Advancing Analytics to Support Nuclear Asset Lifecycle Management through Explainable Data Pipeline Design and Complex Dependency Modelling

PhD Thesis

Jennifer Blair

Institute for Energy and Environment

Department of Electronic and Electrical Engineering

University of Strathclyde, Glasgow

June 7, 2025

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

Nuclear plant operators require trusted data analytics tools to support the management of asset health throughout their operating lifetimes. Management of the data pipeline that serves data analytic tools, alongside the development of the analytic tools themselves, creates an ecosystem whereby operators can more effectively access the risk associated with the utilisation of data-driven systems within their decision-making processes. Prognostics and health management, and structural health monitoring practices allow nuclear power plant operators to monitor the state of assets and structures in the plant to avoid the financial strain and loss of generation from unexpected faults. However, for these technologies to be adopted, they must have high accuracy to prevent false alarms or missed faults, which can degrade operator trust in these tools. There is a need for trustworthy analytics across the nuclear sector, with analytic tools capable of uncertainty quantification to attribute risk to analytic outputs, and an understanding of uncertainty sources in the full data pipeline serving these analytics.

This work firstly investigates the impact of data pipeline design on analytic performance by using a SHAP-based explainability tool to form part of a novel pipeline design interrogation framework. This framework identifies the highest impact positive and negative performance drivers, providing informed design decisions for data pipelines to improve performance of analytic tools within these pipelines. The process was shown to be transferable to the data pipeline designs of similar assets with less available design data, leveraging insights from one system to reduce the uncertainty sources within designs across other systems for improved fleetwide monitoring.

Secondly, this work demonstrates the development of a novel copula-based calibration module within a hierarchical modelling structure which is used to improve pre-

dictions of transparent, but interchangeable, base models that are commonly applied within the highly regulated nuclear sector. The approach has the additional benefit of uncertainty quantification which attributes risk to the final prediction. The procedure was shown to be effective for spatial and temporal data, demonstrating applicability to a diverse set of engineering applications.

The methods developed in this work have made progress towards providing trustworthy data analytic tools and data pipeline designs to provide nuclear operators with the risk associated with applying such tools to the management of the health and maintenance of their assets.

Contents

List of Figures

List of Figures

List of Tables

Acknowledgements

A PhD isn't a journey undertaken and completed alone. We rely on guidance and support from our supervisors, mentors and trainers. We rely on the camaraderie and love from our peers, friends and family. I have a lot of people to be grateful to over the years spent working towards my degree, and hopefully I have in turn made a positive impact in other people's lives.

To my supervisor, Bruce, I thank you for your patience, interest, guidance, kindness and humour. My development is thanks to your care and it was your unique way of sharing your subject ("Is it a possum or a rock?") that caused me to pursue further education in the first place! You taught me that perceived failures are a part of the process, and to not shy away from them. I've learned how to see them as opportunities for further questions and to try again, which has impacted my life beyond my research. I have thoroughly enjoyed our Tinderbox discussions, and your mentorship over ... 6 years?! I truly appreciate all the time you've spent working with my style of learning to make sure I get the most out of my degree (undergraduate, internships and PhD!). To my supervisory team, Blair and Alistair, I am grateful for your understanding and praise, through all the good and the bad updates. It softened the blow of disappointments, and allowed me to try again with newfound resolve. During, and after, my time as an NPL PGI Communications Ambassador, Linden and Leah have made it their mission to make sure their students feel welcomed, valued and appreciated, which has improved my confidence and creativity immensely. My few poster prizes are thanks to the training and experience you passed on to me!

To my friends, and peers in the PhD, your friendship and understanding has meant the world to me! All of the discussions and distractions have made me both a better

and happier person over my time at the University. To those who partook, I hope you enjoyed the lunches, movie nights, games nights, and board games at the Union as much as I did.

My family don't understand what I'm doing or why I'm doing it... but are happy for me, regardless! They have celebrated my victories and commiserated my losses, purely because they see how important it is to me. I am grateful for their understanding and support; and for the trips, meals and chats to bring me back out into the real world where PhD problems don't matter.

To my partner, Tim, there is not much I can say here, except perhaps dedicate this thesis to you. You make me want to be a better person and scientist, and with your love, dedication and support throughout this, I hope I can say I have been successful. Your outlook on life and research is inspiring, and I hope I am able to look at life from your perspective one day.

Thank you to everyone who has made this work possible, and more importantly, made my life enjoyable throughout the process. I hope you keep spreading kindness and look out for one another, to make academia a fulfilling place to be.

'If you do the right thing in the here and now, the future has a way of taking care of itself'

Dolly Parton, The Orville

'We all do better when we all do better'

Captain Ed Mercer, The Orville

Chapter 1

Introduction

1.1 Quantifying uncertainty in data analytics for nuclear applications

In power plants, operation and maintenance (O&M) account for a large portion of generation costs, estimated to be between 40 - 70 % across different areas of the world [1]. Streamlining and supporting this process has been given great attention over the lifetime of the nuclear sector, which benefits from innovation exchange between similar industries. Originally, most maintenance processes were conducted reactively once a failure was evident and relied entirely on expert knowledge to diagnose and rectify. This is the most expensive course of action [2], as assets must be taken offline with little notice and the asset can be impacted by cascading faults, where a failure in one part of the asset creates issues in wider subsystems. This incurs delays while the faults are diagnosed, parts are sourced and arrangements are made to work around the interruption, which has likely resulted in lost revenue [2]. With the introduction of condition monitoring (CM) and condition based maintenance (CBM) [3], the health status of assets are more closely monitored to allow maintenance to be scheduled with equipment and interruptions planned for in advance, streamlining the process. In power plants, such maintenance may involve replacing or lubricating bearings in rotating plant such as pumps or motors [4]; identifying corrosion in piping [5] or cracks in civil infrastructure [6]; or preventing electrical faults through replacing degraded wiring [7].

To accurately assess the health status of monitored assets or identify developing faults, data pertaining to important health indicators must be collected, which vary from asset to asset and fault type to fault type. For example, this can include temperature [8], or vibration monitoring [9], which may be measured by different types of sensors. Data from these sensors can then be utilised to not only assess the health of the asset at the current time, but be used to identify developing trends which can permit predictions to be made on when the asset is likely to fail, known as predictive maintenance [10]. Predictive maintenance utilises data streams monitoring different process variables with data analytics to detect or diagnose developing faults, or predict when the fault will develop into a system failure.

Data pipelines describe the flow and transformation of data through important stages in a data acquisition system, from collection by the measurement system, through to life-long storage and management. Accuracy in data pipelines can be eroded by data quality, which can be circumvented with appropriate system design and calibration [11]. The form of (and range of) uncertainty in the CM data acquisition lifecycle and its impact on data analytics within the pipeline is currently not fully understood and hence is unquantified, necessitating a comprehensive study of uncertainty and its propagation in data acquisition systems.

To incorporate the output of data analytics into maintenance decisions, operators require thorough understanding of the risks incurred from the data quality of the data pipeline and the type of analytics applied to the desired predictive problem. This provides flexibility and agency in how the outputs of the analytics are utilised to prevent operators losing confidence in their data analytic tools. Investigating sources of uncertainty and incorporating models which are capable of uncertainty quantification is one method of attributing this risk to the data-driven system, or removing barriers to the models success in the first place by improving data quality.

1.2 Scope and objectives

In this thesis, the uncertainty sources in the lifecycle of power plant asset data, alongside methods and models able to quantify or explain this uncertainty, are investigated to

support analytics for electrical, mechanical and civil engineering applications of interest to the nuclear sector. This is approached through a variety of questions that are subsequently investigated in the following chapters. Of interest to this thesis is the impact of pipeline design on data analytics performance, the choice of analytics and modelling strategy, and their effectiveness across diverse data types and applications, as covered by the following research questions:

- What impact, if any, does the pipeline design have on analytic performance and behaviour? If design changes degrade or improve model performance, can these be reliably identified and explained?
- Can additional uncertainty quantification be built into the analytics stage, and models or modelling approaches be combined to improve system robustness?

These questions have additional opportunities for investigation depending on the outcome. This includes how different types of machine learning model may be impacted differently under the same circumstances within a data pipeline design, potentially adding another source of uncertainty to consider. Additionally, how the utilisation of explainable models and explainability tools may affect analytic system transparency, or allow explanations of one system pipeline to potentially be utilised for developing designs in other systems. Lastly, to provide generalised approaches which are suitable for multiple applications within data-based systems of relevance to the nuclear industry, the approaches developed should be flexible to cover temporal and spatial data applications.

1.3 Research novelty and contribution

Current data-based systems are at risk of deteriorated performance from so called "rubbish in - rubbish out", whereby any issues with data quality, data system design or data analytics development can understandably produce untrustworthy outputs. Understanding sources of uncertainty within the pipeline design and on the outputs of machine learning model predictions can provide flexibility to users of these systems to make decisions based on the level of provided risk. The research questions laid out

in Section 1.2 aim to cover this risk from the data pipeline and the data analytics within the pipeline, and the contributions of this thesis provide potential solutions. A detailed breakdown of the novelty and contribution can be found at the beginning of each technical chapter, but can be summarised as follows:

- In this work, a methodology is developed to demonstrate the impact of data pipeline design on data-based and hybrid models across prognostic, diagnostic and detection applications which is showcased on bearing fault data. This methodology is able to decouple and explain which design choices improve or degrade analytic performance, and quantify the strength of this impact, allowing for the informed design of condition monitoring systems involving data analytics. This methodology is able to leverage design insights from one asset pipeline to those of similar systems, promoting more efficient fleetwide monitoring.

- Focusing on the analytics stage of the pipeline, a hierarchical modelling methodology is developed to combine the advantages of explainable models with models capable of uncertainty quantification to provide robust timeseries forecasting, demonstrated in this case on a temperature monitoring application.

- The hierarchical approach comprised of explainable and uncertainty quantification capable models were generalised further and applied to spatial data to capture uncertainty in the error of a structural health monitoring rig. As part of this process, a structural health monitoring spatial dataset based on a photometric stereo rig was collected, curated and released. The dataset covered civil engineering structural damage and materials while diversifying using household objects to assess wider applicability across diverse geometries.

Each contribution is directly associated with answering and exploring beyond the provided research questions in Section 1.2. The first contribution (associated with Chapter 3) is associated with the first question, while the second and third contribution (associated with Chapter 4 and Chapter 5, respectively) investigate the second research question.

This thesis is conducted as part of a collaboration between the National Physical Laboratory, the University of Strathclyde, and partners of the Advanced Nuclear Research Centre. This agreement has supported this research by providing access to diverse expertise, equipment and industrial data which has been collected under realistic operating environments. At time of writing, there are three publications associated with this thesis: Chapter 3 has a published conference paper [12], and a published journal paper [13]; and Chapter 5 has a published data article [14] with open source data released.

1.4 Organisation and structure

This thesis is organised as follows: Chapter 2 contains the literature review covering topics of interest across several chapters, including data analytics for condition monitoring in engineering, uncertainty quantification and trustworthiness, and dependency modelling. Each technical chapter includes a brief literature review of topics pertaining to that chapter. In Chapter 3, the methodology used to provide the impact, explanation and leveraging of uncertainty sources in data pipeline design is presented; Chapter 4 brings focus to the analytics stage of the data pipeline where a hierarchical modelling structure with uncertainty quantification is presented on a forecasting application for the monitoring of nuclear power plant heat exchanger temperature data; in Chapter 5, the hierarchical modelling approach with uncertainty quantification was adapted for handling spatial data from a structural health monitoring rig. Finally, the thesis is concluded in Chapter 6, with a summary and future work discussion.

Chapter 2

Literature: Uncertainty quantification in prognostics and health management for asset condition monitoring

2.1 Prognostics and health management and structural health monitoring for asset condition monitoring

Historically, without insight into the health status of an asset, maintenance of industrial equipment was scheduled reactively [15] where technicians diagnose and perform maintenance after a failure has occurred. This incurs lengthy and expensive downtimes of equipment and in some scenarios can pose safety concerns. With the development of expert knowledge, maintenance could be performed proactively where known fault symptoms could be identified and addressed before developing into a system failure [16].

2.1.1 Types of maintenance strategies and methodologies

Experience-based maintenance informed time-based maintenance schedules to enable timely replacements or upkeep of parts before fault symptoms emerged [17]. This was

especially important for safety critical assets and improving maintenance schedule costs while avoiding run-to-failure scenarios. The development of affordable sensor systems allowed for condition monitoring (CM), providing more diverse information about the current performance and health status of an asset. With greater understanding of degradation modes in assets, condition-based maintenance (CBM) [18] became possible, whereby the nature and severity of the developing fault can be established through data-based or physics-based models before removing the monitored asset from use. This can allow operators to forecast potential maintenance schedule options to optimise asset management actions and prevent the replacement of healthy parts until it is required. This eventually led to prognostic and health management (PHM) processes, where the combination of historical sensor data and analytics can forecast potential degradation scenarios, or diagnose developing faults based on the assets condition [19]. For 'passive' assets, such as piping, asset housing or concrete, structural health monitoring (SHM) [20] is applied to estimate asset condition and detect or diagnose fault scenarios. If an asset experiences common failure modes, aggressive maintenance techniques such as 'design-out' strategies, involve design changes in the next generation of the asset to mitigate the development of specific failure types [21, 22].

2.1.2 Condition based maintenance cycle

The CBM system contains five stages from collecting sensor data from the monitored asset through to informed decisions being made to optimise maintenance actions [23,24]:

1. Acquire condition monitoring data
2. Signal processing and data preparation
3. Condition Monitoring
    - Feature selection
    - Statistical Models/Machine Learning models
    - Fault Diagnosis via operators/machine learning models
4. Prognostics and Health Management

- Calculate Remaining Useful Life and Probability of Failure, potentially using Physics of Failure models.

5. Operation and Maintenance systems

- Condition based maintenance using cost-benefit analysis and from Remaining Useful Life and Probability of Failure estimates to inform risk budgets and maintenance scheduling.

Diagnosis seeks to identify the location of the fault (isolate fault and trace to a specific component) and the type of fault occurring (identify fault and type of damage to the component). Fault symptoms can be signatures from sensed data, anomaly detection results, residuals from system monitoring or features extracted from sensed data. Expert systems [25] or classification algorithms [26] can be used to diagnose faults and estimate fault severity from these sources. Prognosis takes the fault diagnosis and attempts to determine the RUL of the system [27]. Physics based models describe the condition of the asset/part by performing computations using the systems underlying degradation and failure behaviour [28]. These are useful in situations with a lack of empirical data (such as cases where assets cannot be run to failure due to cost or safety factors), however limited understanding of the physics of failure, uncertainties in the model and the potential oversimplification of the model physics can make them inaccurate [29]. Improvements in accuracy and sophistication of these methods also incur greater computational expense. Empirical models model the relationship between normal real world system behaviour and different types of fault behaviour observed in historical collected data [30]. These are less complex and easier to develop, however require representative historical data to build robust models. In many applications, equipment has not (or cannot) been run to failure in order to collect comprehensive and diverse fault data [31], limiting the model based on available training data and known fault modes which may have been collected from test rig equipment under non representative operating conditions [23]. With new technologies and system integrations becoming increasingly complex, a single method of informing maintenance may not be enough, inspiring the creation of different modelling types and strategies.

2.1.3 Condition monitoring for PHM and SHM in engineering applications

Maintenance is required to ensure the health and reliability of many assets across different engineering disciplines: from conveyor system imbalance in smart manufacturing [32]; to broken railway tracks in the transport industry [33]; to wind turbine blade erosion in the energy sector [34]. In this section, some examples of where maintenance may be required in the aeronautical, power systems and power generation industries are briefly discussed to demonstrate the diversity of fault types and where maintenance strategies may be shared across different disciplines.

Aeronautical engineering

Aircraft require a high level of reliability in their electrical and mechanical systems, as they must operate autonomously in highly dangerous environments. To ensure personnel and customer safety, maintenance schedules must be balanced to maintain high performance of the assets while ensuring the industry remains profitable.

In manned aircraft, fault management systems can be used to protect the electrical systems using fault isolators to separate the faulted sub-system from healthy sub-systems; current limiters to reduce current spikes to prevent overloading other components; and current divertors to protect sensitive subsystems from fault currents [35]. Redundancy in the internal systems can keep aircraft operational despite occurring faults, allowing the system to reconfigure to maintain functionality until maintenance can be performed [36]. The performance of aircraft engines can degrade over time due to the accumulation of wear and impact of various stresses from the different operating environments and conditions [37]. Predicting the remaining useful life of engines can provide a risk of failure to be estimated based on the degradation trends of the engine, to allow maintenance to be scheduled before faults occur [38]. Aircraft rely heavily on sensors to detect abnormal operation in subsystems, but these sensors themselves can become faulty. Sensors can experience excess noise, oscillations, excessive drift or become stuck at a given output which can create issues in flight control systems [39]. To mitigate this, redundancy is also applied to sensors, whereby different voting mecha-

nisms can be employed to reduce the impact of incorrect or unstable measurements and detect sensor issues [40]. Due to the fossil fuel consumption in the aviation industry, there is a push for hybridisation or electrification of their future designs [41], which will create new fault modes and maintenance requirements.

Unmanned aerial vehicles (UAVs), such as drones, can also be a tool in maintenance strategies for other assets by delivering non-destructive testing sensors to testing location sites. The use of drones for structural health monitoring applications allows maintenance crews to conduct their work remotely without the safety risks of working at height [42]. Visual inspection can be conducted through UAVs equipped with cameras, which can allow for: the detection of cracks, corrosion and impact damage from birds or hail on aircraft exteriors [43]; for the diagnosis of erosion, mechanical damage and lightning damage in wind turbine blades [44]; and, for the assessment of civil infrastructure such as bridges, as either a part of routine inspection or for safety assessments after natural disasters or extreme weather events [45]. Alternatively, sensors requiring surface contact can be deployed to detect subsurface defects which are not detectable through visual assessments. Ultrasonic sensors use reflections of pressure waves to detect subsurface defects such as the corrosion of storage tanks or offshore platforms in the oil and gas industry [46]. Eddy current sensors induce magnetic fields in metal components which can diagnose subsurface faults, such as wall thinning in pipes [47].

Power systems

Power networks in Great Britain consist of generation sources and high voltage transmission networks connected to medium-low voltage distribution networks via transformers to step the voltage, and protection devices designed to prevent, isolate or clear faults. Each of these sections consist of many components requiring maintenance. Electrical and thermal stresses inside transformers can weaken the insulation [48], putting the system at risk of overheating and partial discharges [49]. Power lines responsible for transporting power over long distances can either be above ground supported by utility poles [50], or buried (on land [51], or undersea [52]). Safe transmission line

operation is limited by thermal ratings. The temperature of cables is influenced by the line loading and environmental conditions, which impact the lifetime of the conductor material [53] and, for overhead lines, the sag of the line [54]. Above ground power lines are additionally susceptible to damage from adverse weather, animal or human involvement, and flora encroachment or damage [55]. The utility poles themselves are also subjected to rot from environmental conditions and wildlife encroachment, such as from fungi or insects [56], necessitating structural health monitoring to detect faults which may not always be visible externally.

Circuit breakers are responsible for isolating faulted sections of the power grid to prevent faults cascading through the network [57]. Circuit breakers are susceptible mechanical faults which can target the internal spring system used to trigger the opening and closing mechanism, such as spring wear or jamming [58], or wearing of the spring dampening mechanism [59]. Abnormal operation of the circuit breaker supply voltage can also prevent normal operation, or even lead to breakdown if it exceeds expected voltage limits [60]. The health of circuit breakers which may not operate for long stretches of time can be inferred through tests during normal grid operation to prevent undetected faults being observed when the circuit breakers are needed [57].

Generation sources can be from power plants, renewables or different forms of battery storage. Wind turbines are some of the most utilised renewables in the UK, contributing 21.3 % of generated electricity in July 2024[1]. Wind turbines contain generators, converters, transformers, gearboxes, and pitch and yaw systems. Generators are responsible for the conversion of mechanical energy to electricity, and are susceptible to bearing faults [61]. Convertors enable the electrical output of the turbine to be compatible with connection to the wider grid [62] and can experience short or open circuit faults due to the long term effects of thermal stress, moisture, debris accumulation or voltage/current spikes [63]. Wind turbines do not rotate at the required speed of electrical generators, and so gearboxes are utilised (in some designs) to compensate for this difference [64]. Variability in lifetime can come from operating environment, material properties and design defects which can result in mechanical failures in the

[1]National Grid, Great Britain's monthly electricity stats, https://www.nationalgrideso.com/electricity-explained/electricity-and-me/great-britains-monthly-electricity-stats

gear teeth, bearings and shafts [64]. External components of interest are mainly on the health of the turbine blades which are exposed to icing or lightning damage, cracks, delamination, fatigue, corrosion and general wear [65].

Thermal generation and nuclear power plants

In a simplified view, power plants are constructed from several key components: combustion engines, boilers or nuclear reactors for heat generation; heat exchangers for steam generation (in most plant types); turbines and motors for energy generation; pumps to circulate the coolant medium, and condensers to restart the cycle. Gas turbines operate under high thermal and mechanical stress conditions which can degrade the performance of the asset [66]. Chao et al [67] developed a calibration and uncertainty quantification method for a hybrid gas turbine model used to predict turbine performance at observed or new operating conditions. Monitoring of the heat exchange between steam and sodium in the super heater of the 'Monju' fast breeder reactor was conducted by Gofuku [68] through a hybrid physics- and data-based diagnostic system. Forecasting the reactor inlet header temperature for monitoring asset aging is also presented in this thesis in Chapter 4. Rotating plant, such as motors and pumps, rely heavily on bearings and gears [69] which are subjected to high speeds, temperatures, and other stresses from friction, which can result in cracks, pitting and wear [70]. Across the variations of these designs and cycles, there are many sensor types across key components in the plant, such as pressure sensors, temperature sensors, or vibration sensors [71]. Sensor systems in plants can be linked through sensor selection to identify logical connections in sensor anomalies which suggest certain faults, for example: anomalies in the flowrate sensor of the inner coolant loop and the inner loop temperature sensor can together suggest a developing fault in the coolant pump [72]. Sensor faults can cause the state of the plant to become uncertain and risks developing faults from going unnoticed due to miscalibration, failure or drift [73]. Online calibration techniques can be used to detect sensor calibration errors while minimising physical intervention from maintenance staff in harsh environments [74]. Such techniques may involve cross-calibration, where a collection of redundant sensors measuring the same

process on the same asset are compared to provide an average estimate of the plant state and a measure of deviation for each sensor [75].

Power plants and surrounding industries also contain many passive components requiring monitoring, such as the reactor housing [76], reactor fuel channels [77, 78], nuclear waste storage containers [79], piping [80], and concrete in supporting civil structures [81].

2.1.4 CBM and PHM in civil nuclear generation: Barriers and opportunities

To meet carbon targets in the energy sector, fossil fuel powered thermal plants are being phased out to reduce the amount of CO_2 being released into the atmosphere [82]. With the increased adoption of distributed renewable energy sources into the energy network, there is a need for a reliable, high power, low carbon alternative to coal plants to supply regional base loads for the energy network. Nuclear is a mature technology with the potential to fulfill this role [83]. In this section, additional context surrounding maintenance in the nuclear sector will be discussed.

Nuclear plants across the globe were commissioned and built in the late 1970s and 1980s with an expected 30-40 year lifespan [84, 85]. As such, many nuclear plants are approaching the end of their initially expected lifespan, but with the cost and planning difficulty associated with commissioning new sites and decommissioning the old, many plant operators are looking at lifetime extension as the route forward [85]. However, aging assets may experience unexpected and lengthy downtimes while faults are diagnosed, and the appropriate maintenance actions and equipment are organised. To improve operating efficiency, prevent penalties and loss of generation revenues due to asset downtime, many companies have turned to prognostic and health management and condition monitoring techniques to monitor the health of aging assets more closely [23]. This will have increased relevance in Great Britain, where growth of nuclear generation is being planned to provide quadruple the current generation capability over the next 25 years[2].

[2] "Biggest expansion of nuclear power for 70 years to create jobs, reduce bills and

In the nuclear sector, retrofitting modern sensors has been slow and expensive to do, with novel technologies posing new opportunities and new problems. For example, a shift to wireless data transfer avoids the need for shielded cables in adverse environments [86], but introduces cybersecurity risks and reliability concerns due to electromagnetic interference [87]. Similarly, new technologies such as industrial control systems, can be vulnerable to cyber-physical attacks which have been increasing in frequency in recent years [88]. These attacks may involve loss of communications or access to systems due to denial of service attacks; and in severe cases, may either hide faulty operation by masking real sensor readings or cause control systems to act unnecessarily due to fake data injected into the system, putting assets at risk of automated tripping and shutdowns [89]. In their 2023 report, The Nuclear Threat Initiative[3], a non-profit global security organisation, gave 17 out of 47 assessed countries the lowest score category in cybersecurity protection for nuclear facilities, with the overall median at 50 %, demonstrating a lack in sufficient cybersecurity measures [90].

It is estimated that 60-70 % of generating cost for US nuclear plants is due to operation and maintenance, however successful PHM schemes are estimated to save up to $1 billion/year by avoiding loss of generation revenue and penalties from unscheduled downtime of equipment [1,91]. Future generations of nuclear plants would benefit from being designed for improved PHM and SHM adoption to avoid issues experienced while upgrading legacy stations [92]. To improve current PHM practices in the nuclear sector, additional barriers identified by Coble et al [23] include: limited understanding of physics of failure preventing improvement to physics based models, sensor choice and placement schemes; limited uncertainty quantification poses problems for risk management; lack of specialised signal and feature extraction techniques, and techniques which are computationally feasible in cases with large datasets; trade-off between sensitivity and accuracy of diagnostic models to developing faults [85]; and robust, affordable sensor technologies and calibration schemes [24].

strengthen Britain's energy security", Department for Energy Security and Net Zero, https://www.gov.uk/government/news/biggest-expansion-of-nuclear-power-for-70-years-to-create-jobs-reduce-bills-and-strengthen-britains-energy-security

[3]https://www.ntiindex.org/about-the-nti-index/

2.2 Machine learning analytics for prognostics and health management

2.2.1 Data-based models

What machine learning models are available, and what domain the models receive the condition monitoring data in are important considerations in implementing analytic tools. The shortcomings and bias of these options will result in different types of error rates which may have more severe consequences in a given industrial application. There are several categories which describe how models are trained which are roughly covered by supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Supervised learning fits a model by learning from provided input-output pairs to either predict a discrete or continuous variable [93]. This is suitable for classification tasks, where the inputs are mapped to a set of discrete labels which represent meaningful groupings or states. For example, this can be used for fault detection or fault diagnosis, where the inputs may contain 'symptoms' of different developing faults which can be identified by the model [94]. Supervised learning is also suitable for regression tasks, where the model fits trends in the data which map to continuous outputs. This is suitable for prediction tasks such as forecasting (where future trends are predicted from observed historical data) [95] and remaining useful life (where the health of the asset is estimated and its degradation trajectory to component failure is predicted) [96]. Unsupervised learning is used to identify structure or trends in the data without known example pairs [97], making it suitable for clustering or anomaly detection tasks. Clustering tasks identify grouping in the data which have a commonality identified based on a combination of features which can be used to identify assets at different health stages or fault states [98]. Semi-supervised learning is a hybrid approach between supervised and unsupervised learning, which may be an appropriate compromise based on the type or amount of labelled data available [99]. Lastly, reinforcement learning allows a model to develop based on set rewards and penalties which it is designed to adapt to maximise and minimise, respectively [100]. This can allow for more flexible prognostics of assets when the data-space fluctuates due to changing

operating conditions [101]. While model task examples were given for each discussed model learning category, the model tasks are not limited to only their allocated learning category. For example, anomaly detection was discussed under unsupervised learning, however this can also be achieved through other learning categories. There is plenty of flexibility to be found depending on the circumstances and priorities, for example, should the model design be task-driven, data-driven or environment-driven [100]. Further discussion on regression models is given in Chapter 4, so data-based classification models will be further discussed in this section.

Families of machine learning models attribute classification boundaries to input data through different methods. Decision Tree models partition the data into progressively smaller sets by learning decision rules based on the predictors which, ideally, lead to the most information gained in the least number of splits. The complexity of tree models can be controlled through limiting the number of splits per node, or the depth the tree can grow to prevent overfitting. Other methods exist, such as pruning, which removes branches in the trained model with the least amount of information gained. Tree models are advantageous as they can capture hierarchical structures in the data and remain interpretable, however they can provide different interpretations of the same data if the data is reordered [102].

Support Vector Machine (SVM) models perform classification tasks by constructing hyperplanes in high dimensional data spaces which aim to maximise the margin between different class samples and capture the maximum number of samples of each class within the boundaries. The kernel function used (linear, polynomial, Radial Basis Function, sigmoid, etc) allows input data to be mapped to a feature space that allows a simpler hyperplane to be fit to the data. The SVM's kernel can allow complex structures within the data to be captured through mapping to a more insightful feature space, however this also makes the model difficult to interpret and the models performance contingent on the suitability of the kernel chosen [103].

Naive Bayes models use density estimates to assign samples to the most probable class. Despite the model's assumption of conditional independence of the predictor variables (which does not hold true in most applications) the posterior distributions of

the classifiers tend to be robust to bias in the density estimates of each class, allowing the Naive Bayes classifiers to often outperform more complex models [104].

The k-Nearest Neighbours (kNN) algorithm will implement a majority vote based on the nearest training data samples to a query point using a chosen distance metric. Different distance metrics (Euclidean, cosine, chebychev, etc) and different distance weightings can result in the same data being clustered differently which may improve the model fit. This mitigates a notable shortcoming of the kNN algorithm which is it's vulnerability to error in cases where the nearest neighbours cover a large distance or there is a large class imbalance, and a smaller number of 'closer' neighbours are more reliable for predicting the class of the current query point [104, 105].

Ensemble models average predictions from many 'weak learners' which are simpler models trained on different subsets or orderings of the training data to reduce bias in the predictions. This allows the relative strength of the models being ensembled (which may be the same type of model or a collection of model types) to be leveraged while, ideally, the impact of the individual model weaknesses are reduced through the averaging of all outputs. Ensemble methods can also allow for epistemic (model) uncertainty estimation [106].

2.2.2 Hybrid models

A survey of 274 prognostic approaches by Lei et al [107] separated works into statistical-, AI-, physics- and hybrid-based approaches, with 56% contribution from statistical based methods, and 26% from AI based approaches which both rely heavily on available CM data. ML or Deep Learning (DL) approaches are gaining increasing popularity as they can handle complex prognosis problems which may be traditionally difficult to create reliable physics or statistical models for, however due to their black-box nature it is difficult to justify their usage in safety critical applications. The approaches which gained the most attention for machine prognosis in Lei et al [107] review were Artificial Neural Networks, Neuro-Fuzzy systems (both DL methods), Support Vector Machines (SVM), K-nearest neighbour (kNN) and Gaussian Process Regression. DL approaches require access to large quantities of high quality, representative data which can be unob-

tainable in some industrial settings, however can produce excellent RUL predictions in return. ML models such as the SVM and kNN methods can provide better performance in cases with limited access to representative data, however are subject to appropriate kernel and parameter selection [103]. Gaussian Process Regression are computationally expensive when utilising large number of samples due to a required matrix inversion, but is a flexible method that can be updated with new data, adapt to limited data and incorporate uncertainties [108].

A single knowledge-, physics- or data- based approach is unlikely to provide effective system coverage for multiple failure modes and fault types. Utilising a combination of approaches aims to leverage the relative advantages of each individual method while limiting the impact of their respective weaknesses [109]. Hybrid modelling strategies utilise combinations of experience (through capturing expert knowledge), data (through collecting sensor measurements) and physics (simulating the expected physical behaviour of the asset) based modelling techniques to enhance understanding of asset degradation modes [21]. Goebel et al [110] found that combining a bearing physics of failure model with an empirical method based on measured data (Dempster-Shafer Regression) produced more accurate RUL prediction results than either method independently. Similarly, Chao et al [111] utilise a hybrid physics and data based deep learning model which was found to provide an extended remaining useful life prediction horizon for turbofan engines, while requiring less training data and suffering less from the limitations of the chosen training data. Kundu et al [112] found that a hybrid physics- and data-based prognostics framework allowed the limitations on gear damage thresholds tied to historical training data can be relieved through a hybrid approach, allowing users more flexibility to detect damage thresholds of interest.

The method of combining two or more of these methods in a hybrid approach varies and tends to be application specific due to the relatively early development stage of the research field as shown by the small (8 %) contribution to the canvassed literature in Lei et al's review [107]. As such, many methods of creating hybrid models are being explored, such as utilising one model to estimate the asset health state and another for RUL estimation; combining the RUL estimates from multiple methods; or utilising

one method for short-term forecasting and another method for long-term forecasting [113]. Of particular interest in this work is the combination of knowledge- and data-based approaches. Incorporating domain knowledge into data-driven approaches allows known trends and rules that govern the degradation patterns to be encoded to support the prognostic tool in identifying and predicting the failure dynamics of well understood failure modes. The data-driven component can provide the needed flexibility to apply and extrapolate these rules into an RUL estimate tailored to the monitored asset, while providing capability to identify new failure modes not included in the encoded expert knowledge [114]. This approach was employed by von Hahn et al [115] who presented a knowledge informed machine learning approach created via the inclusion of a Weibull-based loss function (derived from the field of reliability engineering) in a neural network model.

2.2.3 Dependency modelling, model calibration and uncertainty quantification

Computer models of physical processes are not able to account for differences due to real world factors, be it manufacturing, environmental or operational differences which accumulate and fluctuate over the lifetime of a monitored asset, or important governing quantities which are difficult to estimate. This causes the models simulated expected behaviour to deviate from the actual behaviour of the asset. Using a hybrid approach, physics based models can be calibrated using measured data to compensate for this difference, and account for some uncertainty in the model. Using a Bayesian approach, Kennedy and O'Hagan [11] were able to incorporate all types of uncertainty previously discussed in the research space while capturing the discrepancies between the model and real application not accounted for by the best parameter predictions. The forms of uncertainty included parameter uncertainty, random effects, model inaccuracy, data collection errors and uncertainty of the unseen code output. In both Chao et al [67] and Hart [116], Gaussian Processes were used to calibrate and provide uncertainty quantification for physics models for gas turbine and wind turbine applications, respectively.

Gaussian Processes are multivariate normal probability distributions over functions.

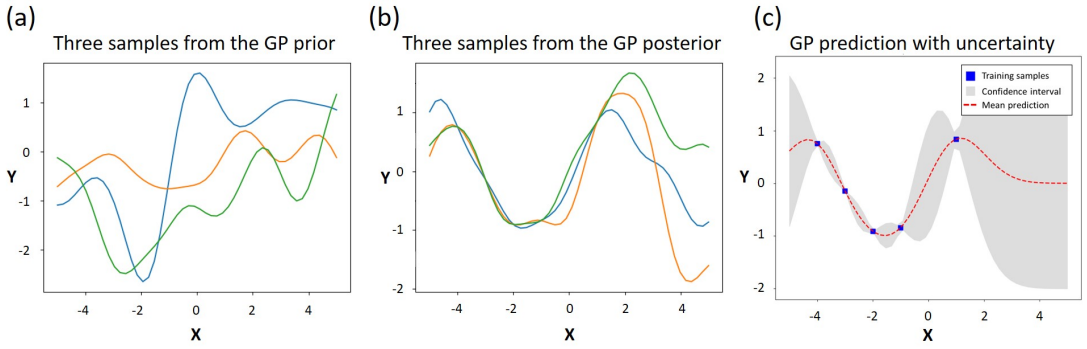

Figure 2.1: a) Three randomly sampled functions taken from the Gaussian Process prior, b) Three randomly sampled functions taken from the Gaussian Process posterior (these are functions which fit well to the training samples), and c) An example output of the Gaussian Process, showing the training samples, the mean prediction (most probable function) and confidence interval which captures uncertainty in the output.

The mean of this probability distribution signifies the most likely function that fits the observed data, while the standard deviation across the probable functions is used to attribute a confidence interval to the fitted distributions. The covariance kernel used can allow prior knowledge of the expected characteristics of the distribution governing the relation between the input and output data to be encoded. This will constrain the potential functions considered as prior distributions and can also be used with the training and testing data to constrain the most probable functions to those that pass through the training data points (noiseless case) or close to the training points (noisy case). The covariance kernel can incorporate a scale factor which governs the strength of a data points influence on its neighbours and the variance of expected noise in the data. Gaussian Processes incorporate and propagate uncertainty estimation in an intuitive manner, with lower uncertainty margins near observed data points, and larger uncertainty margins far from observed data points where there is no information to constrain the likely functional relation [117, 118]. An example of the Gaussian Process priors, posteriors and outputs are shown in Figure 2.1. Gaussian Processes are powerful tools in uncertainty capture and model calibration, however have known scalability issues due to potentially large matrix inversions during computation [119], making them currently unsuitable for large datasets or very high dimensional modelling.

Uncertainty quantification can be provided across several model types due to tech-

niques such as conformal prediction. Conformal prediction is a technique which uses a non-conformity measure to convert input samples into prediction regions around the model output with a given confidence (usually chosen to be 95 %) [120]. For input samples which are similar to already observed samples, the confidence is higher, resulting in a smaller prediction region, and vice versa for samples which deviate greatly from previous observations. The use of such tools for uncertainty quantification can be used for additional insights into data quality and systematic issues. In Olsson et al [121], conformal prediction was used to detect systematic differences in data collection systems which led to degraded predictions, limit miss-classifications compared to the base model alone, and identify cases most at risk for miss-classification in a cancer diagnosis application. Conformal prediction is capable of providing region predictions for traditional point estimator models across regression and classification tasks; is capable of operating in different states as required, from offline to online systems; and provides limited assumptions except assuming samples are independent and identically distributed [122]. However, it has been shown that limiting assumptions on the underlying distributions has disadvantages, such as preventing conditional coverage guarantees and limiting validity estimates for individual observations [123].

Copulas are a type of statistical model capable of dependency modelling, prediction and uncertainty quantification tasks. They are a general model that can be adapted to a wide variety of applications, are scalable to high dimensions and capable of capturing linear and non-linear dependencies. Copulas are applied in Chapter 4 and Chapter 5, and will be discussed in more detail in Section 2.2.3. In Section 2.2.3, a brief introduction to the fundamental theory of copulas is provided, along with a discussion on the assumptions and limitations of fitting marginal univariates and common copula families. This is followed by their extension to higher dimensions and their use with respect to uncertainty quantification in other fields, such as biomedical [124], earth science [125] and financial applications [126].

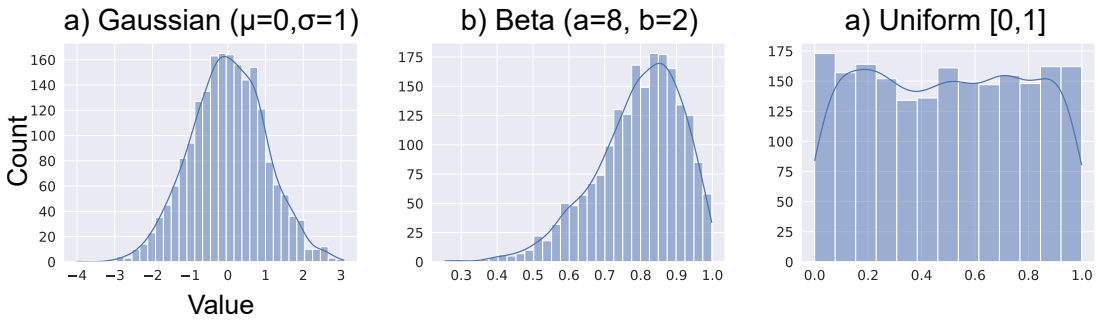

Figure 2.2: Example of common univariate distributions showing the histograms and KDE estimates of a) Gaussian, b) Beta and c) Uniform distributed variables.

Copulas

Sklar's theorem [127] is used to describe a joint distribution function in N-dimensions, G_N, over random variables $x_1, ..., x_N$ as a function of univariate uniform marginals with interval [0,1] (given by passing x_i through its cumulative distribution function (CDF), F_i, for $i = 1, ..., N$) and a unique copula, C_N:

$$G_N(x_1, ..., x_N) = C_N(F_1(x_1), ..., F_N(x_N)) \tag{2.1}$$

This has the convenience of allowing the joint distribution to be specified separately in terms of its dependency and marginals.

To utilise the appropriate cumulative distribution function to transform the marginals, the marginals $(x_1, ..., x_N)$ must first be fitted by the appropriate univariate distribution. This may be from common families of distributions, such as Gaussian or Beta distributions, or empirical methods such as Kernel Density Estimates (KDE). An example of common univariate families are shown in Figure 2.2, which shows Gaussian, Beta and Uniform histograms and KDE estimations. Non-parametric methods such as KDE impart minimal assumptions on the properties of the distribution which allow for capturing important detail in the data (for example, bi-modal structures) which may not be represented in common parametric methods, however, the accuracy of non-parametric methods may vary depending on the sample size of given data [128].

With fitted univariate marginals, the CDF can be computed via equations or estimated empirically to provide uniform marginals to fit the copula. As with the univariate

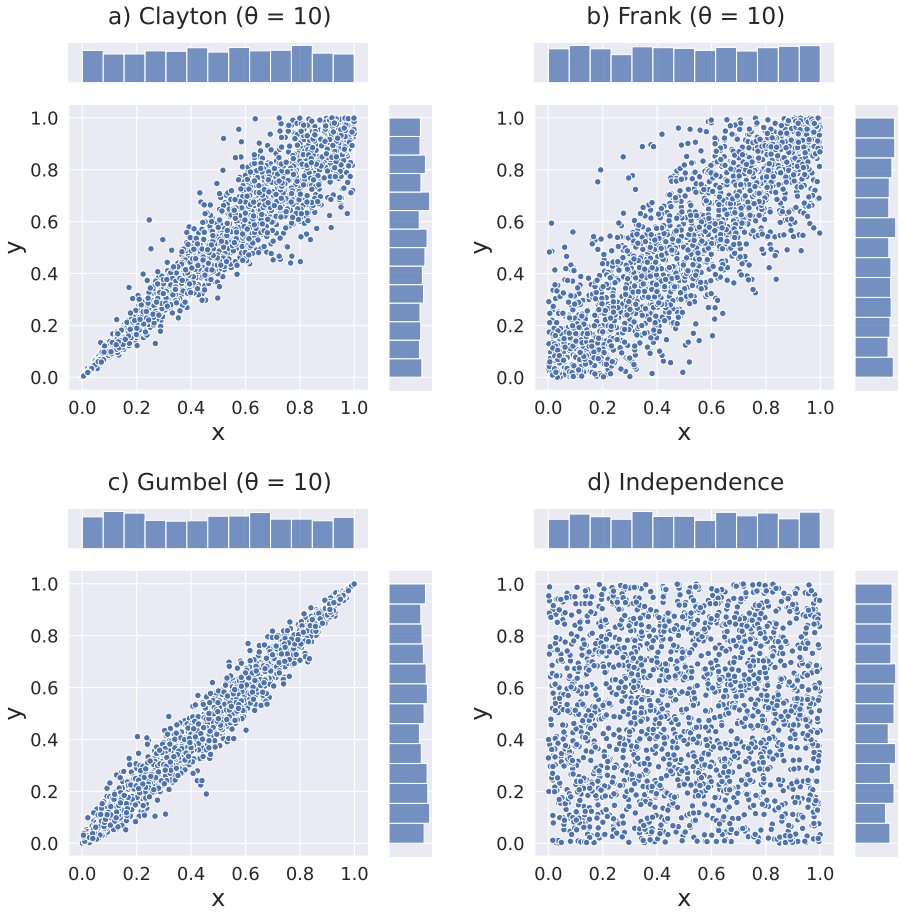

Figure 2.3: Examples of common copula families showing a) the Clayton copula (asymmetric, strong lower tail dependence), b) the Frank copula (symmetric, low upper and lower tail dependence) the Gumbel copula (strong upper and lower tail dependence), and d) independence (no dependency relation between variables)

case, there are well-known families of copula which capture different behaviours: such as Clayton copulas with strong lower tail dependence; Gumbel copulas with strong upper and lower tail dependence, and Gaussian copulas which capture elliptical dependence. These well-defined cases are limited to two dimensions. An example of common copula families are shown in Figure 2.3. Contour plots are a visualisation method which can be used to assess the shape of the fitted copula. This can rule out obvious issues with the fitted copula model, and provide a general sense of how difficult the underlying data structure may be to capture. An example contour plot is shown in Figure 2.4 which includes a good and poor copula fit compared to example target data.

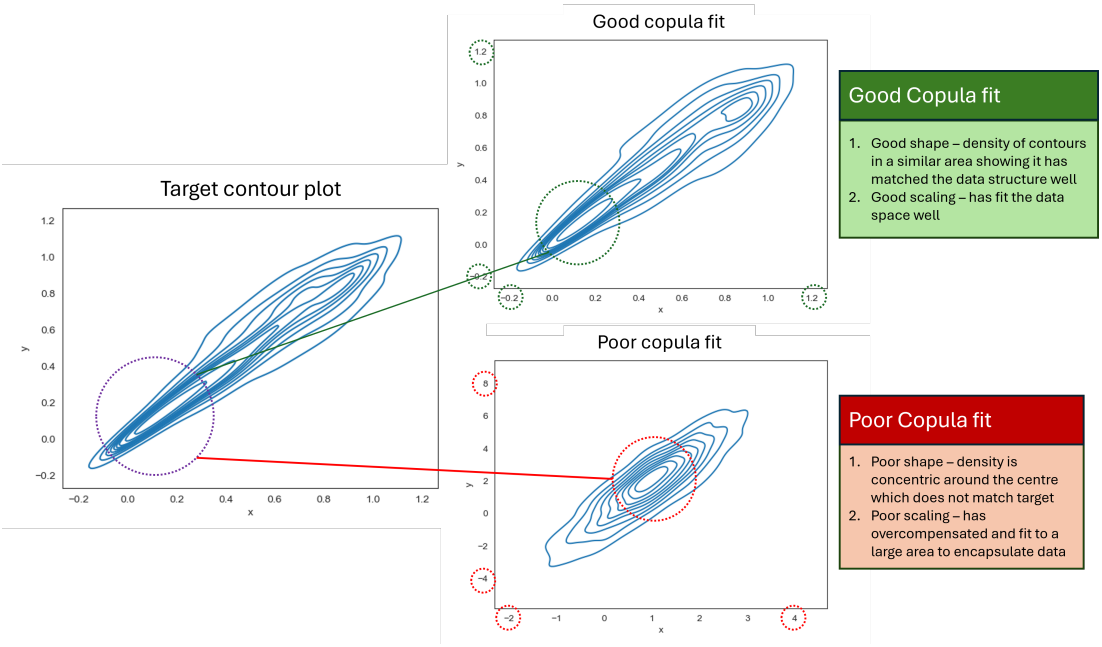

Figure 2.4: Example contour plots to visually inspect copula fitting. An example is given of a good and bad fit compared to the target data contour plot.

While limited to two dimensions, bivariate copulas can be conditioned on other variables of interest which has seen application to financial timeseries prediction [126] and in neuroscience where the changing relationship between stimuli and neuron behaviour are captured by copula-based models [124].

As the copula is a joint probability function between two random variables, when the value of one variable is known, the density of the copula can be used to estimate the value and uncertainty in the estimate for the other variable. This provides a valuable predictive tool for both the estimation of unknown marginal values and the uncertainty attached to that estimate. The conditional density is useful to provide this predictive capability, whereby values of known variables can be used to provide a distribution for the potential values of unknown variables. The conditional density can be described as follows [129]:

$$f(x_N|x_1,...,x_{N-1}) = f_N(x_N)\frac{c(F_1,...,F_N)}{c(F_1,...,F_{N-1})} \tag{2.2}$$

Where $f_i(x_i)$ is the density of F_i, and $c(F_1,...,F_i)$ is a copula density defined by the

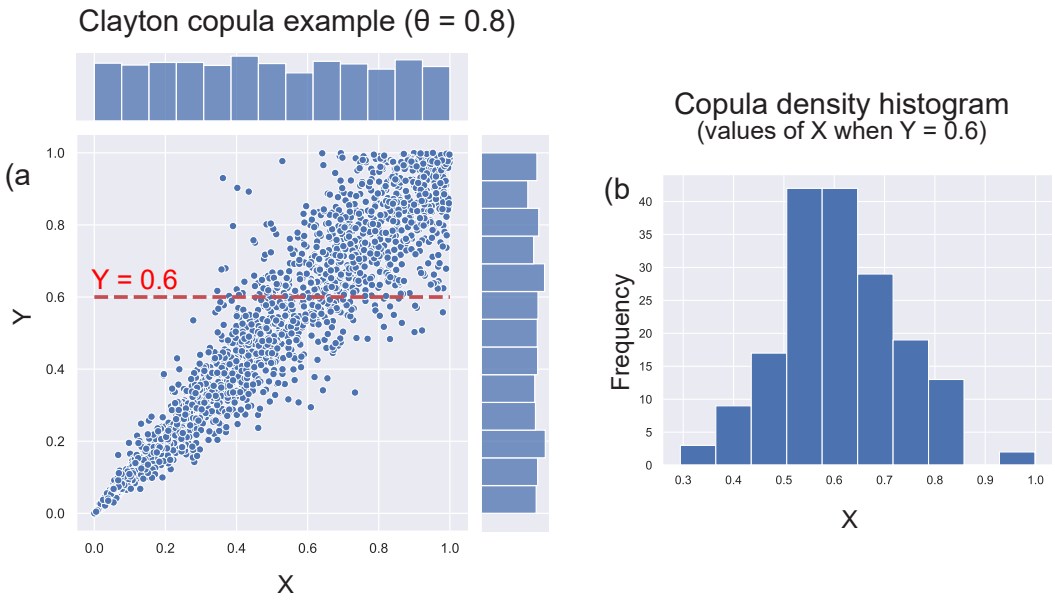

Figure 2.5: (a) Example showing a Clayton copula with its uniform marginals, and (b) how the copula density may be used to predict the range of plausible values of an unknown variable (X) when the value of the other variable is known (Y = 0.6).

derivative of the copula, C. Figure 2.5 shows an example of this process, where the value of Y is known to be 0.6 and the copula density provides probable values of X. For example, at Y = 0.6, X is likely between [0.3,1] with the most probable value between [0.5,0.6]. The cumulative distribution function is used to find the expected value (most probable value), and the value of the 5th and 95th percentile. This provides a prediction, and upper and lower prediction bounds for uncertainty quantification. An example of this is shown in Figure 2.6 which builds on part (b) of Figure 2.5 to get a prediction and upper and lower prediction bounds from the copula conditional density.

Copulas were applied in [130] to quantify uncertainty in rainfall measurement data, and shown by [131] that they are able to implement uncertainty propagation as described by the theoretically rigorous "Guide to the Expression of Uncertainty in Measurement". In this literature review, the copulas are fit to general variables, X and Y. In practice these are generally multivariate data. In this thesis, they are used for lagged spatial and temporal data. In Chapter 4, the variables are time-lagged residuals from a base model prediction for a temperature forecasting scenario. While in Chapter 5, the variables are prediction residuals from a base model on neighbouring points for

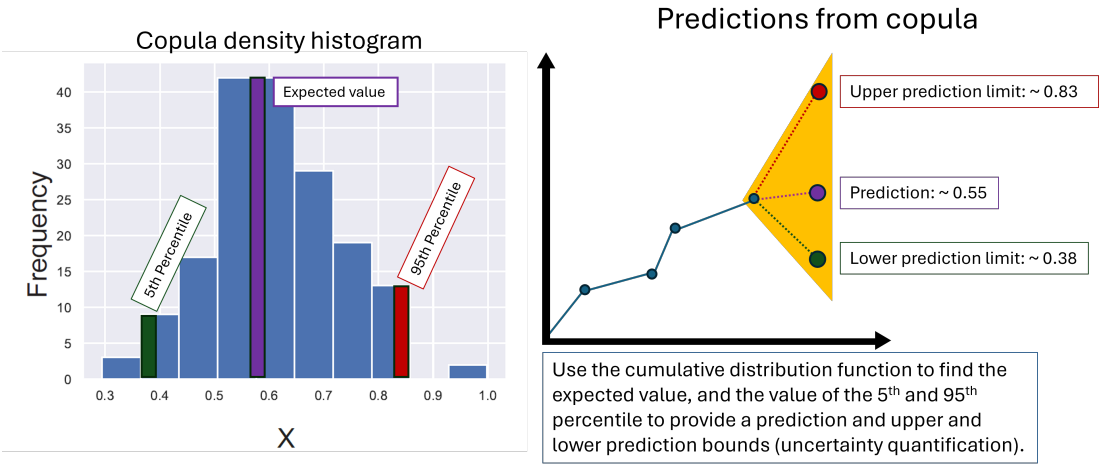

Figure 2.6: Predictions and uncertainty quantification from copula conditional density (building on (b) of Figure 2.5)

the correction of a structure surface reconstruction application.

Multivariate Gaussian copulas

Copula models have been extended to facilitate their application to higher dimensional data in several ways, two of which are the Multivariate Gaussian copula, and vine copulas. Multivariate Gaussian copula, C, is given by:

$$C(u_1, ..., u_N; R) = \Phi(\phi^{-1}(u_1), ..., \phi^{-1}(u_N); R) \tag{2.3}$$

Where u_i are uniform univariate marginals for $i = 1, ...N$, which can be fitted using common families of parametric distributions such as Gaussian, Gamma or Beta distributions; Φ is Multivariate Gaussian cumulative distribution function (CDF); ϕ^{-1} is the inverse Gaussian CDF, and R is a $N \times N$ correlation matrix between the marginal variables. The correlation matrix in the Multivariate Gaussian copula captures the dependencies between all variables which alleviates the dimensional limitations from bivariate copula models, however, the Gaussian base copula is limited to capturing elliptical behaviour which may not always be the most appropriate choice for a given application.

Vine copulas

Regular Vines are graphical models constructed of interconnected tree structures created with nodes and edges. The graphical structure utilises bivariate and conditioned bivariate copulas [124] to capture the dependency between all variables in a pairwise manner. By linking smaller graphical networks (trees) together, complex dependencies can be modelled using a variety of well-known copula families able to capture diverse behaviour. Centre Vines are a subset of Regular Vines where each tree is created by branching from a single, central, node. An example factorization of the joint dependency for three random variables for a Regular Vine is given in [132] where the resulting dependency, $f(x_1, x_2, x_3)$, is:

$$\begin{aligned} f(x_1, x_2, x_3) = & f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \\ & \cdot C_{1,2}\{F_1(x_1), F_2(x_2)\} \\ & \cdot C_{2,3}\{F_2(x_2), F_3(x_3)\} \\ & \cdot C_{1,3|2}\{F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)\} \end{aligned} \tag{2.4}$$

Where $f_i(x_i)$ is the marginal probability density for $i = 1, 2, 3$, where x_i is arbitrarily distributed, and C are the bivariate or conditioned bivariate copulas between the subscripted variables. Vines provide the advantage of allowing different families of bivariate copulas to be used to best fit the dependency between nodes, however, some assumptions are often present to simplify tree construction which can result in important conditional relationships being ignored [133]. Vine copulas have been applied for high dimensional financial timeseries forecasting [133] and applied to spatio-temporal data for streamflow prediction in earth science applications [134].

Hierarchical modelling

Hierarchical modelling structures involve the linking of the outputs of one model to the inputs of another, facilitating the chaining of multiple specialist models and for the calibration of model outputs to improve predictions. In Bull et al [135], a population model is created using a hierarchical Bayesian approach for windfarm and truck

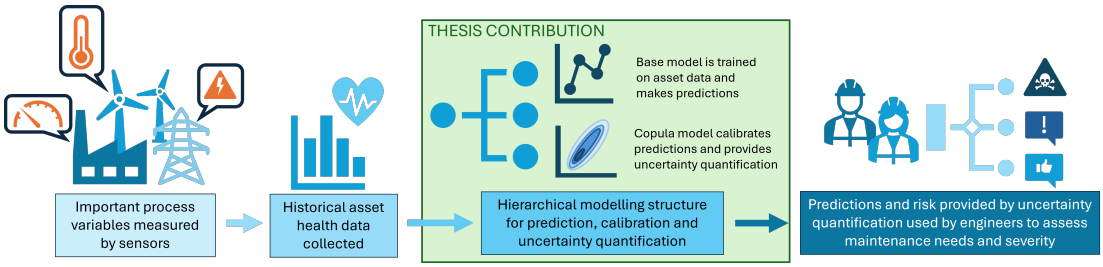

Figure 2.7: Simplified diagram showing the thesis modelling contribution in an engineering context: from sensor measurement; historical data collection; modelling; and utilising model outputs to inform maintenance decisions.

fleet monitoring applications. Predictions were improved through the sharing of data and model parameterisation with similar assets in subgroups within the hierarchical structure, which benefited from the transfer of data between data-rich and data-sparse assets. The calibration of outputs from a hybrid (physical and statistical) solar forecasting model in Schulz et al [136] was found to provide improved performance at 48 hour forecasting horizons which is important for estimating solar generation for integration with the electrical grid.

Copulas can also provided calibration and uncertainty quantification in hierarchical modelling structures. In Stephen et al [137], Multivariate Gaussian copulas formed part of a hierarchical modelling approach that provided a means of capturing the dependency between model residuals to calibrate and provide uncertainty quantification for low voltage load forecasting. A similar structure was used by Möller et al [125] for weather forecasting, where the Multivariate Gaussian utilised the dependency structure between weather variables to calibrate the upstream model predictions. Multivariate Gaussian copulas and Centre Vine copulas are applied in a hierarchical modelling structure in Chapter 4 and Chapter 5 of this thesis. The copulas are used to calibrate and provide uncertainty quantification for the predictions of a simple base model. This base model is an ordinary least squares model for a temperature forecasting application in Chapter 4 and a polynomial regression model in Chapter 5.

The modelling process is shown with engineering context in a simplified diagram in Figure 2.7, showing where the analytics are applied, the hierarchical structure and how the outputs would be utilised across a diverse range of industrial settings.

2.3 Trustworthy analytics and uncertainty

2.3.1 Robustness and transparency

Due to the prevalence and novelty of many artificial intelligence (AI) technologies being applied across different industries, there is an ongoing conversation between policy makers, businesses and academics regarding consensus on definitions, safety and legal considerations when applying 'AI'. From a policy stand point, emerging AI technologies present many economic and societal opportunities and threats, which in turn require regulation through legislation. Policy makers must be aware of novel technologies emerging from academic or business spaces, and account for the potential breadth forms of new AI technologies can take. For businesses, AI offers new ways to improve productivity, development of consumer bases and forms of client or customer interaction, which can prove profitable in many scenarios. Businesses are concerned with adhering to laws and regulations, protecting their IP and reputation, as a failure to do so will severely impact their finances. For academic spaces, researchers aim to innovate and develop the state of the art to provide useful outputs and expertise for their partners and remain competitive in such a fast moving field. As such, many stakeholders in the field of AI are racing to provide clear, encompassing definitions which are able to match both existing and developing research. This has led to a lack of true consensus on certain definitions which are applicable to this thesis. While this thesis makes use of machine learning rather than AI technologies, there is overlap in the concerns around the use and reliance on new AI technologies which can be applied to the deployment of data analytics.

Where definitions have been proposed, an umbrella term is usually used to cover a series of more specific sub-terms, of which the number is also not yet agreed upon. For example, **trustworthy AI** (EU Commission [138], UK House of Lords [139], White

House OSTP [140], MIT Technology Review [141], ISO[4], OECD[5], NIST[6], IBM[7], and Deloitte US[8], etc.); **ethical AI** (Floridi et al [142], and UNESCO[9], etc.); and **Responsible AI** (UK House of Lords [139], University of Montreal [143], Ministry of Science and Technology of the People's Republic of China [144], NIST, Google[10], etc) are possible umbrella terms being used by various government bodies, academic institutes and businesses. Thiebes et al [145] published a review paper on new definitions proposed by different parties, from Government organisations to individual researchers, and summarised their findings into: Trust; Beneficence; Non Maleficence; Autonomy; Justice; and Explicability. However, each of these terms or ideas were not always present within each piece of literature that were reviewed. For example, it was found that 'trust' was interpreted in different ways:

- Trust is Lawful, Ethical and Robust [138]
- Trust is Performance, Purpose and Process [146]
- Trust is Functionality, Helpfulness and Reliability/Predictability [147, 148]
- Trust is Competence/Ability, Benevolence and Integrity [149, 150]

The commonality between all proposed versions are the desire to build reliable AI technologies that sustainably benefit and protect human society, globally. Individual parties generally tend to select which definition to use based on which terms align well with their individual interests or prevent complications from overlapping jargon. The chosen themes, in relation to this thesis, are considered most aptly described by the definition of 'Trustworthy AI' and associated principles proposed by the European

[4]"Towards a trustworthy AI", International Organization for Standardization (ISO), https://www.iso.org/standard/77608.html

[5]"Policies, data and analysis for trustworthy artificial intelligence", Organisation for Economic Co-operation and Development (OECD), https://oecd.ai/en/ai-principles

[6]"Trustworthy and Responsible AI", National Institute of Science and Technology (NIST), https://www.nist.gov/trustworthy-and-responsible-ai

[7]"Trustworthy AI", IBM, https://research.ibm.com/topics/trustworthy-ai

[8]"Trustworthy AI™: Bridging the ethics gap surrounding AI", Deloitte US, https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html

[9]"Global AI Ethics and Governance Observatory", UNESCO, https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

[10]"Responsible AI practices", Google, https://ai.google/responsibility/responsible-ai-practices/

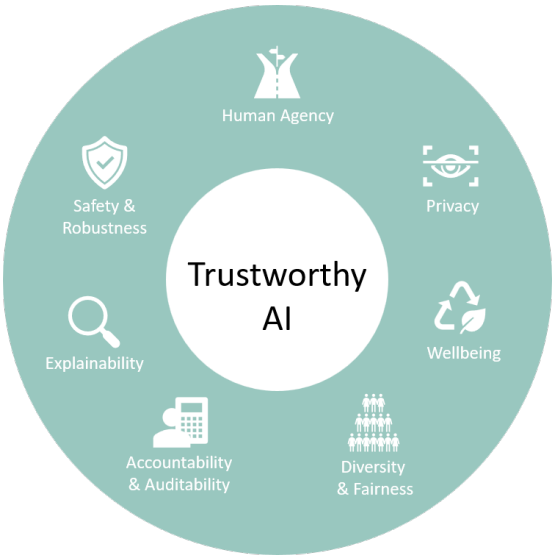

Figure 2.8: Wheel of trustworthiness principles as described by the European Commission's High Level Expert Group on Artificial Intelligence [138].

Commission, based on: relevancy (up to date with new technologies); diversity and coverage of themes, and associated terminology; from a local (geographically) governing body which would likely influence policy in the United Kingdom and neighbours; and, actionable guidance and checklists to provide references for AI users.

The European Commission guidelines are summarised in Figure 2.8, with their definitions summarised as follows:

- **Human agency**: Concerned with ensuring humans maintain their rights, freedom of choice and decision making ability, and ability to oversee AI technologies.
- **Safety and robustness**: Concerned with resilience to attacks, safeguards in the face of failure, accuracy, reliability and reproducibility.
- **Accountability/Auditability**: Minimisation of, and reporting of negative impact, trade-offs and redress, to manage who is responsible for actions made on AI outputs.
- **Privacy and data governance**: Respect for privacy and data access, quality and integrity.

- **Diversity and fairness**: Prevent discrimination and bias, ensure it is accessible, universal and includes stakeholders.
- **Wellbeing**: Sustainable, environmental and social societal wellbeing.
- **Transparency**: Traceability, explainability and communication.

Of particular relevance to this thesis is the terms covered by robustness, essentially ensuring data analytics are able to do the task they are designed for, do it well, and do it consistently. Of additional interest is the explainability, traceability and communication of analytics, which is covered under the transparency theme. This would allow the outputs of analytics to be interpreted by a human user based on an intuitive relationship between the model inputs and outputs (such as with linear regression models), or for this process to be explained to users through some additional means.

