as Good Modelling Practices tool for the identification of the most influential process parameters of the primary drying step during freeze-drying. European Journal of Pharmaceutics and Biopharmaceutics 123, 108–116. URL: https://doi.org/10.1016/j.ejpb.2017.12.006, doi:10.1016/j.ejpb.2017.12.006.

Van Horenbeek, A., Pintelon, L., 2013. A dynamic predictive maintenance policy for complex multi-component systems. Reliability Engineering and System Safety 120, 39–50. URL: http://dx.doi.org/10.1016/j.ress.2013.02.029, doi:10.1016/j.ress.2013.02.029.

Vassiliadis, C.G., Pistikopoulos, E.N., 2001. Maintenance scheduling and process optimization under uncertainty. Computers and Chemical Engineering 25, 217–236. doi:10.1016/S0098-1354(00)00647-5.

Vazquez-Arenas, J., Gimenez, L.E., Fowler, M., Han, T., Chen, S.K., 2014. A rapid estimation and sensitivity analysis of parameters describing the behavior of commercial Li-ion batteries including thermal analysis. Energy Conversion and Management 87, 472–482. URL: http://dx.doi.org/10.1016/j.enconman.2014.06.076, doi:10.1016/j.enconman.2014.06.076.

Viana, F.A., Gogu, C., Goel, T., 2021. Surrogate modeling: tricks that endured the test of time and some recent developments. Structural and Multidisciplinary Optimization 64, 2881–2908. doi:10.1007/s00158-021-03001-2.

Villa-Vialaneix, N., Follador, M., Ratto, M., Leip, A., 2012. A comparison of eight metamodeling techniques for the simulation of N 2O fluxes and N leaching from corn crops. Environmental Modelling and Software 34, 51–66. URL: http://dx.doi.org/10.1016/j.envsoft.2011.05.003, doi:10.1016/j.envsoft.2011.05.003.

Vogel, L., Peukert, W., 2005. From single particle impact behaviour to modelling of impact mills. Chemical Engineering Science 60, 5164–5176.

Wang, K., Gasser, T., 1997. Alignment of curves by dynamic time warping. Annals of Statistics 25, 1251–1276. doi:10.1214/aos/1069362747.

Wang, L.G., Morrissey, J.P., Sousani, M., Barrasso, D., Slade, D., Hanley, K., Ooi, J.Y., Litster, J.D., 2019. Model driven design in particulate products manufacturing, in: International Granulation Workshop, Lausanne, Switzerland.

Wang, Q., Ping, P., Zhao, X., Chu, G., Sun, J., Chen, C., 2012. Thermal runaway caused fire and explosion of lithium ion battery. Journal of Power Sources 208, 210–224. doi:10.1016/j.jpowsour.2012.02.038.

Wang, S., Lu, L., Liu, X., 2013. A simulation on safety of $LiFePO_4$/C cell using electrochemical-thermal coupling model. Journal of Power Sources 244, 101–108. doi:10.1016/j.jpowsour.2013.03.100.

Wang, Y., Yang, M., Wei, G., Hu, R., Luo, Z., Li, G., 2014. Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy. Sensors and Actuators, B: Chemical 193, 723–729. URL: http://dx.doi.org/10.1016/j.snb.2013.12.028, doi:10.1016/j.snb.2013.12.028.

Wang, Z., Escotet-Espinoza, M.S., Ierapetritou, M., 2017. Process analysis and optimization

of continuous pharmaceutical manufacturing using flowsheet models. Computers and Chemical Engineering 107, 77–91. URL: https://doi.org/10.1016/j.compchemeng.2017.02.030, doi:10.1016/j.compchemeng.2017.02.030.

Weron, R., 2006. Modeling and forecasting electricity loads and prices: A statistical approach. Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. Cited By :602.

Weron, R., 2014. Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting 30, 1030–1081. URL: http://dx.doi.org/10.1016/j.ijforecast.2014.08.008, doi:10.1016/j.ijforecast.2014.08.008.

Wiener, N., 1949. Extrapolation, Interpolation and Smoothing of Stationary Time Series. MIT Press.

Williams, C.K.I., Rasmussen, C.E., 1996. Gaussian Processes for Regression. Advances in Neural Information Processing Systems , 514–520.

Williams, C.K.I., Rasmussen, C.E., 2006. Gaussian processes for machine learning. The MIT Press, London. URL: http://www.gaussianprocess.org/gpml/chapters/.

Wilmer, C.E., Leaf, M., Lee, C.Y., Farha, O.K., Hauser, B.G., Hupp, J.T., Snurr, R.Q., 2012. Large-scale screening of hypothetical metal-organic frameworks. Nature Chemistry 4, 83–89. doi:10.1038/nchem.1192.

Wipf, D.P., Nagarajan, S., 2007. A New View of Automatic Relevance Determination, in: Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS'07, Curran Associates Inc.. pp. 1625–1632. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.8009{&}rep=rep1{&}type=pdf.

Wu, H., Xie, Z., Wang, Y., Zhang, P., Sun, L., Lu, C., Ma, Z., 2019. A constitutive model coupling irradiation with two-phase lithiation for lithium-ion battery electrodes. Philosophical Magazine 99, 992–1013. doi:10.1080/14786435.2019.1569767.