2.3.2 Trustworthy AI applications

Scenarios which fall under the key themes present in Figure 2.8 span across many technical disciplines which impact society in low impact (e.g movie recommender systems) and high impact (e.g healthcare diagnosis) ways. Despite the diversity of disciplines, most research goals align on the developmental outcomes being worked towards. For example, in safety and robustness, machine learning and AI models are required to be robust to small perturbations, robust against adversarial attacks, and also ensure that the negative impact of such perturbations (malicious or otherwise) are minimised. This can be shown in work by Eykholt et al [151] who propose a testing methodology designed to generate highly impactful adversarial examples for deep neural networks which, in their study, involves road signs. Deep neural networks are frequently used in computer vision technologies which are the basis of self-driving vehicle technologies - a highly safety critical application. For diversity and fairness, models must avoid any bias and discrimination against particular groups or protected characteristics, which may be introduced via the data, algorithm or evaluation methodology. Open source tools, such as FairTest[11] and FairLearn[12], are being released to provide ML developers access

[11] FairTest [152], https://github.com/columbia/fairtest

[12] "Improve fairness of AI systems", FairLearn, https://fairlearn.org/

to 'fair' evaluation criterion to provide insight into the presence of bias in their models. This has been observed in medical diagnosis where seemingly highly performing classifiers underperform for certain patient groups [152]. Menon et al [153] investigated the trade-offs and implications between model accuracy and fairness from a modelling perspective, with an example given on gender-based assumptions and their potential impact on loan applications and model bias. For transparency, there is a drive to ensure the outputs of models are explainable to ensure that users are able to understand how decisions made by ML- or AI-based systems are formulated [154]. This is a very active area of research in deep learning as there is a need to explain how 'black box' models derive their outputs, such as in sentiment analysis in natural language processing systems [155, 156]. Privacy principles are focused on the security and protection of sensitive data, which can be important when training ML tools which may require access to sensitive data from unrelated providers. This is being addressed through modelling methodologies such as federated learning which allow participants to provide training data to a global model while data is kept locally and separate [157]. This methodology can be applied across all types of service providers where there are data silos that cannot be openly aggregated and shared; from healthcare institutes, online shopping recommendation systems or for the development of urban and transport planning in smart cities [158]. The final two themes are accountability and human-agency which describe how we govern AI and take responsibility for utilising their outputs in our decision making processes. These principles rely on a human's ability to effectively engage, oversee and interface with AI technologies, perhaps through human-in-the-loop [159], and for there to be consistent auditing of utilised technologies [160].

2.3.3 Robustness and transparency tools

'Trustworthiness' tools are available for most stages in the data analysis pipeline, which can most generally be described as pre-model (data preparation or collection stage), model, and post-model stages (performance metrics, model decisions and outputs) shown in Figure 2.9. Due to the relevance of robustness and to some extent explainability principles to this work, this will be further focused on. In the pre-model

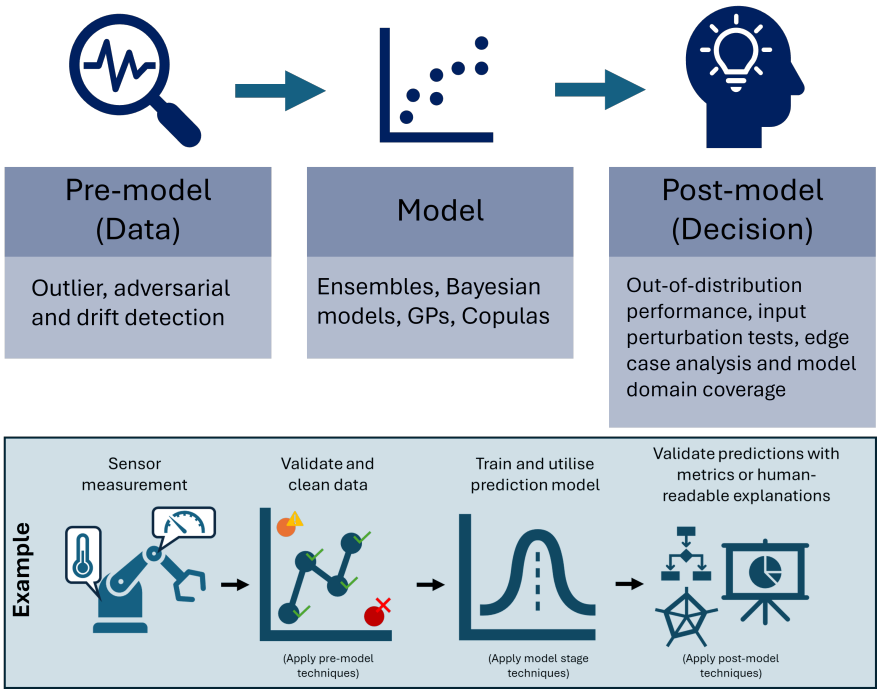

Figure 2.9: Examples of robustness tools and evaluation techniques in the pre-model, model and post-model stages.

stage, open source toolboxes exist to assess incoming data for any domain shift or model weaknesses to certain types or values of inputs. For example, GreatExpectations[13] is a toolbox which specialises in ensuring if the input data is of the required form and quality before it is passed further through the data pipeline. Evidently[14] is a toolbox for pre-model and model stage which is designed to provide sentiment and toxicity monitoring for large language models; data quality evaluations through tests for missing data, data correlations and new categories in the data; data drift which implies changing data distributions; and traditional model evaluation with automated testing and reporting for model development. Other toolboxes, such as Alibi Detect[15], AdvBox[16] and Adversarial Robustness Toolbox[17] are designed to generate, detect and improve protection from adversarial examples, which can be used maliciously to attack

[13] GreatExpectations, GX OSS, `https://github.com/great-expectations/great_expectations`
[14] Evidently, https://github.com/evidentlyai/evidently
[15] Alibi Detect, `https://github.com/SeldonIO/alibi-detect#adversarial-detection`
[16] AdvBox, https://github.com/advboxes/AdvBox
[17] Adversarial Robustness Toolbox, https://adversarial-robustness-toolbox.org/

ML and AI models to produce incorrect outputs. This can be particularly serious where ML or AI tools are used within control loops which can cause purposeful equipment malfunction and risk personnel safety [161].

At the model stage, there are several options to improve robustness which ties with uncertainty quantification. Capturing uncertainty in ML outputs provides a level of risk to that output, which can alert developers to model deficits and allow ML users more flexibility in how much to trust given outputs. Uncertainty is often characterised into two broad categories, Type A (aleatoric) [162] where the uncertainties are driven by randomness, or Type B (epistemic) [163] where the uncertainties are driven by lack of knowledge and so can feasibly be driven down through improved understanding or measurement [164]. Examples of this in a data acquisition system could be sensor noise (aleatoric), where further measurement can reduce the uncertainty to a certain extent by averaging out random effects, versus increased fault observations (epistemic) where more examples would provide a ML model with more characteristic information [165]. Ensembles involve training a collection of models on different data subsets [166] or different model types [167] (or model hyper parameter initialisations [168]) to create a prediction distribution to capture the uncertainty in the output. In such cases, the individual models tend not to be able to attribute uncertainty to outputs on their own (hence the use of the ensemble method). For models such as Gaussian Processes [169], Bayesian modelling [170] and copulas [127], uncertainty quantification is an inherent feature. The models are able to propagate uncertainty and create uncertainty bounds on predictions which are intuitive, for example, the models are more confident close to observed data points and less confident further from these points [117]. Lastly, for monitoring models 'post-hoc', there are a variety of explainability and monitoring tools available, such as Dalex[18], SHAP[19] and Efemarai[20] which provide methods to maintain and update deployed models, test for edge cases and domain coverage to understand model robustness, and explain model outputs through perturbation tests.

[18]Dalex, https://dalex.drwhy.ai/
[19]SHAP, https://shap.readthedocs.io/en/latest/index.html
[20]Efemarai Continuum, https://www.efemarai.com/

2.4 Implications for trustworthy analytics for condition monitoring in nuclear plants

Coolant pumps, steam generators, reactor buildings and sensing systems are a few components within core sub-systems in a nuclear plant that require maintenance intervention to ensure reliability of the full plant. Condition monitoring of these assets within a PHM or SHM maintenance framework supports high performance of sub-systems while maintaining personnel and asset safety. Understanding the health condition of key assets in the plant allows for further lifetime extension of the plant so that current stations may continue production. Understanding the degredation modes of monitored assets allows for more flexible and dynamic maintenance scheduling which is a more affordable alternative to reactive strategies which require failures to occur before action is taken. Predictive maintenance supports better financial management in nuclear plant operation and maintenance budgeting, but only if the fault prediction or diagnosis tools are proven to be reliable. Prognostic tools rely on data from sensor systems which are faced with a number of barriers in the nuclear industry. The harsh environments some sensors are exposed to, such as high temperatures or radiation, can degrade the performance and quality of the collected data. The costs involved with retrofitting these systems into legacy plants and their upkeep or replacement can also be prohibitive. Fault diagnosis or diagnostics tools incorporated into maintenance planning and risk management must adhere to strict regulations to ensure traceability and accountability can be taken for decisions incorporating suggestions from these tools. All of these barriers make the adoption of prognostic tools slow in the nuclear sector compared to other, less restricted industries. To ensure adopted tools have high reliability and can adhere to risk management requirements, trustworthiness principles can be adopted. Adopting modelling practices with inherent uncertainty quantification or transparency can ensure users of the tools can justify the model outputs with an appropriate confidence attached. Model calibration, hybrid models or hierarchical modelling structures can build on the strengths of individual models to improve the robustness of the prognostic tool. Additionally, understanding sources of uncertainty from the sensor system which

contribute data to these models can alert to data quality issues which may impact the performance of prognostic tools. The use of 'trustworthy' fault diagnostic or prognostic tools can ease their integration into the nuclear sector.

Chapter 3

Uncertainty in design: Transferring quantified uncertainty in data pipeline design to new systems

3.1 Explainable, transferable data pipeline design for improved analytics performance and fleet-wide monitoring

Most often, asset maintenance is conducted reactively, where-by corrective maintenance is conducted once a failure has occurred [171]. In power plants, an unexpected outage of an asset can be expensive due to lost revenue from interrupted generation with downtimes being potentially lengthened by the requirement to: retrospectively identify the root of the fault, source required components and perform the maintenance action. With many nuclear power plants (NPP) coming to the end of their designed lifetime, many operators are utilising condition monitoring (CM) and condition based maintenance (CBM) techniques to justify and manage NPP lifetime extensions and to avoid unplanned outages [23]. This requires aging assets to be closely monitored to estimate

asset health and ensure extension plans are affordable.

A common asset in power plants are rotating plant (e.g. motors, turbines, centrifugal pumps, fans), which are prone to bearing failure [172]. These could be turbine or motor driven pumps which form part of a larger generation or cooling system. Despite being relatively simple components, bearings are largely responsible for the reliable operation of rotating plant by supporting huge loads to reduce friction on downstream components [173]. As such, if bearing faults are left untreated, damage could propagate through the drivetrain and create wider system complications in more expensive components, such as the gear box [174]. Cascading failures would lead to expensive and lengthy maintenance intervention which would cause disruption to plant generation and incur additional regulatory reporting overhead.

CBM and data based analytics can be used to estimate the RUL of rotating plant bearings which, if effective, can provide sufficient warning of an impending failure, with diagnostic tools providing an indication of the type of failure developing. An operator can incorporate this into their resource scheduling and budgeting actions to ensure the asset is taken offline and serviced while minimising disruption to plant operation. However, developing and applying this approach requires access to data. Systems within NPP's were designed before modern digital sensing and monitoring techniques were available and capable of operating within the hostile environments they may be installed in, which is an additional consideration that can impact upon the associated data acquisition components. This can result in operators making decisions on unhealthy or unstructured data collected from NPP's which are not ideally designed for modern sensing systems, adding additional uncertainty to maintenance plans or processes.

Sources of uncertainty can impact the data acquisition pipeline at every stage, including: the choice of sensor type and placement; the chosen sampling rate; data pre-processing steps to present the data in a specific format; and, the metric(s) used by the analytics to convey information to an operator. Design choices at each of these stages offer a trade off, which will incur uncertainty in the output of the pipeline at each stage and can be compounded by the interaction between upstream and downstream

pipeline stages. In addition to this, data based analytics generally do not attribute a measure of confidence in their output, making it difficult to determine if the analytics are performing poorly in a sub-optimal pipeline. This makes ML outputs difficult to trust for inclusion in risk and cost assessments. They also do not provide the operator with relevant information that could allow future improvements to the pipeline to be made.

One method to quantify and account for operational uncertainty is calibrated hybrid models, employing physics, knowledge or data driven methods to improve model accuracy and robustness. Hybrid models allow known physical relations to offset full reliance on potentially untrustworthy data, whilst reducing the need for an abundance of representative historical data to reliably identify the monitored asset's underlying behavioural trends. Calibration of the model then ensures the model is updated and representative of the real monitored asset by accounting for differences between the physics or knowledge model and CM data.

In Section 3.4, an open-source bearing knowledge informed machine learning (ML) model and CM datasets are utilized in an illustrative bearing prognostic application. The uncertainty incurred by the decisions made at key stages in the development of the model's data acquisition and processing pipeline is assessed and demonstrated by the resultant impact on RUL prediction performance. It is shown that design decisions could result in multiple valid pipeline designs which generated different predicted RUL trajectories, increasing the uncertainty in the model output.

This analysis is extended in Section 3.5, to the explanation of how the design impacts analytics, allowing an operator to make comparative and informed decisions on the selection of pipeline features. To achieve this, a SHAP-based human-readable explainable AI (XAI) framework was used to rank and explain the impact of each choice in a data pipeline on the analytics, allowing the decoupling of positive and negative performance drivers. The explanations of a fully-observed asset facilitate the successful selection of highly-performing pipelines. In Section 3.6 this operational insight is then leveraged to utilise knowledge gained from the fully-observed data pipeline to a similar, under-observed case. The transfer of uncertainties can provide insight into

uncertainty drivers across a fleet of similar assets without repeating the computational or cost overhead of fully redesigning the pipeline for the new systems. This empirical approach presented across Section 3.5 and Section 3.6 is demonstrated on bearing fault classification case studies, using well-understood open-source data.

3.1.1 Contribution and novelty

This chapter is covered by two publications. The first case study investigating the impact of pipeline design on hybrid model performance was published and presented at PHM Society European Conference 2022 [12]. The second case study is published in Data-Centric Engineering [13].

The contribution of the first part of this chapter is not the creation of a novel RUL technique, but to demonstrate and quantify the confidence associated with the application of existing hybrid RUL approaches with the associated data acquisition pipeline decisions. Confidence can be undermined by these choices, which impact the performance of the underpinning model and can reduce the operators trust in the whole decision support system. Without sufficient trust, especially in the heavily regulated nuclear engineering environment, decision support tools will not be utilised to support maintenance scheduling activities. As such, the methodology presented in Case Study 1 is concerned with investigating the uncertainty in analytic design and deployment by capturing the sources of uncertainty and demonstrating how these impact on an uncertainty budget for the whole data to decision pipeline rather than just the output of the ML model. The uncertainty in the model performance due to the whole pipeline design is captured by analysing the quantiles of the model outputs under different data acquisition pipeline designs. To evaluate data pipeline uncertainty, evidence is presented from open-source, curated test rig datasets (to reduce the impact of excessive operational noise).

The motivation for the second part of this Chapter in Case Study 2A and 2B is to demonstrate how the uncertainty associated with data pipeline design choices can be identified, quantified (in terms of the choice's contribution to analytic performance) and leveraged as transferable knowledge when designing pipelines for similar engineering

applications.

This contribution can be summarised in three ways:

- Demonstrating how the uncertainty from the imposed pipeline design constraints can compound to create improved or deteriorated performance in fault diagnostic systems.
- Identifying highly or under-performing system design options using a human-readable XAI framework, leading to better or worse system performance, respectively.
- Identification of uncertainty sources learned from a fully-observed pipeline system design (source) to an unseen pipeline (target). This allows insight into the target system's pipeline without the computational overhead of fully observing all possible pipelines or fully re-designing the new pipeline.

The first case study lays the groundwork for the first contribution, while the second case study addresses all three.

3.2 Literature: Bearing prognostics, explainability tools and transfer learning

The literature review covers research trends in bearing prognostics applications, which are the key engineering problem covered by both case studies. This is followed by a discussion on explainable AI which forms a central part of how the impact of the pipeline design impacts the model outputs in Case Study 2; and finally, transfer learning which is how the information gained in Case Study 2A is leveraged to another, under-observed system in Case Study 2B.

3.2.1 Fault classification and Remaining Useful Life prognostics for bearings

Bearings are subject to high stress operating conditions which makes failures common. These can manifest due to overloading, imbalanced loading, or lubrication issues due

to insufficient lubrication, contamination or sealing failures. Bearings are mechanical faults and mechanical failures are most commonly monitored via vibration monitoring, although have been approached using temperature, oil analysis and accoustic emission approaches [175]. Vibration monitoring, while subjected to the robustness and cost of the sensor system, allows changes in bearing health to be observed immediately and has been proven as a reliable method for bearing fault prognosis. Temperature based schemes are most useful for end of life where the fault has progressed significantly, oil analysis methods require the bearings to have a dedicated supply system and acoustic emission requires access to high quality measurements [173].

3.2.2 Explainable AI

Explainable AI (XAI) tools are being adopted into industrial fault diagnosis systems as a way of improving the transparency of ML outputs for maintenance applications [176]. XAI tools can be integrated into a ML model pipeline at several stages, and used to provide different levels of explanations. Tools such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding, can be applied pre-model where they provide insight into the structure of the collected data before any model is selected for training [177], but inherently cannot explain how a chosen model generates predictions from said data. Gaining interpretability during the model stage can be achieved by using inherently transparent models where the underlying decision processes are well understood, as with linear regression, generalized additive models or decision trees [178], however this limits the selection of models which can be utilised.

The most flexible approach is applying 'post-hoc' techniques where the decision making process for a trained model is explained after training and producing predictions. This includes techniques such as gradient-based methods (such as saliency maps [179] which can visually convey what the model has identified as the most important aspects of image-based input data), surrogate modelling (such as in [180], where an inherently transparent model is used to approximate the behaviour of the model of interest) or pertubation-based methods (such as Shapley Additive exPlanations (SHAP) [181]). SHAP is based on Shapley values [182], a coalition game theory

method, which computes the contribution of model input features to the obtained output to provide insight into the importance of each feature. This importance can be provided on a local or global scale, where global explanations provide an overview of the models learned relationship between all input variable instances and resulting output predictions, and local explanations provide insight into the contributions of one observation or region in the model space [177].

Due to its flexibility in explanation levels and robust theoretical grounding, SHAP is becoming an increasingly popular and widely adopted XAI approach [183]. SHAP is used in [184], where a human-readable XAI framework utilises SHAP to explain the outputs of 3 regression models predicting: power output in a combined cycle gas plant; gearbox vibration; and, bearing wear in feed-water pumps. The human-readable outputs are generated through encoding text explanations of SHAP outputs to aid non-ML experts in engaging with their predictive tools in a more intuitive manner. In [176] SHAP is used then compared with Local Depth-based Feature Importance for Isolation Forest (Local-DIFFI) to explain feature importance of machine learning based fault diagnosis of rotating machinary. More generally [185] investigates the application of XAI techniques when using deep neural nets in predictive maintenance within the field of aerospace integrated vehicle health management.

The literature has a general consensus that XAI techniques can contribute towards establishing trust in automated machine learning processes. However, the literature also denotes limitations in state-of-the-art XAI techniques related to the complexity and human biases inherent in the XAI models themselves, in addition to XAI end-users' ability to understand the explanations provided. However, these limitations are not the focus of this work, where SHAP, due to its robust theoretical foundation, is used to explain the outputs of a decision tree trained to identify which features (pipeline design options) lead to the success or failure of fault detection and diagnostic systems, as measured through the model classification error.

3.2.3 Transfer Learning

Industrial fault diagnosis has evolved through several stages. Initially, fault diagnosis relied entirely on expert knowledge until data collection and processing became more widely available. Access to historical, labelled data allowed the development and resultant popularity of Machine Learning-based, and then Deep Learning-based (DL) approaches as Big Data became possible. Both ML and DL approaches are heavily impacted by the availability of data and labels, and incur high computational overheads when retraining on new datasets, pushing the need for transfer learning (TL). Transfer learning can allow available data or pre-trained models to be adapted to a new compatible task or application [186].

The lack of sufficient labelled data is a common problem in supervised learning. However, the utilisation of knowledge gained from available (source domain) data can often improve model performance in a situation where the new application (target domain) may lack sufficient labelled data [187]. TL can be categorized in many ways: from considering this lack of sufficient labelled data to also considering the similarities between data domain or modelling tasks [188]. TL is usually defined by three high-level categories: inductive; transductive; and, unsupervised, with some authors [188] including a fourth to consider negative transfer. Inductive TL covers scenarios where a degree of labelling is available in both domains or where tasks and data modalities are the same. Transductive TL is where labelled data is only available from the source domain or where the tasks are different but relevant to each other. Unsupervised TL is where there is no labelled data from either domains or the domains and tasks are different. Lastly, negative transfer considers when the transfer of knowledge hinders the performance of the model in the target domain. TL can be further broken down into the types of learning conducted and whether the transfer of knowledge focuses solely on data [187], such as learning on instances. Learning on instances applies where there is access to a small sample of labelled data in the target domain, with relevant data able to be adapted from the source domain [188], i.e. is a sub-category of inductive TL. Additionally, further sub-categorises for TL can be based upon data heterogeneity [189]. Finally, in [190], a different approach to categorisation is presented where four industrial

transfer learning concepts are proposed, which focus on the description of the intended application instead of the technique used to conduct the transfer.

As this work focuses on the industrial engineering application of transfer learning, the concept of "cross-entity" transfer learning (from [190]) will be adopted. Cross-entity transfer specifically focuses on the scenario where knowledge is transferred between two similar assets, with similar functionality and faults. Inductive TL relates to this concept as labelled data is available for both source and target assets/domain. However, the labelled data in question varies in terms of different data quantity and data diversity.

Other works which combine TL and uncertainty quantification mainly focus on the uncertainty of available data or model predictions. In [191] the authors use the variance in the outputs of an ensemble of pre-trained, re-weighted convolutional neural networks to calculate epistemic uncertainty in the diagnosis of COVID-19 from medical images. In [192], the authors use uncertainty quantification to identify the most uncertain samples in available industrial elevator usage data while transferring their hybrid digital twin and Neural Network architecture to new elevator usage scenarios. However, in [193], the meta data for simulations used to characterise the design of a previous crash box is used to predict the behaviour of a new, similar, early stage design with limited simulations. The TL in [193] was designed to permit further uncertainty quantification in new designs, but does not perform this analysis or explore how potential designs may be compared.

3.3 Data pipeline stages and uncertainty sources

In this section, the considerations and selection of pipeline stages are discussed. This process is situation dependant, but some examples and justifications for those choices are presented. The methodology used to generate pipeline designs is presented, which creates a new data set which can be analysed in various ways, as presented in the later case studies.

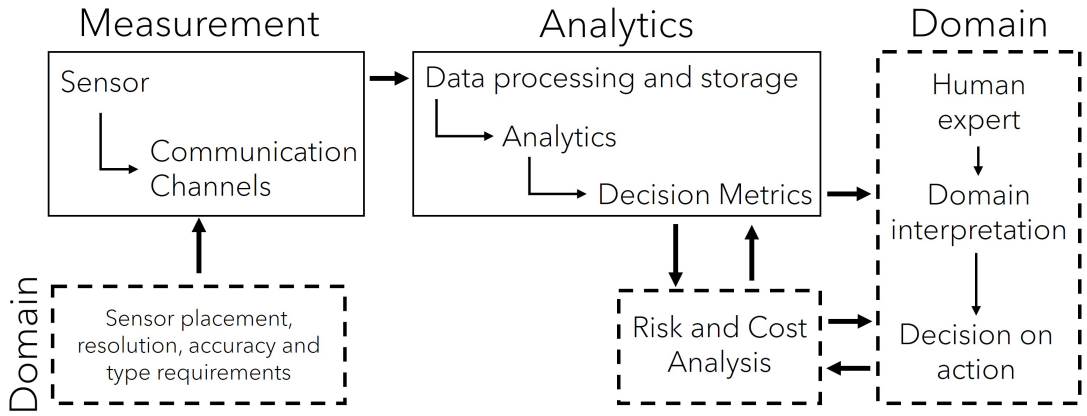

Figure 3.1: Illustration of the major stages and flow of data in a simplified industrial data acquisition pipeline. (Domain expertise sections are shown in boxes with dashed lines.)

3.3.1 Uncertainty quantification in data pipelines

Uncertainty is often characterised into two broad categories, Type A (aleatoric) where the uncertainties are driven by randomness, or Type B (epistemic) where the uncertainties are driven by lack of knowledge and so can feasibly be driven down through improved understanding or measurement [162–164]. Examples of this in a data acquisition system could be sensor noise (aleatoric), where further measurement can reduce the uncertainty to a certain extent by averaging out random effects, versus increased fault observations (epistemic) where more examples would provide a ML model with more characteristic information [165]. The general components of the data pipeline in industrial process condition monitoring, shown in Figure 3.1, are the sensor systems, data communications channels, data processing and storage, data analytics and decision metrics. At each of these stages, uncertainty sources impact the quality of the derived information. Quality issues can arise from sensor calibration issues [194], the method and standards for the data communication channels [195] and possible problems due to data storage and bandwidth limitations [196], all of which increase uncertainty contributions associated with the data. These uncertainty contributions often cannot be accurately corrected for, resulting in a change in the confidence interval on the final output [197]. Additionally, many ML models cannot attribute a confidence margin to an

output decision which makes risk and cost assessments involving the output of machine learning models problematic [198]. This inability to quantify prediction confidence results in either a lack of adoption of machine learning techniques or poorly-informed decisions being made due to the poor quality of the data sources.

Many legacy industrial process plants (such as power plants), even when retrofitted with modern digital data collection and acquisition systems, are likely to have sub-optimal data pipelines which limits the information available for diagnostic and fault detection systems. Hence, the automated design of data pipelines for machine learning applications in data science spaces is an active area of study [199–201]. This has arisen due to the increasing popularity and desire for 'off the shelf' machine learning techniques. [199] proposes 'Auto-sklearn' and improves on automated pipeline design by taking account of historical performance and constructing ensembles evaluated during Bayesian optimisation. [200] builds on ADMM (Alternating Direction Method of Multipliers) optimisation by decomposing the pipeline optimisation problem into easier sub-problems and incorporating constraints on the objective, results in performance improvements. [201] introduces a new greedy design protocol to gather information about a new pipeline dataset efficiently and proposes 'TensorOboe', which uses low rank tensor decomposition as a surrogate for efficient pipeline search. However, these state-of-the-art techniques still require calibration, albeit with less expert intervention. Additionally, these techniques tend to focus solely on delivering the 'best' performing pipeline within a given search space but do not provide the developer with an understanding of *why* the proposed pipelines should be accepted or rejected. An associated explanation of the interactions between engineering sub-systems and how they drive the performance of the overall system is also lacking. The work in this Chapter aims to provide insight into these elements of pipeline design.

To improve the trustworthiness of a data-driven ML system, it is desirable to reduce uncertainty during the data pipeline design phase, as design changes are easiest at this time. During the initial stages of the pipeline design, a developer will be presented with many design choices, each incurring a different performance trade-off. Firstly, inter-stage pipeline uncertainties may compound and propagate to mislead the analytic

in unforeseen ways, preventing a developer from initially identifying under performing choices. Additionally, pipeline design choices may be constrained by the desired system functionality (consider the purpose of a fault diagnostic rather than a fault detection system), and some sources of data may degrade the performance (e.g. due to poor sensor positioning or calibration). Rapid design or automation of design for analytic pipelines can assist in presenting a developer with explanations for the uncertainty sources in possible designs, to help them identify important drivers of system performance. This can identify areas to focus investment to reduce the overall uncertainty, and so risk, in the system. A flowchart of the full approach proposed in this work to support pipeline design is shown later in Figure 3.9, and will be described in subsequent sections.

3.3.2 Identifying key pipeline stages and design options

As identified in [202], the discretisation of a pipeline into stages may be dictated by project, budget or domain expertise of the contributing engineers. Considering the industrial data acquisition pipeline in Figure 3.1 as the basis for a fault detection or diagnosis system, there are several key stages along with the consequences of the resulting design trade-offs. Some stages may be less important or less flexible, necessitating effort to be directed at elevating performance in other stages. Some options to consider are:

Sensors and Process Variable Measurements

This stage is primarily concerned with the fidelity of instrumentation/sensing coverage on the asset and how the associated physical phenomena are measured, taking into account the cost of installing and maintaining the measurement system. For example, a rotating plant asset may be monitored by vibration, temperature, current, oil or acoustic sensors [173], all of which provide different levels of insight into faults of interest. Additional sensing points may provide more information about a developing fault across different axes or, without suitable access to the fault location, may fail to detect meaningful information allowing faults to develop unnoticed, but this may come at an additional cost.

Data collection

This stage covers the trade-off between the running and upfront cost of the sensors against the resolution and sampling rate of the data acquired [195]. A high precision sensor or high sampling rate may allow fault behaviour to be observed in great detail, potentially allowing early fault detection, but will result in the collection of a large amount of non-fault data which must also be handled/stored.

Data transmission

This stage covers the amount of data that can be reliably transmitted by the chosen data transmission system against the cost of increased bandwidth [196]. Even if the measurement system is capable of generating high frequency, high fidelity measurements, the data transmission system may not be able to transmit this with sufficient speed or quality to a centralised storage location. The amount of data that can be transmitted as input to the data pre-processing stage may influence the analytic further downstream by capturing insufficient asset behaviour to perform reliable analysis. However, transmitting too much data associated with relevant asset behaviour may be averaged out by the pre-processing stage, obscuring the fault from the analytic.

Data processing and storage

This stage is mainly concerned with the trade-off between the computational cost of processing and storing data against the quantity and type of useful information preserved within the data [196]. This can include the format the data will be presented to the analytic stage, such as a time series signal being translated to time, time-frequency or frequency domain. Also, the trade-off between the amount of preserved information against the dimensionality of the data during dimensionality reduction procedures should be considered. The data may be amalgamated for long term storage to reduce the time resolution of measurements while keeping the storage requirements low. Furthermore, some data resolution or meta data concerning the data that was collected may be lost during conversion to a designated storage format, impacting its usefulness and trustworthiness. In [203], the high sampling rate time series data used to capture

electrical faults actually carried it's predictive value in the proportion of energy in frequency domain subsets at relatively low resolution. In this instance, this stage of the pipeline would have featured sharply characterised uncertainty across the choices of pre-processing.

Analytics

This stage covers the trade-off between more informative model tasks (such as detailed fault diagnosis) and model family bias against the required model accuracy, model complexity and computational limitations [204]. Different model families identify relationships within provided data by different means, all of which incur specific biases that may only be suitable for limited applications. The computational requirements for model hyperparameter tuning differs depending on the amount of data available and type of model used. A developer may require the model choice to be transparent to allow the decisions made by the model to be explainable. Additionally, the model task can deteriorate the model performance through different ways. Some faults may be more difficult to identify or separate from others, and the labels used to group data may cause dissimilar samples to be grouped which can also deteriorate model performance. However, access to more descriptive fault warnings can provide an operator with more useful information to inform maintenance activities.

Decision metrics

The most informative decision metrics are generally application specific and their selection usually driven by the cost of different types of model failure. False alarms can cause healthy equipment to be taken offline to perform unnecessary maintenance, while missed faults can cause unexpected outages of equipment. Both reduce trust in the fault diagnostic system and can eventually result in the model recommendations being ignored altogether. Metrics like accuracy cover the model's general performance, but can fail to represent the model's skill in identifying different types of failures. The utilisation of different metrics for model selection can lead to different models being favoured, or even unacceptable models being selected if the validation requirements are

not carefully considered [12].

Domain interpretation

The outputs of the ML-based system should be compatible with the linguistic terms and procedures used in the intended application domain to support communication and build trust with the system [205]. A system that can provide different levels of explanation alongside the ML outputs can give the user more agency in engaging with the system, better informing their resultant decisions. This could involve providing access to raw data, generating visualisations or applying XAI tools to enhance explanations of the ML outputs.

3.3.3 Data pipeline construction

Once the relevant pipeline stages and the choices of interest at each of these stages have been identified, pipeline designs observing all combinations of each choice are constructed following Algorithm 1. Note that the stages must be built sequentially in order of the flow of data to ensure the pipeline is built in the required order (i.e. the input sensor data must be chosen before the input data can be pre-processed) and for stages that have a fixed design, the loops in Algorithm 1 will have only one iterative loop. Lastly, the term Error, E, is the validation metric of choice. Algorithm 1 is an exhaustive/greedy search and is needed to provide full observation of the system and the interactions between stages and choices. This permits the prominent and important relationships and dependencies to be uncovered in the model errors. The design, D, and the model metric, E, now form a new set of data to perform meta-analysis on.

3.4 Case Study 1: Impact of pipeline design on data-based and hybrid models

The proposed methodology to assess the uncertainty is presented in the form of a case study that investigates the impact of decisions made in the data acquisition and processing pipeline through the resulting uncertainty in the RUL prediction for motor

Algorithm 1 Pipeline Construction

Variables:
Stage, S(1,N)
Choices, $C((1, C_{S_1}), (1, C_{S_2}), ..., (1, C_{S_N}))$ where C_s is a vector of (1,C_{S_s}), choices for stage, s
Error, E(1,Z) where $Z = \prod_{i=1}^{N} \sum_{j=1}^{C_{S_i}} j$
Design, D(1,Z)
procedure BUILDPIPELINES(S, C)
    % Construct all design combinations
    z = 1
    for choice k, $1 \leftarrow C_{S_1}$ **do**
        for choice m, $1 \leftarrow C_{S_2}$ **do**
            ...
            for choice p, $1 \leftarrow C_{S_N}$ **do**
                % Build pipeline of design D_z
                % Get model error, E_z
                $D_z = (S_1(k), S_2(m), ..., S_N(p))$
                $E_z = f(S_1(k), S_2(m), ..., S_N(p))$
                z += 1
            end for
            ...
        end for
    end for
    return D, E
end procedure

bearing prognostics.

3.4.1 Condition Monitoring Datasets

Two open source bearing prognostics datasets are used in this work: NASA IMS [206] and NASA FEMTO [207]. Both datasets observe run to failure experiments for bearings with no initial defects. Each data set has visibility of the bearings failures by vertically and horizontally mounted accelerometers (termed 'x-axis' and 'y-axis' respectively), with limited access to the vertical data for the IMS dataset. Four distinct bearing failures are observed in the IMS dataset, with two occurring concurrently, while the FEMTO dataset contains 17 run to failure examples. The IMS failures were accelerated due to intensive, but in specification, bearing loading conditions, while the FEMTO dataset was created using the PRONOSTIA test rig which artificially overloaded the bearings to further accelerate wear.

3.4.2 Existing Hybrid Model

Combining Knowledge- and Data-driven Components

An open source hybrid RUL model consisting of a novel Weibull-based loss function for Neural Networks (NN) by [115] was chosen as the basis for this study. Utilising a Weibull distribution to capture domain knowledge from the field of reliability engineering, the authors create 9 NN loss functions to evaluate the success of their knowledge informed ML model for bearing prognosis on the IMS and FEMTO datasets. The knowledge component of the hybrid model is captured by using the data to calibrate the Weibayes equation [208] shown in equation 3.1. The one parameter Weibayes has been shown to produce accurate results for a small number of failures (<20) where the estimated value of shape parameter, β, is representative of the true system behaviour [208]. The value of β was fixed at a value of 2 in [115] due to model stability concerns, and this value being deemed a reasonable shape estimate for ball bearing failures [208]. The values of η and β are used to calculate the Weibull cumulative distribution function (CDF) in equation 3.2. The 9 loss functions are shown in figure 3.2 and are incorporated into the model as the loss function to be minimised by the NN in

Loss Function	Equation
MSE Loss ($\mathcal{L}_{\text{MSE}}$)	$\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2$
RMSE Loss ($\mathcal{L}_{\text{RMSE}}$)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2}$
RMSLE Loss ($\mathcal{L}_{\text{RMSLE}}$)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(t_i + 1) - \log(\hat{t}_i + 1))^2}$
Weibull Only MSE Loss ($\mathcal{L}_{\text{Weibull-MSE}}$)	$\lambda\frac{1}{n}\sum_{i=1}^{n}(F(t_i) - F(\hat{t}_i))^2$
Weibull Only RMSE Loss ($\mathcal{L}_{\text{Weibull-RMSE}}$)	$\lambda\sqrt{\frac{1}{n}\sum_{i=1}^{n}(F(t_i) - F(\hat{t}_i))^2}$
Weibull Only RMSLE Loss ($\mathcal{L}_{\text{Weibull-RMSLE}}$)	$\lambda\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(F(t_i) + 1) - \log(F(\hat{t}_i) + 1))^2}$
Weibull-MSE Combined Loss ($\mathcal{L}_{\text{Weibull-MSE-Comb}}$)	$\mathcal{L}_{\text{MSE}} + \lambda\mathcal{L}_{\text{Weibull-MSE}}$
Weibull-RMSE Loss ($\mathcal{L}_{\text{Weibull-RMSE-Comb}}$)	$\mathcal{L}_{\text{RMSE}} + \lambda\mathcal{L}_{\text{Weibull-RMSE}}$
Weibull-RMSLE Loss ($\mathcal{L}_{\text{Weibull-RMSLE-Comb}}$)	$\mathcal{L}_{\text{RMSLE}} + \lambda\mathcal{L}_{\text{Weibull-RMSLE}}$

Figure 3.2: Loss functions from [115]

the back-propagation step.

$$\eta = \left[\sum_{i=1}^{N}\frac{t_i^{\beta}}{r}\right]^{\frac{1}{\beta}} \tag{3.1}$$

$$F(t) = 1 - e^{-(\frac{t}{\eta})^{\beta}} \tag{3.2}$$

Where

- t = time or cycles,
- r = number of failed units,
- N = total number of failures plus currently running units (incomplete failures)
- η = maximum likelihood estimate of the unit characteristic life (63.2 distribution percentile)
- β = Weibull shape parameter, and

Dataset	Train.	Val.	Test.
IMS	Run 2 (B 1) Run 3 (B 3)	Run 1 (B 3)	Run 1 (B 4)
FEMTO	Bearing1_1 Bearing2_1 Bearing3_1	Bearing1_2 Bearing2_2 Bearing3_2	Bearing1_3 Bearing2_3 Bearing3_3

Table 3.1: Data split between training, validation and testing

Parameter	Selection Choice
Batch size	32, 64, 128, 256, 512
Learning rate	0.1, 0.01, 0.001, 0.0001
Lambda	Floating point number 0-3
Number of layers	Integer between 2 and 7
Number of units per layer	16, 32, 64, 128, 256
Probability of dropout	0.1, 0.2, 0.25, 0.4, 0.5, 0.6

Table 3.2: NN Architecture Hyperparameter Options Table from [115]

- F(t) is the Weibull CDF

RUL Estimation Procedure

The following process was conducted by [115] to generate RUL estimates for both the IMS and FEMTO datasets. First, the input data from the horizontal sensors was processed into spectrograms to obtain the frequency representation of the vibration data. The number of input features was reduced by 'binning' the spectrogram into 20 bins, where the maximum value of the frequency bands included in each bin is taken as the value for that bin, repeated for each timestep. The response variable was the lifetime percentile status of the bearing, with 0 % being healthy bearing at the start of the experiment, to 100 % signifying the failure of the bearing at the end of the experiment. The training, validation and testing split of the datasets are shown in Table 3.1.

The Weibayes equation was calibrated with the training data to be incorporated into the loss functions. To initialise and optimise the NN architecture, a random search was conducted to select from the hyper parameters shown in Table 3.2 for each of the loss functions, which the authors set to 1000 in their study. The coefficient of

Pipeline Stage	Parameter Settings
Dataset	IMS or FEMTO dataset
Sensor channel	Horizontal or Vertical aligned
Subsampling	Lose 1/8, 1/4, 1/2 or no data
Spectrogram Bins	10, 20 or 40 bins
Hyperparam. Opt.	Random search of 10, or 100
Model Choice	NNs or Linear Regression (LR)

Table 3.3: Summary of pipeline stages and parameters

determination (R^2) and Root Mean Squared Error (RMSE) were used to discard models that performed poorly, with models with a $R^2 > 0.2$ and $RMSE < 0.35$ progressed to the testing stage. After testing, the models were filtered again by the R^2 and $RMSE$ bounds before selecting a subset of the top performing models based on the R^2 metric. The authors found that the top performing loss function for the IMS dataset was the Weibull-RMSLE combined, and the Weibull-MSE combined for the FEMTO dataset, both containing the knowledge informed loss function.

3.4.3 Pipeline Design Uncertainty

For this sensitivity study, the data acquisition pipeline design was varied, considering the following stages and settings, also summarised in Table 3.3.

Dataset

The FEMTO and IMS datasets were chosen due to initial bearing states with no faults; their curated, open source nature; but also their differences in aging methods, timescales and number of recorded failures. In the IMS dataset, the bearings are operated under their maximum specified operating condition limits and failed after their design lifetime (in number of revolutions). This represents scenarios where the bearings are operated in an unhealthy but within technical specification manner. However, as it took weeks to months to observe these failures, only 4 failures over 3 runs were observed, severely limiting the analytic's scope to learn from a diverse sample of run-to-failure trajectories. This issue is reversed for the FEMTO dataset, where 17 distinct failures were observed due to the run-to-failure process taking several hours. However, the condi-

tions the bearings were operated in would not be practical in an industrial setting as the operating environment is purposefully designed to damage and wear the bearings down at an escalated rate. Data pipeline choices at this stage investigate the impact on the analytics RUL performance due to the amount and nature of the failures observed, and how the analytics perform on the different methods of accelerated lifetime testing.

Sensor Channel

Both datasets have access to vertically and horizontally aligned vibration sensors (noting limited availability for the IMS dataset). Depending on the nature of the fault, ML models may be more successful in identifying failure signatures in one axis over another, leading to more reliable RUL estimates *if* measurement data is available for this orientation. However, it is not always feasible or maintainable to retrofit assets with extensive sensor coverage, meaning the developing failure may not be measured from the most suitable angle. With no prior knowledge of the bearing failure, data pipeline choices at this stage investigate the consequences on the RUL estimate of having limited, and potentially inadequate, sensor coverage of an impending failure.

Data Sampling

In an ideal scenario, condition monitoring would consist of high resolution, continuous measurement to ensure that as much data is available to the prognostic algorithms as possible. In practice, this would generate enormous volumes of data that would be impractical to transmit, process and store, while potentially providing diminishing returns on the useful information contained in the data streams. Communications and storage infrastructure is limited in an industrial setting where fleets of assets are expected to be monitored simultaneously. At this stage of data pipeline uncertainty assessment, comparisons are made for RUL estimates where 1/8, 1/4, 1/2 and no data is lost due to these constraints.

Spectrogram Bin Count

The spectrogram binning process from [115] allows the frequency domain information from the full spectrogram to be used while condensing this information into a more manageable number of input features to the ML stage. This forms a trade off between the amount of information lost in the binning process, and the dimensionality. The spectrogram bin count is chosen to be 10, 20 (as original author) and 40, to compare how the RUL is impacted by this trade off.

Hyperparameter Optimisation

NNs are computationally expensive to train, and it may be infeasible to evaluate a large selection of models in order to optimise the selected hyperparamters. Selecting a sub-optimal model will impact the quality of the RUL estimate. The original author runs a parameter search by selecting n combinations of model hyperparameters (table 3.2), then filtering out models with unsatisfactory performance. Computational limitations may make training many models to allow the most optimal hyperparameters to be chosen an unfeasible action to take. This stage of the pipeline design process investigates the impact on the RUL estimate when the best 10 models are selected from a random search of 10 (90 unique models based on 10 random hyperparameter initialisations for each of the original authors 9 loss functions) and a random search of 100 (900 unique models),

Model Choice

The original author utilises NNs in their study which are black box and computationally expensive. This can undermine the operators trust in the chosen analytic as outputs can not be explained by the model, increasing the risk associated with incorporating model suggestions into decision making processes. Linear Regression (LR) models reside at the other end of the model complexity spectrum as they are cheap to train and simple to understand. However, NNs are able to tackle complex data problems with complicated underlying relationships which cannot be captured by the LR model. In this stage of the pipeline design process, the chosen models are NNs and LR models to

compare the RUL prediction between computationally expensive, sophisticated models and interpretable, computationally inexpensive models.

Evaluating Uncertainty

To evaluate the effect of uncertainty in data pipeline design based upon the design choices described in Section 3.4.3, the original data for each dataset was processed to remove every 8th, 4th or 2nd data point for every datafile in the dataset and resaved; and this process was repeated for each sensor channel. This ensured all combinations of dataset, data sampling and sensor channel were available to train the models. Each model type was trained on all combinations of dataset, data sampling and sensor channel, with the data preprocessed for each selected bin count. For each of these combinations, the NN model hyperparameters were chosen with a random search of 10 or 100, with the model and metrics saved for later processing. The metrics chosen to validate the models were R^2, mean squared error (MSE), RMSE, mean squared log error (MSLE), root mean squared log error (RMSLE), in line with those chosen by [115]. The conditions for successful models to be progressed to the testing stage were a training (and for NN models, validation) performance of $R^2 > 0.2$ and $RMSE < 0.35$, which was applied again after the testing stage to shortlist the top models. To obtain the quantiles, the testing data was run through each of the top models to obtain their RUL predictions, where the 5 %, 25 %, mean, 75 % and 95 % percentiles were calculated for each timestep. The choice of testing data was Run 1, Bearing 4 for IMS and Bearing 1_3 for FEMTO, as the original authors method performed well on these and was decided to be a good point of comparison. This process generated results for all combinations of the 2 datasets, 2 sensor channels, 4 data sampling regimes, 3 spectrogram bin counts, 2 hyperparameter optimisation searches and 2 model choices, resulting in 192 distinct pipeline designs. For each pipeline, the maximum number of models to analyse is the top 10 NNs and a LR model, however not all combinations produced this amount of models that successfully passed the metric bounding criteria.

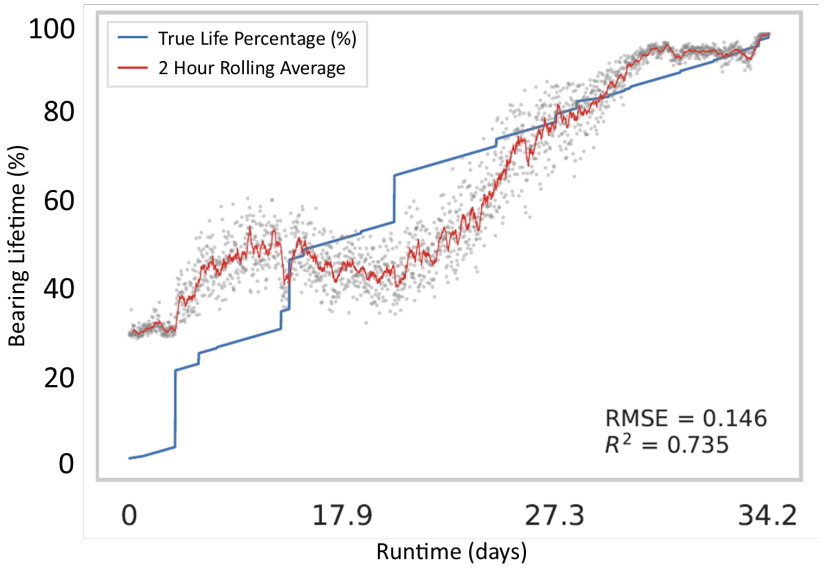

Figure 3.3: IMS Run 1, Bearing 4 Test Results from [115]: $R^2 = 0.735$, RMSE = 0.146

3.4.4 Case Study 1: Results

As mentioned in section 3.4.3, the case for comparison between [115] and this work was Run 1, Bearing 4 testing data from IMS and Bearing 1_3 testing data for FEMTO.

IMS Results

The RUL prediction shown in Figure 3.3

shows [115] results for their best performing model on the IMS dataset. This NN model has a Weibull-RMSE Combined loss function, 4 layers with 32 units per layer, 0 % dropout probability, lambda of 0.53, Weibull shape parameter (β) of 2 and characteristic lifetime (η) of 63.9 days. In Figure 3.3, the bearing lifetime extends from 0 % to 100 %, where the jumps are due to the gaps in data collection from the original IMS experiment. The NN predictions are smoothed using a 2 hour rolling average to more clearly demonstrate the trends in the prediction. As shown, the model fits this data well, with a low RMSE score of 0.146, and a high R^2 score of 0.735.

Figure 3.4 shows the quantiles and mean RUL estimate from the top NN models across all IMS pipelines which met the training and validation metric bounding criteria. The quantiles are calculated on the models performance on Run 1 Bearing 4 testing

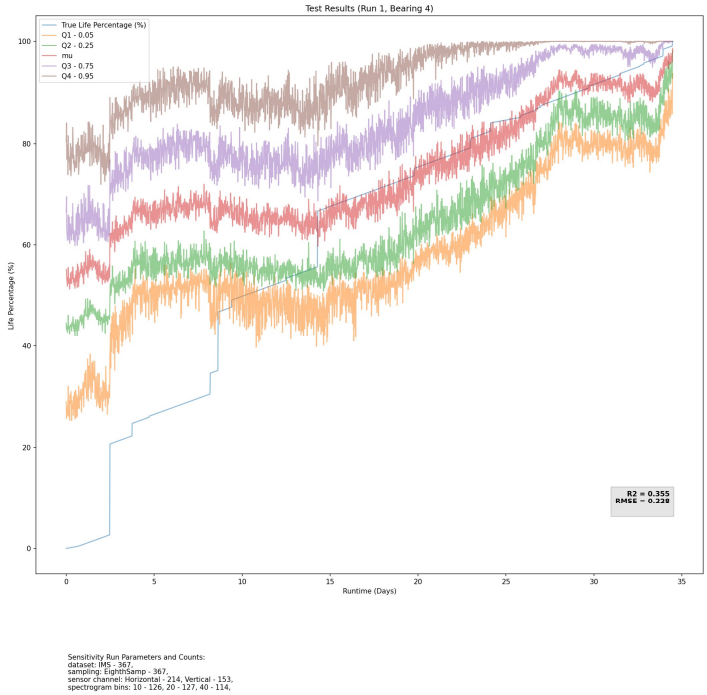

Figure 3.4: IMS Run 1, Bearing 4 Test Result Uncertainty (NN Model): $R^2 = 0.355$, RMSE = 0.228

data from the IMS dataset and the mean of these predictions result in a R^2 of 0.355 and RMSE of 0.228. From approximately 50 % bearing lifetime the quantiles bound the actual lifetime percentage until failure, with the mean fitting the true lifetime percentage well from 60 % lifetime onwards. As shown, the models do not predict early-mid life with any success, which may mislead an operator incorporating the model into a maintenance decision as the model cannot distinguish between any states $<$ 50% lifetime. Some of this deviation may be explained by the large jumps in lifetime % within the first 10 days of the experiment, compared to the much smoother data collection from day 15 to failure, regardless, this still undermines confidence in the predictions.

The results for the IMS LR models are shown in Figure 3.5. While the quantiles bound the true lifetime from experiment start to end, the lack of incorporated knowledge allows the models to expand to many multiples of bearing lifetime and into negative values. This results in a mean R^2 score of -0.223 and RMSE of 0.314, despite

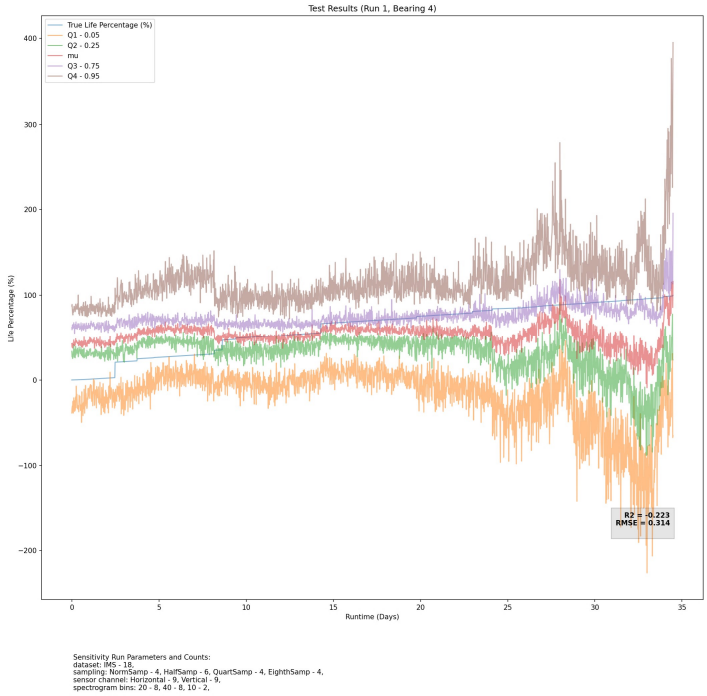

Figure 3.5: IMS Run 1, Bearing 4 Test Result Uncertainty (LR Model): R^2 = -0.223, RMSE = 0.314

all of the models successfully meeting the R^2 and RMSE bounds in the training stage. Additionally, the RMSE for the testing results is still within [115] 0.35 boundary while producing unreliable predictions, suggesting other forms of validation are required in tandem to discount unsuitable models. This demonstrates that applying regression models that minimise computational cost or maximise interpretability cannot always perform the required task, and further demonstrates the need for hybrid modelling approaches to incorporate known behaviour.

The pipeline design parameter summary is shown in Table 3.4 which shows the breakdown of pipeline stage parameter counts in the final model selection. The maximum number of models is the top 10 NNs from the 46 IMS NN pipelines, and single LR model for each of the 46 IMS LR pipelines if all of these models trained successfully. This results in an acceptance rate of 79.8 % for the NNs (367 out of potential 460 models were successful) and only 39.1 % for the LR (18 out of potential 46 models were successful), demonstrating that the NN is more likely to be successful at this prognostic

Pipeline Stage	Value	NN	LR
Max Models	NN - 460	367	-
	LR - 46	-	18
Sensor	Horizontal	214	9
	Vertical	153	9
Sampling	Normal	101	4
	- 1/8	95	4
	- 1/4	86	4
	- 1/2	85	6
Spec.Bins	10	126	2
	20	127	8
	40	114	8
HyperParam Search	10	137	-
	100	230	-

Table 3.4: Summary of IMS pipeline settings for LR and NN models (by successful model counts)

task. For the 18 successful LR models, the sensor alignment choices are split evenly, implying the sensor orientation neither hindered nor helped the models performance, while the NNs tended to favour the horizontal channel as chosen by [115]. Interestingly, for the sampling regime the LR models favoured learning from the least data and fared equally amongst the other options. The NNs were also fairly evenly spread amongst the sampling options, favouring the maximum amount of data. The LR models selected the most condensed spectrogram the least, implying the higher dimensional representations provided more useful degrees of freedom to the model. Conversely, the NNs were more evenly spread across the bin options, suggesting all options provided the NNs with enough information. To summarise, it appears that on the IMS dataset, the most influential design parameter was the dimensionality of the input data for the LR models as shown by the aversion to the 10 bin spectrogram, and the time available to optimise the hyperparameters for the NN as this displayed the largest diversion by model contribution in favour of larger number of searches.

FEMTO Results

The RUL prediction shown in Figure 3.6 shows [115] results for their best performing model on the FEMTO dataset, with a Weibull only RMSLE loss function, 2 layers with

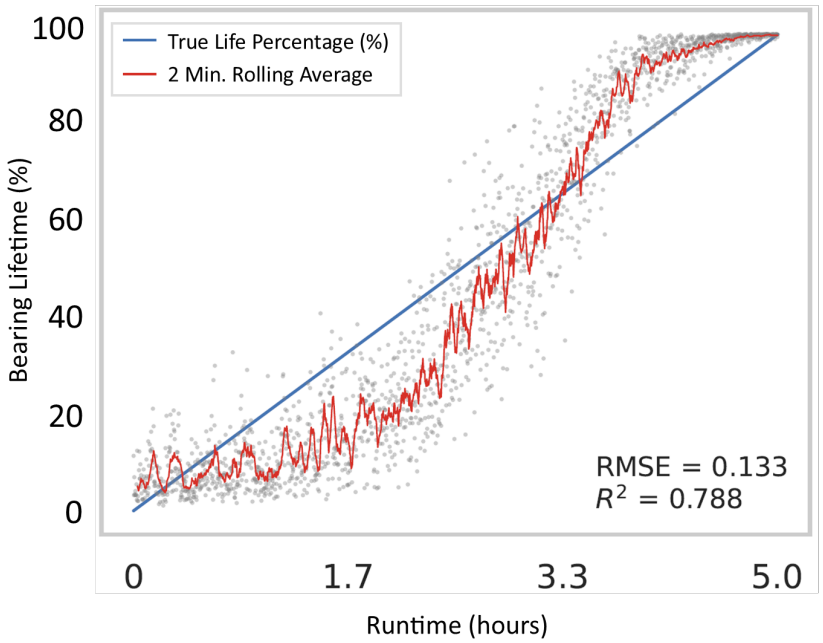

Figure 3.6: FEMTO Bearing 1_3 Test Results from [115]: $R^2 = 0.788$, RMSE = 0.133

32 units per layer, 0.25 % dropout probability, lambda of 2.28, Weibull shape parameter (β) of 2 and characteristic lifetime (η) of 4.8 hours. The trend of the predictions is shown by a 2-minute rolling average with straight line from 0 - 100 % demonstrating the bearing lifetime. This NN fits the data well as shown by the low RMSE of 0.133 and high R^2 of 0.788.

The NN FEMTO uncertainty plot is shown in Figure 3.7, which shows the quantiles bounding the whole bearing lifetime, but does not narrow as much as the IMS results at end of life. This larger spread in predictions demonstrates the volatility of the NN predictions on this dataset, as depending on the model, the prediction could be anywhere between 0 and 60 % at start of life and 50-100 % at end of life. The mean prediction has R^2 of 0.729 and RMSE of 0.15 which suggest the mean has a decent fit, however, it can be seen that the models tend to overestimate degradation early-mid life and underestimates mid-end life. If used to inform maintenance schedules, the start of life predictions could result in actions being taken too early where still usable components are prematurely replaced. Actions taken based upon the end of life predictions could be left too late, putting operators at risk of unplanned outages.

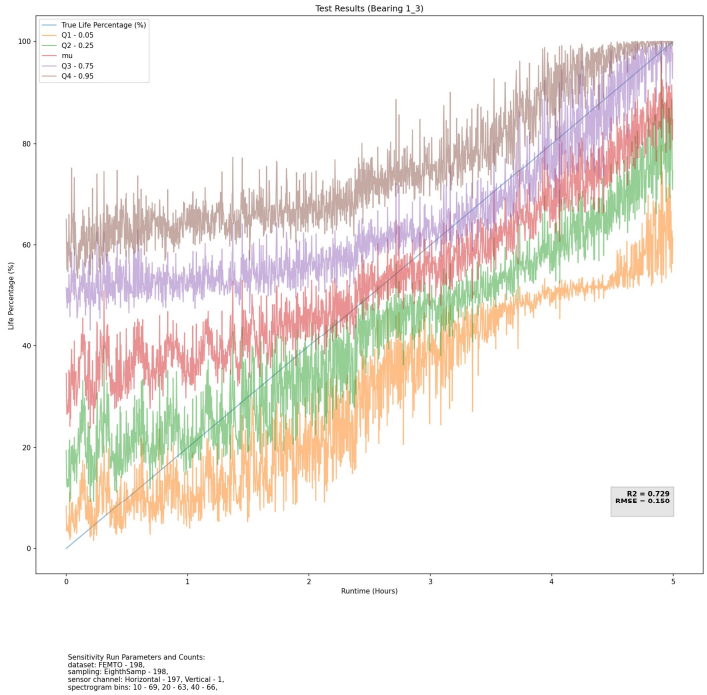

Figure 3.7: FEMTO Bearing 1_3 Test Result Uncertainty (NN Models): $R^2 = 0.729$, RMSE = 0.150

The fit of the LR models in Figure 3.8 shows consistent estimations early-mid life, then a huge divergence of multiple lifetimes in positive and negative direction is observed in the final stages of the bearing life. This may be due to the rapid decay of the bearings, as the spectrograms show a rapid increase in vibration for some of the training data in the later stages of the experiment. As the end of life prediction is arguably the most crucial aspect of prognostics, these LR models could be considered a risk for any operator to employ in maintenance activities.

In the pipeline design summary in Table 3.5, both the NN and LR models have relatively even contributions to the 198 successful NN models and 24 LR models from all settings for the sampling and spectrogram bin options, suggesting these do not have a great influence on the model performance. This is also true for the sensor alignment for the LR models, while for the NN models there is almost entirely self selected horizontal channel, as in [115]. This suggests that the horizontal sensor provides the most useful information for the NN model. Additionally, the NN has strong contribu-

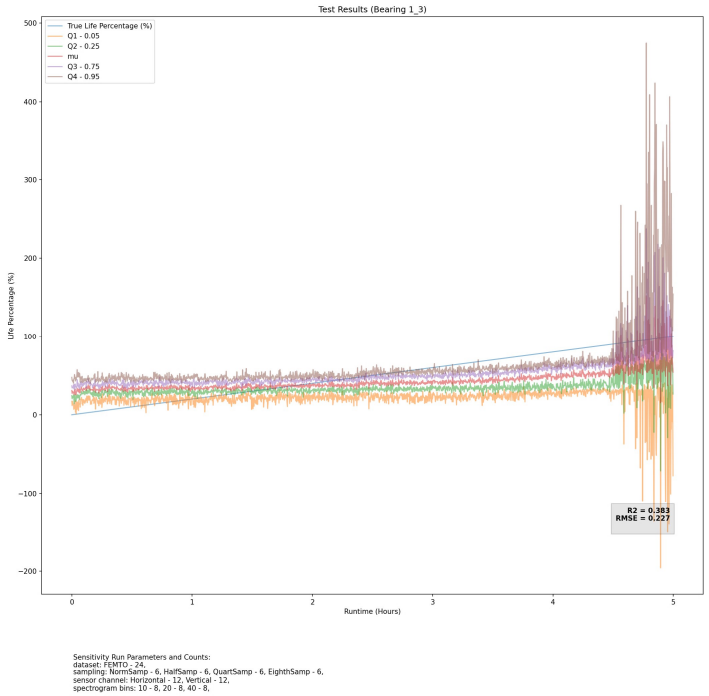

Figure 3.8: FEMTO Bearing 1_3 Test Result Uncertainty (LR Models): $R^2 = 0.383$, RMSE = 0.227

Pipeline Stage	Value	NN	LR
Max Models	NN - 460	198	-
	LR - 46	-	24
Sensor	Horizontal	197	12
	Vertical	1	12
Sampling	Normal	45	6
	- 1/8	43	6
	- 1/4	55	6
	- 1/2	55	6
Spec.Bins	10	69	8
	20	63	8
	40	66	8
HyperParam Search	10	77	-
	100	121	-

Table 3.5: Summary of FEMTO pipeline settings for LR and NN models (by successful model counts)

tion from the larger hyperparameter search with a majority of models being chosen by the random search of 100. Finally the NN models have an acceptance rate of 43.0 % while the LR models have an acceptance rate of 52.2 %. Interestingly, while the mean NN performance produces better results for R^2 and RMSE, the LR models are more consistently performing above the set metric boundaries and being accepted into the testing stage. Despite their unsuitable design, the choice of metrics and bounds used to assess these models suggest they should be accepted, again suggesting that models require more diverse validation to determine their general suitability, or what situations they may be best suited for. This may also require an appreciation of the similarity of the training and testing data, as models that succeed at the training stage should be trusted to succeed in the testing or online monitoring stage.

This sensitivity analysis has demonstrated that the approach taken to data pipeline definition can have a significant impact on the accuracy of prognostic algorithms, with evidence for a specific bearing vibration case-study provided. This case-study suggests that when developing a data pipeline for this purpose valid models can be selected from a variety of plausible data pipeline configurations while resulting in a diverse range of learned RUL trajectories.

3.4.5 Case Study 1: Result summary and discussion

Civil nuclear is a safety critical industry which cannot readily deploy data-driven analytics in decision-making processes without quantification of the uncertainties involved. Consequently, in this work an analysis of the impact of data acquisition pipeline design decisions on the performance of an existing hybrid RUL model for bearing prognostics was conducted. It was shown that the design decisions made at key stages of the data acquisition pipeline can create a large variance of potential RUL trajectories for both NN and LR models on both of the bearing run-to-failure datasets utilised in the study. The models were more sensitive to some design decisions than others, such as the available number of hyperparameter optimisation searches for the NN or the dimensionality of the input features for the LR model (on the IMS dataset). The presence of incompatible design decisions was not suggested by the results as many stages produced an equal

number of successful models across the different design options. This suggests that valid models could be generated from completely different pipeline designs, which result in an entirely different learned RUL trajectory. Understanding how the data acquisition pipeline can impact on hybrid prognostic tools can allow nuclear plant operators to justify utilising resources towards reducing high uncertainty areas in the pipeline design to provide more confidence in applying these tools to support maintenance processes. This is of particular concern in the nuclear industry as ML algorithms applied to rotating plant deployed in nuclear engineering environments experience unique operating challenges, such as legacy data acquisition systems that have been upgraded over time without emphasis on the data that will be used for ML purposes.

The models were filtered by a requirement of $R^2 > 0.2$ and $RMSE < 0.35$ to remove unsuitable models before progressing to the testing stage, as in [115]. The results showed that the chosen metrics are not sufficient to definitively identify unsuitable models and are not descriptive enough to show the operator where model application should and should not be trusted. Additionally, the chosen training and testing data may not have been sufficiently comparable for LR type models, as shown by models that had been deemed acceptable in the training stage performing poorly on IMS testing data in Figure 3.5.

To further develop this work, more analysis would be conducted on the impact of metric bias in the model selection process. Models were selected and ranked based on their R^2 and RMSE scores, but a different selection of shortlisted models may have been generated if different metrics had been used or prioritised. Additionally, if it was discovered that some models were more accurate for end of life predictions while other models are more suited for early-mid life, this may not be captured by summary statistics used to qualify the overall model usefulness. Additional methods to describe where the model is successful is needed to further justify the models use for specific prognostic stages, which could be aided by the application of explainability tools. Finally, for a more robust comparison, knowledge would be incorporated into different model types. This would provide more hybrid combinations to compare against, while investigating how model bias impacts the RUL prediction.

3.5 Case Study 2A: Quantifying design uncertainty in data pipelines

Bearings are prolific in industrial applications, and are a common point of failure in rotating plant [173]. Accordingly, two well-understood open-source bearing fault test rig data sets will form the basis of a transfer learning task, with Section 3.5 covering the creation of the source domain pipelines, their diagnostic process choices and the subsequent identification of negative and positive performance drivers for this asset. For the source domain, the publicly available Case Western Reserve (CW) rotating plant dataset [209] was chosen, which includes a selection of seeded bearing faults introduced to key locations in a motor. Faults are present in the ball, inner race and three locations in the outer race at both ends of the motor. The faults are monitored by vibration sensors present at the motor fan end, drive end and base plate. The motor operating conditions are varied between 0-3 HP, with data collected at 12 or 48 kHz (down-sampled to 12 kHz for consistency). This dataset has been extensively utilised to demonstrate machine learning models for engineering applications, and in this instance acts as a demonstrative baseline to showcase the framework rather than the classification capability of a particular machine learning model.

To enhance the domain interpretation of the system, a SHAP-based human readable explanation framework was constructed, as in [184]. This was created to explain how the design choices in the pipeline relates to the uncertainty, in the form of the classification error produced by the model in that pipeline. The design of the pipeline is 'one hot encoded', where categorical variables with n classes are converted into n separate binary variables, with a 0 representing an absence of the choice in the design and 1 representing the inclusion of the choice in the design. A decision tree regressor is fit over all available data [210] with the encoded pipeline design as predictors and the model error as the target. As the tree model is used to find relationships between the pipeline stages and error, all available data is used for training [210].

To aid the understanding of SHAP plots produced by the SHAP-based framework, human readable explanations are generated to explain the significance and type of influ-

ence each choice has on the detection or diagnostic system performance. Explanations provided by the framework elaborate on the choice's individual impact on the system error using domain relevant language. In cases where the SHAP tools have identified the choice as a 100 % positive or negative influence, further information is provided using the information from the raw pipeline design data. For such cases, the framework allows for local explanations which provide further analysis of anomalous results. From the human readable explanations, the most impactful positive and negative performance drivers for each pipeline stage can be selected with the aim of producing the 'best' and 'worst' performing pipelines to validate the SHAP-informed pipeline configurations. To construct a highly performing pipeline, design decisions with a positive influence and highest impact are selected for each pipeline stage.

The methodology followed throughout Case Study 2 (part A and B) is described in Figure 3.9. This details the pipeline stage and design choice identification stage; design construction process; applying the SHAP-based framework to explain the impact of each design choice; creating improved pipeline designs from provided insights; and finally, transferring these insights to designs for new systems. Each of these stages will be discussed in the following sections.

3.5.1 Source domain pipeline design stages and decisions

Six key pipeline stages and their potential design decisions are identified with the choices at these stages given in Table 3.6. The chosen stages reflect key areas in the pipeline design for this dataset where a diverse range of decisions are available, and the optimal solution is not immediately clear. The model choices are chosen similar to [204] due to the range of classification boundary attribution methods, while the window length of the vibration data and diagnostic task is decided based on the number of samples and included fault types. The row ordering in Table 3.6 reflects the ordering of pipeline stages that was applied in Algorithm 1.

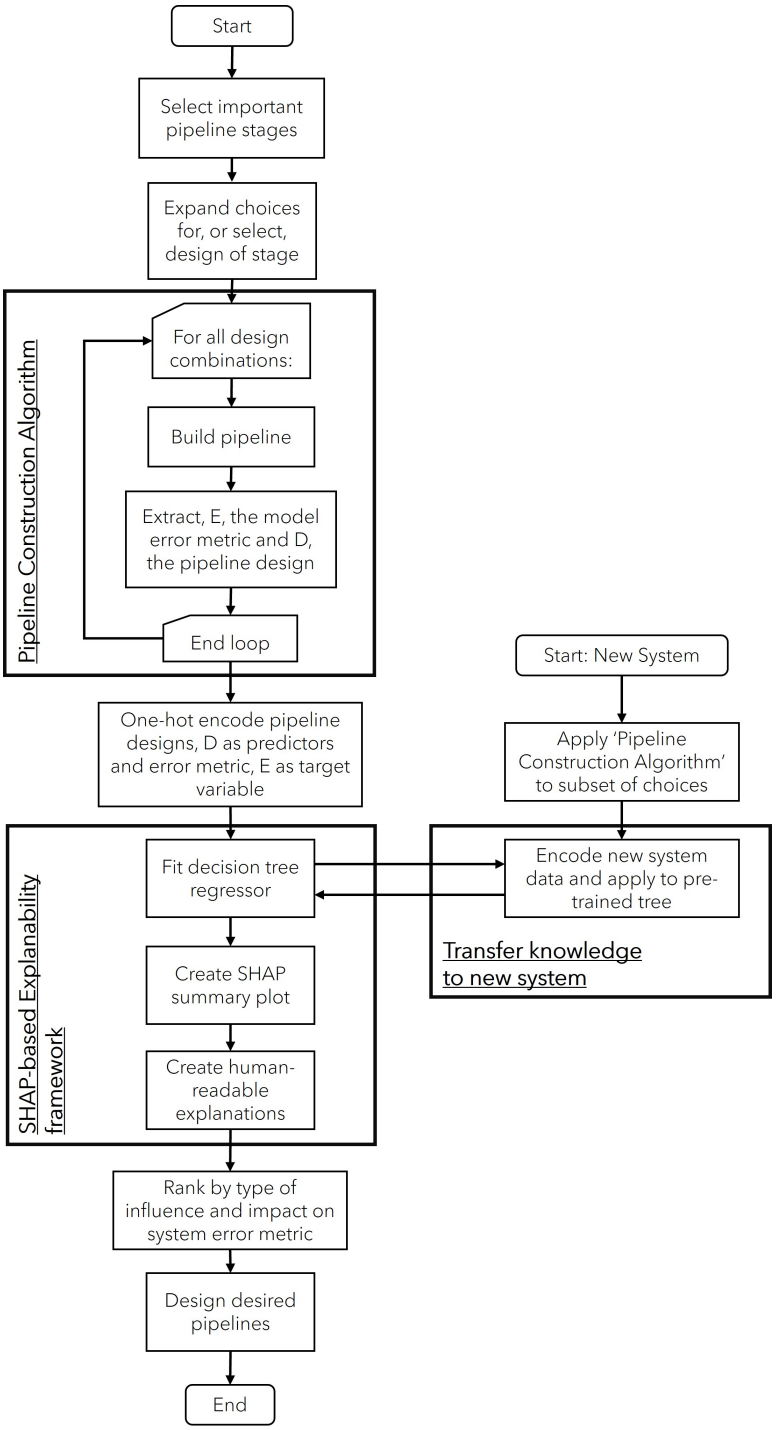

Figure 3.9: Flowchart of the pipeline design, construction, explanation and transfer to new systems

Table 3.6: Summary of pipeline stages, choices and their rankings for source and target datasets (averaged over 5 runs)

Pipeline Stage (num. choices)	**Pipeline Choice**	**Rank (Source Dataset)**	**Rank (Target Dataset)**
Sensor, (4 **(1*)**)	Fan end only sensor	6.0 +	(N-A)
	Drive end only sensor*	**13.0 +**	**8.8 +**
	All sensors	15.4 -	(N-A)
	Drive end and base plate sensors	24.6 -	(N-A)
	Fan end and base plate sensors	29.4 +	(N-A)
	Drive end and fan end sensor	34.4 -	(N-A)
Window, (2)	0.5s window	26.8 +	20.6 +
	1 s window	30.4 -	22.0 -
Model Task, (6 **(1*)**)	**Task: Binary (TB) (0, healthy or 1, fault)***	**1.0 +**	**1.0 +**
	Task: End (TE) (healthy, Fan end or Drive end faults)	3.0 +	(N-A)
	Task: Location (TL) (healthy, ball, inner race or outer race faults)*	**31.0 +**	**20.4 +**
	Task: All (TA) (healthy, fan end ball, inner race, outer race or drive end ball, inner race, outer race faults)	34.6 -	(N-A)
Data Domain, (3)	Wavelet scattering with principle component analysis (PCA)**	2.0 +	2.0 +
	Frequency (Power Spectral Density) with principle component analysis (PCA)**	7.6 -	6.6 -
	Timeseries Statistics (mean, median, standard deviation, root mean squared (RMS), peak, skewness, kurtosis, crest, shape and impulse)	20.8 +	23.4 +
Data Alloc. (2)	Random Allocation	38.4 -	38.4 -
	Prevalence Allocation	38.6 +	38.6 +
Model, (22)	ESD (Ensemble subspace discriminant model)	4.0 -	4.8 -
	GNB (Gaussian Naive Bayes model)	5.0 -	6.0 -
	CGSVM (Coarse Gaussian support vector machine) model	7.8 -	10.2 -
	CT (Coarse Tree) model	9.0 -	19.0 -
	CSVM (Coarse support vector machine) model	10.2 -	17.2 -
	FKNN (Fine K-Nearest Neighbours) model	11.2 +	7.2 +
	MKNN (Medium K-Nearest Neighbours) model	12.4 +	9.0 +
	MGSVM (Medium Gaussian support vector machine) model	15.4 -	22.6 -
	KNB (Kernel Naive Bayes) model	15.6 -	26.6 -
	CsKNN (Cosine K-Nearest Neighbours) model	16.6 +	12.0 +
	WKNN (Weighted K-Nearest Neighbours) model	18.0 +	13.2 +
	CbKNN (Cubic K-Nearest Neighbours) model	19.6 +	15.2 +
	EBgT (Ensemble Bagged Tree) model	20.6 +	15.8 +
	CKNN (Coarse K-Nearest Neighbours) model	22.6 -	23.4 -
	LSVM (Linear support vector machine) model	23.4 -	35.0 -
	QSVM (Quadratic support vector machine) model	26.2 +	34.0 +
	MT (Medium tree) model	27.0 -	33.4 -
	FT (fine tree) model	27.6 +	21.2 +
	ESKNN (Ensemble subspace K-Nearest Neighbours) model	28.8 +	15.8 +
	ERUSBT (Ensemble Random undersampling Boosted Tree) model	31.6 -	23.0 +
	EBoT (Ensemble Boosted Tree) model	34.2 -	26.2 +
	FGSVM (fine Gaussian support vector machine) model	36.2 -	30.6 -

* Subset of choices used for the target dataset, for all other stages the choices are the same. ** Number of principle components chosen to explain at least 95% variance.
Choices are marked to indicate if they have an overall (+) positive or (-) negative influence on the system performance. (N-A) marks choices unavailable to the target dataset.

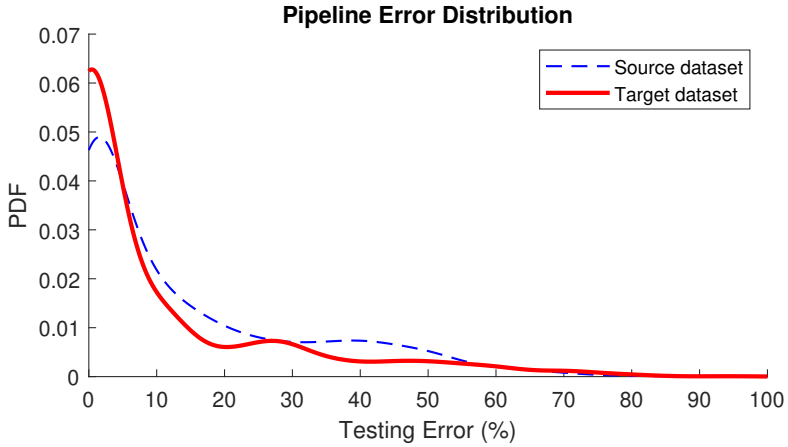

Figure 3.10: Probability distribution of classification errors across the source and target data sets. The source data set consists of 31680 CW pipelines. Many pipelines result in low, near 0 % errors and the distribution has a heavy tail at higher errors with a peak around 40 %. The target data set consists of 2640 generated pipelines. Many pipelines result in low errors, with a heavy tail towards high errors. There is another peak near 30 % error, similar to the source domain

3.5.2 Uncertain pipeline construction, explanation and selecting improved pipeline designs

Classification testing error was collected for all combinations of variable settings for pipeline design variables, with 5-fold cross-validation for each combination resulting in 31680 total pipelines. Varying the chosen design variable settings resulted in a wide range of performance output, as shown by the testing error distribution in Figure 3.10. Two dominant modes are present in the error distribution at ~ 0 % and ~ 40 % testing error, with a heavy tail towards the upper end of the error range, representing setting combinations which contribute to improved or degraded performance, respectively. The fitted decision tree regressor model used to identify pipeline to error relations in the SHAP-based framework resulted in a R2 score of 0.9788, with an ideal value of 1, showing that the model fit the data well but still incurs some error. To account for this, the model is retrained 5 times with each trained model being passed through the human readable XAI framework.

Applying the SHAP tools within the human-readable XAI framework generated magnitude plots which describe the ranking of design choices based on the magnitude

of their impact on the bearing fault system error (taken from the trained decision tree output), and summary plots which show the direction of this impact (i.e. do the design choices drive the system error up or down). Example figures are included within Appendix B. The SHAP tools and human readable explanation tools within the human-readable XAI framework are used to order the design options based on the strength of their impact on the performance outcome (which could be positive or negative). Each design choice is ranked 1 to 39 (the total number of choices) for each retrain of the tree, which is then averaged to give an overall score. The average ranking of all pipeline choices across all stages for 5 decision tree model runs, shown in Table 3.6.

The high importance and positive influence of 'TB' (Task: Binary) and 'TE' (Task: End) choices mean models are more likely to perform well on the simpler binary (TB) fault detection case and can differentiate if the faults are present on the drive or fan end of the motor (TE), suggesting the data is quite easily separable using these groupings. Generally, the models seem to struggle with the 'All' (TA) choice, the task containing the highest diagnostic information and complexity, suggesting that separating the data based on both motor cross section and end of the motor is difficult for the models to separate accurately. The K-Nearest Neighbour (KNN) models are the most important, positively performing model group, which attribute classification boundaries by clustering, suggesting this method is the most appropriate for this case study. The least successful models (ESD, GNB, CGSVM, CT, CSVM) have diverse boundary attribution methods, but tend to be the 'coarse' equivalent which restricts the amount of detail that can be captured by the model. This suggests more complex models are required to adequately capture the fault characteristics. The 'Wavelet' (time-frequency domain) method was the most impactful and positive influencing choice for the data domain stage of the pipeline. This supports standard engineering understanding that providing models with time and frequency information of a developing fault leads to the strongest identification of the type of developing fault, instead of using only time or frequency domain representations [211]. For the sensors, the combinations containing drive end sensor or baseplate sensor in tandem with other sensors tend to perform negatively and the most important sensor combinations (fan end only with positive

Table 3.7: Summary of 'best' and 'worst' pipeline choices per stage for source and target datasets

Pipeline Stage	**CW Dataset**		**Target Dataset (MFPT)**	
	Best Choices	**Worst Choices**	**Best Choices**	**Worst Choices**
Sensor	Fan end sensor only	All sensors	Drive end only*	Drive end only*
Window	0.5 seconds	1 second	0.5 seconds	1 second
Data Allocation**	Both (label prevelance and Random)	Both (label prevelance and Random)	Both (label prevelance and Random)	Both (label prevelance and Random)
Data Domain	Wavelet (time-frequency)	Frequency	Wavelet (time-frequency)	Frequency
Classification task	Binary (fault detection)	All - specific motor cross section location and motor end	Binary (fault detection)	Motor cross section location
Classification Models	Fine, Medium, Cosine, Cubic and Weighted KNN models	Ensemble Subspace Discriminant, Gaussian Naive Bayes, Coarse Gaussian SVM, Coarse Tree and Coarse SVM models	Fine, Medium, Cosine, Cubic and Weighted KNN models	Ensemble Subspace Discriminant, Gaussian Naive Bayes, Coarse Gaussian SVM, Coarse Tree and Coarse SVM models
Error	0.057 % (max), 0 % (min)	82.4 % (max), 50.8 % (min)	0 % (max), 0 % (min)	70 % (max), 0 % (min)

* There is only one sensor option, so must be selected for both best and worst pipelines. ** The data allocation stage was equally unimportant with a neutral influence on performance between both datasets, so both options are included for selecting best and worst pipelines.

performance, drive end only with positive performance, all sensors with negative performance) also have the most extreme positive or negative influences which drives up their impact on the system performance. The fan end sensor may experience less noise due to being slightly removed from the driving force, which may explain its generally positive influence on the system performance. Lastly, the least significant contributing features towards pipeline error were the design choices for the training and testing set data allocation method stage (Random or Prevalence allocation). Due to the large supply of fault examples in this curated data set, both methods may ensure a variety of fault types would be visible to the models to allow for more consistent training.

Designing the best pipeline based upon the human readable explanations from the SHAP plots identified the top choices as shown in Table 3.7. The original pipeline data

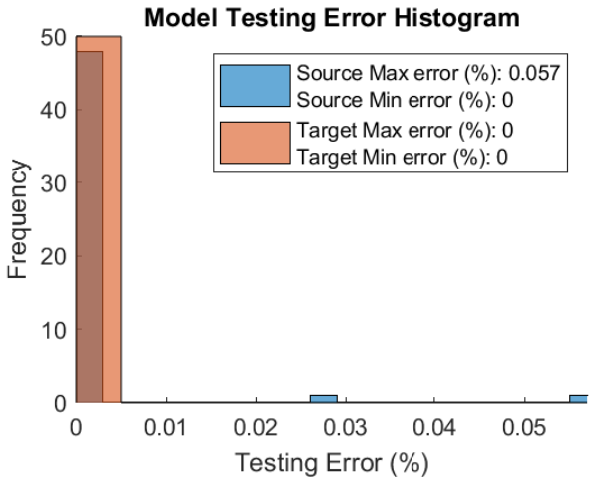

Figure 3.11: Histogram of classification errors from SHAP recommended 'best' pipeline design choices for the source and target domains. All chosen source domain and target domain pipelines have very low classification error of maximum 0.057 % and 0 % error, respectively

was filtered using the chosen design choices, and the resulting error histogram for the top 50 pipeline designs is shown in Figure 3.11 with the maximum error of 0.057 %.

3.6 Case Study 2B: Transferring quantified uncertainty to another system

3.6.1 Transferring knowledge of design uncertainty to new systems

There is potential for knowledge gained from one system to be transferred to gain insight into the uncertainty drivers in the design of a similar, less observed system. Fully observing one system to an extent that provides enough information to meaningfully compare pipeline designs and the drivers of uncertainty can require significant overhead. Suitable sensor coverage, fault measurements and computational resources to compute all design combinations must be provided, which may be possible on a test rig but may be difficult to translate into a practical environment.

To transfer instances from the target domain to the source domain, there must be alignment potential between the pipeline classification error between the source and target domains, and the alignment potential between the pipeline design choices,

ensuring the design of the pipelines should have the same or fewer design choices. This can be determined through various metrics or visualisations, as demonstrated later in this section. Instance transfer is conducted by converting the design of the target domain pipeline to align with that of the source domain. Having the same or fewer choices allows the instances from the target domain to align with the source domain, as design options that are not observed in the source domain cannot be leveraged to the target domain. This instance transfer will allow the pre-trained tree model in the SHAP-based framework from the source domain to be applied to any of the target domain pipelines with consistent encoding of the model inputs. As the model has learned from the source domain, this knowledge can provide insight into the behaviour of the target system pipelines.

A second system may be under-observed as the time and cost overhead required to augment the asset to collect the same diversity of measurements and re-certify the asset may be infeasible. Instead, with the construction of a source system completed, the knowledge of the present uncertainty drivers can be transferred to the under-observed target system. The open-source bearing test rig data set for the target system is the Society For Machinery Failure Prevention Technology (MFPT) dataset [212].

The target dataset contains bearing faults introduced into the ball, inner race or outer race of a motor driven bearing housing which is monitored by one vibration sensor. The operating loading conditions are varied for each bearing health state: healthy data is collected at 270 lbs (sampled at 97.656 kHz), inner race faults are collected for 0-300 lbs (sampled at 48.828 kHz) and outer race faults collected for 25-300 lbs (sampled at 48.828 kHz) and 270 lbs (sampled at 97.656 kHz). The input shaft is driven at 1500 rpm with each experiment lasting 3 or 6 seconds. Due to being a smaller dataset, the 48 kHz cases were upsampled for consistency.

While including similar bearing faults, the target dataset includes less sensor coverage and fault locations than the source system, which limit the amount of pipelines that can be designed. This specifically affects the sensor and classification task options, as there is no access to fan end and base plate sensors or fan end faults. The summary of pipeline stages and choices for the target dataset is shown in Table 3.6.

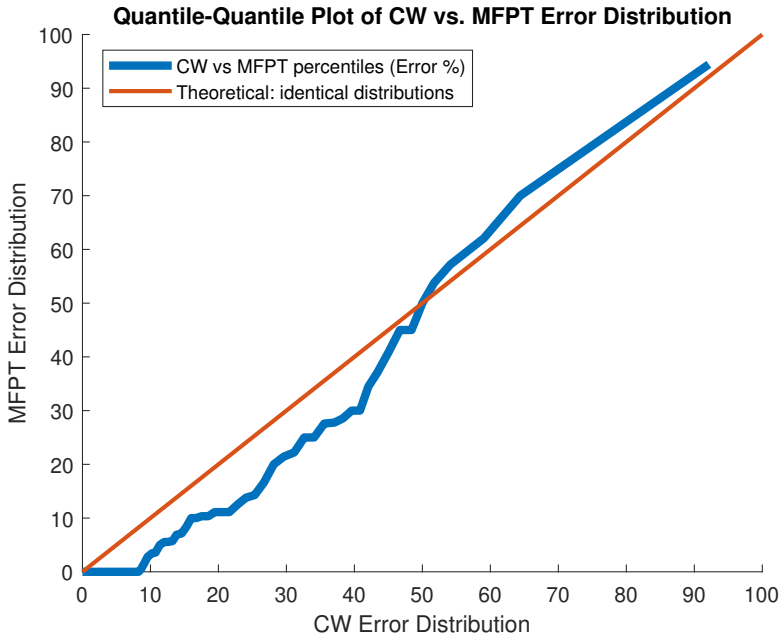

Figure 3.12: Quantile-Quantile Plot of the source and target pipeline error distributions. As the two systems align well with the theoretical plot, information gained from one system would provide useful insight into the behaviour of the other. This is effectively transference of the expected errors between the source and target systems

3.6.2 Target data pipeline design

To validate the methodology, the pipelines for the combination of all available choices were constructed with the classification error collected for each pipeline with 5-fold cross-validation, as in Section 3.5, resulting in 2640 total pipelines for the target case. A subset of this will be provided to the human-readable XAI framework trained in Section 3.5. The histogram of errors over all pipelines is shown in Figure 3.10, with a peak at ~ 0 % and ~ 30 %, and a long tail towards high classification errors. The histogram demonstrates the target domain behaves similarly to the source domain. A Quantile-Quantile plot in Figure 3.12 is used as a comparison between the source and target error histograms over all pipeline designs and captures the errors at the design extremity. The Pearson correlation coefficient between the two distributions is 0.97 showing a strong linear correlation between the quantiles of the two error distributions. The similar behaviour (identical distributions follow the plotted theoretical line) suggests that the

knowledge in the source pipelines is a good candidate to support the explanations of the target case.

The data is one hot encoded to align with the source data and a subset of 110 pipelines are provided to the human-readable XAI framework (that as been trained on pipeline data from Section 3.5), to explain the impact of each choice. The subset chosen were those with a window length of 1 s, classification task of cross section location (TL), data allocation method of random and the data domain of frequency with all models and cross validation samples taken. These were chosen due to design complexity (frequency domain and classification task offer increased complexity in their pipeline stages), dataset limitations (only one sensor option), or randomly (there was no obvious advantage within the choices for the dataset allocation or window length stages).

The average ranking of each choice from the human-readable XAI framework is shown in Table 3.6 (right hand column), allowing for a comparison with the source case.

Many of the recommendations align with the source system, such as the allocation methods being equally unimportant due to the balanced ratio of faulty to healthy data; the high performance on the binary classification task; and processing data in the wavelet (time-frequency) domain providing more useful information to the models. The target case has several choice limitations compared to the source system, such as the limited diagnostic tasks and sensor combinations, however the influence of the available choices are in agreement with the source case (both contributing positively to system performance). Due to the limited samples provided to the human-readable XAI framework, the influence of some model types has changed, such as the Ensemble Boosted Tree and Ensemble RUS Boosted Tree which perform well for the limited pipeline samples provided for the target case. Lastly, the window lengths have increased in importance but result the same influence. The target dataset has less data samples than the source domain and the use of the 0.5 second window length can increase the number of observations used to train the models.

3.6.3 Selecting improved pipeline choices for the target system

Using the human readable explanations to rank the choices in each stage for the target case, the 'best' pipelines can be extracted from the all the target pipeline data, including the held out design data. For the 'best' pipelines, the choices presented in Table 3.7 are chosen. Aside from the limited sensor options and classification task, the rest of the recommendations for the 'best' pipeline choices align with the recommendations for the source system. The histogram of errors for the 'best' 50 pipeline designs are shown in Figure 3.11 with the maximum error of 0 %. Applying the 'worst' choices highlighted by the source domain do lead to degraded performance in the target domain data as shown in Appendix B.

3.6.4 Case Study 2: Result summary and discussion

Highly regulated industries require complete understanding of the level of trust they can have in their data acquisition and information systems. Trustworthiness can be eroded along the entire length of a data pipeline, from sensor placement to the analytic performing the fault diagnostics. The empirical study presented in this chapter has contributed a means of explaining how uncertainties in each design stage of a pipeline influence the overall error, and how this understanding can be transferred to new target systems which may lack abundant labelled data. Once a source domain is created, the explanation framework can then be used to prescribe high performing pipeline designs in other, similar monitoring situations without repeating the exhaustive evaluation of design choices. The approach in this work successfully demonstrated how constraints placed on a bearing fault classification system can improve or degrade the performance of the system through interactions between choices made at different data pipeline stages. The SHAP-based framework explanations were able to adequately identify choices which lead to the construction of better bearing fault classification pipeline performance, even across difficult to compare data processing stages, providing an operator with insight into the uncertainty drivers in their data acquisition systems during the design stage.

The approach presented is flexible, it can allow developers to consider where de-

sign decisions are made in their proposed pipelines and to investigate the impact their choices can have on a data acquisition pipeline performance. In this presentation of the work, no constraints were placed on the design of the pipeline, however if a developer must work to certain specifications, this methodology could allow them to design around these constraints in order to improve system performance and understand the resultant limitations. Additionally, the proposed methodology can be expanded to handle more stages (or settings per stage) to suit the application and the human-readable explanations can be customised to present end-users with more familiar language to further enhance understanding of the outputs. Lastly, there is the potential to transfer learned uncertainty drivers with human-readable explainability frameworks trained on fully observed systems to gain insight into new, comparable cases which may allow experimental test rigs to translate to practical assets. This was demonstrated successfully by decoupling the positive performance drivers in a target system using a smaller subset of pipeline designs.

There are several notable limitations to the work, which would be high priority for future work. Firstly, investigating if sufficient pipeline performance alignment can be detected from the source and target input datasets would provide an early warning system for the potential of negative transfer, where the insights from the source domain may not apply to the target domain, before the pipelines are constructed. Additional analysis on the sensitivity of the method to the alignment between the pipeline design performance would further support the detection of negative transfer, which could be conducted through sourcing a third, less applicable dataset or investigating the impact of noise on the currently applied datasets. The SHAP framework applied in this work depends on high accuracy from the decision tree used to explain the impact of the pipeline designs. More investigation into the failure modes of this model type, with specific attention paid to local level explanations where pipeline design performance may deviate from identified patterns, would provide more insight and control to those responsible for pipeline design. Currently, to demonstrate the explanations of the pipelines applies to all designs explored within this work, an exhaustive search was used to ensure all design choices were compared. This can be computationally expensive

and depending on the impact of the insights may not provide additional benefits if the design choices have little impact. In future, more intelligent algorithms can be explored with a more cyclical approach, whereby the methodology discussed in this work could provide updates when new pipeline design results are available to test the impact of the new choices before they are exhaustively compared. A more intelligent search algorithm could highlight the most informative design choices to create pipelines for, making more effective use of the search space.

The application of this approach could be useful in cases requiring rapid design, or automation of design for analytic pipelines for similar assets in a fleet, while providing developers with information on where to focus resources to reduce diagnostic system risk. Future work could involve investigating the trade-off between the cost of implementing recommended design choices and savings generated by enhanced performance of maintenance tools over time.

3.7 Conclusion, contribution and future work

In conclusion, in energy sectors, downtime in key assets can be costly for operators, incurring lost generation revenues and associated penalties. To mitigate these costs, many operators have turned to prognostic and health management (PHM) and condition monitoring (CM) techniques to monitor the health of assets more closely [85]. PHM and CM techniques can utilize sensor data and machine learning (ML) models to detect the onset of faults [23] or predict the remaining useful life (RUL) of assets, and when these diagnostic approaches are applied to critical operational components they can decrease maintenance efforts and expedite return to service - but only if they are known to deliver high predictive accuracy.

Power plant operators are required to understand the uncertainties associated with the deployment of detection, diagnostic or prognostics tools in order to justify their inclusion in operational decision -making processes and to satisfy regulatory requirements. This is especially pertinent in the nuclear sector, where safety requirements are suitably strict. Operational uncertainty can cause underlying detection, diagnostics or prognostics models to underperform on assets that are subject to evolving impacts of

age, manufacturing tolerances, operating conditions, and operating environment effects, of which may be captured through a condition monitoring (CM) system that itself may be degraded. Many industries require a high level of transparency when utilizing data analytics to ensure plant operation is well informed and traceable to trustworthy evidence for subsequent reasoning and decision making. It is vital to know when to trust the output of analytics, and to understand where the largest uncertainty contributions occur along the data pipeline. Knowledge of the uncertainty present in a data pipeline de-risks the system by providing the operator with confidence in the quality of data and decisions being made by a fault detection, diagnostic or prognostic system.

In this chapter, three main outcomes have been collated over two case studies involving bearing fault prognostics or detection and diagnosis, the contribution of which are summarised as follows:

1. **Demonstrating that pipeline design impacts analytic performance within fault diagnostic systems**

    - In Case Study 1, the data pipeline design was shown to impact the outputs of data-based and hybrid models in a bearing prognostics application. This impact could improve or degrade the predictive performance of the prognostic system with data-based models being more negatively affected by this degradation. Hybrid models which incorporated domain knowledge were shown to be more robust to the negative impacts of poor pipeline design.

2. **Identifying and explaining highly performing design options using a human-readable XAI framework**

    - In Case Study 2A, explainability tools (SHAP) were applied to datasets created from encoding pipeline design choices as predictors of analytic error for a bearing classification application. This identified highly-performing or poorly-performing design options, with the additional ability to compare the impact of choices *across* pipeline stages to identify the most important stages of the pipeline. This provides an operator with information on where resources can be deployed to improve fault detection or diagnostic

system performance by, for example, justifying the cost of installing higher fidelity sensors or commissioning additional fault data collection projects. High predictive accuracy was achieved when utilising the design recommendations of the framework, resulting in a mean predictive error of $< 0.1\%$, while using the detrimental design options flagged by the framework resulted in mean predictive error of $> 50\%$. Distinguishing highly and poorly performing pipeline design choices allows for the improvement of pipeline design in cases of both very flexible design, where the choices are able to be chosen to produce the best performing analytic; and in cases where the design is constrained by fixed pipeline choices, where an associated risk can be attached.

3. **Improved fleetwide monitoring by leveraging insights from previous designs to new systems**

   - In Case Study 2B, the insights gained from the fully observed system investigated in Case Study 2A were used to design pipelines for a similar system with much fewer observations. Where systems are deemed similar (as shown by the Q-Q plot in Figure 3.12), the time taken to exhaustively test one asset can benefit the fleet of assets with less intensive testing required to acquire the same insight. This can improve fleetwide monitoring by providing pipeline design suggestions leveraged from the combination of previous designs and specific insights provided from minimum observations on the new system.

The pipeline design approach was applied to bearings, a common failure point across rotating plant, and shown to be successful across different maintenance tasks from fault classification to remaining useful life prediction. The case studies presented were composed of diverse types of pipeline stages which the framework was shown to accomodate, which could be tailored to many different assets. The transferability of insights between assets can save time and funds by reducing the amount of data collection and computational expense in computing the same amount of pipeline design observations

for a whole fleet of assets. However, for an operator to benefit from these insights, the assets are required to behave similarly and require enough fault samples to make meaningful predictions (a limitation of the choice of analytic model).

3.7.1 Future work

In future, considerations could be made to incorporate further pipeline stages noted in Figure 3.1, such as how operators can engage with the outputs of analytics. Additionally, due to this chapter focusing on open source datasets, the design and implementation of a customisable analytic pipeline rig would provide more flexibility and control to collect data under different conditions to provide more comprehensive coverage of key pipeline stages and their interactions. In this work, different models were utilised to provide coverage of different model 'families', with some work done to utilise hybrid rather than purely data-based models. This could be expanded to consider more complex models, and models with inherent uncertainty quantification. Additionally, to fully observe a system required all pipelines to be constructed and evaluated to collect data to be able to apply the explainable tools. This is a time consuming process, even when no physical equipment is involved. There are methods being designed in data labelling spaces which specialise in identifying the most impactful observations to be labelled to gain the most information for the fewest labelled samples [213]. A similar approach could be taken to identify the most impactful pipeline design combinations needed to provide sufficient insight into the system performance drivers without requiring exhaustive testing. This would improve the efficiency, speed and practicality of this methodology. Lastly, the impact of transferring pipeline designs was discussed in this work, but there was no consideration of transferring already trained models within a pipeline. New models were trained within each pipeline which could result in a large number of individual analytics requiring monitoring and maintenance across a fleet of assets. In Appendix A, a short study into how pretrained bearing fault classification models react to similar assets when trained on only one asset's data versus a mixture of both asset data is discussed. It was found that classifiers trained on only one asset can perform poorly on the other asset data due to domain shift caused by operating

condition differences between the assets. When the classifier is trained on a mixture of both assets' data, this effect is removed, and the models perform well on observations from both assets.

Chapter 4

Uncertainty in time: Quantifying temporal uncertainty in timeseries data for trustworthy temperature forecasting

4.1 Providing confidence in timeseries predictions for asset temperature monitoring

Maintenance interventions are required to keep power generation component temperatures within prescribed guidelines but with the consequence of lost generation days [1]. Understanding temperature increases caused by aging processes is critical to maintain safe operation but avoid needless interventions, particularly where power plants are approaching the end of their planned operational lifetime when assets may not operate as efficiently [85]. Temperature measurements can be subject to a variety of uncertainty and noise stemming from plant configuration, sensor calibration changes and the general variability of component aging. The capability to provide confidence bounds on the predicted temperatures in the presence of measurement noise can permit maintenance decisions to be made with sufficient certainty on lead time to select the best course of

maintenance action given operational or financial constraints.

This chapter presents an approach for identifying the rate at which temperatures can increase over a given operational horizon and presents a predictive distribution of the error that may result from that estimate. A framework utilizing the dependency structure between propagated measurement and modelling uncertainty is developed through investigating a series of increasingly detailed copula based approaches applied to the residuals from data-based predictive models. The generalised temperature timeseries forecasting methodology is demonstrated on both synthetic exemplar, open source bearing temperature data and real operational data provided by industrial partners for a heat exchanger in a nuclear power plant.

4.1.1 Contribution and novelty

The work presented in this chapter was conducted in partnership with Bruce Power, Canada, who provided timeseries temperature data from the inlet headers which provide coolant to reactor cores. In addition to this proprietary dataset, the methodology was also demonstrated on open source data to provide a point of comparison to the wider research community.

Many analytics tools struggle to provide capability at all temporal scales which result in a trade off between short-term and long-term accuracy. This compromise can be managed through the use of hierarchical modelling where models specialise in different tasks which can improve predictions over long and short horizons. Providing more accurate estimates of an assets future state, along with the uncertainty, and so risk, in this estimate, provides more flexibility in the monitoring and maintenance of the asset. Additionally, existing methods such as autoregressive models are capable of timeseries prediction, and some of these can provide uncertainty quantification. For example, Kalman filters can be applied to temperature forecasting but assume that errors are Normally distributed. For data with more complex dependency structures than linear, i.e. Gaussian distributed, this may result in degraded performance for autoregressive methods. This demonstrates a need to consider more complicated dependency structures for temperature forecasting that are baselined against a multivariate

Gaussian approach - contained herein as Multivariate Gaussian (Gaussian Marginals). The difference between Vine copulas and Multivariate Gaussian copulas in terms of their ability to capture linear and non-linear behaviour are shown in a short example in Appendix C. Copulas were originally created to incorporate marginal distributions with extreme tail behaviors into multivariate distributions [127] – in the context of this work, extreme temperature prediction error variations. These extreme behaviors are not necessarily expected values centered around the central mean but could also represent non-symmetrical extremes and multi-modal behaviors. Copulas permit the adopter to capture these extremes in a multivariate distribution and permit the dependency behavior to be specified individually. Thereafter, using the joint distribution, the derived conditional form can be used as a predictive model – in the context of this work, across multiple predictive horizons. This chapter presents a hierarchical copula-based modelling process for temperature timeseries forecasting with three key factors:

1. The underlying machine learning model is transparent and explainable which allows the method to be more appropriate for highly regulated environments (such as nuclear power plants).

2. The hierarchical approach combines the benefit of the long-term forecasting model and the calibration capability of the short-term forecasting model.

3. The copula approach incorporates uncertainty into predictions which allows the level of 'risk' in the forecast to be acted upon.

This chapter is organised as follows: Literature covering timeseries forecasting and the importance of temperature monitoring in PHM applications are presented in Section 4.2. This covers the landscape of timeseries prediction in PHM, discussing the impact of chosen forecast horizons, selected model inputs and managing sources of uncertainty. The synthetic data case study, methodology and results are presented in Section 4.3.2. The methods application to open source operational and real industrial data are presented in Section 4.3.3 and Section 4.3.4, respectively. Finally, future work and a summary of the results presented in this work are covered in Section 4.4.

4.2 Literature: Temperature monitoring applications and hierarchical timeseries forecasting

4.2.1 Temperature monitoring in prognostics and health management

Temperature measurements are an important indicator of health for many assets across engineering applications [19, 214]. Temperature changes can alert to early signs of increasing friction, component wear or reduced efficiency which, with lack of intervention, can result in failures [215, 216]. Intervention may be in the form of simple maintenance without need for replacement, such as lubrication to reduce friction in rotating plant [217], or cleaning to improve efficiency in heat exchanging components [218], extending the operational lifetime of components. However, long term data trends can provide important insight into the impacts of asset aging, which requires much more extensive monitoring, careful operation and detailed planning to justify continued operation or replacement [107]. Timeseries forecasting can facilitate such planning through providing predictions of the state of the plant at some future horizon based on previous observations.

Heat exchangers play a key role in the generation capability of nuclear power plants as their operating efficiency sets additional limitations on the reactor output. Without efficient heat exchange, the amount of generation the plant can produce may be limited, affecting production targets. Aging heat exchangers can significantly impact forecasted operational and maintenance costs of plants as additional planned outages may be required to conduct cleaning or other interventions to improve the performance of the asset. When considering plant operation in its final operational years, the effects of aging can be evident in the behaviour of assets which require close monitoring to continue meeting plant generation targets [85]. There are several methods in place to allow plants to operate efficiently despite the impact of aging, however this can lead to unnecessary intervention depending on the uncertainty in the asset monitoring process. Sensor calibration, noise and non-linear aging effects add uncertainty to the monitored state of an asset and should be accounted for in any maintenance planning processes to make best use of resources [219].

4.2.2 Timeseries forecasting

One option to support short term maintenance planning is data-based timeseries forecasting to provide an estimate of the state of the plant ahead of time [23], [220], [221]. This can provide warnings for when operational limits may be exceeded or may identify opportunities to increase production. Providing uncertainty alongside these forecasts captures the associated 'risk' in the estimate and allows the comparison of expected versus best- and worst-case scenarios to allow for more flexible planning [222].

Hierarchical models can allow for the capabilities of long-term forecasting models and short-term forecasting models to be combined to provide more accurate estimates of asset health [223], [224]. This can be achieved by training a data-based model on data spanning many operational years to fit a general trend to the assets aging process. The residuals of this model contain information which impacts the asset on a timescale of weeks to months, such as sensor noise, sensor calibration intervention or maintenance actions. This behaviour can then be learned by a secondary model, and its predictions can be used to calibrate the final estimates of asset behaviour. Providing uncertainty estimates alongside the expected value demonstrates the model's confidence in the prediction [225]. There are a wide range of candidate models for this process, however, the nuclear domain has an understandable preference for transparent modelling procedures [91]. Analytics which contain black-box models cannot be easily explained, and so any decisions based on these approaches cannot be easily actioned on in such a highly regulated environment. In this work we utilise a simple linear regression model to act as the transparent, long-term model candidate and several copula-based approaches to calibrate over the short-term.

4.3 Complex temperature timeseries dependency modelling with copulas

4.3.1 Data processing and hierarchical modelling structure

The hierarchical modelling structure with the training, validation and testing set paths are shown in Figure 4.1 The modelling process involves two linked models, where the base model is trained on a training set and tested on a validation set. The copula-based model is trained on the residuals of the base model's predictions on the validation set. Both models are then able to contribute to the predictions on a held out testing set. In this chapter, the base model is a linear regression trained on temperature and time, while the copula-based model is a selection of high dimensional copula models which are swapped out to investigate the limitations of key choices made on assumptions and complexity in the copula model design. It should be noted that, while linear regression is used as the base model, this can be interchanged with another model type, which may be more suitable to different applications.

Seven copula-based approaches are compared on three datasets, with varying levels of complexity and assumptions to cover several scenarios in model capability and marginal assumptions. These are grouped by model type, where the capability of multivariate Gaussian copula and two types of vine copula (regular and centre vine) will be demonstrated. Simplifications and increased flexibility on each models marginals will also be tested to demonstrate how much complexity may be necessary to capture the dependency structures within the case study data. Investigating different models and marginal assumptions explores the trade-off between simplified modelling strategies which are more easily managed, versus modelling strategies with increased complexity but enhanced flexibility. This process is designed to answer: "Is it worth it to design, manage and maintain a complex modelling strategy to enhance predictive performance for our maintenance system?". All vine models allow non-linear dependency in various ways but will be compared with the same variety of marginal limitations as for the Multivariate Gaussian models. The seven model variations are:

1. **Multivariate Gaussian Copula**

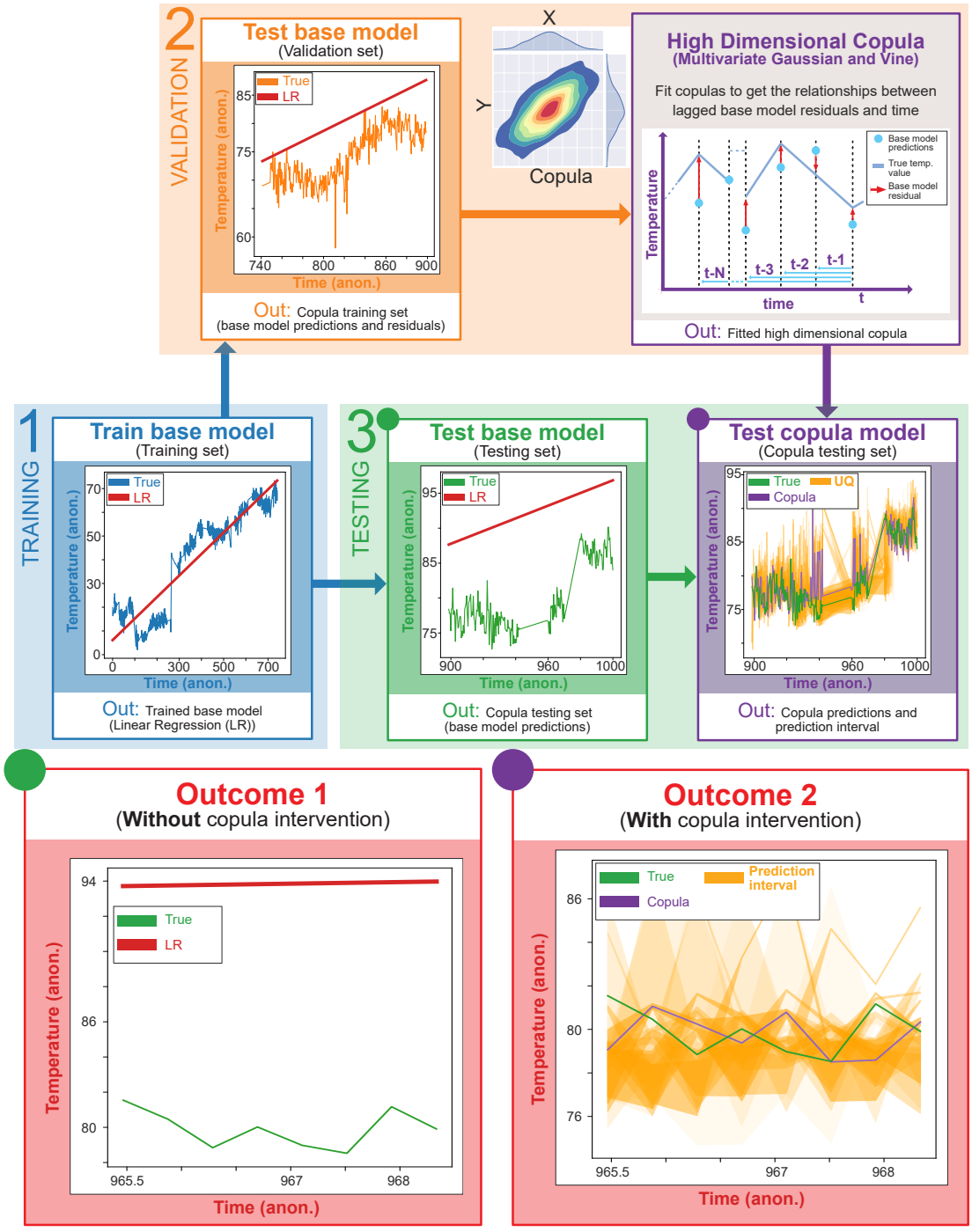

Figure 4.1: Summary of the process, and subset of example results presented in Chapter 4, for the industrial partner dataset. The base model used in this example is a Linear Regression (LR) and the chosen copula is the Centre Vine with Gaussian marginals.

(a) **Multivariate Gaussian Copula with Gaussian Marginals** (Simple variation): Showcasing the regular Multivariate Gaussian with no tail dependency and all Gaussian marginals. Tail dependency captures behaviours at the extremes which may represent high risk outcomes, such as large temperature fluctuations, which these model simplifications may not account for.

(b) **Multivariate Gaussian Copula with kernel density estimated marginals with Gaussian kernel** (Complex variation): Showing the potential value of heterogenous marginals which provides more flexibility. This method is non-parametric, which imparts minimal assumptions on the properties of the distribution to capture important detail in the data (for example, bi-modal structures) which may not be represented in common parametric methods, however, the accuracy can vary depending on sample size [128].

(c) **Multivariate Gaussian Copula with best fit marginals** (Complex variation, increased flexibility): Compares the fit of Gaussian, Beta, Gamma, kernel density estimate with Gaussian kernel, or truncated Gaussian univariate distributions. This expands the simple variation to demonstrate heterogenous marginals but with an additional selection of parametric distributions alongside the non-parametric option. This range of options is able to capture a wide variety of behaviour but results in higher modelling complexity.

2. **Vine Copula**

(a) **Regular Vine**

i. **Regular Vine with Gaussian Marginals** (Simple variation): Limited to Gaussian marginals only.

ii. **Regular Vine with best fit marginals** (Complex variation): Flexible marginal selection from the choice of Gaussian, Beta, Gamma, kernel density estimate with Gaussian kernel, or truncated Gaussian univariate distributions.

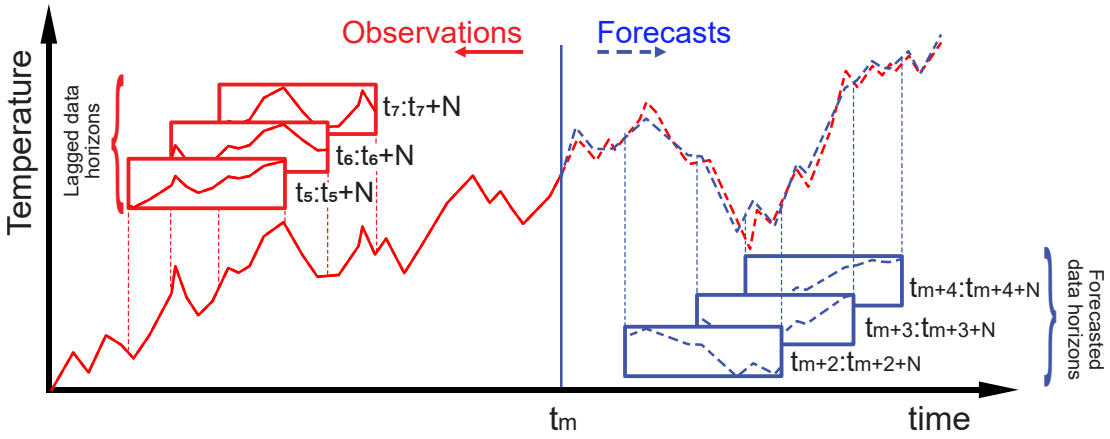

Figure 4.2: Diagram showing the method of using lagged data windows of N timesteps to train models to forecast up to N timesteps.

(b) **Centre Vine**

i. **Centre Vine with Gaussian Marginals** (Simple variation) : Limited to Gaussian marginals only.

ii. **Centre Vine with best fit marginals** (Complex variation): Flexible marginal selection from the choice of Gaussian, Beta, Gamma, kernel density estimate with Gaussian kernel, or truncated Gaussian univariate distributions.

For each dataset the copula models are trained on the machine learning residuals from a validation set. The error residuals are lagged from e_t to e_{t-N}, where N is the prediction horizon as described in Figure 4.2. This creates an $m \times (N+1)$ matrix with $m = L - N$, where L is the length of the validation matrix, to account for missing data due to the lags. The different Multivariate Gaussian copula and Vine models are trained on the $m \times (N+1)$ lagged error data. The autocorrelation plots of the data used to train the copula models are shown in Appendix D. The data space is scaled to a range of [0,1] and histogram equalisation used to create uniform marginals for compatibility with the copula models. The scaling parameters required to complete this transform are collected to enable values returned from the copula models to be scaled back into the original data space to retain their original engineering context (e.g

to allow for the presentation of temperatures in real values).

Predictions are computed on the testing set by providing the last known error between the target timeseries and the machine learning prediction to the trained copula-based models to predict the most likely correction value for timesteps up to a horizon of $N-1$. This is similar to the cross-validation of autoregressive models provided by [226]. Conditional relationships are used as the prediction window is stepped through the testing data set, where predictions from longer horizons can be used to inform the newest updated prediction as shown in Figure 4.2. This provides N predictions per timestep with varying ranges. The cumulative distribution (CD) value is computed for the possible correction values with values corresponding to < 0.05 and > 0.95 used to provide an upper and lower uncertainty estimate at each timestep. The predicted value of e_{t+1} is taken at each e_t as the prediction factors in the conditional behaviour across the full prediction horizon, N. The copula confidence bounds provide an estimated upper and lower temperature correction which are converted into a prediction interval by adding to the machine learning outputs. This can be interpreted as best and worst case risk for different maintenance scenarios. An example of a simplified case for a good and poor performing timeseries is shown in Figure 4.3. The poor performing timeseries plot shows that the example model predictions do not capture the true behaviour well, and the prediction intervals are wide which does not provide useful information. This behaviour will be assessed in the case studies of this chapter to distinguish models which have performed well, or poorly.

To evaluate model performance, three metrics are chosen: mean absolute error (MAE), continuous rank probability score (CRPS) and interval score. The MAE and percentage improvement in MAE are used to generalise the model performance over the timeseries and provides a direct point of comparison to the benchmark case where only the machine learning model predictions are used.

The Continuous Rank Probability Score (CRPS) [227], is the generalisation of the MAE which provides a comparable metric for the evaluation of probabilistic forecasts rather than point forecasts. When the CRPS is arranged in negative orientation, as in equation 4.1, it provides results in the same unit as the data space for direct comparisons

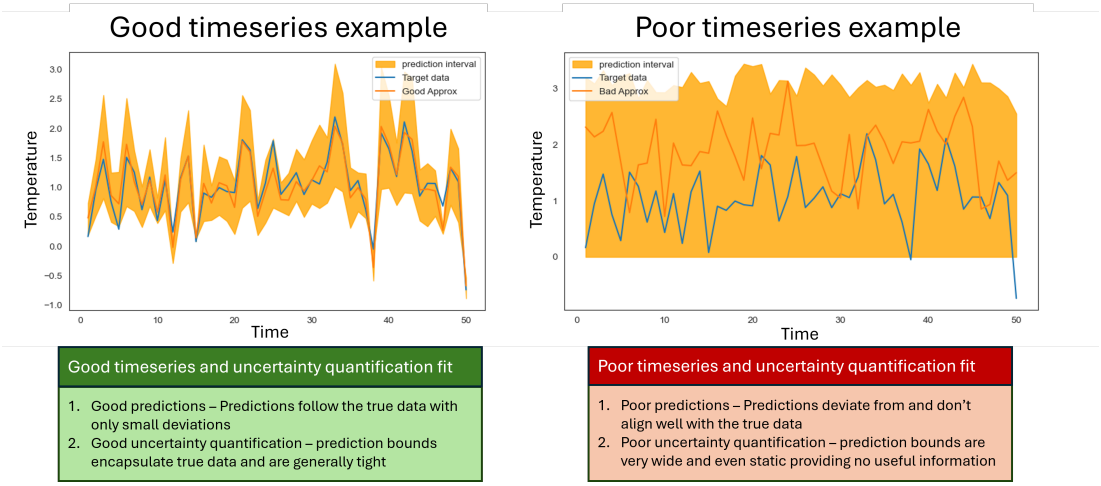

Figure 4.3: Example timeseries plot showing poor and good performing example models. The prediction bounds should encapsulate the true data and be as small as possible to provide useful risk information about the prediction.

[228]. In general terms, as the CRPS provides the MAE for probabilistic rather than point forecasts, lower values are preferred.

$$CRPS^*(F,x) = E_F|X - x| - \frac{1}{2}E_F|X - X'| \tag{4.1}$$

Where $CRPS*(F,x)$ is the negative orientation of the CRPS, F is the distribution function being evaluated at observations x, E_F is the expected value, or mean function [125] and X and X' are two independent random variables drawn from the distribution function, F.

The interval score [228] evaluates the effectiveness of the upper and lower predicted quantiles of interval forecasts. This is applied here to the upper and lower 90 % uncertainty bounds of the copula predicted corrections. For all true observations, x, the interval score, S^{int}, is evaluated using the forecasted upper, u, and lower, l, limits at a quoted confidence level $(1-\alpha)\times 100\%$. For a 90 % confidence interval, $\alpha = 0.1$. The interval score is calculated by:

$$\begin{aligned} S_\alpha^{int}(l,u;x) =& (u-l) + \frac{2}{\alpha}(l-x)\mathbb{1}\{x<l\} \\ &+ \frac{2}{\alpha}(x-u)\mathbb{1}\{x>u\} \end{aligned} \tag{4.2}$$

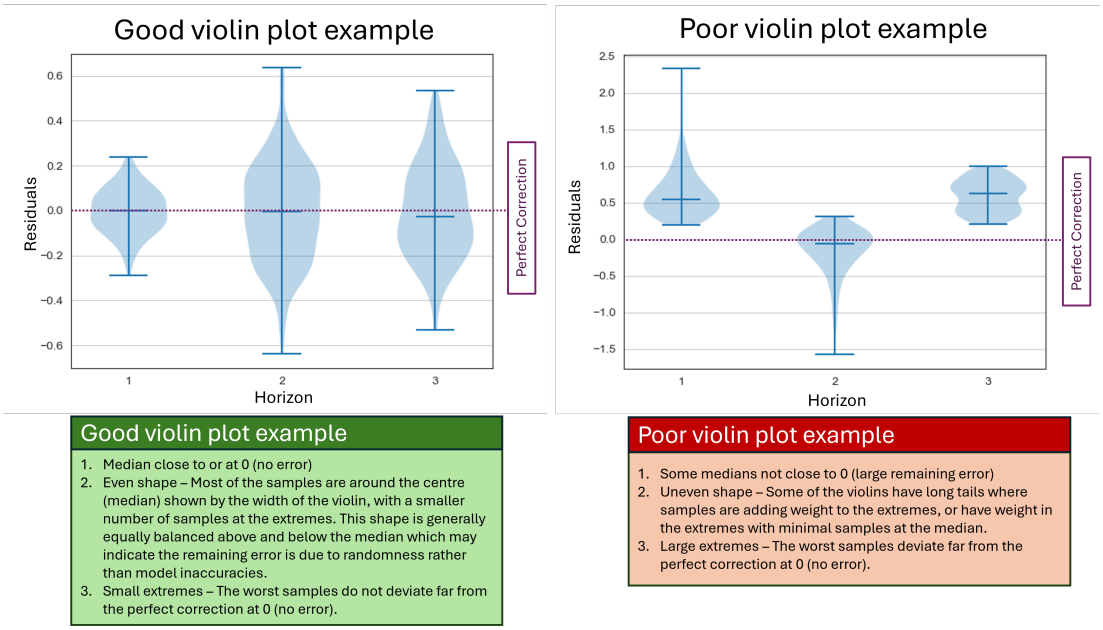

Figure 4.4: Example depiction of good and poor results on a violin plot

Where $\mathbb{1}\{condition\}$ defines the indicator function which takes a value of 1 if the condition is met, and 0 otherwise. The indicator functions apply an additional penalty where the bounds are violated by the true observation, with better forecasts covering the true values with small bounds.

The predictions and prediction bounds are also visually assessed with violin plots. An example case showing good or bad performance for a violin plot is shown in Figure 4.4. In this example case, the violin plots show residuals, whereby great performance is signified by the 0 on the Y axis. If most points align with this perfect prediction line, then the model predictions are high performing and generating minimal errors. The poor performing plot shows data with large remaining errors, where the median and much of the samples are not near 0. This can be used to compare how models perform across prediction horizons to identify forecast horizons which produce reliable, or unreliable results. It also allows comparison between models.

4.3.2 Case Study 1 - Synthetic data: Benchmarking case on Clayton copula synthetic data

Synthetic case study organisation and methodology

A benchmarking case based on synthetic data is presented to demonstrate the approach on an idealised case without the presence of unquantified operational noise and uncertainty sources. The synthetic dataset is comprised of two components: a target temperature timeseries and a machine learning approximation. The target temperature timeseries is comprised of a linear rising trend (with slope = 15 and intercept = 0.05), Gaussian noise (with $\mu = 0$ and $\sigma = 0.05$) and consecutive error terms sampled from a Clayton copula in range [0,1] with $\theta = 8$ (to preserve the lower tail relationship between lagged errors). The machine learning approximation is a linear rising trend with slope = 15 and intercept = 0.05. The machine learning approximation represents a simple linear model attempting to learn the data, where the resulting error relation between the target timeseries and the machine learning approximation is the copula samples and noise. In this case study, there is no requirement for a held out training set as our base model is a simple linear trend which does not require training. As such, the data is split into validation and testing set at a ratio of 75:25, with a total of 1000 samples. Figure 4.5 shows the target timeseries and the machine learning approximation from the validation and testing set of the synthetic data. The horizon, N, is chosen as 5, and the lagged errors on the validation set are used to train the 7 copula models. The testing procedure covered in Section 4.3.1 is completed to obtain the N horizon predictions on each timestep.

Synthetic case study results

The MAE and percentage improvement in MAE for each copula-based correction method and the benchmark case of no corrections are shown for each dataset in Table 4.1 and Table 4.2, respectively. The MAE values shown in Table 4.1 for the synthetic dataset show that all copula models reduce the error down to the approximate level of noise added to the signal. For Gaussian noise with $\sigma = 0.05$, 95.4% of samples will

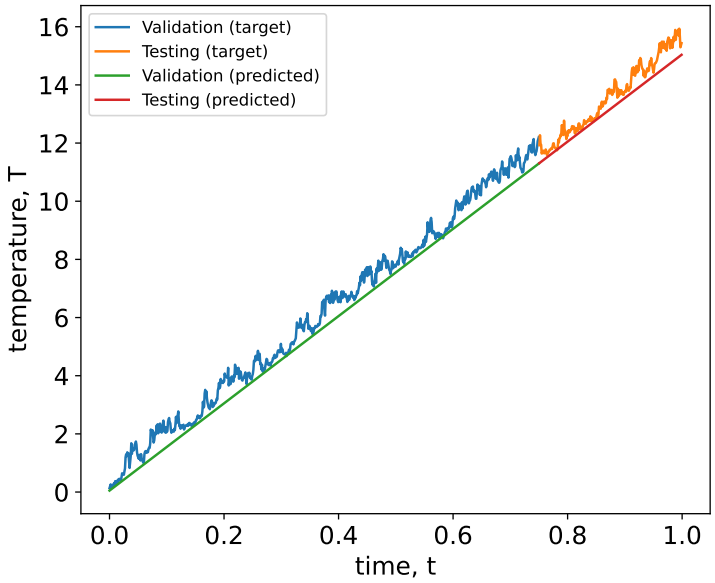

Figure 4.5: Validation and testing split of the synthetic timeseries, showing both the target signal and the linear trend representing the predictions of a simple base model attempting to learn the data.