Wu, Y., Maravelias, C.T., Wenzel, M.J., ElBsat, M.N., Turney, R.T., 2021. Predictive maintenance scheduling optimization of building heating, ventilation, and air conditioning systems. Energy and Buildings 231, 110487. URL: https://doi.org/10.1016/j.enbuild.2020.110487, doi:10.1016/j.enbuild.2020.110487.

Xie, X., Schenkendorf, R., Krewer, U., 2019. Efficient sensitivity analysis and interpretation of parameter correlations in chemical engineering. Reliability Engineering and System Safety 187, 159–173. URL: http://www.sciencedirect.com/science/article/pii/S0951832018300541, doi:https://doi.org/10.1016/j.ress.2018.06.010. sensitivity Analysis of Model Output.

Xu, C., Gertner, G.Z., 2008. Uncertainty and sensitivity analysis for models with correlated parameters. Reliability Engineering and System Safety 93, 1563–1573. doi:10.1016/j.ress.2007.06.003.

Xu, J., Lan, C., Qiao, Y., Ma, Y., 2017. Prevent thermal runaway of lithium-ion batteries with minichannel cooling. Applied Thermal Engineering 110, 883–890. URL: http://dx.doi.org/10.1016/j.applthermaleng.2016.08.151, doi:10.1016/j.applthermaleng.2016.08.151.

Yan, J., Meng, Y., Lu, L., Li, L., 2017. Industrial Big Data in an Industry 4.0 Environment: Challenges, Schemes, and Applications for Predictive Maintenance. IEEE Access 5, 23484–23491. doi:10.1109/ACCESS.2017.2765544.

Yan, X., Chowdhury, N.A., 2013. Mid-term electricity market clearing price forecasting: A hybrid LSSVM and ARMAX approach. International Journal of Electrical Power and Energy Systems 53, 20–26. URL: http://dx.doi.org/10.1016/j.ijepes.2013.04.006, doi:10.1016/j.ijepes.2013.04.006.

Yeardley, A.S., Bellinghausen, S., Milton, R.A., Litster, J.D., Brown, S.F., 2021. Efficient global sensitivity-based model calibration of a high-shear wet granulation process. Chemical Engineering Science 238, 116569. URL: https://doi.org/10.1016/j.ces.2021.116569, doi:10.1016/j.ces.2021.116569.

Yeardley, A.S., Bugryniec, P.J., Milton, R.A., Brown, S.F., 2020a. A study of the thermal runaway of lithium-ion batteries: A Gaussian Process based global sensitivity analysis. Journal of Power Sources 456, 228001. URL: https://doi.org/10.1016/j.jpowsour.2020.228001, doi:10.1016/j.jpowsour.2020.228001.

Yeardley, A.S., Ejeh, J.O., Allen, L., Brown, S.F., Cordiner, J., 2022a. Predictive Maintenance in the Digital Era, in: 32nd European Symposium on Computer Aided Process Engineering, Elsevier B.V.

Yeardley, A.S., Milton, R.A., Moghadam, P.Z., Cordiner, J., Brown, S.F., 2022b. Active subsets as a tool for structural characterisation and selection of metal-organic frameworks. Chemical Engineering Research and Design 179, 424–434. URL: https://linkinghub.elsevier.com/retrieve/pii/S0263876222000533, doi:10.1016/J.CHERD.2022.01.045.

Yeardley, A.S., Roberts, D., Milton, R., Brown, S.F., 2020b. An Efficient Hybridization of Gaussian Processes and Clustering for Electricity Price Forecasting, in: 30th European Symposium on Computer Aided Process Engineering, Elsevier B.V.. pp. 343–348.

Yin, S., Kaynak, O., 2015. Big Data for Modern Industry: Challenges and Trends. Proceedings of the IEEE 103, 143–146. doi:10.1109/JPROC.2015.2388958.

Zhang, L., Lyu, C., Hinds, G., Wang, L., Luo, W., Zheng, J., Ma, K., 2014. Parameter Sensitivity Analysis of Cylindrical LiFePO 4 Battery Performance Using Multi-Physics Modeling . Journal of The Electrochemical Society 161, A762–A776. doi:10.1149/2.048405jes.

Zhang, W., Cheema, F., Srinivasan, D., 2018. Forecasting of electricity prices using deep

learning networks. Asia-Pacific Power and Energy Engineering Conference, APPEEC 2018-October, 451–456. doi:10.1109/APPEEC.2018.8566313.

Zhang, Z., Wang, C., Peng, X., Qin, H., Lv, H., Fu, J., Wang, H., 2021. Solar radiation intensity probabilistic forecasting based on k-means time series clustering and gaussian process regression. IEEE Access 9, 89079–89092. doi:10.1109/ACCESS.2021.3077475.

Zhao, R., Gu, J., Liu, J., 2014. An investigation on the significance of reversible heat to the thermal behavior of lithium ion battery through simulations. Journal of Power Sources 266, 422–432. doi:10.1016/j.jpowsour.2014.05.034.

Ziel, F., Steinert, R., 2018. Probabilistic mid- and long-term electricity price forecasting. Renewable and Sustainable Energy Reviews 94, 251–266. URL: https://doi.org/10.1016/j.rser.2018.05.038, doi:10.1016/j.rser.2018.05.038, arXiv:1703.10806.

Zonta, T., da Costa, C.A., da Rosa Righi, R., de Lima, M.J., da Trindade, E.S., Li, G.P., 2020. Predictive maintenance in the Industry 4.0: A systematic literature review. Computers and Industrial Engineering 150. doi:10.1016/j.cie.2020.106889.