Table 4.1: MAE for each copula-based correction method and the benchmark case of no corrections on each dataset. The lowest MAE for each dataset is highlighted in bold text.

Dataset /model	**Synthetic**	**Open Source**	**Industrial**
No Correction (Benchmark)	0.4188	5.5327	12.2061
Regular Vine (Gaussian)	0.2516	4.6166	2.9275
Regular Vine (Best fit)	0.2738	4.9614	3.5263
Centre Vine (Gaussian)	0.2080	4.2258	**2.1764**
Centre Vine (Best fit)	0.1785	**4.030**	3.0277
Multivariate Gaussian (Gaussian)	0.1642	5.041	2.8993
Multivariate Gaussian (KDE)	**0.1622**	4.3865	3.5687
Multivariate Gaussian (Best fit)	0.1755	4.1098	3.0339

Table 4.2: Percentage improvement over no corrections MAE for each copula-based correction method on each dataset. The largest percentage improvement for each dataset is highlighted in bold text.

Dataset /model	**Synthetic**	**Open Source**	**Industrial**
Regular Vine (Gaussian)	39.91	16.56	76.02
Regular Vine (Best fit)	34.61	10.33	71.11
Centre Vine (Gaussian)	50.34	23.62	**82.17**
Centre Vine (Best fit)	57.38	**27.16**	75.19
Multivariate Gaussian (Gaussian)	60.79	8.89	76.25
Multivariate Gaussian (KDE)	**61.27**	20.72	70.76
Multivariate Gaussian (Best fit)	58.09	25.72	75.14

be within $\pm 2\sigma$, which is roughly 20% of the scale of samples generated by the Clayton copula on a scale of [0,1]. The minimum improvement made by the copula approaches over the no corrections case shown in Table 4.2 is 39.9% by the Regular Vine with Gaussian marginals (RVG), and the maximum improvement is 61.3% by the Multivariate Gaussian copula with KDE marginals (MGK). For the three categories of models, the Gaussian marginals were the highest performing Regular Vine model (RVG), the best fit marginals were the highest performing Centre Vine model (CVB), and the KDE marginals were the highest performing Multivariate Gaussian model (MGK), based on MAE.

The target timeseries, machine learning approximation and corrected timeseries with uncertainty bounds given by the CD values for the testing set of the synthetic data are shown in Figure 4.6. In Figure 4.6, the Multivariate Gaussian (identified by MGG, MGK or MGB) and Centre Vine models (identified by CVG or CVB) are able to track the target signal (blue) with less noise in the chosen correction (red) and the prediction intervals (orange) than the Regular Vine models (titled RVG and RVB). In this case study, the largest barrier to the copula predictions is the Gaussian noise added to the target signal, in spite of which, the Multivariate Gaussian and Centre Vine models have performed well (both groups of models have a lower MAE than the Regular Vine group as provided in Table 4.1) which has also led to having narrower prediction intervals, showing increased confidence in the predictions compared to the Regular Vine group.

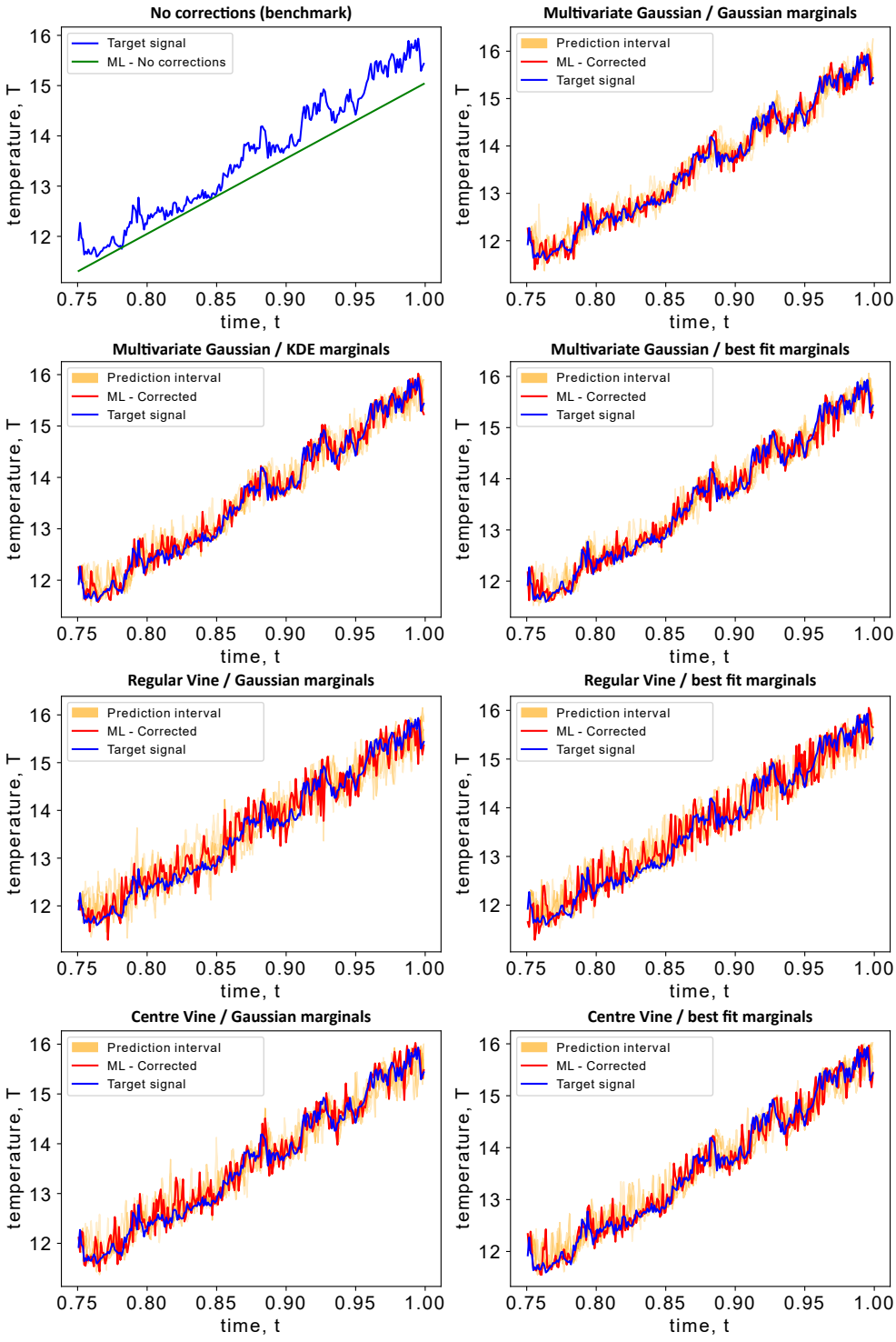

Figure 4.6: (Top left) Synthetic data timeseries of the target signal and the ML approximation for the testing set; and timeseries plots of the target data, copula corrected timeseries and prediction interval on the copula corrections.

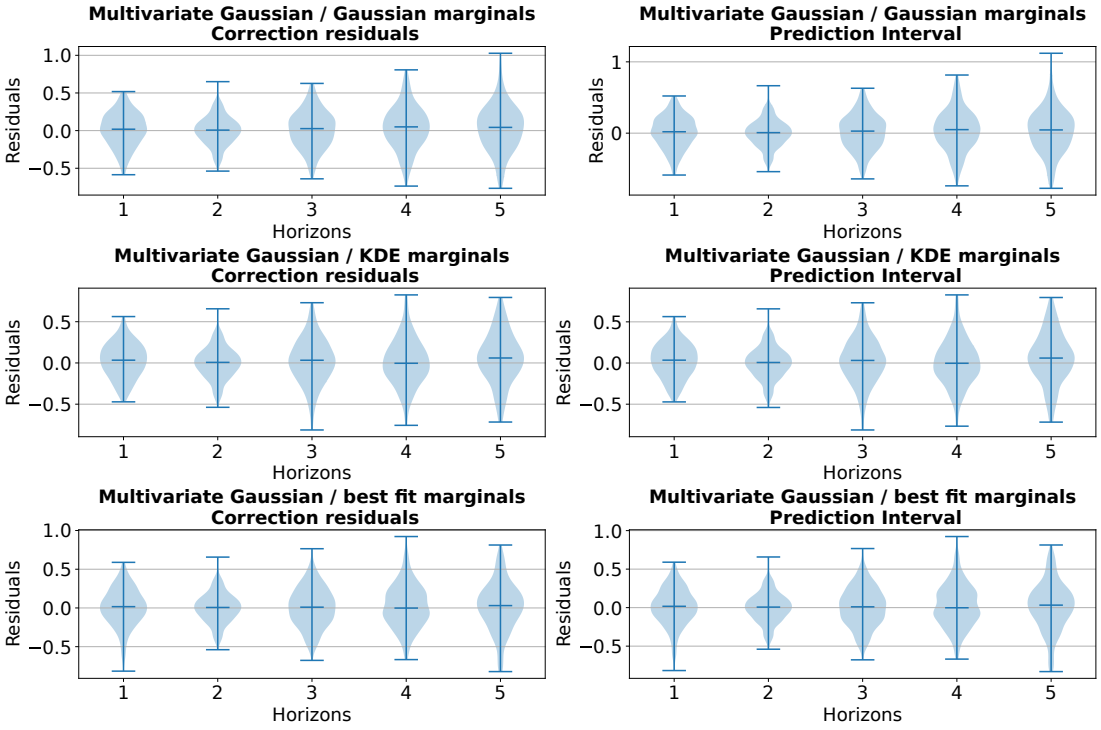

Figure 4.7: Violin plot of the $N = 5$ prediction horizons showing the residuals of the Multivariate Gaussian corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon on the synthetic data.

The residuals of the corrected timeseries (where a perfect correction results in 0 error) by prediction horizon, and the spread of the prediction interval are shown in the violin plots seperated by copula type. The Multivariate Gaussian violin plot is shown in Figure 4.7. The horizon CRPS and percentage improvement over the reference horizon 1 is shown for all models and horizons for the synthetic dataset in Table 4.3. At $N = 2$, all copula models across the Multivariate Gaussian, Regular Vine and Centre Vine groups have a very similar MAE at a resolution of 4 decimal places of 0.1497. The only deviation is CVG at 0.1498 which suggests that there would be additional distinction of more decimal places were included. All Multivariate Gaussian models have the largest CRPS at the longest horizon ($N = 5$) which is expected as this horizon has the least information. All models have improved CRPS at $N = 2$ over $N = 1$, which may be due to improved model generalisation if the closest horizon is encouraging the model

Table 4.3: CRPS and percentage change for each forecast horizon copula-based correction method and the benchmark case of no corrections for the synthetic dataset. Horizon 1 is the reference for the percentage change with positive percentage changes as improvements.

Model/ Metric	MGG	MGK	MGB	RVG	RVB	CVG	CVB
Horizon 1 CRPS	0.1829	0.174	0.1905	0.2606	0.2844	0.2251	0.1923
Horizon 1 % improv.	0	0	0	0	0	0	0
Horizon 2 CRPS	0.1497	0.1497	0.1497	0.1497	0.1497	0.1498	0.1497
Horizon 2 % improv.	18.17	13.97	21.44	42.55	47.36	33.47	22.16
Horizon 3 CRPS	0.202	0.217	0.2065	0.2519	0.2596	0.2493	0.2428
Horizon 3 % improv.	-10.43	-24.67	-8.42	3.31	8.72	-10.73	-26.25
Horizon 4 CRPS	0.2299	0.221	0.2226	0.3388	0.3401	0.2296	0.2406
Horizon 4 % improv.	-25.71	-27.0	-16.84	-30.02	-19.55	-1.98	-25.08
Horizon 5 CRPS	0.243	0.2477	0.2431	0.3469	0.3472	0.2666	0.2575
Horizon 5 % improv.	-32.88	-42.34	-27.62	-33.12	-22.07	-18.45	-33.89

The model acronyms are MGG - Multivariate Gaussian (Gaussian), MGK - Multivariate Gaussian (KDE), MGB - Multivariate Gaussian (Best fit), RVG - Regular Vine (Gaussian), RVB - Regular Vine (Best fit), CVG - Centre Vine (Gaussian) and CVB - Centre Vine (Best fit).

to overfit to previously observed behaviour. The smallest extremes for the MGG and MGK models are at $N = 1$, and $N = 2$ for the MGB model. This trend matches the uncertainty bounds of each model, showing that model confidence matches the model performance. Based on prediction horizon behaviour, MGK copula is the best for shorter horizons, while the MGG model is the best performing of the MG models at the furthest horizon. The violin plots for the Regular Vine models is shown in 4.8. The Regular Vine models tend to overestimate the value of the target signal as shown in Figure 4.8, where the median value increases above 0 for farther out prediction horizons, where a value of 0 is a perfect correction. The Regular Vine models also have the largest spread of residuals after correction for closer prediction horizons where a higher accuracy would be expected, which is visible in the corrected timeseries as it often over and under estimates the target signal. For both Regular Vine models, the highest (worst) CRPS is at the longest horizon, $N = 5$, while both models improve at $N = 2$ and slightly at $N = 3$. The smallest extreme residuals for both Regular Vine models are at $N = 2$ with this matching the models uncertainty bounds on the correction. This would suggest the model can accurately anticipate areas of higher

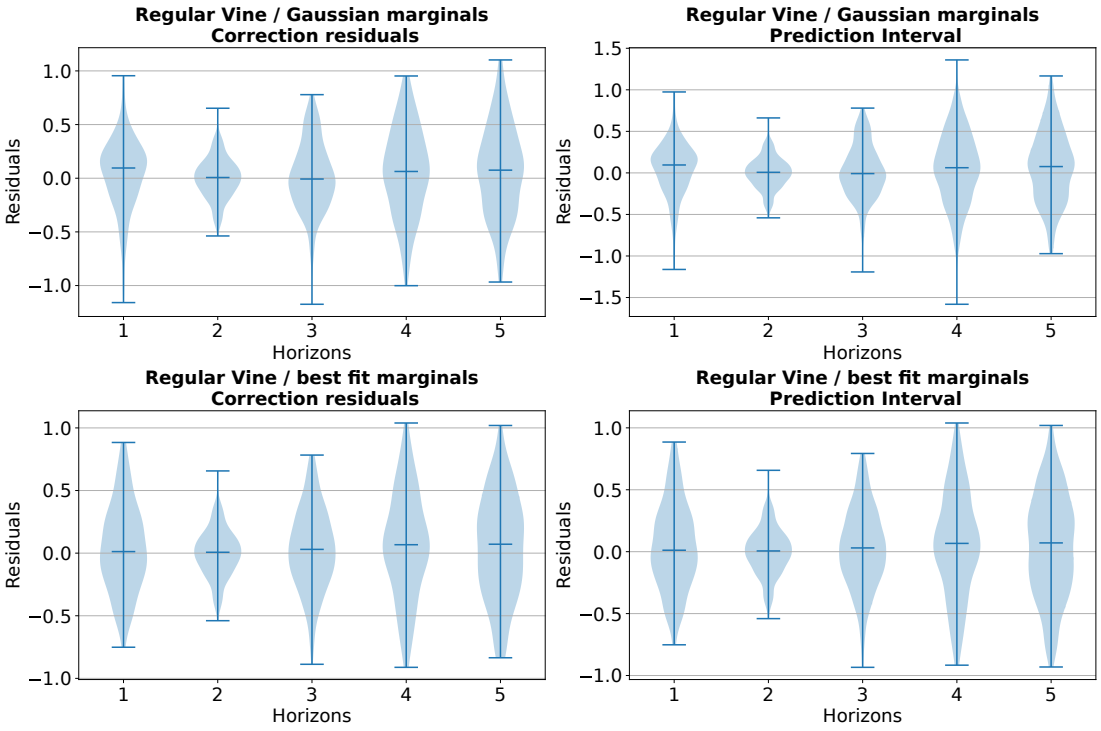

Figure 4.8: Violin plot of the $N = 5$ prediction horizons showing the residuals of the Regular Vine corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon on the synthetic data.

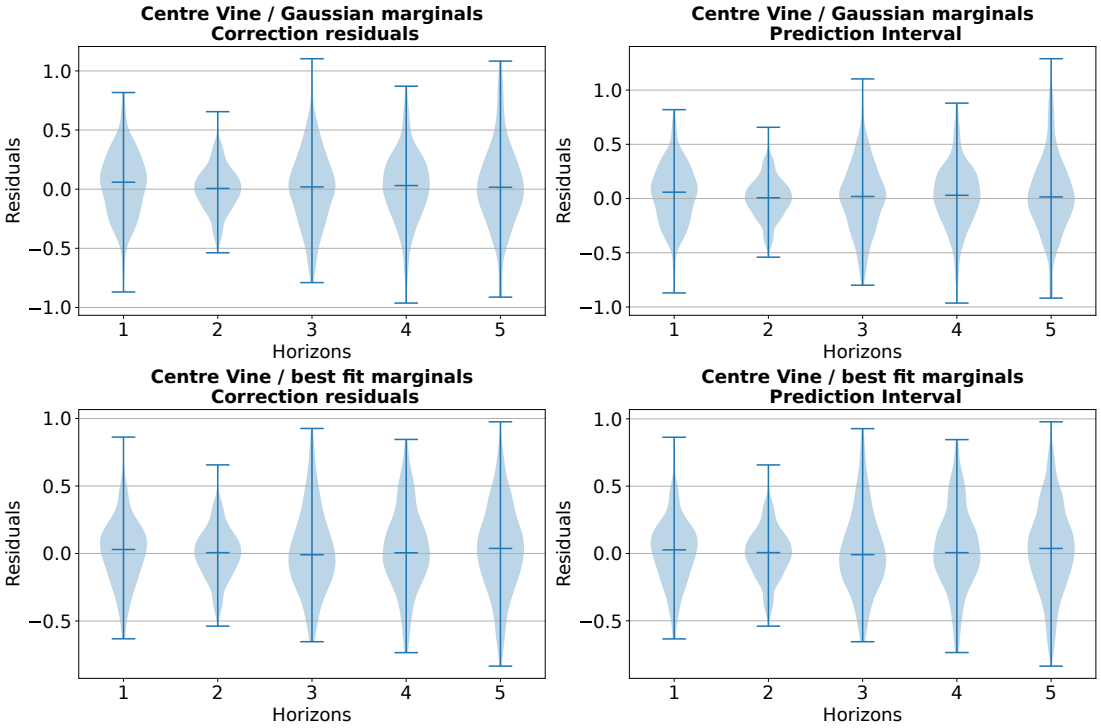

Figure 4.9: Violin plot of the $N = 5$ prediction horizons showing the residuals of the Centre Vine corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon on the synthetic data.

uncertainty, which follows for the RVB model with the largest uncertainty bounds at $N = 4$ and the largest extreme residuals at $N = 4$. However, for the RVG model, the largest uncertainty bound extremes are at $N = 4$ while the largest extreme residuals are at $N = 3$, suggesting poorer uncertainty estimation. The Centre Vine violin plot is shown in Figure 4.9. As with previous copula model groups, the worst CRPS is at the furthest horizon ($N = 5$) and the best CRPS values are at $N = 2$, rather than $N = 1$ (the closest horizon). The smallest extreme residuals follow the same trend as the CRPS, with the smallest extreme residuals for both CVG and CVB model at $N = 2$. The narrowest uncertainty bound also occurs at $N = 2$, showing the model confidence is aligned with its performance. Similarly, the largest extremes are once again at $N = 5$, where the largest uncertainty bounds on the correction occur. Overall, for the Centre Vine group, the CVB model has the lowest CRPS across most horizons, covering a decent amount of close and far horizons.

Table 4.4: CRPS for corrections over different horizons showing the mean (μ) and standard deviation (σ) for each model, the minimum CRPS, maximum CPRS and their associated horizons.

Dataset	Synthetic			Open Source			Industrial		
Model	$\mu \pm \sigma$	minimum (horizon)	maximum (horizon)	$\mu \pm \sigma$	minimum (horizon)	maximum (horizon)	$\mu \pm \sigma$	minimum (horizon)	maximum (horizon)
MGG[1]	0.2015 ± 0.0334	0.1497(2)	0.243(5)	5.8211 ± 0.6167	4.6183(2)	6.5989(4)	3.6409 ± 0.5954	2.0842(2)	4.2552(7)
MGK[2]	0.2019 ± 0.0352	0.1497(2)	0.2477(5)	5.6903 ± 0.7449	4.3865(1)	6.8799(4)	3.898 ± 0.6317	1.8188(2)	4.4512(15)
MGB[3]	0.2025 ± 0.0316	0.1497(2)	0.2431(5)	5.236 ± 0.5779	4.1098(1)	6.001(8)	3.7214 ± 0.7118	1.7309(2)	4.5808(12)
RVG[4]	0.2696 ± 0.0715	0.1497(2)	0.3469(5)	6.0315 ± 0.8703	4.5783(2)	6.8688(5)	4.367 ± 1.0339	1.7375(2)	5.4438(9)
RVB[5]	0.2762 ± 0.0714	0.1497(2)	0.3472(5)	5.8636 ± 0.6781	4.623(2)	6.547(8)	5.1002 ± 1.1499	1.7881(2)	6.0263(11)
CVG[6]	0.2241 ± 0.04	0.1498(2)	0.2666(5)	5.7587 ± 0.8109	4.2258(1)	6.5333(5)	2.6104 ± 0.284	1.8126(2)	2.9193(15)
CVB[7]	0.2166 ± 0.04	0.1497(2)	0.2575(5)	5.4988 ± 0.7332	4.0303(1)	6.188(4)	2.9939 ± 0.3468	1.8103(2)	3.352(13)

[1] Multivariate Gaussian with Gaussian marginals (MGG)
[2] Multivariate Gaussian with KDE marginals (MGK)
[3] Multivariate Gaussian with best fit marginals (MGB)
[4] Regular Vine with Gaussian marginals (RVG)
[5] Regular Vine with best fit marginals (RVB)
[6] Centre Vine with Gaussian marginals (CVG)
[7] Centre Vine with best fit marginals(CVB)

The CRPS for the suggested copula corrections for each model and the interval score for the 90 % uncertainty bounds on the copula correction are summarised in Table 4.4 and Table 4.5, respectively. The MGG model has the lowest mean CRPS over all horizons with the MGB model having the lowest standard deviation across all horizons, as shown in Table 4.4. The lower standard deviation suggests that the performance of the models has lower variance across all horizons and will perform more consistently, however at a slightly higher CPRS value. The RVB model has the highest CRPS and the RVG model has the highest standard deviation, showing this model group has high variability across horizons and generally leaves higher residuals. All models have their maximum and minimum CRPS horizons occur at the same points, at $N = 5$ and $N = 2$, respectively. In this simplified example, this is expected as the furthest horizon has the least information to narrow down potential behaviour while the horizon of $N = 2$ may allow the model to benefit from larger amounts of information while keeping good generalisation. With the interval score, the models are penalised whenever the true value falls outside the suggested interval and on the difference between the provided interval, with narrow intervals awarded. The Regular Vine models have the highest interval score at 5.4099 for the RVB model, as shown in Table 4.5, which means the 90

Table 4.5: Interval score for all model 5 % and 95 % uncertainty bounds on the copula correction, showing the mean (μ) and standard deviation (σ) for each model, the minimum interval score, maximum interval score and their associated horizons.

Dataset	**Synthetic**			**Open Source**			**Industrial**		
Model	$\mu \pm \sigma$	minimum (horizon)	maximum (horizon)	$\mu \pm \sigma$	minimum (horizon)	maximum (horizon)	$\mu \pm \sigma$	minimum (horizon)	maximum (horizon)
MGG[1]	**3.8972 ± 3.1349**	**2.9315 (2)**	**4.7685 (5)**	74.4523 ± 66.0736	62.0659 (2)	82.5907 (7)	75.846 ± 52.567	54.5494 (2)	87.9467 (7)
MGK[2]	3.9419 ± 3.2756	2.9379 (2)	4.8548 (5)	76.4303 ± 68.1522	59.9617 (2)	85.9058 (8)	81.1287 ± 56.2551	47.347 (2)	90.698 (15)
MGB[3]	3.9732 ± 3.2528	2.9383 (2)	4.7956 (5)	**70.8578 ± 66.5586**	61.0319 (2)	87.6661 (8)	78.1224 ± 54.2014	46.33 (2)	92.8979 (12)
RVG[4]	5.1428 ± 4.2919	2.9428 (2)	6.7423 (5)	78.3235 ± 66.4062	61.81 (2)	88.4969 (7)	89.1464 ± 66.8781	39.8679 (2)	109.9289 (9)
RVB[5]	5.4099 ± 4.3648	2.9537 (2)	6.7547 (5)	75.6974 ± 65.2588	58.742 (2)	86.6243 (8)	102.7755 ± 72.2326	**37.7329 (2)**	121.6813 (11)
CVG[6]	4.3391 ± 3.7619	2.9393 (2)	5.1442 (5)	72.0799 ± 62.18	61.9277 (2)	**81.0617 (5)**	**58.3631 ± 49.2017**	44.3172 (2)	**65.476 (15)**
CVB[7]	4.2373 ± 3.5929	2.9421 (2)	5.0459 (5)	70.9053 ± 65.0337	**58.7243 (1)**	89.4956 (8)	62.9259 ± 50.8054	39.4543 (2)	73.0536 (1)

[1] Multivariate Gaussian with Gaussian marginals (MGG)
[2] Multivariate Gaussian with KDE marginals (MGK)
[3] Multivariate Gaussian with best fit marginals (MGB)
[4] Regular Vine with Gaussian marginals (RVG)
[5] Regular Vine with best fit marginals (RVB)
[6] Centre Vine with Gaussian marginals (CVG)
[7] Centre Vine with best fit marginals(CVB)

% uncertainty bounds on the correction were not as accurate as for other models, such as the MGG copula model which had the lowest score at 3.8972. For all models, the uncertainty bounds on the $N = 2$ horizon produced the lowest interval score (better bounding on the true prediction) and all models had the highest interval score at the largest horizon $N = 5$, showing that they provide better uncertainty estimates for closer horizons which have been calibrated using additional information gained over more forecast horizons. The CRPS and interval score behaviour aligns, showing the models are generally capable of providing uncertainty quantification which aligns with their performance in practice.

The changes in histogram shape from the benchmark case in Figure 4.10 show that all copula models have moved the residuals to be centred around 0 and are unimodal in shape, which has improved upon the "no corrections" case. The "no corrections" case entirely underestimates the value of the target signal (as designed), and all models can recover this shift. The copula correction histograms seem to have improved the Normality of the 'no corrections' case, but to confirm this observation, the skewness and kurtosis measurements of each distribution are shown in Table 4.6. If the model

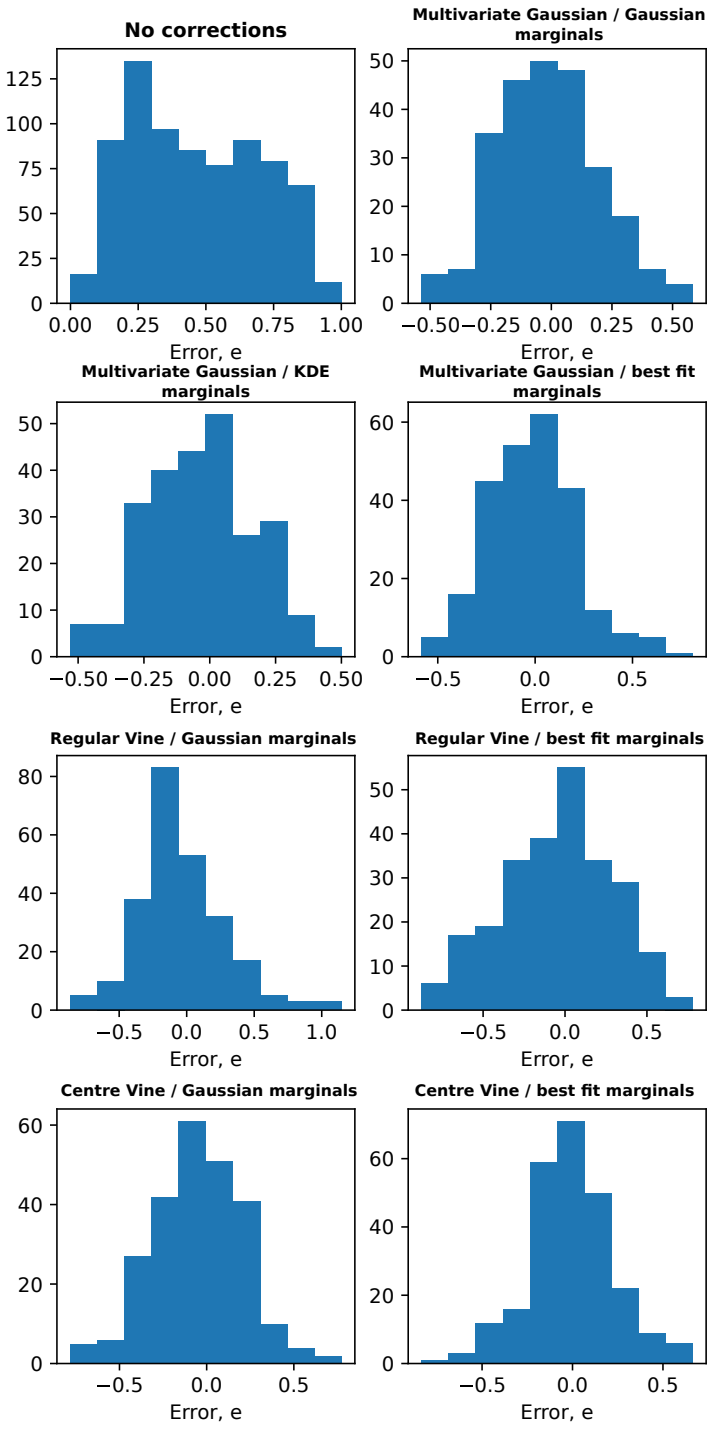

Figure 4.10: Residual histograms for the seven copula correction methods and the benchmark case of no copula corrections for the synthetic dataset.

Model	Kurtosis	Skewness
No Corrections	-0.7558	**0.6183**
Multivariate Gaussian / Gaussian marginals	*-0.1205*	0.2674
Multivariate Gaussian / KDE marginals	-0.4336	0.0451
Multivariate Gaussian / best fit marginals	0.6107	0.4071
Regular Vine / Gaussian marginals	**1.1163**	0.6175
Regular Vine / best fit marginals	-0.4841	-0.1652
Centre Vine / Gaussian marginals	0.2351	*0.0155*
Centre Vine / best fit marginals	0.6482	-0.0182

Table 4.6: Skewness and kurtosis values for the synthetic data model residual histograms. The values for Gaussian distributions are 0 for kurtosis and skewness, as reference. The values furthest from 0 are shown in bold while those closest to 0 are in italics.

residuals are Gaussian and centred around 0, this implies that much of the relationship in the data has been captured, except for random noise. For the distributions to be Gaussian, the value of skewness and kurtosis would be 0. The largest value (furthest from 0) for kurtosis is the RVG model, and the largest skewness is the 'no corrections' case. This means all models improved the skewness value, moving the residuals to more Gaussian behaviour. However, for kurtosis, the RVG model deteriorated that of the 'no corrections' case. The lowest kurtosis value is from the MGG model at -0.1205, and lowest skewness is from the CVG model at 0.0155. All models, except the RVG model, improve on both the skewness and kurtosis of the 'no corrections' case, suggesting they have accounted for some remaining information available in the base model residuals.

Quantile plots (Q-Q Plots) are a visual technique used to compare samples or distributions [229]. Identical distributions would result in a linear line intercepting 0 due to aligned quantile values. As shown in the Q-Q plot in Figure 4.11, the Multivariate Gaussian copula and Centre Vine models match the quantiles of the target signal more accurately than the Regular Vine models which have larger deviations in the lower quantiles, where it overestimates the values. This is also visible in the timeseries plot in Figure 4.6 where the Multivariate Gaussian and Centre vine models track the target signal with less noise in the chosen correction (red) and the prediction intervals (orange) than the Regular Vine models. For this case study, the chosen data is meant to provide a benchmark which is designed to be relatively simple for the models to fit to,

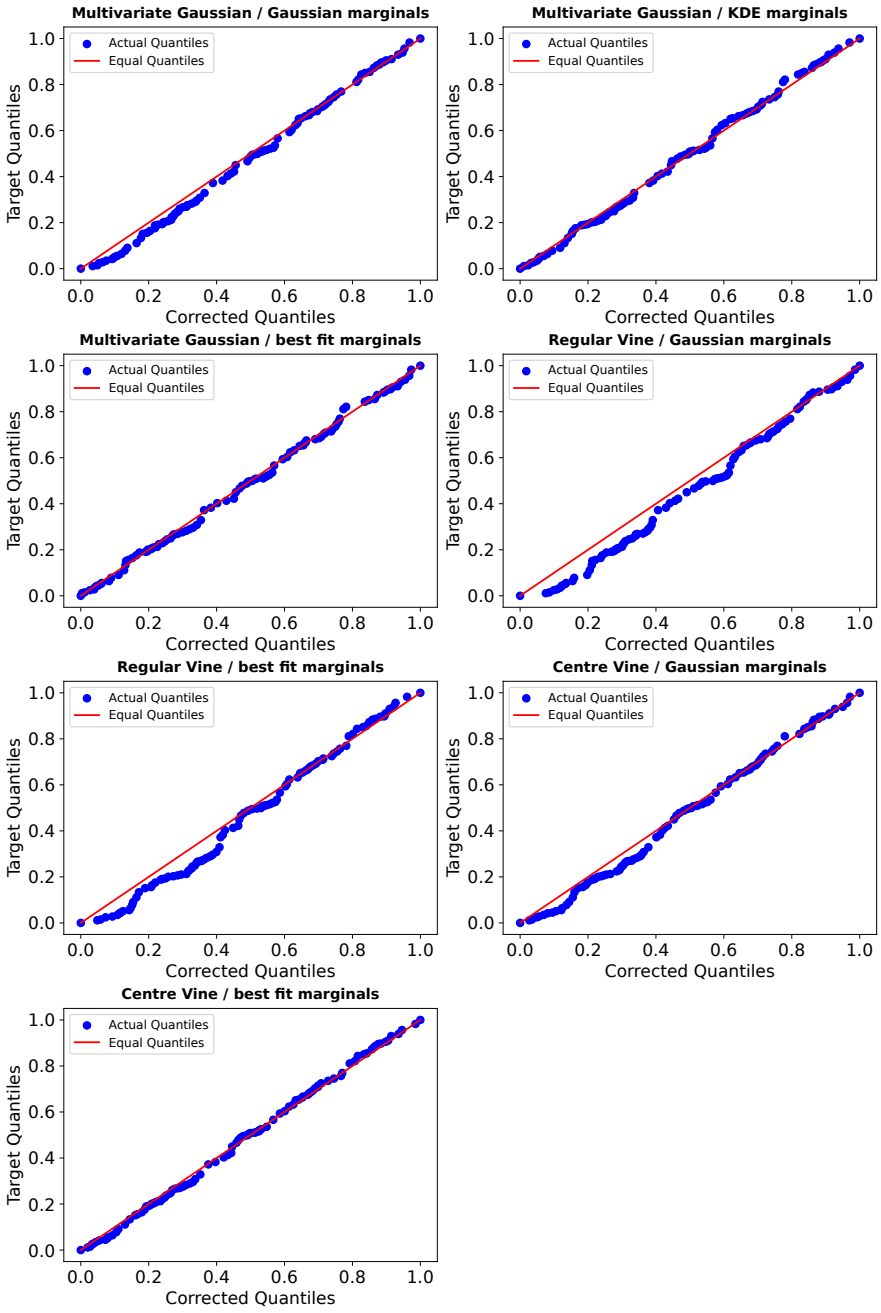

Figure 4.11: Quantile-quantile plot of the target signal quantiles against the corrected signal quantiles for the synthetic dataset. Identical distributions result in a straight line.

and so the successful models' predictions being able to match the target distribution is expected. However, even in this simplified example, the Regular Vines are lacking in ability at lower quantiles.

The copula fitting shown in Figure 4.12 presents how the models capture the relationship between e_t to e_{t-1} which have been designed in this case study to specifically have the strongest relationship. For each model, 1000 samples are taken. The target relationship is shown in the first plot (top left) which is the form of a Clayton copula with Gaussian noise added. The Multivariate Gaussian copulas are based on the Gaussian copula which captures elliptical behaviour, which generalises well to the noisy Clayton. However, due to the constraints of the elliptical shape, the Multivariate Gaussian copulas capture a tighter relationship in the upper tail than is present in the target data. Additionally, the MGG model fails to capture the bi-modal structure in the target density. The RVG, and more so the RVB, covers a large area without capturing much detail. Between both models and in this selection of samples, the RVG model seems to have more density in the centre (around [0.5,0.5]) while the RVB is more dense in the lower tail, which is an important feature of Clayton behaviour. The CVG model seems to have taken a more conservative estimate of the target behaviour, without much distinction between the upper and lower tail behaviour, while the CVB has distinguished the tight lower tail relation.

4.3.3 Case Study 2 - Open source data: Wind turbine generator bearing temperature forecasting

Open source case study organisation and methodology

To demonstrate the method on operational data, an open source temperature timeseries dataset is chosen comprising of the rear bearing temperature of a turbine generator on Penmanshiel wind farm [230] located in Scotland. Bearings are a common point of failure in rotating plant due to stress and wear during operation, and are found across a vast range of industrial applications [173]. As per manufacturer advice, temperature and vibration measurements allow the health of bearings in key assets to be monitored [231], to provide advanced warning of early signs of degradation or faults. The

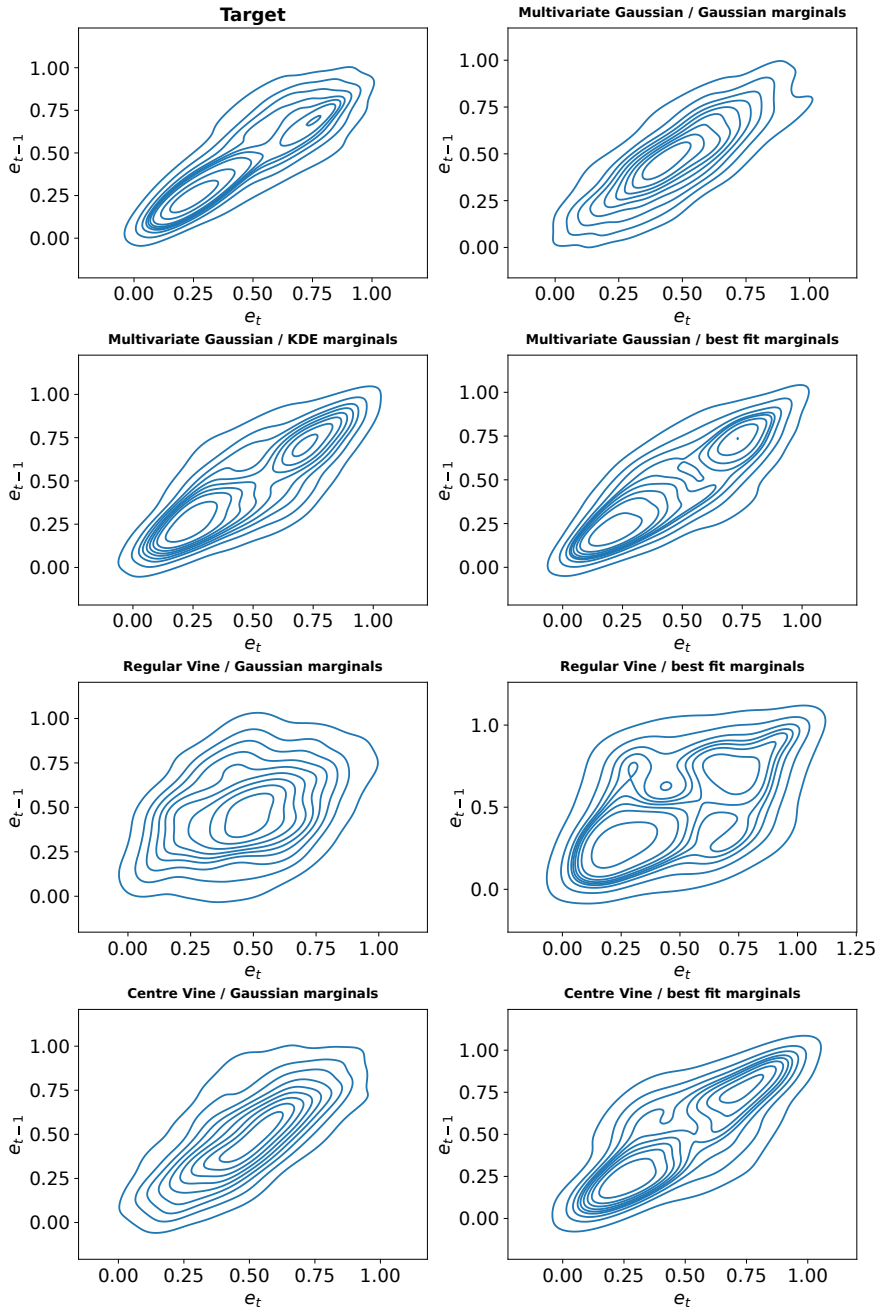

Figure 4.12: Synthetic dataset relationship between e_t to e_{t-1} for the target data and sampled copulas.

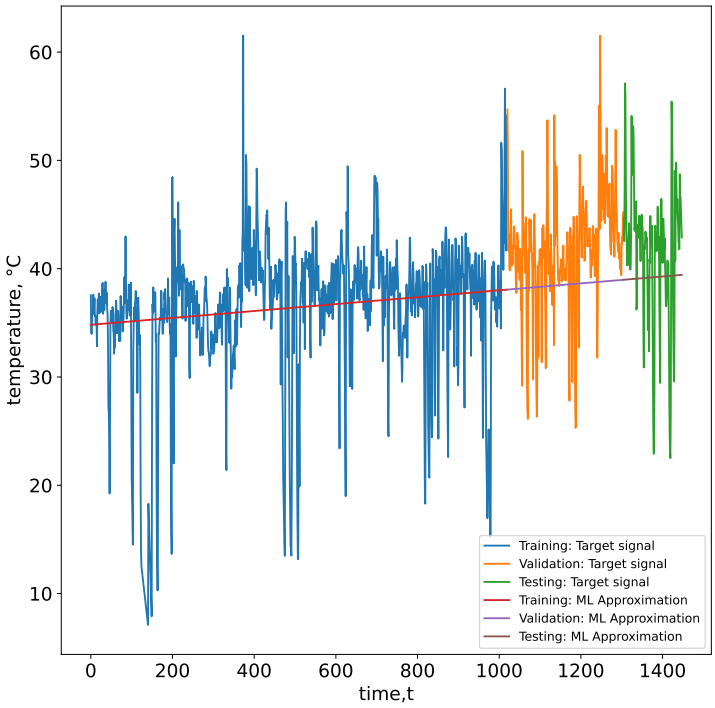

Figure 4.13: Training, validation and testing split for the Penmanshiel wind turbine rear generator bearing temperature. The data is 3 hourly from 01/01/2021 to 30/06/2021.

turbine generator is an important generation asset located in an environment where operational stress and ambient temperatures may cause the bearing temperatures to reach unacceptable limits. The Penmanshiel's wind farm timeseries runs from 1st January 2021 to the 30th June 2021, with data collected every 10 minutes. To permit day ahead forecasting, the data is processed to contain data collected every 3 hours, on the hour, with a day ahead horizon, N, chosen as 8. A linear regression model is used as the long-term prediction model and the data is split into training, validation, and testing sets at a ratio of 70:20:10 with 1426 total samples. The target signal and linear regression predictions are shown for the training, testing and validation sets in Figure 4.13.

The errors from the linear regression model predictions on the validation set are lagged from e_t to e_{t-N}, where $N = 8$, and used to train the copula models. Following the process stated in Section 4.3.1, the testing predictions and corrections are collected for each copula model.

Open source case study results

The recorded MAE and percentage improvement in MAE for each copula model are given in Table 4.1 and Table 4.2, respectively. The baseline improvements for all copula models are much lower than for the synthetic dataset, as shown by the MAE values in Table 4.1. This is expected as the synthetic dataset was designed to demonstrate the capability of the copula methods and is not operational data, however, all models do improve on the 'no corrections' case. The best improvement shown in Table 4.2 is 27.16 % by the CVB model, while the worst *improvement* is 8.89 % by the MGG model. For the three categories of models, the Gaussian marginals were the highest performing Regular Vine model, the best fit marginals were the highest performing Centre vine model, and the best fit marginals were the highest performing Multivariate Gaussian model, based on MAE. In this case, the more complex marginals allowed most categories to outperform those with simpler assumptions, with the KDE marginals supporting this by being the second best performing Multivariate Gaussian model based on MAE.

The target timeseries, linear regression prediction and corrected timeseries with uncertainty bounds given by the 90 % copula CDF values are shown in Figure 4.14. As shown in Figure 4.14, the target signal (blue) has several large peaks and troughs which the CVG and MGG model corrections (red) tend to underestimate compared to the more complex marginal models. This does not hold for the RVG model, which does seem to capture the troughs and peaks better than its more complicated counterpart, RVB. The CVB model does well to capture the large 'trough - peak - trough' pattern at the end of the timeseries, but also tends to predict such features when they do not appear, such as the peak in the latter half of the timeseries. The MGB and CVB models have the highest percentage MAE improvement compared to the 'no corrections' case and also have much smaller prediction intervals (orange) than their less complex marginal counterparts. The models that struggled more with this dataset have correspondingly wider prediction intervals.

The violin plot of the predictions and prediction interval over each horizon for the open source dataset are separated by model type. The Multivariate Gaussian models are shown in Figure 4.15. The horizon CRPS and percentage improvement over horizon

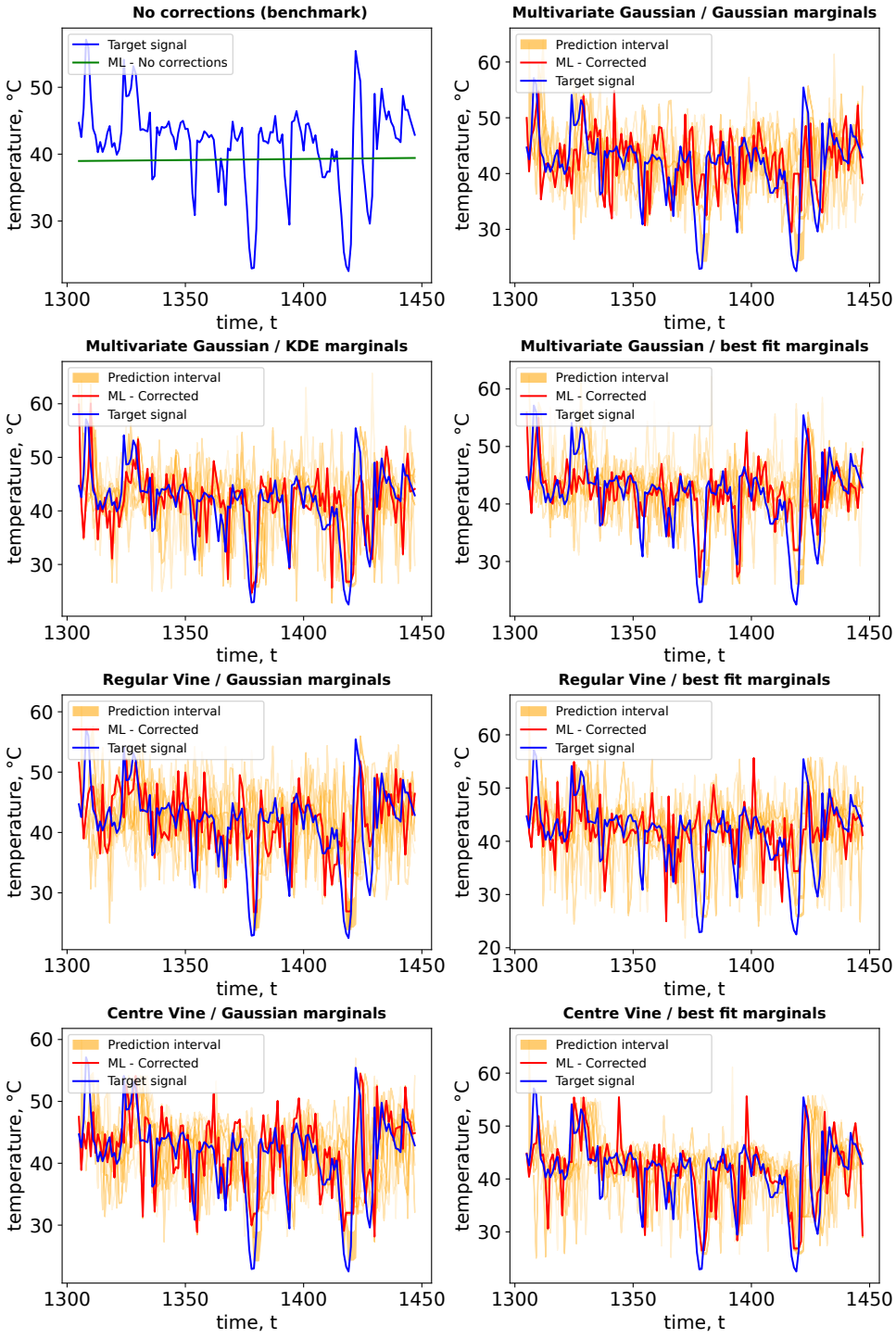

Figure 4.14: (Top left) Rear bearing generator temperature timeseries for Penmanshiel wind farm dataset; and timeseries plots of the testing data, corrected timeseries, prediction interval on the copula corrections for the 7 copula models on the open source wind turbine data.

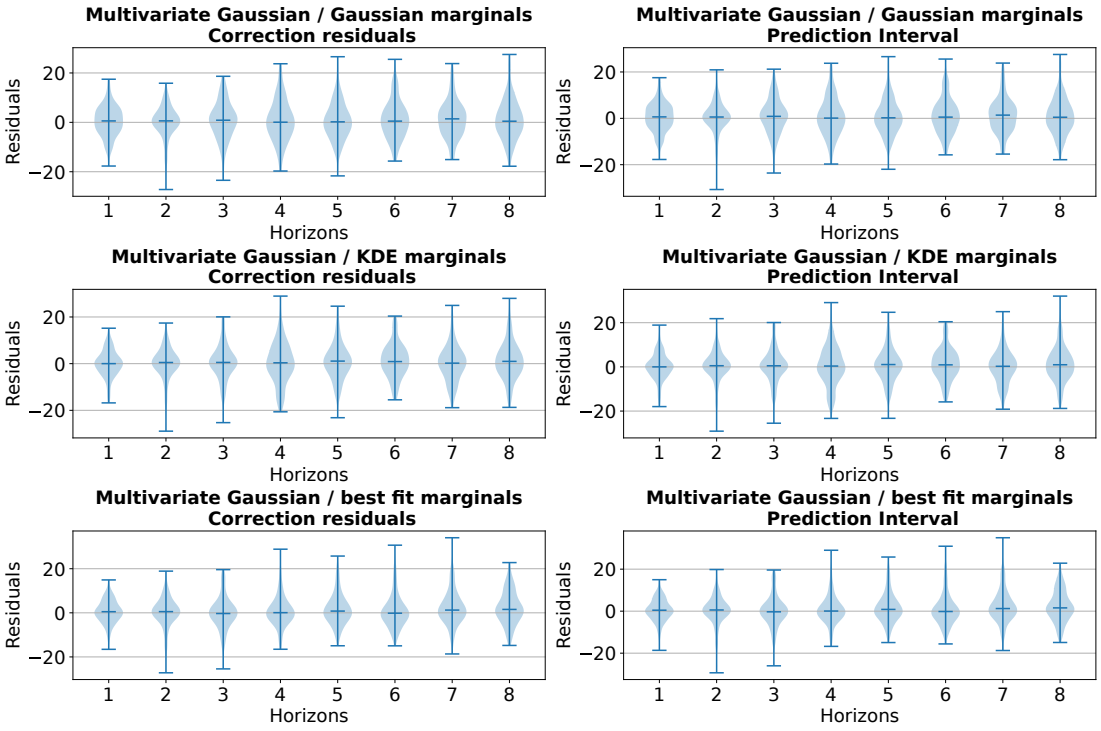

Figure 4.15: Violin plot of the $N = 8$ prediction horizons for the Multivariate Gaussian models on the open source wind turbine bearing data, showing the residuals of the linear regression model and corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon.

Table 4.7: CRPS and percentage change for each forecast horizon copula-based correction method and the benchmark case of no corrections for the open source wind turbine generator bearing dataset. Horizon 1 is the reference for the percentage change with positive percentage changes as improvements.

Model/ Metric	MGG	MGK	MGB	RVG	RVB	CVG	CVB
Horizon 1 CRPS	5.0408	4.3865	4.1098	4.6166	4.9614	4.2258	4.0303
Horizon 1 % improv.	0	0	0	0	0	0	0
Horizon 2 CRPS	4.6183	4.7419	4.7456	4.5783	4.623	4.5701	4.6219
Horizon 2 % improv.	8.38	-8.1	-15.47	0.83	6.82	-8.15	-14.68
Horizon 3 CRPS	6.003	5.5501	5.3454	6.1034	5.6124	5.8719	5.5292
Horizon 3 % improv.	-19.09	-26.53	-30.07	-32.2	-13.12	-38.95	-37.19
Horizon 4 CRPS	6.5989	6.8799	4.9579	6.1445	6.2297	6.2926	6.188
Horizon 4 % improv.	-30.91	-56.84	-20.64	-33.09	-25.56	-48.91	-53.54
Horizon 5 CRPS	5.9088	5.9711	5.3989	6.8688	6.2112	6.5333	5.8279
Horizon 5 % improv.	-17.22	-36.13	-31.37	-48.78	-25.19	-54.6	-44.6
Horizon 6 CRPS	6.0148	5.8614	5.448	6.6824	6.2144	6.3488	6.0916
Horizon 6 % improv.	-19.32	-33.62	-32.56	-44.75	-25.25	-50.24	-51.15
Horizon 7 CRPS	6.0744	5.9804	5.8815	6.8402	6.5097	6.096	5.5408
Horizon 7 % improv.	-20.5	-36.34	-43.11	-48.16	-31.21	-44.26	-37.48
Horizon 8 CRPS	6.3102	6.1515	6.001	6.4179	6.547	6.131	6.1607
Horizon 8 % improv.	-25.18	-40.24	-46.02	-39.02	-31.96	-45.09	-52.86

The model acronyms are MGG - Multivariate Gaussian (Gaussian), MGK - Multivariate Gaussian (KDE), MGB - Multivariate Gaussian (Best fit), RVG - Regular Vine (Gaussian), RVB - Regular Vine (Best fit), CVG - Centre Vine (Gaussian) and CVB - Centre Vine (Best fit).

1 is shown in Table 4.7. The CRPS slightly improves with predictions taken at $N = 2$ for MGG, but $N = 1$ remains the most successful forecasting horizon for MGK and MGB. The smallest extreme residuals for all Multivariate Gaussian models are at a horizon of $N = 1$, which is also the case for the uncertainty bounds for the MGG and MGB models, suggesting the models are more successful and more confident in their predictions at this range. The MGK model has the smallest uncertainty bound extremes at $N = 5$, not aligning with its CRPS performance. The largest extreme residuals are at $N = 5$ for the MGG model, $N = 4$ for the MGK model and $N = 7$ for the MGB. By being more successful at shorter horizons compared to longer, it suggests the models benefit greatly from the increased information of additional observations for this case.

For the Regular Vine models, the violin plots are shown in Figure 4.16. Both models

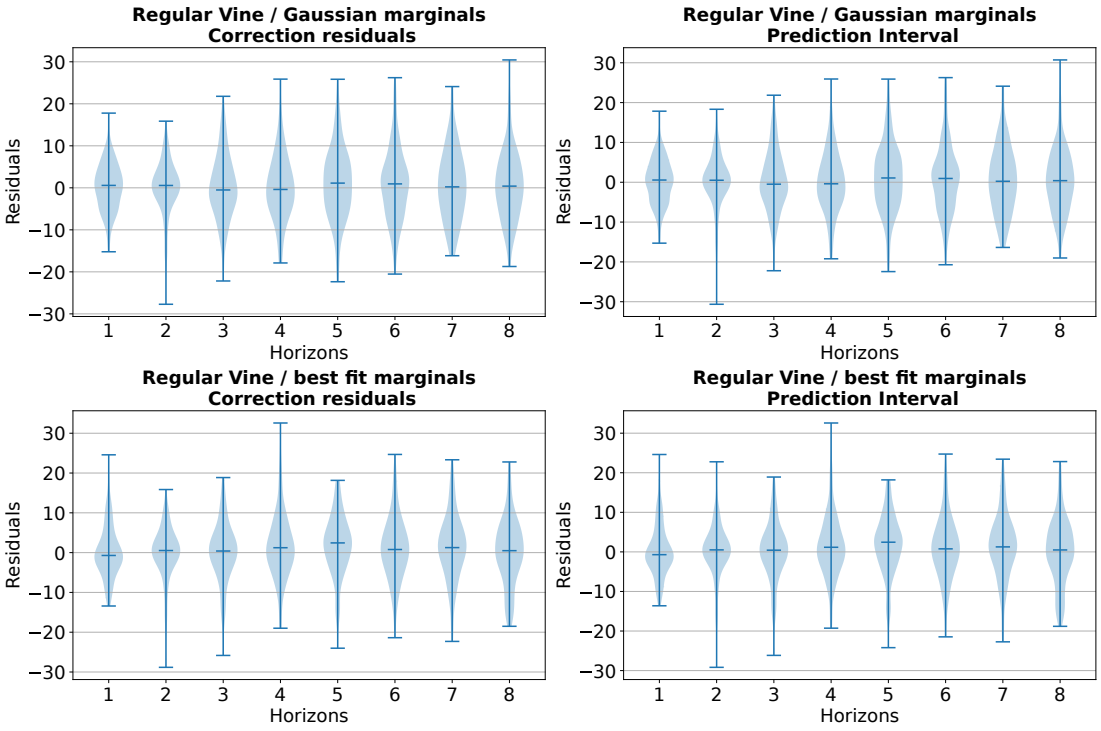

Figure 4.16: Violin plot of the $N = 8$ prediction horizons for the Regular Vine models on the open source wind turbine bearing data, showing the correction residuals against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon.

see slight improvement in CRPS moving to a different prediction horizons of $N = 2$. The worst prediction horizon for each model is $N = 5$ for RVG (15 hours ahead) and $N = 8$ for RVB (24 hours ahead). The reduced performance at the farthest horizon is expected due to the lack of any updated information, however the worst performance occurring at the middle horizon may be due to external factors in the data, such as the impact of environmental temperature fluctuations (for example, moving from day to night). This is especially of concern for exposed plant, such as wind turbines, who are more exposed to daily and seasonal temperature fluctuations. One component of future work which such cases could benefit from is the inclusion of other environmental variables in the prediction, such as weather variables. The most extreme residuals are at a horizon of $N = 8$ (24 hours ahead) for the RVG model and $N = 4$ (12 hours ahead) for the RVB model. This aligns with the largest extremes in the uncertainty bounds on the correction for the RVG model, but the largest uncertainty bound for the RVB model instead occurs at $N = 2$. Additionally, the smallest extreme residuals for both models on the correction and uncertainty bounds occur at $N = 1$. For both best and worst cases, the uncertainty bound aligns with the model performance for the RVG model, showing that the risk communicated by the uncertainty bounds reflect the state of the model performance.

The violin plot in Figure 4.17 shows the predictions and prediction intervals of the Centre Vine models over all horizons. For both Centre Vine models, the best MAE is at a horizon of $N = 1$ with the worst for CVG at $N = 5$ (15 hours ahead), and $N = 4$ (12 hours ahead). Both models struggle at far horizons, but this range of 12-15 hours ahead is particularly challenging for this model. For both models, the uncertainty quantification capability matches the model performance. For example, the models are more confident (narrower uncertainty bounds) at a horizon of $N = 1$ where the model performs the best in terms of minimising extreme residuals, while the largest bounds are at a horizon of $N = 4$ where the models experience the worst extreme residuals.

The interval score assesses the models uncertainty bounds over all horizons, penalising estimations where the required correction falls outwith the recommended interval and rewarding tighter, more accurate intervals. As such, lower values are preferred.

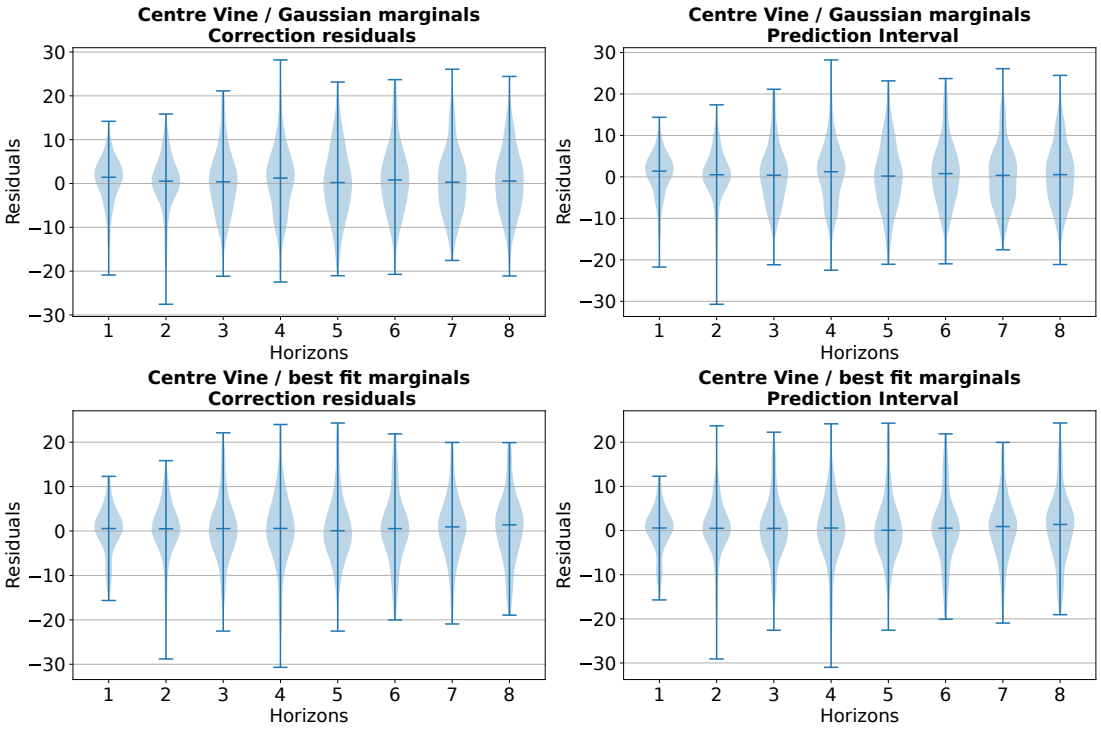

Figure 4.17: Violin plot of the $N = 8$ prediction horizons for the Centre Vine models on the open source wind turbine bearing data, showing the residuals of the linear regression model and corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon.

The CRPS is the expansion on MAE for probabilistic forecasts across all horizons, with lower values being preferable. The highest performing model (CVB) has the lowest interval score for the uncertainty bounds on the correction at a horizon of $N = 1$, and the second lowest interval score across all horizons. This implies that the uncertainty bounds are estimated more appropriately than other models, which have been more heavily penalised for wider and, or, more inaccurate intervals. For all models except CVB, the horizon with the lowest interval score, and so most accurate uncertainty bounds on the copula correction, is $N = 2$, with CVB at $N = 1$. This does not always occur alongside the lowest CRPS for each model, with MGG, RVG and RVB having the lowest CRPS at $N = 2$, matching their lowest interval score horizon, while MGK, MGB, CVG and CVB have the lowest CRPS at $N = 1$. In this case the most accurate uncertainty quantification does not necessarily reflect where the model is performing most accurately. This also applies to the maximum, where the maximum interval score (most penalised uncertainty bounds) occurs at $N = 5$ ((15 hours ahead) for the CVG, $N = 7$ (21 hours ahead) for MGG and RVG, and $N = 8$ (24 hours ahead) for MGK, MGB, RVB and CVB (all models with the more complicated marginals)). The models where the worst performance in terms of CRPS align with their largest interval score are MGB, RVB and CVG, with the rest instead occurring at $N = 4$ or $N = 5$ (12 - 15 hours ahead).

Histogram plots of each method's residuals are shown in Figure 4.18. All of the copula correction methods have shifted the 'no corrections' case to be more centred on 0, with the Multivariate Gaussian model group, the RVG model and CVB model reducing the scale of the maximum errors. To test if the copula methods have improved the Normality of the residual distribution (thus implying the remaining error is approaching Gaussian noise), the skewness and kurtosis values are calculated and presented in Table 4.8. The CVG has the highest kurtosis and skewness at 1.4699 and 0.8503, respectively. This is even larger than the 'no corrections' case with a kurtosis of 1.4352 and skewness of -0.8046, although with a flipped sign. This suggests that the CVG model has even less weight in the distribution tails but has changed the direction of the bulk of the errors to skew negatively. The MGG model has the lowest kurtosis and skewness of all methods,

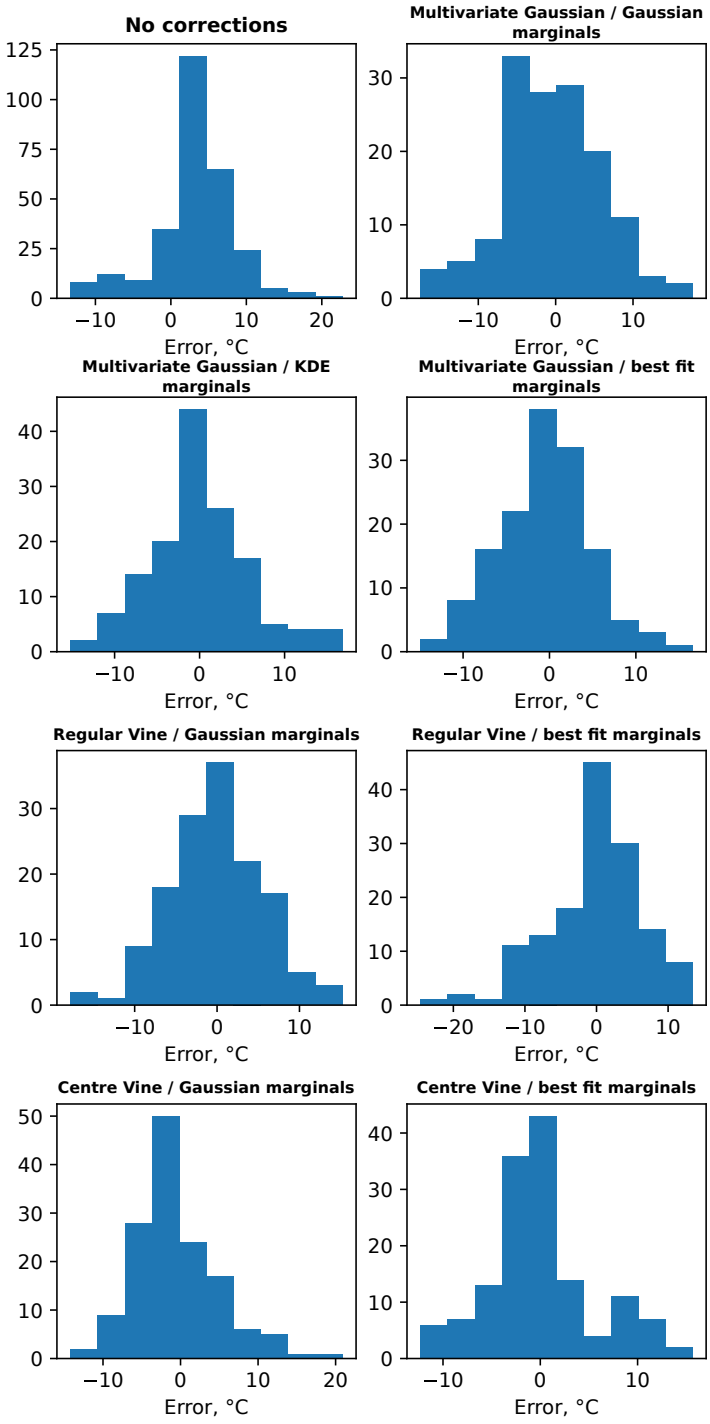

Figure 4.18: Residual histograms for the seven copula correction methods and the benchmark case of no copula corrections for the open source bearing dataset.

Model	Kurtosis	Skewness
No Corrections	1.4352	-0.8046
Multivariate Gaussian / Gaussian marginals	0.2507	0.0261
Multivariate Gaussian / KDE marginals	0.5388	0.2969
Multivariate Gaussian / best fit marginals	0.4414	0.1032
Regular Vine / Gaussian marginals	0.2907	-0.0514
Regular Vine / best fit marginals	0.8233	-0.6441
Centre Vine / Gaussian marginals	1.4699	0.8503
Centre Vine / best fit marginals	0.3905	0.5206

Table 4.8: Skewness and kurtosis values for the open source data model residual histograms. The values for Gaussian distributions are 0 for kurtosis and skewness, as reference.

suggesting it's residuals are the most Gaussian, and so most useful information has been captured by the model.

The Q-Q plot of the timeseries corrections are shown in Figure 4.19. Understandably, the models deviate much more than the linear line depicting equal quantiles than for the simplified synthetic dataset. The models with the most balance across all quantiles are the MGB and CVB models which have done well to match the target data. This would result in more accurate predictions across the full range of temperatures experienced. The models with the highest deviations are the MGG, CVG and RVB at lower quantiles, where the RVB model overestimates the same quantiles while the MGG and CVG underestimate. The RVB has the highest deviation at higher quantiles where it underestimates the target quantile values. This has different consequences in practice, especially underestimating higher quantile values as this will potentially allow assets to experience higher temperatures than expected. This may mean potential intervention may have to be conducted at much shorter notice than desirable. For cases such as wind turbines, generation may have to be curtailed to prevent temperatures in key assets rising higher than permitted levels, or parts may have to be replaced at more frequent intervals due to the increased thermal stress.

The lagged target samples from e_t to e_{t-1} and 1000 samples from the fitted copulas are shown in 4.20. The target samples are located in a dense cloud in the centre at [0.5,0.5], with sparse behaviour around this value in the lower tail. This is reflected in

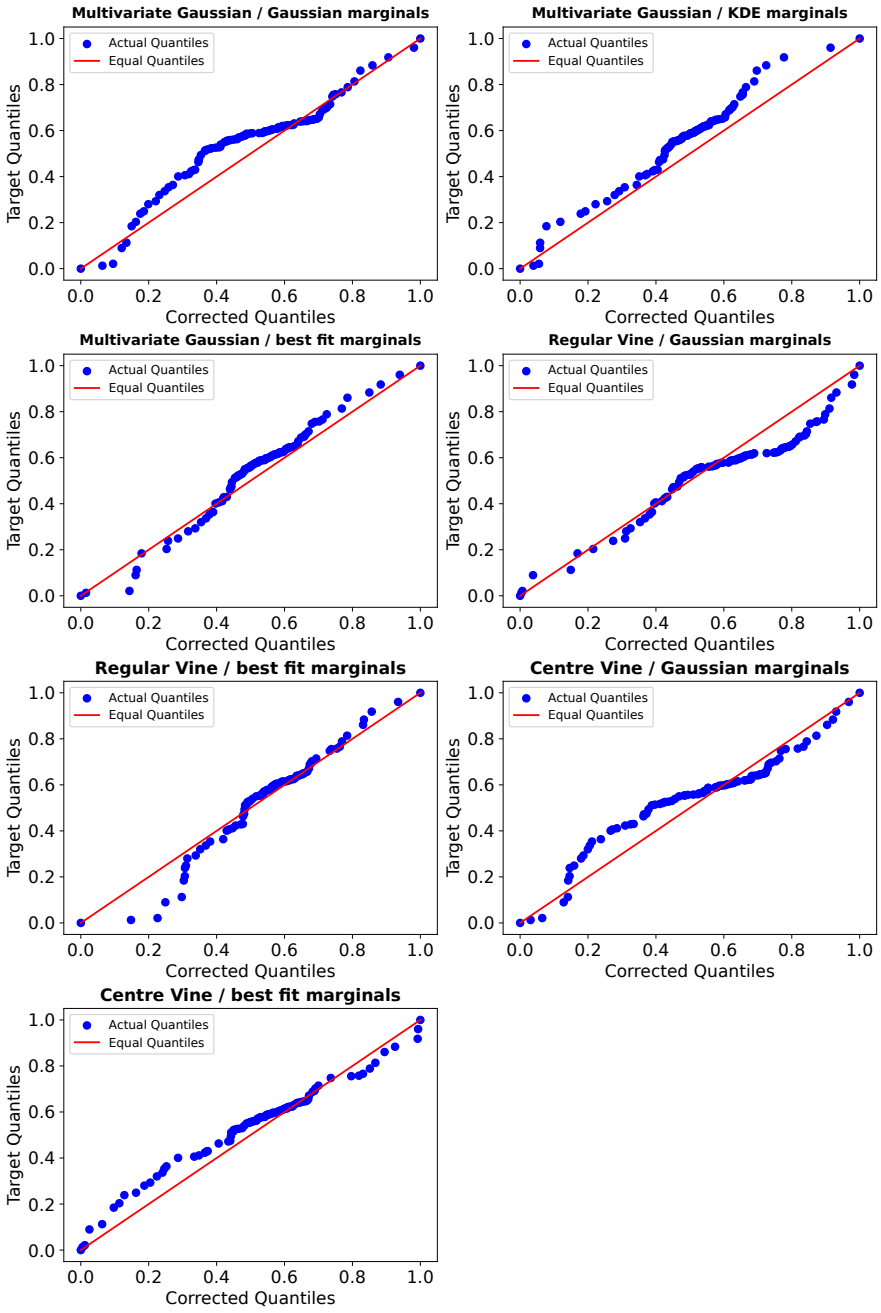

Figure 4.19: Quantile-quantile plot of the target signal quantiles against the corrected signal quantiles for the open source wind turbine bearing data. Identical distributions result in a straight line.

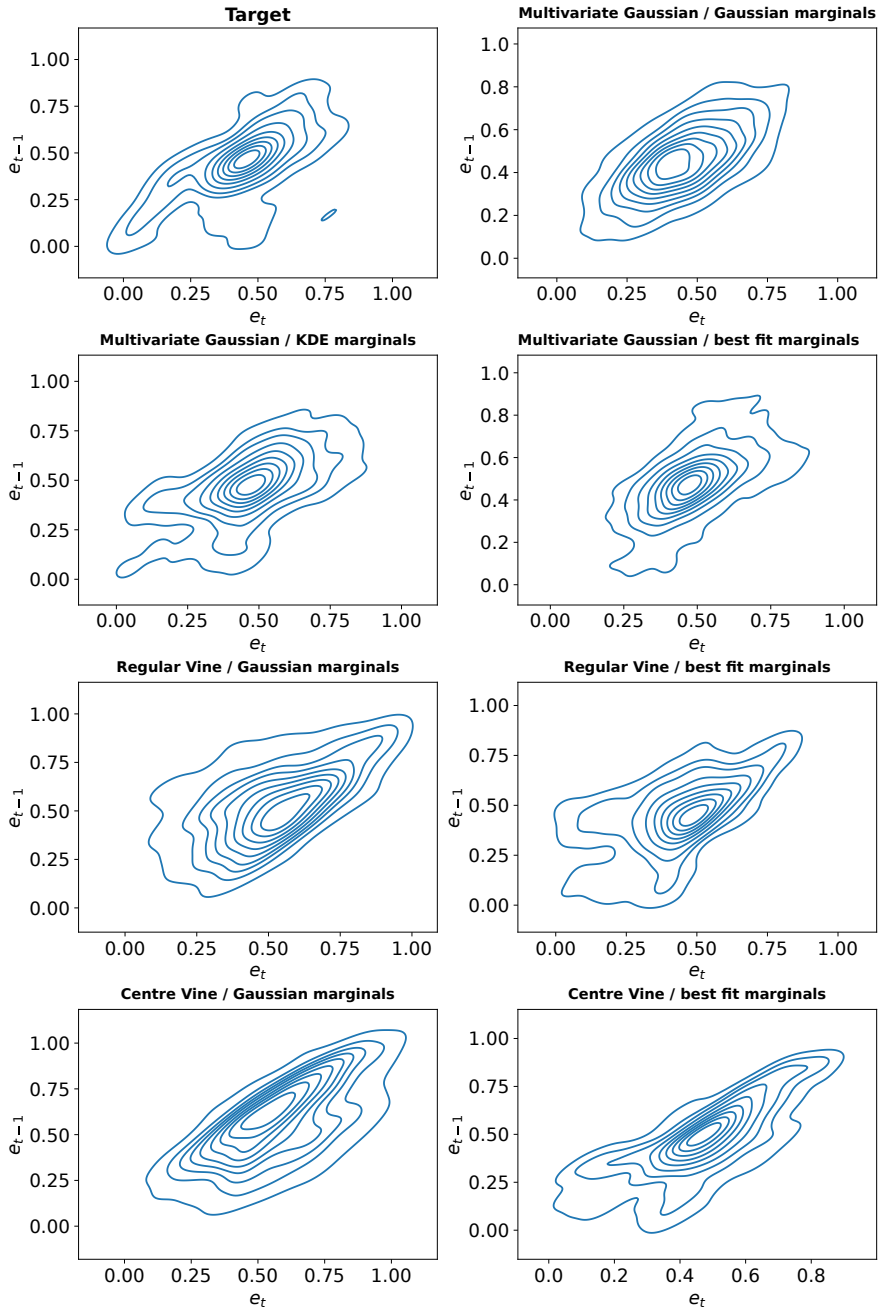

Figure 4.20: Open source wind turbine bearing dataset relationship between e_t to e_{t-1} for the target data and sampled copulas.

the MGK samples where the lower tail behaviour is present around a dense centre. The MGB model behaves similarly, as it has also captured the sparsity around the dense centre but with reduced attention at the extremes. The MGG, RVG and CVG models have generalised the trend to be more elliptical, with the RVG increasing the sparsity in the upper half of the plot (increased density in the lower half of the e_{t-1} samples) and the CVG reflecting this behaviour to have higher density in the upper half of the plot (higher values of e_{t-1}). The RVB and CVB depict similar behaviour where both models have generalised the target data to have a tight upper tail dependence with sparsity in the lower tail.

4.3.4 Case study 3 - Industrial partner data: Nuclear reactor coolant temperature forecasting

Industrial case study organisation and methodology

The seven copula-based methods are applied to anonymised temperature data from a nuclear plant heat exchanger. The data is pre-processed to include only effective full power days (EFPD) and 30 EFPD post-outage are discarded to remove any settling behaviour. This process ensures the models are trained on aging behaviour rather than other modes of plant operation. The data is anonymised by normalising the temperature to a scale of [0,100] and the EFPD is anonymised by masking the timeseries from 0 to 1000 to preserve partner data privacy. A linear regression model is used to predict the temperature timeseries over time. The target timeseries represents reactor inlet header temperature which is created from empirical relationships between measured header temperature, boiler pressure and feedwater temperature at different quadrants of the reactor. The data is split into training, validation, and testing sets at a ratio of 70:20:10 with 1840 total samples, and the timeseries for the inner zone temperature sensor and machine learning prediction is shown in Figure 4.21. The jumps in the linear regression model occur due to the masking of the sensor timeseries to remove gaps in the the operational data due to outages or missing data.

For the inner zone temperature timeseries, a linear regression model is trained on a training set and residuals are created from the linear regression predictions on a

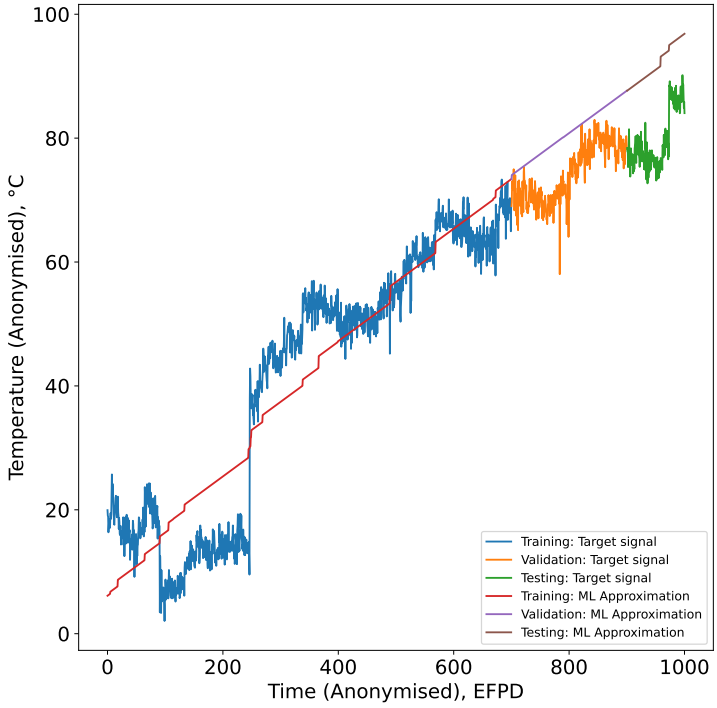

Figure 4.21: Anonymised inner zone temperature in a nuclear reactor with the linear regression predictions on the training, validation and testing sets. The data is remasked to anonymise gaps in the data after outages or sensor failures and to make the data continuous (this process causes the linear regression predictions to no longer look straight).

validation set. The errors are lagged from e_t to e_{t-N}, where $N = 15$, to train the seven copula-based methods. Prior to training the copulas, the errors are normalised from [0,1] and the scaling parameters are kept so that the copula correction samples may be scaled into the appropriate scale of the data space. The testing set predictions are created using the conditional relationship between the last known error up to the forecast horizon of $N-1$. Stepping through each timestep updates the predictions over the new horizon to result in N correction predictions per data point.

Industrial case study results

For each copula-based method, the recorded MAE and percentage improvement in MAE are given in Table 4.1 and Table 4.2, respectively. The industrial dataset MAE values in Table 4.1 show that all copula models improve on the "no corrections" benchmark, with the worst improvement shown in Table 4.2 from the MGK model cutting the MAE by 70.8%, to the best improvement from the CVG marginals cutting the MAE by 82.2%. For the three categories of models, the Gaussian marginals were the highest performing Regular Vine model, the Gaussian marginals were the highest performing Centre Vine model, and the Gaussian marginals were the highest performing Multivariate Gaussian model based on MAE.

The target timeseries, linear regression prediction and corrected timeseries with uncertainty bounds given by the 90% copula CDF values are shown in Figure 4.22 for each copula model. In Figure 4.22, the Regular Vine and Multivariate Gaussian models seem to track (red) the target signal (blue) more accurately towards the end of the testing set where a jump in the data is present. The Centre Vines, however, adapt to this jump and manage to track the target signal throughout the testing dataset. The Centre Vines models have very wide prediction intervals (orange) before the jump in the data, which reduces once the jump has occurs and the models become less uncertain about the state of the plant.

The violin plot of the predictions and prediction interval over each horizon for the industrial heat exchanger dataset are separated by model type. The Multivariate Gaussian models are shown in 4.23. The breakdown of the horizon CRPS and percentage

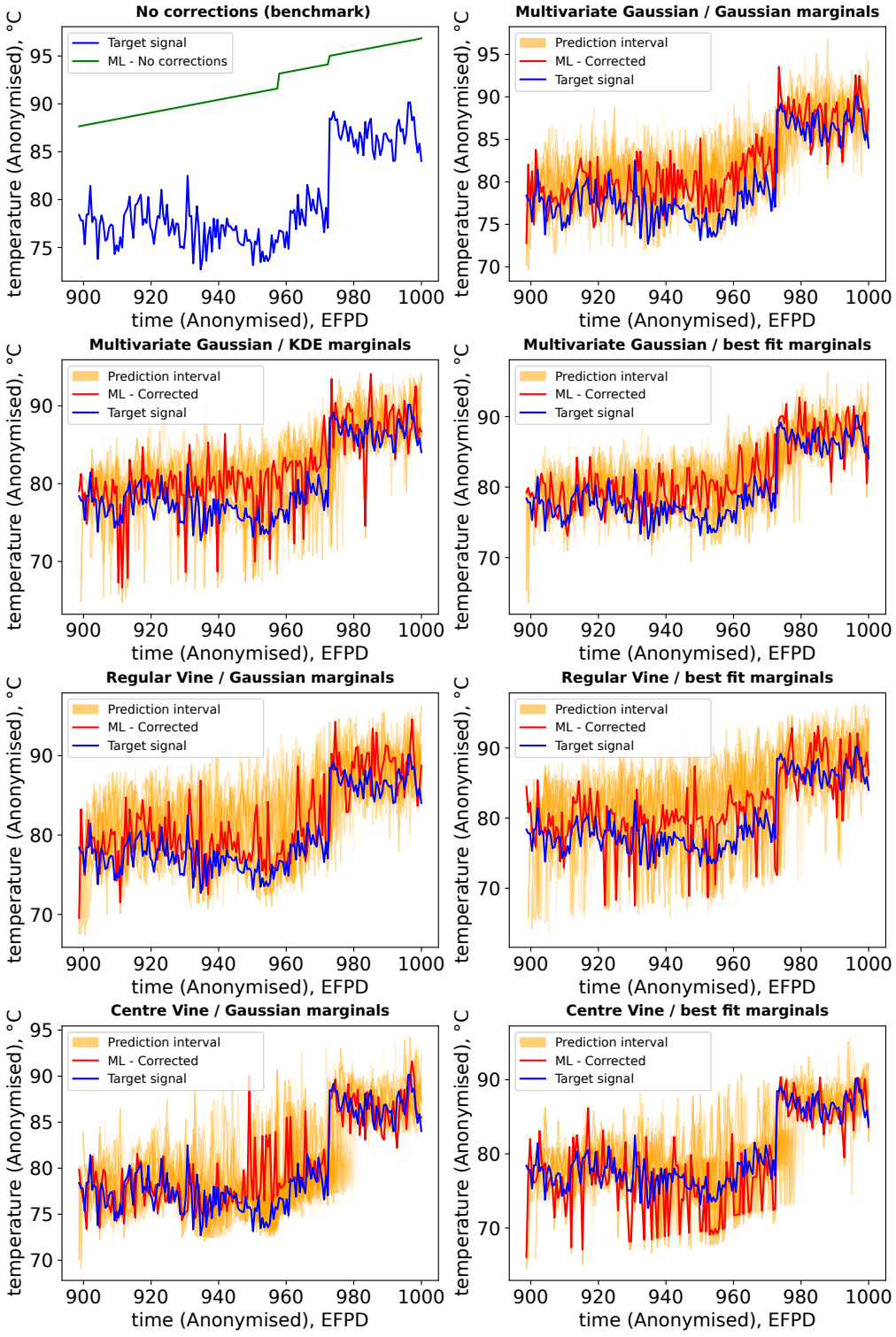

Figure 4.22: Timeseries plots of the seven copula models with the true sensor testing data, corrected timeseries and prediction interval on the copula corrections.

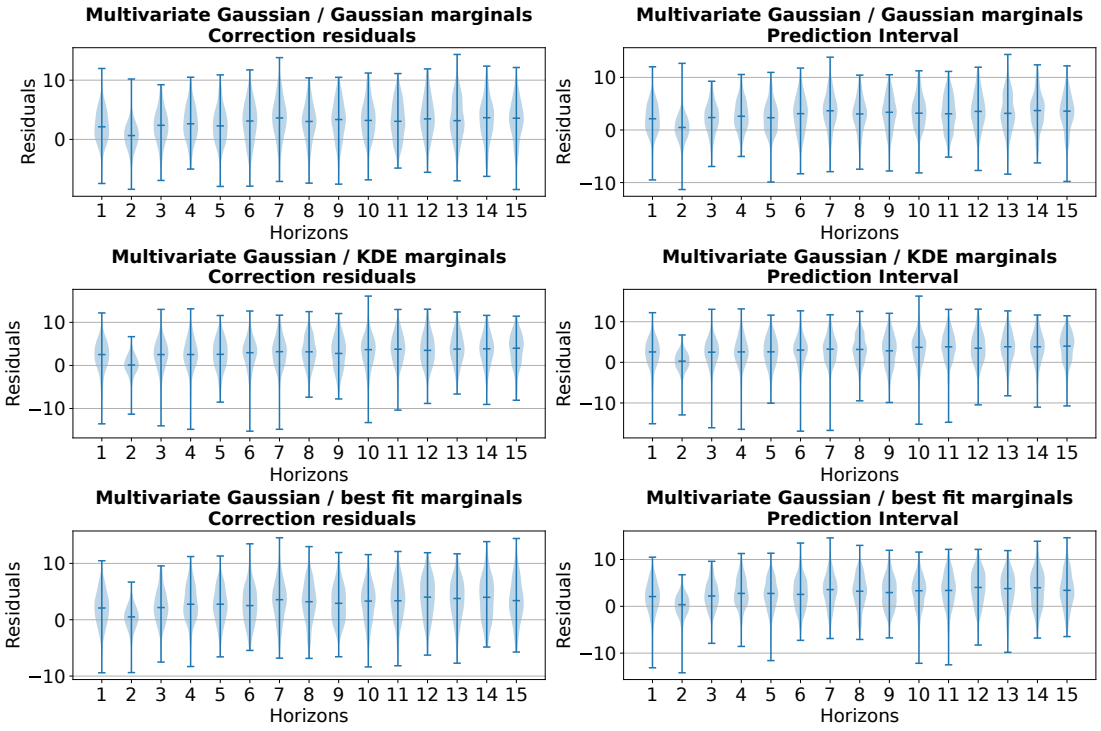

Figure 4.23: Violin plot of the $N = 15$ prediction horizons for the Multivariate Gaussian models on the industrial heat exchanger data, showing the residuals of the linear regression model and corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon.

improvement over horizon 1 is shown in Table 4.9. All Multivariate Gaussian copula models have a percentage improvement over the reference horizon of $N = 1$ at $N = 2$, with the MGK model having the largest improvement of 49.03 %. The worst CRPS occur at different horizons for each model with MGG at the shortest horizon of $N = 7$ (week ahead) and MGK at the furthest horizon of $N = 15$. The furthest horizon in this case study is $N = 15$, and so the poorest performance of the MGK model at this forecast horizon makes sense due to the lack of information to make more informed predictions, however, for the MGG model, there is behaviour being learned that misleads the model at this week ahead mark which makes it perform worse than further out horizons where less information is available. The model uncertainty bounds demonstrate that the model is able to recognise where it is not confident in areas where the uncertainty intervals are larger and the model corrections under perform, as this means more risk is being assigned to areas that are shown to possess more risk. The same is true for the inverse, where the model is shown to perform well in areas it is more confident in the correction. Looking at the extremes in both uncertainty bounds and correction residuals, the MGK model has both of these aligned, performing the best at $N = 2$, and worst at $N = 10$, with the uncertainty quantification in line with this performance. The MGG model has aligned minimised uncertainty bounds and correction performance extremes at $N = 4$, and the MGB model does not have any aligned maximum or minimum extremes across horizons. The violin plots for the Regular Vine models are shown in Figure 4.24. For both models, there is significant improvement at $N = 2$ over the chosen correction horizon, with the RVG model also showing slight improvement at $N = 3$. This again, may be due to better generalisations at these horizons. The minimum extremes in the correction residual and uncertainty bounds align on the same forecast horizon for both models, with that horizon being $N = 1$ for the RVG model and $N = 2$ for the RVB model. For the RVB model, the minimum extremes occur on the same horizon with the best CRPS, while the RVG minimises the extremes at the chosen horizon in the current methodology. Minimising extreme residuals may be a more desirable metric than general performance (measured by CRPS, for example) in a practical application. There may be threshold alarms on an asset which, if not

Table 4.9: MAE and percentage change for each forecast horizon copula-based correction method and the benchmark case of no corrections for the industrial heat exchanger dataset. Horizon 1 is the reference for the percentage change with positive percentage changes as improvements.

Model/ Metric	**MGG**	**MGK**	**MGB**	**RVG**	**RVB**	**CVG**	**CVB**
Horizon 1 MAE	2.8993	3.5687	3.0339	2.9275	3.5263	2.1764	3.0277
Horizon 1 % change	0	0	0	0	0	0	0
Horizon 2 MAE	2.0842	1.8188	1.7309	1.7375	1.7881	1.8126	1.8103
Horizon 2 % change	**28.11**	**49.03**	**42.95**	**40.65**	**49.29**	**16.72**	**40.21**
Horizon 3 MAE	3.0687	3.6265	3.0692	2.8758	3.8826	2.5364	3.1815
Horizon 3 % change	-5.84	-1.62	-1.16	1.77	-10.1	-16.54	-5.08
Horizon 4 MAE	3.3624	3.7306	3.5424	4.016	4.6081	2.5501	2.8019
Horizon 4 % change	-15.97	-4.54	-16.76	-37.18	-30.68	-17.17	**7.46**
Horizon 5 MAE	3.1352	3.7116	3.4275	4.1421	5.1906	2.7202	2.8882
Horizon 5 % change	-8.14	-4.0	-12.97	-41.49	-47.2	-24.99	**4.61**
Horizon 6 MAE	3.7932	3.9238	3.5594	4.3876	5.8578	2.4773	3.0453
Horizon 6 % change	-30.83	-9.95	-17.32	-49.88	-66.12	-13.83	-0.58
Horizon 7 MAE	4.2552	3.9436	4.1059	5.3412	5.4259	2.4607	2.9818
Horizon 7 % change	-46.77	-10.51	-35.33	-82.45	-53.87	-13.06	**1.52**
Horizon 8 MAE	3.724	3.9082	4.019	5.1606	5.759	2.8213	3.1938
Horizon 8 % change	-28.44	-9.51	-32.47	-76.28	-63.32	-29.63	-5.49
Horizon 9 MAE	3.8917	3.974	3.5323	5.4438	5.7873	2.7196	3.2284
Horizon 9 % change	-34.23	-11.36	-16.43	-85.95	-64.12	-24.96	-6.63
Horizon 10 MAE	3.9411	4.2645	3.8654	5.1837	6.0124	2.85	3.0407
Horizon 10 % change	-35.93	-19.5	-27.41	-77.07	-70.5	-30.95	-0.43
Horizon 11 MAE	3.807	4.4194	4.2213	5.0281	6.0263	2.8108	3.0175
Horizon 11 % change	-31.31	-23.84	-39.14	-71.75	-70.9	-29.15	**0.34**
Horizon 12 MAE	4.2056	4.2943	4.5808	4.8511	5.4499	2.7869	3.1681
Horizon 12 % change	-45.06	-20.33	-50.99	-65.71	-54.55	-28.05	-4.64
Horizon 13 MAE	4.1796	4.413	4.4129	4.8071	5.6946	2.7684	3.352
Horizon 13 % change	-44.16	-23.66	-45.45	-64.2	-61.49	-27.2	-10.71
Horizon 14 MAE	4.1659	4.4223	4.4576	4.7947	5.9275	2.7463	2.9398
Horizon 14 % change	-43.69	-23.92	-46.93	-63.78	-68.09	-26.19	**2.9**
Horizon 15 MAE	4.0999	4.4512	4.2626	4.8092	5.5663	2.9193	3.2321
Horizon 15 % change	-41.41	-24.73	-40.5	-64.28	-57.85	-34.13	-6.75

The model acronyms are MGG - Multivariate Gaussian (Gaussian), MGK - Multivariate Gaussian (KDE) and MGB - Multivariate Gaussian (Best fit), RVG - Regular Vine (Gaussian), RVB - Regular Vine (Best fit), CVG - Centre Vine (Gaussian) and CVB - Centre Vine (Best fit).

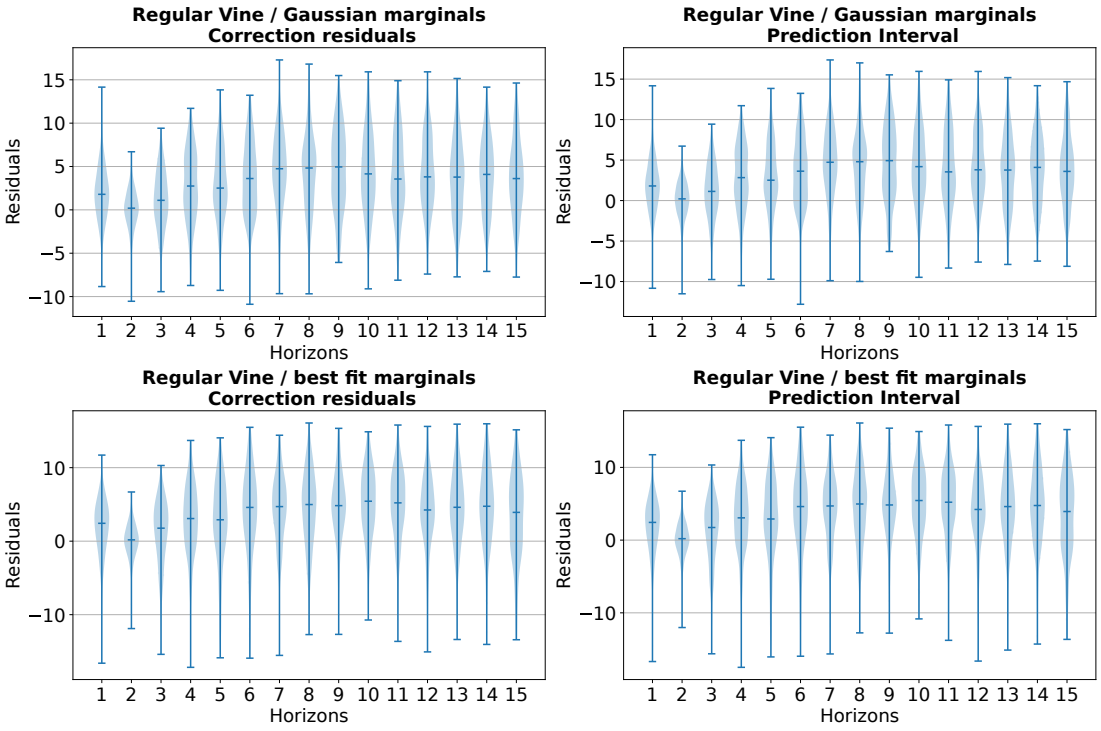

Figure 4.24: Violin plot of the $N = 15$ prediction horizons for the Regular Vine models on the industrial heat exchanger data, showing the residuals of the linear regression model and corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon.

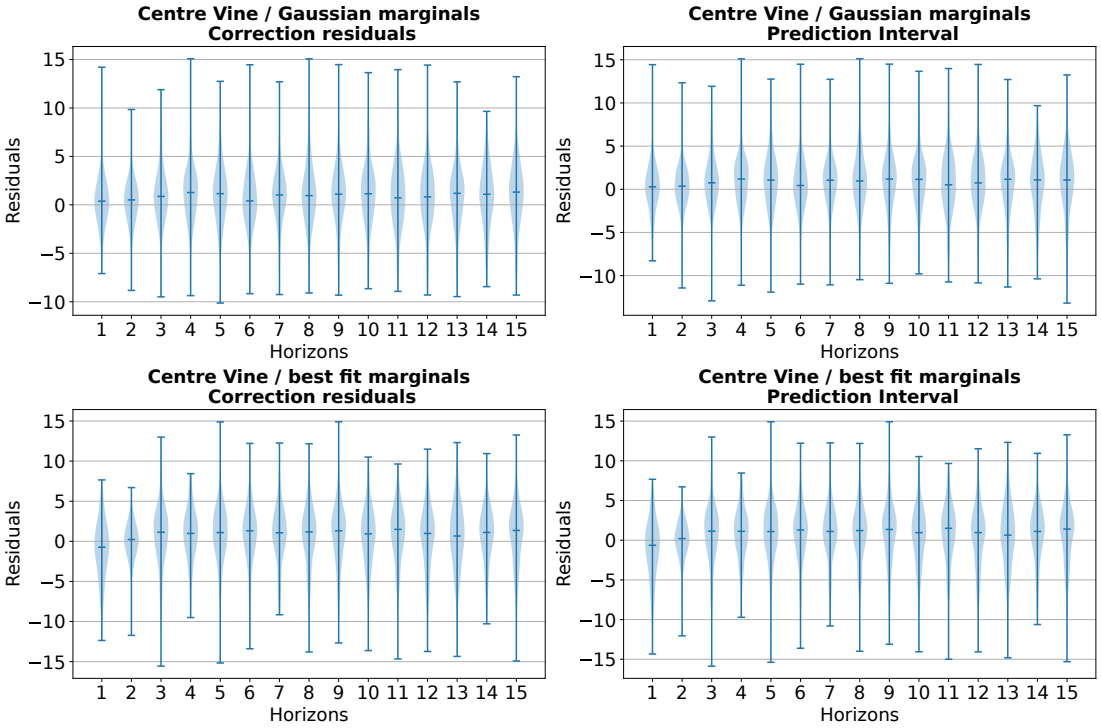

Figure 4.25: Violin plot of the $N = 15$ prediction horizons for the Centre Vine models on the industrial heat exchanger data, showing the residuals of the linear regression model and corrections against the target signal (a perfect correction would result in 0 residual) and the spread of the prediction interval over each horizon.

expected to trip, have more consequences for the immediate intervention required on the asset than being within a wider, but acceptable margin of error within an allowed temperature threshold. The maximum extremes for both the correction residuals and uncertainty bounds align on $N = 7$ for the RVG model, but not for the RVB model. Again, this means that the RVB model is appearing more confident on a horizon which is incurring the models most extreme residuals. Lastly, the violin plots for the Centre Vine models are shown in Figure 4.25. The CVG model CRPS improves at a horizon of $N = 2$, as with previous models, but the CVB model experiences interesting behaviour in improvement across horizons. For previous models that have improved at multiple horizons, it tends to occur chronologically, with improvement percentage lessening at longer horizons. For the CVB model, improvements occur at $N = 2, 4, 5, 7, 11, 14$, with the percentage improvement generally lessening the larger the horizon. This means that, taking any of these horizons, the corrections would be better than for the horizon

which provides the model with seemingly the most information. The largest improvement is at $N = 2$, which aligns with all other models in this case study. The CVB model also has aligned correction residual and uncertainty bound extremes for both the maximum and minimum cases, demonstrating a match in demonstrating when the model believes it is confident or not, and when the model correction performance reflects this belief.

The CRPS is given in Table 4.4 for each model over all horizons and the maximum and minimum with corresponding horizons, and the interval score on the prediction interval on the copula correction are shown in Table 4.5 across all horizons for each model, with the horizon associated with the maximum and minimum interval score provided. On the industrial heat exchanger case study, the CVG model has the lowest mean and standard deviation for both the average CRPS across all horizons and the average interval score on the uncertainty bounds across all horizons. The RVB model, in contrast, has the highest (and so worst) mean and standard deviation across the average CRPS and interval score on all horizons. All models have their lowest CRPS and interval score occur at a horizon of $N = 2$, which, as discussed in previous case studies, is likely due to a balance of model information across the number of available forecasting horizons while preventing overfitting, preserving model generalisation. At this horizon of $N = 2$, the MGB model had the best CRPS while the MGG had the worst. For the the interval score, the RVB model had the lowest interval score of all models at $N = 2$, while the MGG model had the highest. However, the worst CRPS and interval score horizon changes per model. For CRPS and interval score, the MGK and CVG models have the worst horizon at the furthest forecasting horizon of $N = 15$, where the model is provided with the least information. The CVB model, on the other hand, has its highest interval score at a horizon of $N = 1$, implying its poorest uncertainty quantification performance occurs at the horizon of our chosen correction where the risk associated with the prediction is most important. For model maximum CRPS and interval score, the RVB model had the worst of all model maximums, while CVG had the lowest.

The changes in histograms in Figure 4.26 show that all copula models have moved

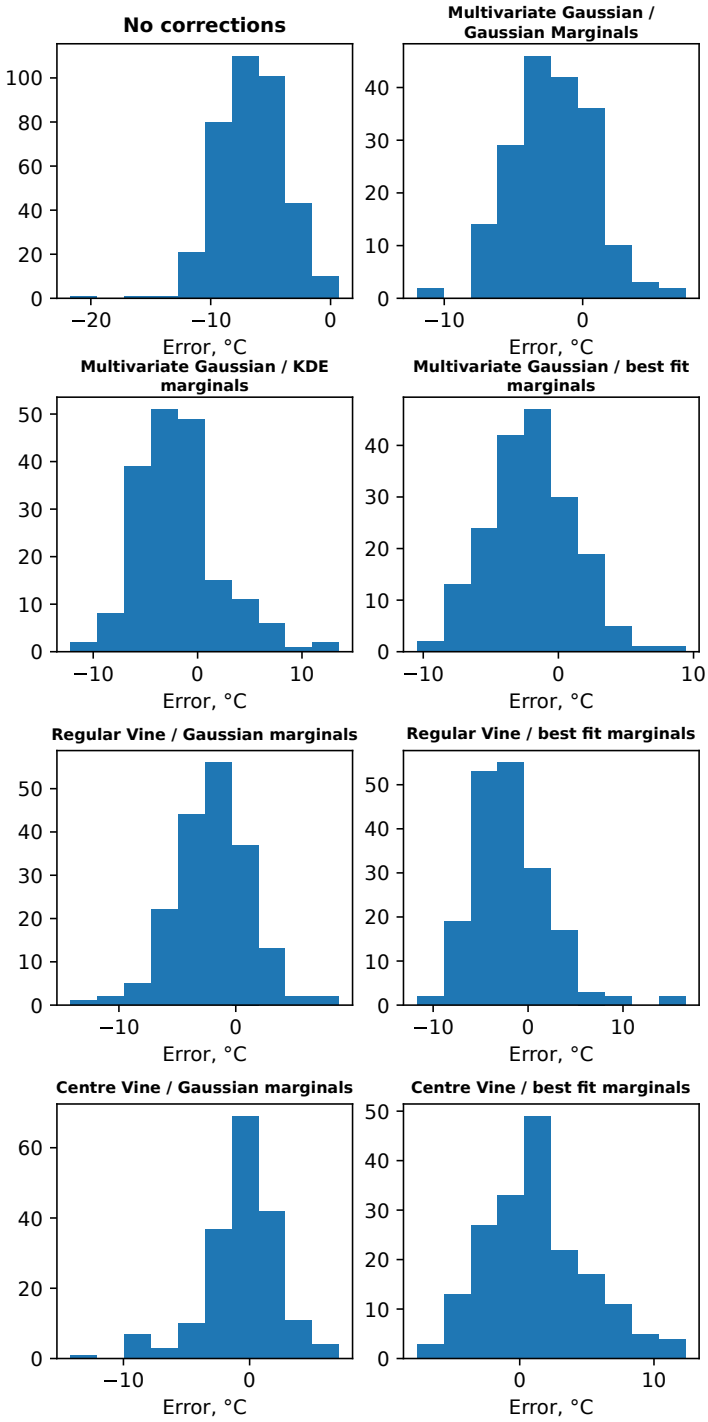

Figure 4.26: Residual histograms for the seven copula correction methods and the benchmark case of no copula corrections for the industrial heat exchanger dataset.

Model	Kurtosis	Skewness
No Corrections	-0.9251	-0.0388
Multivariate Gaussian / Gaussian marginals	0.5598	0.0443
Multivariate Gaussian / KDE marginals	1.534	0.8597
Multivariate Gaussian / best fit marginals	0.2998	0.2858
Regular Vine / Gaussian marginals	1.4171	-0.1521
Regular Vine / best fit marginals	3.2703	1.2128
Centre Vine / Gaussian marginals	2.7392	-1.0488
Centre Vine / best fit marginals	0.1965	0.4619

Table 4.10: Skewness and kurtosis values for the industrial data model residual histograms. The values for Gaussian distributions are 0 for kurtosis and skewness, as reference.

the errors to be centred much closer to 0 than the benchmark 'no corrections' case. The 'no corrections' case entirely overestimates the value of the target signal, and all models can recover this shift. To test the Normality of the residuals after each correction, the skewness and kurtosis values are shown in Table 4.10

. The 'no corrections' case has a negative kurtosis, which implies heavier tails than a Gaussian distribution. All the copula models change the kurtosis to a positive value, which implies less weight in the tails than expected for a Gaussian with the bulk of the residuals centralised in the distribution. The only models with a smaller value of kurtosis than the 'no corrections' case are the MGG, MGB and CVB models, which have made the 'no corrections' case closer to Gaussian. The 'no corrections' case has a small negative skewness value which means the left tail is heavier than the right. None of the copula models reduce the value of skewness, which implies some transform to the data which has not resulted in leftover Gaussian noise. The RVG and CVG remain negatively skewed, while the other models have shifted the skewness to a positive value. The model with the highest kurtosis and skewness is the RVB model, with the RVG model closely behind. The model with the lowest kurtosis is the CVB (3rd lowest skewness), and lowest skewness is the MGG (3rd lowest kurtosis).

Compared to the synthetic dataset, the deviations on the Q-Q plots in Figure 4.27 are much more evident, with MGK, RVG, RVB and CVB models showing the highest deviation in the mid quantile predictions where the models tend to overestimate the

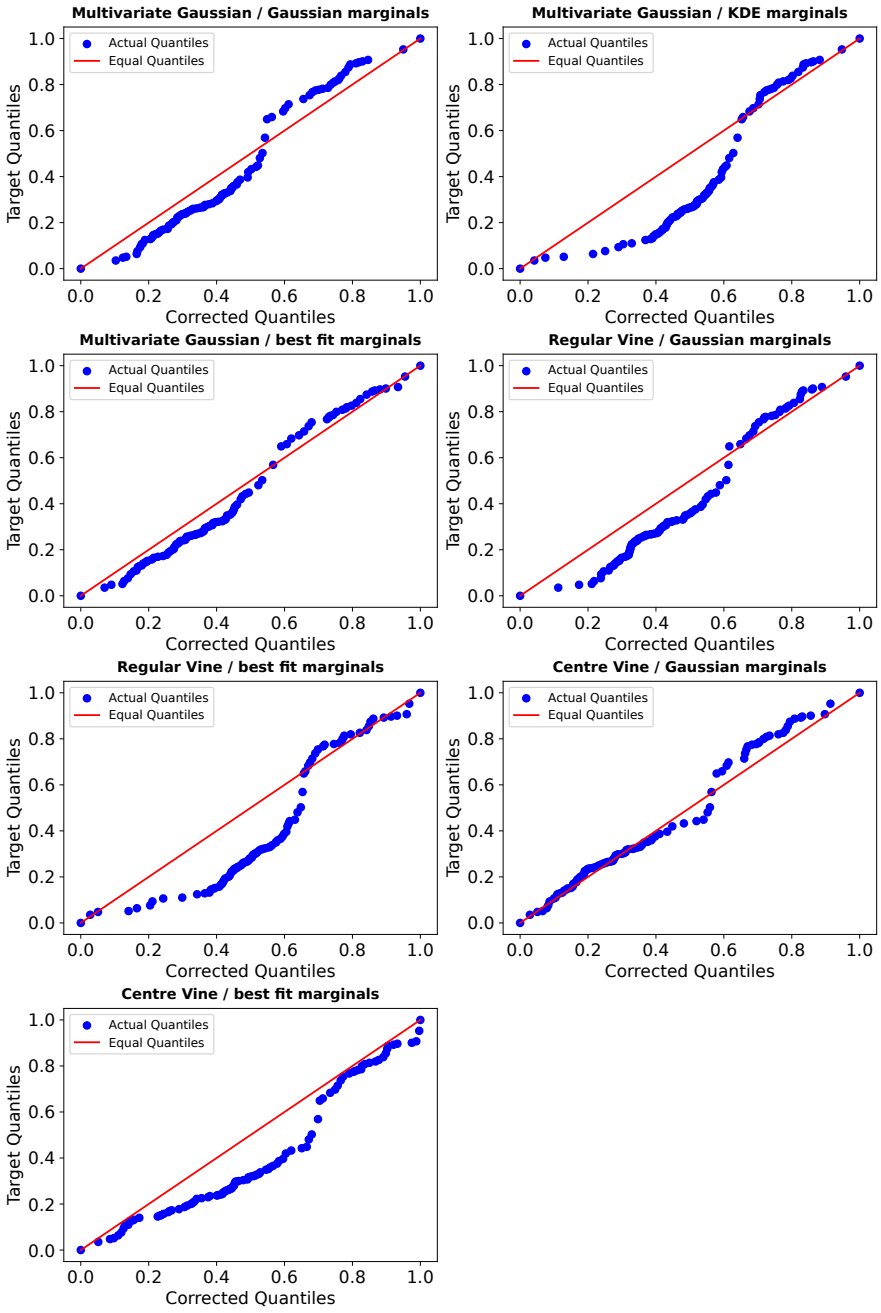

Figure 4.27: Quantile-quantile plot of the target signal quantiles against the corrected signal quantiles for the industrial heat exchanger data. Identical distributions result in a straight line.

value of the target signal. The MGG, MGB and CVG models match with the target signal on the Q-Q plot more accurately. The most severe deviations across the models overestimate the target quantiles which in a practical setting could result in forecasts being utilised to advise emergency maintenance or changes in plant operation to prevent temperature thresholds from being violated. However, most of these deviations occur at the mid quantiles which may not have as much impact as a deviation at an upper extreme. Interestingly, all models seem to have decent calibration at the extreme quantiles which is where, depending on the application, the most severe consequences of temperature occur.

The copula fitting shown in Figure 4.28 shows that the more complex marginal fitting processes (KDE or best fit) tended to capture the behaviour occurring in the upper tail in the target residuals (top left of Figure 4.28), whereas the models with Gaussian marginals tend to widen their focus, exhibiting longer lower tail behaviour. However, based on other metrics, capturing more general behaviour may have allowed the Gaussian marginal methods to succeed over the more complicated marginal methods.

4.3.5 Result overview and discussion

A variety of metrics and visualisations have been chosen to examine the model performance from several view points, with success and failure in each metric resulting in a different operational consequence. The percentage improvement in MAE at the chosen horizon over the linear regression model justifies the additional computational and theoretical complexity of the hierarchical copula approach. The CRPS, interval score and violin plots analyse the copula model competency in providing suitable uncertainty quantification that aligns with the performance of the model corrections. It is expected that the model should provide a tight prediction interval on corrections that are deemed to have less risk, and this should align with more accurate corrections. Additionally, analysing the model performance over its forecast horizons can identify barriers to model performance which may reveal areas for additional model development. For example, it would be expected that forecasting to the furthest available

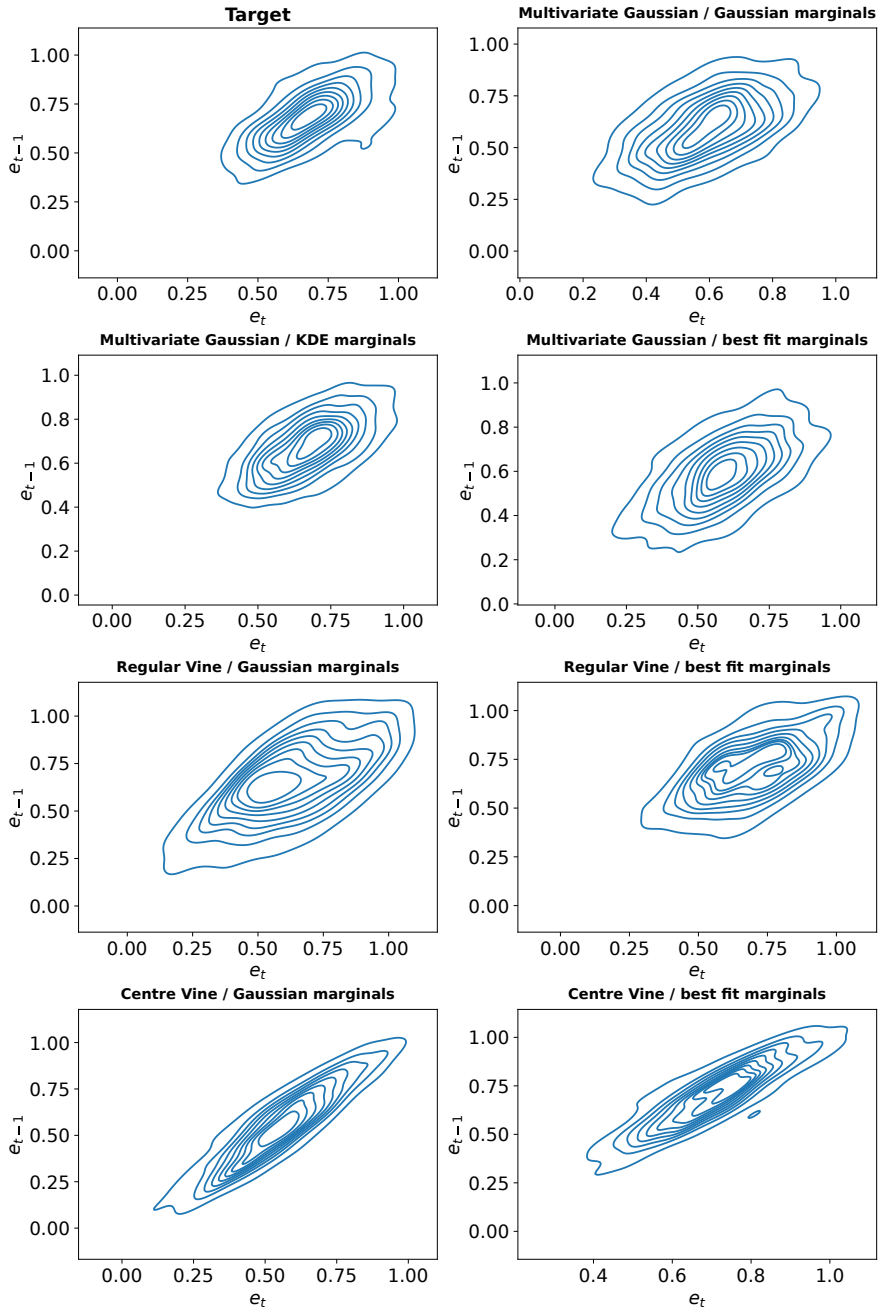

Figure 4.28: Industrial heat exchanger dataset relationship between e_t to e_{t-1} for the target data and sampled copulas.

horizon would be the most uncertain and inaccurate due to limited information being available to the model. It would follow that closer horizons should be expected to be more accurate due to benefiting from the additional observations. Where models do not experience this behaviour, there may be additional relations in the data that either misleads the model; information that is insufficiently related to the target variable; or behaviour that the model is unable to capture. This may be permissible if the model is able to accurately communicate the risk in the correction at these points, and if the horizons affected have less impact on operational decisions. The histogram, timeseries plot and Q-Q plot are three variations on visualising the model performance over the testing data. The histogram provides the distribution of remaining residuals after the corrections are applied, showing where the mode falls, the tendency and shape of the distribution and the extreme values. The extremes provides a sense of the worst case residuals while the timeseries plots demonstrate where these worst case residuals occur. Combined, the histogram provides an overview of *how* wrong the model is, and the timeseries plot shows *where* the model goes wrong. To supplement this function of showing where the model goes wrong, the Q-Q plot presents how well the model captures the target distribution. Where the model deviates from the target distribution, and if the model under or over estimates the value, can provide additional understanding of risk in applying the model. If the model continuously underestimates the upper extremes, the asset may be experiencing more severe temperatures than is being communicated by the model. This may result in more severe wear over time, or the unexpected tripping of alarms requiring immediate operation changes. Lastly, the copula density plots provide a sanity check on how each copula model approaches the first lagged temperature variable. The density plots demonstrate what features the copula models have identified and prioritised, and also show where the copulas are fitting poorly.

For the synthetic dataset, the MGG model was, across the different metrics, found to be the best performing model with the most balance across the CRPS, interval score and histogram metrics. The performance in these metrics was also visible in the timeseries plots where the MGG model corrections were more able to track the

target signal with less noise and peaks. The MGG also had the second best percentage improvement in MAE over the linear regression benchmark, with third place going to the MGB. The MGB model additionally performed well in the violin plots, with the most balanced CRPS performance over all horizons. The MGB model also most closely captured important detail in the lower tail of the copula density plots. The worst performing model was the RVB model, closely followed by the RVG model. The RVB model performed poorly across the percentage improvement in MAE over the linear regression, had the most visible timeseries noise and had the highest CRPS across all forecast horizons. For the CRPS summary, the RVB model had the largest mean of all models for the averaged horizons and the largest CRPS across all models worst performing horizon. The RVB uncertainty quantification ability was the poorest of all models, with the highest mean and standard deviation of the averaged horizon interval scores, and the largest interval score across all models best and worst horizons. The RVG model performed poorly on the histogram metrics, with the highest kurtosis and second highest skewness, showing poor improvement on improving the linear regression residual Normality. It also had the largest deviations in the Q-Q plot at the lower and mid quantiles, and the copula density showed a noisy circular behaviour with poor conditioning to the target sample shape. In this case study, with the MGG and MGB models as the top performing models, and the RVB and RVG models as the bottom performing group, the Multivariate Gaussian has outperformed the Regular Vine approach, regardless of marginal assumption choice.

For the open source wind turbine bearing temperature dataset, the MGB model performed the best, tied jointly with the CVB model. The MGB provided balanced performance across all forecasting horizons, minimising errors at longer timescales compared to the other models. The mean and standard deviation on the averaged horizon CRPS was the lowest for the MGB model, which also had the lowest maximum CRPS of all models. Other than deviation at the lower quantiles, the MGB model experienced the lowest deviation from the target quantiles in the Q-Q plot. The CVB model slightly outperformed the MGB in percentage improvement over the linear regression, had the lowest mean interval score averaged across all horizons and lowest minimum interval

score across all horizons, hence achieving the second place spot. The worst models were a tie between the MGG and RVG models. The MGG model underperformed in the percentage MAE improvement over the linear regression base model and had vsibly poorer tracking of the target signal in the timeseries plots, expecially the large peaks and troughs. With the copula density plots, the MGG model opted for a large, general coverage with less details captured compared to the other models. The RVG model had poor CRPS balance across all horizons and underperformed in CRPS and interval score. The RVG model had the largest mean and standard deviation in CRPS averaged across all horizons and the highest CRPS of all models worst horizon. Additionally, the RVG model had the highest mean interval score averaged across all horizons. In this case, the complex marginal assumptions outperformed the simplified, with the MGB and CVB outperforming the MGG and RVG. Once again, the Regular Vine features in the underperforming models.

For the final case study on the partner industrial heat exchanger dataset, the CVG model is the best performing, with no distinct second best model. The CVG performs well across the percentage improvement in MAE compared to the linear regression model and also has the most alignment with the target quantiles in the Q-Q plot, despite some deviation in the upper quantiles. It has the best general performance across all horizons, with the lowest mean and standard deviations in CRPS and interval score averaged across all horizons. The CVG model has also minimised the maximum CRPS and interval score at the worst performing horizons compared to the other models. The worst performing models are the RVB, then MGK. The RVB underperforms across most metrics: having poor performance across all horizons, the highest mean, standard deviation and worst case maximum CRPS and interval score; the highest skewness and kurtosis for the residual histograms; and lastly the largest Q-Q Plot deviations in the lower quantiles. The MGK model only performs worse than the RVB model in the percentage improvement in MAE over the linear regression base model and is visually noisier with larger peaks in the timeseries plot. However, the MGK model seems to capture the target behaviour most appropriately in the copula density plot. In this case, the findings are less clear cut. This is the second time the Centre Vine models

have appeared in the top group, the second time the Multivariate Gaussian models have appeared in the bottom group, and the third time the Regular Vine models have appeared in the bottom group, cumulatively, across all case studies. In this example, the simpler marginal assumption model was a success, with the complex marginal methods appearing in the bottom group. Additionally, across all case studies, only the Centre Vine models have appeared in the top group exclusively. The Regular Vine models have exclusively appeared in the bottom group, while the Multivariate Gaussian model appears in both top and bottom groups cumulatively across all three case studies.

An overview figure emphasizing the process and some selected results from the industrial heat exchanger dataset was presented prior to the case studies in Figure 4.1 to further highlight the practical benefit of this approach, and the consequences of no intervention from the copula models. As shown in the outcome section of Figure 4.1, the true measurement is much closer to the copula predictions and within the uncertainty bounds estimated by the copula model, while the base model (linear regression) entirely overestimates the true measurement by over 10 °C.

4.4 Conclusion, contribution and future work

Cost-effective maintenance for critical assets requires sufficient time margins and an accurate assessment of the health of assets to prevent unnecessary interventions which incur loss of revenue. One method of accomplishing this is through hierarchical modeling to predict the temperature increases over a suitable time horizon. In this paper, complex dependency modeling was used to calibrate and provide uncertainty quantification for a base model in a hierarchical modeling approach applied to temperature forecasting of critical infrastructure. A simple, interpretable data-driven linear regression model was used to generate the initial temperature forecasts and statistical Copula models were used to calibrate the predictions over short-term horizons. The applicability of the approach to industrial data was the primary factor influencing the adoption of Copulas, in order to represent time-series behavioral patterns. For example, data quality and malfunctioning sensors may result in noise which is neither linear dependent nor Gaussian distributed. In this chapter, this hypothesis has been

executed through the demonstration of different multivariate distributions considered for uncertainty propagation over time horizons (as detailed in tables 4.4 and 4.5). A variety of high dimensional Copula models were compared in this study, investigating the differences in Multivariate Gaussian and Vine Copulas with different levels of complexity in the assumptions of their marginal distributions. The density in the Copula models were used to provide corrections with a measure of uncertainty based on the last known error between the linear regression model and the true state of the plant, informed by predictions made at previous horizons. The uncertainty bounds provide a measure of risk in the correction, showing where the Copula model is confident in its calibration. Results demonstrate the Copula-based approach is robust against some industrial data quality issues. For example, considering Figure 4.22, discrete changes in the data due to maintenance or post-outage behavior did not degrade the methods performance. Additionally, in the case of sensor noise (included as Gaussian noise) in the synthetic dataset, the proposed method performed better than the base model in all scenarios. Both these data quality issues are addressed as the proposed method is robust to both the form of noise distribution and the form of dependency that dictates the propagation of uncertainty.

Three datasets were used to demonstrate the methodology: the first was a synthetic dataset designed to demonstrate the ability of the Copulas on a purpose built scenario; the second was operational data of the rear bearing temperature of an operational turbine, which is an openly available dataset showing the applicability of the method to other industrially relevant timeseries; and lastly, the operational data of a nuclear plant inner zone temperature used to estimate the effects of aging by evaluating the heat exchange between the reactor and coolant loop. In each dataset, long-term predictions were provided by the base model and where the datasets were split into training, validation, and testing sets. As all models were validated against held-out data from real operational environments, the models have not been previously exposed to this data, thus the performance of the models have been evaluated without contamination. All Copula models improved on the benchmark case which was the linear regression model with no corrections. The Center Vine models out competed the Multivariate Gaussian

and Regular Vine models on the operational datasets, with the best fit marginals providing 27.16 % improvement in MAE over the benchmark for the wind turbine dataset and Gaussian marginals providing a 82.17 % improvement in MAE over the benchmark for the nuclear plant data. For the synthetic dataset, the Multivariate Gaussian with KDE marginals was the highest performing model with a 61.27 % improvement in MAE over the benchmark. Overall, all models provided useful corrections to the base model, which has the potential to significantly improve in cases where more complex, black box base models would be acceptable.

4.4.1 Future Work

There are opportunities for the presented method to be developed and refined. This work calibrated long-term predictions provided from a simple base model due to the requirement to reflect a transparent and well-studied model that is used by the heavily regulated nuclear sector. These constraints may be more flexible in other industries with looser regulations and there exists the opportunity to use alternative, more advanced, base models that may lead to further performance improvements. Furthermore, the computational efficiency of this method could be improved as discussed within section 4.3.1 by the use of a surrogate model to replace computationally expensive CDF calculations. This would allow more effective support for larger models to either include other variables of interest or support longer prediction horizons. Other extensions to this method may include the inclusion of other relevant system parameters and the use of multi- or cross-modal data.

The generality of the Copula-based approach presented in this work imparts to it a wide domain of applicability. The work builds on applications in finance [126], biotechnology [124], hydrology [130], and power distribution [137]. In this work, the method was demonstrated on bearing data (section 4.3.3) which is critical to many other types of rotating plant, e.g., motors, pumps, or generators, found across a range of industries ranging from manufacturing, mining, or oil and gas. In addition, temperature is a crucial measurement for process control often encountered within chemical engineering or food processing equipment which may benefit from the type of predictive model

presented in this work. In future, the proposed method could also be applied to spatial or spatio-temporal data, for example, where the lagged timeseries data featured in this work could be replaced with spatially adjacent data on a graph network and the dependency structure between these adjacent points can be captured through the Copula. Such scenarios may be relevant for tasks in the field of structural health monitoring.

Chapter 5

Uncertainty in space: Quantifying spatial uncertainty to validate affordable structural health monitoring test rigs

5.1 Photometric stereo test rigs for structural health monitoring

Many industries rely on the integrity of large, complex structures to ensure safe and reliable operation. Structural health monitoring (SHM) allows the condition of an asset to be monitored, supporting early detection and assessment of structural damage which is essential for the development of cost-effective maintenance strategies. Non-destructive SHM solutions exist which can allow surface damage to be recreated virtually for more indepth assessment and the monitoring of the rate of degradation over time. One such solution is photometric stereo. Photometric stereo photographs surfaces under various lighting conditions to determine the topology of the surface, allowing detailed recreation of damage shape and size. The virtual meshes created with this method can allow operators to make complex decisions based on the location and geometry of the

surface damage. However, inaccuracies or limitations in the capability of photometric stereo based assessments can result in misinformed maintenance decisions, potentially leading to either unnecessary maintenance being conducted or the worsening of surface damage from lack of intervention.

Across many engineering applications, monitoring is required in areas that may be hazardous, space-constrained or time-consuming to reach. In some cases, types of contamination from severe environments may require the monitoring device to be disposed of or discarded after only a few uses. As such, potential visual inspection rigs must be portable, lightweight and affordable to promote usage in industry. However, a compromise may have to be made in terms of accuracy of the rig due to the constraints placed on its cost or size. The limitations introduced by these constraints can be addressed to reduce the impact on the confidence of the rigs measurement of damage shape and depth through the use of additional validation procedures using data-based models and uncertainty quantification, which is the focus of this chapter.

In this work, the analysis and quantification of uncertainty sources in a photometric stereo test rig was conducted through an intersystem comparison against a well-characterised, laser-based method (coordinate measurement machine (CMM)). An open source dataset was collated and published containing a variety of physical objects measured by the CMM and photometric stereo rig, and also a virtualised version where the virtual rig, rendering process and virtual objects were curated for further analysis by other researchers or future work. Three calibration methods were applied to quantify the rig error and uncertainty with various levels of complexity: from the self-calibration of the rig on a blank background; to residual dependency modelling using polynomial regression; and high dimensional copula models. Within non-contact optical measurement, a calibration object is often used to calibrate equipment [232]. This process is accounted for in this work through calibrating the rig against a flat surface. Deep learning models capable of uncertainty quantification have been applied to spatiotemporal data applications including air quality and epidemic modelling [233]. However, as discussed in Chapter 2, while the explainability of deep learning models are being actively improved, they are still currently unsuitable for highly regulated environments, such as

the nuclear industry. In geology applications, Kriging (based on Gaussian Process regression) is used to provide interpolations between samples while providing uncertainty quantification for 2D or 3D applications [234]. However, this suffers from the same limitations discussed in Chapter 2; while the scalability of Gaussian Process Regression is being developed, it is currently restrictive to smaller dataset sizes. The scale of the datasets in this application is much larger, at $\tilde{3}00$ k data points per object. Copula models have been utilised in spatial data applications to capture dependence between weather variables [125], or spatial and temporal dependence for rainfall prediction [235]. The limitations of the other methodologies with respect to the constraints imposed by application domain and dataset size, alongside the successful demonstration of copulas for spatial dependence modelling, supports the use of copulas in this work. Specifically, copulas are used to capture the dependency between lagged base model prediction error to provide model calibration and uncertainty quantification on a spatial dataset.

5.1.1 Contribution and novelty

Part of the work presented in this chapter is a published article [14] which contains the extended discussion of the experimental design and data collection process, with the data made open source for other researchers benefit. The work was conducted in collaboration with the Civil Engineering group at University of Strathclyde, UK, who provided access to the photometric stereo rig and associated software, while the CMM access was provided and supervised by National Physical Laboratory at Huddersfield, UK. Modelling and analysis work was jointly supervised by the Industrial Informatics group at the University of Strathclyde and the Data Science team at the National Physical Laboratory, Teddington, UK.

The contribution of this chapter broadly covers the data set collection and curation, and the development of a data-based analytics framework for the collected data.

- Dataset creation contribution:
  - The dataset collected and curated as part of this work provides a set of benchmarks for understanding the uncertainty sources between 3D geome-

tries and their 2D renderings under various lighting conditions.

  - Objects included in the dataset are collected from built environments with artefacts of damage, natural degradation, and high frequency surface textures relevant to engineering disciplines. This includes the presence of cracks and spalling damage which, depending on the fidelity of the applied photometric stereo method, would result in different consequences in civil engineering maintenance applications.
  - The materials represented in the dataset are relevant to civil applications (concrete), and further diversified to additional disciplines with inclusion of clay and common plastic used in 3D printing (PLA).
  - Additionally, the dataset contains synthetic data with high resolution concrete textures to allow analysis and comparison of experiment virtualization processes with lab collected data.
  - Potential further applications of the dataset beyond those developed in this work apply to researchers interested in developing methods for improving the accuracy of photometric stereo assessments and 3D printed objects. The data could be used to validate new photometric stereo algorithms by providing photometric stereo input information and their 'ground truth' mesh comparisons or developing new methods of generating and validating synthetic data.

- Modelling and analysis contribution:

  - Analysis in this work addresses the unknown inaccuracy of an affordable, portable test rig designed for visual inspection of civil engineering assets. This is achieved through the application of data-based models in a hierarchical modelling structure designed to calibrate predictions made to quantify the rig error and uncertainty.
  - The rig geometry was embedded into useful feature data to perform data based analysis using an explainable base model (polynomial).

- The copula models utilised predictions made on nearest point cloud points to predict the polynomial residual at a point of interest, providing calibration for improved predictions and uncertainty quantification.
- The statistical models in the hierarchical modelling structure allow risk to be assigned to both the data-based model predictions and the rig overall performance. The approach used is translatable to other rig designs where the trade off between cost and accuracy may result in a compromise, and where rich data reserves have been produced for use to improve predictions post rig application.

The rest of this chapter is organised as follows: Section 5.2 covers prerequisite literature including visual inspection in structural health monitoring applications and photometric stereo fundamentals. Section 5.3 covers the intersystem comparison and experiment virtualisation processes undertaken as part of the data collection process, along with the pre-processing steps for data curation, and the objects chosen for inclusion in the dataset. Section 5.4 covers the analytics approach developed to provide further analysis of the errors present in the rig, with discussions on the modelling results and computational considerations. Finally, Section 5.5 contains the conclusion and future work.

5.2 Literature: Photometric stereo and visual inspection for structural health monitoring

5.2.1 Visual inspection in structural health monitoring

SHM [236] allows the condition of an asset to be assessed to support maintenance decision-making across a wide range of industries. Visual inspection is a non-destructive testing method [237] focused on detecting visible surface damage which may indicate the need for intervention due to compromised structural integrity based on the location, damage severity and type of damage present. Common surface conditions include cracks or spalling [238, 239], corrosion [240], or deposits which may be due to a build up of

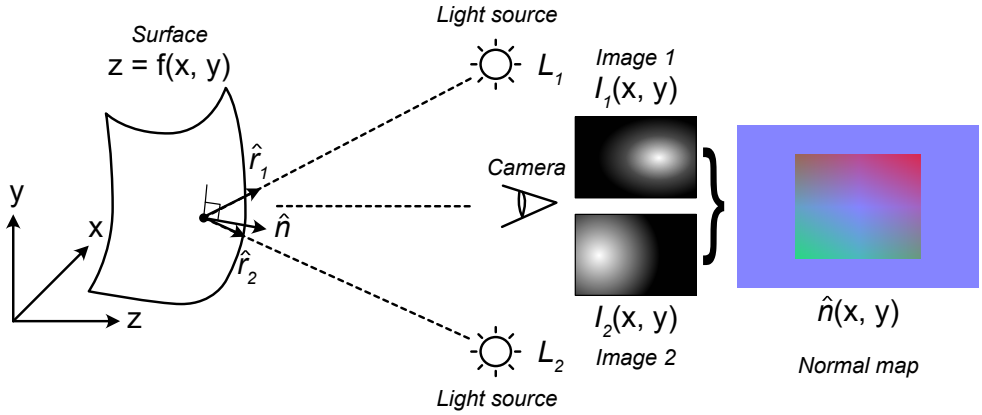

Figure 5.1: Diagram of the basic photometric stereo process, based on images captured from a constant viewing angle using multiple light sources. The image intensity data can be used to reconstruct the surface normals.

foreign material or wildlife encroachment [241]. In aviation, there is a growing interest in visual inspection procedures due to the deployment of new composite materials in aircraft which may experience novel defect modes, requiring further study [242]. In naval applications, visual inspection based methods are required to automate ship hull inspection processes [243]. Visual inspection techniques can be combined with data-based approaches to allow for improved detection of defects, such as in [77] where image stitching and processing was used to provide defect detection in nuclear plant fuel channels; or in [244] where classifiers were used to differentiate treatment between nuclear waste materials. In [245] and [246], a directional lighting rig was used to enhance the detection of surface defects in concrete through the application of data-based convolutional neural network models.

5.2.2 Photometric stereo

With a stationary camera and multiple light sources of known direction, photometric stereo algorithms can construct 3D representations of an object or surface from a collection of 2D images [247]. An outline of the photometric stereo process is provided in Figure 5.1. The morphology of a surface, $z = f(x, y)$, is characterised by a set of unit normal vectors, $\hat{n}(x, y)$, which describe the local orientation of the surface. These unit normals also largely determine the direction that incident light will be re-

flected. The appearance of a surface under directional lighting is the combination of the surface geometry, the lighting type and quality, and the surface material reflectance behaviour [248]. Combining pixel intensity, $I(x, y)$, across several images with the known incident light angles allows the determination of the surface structure based on several assumptions related to the surface reflectance behaviour and type of lighting. The captured intensity values encode the orientation of surface normals based on how the light was reflected alongside knowledge of the location of the light source relative to the surface (characterised in Figure 5.1 as the unit directions $\hat{r}_1$ and $\hat{r}_2$). In this way, a normal map can be estimated directly from the intensity mappings. Traditional photometric stereo algorithms assume light sources are point sources, which are far enough from the surface of interest to result in light rays which are parallel to the surface [249], additionally, it is assumed that the surface possesses Lambertian reflectance [250] qualities (diffuse reflectance which is not dependent on view point), both of which are often violated in practice [251]. However, addressing issues from unknown lighting direction, different surface reflectance characteristics or even moving objects remain active areas of research [252, 253]. Due to this increased practicality, photometric stereo has been applied across many diverse applications which desire to recreate surfaces in high detail using low-cost, easy to use and portable equipment. For example: capturing historical artwork in remote caves for heritage digitization [254]; diagnosing skin conditions in dermatology [255]; component defect detection for improved quality control in manufacturing [256]; or a proposed inspection method for structural health monitoring of the reactor pressure vessel in nuclear power plants [257].

5.3 Spatial data collection and experimental design

The data collection methodology covers two routes detailed in Figure 5.2: the intersystem comparison process covers the collection of data for objects under laboratory conditions using a highly characterized CMM and a photometric stereo test rig intended for use in SHM applications; and the experiment virtualization method develops a virtual, idealised version of the photometric stereo test rig in 3D software to generate renders of virtual objects [14]. The advantage of exploring both routes of data collection is the

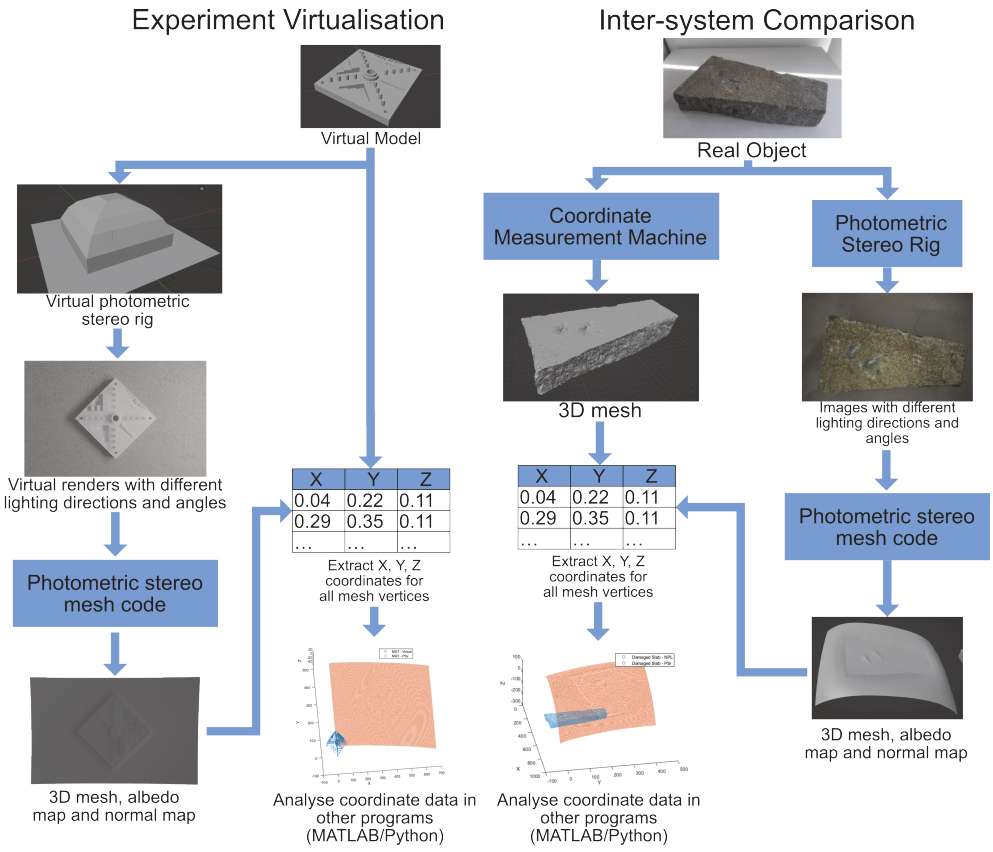

Figure 5.2: Diagram of intersystem comparison and experiment virtualization workflows

decoupling of potential limitations of the photometric stereo rig implementation (hardware) and chosen photometric stereo algorithm (software). Additionally, the processes and outputs of both routes are symmetrical to facilitate further comparisons between the full rig system and the software components. In this work, attention is focused on the outputs of the intersystem comparison to initially analyse the full system, however the choices made in the experimental design allows for further, more expansive studies in future.

5.3.1 Intersystem comparison

Physical object selection

The intersystem comparison covers data collected by the CMM and the photometric stereo test rig. The study covered 9 physical objects with 5 different materials, shown in Figure 5.3. The objects were chosen to represent a range of surface features and geometries, such as the plaster cylinder and sphere covering primitives; the plastic 3D printed NIST Additive Manufacturing test artifact [258] acting as a feature reconstruction test; the ceramic household objects such as the rabbit, train and coral, covering intricate, domestic objects; and the concrete damaged slab, chimney liner segment (unknown material) and broken brick covering structural components of interest to civil applications.

Primitives are described by simple mathematical equations and are some of the simplest "building blocks" which can be used to approximate real world objects [259]. The addition of primitive shapes, such as cylinders and spheres, allows the rig to be tested on basic geometries before introducing the complexity of application specific objects. For example, the chimney liner can be reasonably approximated as part of a cylinder. Additionally, the generalisation of the data set was improved through the addition of "household" objects, such as the rabbit, train and coral, which parallel the standard reference objects in photometric stereo algorithm development and wider computer graphics applications. For example, the Stanford bunny and Stanford dragon [260] are part of a 3D scanning repository[1] widely used in computer graphics for applications

[1] https://graphics.stanford.edu/data/3Dscanrep/

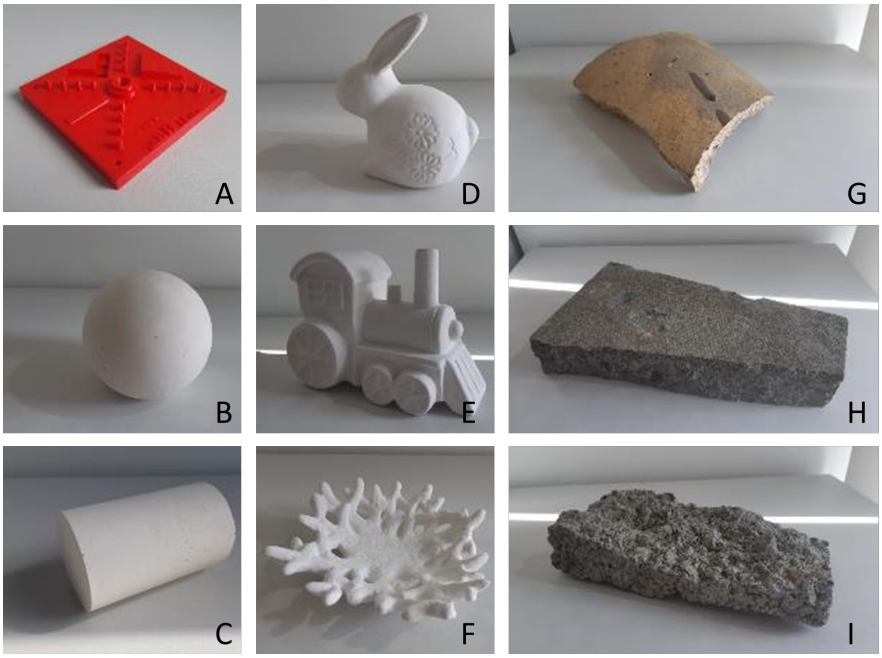

Figure 5.3: Objects used in the intersystem comparison study, A) 3D printed NIST Additive Manufacturing test artifact (2.85 mm PLA) (approx. 99 x 99 x 17 mm), B) Plaster of Paris sphere (approx.. 97 mm diameter), C) Plaster of Paris cylinder (approx.. 60 mm diameter, 97 mm length), D) Ceramic rabbit (max dimensions approx.. 110 x 67 x 115 mm), E) Ceramic train (max dimensions approx.. 150 x 85 x 112 mm), F) Ceramic coral (max dimensions approx.. 150 x 145 x 42 mm), G) Chimney liner (unknown material) (max dimensions approx.. 200 x 145 x 20 mm), H) Damaged concrete slab ((max dimensions approx.. 270 x 144 x 50 mm), I) Broken concrete brick (max dimensions approx. 212 x 94 x 45 mm)

including surface reconstruction algorithms [261], physics simulations of material fracturing behaviour [262] or artistic rendering of trees and fur [263]. The trend of using such figurines and statue objects persists in more recent photometric stereo datasets such as in [264] with cats and frogs[2], and in [252] with bears and teapots[3]. To specialise the dataset for the intended civil application, the chimney segment, broken brick and damaged slab were included. These present typical wear and tear from exposure to outdoor environments along with manually added impact damage to include more elements of interest to the surface (such as for the damaged slab). Additional debris such as loose material, dirt, plant matter or insects were removed prior to any measurements to improve consistency between the CMM and PS. However, there may be opportunity to further investigate how additional surface contamination may impact the photometric stereo process due to different light reflectance behaviour. Most objects in the study present Lambertian qualities, where the surface is matte and diffuse. The closest the dataset comes to including more specular qualities is the PLA plastic used to print the NIST Additive Manufacturing test artifact as it is more reflective than the other materials used. Fully specular data is being investigated in other data sets as in [265, 266] where metals are included.

Photometric stereo rig

The photometric stereo rig consists of a plastic hood surrounding a camera (Blackfly USB bfs-u3-200s6c, 8 mm lens) with 4 white LED array strips on each of the 4 sides (resulting in 16 LED strips). A box attached to the rig hood contains the electronics required to automate, run and store the collected data, which includes a single board computer and interchangeable battery packs. The rig is portable with dimensions of 54 x 54 x 27.5 cm, and lightweight, with much of the weight made up of the battery packs and camera, allowing it to be maneuvered fairly easily. The rig is made of affordable materials which are widely available, ensuring any repairs or part replacement can be easily achieved. Lastly, the rig is easy to use, as the data collection process is automated with the camera calibrated before each collection to allow for consistency between runs;

[2] https://vision.seas.harvard.edu/qsfs/Data.html
[3] https://sites.google.com/site/photometricstereodata/single

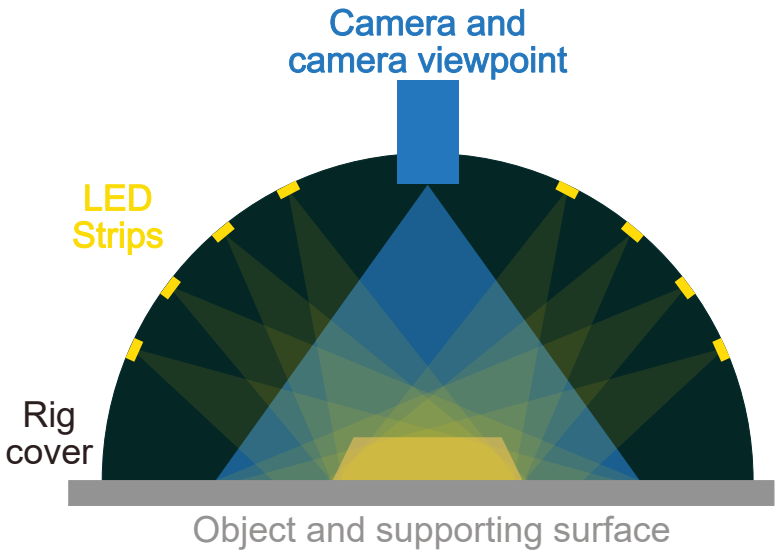

Figure 5.4: Simplified diagram of the cross section of the photometric stereo rig showing its key features: rig cover; LED strips and their lighting path; vertical camera and its viewpoint; and, the object being scanned on a supporting surface.

and the easily interchangeable battery packs are accessible to prolong battery life. An annotated diagram of the cross section of the photometric stereo test rig is shown in 5.4, with the virtual version shown in Figure 5.8 showing the rig in 3D.

The camera and lighting are trained on the centre of the resting surface of the rig, directly below the camera, with working distances of 250 mm. The positioning of the LED strips allows the object to be illuminated from 4 sides, at 4 angles (10, 30, 50 and 70 degrees to the horizontal). For each experiment, the camera is calibrated under 70 degree diffuse lighting, where each lighting direction at the same angle of 70 degrees illuminates at once. An image is taken for the object illuminated by each LED strip at each angle, and an image under diffuse lighting is taken for each level, where the 4 LED strips at the same angle light simultaneously. This produces 20 images – 4 for each 'ring' of LED strips at the same angle to create the diffuse images and 4 different angles on 4 different sides (an image for each of the 16 LED strips). Example images are shown in Figure 5.5. These images, along with information concerning the rig and camera set up are given to a proprietary photometric stereo algorithm to generate the

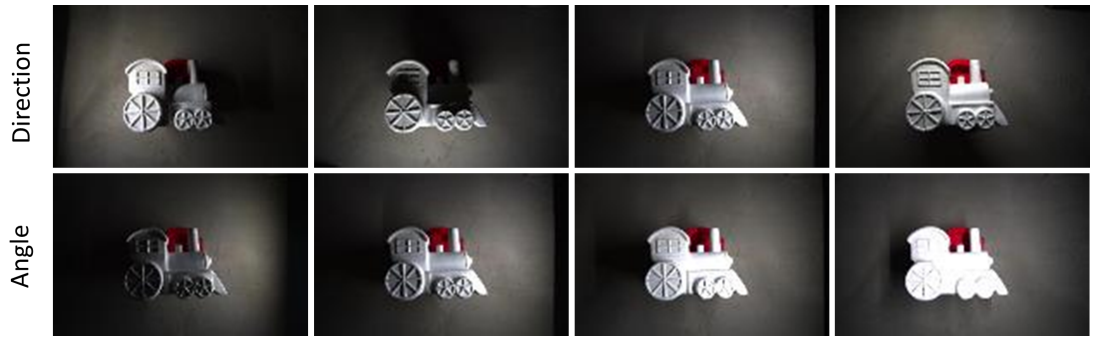

Figure 5.5: Example output images from the photometric stereo test rig with the train object. The top row shows the different lighting directions (images left to right): lighting from left, bottom, right, top. The bottom row shows the lighting angles (images left to right): object lit from the right by lighting at 10 degrees, 30 degrees, 50 degrees and 70 degrees to the horizontal.

surface meshes. At the time of writing, the software provided by the Civil Engineering Department at the University of Strathclyde used to generate the mesh output is not open source, however, it is feasible to apply other photometric stereo algorithms to the data collected due to the lighting and image data provided. Further information on the camera and lighting calibration can be found in previous work [267].

The experimental set up for the rig involved elevating the rig on 10 cm high supports at each corner above a flat concrete slab. Objects are placed centrally under the rig, directly below the camera. The rig supports allowed objects to be removed from the rig easily while maintaining alignment. Due to the height of the supports, objects were also placed on supports to allow the area of interest to be within the working distance of the camera and lights. After the objects were placed, a dark shroud made of thick black fabric was placed over the test rig to prevent external light leakage. The test rig automated the lighting regime and image collection process to generate the image data. The number of runs per object, different support elevations or object orientations are shown in Table 5.1.

Limitations due to the photometric stereo test rig include the nature of the camera focus which introduces a depth limitation to capturing sharp images. If the depth of the object exceeds this range, parts of the image may be unfocused which will impact the output mesh. Additionally, the camera and lights on the test rig are focused on a

Object name	Number of runs	Object elevations	Object orientations
Sphere	3	1	1
Cylinder	2	1	2
NIST Additive Manufacturing test artifact	1	1	1
Rabbit	2	1	2
Train	2	1	2
Coral	2	2	1
Chimney liner	2	1	2
Damaged slab	2	1	2
Broken brick	2	2	1
Blank background (no objects)	2	2	1
Blank background (steel rule)	2	2	1

Table 5.1: Number of runs and number of experiments with different object orientations for photometric stereo data on real objects.

certain distance where the objects were elevated to by supports. Any inaccuracies in the height of these supports may impact the camera focus and lighting quality. For the camera, it was found in [267] that 200 mm - 450 mm provided acceptable clarity. For consistency between objects and to minimise the impact of the lighting misalignment, before each experiment, the supports were adjusted to suit the object being measured and the height to the top surface was measured to be as close to 10 cm as possible. This approach is sufficient for objects like the slab with a relatively flat surface, but becomes more ambiguous for object like the rabbit which are round. After each experiment, the output images were previewed to check for obvious blurring or illumination issues. In situ, this would be less likely to occur because the rig is built to focus the camera and lighting on the surface it is placed on, making it much more suitable for large, flatter surfaces such as walls, floors or supports. While curved surfaces are covered in this dataset by the primitives or chimney objects, further experimentation could be done with the rig either physically or virtually to understand its limitations, if any, on larger curved surfaces such as containers or cylindrical supports.

Object name	Number of runs	High/low user experience level	Object orientations
Sphere	7	6/1	1
Cylinder	5	1/4	1
NIST Additive Manufacturing test artifact	1	1/0	1
Rabbit	1	1/0	1
Train	1	1/0	1
Coral	1	1/0	1
Chimney liner	2	2/0	2
Damaged slab	2	2/0	2
Broken brick	1	1/0	1

Table 5.2: User experience level, number of runs and number of experiments with different object orientations for CMM data.

Coordinate measurement machine

The CMM used in this study is an articulated arm, model Hexagon Absolute Arm 7-Axis, with a laser scanner end effector, model Hexagon Absolute Scanner AS1, as shown in the manufacturing guide[4] in Figure 5.6. This type of CMM records the location and orientation of the end effector by measuring the rotational position of the joints using precision encoders and subsequently inputs that information into a kinematic model of the arm. Uncertainties in measurements are attributed in line with ISO 10360-8 annex D, and data was collected directly into Polyworks[5] and converted to a polygonal mesh at the point of data collection. The data is exported as an .STL file which can be imported and processed in Blender [268] alongside the output meshes from the photometric stereo test rig.

The experimental set up involved the CMM in range of a support bench where the objects could be placed during the experiments. Experiments were taken multiple times for certain objects to capture the variance between measurements of the same object and were taken by operators with differing levels of experience to capture user error. The information on user experience level, number of runs and number of different orientations for each object is shown in Table 5.2.

[4]Hexagon AB, "Absolute Arm 7 axis", https://hexagon.com/products/absolute-arm-7-axis
[5]Innovmetric, "PolyWorks", https://www.innovmetric.com/

As shown in Figure 5.6, the CMM is a larger measurement device than the photometric stereo rig, as it is capable of a measurement range of over 2 m^3 with the weight estimated from the product page as an upper limit of 10.5 kg. With different support options, the portability of the CMM can be improved, however the base stability is highly important to the measurement accuracy and additional systems are in place within the CMM to raise alerts when the device is unstable. As this device is a complete product, there may be options for repair and replacement parts from the manufacturer but is unlikely to be easily sourced externally. Additionally, due to the fidelity and performance requirements of the CMM, it is likely to be much more expensive than the photometric stereo rig. The real time visualization of collected data allows for easier identification of gaps or poor coverage of the object, which assists new users in adapting to the initial complexity of operating the device. Additionally, the CMM has the ability to attribute uncertainty to collected data points and is able to collect multiple measurements of same place, replacing the points with higher accuracy measurements. This ensures repeatability across and within the same measurement process.

Physical data collection and pre-processing

The objects were transported between the photometric stereo rig at the University of Strathclyde, Scotland, UK to the CMM at the National Physical Laboratory in Huddersfield, England, UK. It is usually recommended to create specified registration points on the objects (such as adding spheres) for registration purposes to allow the meshes to be scaled and aligned. To allow safe transportation without damage to the objects or any movement of these added registration points, it was decided that spheres would not be added to the surface of the objects. The approach to registration taken in this work is handled in Section 5.4, but it is worth mentioning that other researchers [252] overcome this problem through the use of features in 3D software, such as Meshlab's mutual information method [269]. To permit the analysis of the meshes in alternative software, all meshes were loaded into Blender (Version 3.4) and a python script was developed to extract the x, y, z coordinates of each vertex in the

Figure 5.6: Manufacturer's buyers guide for the model of CMM used in the intersystem comparison.

mesh. This resulted in a point cloud of data points which represent the mesh vertices.

Traditionally, normal maps are of more interest when evaluating photometric stereo *algorithms* [252, 270, 271], however as the rig as a whole is being evaluated in this case, point clouds are of more interest due to the wider applicability and transfer of the methodology to other spatial data representations outside of PS, such as aligning with those used in the wider field of dimensional metrology [272]. The original mesh files and vertex coordinates were curated in the dataset [14], along with normal map representations of the output data for the photometric stereo rig outputs. It is also possible for normal maps to be extracted for the CMM and virtual objects through Blender (Version 3.4), as all mesh files generated in this process were compatible with this 3D software. The trio of point clouds, 3D meshes and normal maps allows for more potential approaches to analysing the rig in future. Alongside normal maps, heatmaps of height error are often used to demonstrate results from traditional photometric stereo algorithm validation [273], which are also utilised in the analysis part of this work in Section 5.4.

5.3.2 Experiment Virtualisation

Another method explored to validate the photometric stereo test rig was to develop a virtual version of the set up under ideal conditions. In the real rig, the positions and angles of the components may not directly align with the information provided to the photometric stereo algorithm due to measurement errors. Thus, investigating the impact of variables of interest on the quality of the output mesh may be time consuming, inaccurate and expensive to achieve on the real rig. However, creating lighting at many angles and intensities, or changing the shape and scale of the rig is possible with a virtual rig using precise dimensions. Generating data from virtual objects can remove the influences from inaccuracies in the rig design (such as faulty LEDs or camera misalignment) to test the performance of the photometric stereo algorithm used to convert images to meshes. Additionally, generating data from virtual objects can allow direct comparison with the ground truth for further validation or the additional exploration of materials.

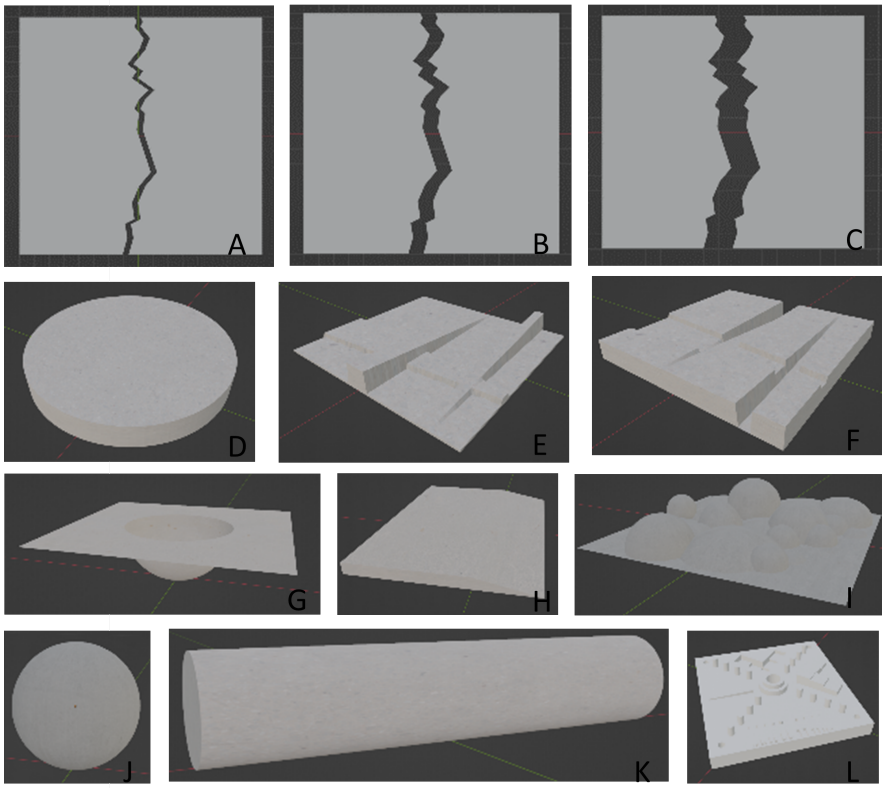

Figure 5.7: Virtual objects in the experiment virtualization study, A) Cracked slab with 0.5 cm gap, B) Cracked slab with 1 cm gap, C) Cracked slab with 2 cm gap, D) Vertical cylinder, E) Extruded channels with varying width, constant slope and varying slope, constant width, F) Indented channels with varying width, constant slope and varying slope, constant width, G) Plane with hemisphere indent, H) Slab with sloped edge, I) Interlocking spherical textured surface, J) Sphere, K) Cylinder, L) NIST Additive Manufacturing test artifact.

Virtual object design

Virtual objects created in 3D software act as the 'ground truth' for this method, as a perfectly accurate photometric stereo algorithm would aim to recreate the virtual object exactly. The 3D modelling software of choice is Blender (Version 3.4). To provide more realistic material behaviour, the albedo maps of 8 K resolution mappings of real concrete surfaces were used for the virtual object texture maps[6]. The 3 concrete materials were assigned over 12 virtual models, as shown in Figure 5.7. The design of the virtual objects were chosen to investigate the test rig performance on a variety of

[6]Quixel, "Megascans", https://quixel.com/megascans/

surface features: 3 objects with a cracked surface of different crack widths to measure the rig capability of capturing damage progression over time; 2 objects with channels of varying width and varying depth (both inset into the surface and extruded out from it) to test the precision limitations of the photometric stereo algorithm to gradual changes in surface deformations; a rectangular sloped slab and slab with spherical deformation were created to approximate spalling damage where the surface has been gouged or worn away; a slab with extruded spherical surfaces to understand the limitations of capturing complex surfaces with many shadows cast, not dissimilar to the broken brick object from the physical objects list; and finally, for consistency with the intersystem comparison, the NIST Additive Manufacturing test artifact [258], sphere and 2 cylinder primitives were created.

The choice of concrete texture maps was to maintain focus on a civil engineering context where concrete is a commonly found material. Additionally, the type of photometric stereo algorithm applied is designed for objects with diffuse, spatially-uniform reflectance (matte) as represented by the physical objects. As texture maps can be easily changed, there are opportunities for other algorithms to be applied and experimented with over a wider variety of materials in future. To further specify the dataset on the intended application, objects depicting simplifications of surface damage were created. The simplification was chosen at this stage as an approximation of much more detailed surface features, such as cracks or spalling. Once positive results were obtained on simplified geometries, more complex depictions of structural faults can be explored, as capturing these features in a reasonable manner is more difficult to justify. If not created digitally, there is also opportunity to have physical representations of these features captured in a 3D mesh (through the CMM, for example). For more general objects (such as the rabbit or train in the physical dataset) which are missing in the virtual dataset, there are a multitude of sources where virtual objects can be bought or procured for free through, for example, the Blender market place[7] or the Unreal market place[8] which are popular sources in digital art and animation.

[7] https://blendermarket.com/
[8] https://www.unrealengine.com/marketplace/en-US/store

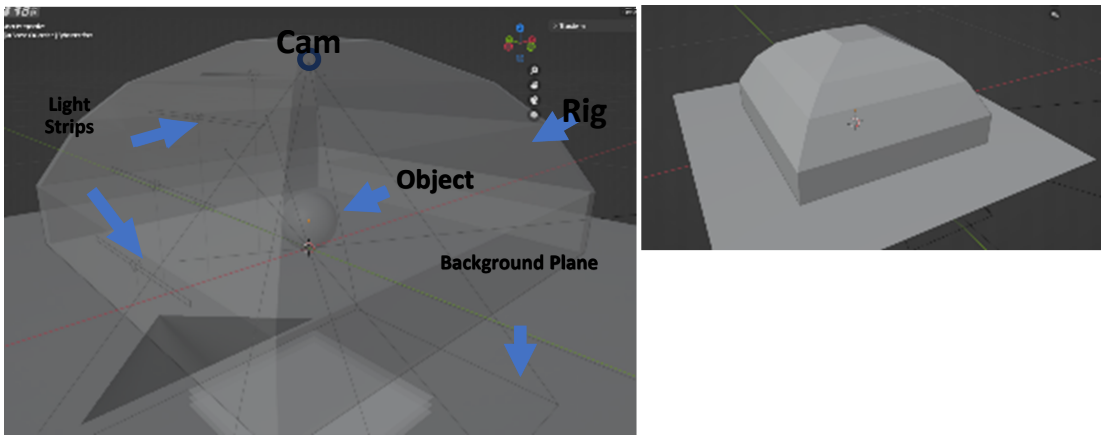

Figure 5.8: Virtual rig in Blender (Version 3.4), (left) annotated rig shown in semi transparent mode for component visibility, the parts of interest are the background plane, the rig cover, camera, lighting strips and virtual object, (right) shows the rig set up.

Photometric stereo rig virtualisation

Blender (version 3.4) was used to design the test rig, virtual objects and produce rendered images. Blender has been utilised in other experiment virtualisation research, such as in [274] where they investigated the utility of a virtual model in the optimisation of a physical multi-camera metrology system. The dimensions of the virtual rig were created from the information provided to the photometric stereo algorithm, namely: the working distance from the camera to the bottom plane of the rig; the LED proximity measured from the LED to the centre of the bottom plane of the rig; LED angle to the horizontal; LED brightness (given as 0-255, with 255 as maximum brightness); camera exposure, gain, red balance and blue balance. The horizontal and vertical distances of the lights were calculated from the working distance of 250 mm at a given angle (10, 30, 50 or 70 degrees to the horizontal) and the camera parameters were chosen to emulate the test rig camera model data sheet. The annotated virtual test rig is shown in Figure 5.8.

There are limitations to representing physical systems in Blender which may impact the behaviour of lighting and material interactions. This was counteracted through the use of high resolution scans of real concrete to define the virtual object materials, the object distances could be precisely defined, and the camera properties were emulated

as closely as possible to replicate the real rig under perfect conditions.

Virtual data collection and pre-processing

The render settings were chosen to balance fidelity and simulation runtime to a resolution which matched the photometric stereo test rig output images. The lighting and render regimes were automated through a python script with the image naming scheme and rig meta data produced to match that created by the real rig for compatibility. The rendered images and data files were processed by the photometric stereo software in an identical process as followed for the physical objects. The vertex coordinates for the virtual objects and the photometric stereo outputs on the virtual objects were extracted through Blender to be further analysed in different software. As mentioned, there are several approaches to overcome alignment issues between the meshes and further opportunity to explore other photometric stereo algorithms and material properties. In future, more work could also be done to diversify the rig design to potentially explore optimal designs or identify design limitations to further improve the physical rig. With the virtual objects there is increased flexibility to increase or decrease the number of vertices in the mesh to meet different data quantity or detail quality requirements. Additionally, there is an opportunity to automate more diverse experiments, such as the automating of crack progression and rendering to create an evolving dataset for future analysis.

5.4 Uncertainty quantification through spatial error modelling

With several methods implemented to provide diverse data for the investigation of the photometric stereo rig errors, an analysis approach was developed. Two data-based approaches were compared to both no intervention (assuming the rig is accurate) and the rig self-calibration on a reference case (no objects) to understand what level of intervention is required to capture the error in the photometric stereo rig. Additional data processing was required to unify and prepare the chosen data formats for further

Figure 5.9: Three industrial objects - slab, chimney and brick (left to right)

analysis and future deployment. Simple metrics were chosen which translate into real world impact, such as: the largest maximum and minimum deviations providing a measure of the most extreme errors; the mean absolute error (MAE) providing a general approximation of comparative error across the methodologies; and the median providing a non-Gaussian and directional estimate of the expected error.

5.4.1 Spatial data processing

The objects chosen for further analysis from [14] are the damaged slab, broken brick and chimney liner, shown in Figure 5.9. These objects are the closest representation to civil infrastructure in the provided dataset. A calibration benchmark of an empty frame was also taken for the photometric stereo rig as the supporting surface is expected to be flat.

Due to the different reference spaces between each method, the vertex coordinates of the CMM and photometric stereo point clouds are of a different scale, rotation and location in 3D space. The conversion from images to mesh results in a scaling factor based on the number of pixels used in the proprietary photometric stereo algorithm. To account for this, an image of a steel rule taken using the photometric stereo rig was analysed to quantify the physical distance captured in a pixel which is proportional to the resulting dimensions of the mesh. To transform and rotate the meshes, 3 or more reference points were taken on each mesh representing areas of interest on the

object, such as the deepest areas of damage, corners, or peaks of deposits on the objects surface. An optimisation was run to align the same reference points on each mesh and the transform and rotations applied to the full point cloud.

As the photometric stereo algorithm creates a mesh based on the full surface visible by the camera, a background surface is present in the meshes around the objects and the side of the object is not visible. For compatibility of the method on new, unseen photometric stereo meshes, a background plane was added the CMM meshes for compatibility and any features not visible from above were removed. Additionally, the CMM point clouds have a much higher density than the photometric stereo point clouds which is limited by the pixel density of the camera. Harmonisation was achieved through creating a flat plane of 400 by 400 points for each mesh and using a search function to sample the nearest point in X and Y, and taking the Z value of that nearest point. This allowed a reduction in the number of points in the CMM meshes to match the amount of points in the photometric stereo meshes, equalising both. The processed X, Y and Z coordinates of the point clouds are exported for further processing with a total of 294007 points per point cloud. An example plot of the CMM and photometric stereo point clouds for the slab before and after alignment are shown in Figure 5.10. As shown, the processed objects are centred on the origin, the photometric stereo point cloud is scaled to real space and aligned to the CMM mesh which has had the sides removed and a background plane added. For the photometric stereo data, the largest deviations are at the edges of the point cloud.

5.4.2 Spatial error modelling case study design

Three approaches are applied to model the error between the CMM and photometric stereo point clouds with varying levels of complexity. The first uses the error in the photometric stereo point cloud of a blank background (which is expected to be flat) to calibrate the other meshes. This is expected to account for certain non-ideal systematic behaviour in the rig due to imperfect lighting sources and positioning by comparing the actual photometric stereo estimate to a theoretical ideal (flat plane). The second method utilises a feature generated from the rig geometry to train a 3rd order poly-

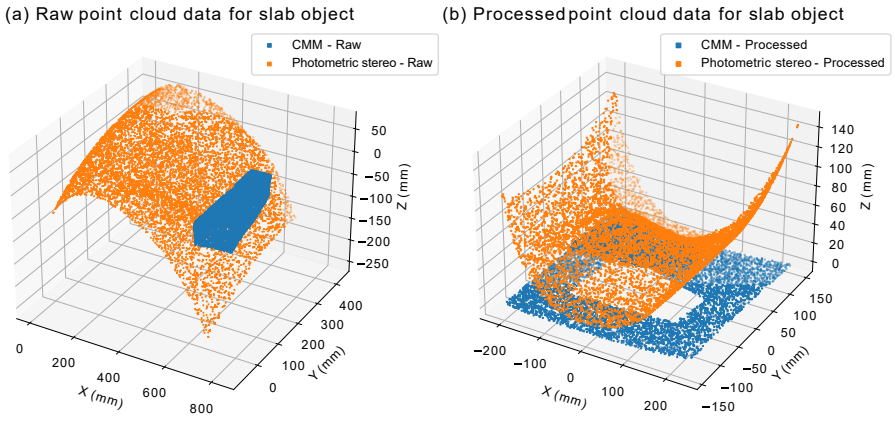

Figure 5.10: The slab object as: (a) the raw point cloud data exported from the CMM and photometric stereo rig scans; and (b), the aligned, processed point cloud data for the CMM and photometric stereo rig.

nomial model to predict the error in the Z coordinate between the photometric stereo and CMM point clouds, which provides additional information about the potential systematic errors from the non-ideal lighting sources. The last method is a hierarchical approach which combines the 3rd order polynomial models with high-dimensional copulas. The copulas are used to calibrate the residuals of the polynomial models by using the relationship between the polynomial predictions at the 4 closest neighbours to the polynomial residual at each given point. The error in the Z coordinate was the chosen target variable for all three approaches as this is significant to the estimation of damage severity - it may result in cracks or spalling appearing less or more severe, which will impact maintenance requirements. The modelling process for the data-based approaches is shown in Figure 5.11. For all heatmap figures discussed in this section, a randomly selected 1 % of points are plotted to reduce figure size due to the large dimensions of the datasets.

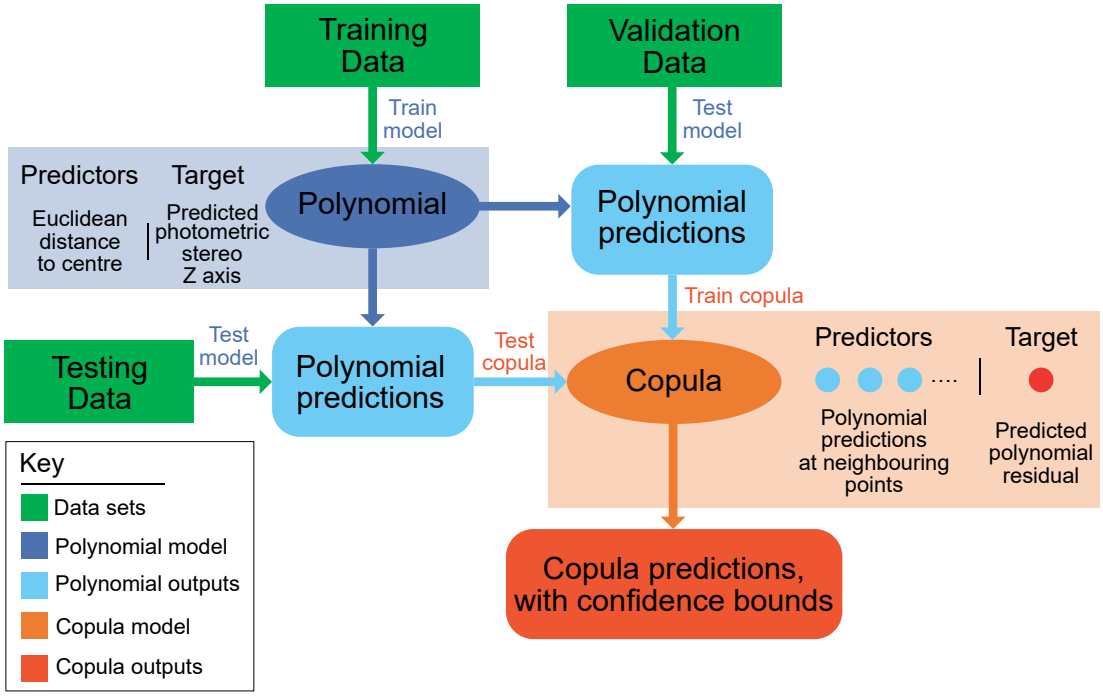

Figure 5.11: Data-based modelling process for the polynomial and hierarchical structure for applying the copula models.

Spatial error prediction using photometric stereo rig estimation of an empty frame

Due to the pre-processing steps taken in Section 5.4.1, the dimensionality between the blank background point cloud and the three industrial objects are identical, allowing 1 to 1 comparison of each point. The blank background point cloud captured by the photometric stereo rig contains many points which should have a Z coordinate of 0. However, large deviations at the perimeter and radial patterns are present, as shown in Figure 5.12. The point cloud is compared to an identical point cloud where all Z coordinates are estimated to be 0, and the calibration using this method becomes the inversion of the photometric stereo rig estimate (the values necessary to apply at each point of the photometric stereo point cloud to result in all Z values returning to 0). This calibration is applied to each point in the Slab, Brick and Chimney datasets and the results are discussed in Section 5.4.3.

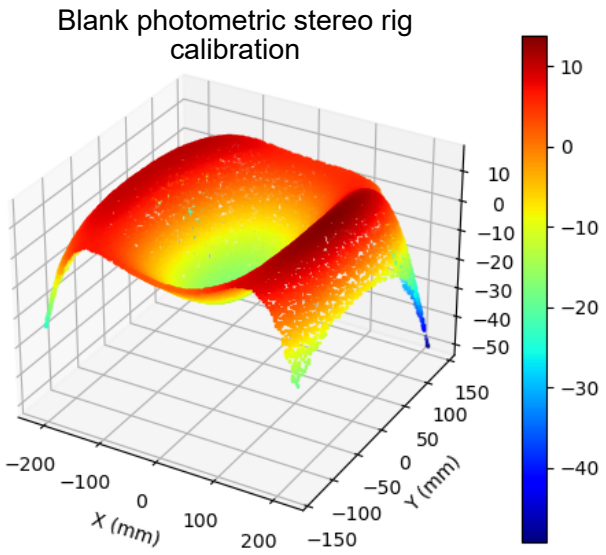

Figure 5.12: PS rig error in the estimate of a flat background (no objects). The true Z is at 0 across the whole point cloud, showing the large deviations and radial patterns in the photometric stereo rig error on a flat plane.

Spatial error prediction using 3rd order polynomial

Several parameters were considered as predictors for the error in the Z coordinate between the photometric stereo and the CMM point cloud which were derived from the design of the rig lighting. The lighting was expected to have a large impact on the resulting mesh error due to its high importance in the photometric stereo algorithm along with radial patterns observed in the errors between the CMM and photometric stereo rig point clouds. To encode the influence of the lights on the mesh, several additional features were created for every point in the point cloud, such as the distance and angle to the lights, or a points radial angle on the X,Y plane. The selected feature was the Euclidean distance between a point's [X,Y] coordinate on the mesh and the centre at [0,0] due to its consistency across all objects and polynomial trend. The Euclidean distance from origin against the error in Z is shown for the chimney object in Figure 5.13. The relationship between all considered geometric features and the Z coordinate error are shown in Appendix E, along with the Euclidean distance from

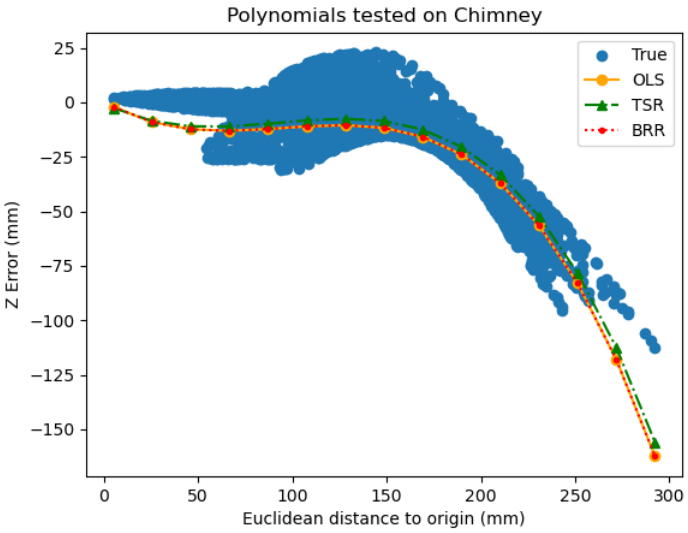

Figure 5.13: Polynomial comparisons tested on the Euclidean distance to origin against Z error for the chimney dataset (trained on combined blank background and slab datasets). The three polynomial models (Ordinary Least Squares, Theil-Sen regression and Bayesian Ridge Regression) follow a very similar trend in the data, and do not meaningfully outperform one another. This plot shows 1 % of the 294007 'true' data points.

origin for all datasets plotted in Figure E.1.

The Euclidean distance from the origin is used as the predictor for Z coordinate error to train a 3rd order polynomial model. Several models of different complexity were chosen to compare - Ordinary Least Squares (OLS), Theil-Sen Regression (TSR) and Bayesian Ridge Regression (BRR) as implemented in the Scikit-Learn python library. OLS is the simplest model, which aims to minimise the square distance between the model fitting and training data points which makes OLS models susceptible to outliers. TSR aims to build upon OLS by removing normality assumptions and limiting the impact of outliers through the use of the median [275] to estimate regression slope, which was made more robust by use of Kendells tau [276]. The ScikitLearn implementation is based on work by [277] and [278] who are credited with generalising the model and median function to multivariate cases, respectively. BRR provides uncertainty quantification through the addition of a predictive distribution [279] and is based on the algorithm in Appendix A of [280] and the parameter update process described in [281].

Table 5.3: Polynomial model candidate comparison metrics

Metric	Model**	Blank and slab*	Chimney	Brick
R^2	OLS	0.248	0.67	0.46
	TSR***	0.236	0.72	0.25
	BRR	0.248	0.67	0.46
MAE	OLS	20.54	10.24	11.05
	TSR***	20.26	9.79	13.33
	BRR	20.54	10.24	11.05

*Polynomial models are trained on blank and slab datasets.
**Differences between OLS and BRR show at scale of 1e-5.
***TSR oscillates between best and worst for each metric.

Table 5.4: Dataset organisation

Training	Validation	Testing
Blank Background, Chimney	Brick	Slab
Blank Background, Slab	Chimney	Brick
Blank Background, Brick	Slab	Chimney

However, the more complicated models were unable to meaningfully outcompete OLS as shown in Table 5.3, and so the simpler model was chosen for this analysis, minimising complexity and computational strain.

The polynomial models were trained on a data set consisting of the Euclidean distance from origin and Z coordinate error for two point clouds, then tested on a held out testing set consisting of one point cloud. The dataset organisation is shown in Table 5.4, with the polynomial method residuals for the slab testing set, brick testing set and chimney testing set discussed in Section 5.4.3. For the uncertainty quantification, an interval of 4σ are used for the BRR model, to represent a coverage of $\pm 2\sigma$, where σ is the standard deviation.

Spatial error prediction using high-dimensional copula calibrated 3rd order polynomial

In this work, four copula models are chosen to combine the complexity of different marginal assumptions and different methods of translating bivariate copulas to high-dimensional applications:

- Multivariate Gaussian with Gaussian marginals (MGG), capturing a standard assumption of Gaussian errors with linear relation.
- Multivariate Gaussian with best fit parametric marginals (MGB) (from Gaussian, Beta, Gamma or Truncated Gaussian univariates), capturing an assumption of non-Gaussian errors but with a linear relation.
- Centre vine copulas with Gaussian marginals (CVG), capturing the assumption of Gaussian errors but with potential tail dependencies between spatial error features.
- Centre vine copulas with Gamma marginals (CVB) (best fit univariate to target variable for all chosen case studies), capturing both Non-Gaussian errors with non-linear relations spatially.

The residuals of the OLS model trained on the Euclidean distance to origin to predict Z coordinate error are used to train high dimensional copulas to provide further calibration. The OLS model is trained on a data set consisting of the Euclidean distance to origin and Z coordinate error for two point clouds then tested on a validation dataset consisting of one point cloud. A K-nearest neighbours algorithm is used on subsets of the point cloud to identify the 4 closest neighbours to each point. The polynomial prediction at each neighbour is taken and the polynomial residual from the validation set is taken for each point. Each high-dimensional copula model is trained on the validation set and tested on the testing set.

To provide predictions of the required testing set calibration, the polynomial prediction at the closest 4 neighbours to the testing point is used to condition the copula model, and the density of the conditioned copula is used to provide a most probable estimate of the polynomial residual at the testing point with a 5 % and 95 % confidence bound on the estimate. The copula confidence bounds provide an estimated upper and lower correction for the polynomial, which are converted into a prediction interval by adding to the polynomial predictions. This captures the best and worst case risk for the estimated height. The dataset organisation is shown in Table 5.4, with the copula calibrated polynomial residuals discussed in Section 5.4.3. For the uncertainty

Table 5.5: Calibration methods residuals metrics

Dataset	Metric	Photometric stereo rig error	Blank rig	Polynomial (OLS)
Brick	MAE	28.498	*34.048*	14.968
	Median±std.	-25.901±17.607	*32.707±14.021*	13.519±10.077
	Max	7.739	*76.557*	44.69
	Min	-110.272	**9.843**	-21.578
Chimney	MAE	18.711	*21.551*	**9.92**
	Median±std.	-10.168±22.476	*15.75±20.531*	**-1.01**±12.419
	Max	23.328	*80.233*	26.872
	Min	-116.557	**-21.555**	-54.069
Slab	MAE	37.655	*47.82*	40.509
	Median±std.	-37.116±27.931	*45.412±20.313*	42.361±12.826
	Max	3.839	*100.7*	77.173
	Min	-146.332	**10.109**	10.451

Dataset	Metric	MG(G)[1]	MG(B)[2]	CV(G)[3]	CV(B)[4]
Brick	MAE	**10.763**	12.095	10.797	12.049
	Median±std.	**-7.84**±10.552	-10.457±10.163	-8.008±10.292	-10.152±**9.961**
	Max	29.786	24.879	23.886	**23.363**
	Min	*-48.653*	-46.626	-42.521	-48.354
Chimney	MAE	13.698	14.473	13.525	14.597
	Median±std.	-10.445±12.513	-11.715±**12.364**	-10.174±12.372	-11.747±12.379
	Max	22.669	22.303	**17.69**	17.887
	Min	-65.226	*-66.193*	-63.267	-65.204
Slab	MAE	20.129	18.028	19.944	**17.708**
	Median±std.	21.093±12.611	17.918±12.707	20.95±12.585	**17.15±12.214**
	Max	60.176	55.503	57.107	**51.197**
	Min	-13.089	*-18.931*	-10.855	-12.697

Values in *red bold italics* are the worst case and values in **blue bold** are the best case for each metric.
All results are in millimeters (mm).
[1] Multivariate Gaussian with Gaussian marginals (MG(G))
[2] Multivariate Gaussian with best fit marginals (MG(B))
[3] Centre Vine with Gaussian marginals (CV(G))
[4] Centre Vine with best fit marginals (Gamma) (CV(B))

quantification, the provided uncertainty bounds represent the 95 % confidence bounds converted to a prediction interval from the copula calibrated polynomials, which are used in the results discussion.

5.4.3 Spatial error prediction results and discussion

The results of the different methodologies are presented on the three case study objects in this section. The chosen metrics are used to compare approaches along with visualisations of the remaining error after each approach. Heatmaps and histograms of the remaining error after the application of each calibration method are shown to demonstrate the shape and scale of the residuals, and where they occur on the point cloud. All metrics are shown in Table 5.5.

Case Study 1: Broken brick

As shown in Table 5.5, the MAE from calibrating the brick object with the photometric stereo blank background is worse than the original error between the photometric

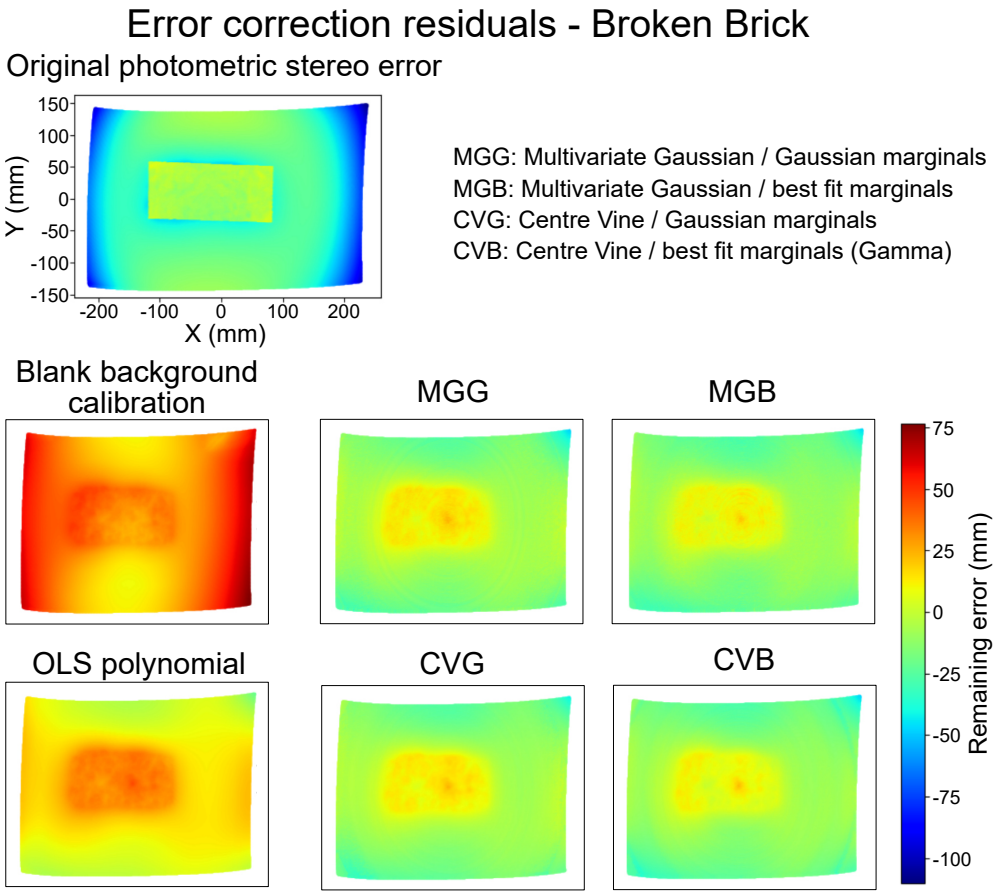

Figure 5.14: Heatmaps for the residuals of each correction method and original error for the brick object showing the spatial distribution of errors. All figures share the same axes. For the brick, the copulas result in the lowest MAE of all methods, with MGG having the lowest MAE of all models.

stereo and CMM point clouds. Both data driven methods (polynomial and copula calibrated polynomial) improve upon the original error between the photometric stereo and CMM point clouds, with all copula calibrated polynomial methods improving upon the polynomial. The polynomial improves upon the MAE by 47.5 % while the best copula model (Multivariate Gaussian with Gaussian marginals) improves by 62.2 %. As shown in Figure 5.14, the copula models are able to smooth out the large perimeter deviations from the original point cloud from dark blue (very large negative errors) to green (slightly negative errors) with the object moved from the orange and red (large positive errors) from the polynomial to yellow and orange (slightly positive errors). As with the slab, the photometric stereo blank background calibration method residuals

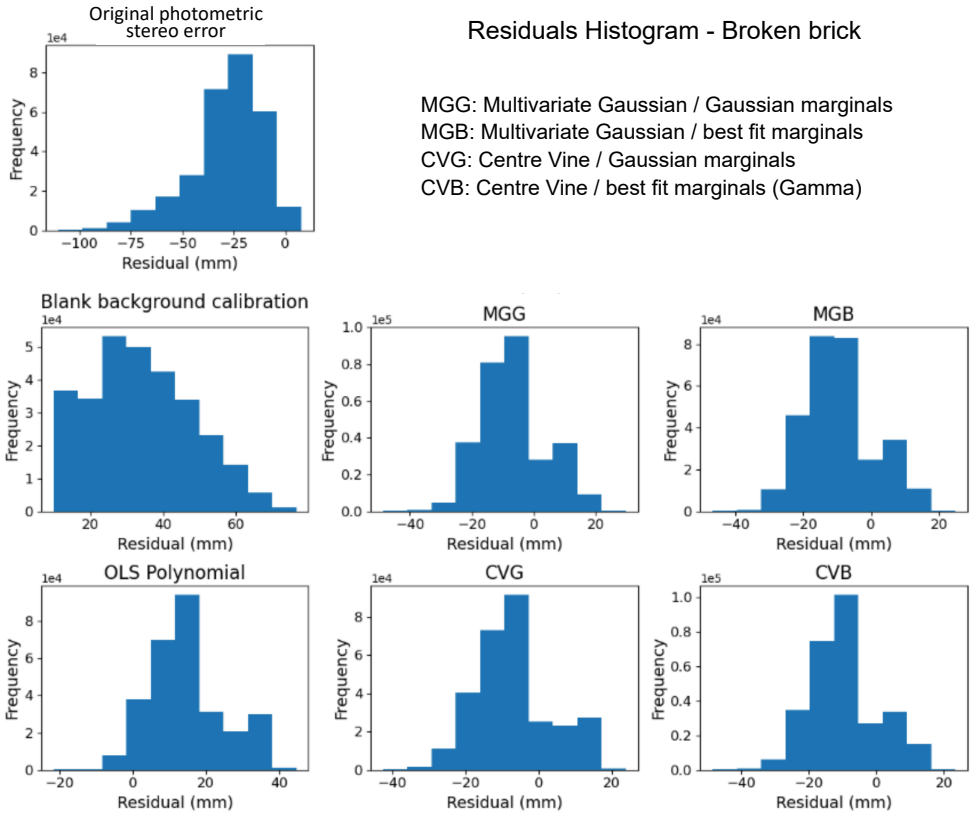

Figure 5.15: Histograms for each of the correction methods and original error for the brick object, showing the error distribution and extreme errors across the point cloud. The copula models have the lowest median value with the MGG copula having the median closest to 0.

also results in the highest median, standard deviation, with the maximum and minimum error both above 0 to prevent accurate predictions. The MGG copula holds the lowest MAE and median of all correction methods, but also the largest minimum error at - 48.653 mm. Meanwhile, the CVB copula holds the lowest standard deviation and maximum residual (23.363 mm). This suggests that, overall, the copula methods are able to recommend better corrections than the polynomial only, but may result in large deviations in areas of high uncertainty, as shown in the corners of Figure 5.16 for the centre vine models, but the centre bands for the Multivariate Gaussian models.

The histograms of the original error and all correction methods are shown in Figure 5.15. The original error is generally negative with a median of -25.9 with extreme errors of [-110.3, 7.7]. The polarity has been reversed for the blank background calibration

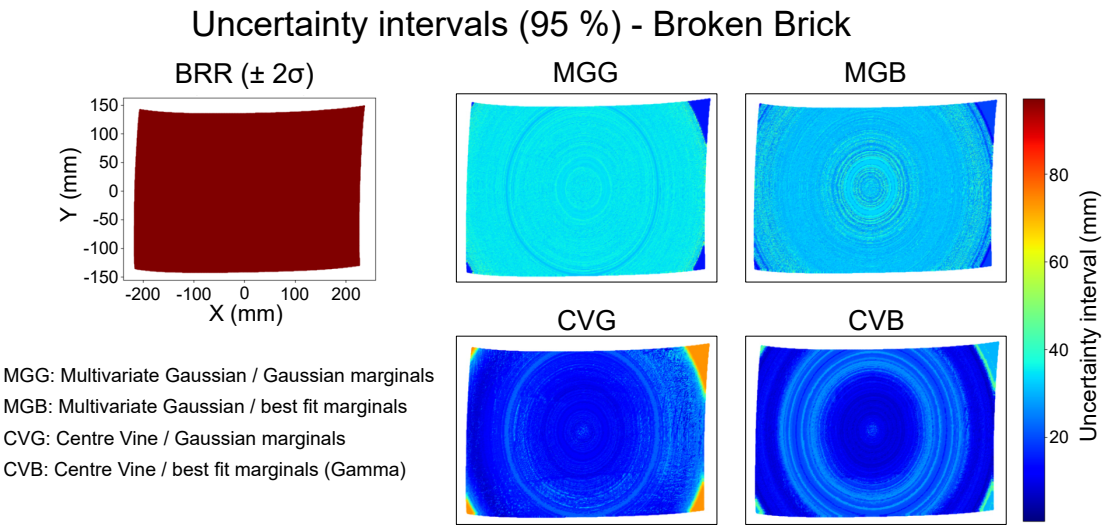

Figure 5.16: Heatmaps of the interval between the 5 and 95 % confidence bounds for the copula models and 4 ($\pm$ 2) σ for the BRR polynomial on the brick object. All figures share the same axes. The corner areas for the CVG copula model is the highest difference between the upper and lower confidence bound of all copula models.

method which has overcompensated and pushed errors into the positive with a higher median of 32.7 but lower extremes of [9.8, 76.6]. Some of the most extreme errors have been reduced, but at the expense of worse performance over the rest of the point cloud. Due to the extreme errors, the blank background calibration method never covers 0, and so is never able to accurately capture the rig error. The polynomial has a median of 13.5 and tends to underestimate the error, with much of the histogram higher than 0, which is remedied by the copula models. The copulas tend to overestimate but with median values closer to 0, with the closest being -7.8 for the MGG copula. All copula models have median values under those for the other methods.

Radial behaviour is present in Figure 5.16 for the copula uncertainty intervals due to the Euclidean distance to origin parameter used to train the polynomials compressing the 2D representation to 1D, where similar Euclidean distance to origin values result in similar uncertainties. This also varies with trends identified by the copulas, for example, all corner points are not equally uncertain, which matches the different behaviour of the original error and OLS polynomial from the left to the right side. The standard deviation value given by the BRR is not a constant value, but varies so little across the

point cloud that the difference is imperceptible on the common colour scale. As the standard deviation value is much larger and more constant than the copula models, it demonstrates that the BRR model is unable to provide more useful or more certain information about the uncertainty across the point cloud than the copula models. The highest uncertainty for the copula models, measured by the largest difference in upper and lower prediction interval, is in the corner regions of the CVG copula model. This trend is similar but less severe for the more flexible CVB copula.

Case Study 2: Chimney liner

The worst performing method in terms of MAE for the chimney object is calibrating using the photometric stereo blank background which has a higher MAE than the original error but is the lowest MAE for this approach across all 3 objects, as shown in Table 5.5. All data driven approaches improve upon the original error between the photometric stereo and CMM point clouds, but for the chimney object, the polynomial only approach has the lowest MAE. This is expected to be due to the photometric stereo rig behaviour which attributes curves to objects (see the error in Figure 5.12), and the chimney object being naturally curved, resulting in a lower error to correct which is more easily captured by the Euclidean distance to origin feature (as shown by the R^2 score for the chimney dataset in Table 5.3 and Figure 5.13). As shown in Figure 5.17, the polynomial smooths out much of the large positive and negative errors in the original point cloud, leaving the most visually obvious errors as slightly positive in the centre of the mesh. While the copulas seem to address the error from the polynomials in the centre to move it more towards green and yellow (slightly positive above 0), the errors in the background of the object have deepened to blue (slightly negative) which would degrade the overall copula performance.

The histograms for the residuals across the whole point cloud are shown in Figure 5.18. The original photometric stereo error has the smallest median compared to all case study objects at -10.168, which is only rivaled by the polynomial median of -1.01. All copula models have a smaller median than the blank background calibration, but are all larger than the original photometric stereo error. The smallest copula model median is the CVG at -10.174. The polynomial method residuals has almost matched the upper value for the original error while limiting the lower value with extremes of [-54.1, 26.9] for the polynomial and [-116.6, 23.3] for the original error.

Figure 5.19 shows the uncertainty quantification from the BRR and copula models. Overall, the centre vines tend to be confident in the centre and uncertain at the corners (where extreme errors tend to appear on the original photometric stereo error), whereas the Multivariate Gaussian models are more confident in their centre regions, shown by the narrower prediction interval. The CVB model has less severe uncertainty at the

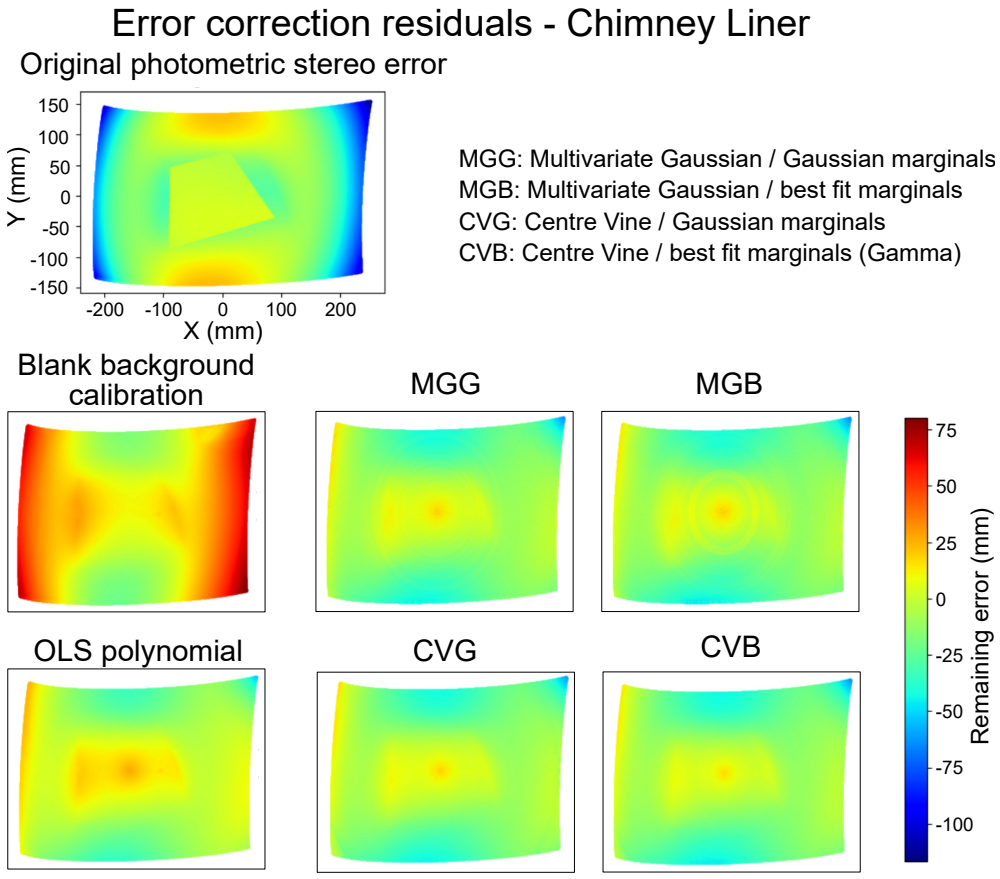

Figure 5.17: Heatmaps for the residuals of each correction method and original error for the chimney object showing the spatial distribution of errors. All figures share the same axes. For the chimney, the OLS polynomial results in the lowest MAE of all methods. The chimney object has the lowest original MAE error of all case study objects.

corners than the less flexible CVG, however, the MGB is less confident in the centre and extremes than its less flexible counterpart, MGG. The CVB is more confident (smaller intervals) but has a higher MAE than its Gaussian assumption counterpart (CVG), while the MGB model is less confident (larger intervals) and higher MAE than its Gaussian counterpart (MGG). In this case, the Multivariate Gaussian models may be able to provide more accurate uncertainty quantification, which aligns with each models performance.

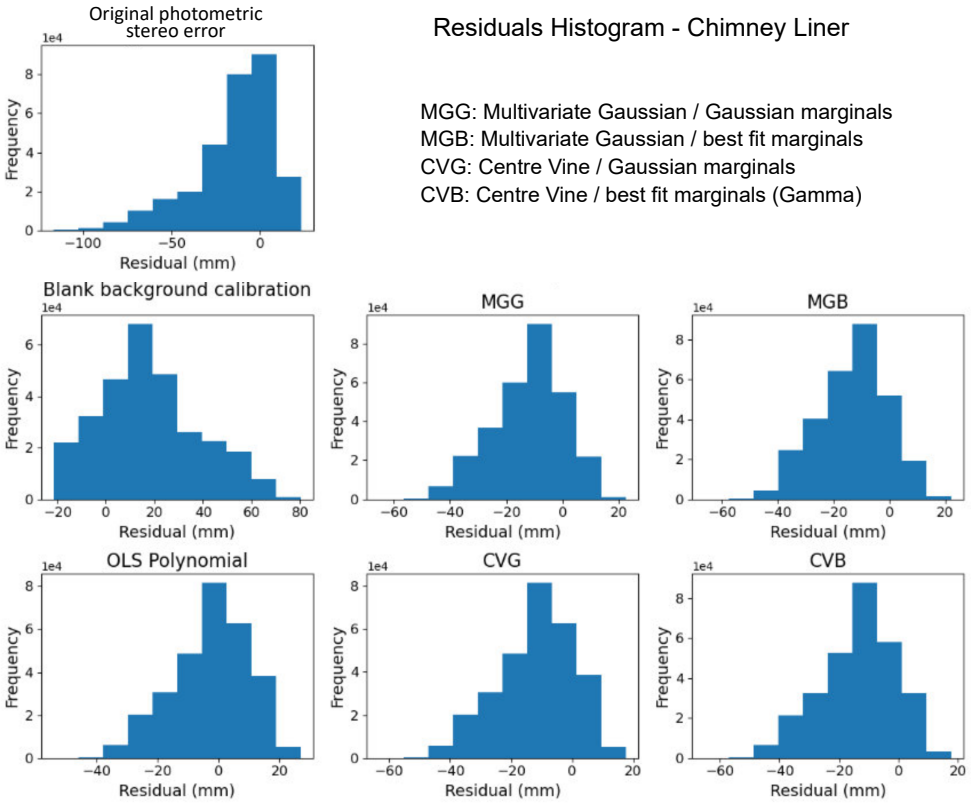

Figure 5.18: Histograms for each of the correction methods and original error for the chimney object, showing the error distribution and extreme errors across the point cloud. The polynomial model has the lowest median of -1.01, which also corresponds to the lowest MAE across the whole distribution.

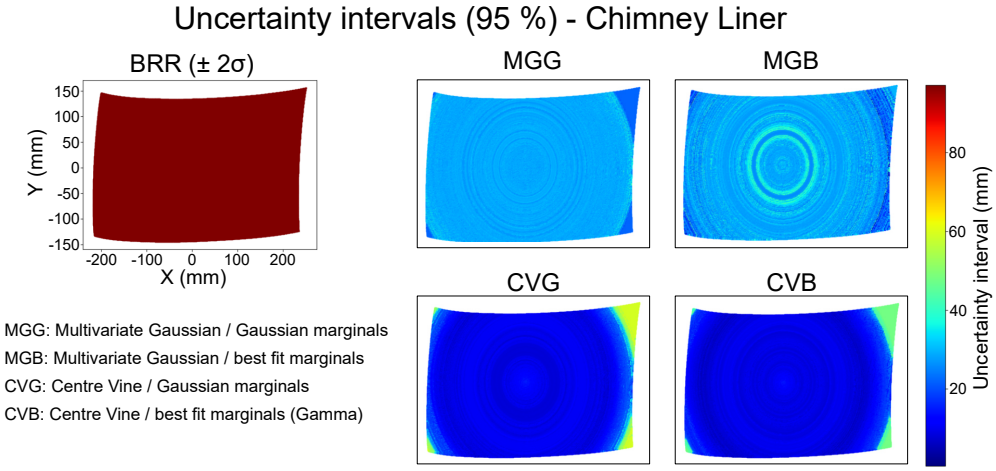

Figure 5.19: Heatmaps of the interval between the 5 and 95 % confidence bounds for the copula models and 4 ($\pm$ 2) σ for the BRR polynomial for the chimney object. All figures share the same axes. The corner areas for the CVG copula model is the highest difference between the upper and lower confidence bound of all copula models.

Case Study 3: Damaged slab

As shown in Table 5.5 for the slab object, calibrating using the photometric stereo blank background has the worst MAE of all calibration methods, which results in a higher MAE than the original photometric stereo to CMM point cloud error, degrading the MAE by 27.0 %. However, the residuals from the polynomial approach also have a higher MAE than the original error, degrading MAE performance by 7.58 %. All copula calibrated polynomial approaches improve upon the original error with the worst performing copula (MGG) *improving* the MAE by 46.54 % and the best performing copula (CVB) improving the MAE by 52.97 % over the original error. This is reflected in Figure 5.20 where the original error has many deep blue (very negative) errors which have been over corrected to deep red or orange (very positive) errors for the photometric stereo blank background and polynomial method, then moved back towards orange to green (above and below 0) by the copula models. The photometric stereo blank background calibration method residuals also results in the highest median and standard deviation, as well as the largest maximum error. This method results in the smallest minimum error (both the maximum and minimum error are positive) which suggests that all of its predictions have been shifted to result in errors above 0 which

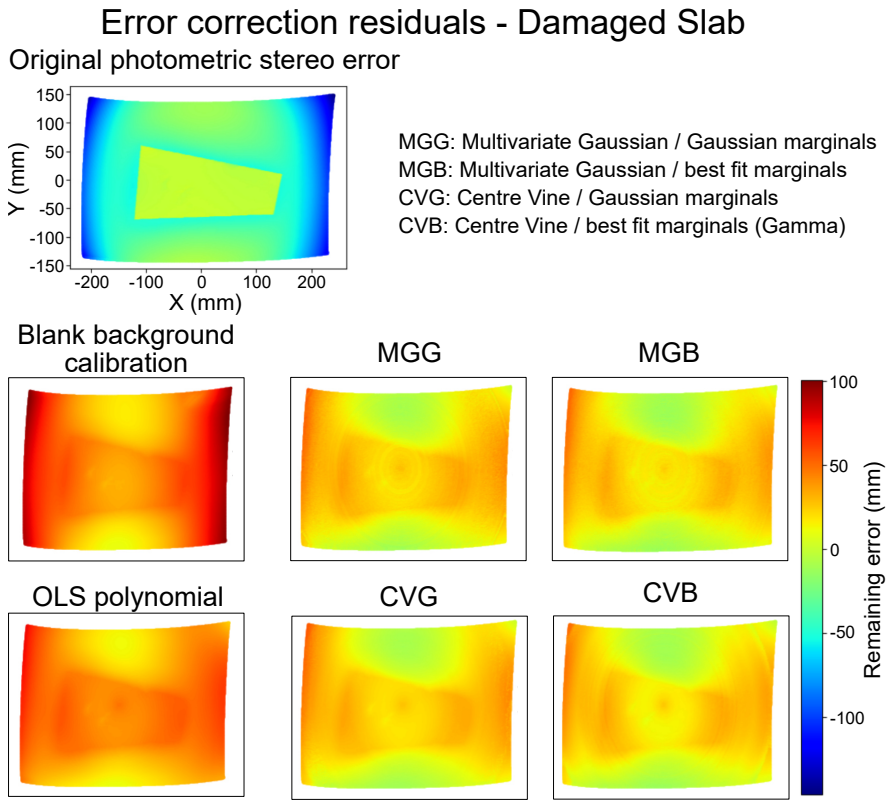

Figure 5.20: Heatmaps for the residuals of each correction method and original error for the slab object showing the spatial distribution of errors. All figures share the same axes. For the slab, the copulas result in the lowest MAE of all methods with the CVB copula having the lowest MAE of the copula models. The slab object has the highest original MAE of all case study objects, with the OLS polynomial resulting in a higher MAE than the original error.

would prevent any accurate predictions. For the slab, the best performing model across 3 metrics is the CVB copula which performs well across MAE, median and standard deviation, and holds the lowest maximum error. This is also reflected in the uncertainty quantification where the centre vine model is generally more confident in its predictions by providing much smaller 95 % uncertainty bounds than the other models in Figure 5.22.

Residual histograms for each method and the original photometric stereo to CMM error are shown in Figure 5.21. The original photometric stereo error has the largest median of all the case study objects at -37.1 which is still lower than the two worst performing methods of calibrating with the blank background (median of 45.4) or the

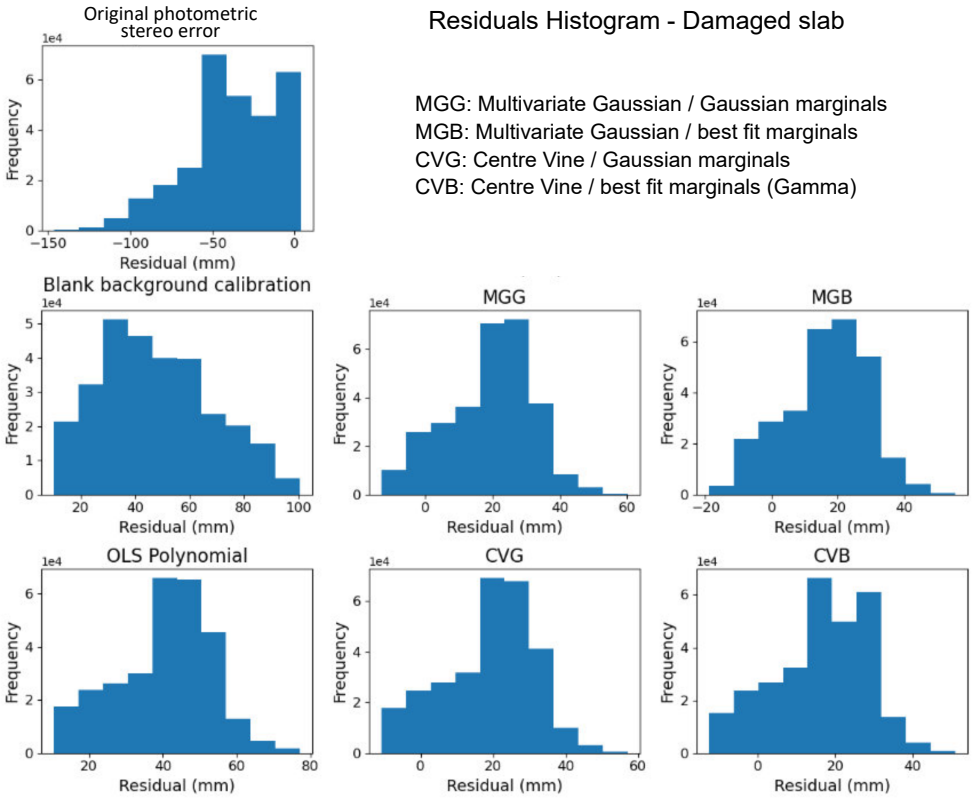

Figure 5.21: Histograms for each of the correction methods and original error for the slab object, showing the error distribution and extreme errors across the point cloud. The copula models have the lowest median of all methods, with the CVB model having the lowest at 17.15.

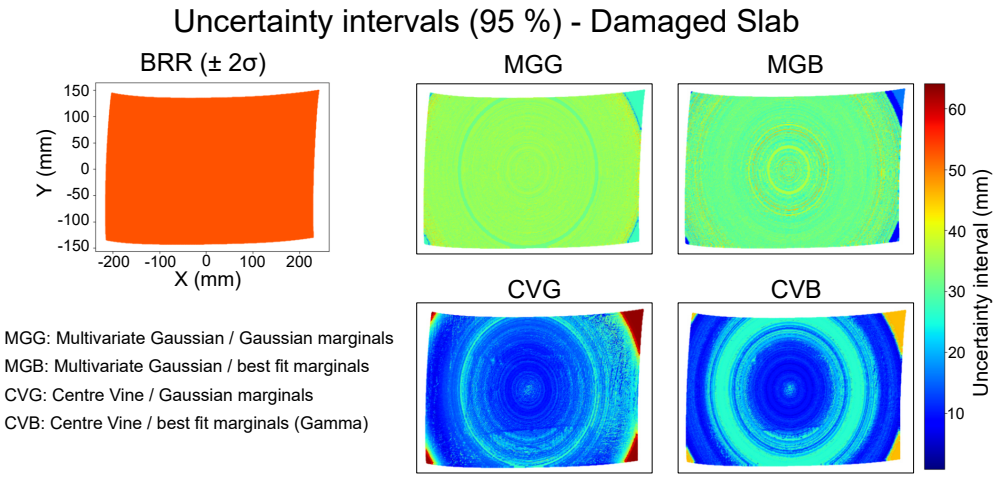

Figure 5.22: Heatmaps of the interval between the 5 and 95 % confidence bounds for the copula models and 4 ($\pm$ 2) σ for the BRR polynomial for the slab object. All figures share the same axes. The corner areas for the CVG copula model is the highest difference between the upper and lower confidence bound of all copula models, even higher than the BRR values.

polynomial (median of 42.4). All copula models have a median lower than the original error, with the lowest at 17.15 for the CVB copula. The CVB copula model has the best MAE, median and standard deviation of all models on the slab, and also limits the extreme error range of the original error from [-146.3, 3.8] to [-12.7, 51.2].

The uncertainty quantification for the different copula models and BRR polynomial are shown in Figure 5.22, where the first instance of the BRR polynomial's large standard deviation being lower than some points on the copula heatmaps is shown. The BRR standard deviation is the lowest value across all case study objects which shows an increase in model confidence, but still lacks any usable discernment across the point cloud at this colour scale. For the copula models, similar trends to previous case studies are present, with the Multivariate Gaussian models being the most confident in the corner areas where extreme errors tend to be present, while the Centre Vine models are more confident in the point cloud centre. The MGB model is overall more confident than the MGG model, as shown by the narrower uncertainty intervals in the point cloud corners and generally across the point cloud, but shows several regions where the uncertainty interval spikes around the very centre to above that of the MGG copula. This may present more fine tuned uncertainty quantification as particular regions are dis-

cerned between without needlessly increasing the uncertainty intervals across the whole point cloud. For the Centre Vine copulas, the CVB model has a lower MAE than the CVG model, but higher uncertainty intervals in some areas, such as the intermediate region between the point cloud centre and corners. The CVG model has the largest interval for all models present in the point cloud corners which is lessened by the CVB model. Again, the more flexible model may be providing more finely tuned uncertainty quantification as it attributes higher uncertainty intervals to areas the CVG model is not able to identify as more uncertain, and provides more reasonable intervals in areas where both models experience high uncertainty.

An overview figure emphasizing the process and some selected results from the damaged slab dataset are presented in Figure 5.23 to further highlight the practical benefit of this approach, and the consequences of no intervention from the hierarchical modelling approach constructed from the simple base model and copula model. As shown in the outcome section of Figure 5.23, the photometric stereo rig has very large deviations from the true measurement (given by the CMM) at the outer edges of the mesh, driving up the error in the Z axis (≥ -125 mm) shown in the heatmap. The second outcome shows the estimations of the Z error from the hierarchical modelling process, which has been used to correct the photometric stereo rig mesh for visualisation purposes. The overall reduction in error from the largest deviations can be observed in the 3D representation of the points and the heatmap of remaining Z error.

5.4.4 Computational discussion

During this work, a large hurdle was the computational time required to evaluate the copula models. This was heavily influenced by the Scipy function 'mvn.mvnun' in the Multivariate Gaussian module of the Copulas package used to calculate the cumulative distribution function of a multivariate normal. The function is based on [282], and utilises a Monte Carlo simulation with defined convergence tolerances. These tolerances were defined by the Copulas package as $1e^{-5}$, while the original paper operated at tolerances of $5e^{-3}$. This manifested in some CDF calculations taking upwards of 30 seconds due to the lack of convergence to the strict tolerances. The tolerances were

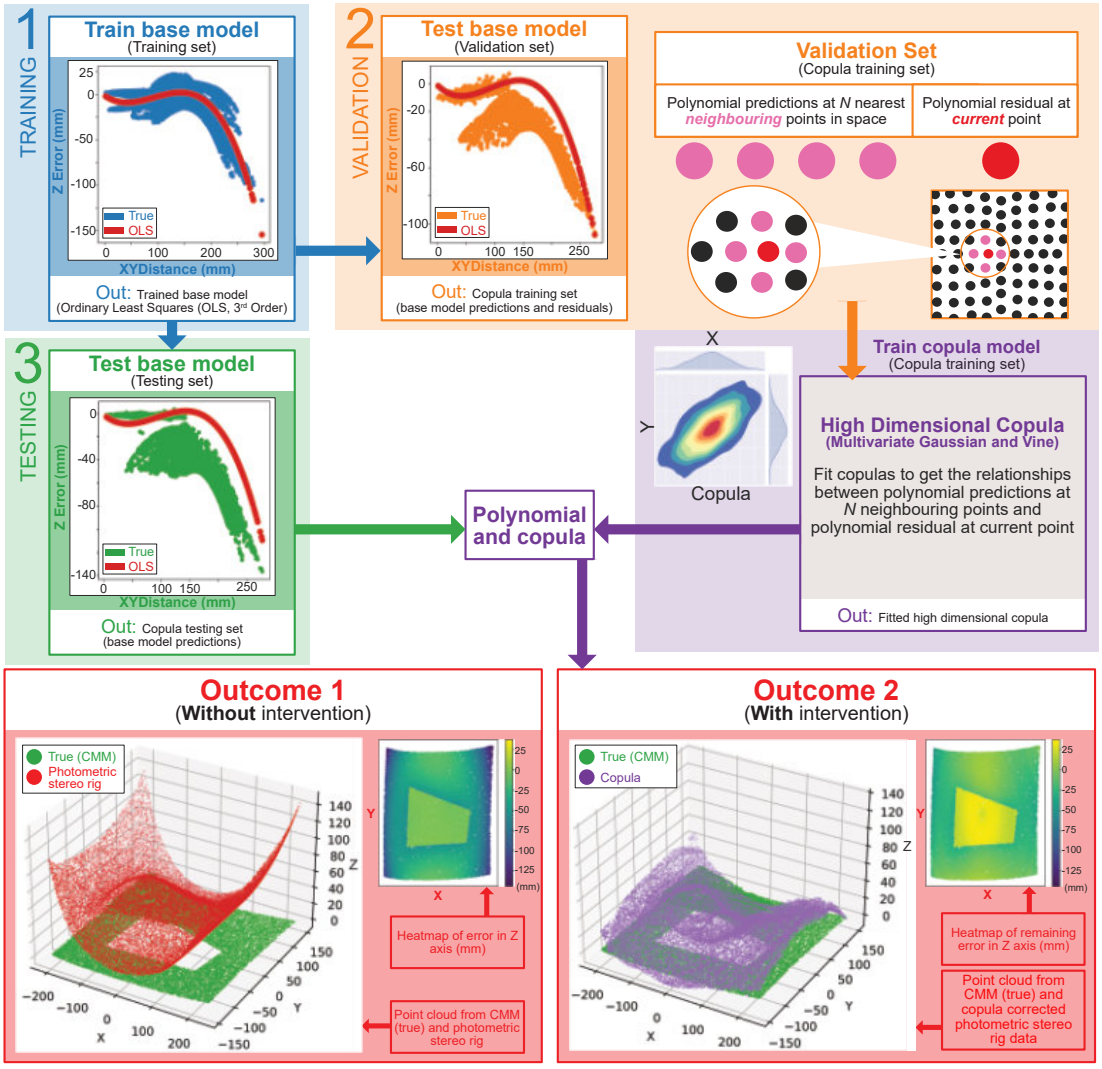

Figure 5.23: Summary of the process, and subset of example results presented in Chapter 5. The chosen datasets show the blank background and chimney as the training set, broken brick as the validation set, and the damaged slab as the testing set. The base model is a 3rd order Ordinary Least Squares (OLS) and the copula results are from the Centre Vine with Gamma marginals (best fit marginals). All figures are plotting a 1% subsample of points due to the large dataset sizes. The range of Z axis errors on the heatmap colour bar are from 25 mm to -125 mm.

changed to $1e^{-3}$ which resulted in greatly improved computational times, with the slowest calculations shifting from over 30 seconds, to approximately 0.5 seconds, with an example change in CDF results being 0.1005452 for $1e^{-5}$ tolerance, to 0.1005371 for the $1e^{-3}$ tolerance. A tolerance of $5e^{-3}$ was also tested, but did not result in significant speed increases, and so $1e^{-3}$ was chosen to preserve a slightly tighter tolerance. As the units of the data being studied is in millimetres, the consequences of setting the absolute tolerance from $1e^{-5}$ mm to $1e^{-3}$ mm was deemed acceptable.

The computational delays were additionally exacerbated by the large dataset sizes (294007 points per dataset) and the dimensions of the copula model (the number of neighbour variables used to predict the target variable). An additional set of experiments were conducted to compare the impact of using 4 or 8 neighbours to investigate any improvements in model predictions against the computational overhead required. The metrics used for comparison across the 4 types of copula model are shown in Table 5.7, with the percentage improvement of the 8 neighbour model over the 4 neighbour model shown in Table 5.6. The chimney dataset results are used here as a condensed example. Positive percentage improvements indicate that the 8 neighbour model has improved on the 4 neighbour model, while negative values indicate the opposite. The values highlighted in the table indicate the largest deviations for each metric, with the Centre Vine models seeing improvements in both MAE and standard deviation by expanding to 8 neighbours, while no models improve for the worst case (maximum) absolute error. The minimum absolute error present in all models is generally very small across the 4 and 8 neighbours as shown in Table 5.7 and so the large percentage improvement or deterioration present is perhaps not as influential as the other categories. Overall, the 4 neighbour models limit the maximum absolute error (extreme model error) and slightly improve the MAE and standard deviation for the Multivariate Gaussian models, while the Centre Vine models see some small improvement in MAE and standard deviation with the 8 neighbour models. In this application, increasing the model complexity through expanding the model dimensionality has not led to significant model improvements. However, this may not hold true for other cases where the model dimensionality may provide necessary performance increases and computational

Metric	MGG^1	MGB^2	CVG^3	CVB^4
MAE	-1.603	-1.446	2.502	**4.841**
Std.Dev	-1.708	-1.760	**2.311**	1.468
Max (abs)	-13.273	-8.857	-4.324	**-23.794**
Min (abs)	48.687	**-509.094**	-85.140	84.836

[1] Multivariate Gaussian with Gaussian marginals (MGG)
[2] Multivariate Gaussian with best fit marginals (MGB)
[3] Centre Vine with Gaussian marginals (CVG)
[4] Centre Vine with best fit marginals (Gamma) (CVB)

Table 5.6: Percentage improvement of 8 neighbour copula models over 4 neighbour models for the chimney segment dataset. Negative values show where 4 neighbour copula models outperform the 8 neighbour model. The largest deviations between model dimensions for each metric are shown in bold.

limitations will require further consideration.

The 4 neighbour (5 dimensional) model and 8 neighbour (9 dimensional) models were trained, tested and evaluated in parallel on two different workstations, and so a direct computational comparison cannot be fully conducted. The workstation processing the 4 neighbour models has a clock speed of 3.8 GHz and 32 GB RAM [9], while workstation processing the 8 neighbour model has a clock speed of 3.4 GHz and 128 GB RAM [10]. The time taken to process the copula corrections across the two workstations is shown in Table 5.7, with the Multivariate Gaussian models (MGG and MGB) taking roughly double the computation time of the Centre Vine models (CVG and CVB) on workstation 1 for the 5 dimensional models. These figures expand dramatically for the 9 dimensional models on workstation 2, with the Multivariate Gaussian models requiring approximately 10 times the processing time of the Centre Vine models, while the Centre Vine models require less than 2 hours more than their lower dimensional counterparts.

To investigate the trade off between model dimension and computation time, a study was conducted on synthetic data with similar properties to the case study data (damaged slab). A selection of 5 candidate univariate distributions were used to select

[9] AMD Ryzen 7 5800X, 8 core processor, 3.8 GHz, 32 GB RAM. The product page can be found at: https://www.amd.com/en/products/processors/desktops/ryzen/5000-series/amd-ryzen-7-5800x.html
[10] AMD Ryzen 9 5950X, 16 core processor, 3.4 GHz, 128 GB RAM. The product page can be found at: https://www.amd.com/en/products/processors/desktops/ryzen/5000-series/amd-ryzen-9-5950x.html

Model	Dimensions	MAE	Std.Dev	Max (abs)	Min (abs)	Duration (hrs)
MGG^1	5	20.144	21.060	57.992	1.524e-4	6.54*
MGB^2	5	19.463	21.516	53.858	8.621e-6	7.10*
CVG^3	5	20.005	21.073	57.174	1.019e-4	2.77*
CVB^4	5	19.196	21.202	50.642	2.055e-4	2.83*
MGG^1	9	20.466	21.419	65.689	7.820e-5	39.22**
MGB^2	9	19.745	21.894	58.627	5.251e-5	55.99**
CVG^3	9	19.504	20.586	59.646	1.886e-4	4.18**
CVB^4	9	18.267	20.891	62.692	3.116e-5	4.51**

[1] Multivariate Gaussian with Gaussian marginals (MGG)
[2] Multivariate Gaussian with best fit marginals (MGB)
[3] Centre Vine with Gaussian marginals (CVG)
[4] Centre Vine with best fit marginals (Gamma) (CVB)

Table 5.7: Comparison of 4 neighbour and 8 neighbour copula models for the chimney segment dataset. All units, except duration measured in hours, are measured in millimetres. Processing duration on workstation 1 are marked with *, while those processed on workstation 2 are marked with **.

the best fit (Beta, Gamma, Gaussian, Truncated Gaussian, Gaussian Kernel Density Estimate), with a beta distribution selected for the neighbour variable, and a gamma distribution selected for the target variable. Samples from both distributions along with their parameter values are shown in Figure 5.24.

Synthetic training and testing sets were created by sampling the beta distribution from 1 neighbour up to 9 neighbours, along with the sampled target variable, resulting in 2 to 10 dimensional models. The data sets consisted of 1000 data points each, and the copulas were sampled 20000 times to allow for the CDF and PDF calculations. The average time taken to perform one copula correction prediction across the 1000 testing samples was measured for the Multivariate Gaussian (Gaussian marginals), Multivariate Gaussian (best fit parametric marginals), Centre Vine (Gaussian marginals) and Centre Vine (Gamma marginals (best fit parametric)). The average iteration time against model dimensions is shown in Figure 5.25. For dimensions under 6, both Multivariate Gaussian models and the simplified Centre Vine model (Gaussian marginals) require less time per iteration than the most complicated Centre Vine model (best fit parametric marginals). However, the general trend of the Centre Vine models is a gently rising linear trend, while the Multivariate Gaussian model is almost exponential.

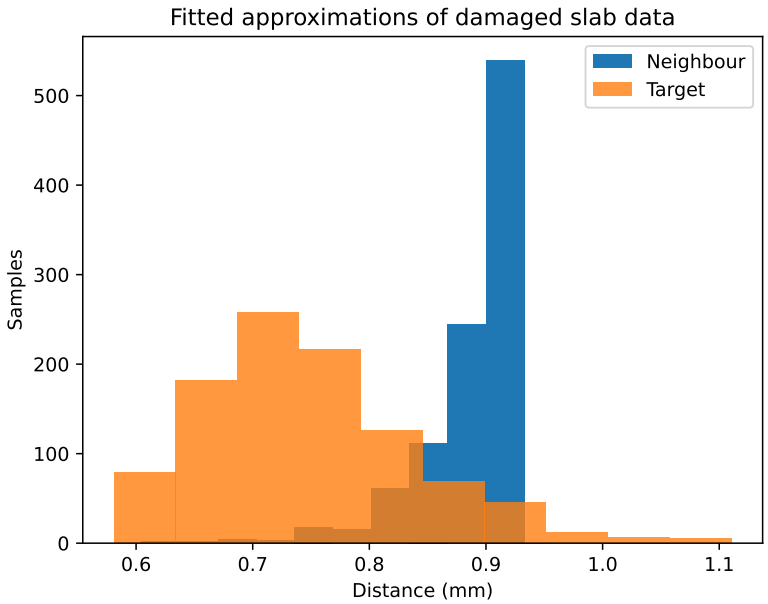

Figure 5.24: Fitted approximations of damaged slab nearest neighbour and target variable using beta and gamma distributions, respectively. The beta distribution is parameterised as: location = 0.0148, scale = 0.918, α = 15.407, and β = 0.759; while the gamma distribution is parameterised as: location = 0.547, scale = 0.0396, and α = 5.029,

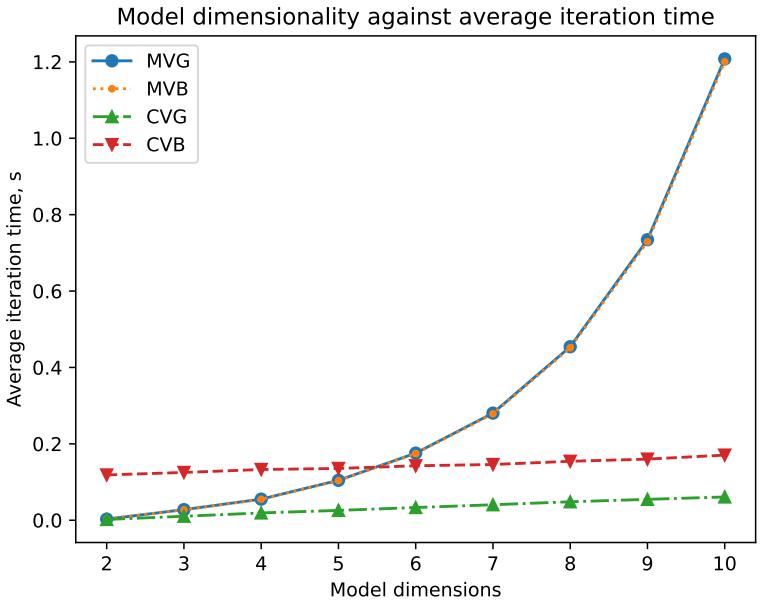

Figure 5.25: Model dimensionality against average iteration time (s) over 1000 testing points for the Multivariate Gaussian with Gaussian marginals (MGG), Multivariate Gaussian with best fit parametric marginals (MGB), Centre Vine with Gaussian marginals (CVG), and Centre Vine with best fit parametric (Gamma) marginals (CVB) on synthetic data based on the damaged slab dataset.

In all cases, a variation of the Centre Vine model may be more appropriate, however become much more viable than the Multivariate Gaussian at high dimensions (≥ 6, in this example).

5.5 Conclusion, contribution and future work

Affordable and portable rigs for non-destructive visual inspection is vital in many applications to performing cost- effective maintenance to maximise asset health and personnel safety. This may be achieved through the application of photometric stereo, a method which uses images of surfaces lit under directional lighting to create 3D meshes of the surface. However, applying such rigs under diverse environments requires some understanding of their behaviour to reduce inaccuracies across different application scenarios. To reiterate the contributions of the work described in Chapter 5:

1. Novel benchmarking dataset created for comparison of real and synthetic surfaces of relevancy to civil engineering

2. High dimensional copula approaches were able to model spatial uncertainty in reconstructed 3D surfaces, with uncertainty quantification providing the confidence in the error estimations

3. The hierarchical copula-based approach was shown to provide improved predictive performance over the base model or rig self-calibration for damaged concrete objects representative of those found in civil infrastructure

In this work, a dataset of 9 physical objects was curated through measurement by a CMM (expensive, complex, and large-scale equipment) to provide a ground truth against a photometric stereo rig (designed to be inexpensive, portable and easy to use). The photometric stereo rig was digitalised through recreation in 3D software (Blender Version 3.4) with 12 virtual objects created and processed by the photometric stereo algorithm. Of this dataset, three industrial objects (slab, brick and chimney segment) were selected for further analysis of the errors between the CMM and photometric stereo rig. Three methods were applied to account for the error between the photometric stereo rig and the CMM: using the photometric stereo rig to self-calibrate based on the error from a blank background measurement; using 3rd order polynomials on feature data generated by the rig geometry; using Multivariate Gaussian or Centre Vine copulas to calibrate the residuals from the polynomial.

In summary of the modelling results, the presence of data-based methods always outperforms using the rig to self calibrate using the blank background information or leaving the error between the photometric stereo rig and CMM without calibration. For the brick, both the polynomial and copula methods provide sufficient performance, with the hierarchical approach through the addition of the copulas outperforming the polynomial on its own. For the chimney, the polynomial outperforms the copulas, which is suspected to be due to the combination of the natural curvature of the chimney object and the rig attributing curves to surfaces resulting in 'less error' to calibrate. This is supported by the chimney having the lowest MAE between the photometric stereo

rig and CMM point clouds originally, and also the photometric stereo rig calibration using the blank background has the lowest MAE of all attempts to use this method. For the slab, the polynomial reduces the performance over providing no intervention, with the original error between the photometric stereo rig and CMM being the highest across all datasets. However, with the addition of the copulas through the hierarchical modelling structure, this approach produces much better results with a minimum MAE improvement of 46.54 %.

In terms of uncertainty quantification, the data-based methods improve upon the rig self calibration, as this method is unable to attribute uncertainty to its corrections. The BRR uncertainty quantification seems unable to discern much variation across the whole mesh and provides a large, almost continuous, uncertainty bound across the whole mesh, which is difficult to action on in practice. The copula methods are able to provide varying confidence bounds across the whole mesh with some expected, and some surprising behaviour. For example, for most objects, the centre vines are more uncertain in the corner regions and some bands around the middle of the mesh which hold the largest errors between the photometric stereo rig to CMM point clouds. These are expected to be the most uncertain as the behaviour of the mesh varies from corner to corner. However, for the Multivariate Gaussian models, more uncertainty tends to be attributed to the centre of the point cloud than the extremes. This may be due to the influence of the object being captured which would change the behaviour at the centre across all datasets.

Within the copula models, the Centre Vine model was the overall highest performing model type across the 3 case studies, outperforming the Multivariate Gaussian copula models. Vine models have additional flexibility to account for dependency in the extremes, which for applications where this is an important feature, can lead to higher performance. By outperforming the other methods, and generally outperforming the other copula methods, non-linear dependency was shown to be an important consideration within the case studies presented in this spatial application. In surface scanning scenarios where the surface texture is rough and surface damage is expected to be present, the surface geometry can present sharp and uneven changes. Add to this

unquantified error and measurement noise from the rig, linear and non-linear relationships within the data are very likely present. Multivariate Gaussian models capture linear dependency between inputs well but struggle capturing non-linear relationships. Demonstrated by their ranking as the second highest performing method, they were generally capable of accounting for a significant part of the dependency behaviour driving the rig error. Overall, the application of data-based models has accounted for a large portion of error between the rig and CMM, additionally providing uncertainty information to aid decision-making on outputs of the rig, supporting the application of this methodology to similar situations. In an application setting, the hierarchical copula based approach provides operators with an improved and affordable structural health monitoring rig with attached risk, to improve decision making flexibility and risk management in the monitoring and maintenance of large infrastructure.

5.5.1 Future work

Much of the usability concerns with the copula models could be improved upon in future work. For example, to alleviate the large dataset sizes from the photometric stereo rig, an investigation into accuracy against the number of points calibrated in a certain region would alleviate computational expense if there is found to be a regional scale where the calibration is equally appropriate. This would mainly rely on the rate of change of the polynomial predictions, as areas with similar Euclidean distances from the origin will have different predictions made across the whole mesh depending on the polynomial slope in that range. Additional computational benefits could be gained from improving the CDF calculation process, which may involve, for example, exchanging the CDF calculation for a look-up table.

In future, more could be done with the curated dataset, some of which has been previously mentioned in Section 5.3. Namely, more work could be done on rendering fidelity and making more use of the advantages of the virtual environment, such as investigating the impact of rig design on mesh accuracy, or automating crack progression, rendering and analysis. In this work, the photometric stereo algorithm used with the rig was proprietary, but there is opportunity to swap this for more cutting edge

algorithms which could be further tested across different scenarios and material types. A comparison of other open source photometric stereo algorithms could see improvement over the provided version which may alleviate inaccuracies within the system. While photometric stereo algorithm development and comparison is not the focus of this work, it may provide additional insights and challenges for the subsequent modelling and analysis work, such as investigating an end to end propagation of uncertainty should the photometric stereo algorithm be uncertainty aware. This ties in to two final points of simply analysing more data, as this work only focused on a small subset, and applying further explainability techniques to explain *where* the rig is going wrong: on what surface features, in what scenarios and under what conditions? The combination of already collected physical data and the potential of the virtual environment could allow for a system study which may pinpoint the major sources of uncertainty from the rig or its environment to allow for more targeted compensation methods to be developed, or for issues to be designed out in future. This would further align the work with Chapter 3.

Chapter 6

Conclusion

There is a drive in industry towards improved predictive maintenance and more efficient plant management. Uncertainty sources impact the ability to accurately measure the state of an asset, and in turn impact an analytics ability to derive the state of health of said asset. Without the ability to then attach risk to decisions from analytics, it becomes difficult to justify actioning their recommendations in industrial applications, especially in environments with strict traceability requirements as in the nuclear sector. As such, uncertainty quantification and explainability for machine learning applications are both highly relevant and active areas of study.

Access to trustworthy data analytics tools capable of attributing risk to their outputs alongside tools able to interrogate the impact of the design of the data pipeline serving these analytics, provides power plant operators with solutions to manage uncertainty within their data-driven asset management processes. To address this need, this thesis presented a framework for explainable pipeline design and an explainable hierarchical copula-based modelling approach which were developed to handle uncertainty sources throughout the full data analytic system.

The contributions of this thesis have targeted issues across different parts of the maintenance process, shown in Figure 6.1. Chapter 3 approached the data pipeline and provided a novel and transferrable pipeline design framework. Chapter 4 and Chapter 5 then delved into the analytics stage and showcased the application of the hierarchical copula-based modelling approach for both timeseries and spatial data. Over

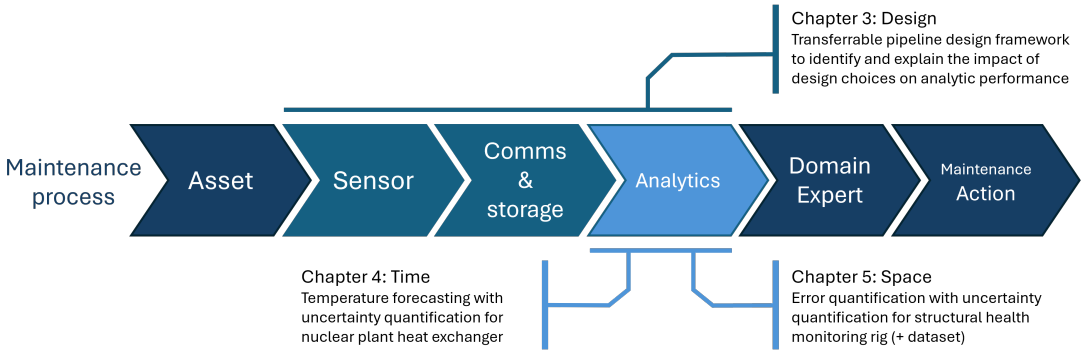

Figure 6.1: Thesis contributions across different parts of a summarised maintenance process flow diagram.

both chapters, the approach was shown to provide improved predictive performance, alongside the ability to add prediction intervals (derived from the copula confidence bounds) to these predictions.

Throughout this thesis, it has been shown that the framework for explainable pipeline design and the developed explainable hierarchical copula-based modelling approach could be applied across different engineering disciplines and were successful for different types of common industrial assets. The applications utilised throughout the case studies included, but are not limited to, applications within electrical, mechanical, and civil engineering. Additionally, the techniques were proven across different machine learning tasks which included the classification of faults, temperature forecasting, and remaining useful life prediction. Lastly, the hierarchical copula-based modelling approach was shown to be generic enough to be applicable to spatial and temporal data, covering an even wider set of potential applications. The following sections summarise the most important conclusions from the technical findings of this thesis.

6.1 Chapter 3 Highlights

The highlighted contributions of Chapter 3, are as follows:

1. Demonstrating that pipeline design impacts analytic performance within fault diagnostic systems

2. Identifying and explaining highly performing design options using a human-readable

XAI framework, leading to systems with better predictive performance

3. Improved fleetwide monitoring by leveraging insights from previous designs to new systems

The first contribution demonstrated how uncertainty sources in the analytics pipeline design could manifest in different ways, with the main result being they could either improve or deteriorate fault prognostic or diagnostic system performance. The detrimental impact of this was shown to be tempered through the application of different modelling strategies, such as using hybrid models which were found to be more robust to these compounding uncertainties.

In a second contribution, the presented explainability framework was used to identify and rank highly-performing or poorly-performing design options, which overall lead to the design of better performing fault detection and diagnostic systems. It was shown that following the recommendations of the framework resulted in a mean predictive error of $< 0.1\%$, while using the detrimental design options flagged by the framework resulted in mean predictive error of $> 50\%$. Importantly, the explanations provided by the framework allow operators to compare between designs and understand what impact their design options might have on the fault detection or diagnostic system. This is especially important in situations where a design option must be used in the finalized system design, where this method provides evaluation and understanding of the risk present in the system.

Finally, insights from the design of one fully-observed system pipeline were leveraged to improve the design of other system pipelines, all without requiring the same level of intensive observation. This has clear benefits where an operator may have a large number of similar assets requiring monitoring, but where it would be infeasible to test each individual asset to the same level.

The framework was shown to be adaptable to different maintenance scenarios (remaining useful life prediction and fault classifications) for motor bearings, and configurable to accommodate different amounts and types of pipeline stages. The flexibility of the approach means it can be applied across different asset types, and the capability of the framework to transfer insights between the same types of asset can save an

operator time and cost by avoiding extensive data collection for all assets. A major limitation of the approach is its reliance on the amount and diversity of fault samples (which is similar to other machine learning approaches), and its requirement for assets to behave similarly for insights to be transferred reliably.

6.2 Chapter 4 Highlights

The summarised contributions of Chapter 4 are:

1. Providing prediction intervals derived from confidence bounds on temperature predictions for critical assets in presence of measurement noise and modelling errors attributes useful risk to model outputs

2. The developed hierarchical copula-based method captures complex dependency structures between time propagated measurement and modelling uncertainties

3. The approach improved performance on real operational data (nuclear heat exchanger and wind turbine generator bearings)

Chapter 4 demonstrated a hierarchical copula-based modelling approach applied to temperature forecasting for three case studies: one synthetic, and two industrial datasets. The hierarchical copula-based modelling approach was shown to improve the predictions of a transparent base model that was previously adopted by the PhD industrial partner. The copula models were applied over short-term intervals to correct the long-term forecasts provided by the base model, with the additional advantage that this novel method could also attribute risk on the short-term horizons. This chapter also involved a comparison of several different high dimensional copula models where the model type and the complexity of their marginal assumptions were varied to investigate the effect this had on the predictions.

Most significantly, the results of this chapter demonstrated that the approach could provide useful prediction intervals on the forecasts even in the presence of measurement noise and modelling error. The case studies covered very different temperature behaviours, including aging over long operational periods with large jumps due to

maintenance and other interventions, or large temperature swings from environmental interference. The models were found to be robust to these effects provided that the data used to train the models was reflective of this behaviour, proving the models flexibility in operational scenarios where temperature changes may be driven by different physical processes. The novel approach was shown to provide superior predictive capability compared to the base model alone, for both synthetic and real operational data, with the maximum mean error improvement being 82.17% on the nuclear plant data. The Centre Vine model was the highest performing model type, followed by the Multivariate Gaussian, for the percentage improvement metric which captures the general predictive performance gain over the base model. However, the Multivariate Gaussian models had the smallest and/or most accurate prediction intervals, followed by the Centre Vine models when evaluated by the interval score metric, which reflects the accuracy of the models attribution of risk to the temperature forecast. The attributes which are most desired by the operator would decide on which model type may be most appropriate. Having access to a transparent, explainable modelling strategy that can also attribute risk to short-term forecasting horizons allows an operator to determine courses of action, be this operational changes to mitigate undesired temperature increases or provide evidence to action maintenance intervention, with sufficient confidence and information. This becomes especially important if assets are operating close to a temperature threshold, where the choice of intervention may avoid financial losses from unexpected outages due to faults.

6.3 Chapter 5 Highlights

The outcomes of Chapter 5 are summarised as:

1. Novel benchmarking dataset created for comparison of real and synthetic surfaces of relevancy to civil engineering

2. High dimensional copula approaches (as in Chapter 4) can be used to model spatial uncertainty in reconstructed 3D surfaces, with uncertainty quantification providing the confidence in the error estimations

3. The hierarchical copula-based approach was shown to provide improved predictive performance over the base model or rig self-calibration for damaged concrete objects representative of those found in civil infrastructure

Chapter 5 presented a study on the verification of a structural health monitoring test rig. The type of damage captured by the test rig and the location of the damage on the concrete infrastructure can be analysed by engineers to prioritise and plan any required maintenance actions. However, inaccuracies in the captured geometry may cause incorrect maintenance decisions to be made. To validate the rig, a custom data set was collected and curated that included physical objects with a range of different attributes and geometries, including examples of worn building objects. The objects were measured in the test rig and compared against the results from a more characterised method. Additionally, virtual 3D models were tested in a virtual recreation of the test rig to represent the case where environmental and measurement noise could be completely removed.

The hierarchical copula-based approach and a transparent base model were applied to the test rig outputs and shown to be capable of correcting the base model predictions while also providing an uncertainty estimate across the surface captured by the rig. The Centre Vine model was the overall highest performing model type across the 3 case studies, outperforming the Multivariate Gaussian copula models, the base model on its own, and the rig self calibrating against a flat plane. The test rig offers a cheaper, portable option for capturing damage in concrete infrastructure, but requires its outputs to be validated to allow the generated meshes to be utilised in maintenance planning. The hierarchical copula-based approach was able to provide this validation, with the highest reduction surface estimate error being 47%, whilst also providing uncertainty estimates across the captured surface to identify areas where the model correction is high or low risk. This provides engineers with more accurate surface geometry estimates and clearer associated risk which may be crucial for structurally significant areas of the damaged infrastructure.

6.4 Future work

While the initial goals of this research presented in Chapter 1 have been either achieved or progressed, there is plenty of scope for future development. In future, the three parts of this work could be more definitively combined, providing analytics capable of uncertainty quantification on spatiotemporal data within an explainable data pipeline, perhaps with further expansion to allow uncertainty sources to be traceable through the entire system. This investigation could be supported through the design and construction of a test rig which is designed to have a flexible and variable data pipeline to observe different faults in a more realistic operating environment, rather than synthesizing them. This would diversify and expand on the use of the structural health monitoring rig utilised in this work and provide the possibility of releasing more open-source data to the research community, with a particular focus on diagnosing data quality issues from the design of the data acquisition system. This could potentially align well with current research on digital twins[1], which are models of real systems used to assist in fault detection and diagnosis. Having access to both the real system and a model of the system provides opportunities to explore both better digital twin design, and also how high fidelity digital twins can be used to improve real systems, either through the early detection of faults or providing operational suggestions to prevent certain faults re-occurring quickly. Although several machine learning tasks and engineering applications were covered in this work, more work could be done within each application to investigate what type of observations (be it of poor data quality, or at a point in the data space with high uncertainty) are most difficult for the analytics, and so put the outputs at most risk. This would fall into the area of anomaly detection or adversarial machine learning, where data points that deviate significantly, or create specific failures in the analytics are studied to provide more robust analytics which are more well defended against these situations.

[1] "Harnessing the power of digital twins", The Alan Turing Institute, https://www.turing.ac.uk/research/harnessing-power-digital-twins

Appendix A

Bearing fault diagnosis across similar assets - Considering the impact of domain shift on pretrained models

A machine learning solution that works well for one asset may not necessarily generalize to a similar plant asset, and separate solutions for every rotating plant being monitored may be infeasible. Different operating or environmental conditions, manufacturing tolerances, maintenance schedules or fault severity cause shifts in the data collected for similar assets with the same developing faults which may cause pretrained analytics to fail. This work is an additional study alongside the pipeline design work in Chapter 3 which has been expanded to consider how pretrained models may transfer successfully or unsuccessfully between similar assets.

To consider how domain shift may impact pretrained classifiers, the Case Western Reserve[1] (CW) and the MFPT[2] bearing fault datasets were chosen to represent data streams from two similar rotating plant. Both datasets contain inner race and outer race faults, however both contain different operating conditions and the MFPT sensor

[1] https://engineering.case.edu/bearingdatacenter
[2] https://www.mfpt.org/fault-data-sets/

locations and motor end containing the faults are not disclosed.

For this analysis, the classifiers were trained and tested on the CW dataset only, then the pretrained classifiers were tested on MFPT data, and vice versa. The classifiers were also trained on a mixed dataset and tested on separated data from each dataset to understand how the classifier performs on each asset. All data combinations involved 5-fold cross validation and the settings chosen were a window length of 0.5 seconds, training and testing set allocation by random (75:25 split), all data domains (timeseries statistics, frequency and wavelet), the classification tasks were fault detection (binary) and cross section location diagnostic with the process as described in Chapter 3 Section 3.3.3. For the CW dataset, only fan end sensor data for fan end faults and normal operation data were included. The models chosen were the Fine Tree, Fine, Medium, Cosine, Cubic, Weighted and Ensemble Subspace k-Nearest Neighbours and the Ensemble Bagged Tree, chosen due to diverse classification boundaries (lines, polynomials, elliptic regions, etc) and methods.

The histograms and kernel density estimates of the testing errors for the 3 cases are shown in Figure A.1, A.2 and A.3. As shown the classifiers perform well on the testing data from the training data domain but perform poorly on data from the other domain, demonstrating that similar assets with different operating conditions and data collection methods can impact the classifiers' ability to identify the same bearing faults. The presence of the peaks at 50 % testing error in the histograms of the non-training data domain suggest that the classifiers have resorted to random predictions, while the prevalence of testing errors above this (60-80 % in Figure A.2, left) suggests the classifier has learnt incorrectly and are more confident in allocating these incorrect predictions. However, when the classifier has access to data from both domains (Figure A.3), the domain shift can be learned allowing the classifiers to generalize better, producing results comparable to those generated by the models trained solely on each dataset with a slight performance deterioration.

Fleet wide monitoring using a smaller set of analytics reduces the amount of maintenance required to upkeep these tools while providing the ability to increase the amount of assets that can be assessed. As innate differences between similar assets cannot be

Appendix A. Bearing fault diagnosis across similar assets - Considering the impact of domain shift on pretrained models

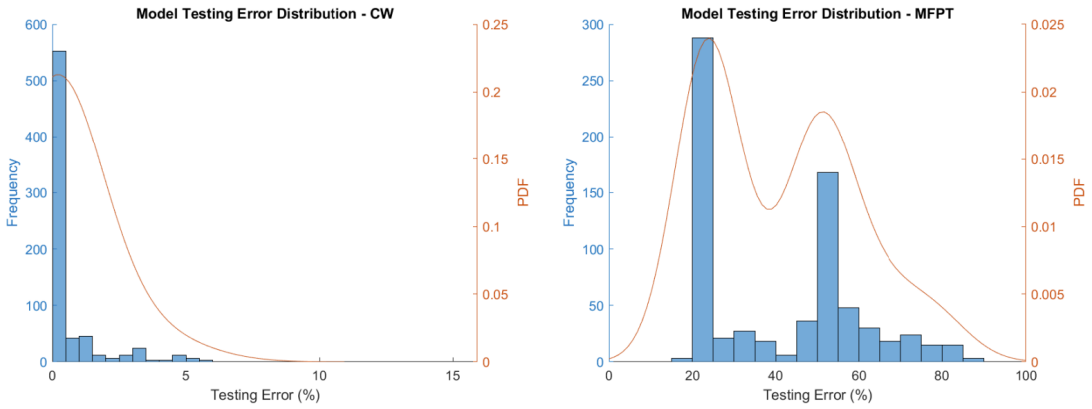

Figure A.1: CW (left) and MFPT (right) testing error (%) histograms for models trained on CW data only

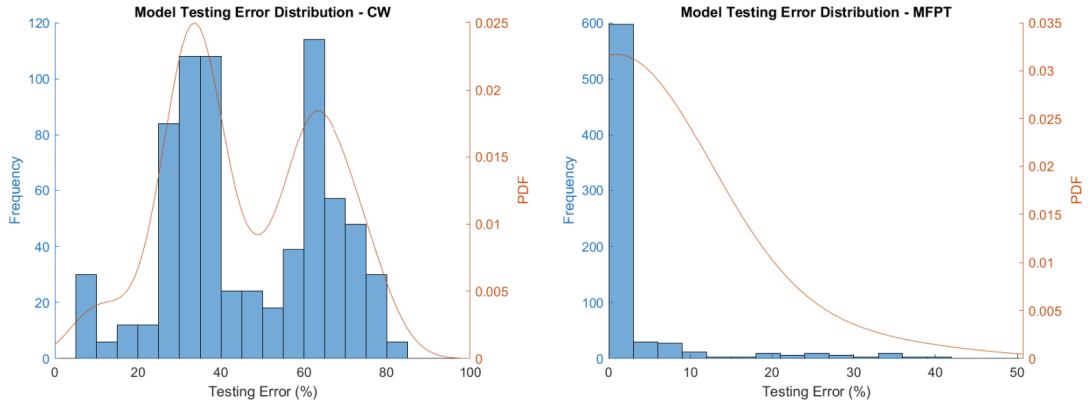

Figure A.2: CW (left) and MFPT (right) testing error (%) histograms for models trained on MFPT data only

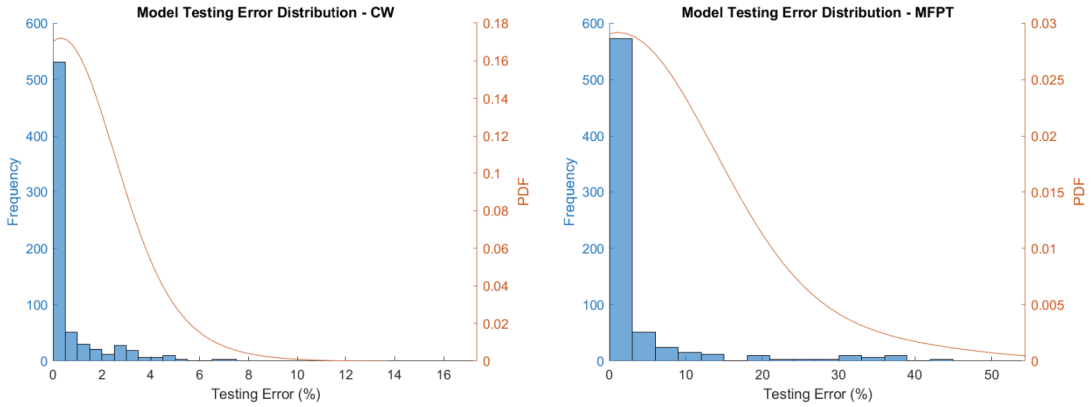

Figure A.3: CW (left) and MFPT (right) testing error (%) histograms for models trained on a mixture of CW and MFPT data.

easily measured, it is important to understand the complication they pose to the analytic tools and how these problems can be mitigated. In this case study, it was enough to provide the classifiers with labelled examples from both domains to generalise the model behaviour and accommodate both assets. However, as more assets are monitored with the same tools, the similarity between the assets in the same family will not be consistent, requiring greater understanding of how domain shift can be successfully quantified and how it impacts the analytics directly to be compensated for. Many methods exist to align the data, align the features or generalise the model to perform transfer learning, which require understanding of the nature of the domain shift and application to advise on the most appropriate method. These options may be explored further in future work.

Appendix B

Pipeline explanations - Example SHAP plots and investigating worst case design choices for additional context

The intermediate stage between encoding the pipeline designs into an additional data source and utilising the ranking of design choice impact is provided by SHAP plots. These plots show the magnitude and direction of the impact the input variables have on the output variable. These SHAP plots are not of direct focus to Chapter 3 and so are instead provided here for additional context. In Chapter 3, the focus was placed on demonstrating the identification and selection of design choices which lead to the highest performance of analytics within the pipeline design, however, the process equally allows the identification of negative performance drivers that should be avoided. This may be of use as further justification for design selections, or may allow inefficiencies within current system design to be identified for further improvement. All data presented in this section can be found, summarized across all runs, in tables within Chapter 3 but are visualized here for additional clarity.

Appendix B. Pipeline explanations - Example SHAP plots and investigating worst case design choices for additional context

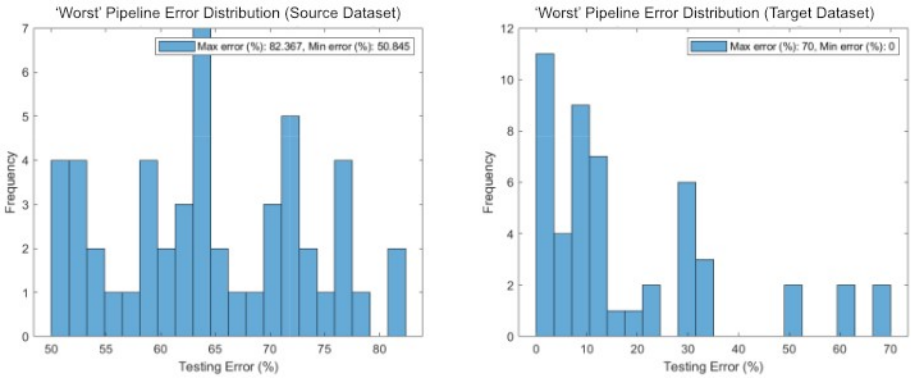

Figure B.1: (Left) Histogram of classification errors for the 'worst' design choices for the source dataset (Case Western reserve dataset). (right) Histogram of the classification errors for the 'worst' design choices identified for the target dataset (MFPT dataset). The 'worst' choices across each stage for each dataset were summarised in Table 3 of the paper.

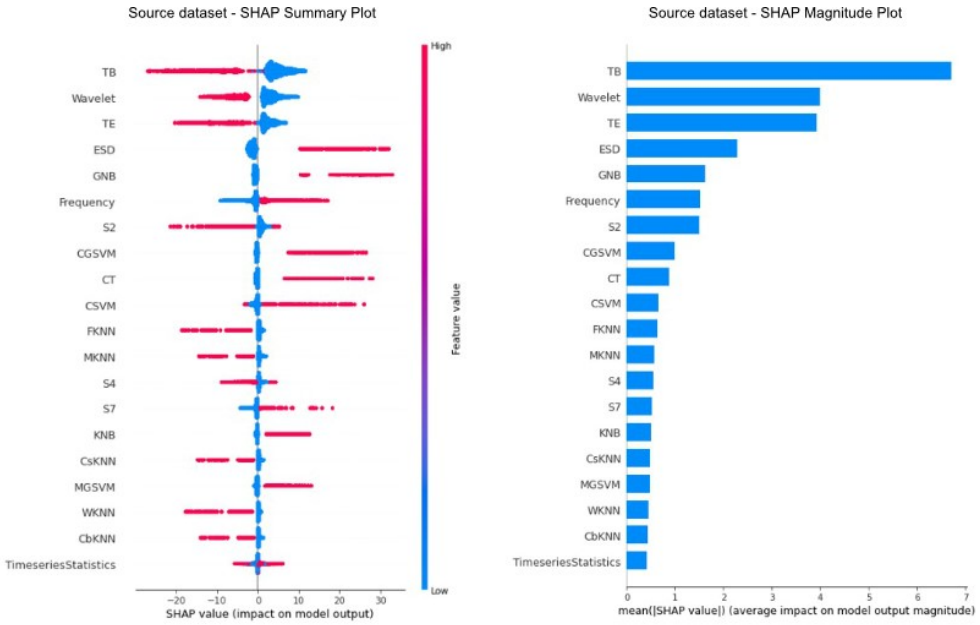

Figure B.2: (Left) Summary plot showing the pipeline choice ranking and direction of impact on the pipeline classification error. When values of the choice is high, the choice is present in the design, while a low value denotes the absence of the choice in a pipeline. Some values, like the timeseries statistics choice, have generally low impact regardless of its presence or absence in the design, while others such as the 'ESD' (Ensemble subspace discriminant) model have little impact when absent, but a large negative impact when present in the design. (Right) Magnitude plot showing the pipeline choice ranking and magnitude of impact on the pipeline classification error. The top 5 most impactful choices in descending order are 'TB' (Binary classification task), 'Wavelet' (data processing domain), 'TE' (motor end classification task), 'ESD' (ensemble subspace discriminant model) and 'GNB' (Gaussian Naive Bayes model). As shown, much of the impact has reduced by the 'S2' (Fan end sensor, only) choice, showing much of the impact is concentrated in a relatively small fraction of the decisions.

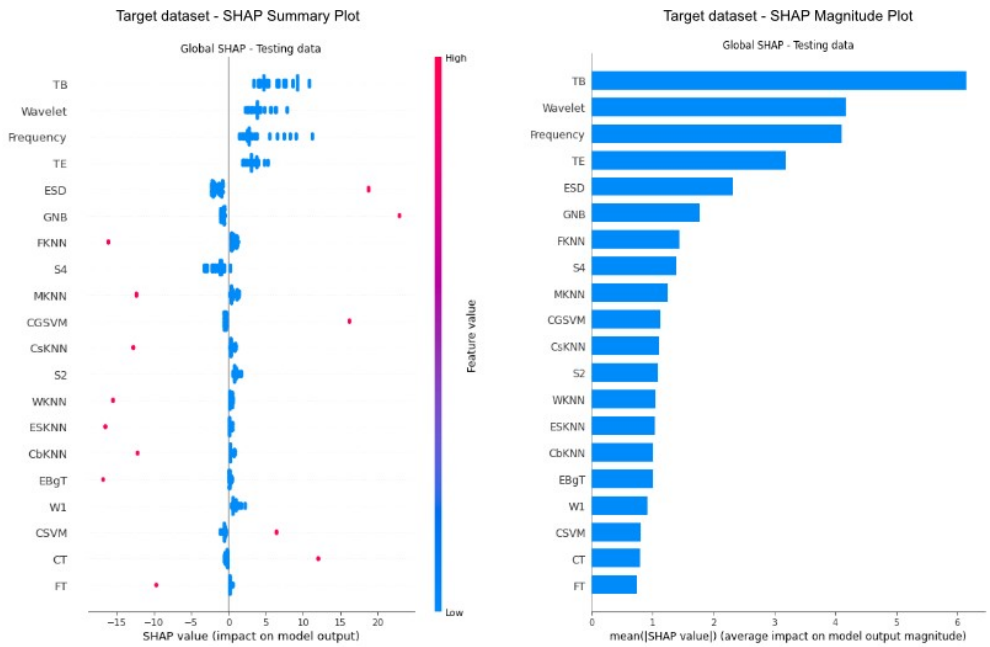

Figure B.3: (Left) Summary plot showing the pipeline choice ranking and direction of impact on the pipeline classification error for the target design subset. For most variables, their absence (low value) are concentrated nearer a SHAP value of 0, while their inclusion (high value) tends to generate a more extreme SHAP value response, showing the model is more sensitive to a design inclusion than absence in this case. (Right) Magnitude plot showing the pipeline choice ranking and magnitude of impact on the pipeline classification error for the target design subset. The top 5 most impactful choices have 4 common choices with source case, with the addition being the frequency data processing domain taking third place. Compared to the SHAP plots for the source dataset, the impact of all variables is higher, however the most impact is still concentrated within the top few variables ('TB'-'ESD').

Appendix B. Pipeline explanations - Example SHAP plots and investigating worst case design choices for additional context

Appendix C

Multivariate Gaussian and Vine Copulas: Linear and Non-linear data structures

High dimensional copula models may be applied to linear and non-linear dependency structures. Vine copulas are generally more flexible and able to handle non-linear dependency, whereas Multivariate Gaussian specialise in linear dependency. Both models have their merits and should be evaluated on a variety of metrics to assess all facets of the models performance for a given application. The application should inform on what type of performance is most valued and the data will inform on potential challenges for different model types. For example, some cases may place high importance on rare events which occur at the extremes, however, both Multivariate Gaussian or Vine models may be sufficient for this purpose depending on the data structures involved. In this appendix, a short example is given for linear and non-linear data to show how Multivariate Gaussian and Centre Vine models perform on these simplified scenarios.

In the linear example, data is generated from Gaussian models which presents as concentric rings on the training data contour plot in Figure C.1. Both the Multivariate Gaussian and Centre Vine model are able to capture this behaviour, as shown by the good overlap of samples on both plots, and similar structure in the contour plots.

In the non-linear example, training data is created by sampling a Frank copula. This

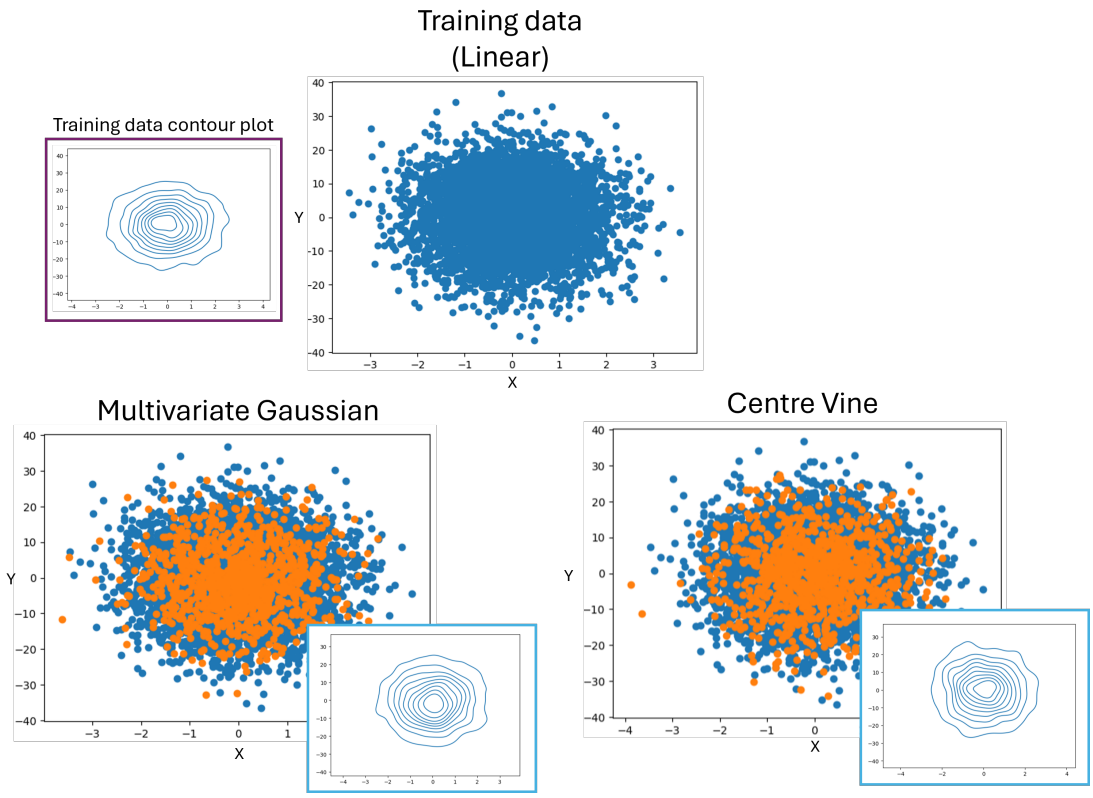

Figure C.1: Multivariate Gaussian and Centre Vine models performance on Gaussian (linear) training data.

type of model has low upper and lower tail dependency, which means there is higher deviation in the extremes due to this reduced dependency. In this case, the relationship between the variables at the extremes is more volatile. In an engineering application, this may be important for assigning risk at a highly uncertain point in the relationship between the variables. Extreme values may be rare, but may be of great importance in an operational sense, where intervention may be required to prevent operational limits being exceeded. Couple this with the uncertainty in the relationship between process variable X and Y at the extremes, and this could be an important scenario to be modeled. The performance of a vine and multivariate gaussian copula on non-linear training data is shown in Figure C.2. The copulas are sampled to provide data for plotting, and the training data used to fit the copula models is shown. Contour plots are provided to show the density of the copulas and training data. The vine copula shows good performance at fitting the data (similar shape of the contour plots between the training data and the vine samples, and good overlap of samples with the training data). The Multivariate Gaussian has overcompensated to encapsulate the training data, resulting in a wider spread of samples. The density on the contour plot is also not as tight, which is not similar to the training data contour plot. However, the Multivariate Gaussian has been able to identify that there are two density structures in the upper and lower tails. The Multivariate Gaussian has captured the general trend of the data, but not the details which were successfully captured by the vine copula.

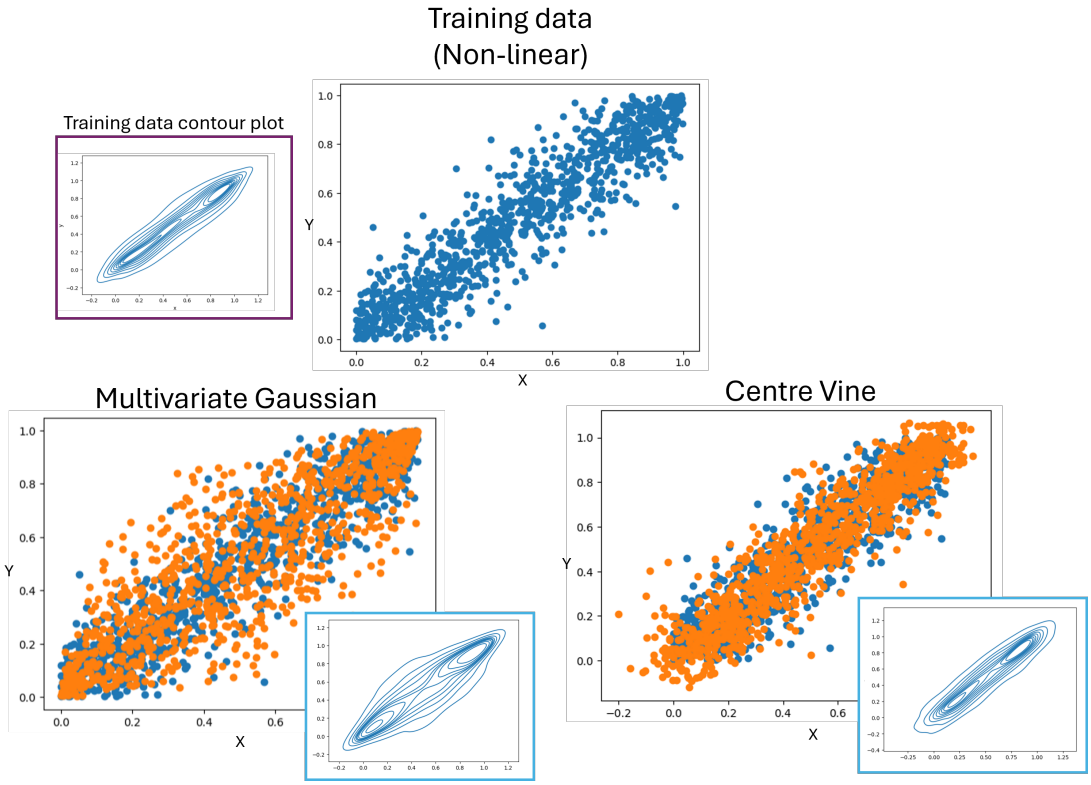

Figure C.2: Multivariate Gaussian and Centre Vine models performance on non-linear training data (low upper and lower tail dependence).

Appendix D

Autocorrelation plots for copula timeseries analysis

Autocorrelation functions [283] measure the correlation (linear relationship) between a timeseries function and itself at different lagged timesteps [284]. This is used to separate out time dependent trends in the data which may be seasonal or periodic. High correlation, whether positive or negative, indicates a strong linear trend. Correlation values within the confidence bounds are considered insignificant and present no linear relationship. While copulas can accommodate dependency structures that are linear and non-linear, autocorrelation function plots are a common method for visualising linear trends in data. In this appendix, the autocorrelation plots for the base model residuals are presented to show the trends present in the data used to train the copula models. The copula models are trained on: 5 lags for the synthetic dataset; 8 lags for the open source dataset (representing a 24 hour forecast horizon); and, 15 lags for the industrial dataset (representing a 2 week forecast horizon). To summarise, the autocorrelation strength across the lags used to train the copula models are sufficiently strong to be considered significant. Interestingly, the open source dataset has a sinusoidal autocorrelation pattern which, across the 8 lags, drops below significance level and then increases again.

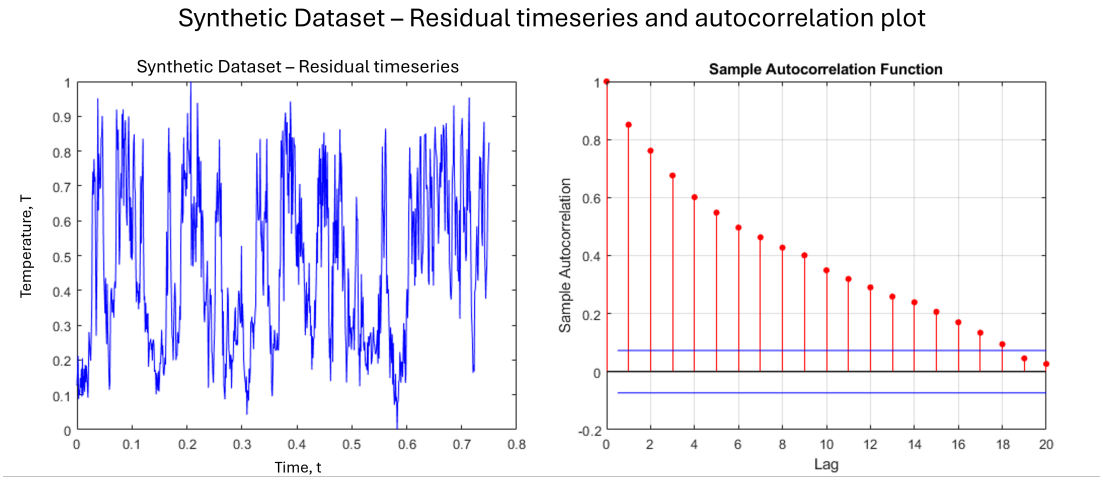

Figure D.1: Base model residual timeseries and autocorrelation plot for the Synthetic dataset. The autocorrelation values are significant until lag 19 where it drops below the confidence bounds. The lags used to train the copula model are up to 5 lags, which present a strong linear trend.

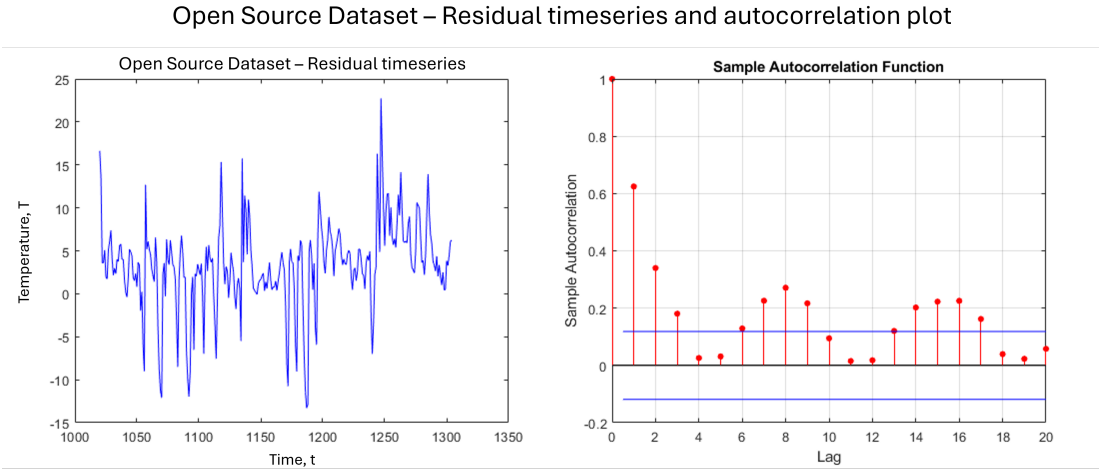

Figure D.2: Base model residual timeseries and autocorrelation plot for the Open Source dataset. The autocorrelation values are significant while above the confidence bound lines. The lags used to train the copula model are up to 8 lags, which means two of the lags used have no linear relationship, while the others present a positive linear relationship.

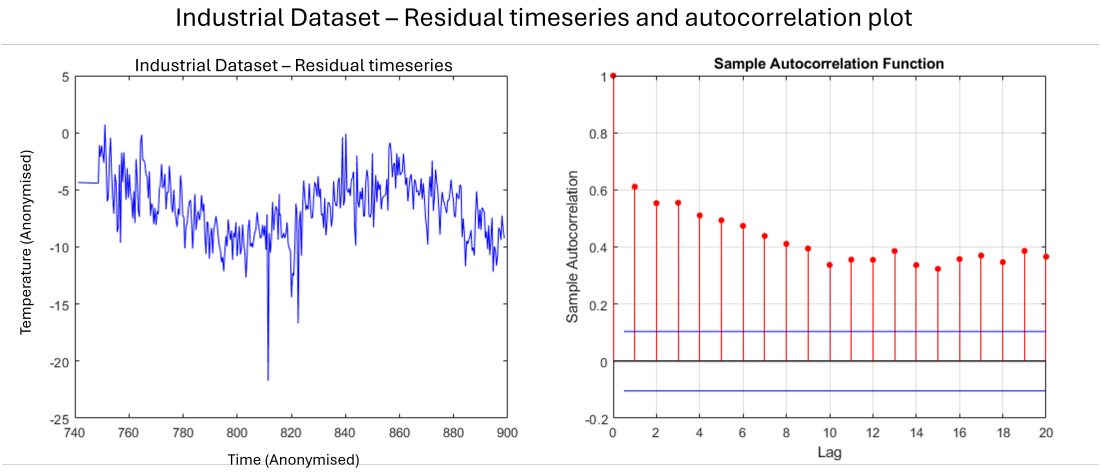

Figure D.3: Base model residual timeseries and autocorrelation plot for the Industrial dataset. The autocorrelation values are significant passed lag 20. The lags used to train the copula model are up to 15 lags, which present a reasonably strong linear trend.

Appendix D. Autocorrelation plots for copula timeseries analysis

Appendix E

Dependency analysis of geometric features for photometric stereo rig error modelling

A total of 36 features derived from the photometric stereo rig geometry were compared to select the feature most suited for modelling the rig error across all objects in the case study. In this section, scatterplots of all 36 features against the Z error between the coordinate measurement machine and photometric stereo rig are included for all objects used in the study. Specifically, the objects utilised are the damaged slab, chimney liner, broken brick and a benchmark dataset of an empty frame (blank background). The features are measured in millimetres for distances and radians for angles, and cover:

- 2 features for the X and Y coordinates
- 2 features for the Euclidean distance to mesh centre (on [X,Y] plane) and the angle to mesh centre (on [X,Y] plane)
- 16 features for the Euclidean distance in 3D to each lighting strip (Directions: North, East, South, West; and angles: 10, 30, 50 and 70 degrees to the horizontal)
- 16 features for the angle from mesh point to each lighting strip (Directions: North, East, South, West; and angles: 10, 30, 50 and 70 degrees to the horizontal)

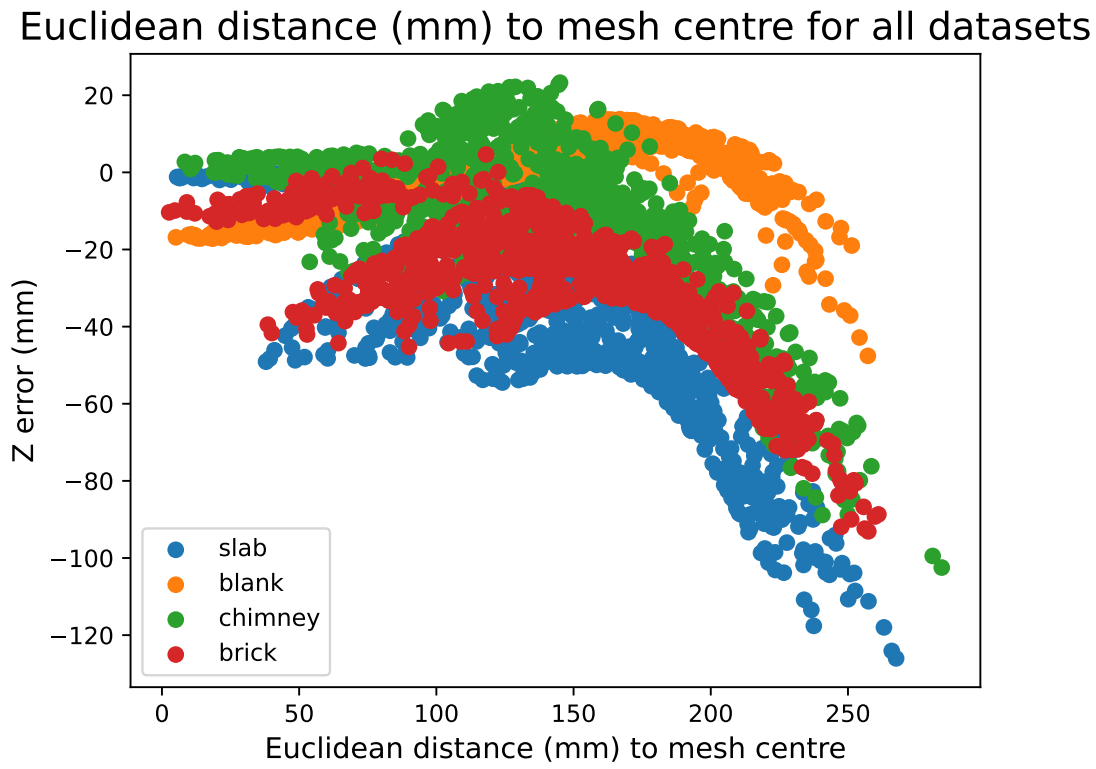

Figure E.1: Chosen feature - Euclidean distance to mesh centre on [X,Y] plane

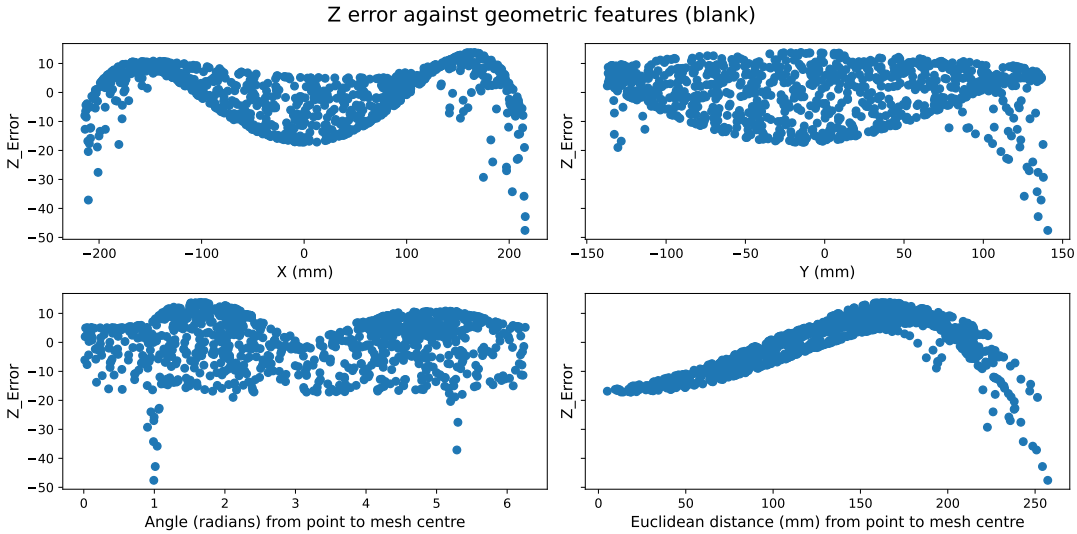

Figure E.2: Blank background - Scatterplots of the X and Y coordinates, 2D angle and 2D Euclidean distance from the mesh centre against the Z error.

Appendix E. Dependency analysis of geometric features for photometric stereo rig error modelling

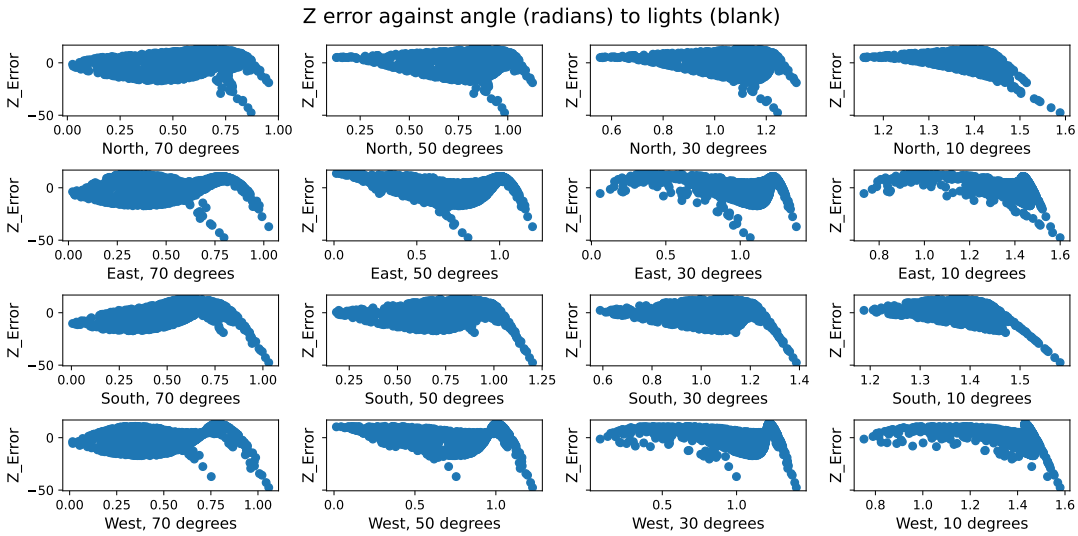

Figure E.3: Blank background - Scatterplots of Z error against the angle to all combinations of lighting direction and lighting angle.

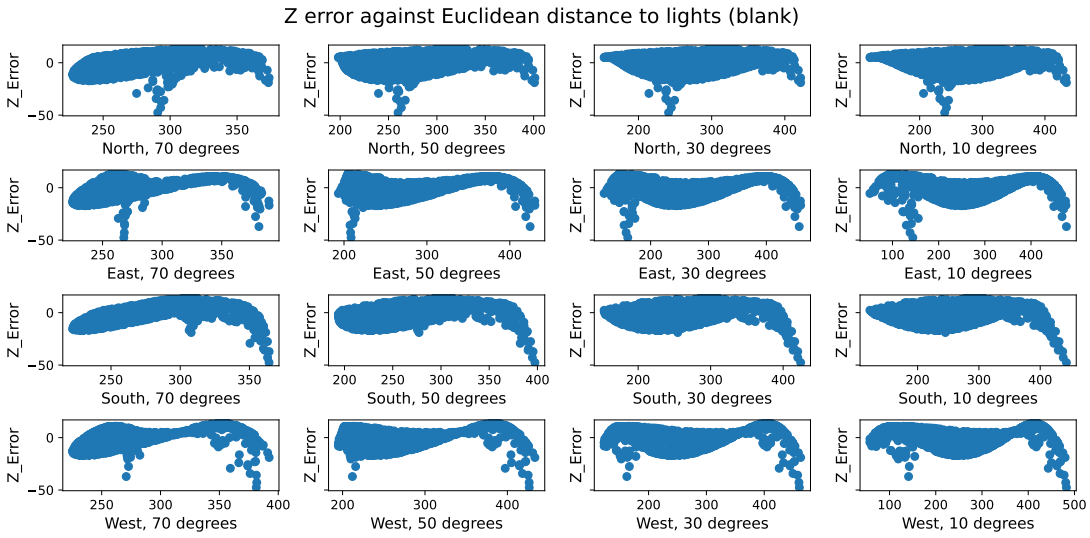

Figure E.4: Blank background - Scatterplots of Z error against the 3D Euclidean distance to all combinations of lighting direction and lighting angle.

Appendix E. Dependency analysis of geometric features for photometric stereo rig error modelling

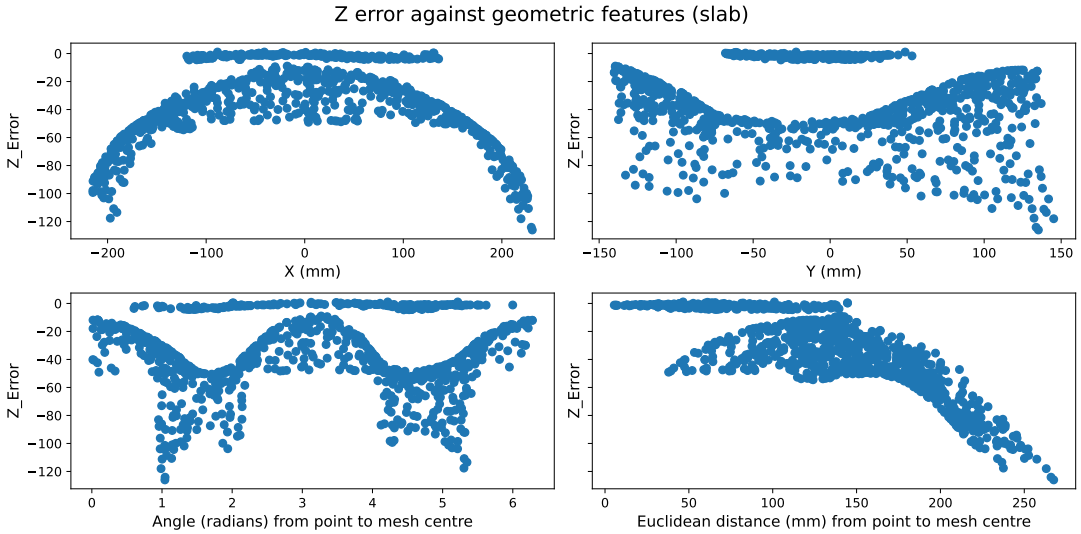

Figure E.5: Damaged slab - Scatterplots of the X and Y coordinates, 2D angle and 2D Euclidean distance from the mesh centre against the Z error.

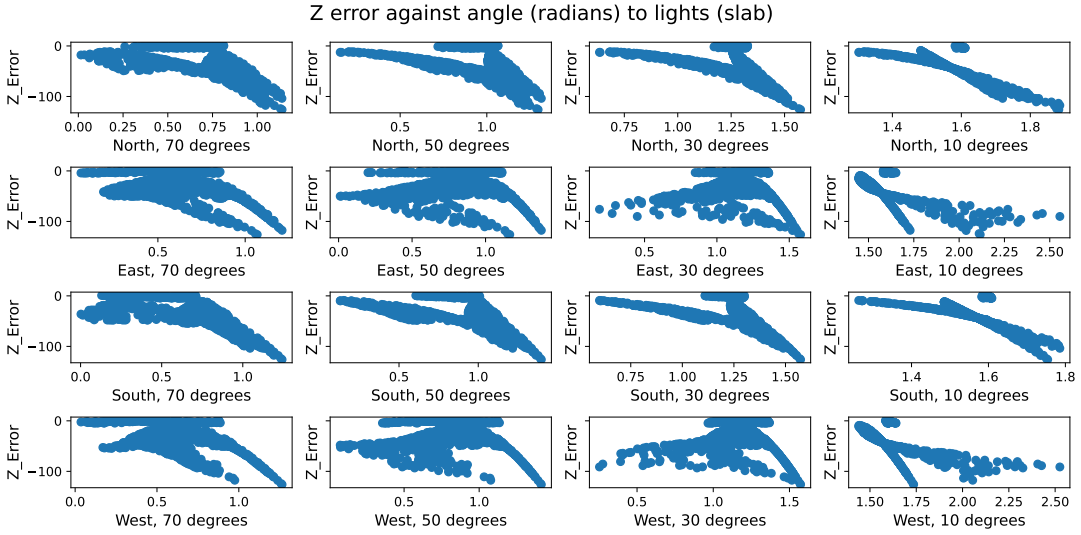

Figure E.6: Damaged slab - Scatterplots of Z error against the angle to all combinations of lighting direction and lighting angle.

Appendix E. Dependency analysis of geometric features for photometric stereo rig error modelling

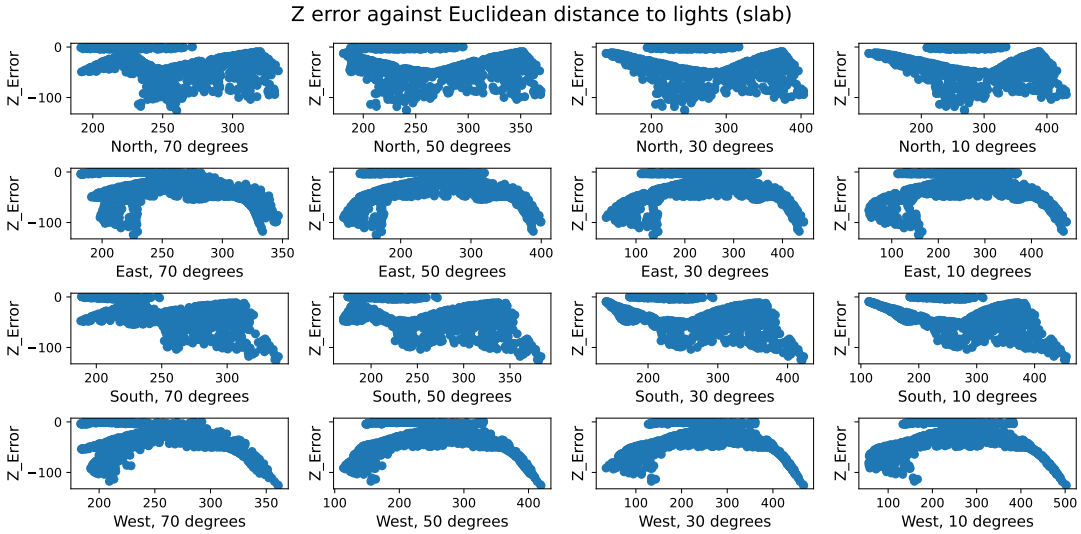

Figure E.7: Damaged slab - Scatterplots of Z error against the 3D Euclidean distance to all combinations of lighting direction and lighting angle.

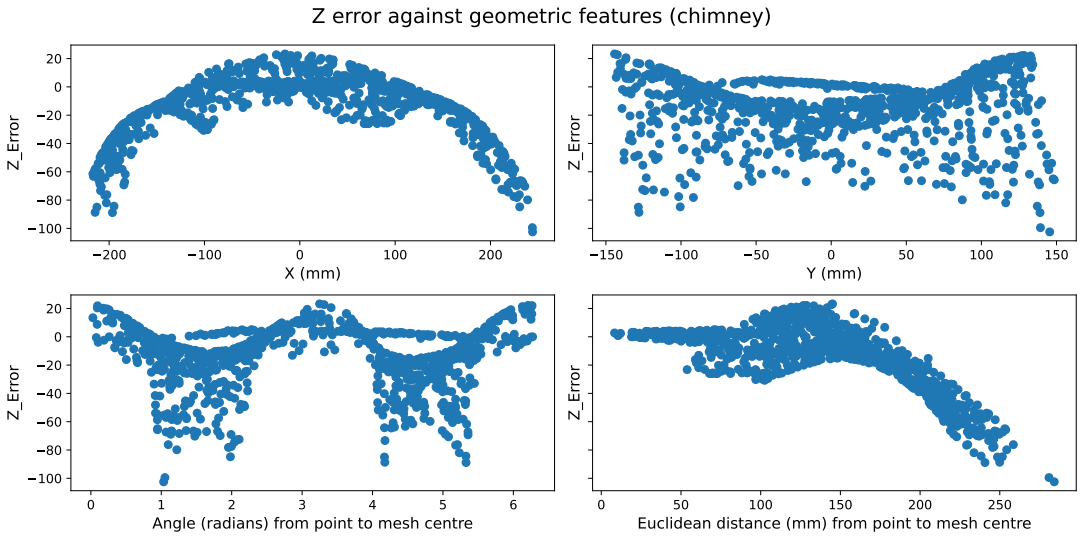

Figure E.8: Chimney liner - Scatterplots of the X and Y coordinates, 2D angle and 2D Euclidean distance from the mesh centre against the Z error.

Appendix E. Dependency analysis of geometric features for photometric stereo rig error modelling

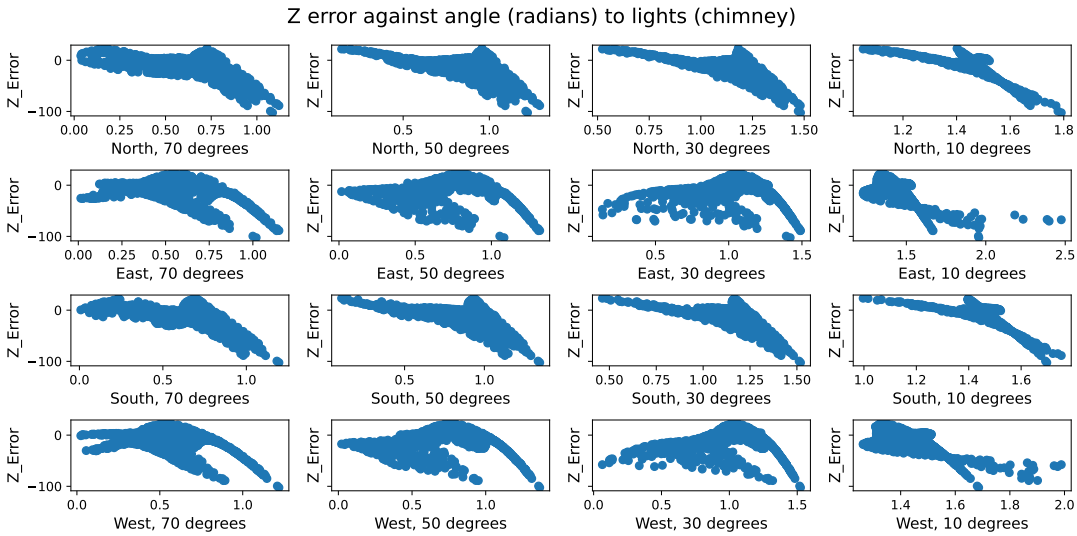

Figure E.9: Chimney liner - Scatterplots of Z error against the angle to all combinations of lighting direction and lighting angle.

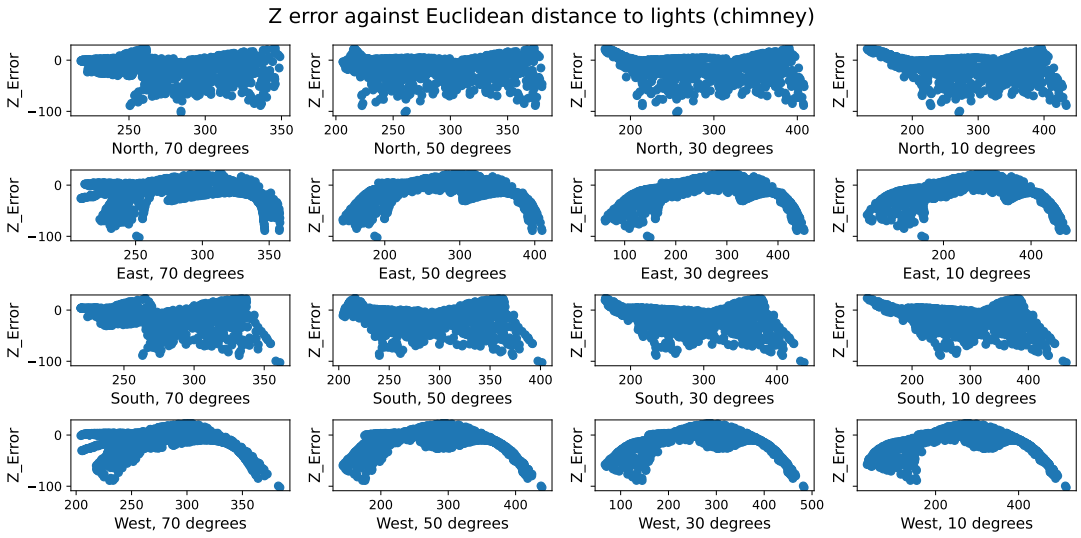

Figure E.10: Chimney liner - Scatterplots of Z error against the 3D Euclidean distance to all combinations of lighting direction and lighting angle.

Appendix E. Dependency analysis of geometric features for photometric stereo rig error modelling

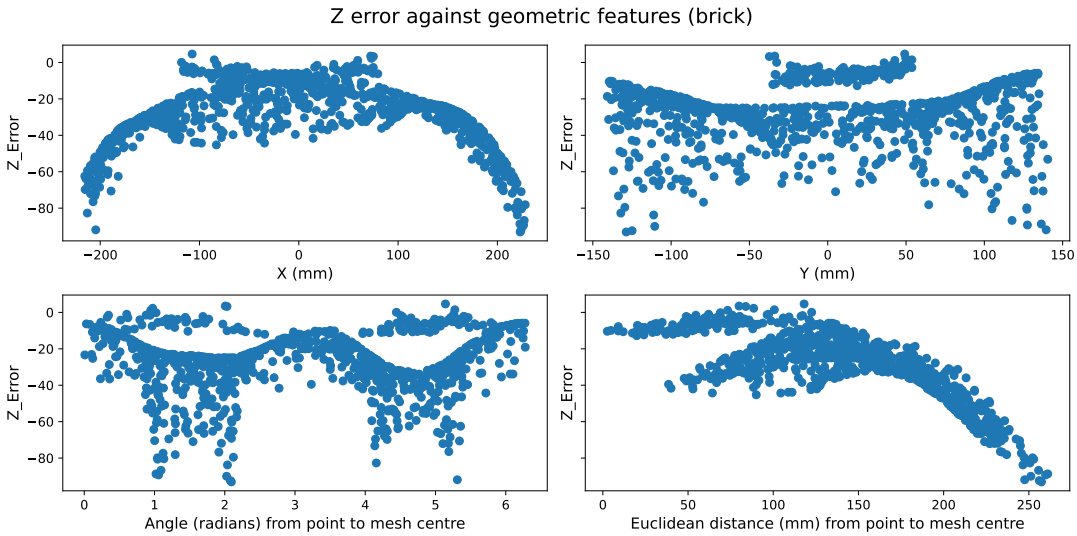

Figure E.11: Broken brick - Scatterplots of the X and Y coordinates, 2D angle and 2D Euclidean distance from the mesh centre against the Z error.

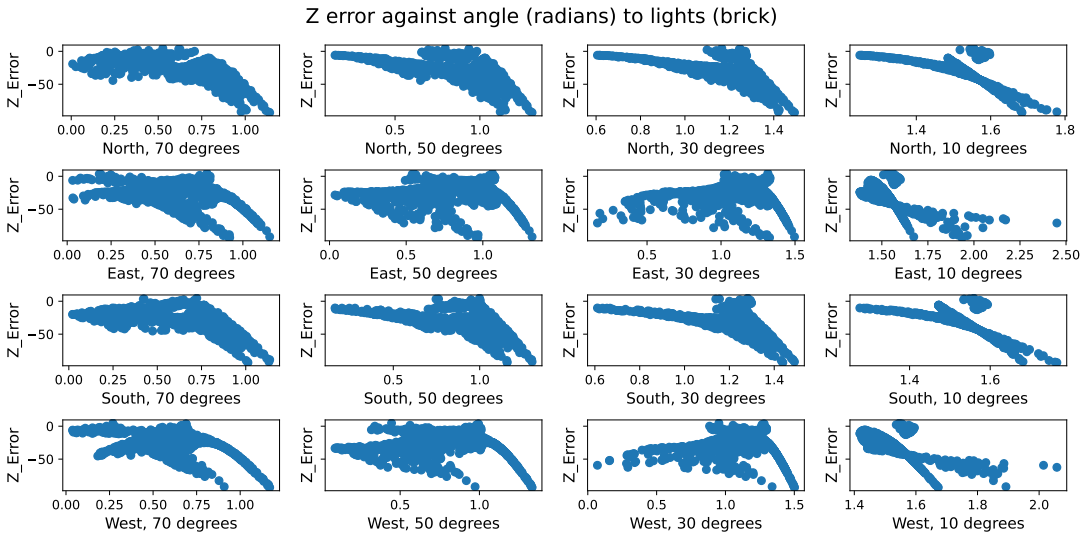

Figure E.12: Broken brick - Scatterplots of Z error against the angle to all combinations of lighting direction and lighting angle.

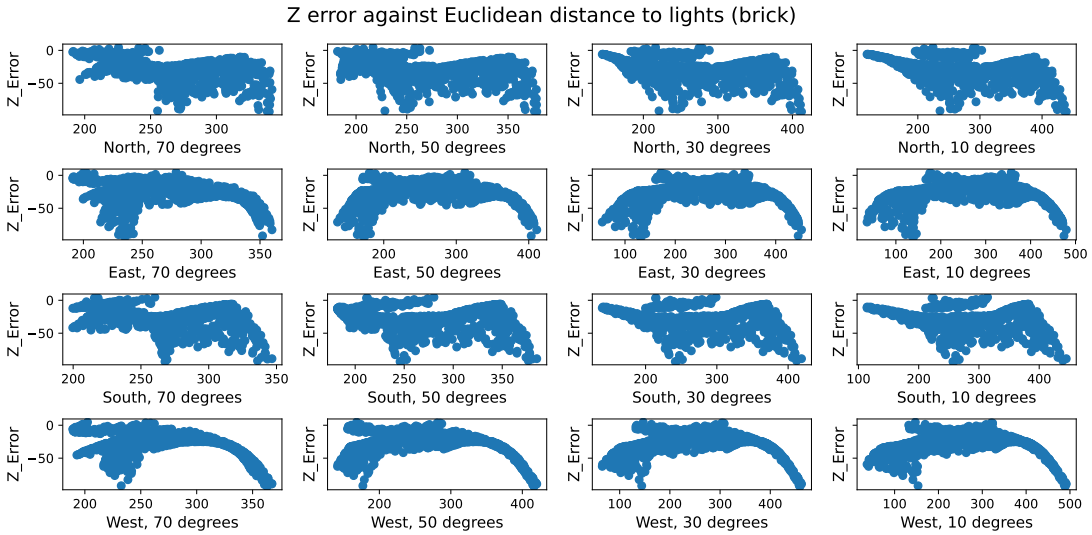

Figure E.13: Broken brick - Scatterplots of Z error against the 3D Euclidean distance to all combinations of lighting direction and lighting angle.

Appendix E. Dependency analysis of geometric features for photometric stereo rig error modelling

Bibliography

[1] L. J. Bond, S. R. Doctor, D. B. Jarrell, and J. W. D. Bond, "Improved economics of nuclear plant life management," in *Second International Symposium on Nuclear Power Plant Life Management*, 2007. [Online]. Available: https://www-pub.iaea.org/MTCD/publications/PDF/P1362_CD/htm/pdf/Keynote%20Speakers/008KS.pdf

[2] D. S. Thomas and B. Weiss, "Maintenance costs and advanced maintenance techniques: Survey and analysis," in *International Journal of Prognostics and Health Management*, vol. 12, no. 1, 2021.

[3] O. Surucu, S. A. Gadsden, and J. Yawney, "Condition monitoring using machine learning: A review of theory, applications, and recent advances," *Expert Systems with Applications*, vol. 221, p. 119738, 2023.

[4] W. Yang and R. Court, "Experimental study on the optimum time for conducting bearing maintenance," *Measurement*, vol. 46, no. 8, pp. 2781–2791, 2013.

[5] X. Tan, L. Fan, Y. Huang, and Y. Bao, "Detection, visualization, quantification, and warning of pipe corrosion using distributed fiber optic sensors," *Automation in Construction*, vol. 132, p. 103953, 2021.

[6] R. Ali, J. H. Chuah, M. S. A. Talip, N. Mokhtar, and M. A. Shoaib, "Structural crack detection using deep convolutional neural networks," *Automation in Construction*, vol. 133, p. 103989, 2022.

[7] G. Shirkoohi, "Modelling of fault detection in electrical wiring," *IET Science, Measurement & Technology*, vol. 9, no. 2, pp. 211–217, 2015.

[8] O. Wallscheid, "Thermal monitoring of electric motors: State-of-the-art review and future challenges," *IEEE Open Journal of Industry Applications*, vol. 2, pp. 204–223, 2021.

[9] Y. Zhang, J. Liu, X. Yang, H. Li, S. Chen, W. Lv *et al.*, "Vibration analysis of a high-pressure multistage centrifugal pump," *Scientific Reports*, vol. 12, 2022.

[10] P. Nunes, J. Santos, and E. Rocha, "Challenges in predictive maintenance – a review," *CIRP Journal of Manufacturing Science and Technology*, vol. 40, pp. 53–67, 2023.

[11] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society Series B*, vol. 63, no. 3, pp. 425–464, 2001.

[12] J. Blair, B. Stephen, B. Brown, A. Forbes, and S. Mcarthur, "Hybrid fault prognostics for nuclear applications: Addressing rotating plant model uncertainty," in *PHM Society European Conference*, vol. 7, no. 1, 2022, p. 58–67.

[13] J. Blair, O. Amin, B. D. Brown, S. McArthur, A. Forbes, and B. Stephen, "The transfer learning of uncertainty quantification for industrial plant fault diagnosis system design," *Data-Centric Engineering*, vol. 5, p. e41, 2024.

[14] J. Blair, B. Stephen, B. Brown, S. McArthur, D. Gorman, A. Forbes *et al.*, "Photometric stereo data for the validation of a structural health monitoring test rig," *Data in Brief*, vol. 53, p. 110164, 2024.

[15] K. M. Sirvio, *Intelligent Systems in Maintenance Planning and Management*. Cham: Springer International Publishing, 2015, pp. 221–245.

[16] R. Ahmad and S. Kamaruddin, "An overview of time-based and condition-based maintenance in industrial application," *Computers & Industrial Engineering*, vol. 63, no. 1, pp. 135–149, 2012.

[17] T. Nakagawa, "Optimal policy of continuous and discrete replacement with minimal repair at failure," *Naval Research Logistics Quarterly*, vol. 31, no. 4, pp. 543–550, 1984.

[18] S. Alaswad and Y. Xiang, "A review on condition-based maintenance optimization models for stochastically deteriorating system," *Reliability Engineering & System Safety*, vol. 157, pp. 54–63, 2017.

[19] Y. Hu, X. Miao, Y. Si, E. Pan, and E. Zio, "Prognostics and health management: A review from the perspectives of design, development and decision," *Reliability Engineering & System Safety*, vol. 217, p. 108063, 2022.

[20] L. Hui and O. Jinping, "Structural health monitoring: From sensing technology stepping to health diagnosis," *Procedia Engineering*, vol. 14, pp. 753–760, 2011, the Proceedings of the Twelfth East Asia-Pacific Conference on Structural Engineering and Construction.

[21] T. Tinga, "Practical issues and challenges in predictive maintenance," 2021, european Conference of the Prognostics and Health Management Society (PHM-E).

[22] P. Muganyi, C. Mbohwa, and I. Madanhire, "Design-out maintenance as a crucial maintenance facet," in *8th Annual International Conference on Industrial Engineering and Operations Management*. IEOM Society International, 2018.

[23] J. Coble, P. Ramuhalli, L. Bond, J. Hines, and B. Upadhyaya, "A review of prognostics and health management applications in nuclear power plants," *International Journal of Prognostics and Health Management*, vol. 6, pp. 1–22, 07 2015.

[24] K. L. Tsui, N. Chen, Q. Zhou, Y. Hai, and W. Wang, "Prognostics and health management: A review on data driven approaches," *Mathematical Problems in Engineering*, vol. 2015, no. 1, p. 793161, 2015.

[25] A. Young, G. West, B. Brown, B. Stephen, A. Duncan, C. Michie *et al.*, "Capturing symbolic expert knowledge for the development of industrial fault detection

systems: manual and automated approaches," *International Journal of Condition Monitoring and Diagnostic Management,*, vol. 25, no. 2, pp. 67–75, jun 2022.

[26] A. Lundgren and D. Jung, "Data-driven fault diagnosis analysis and open-set classification of time-series data," *Control Engineering Practice*, vol. 121, p. 105006, 2022.

[27] R. Jiao, K. Peng, J. Dong, and C. Zhang, "Fault monitoring and remaining useful life prediction framework for multiple fault modes in prognostics," *Reliability Engineering & System Safety*, vol. 203, p. 107028, 2020.

[28] K. Huynh, A. Grall, and C. Bérenguer, "Assessment of diagnostic and prognostic condition indices for efficient and robust maintenance decision-making of systems subject to stress corrosion cracking," *Reliability Engineering & System Safety*, vol. 159, pp. 237–254, 2017.

[29] N. G. N. Irias, F. A. L. Souza, H. de Paula, and L. A. R. Silva, "Challenges in using the physics-of-failure approach in practical applications," in *2017 IEEE Industry Applications Society Annual Meeting*, 2017, pp. 1–8.

[30] G. Y. Heo, "Condition monitoring using empirical models: Technical review and prospects for nuclear applications," *Nuclear Engineering and Technology*, vol. 40, no. 1, p. 49 – 68, 2008.

[31] J. Ma and J. Jiang, "Semisupervised classification for fault diagnosis in nuclear power plants," *Nuclear Engineering and Technology*, vol. 47, no. 2, pp. 176–186, 2015, special Issue on ISOFIC/ISSNP2014.

[32] K. S. Kiangala and Z. Wang, "An effective predictive maintenance framework for conveyor motors using dual time-series imaging and convolutional neural network in an industry 4.0 environment," *IEEE Access*, vol. 8, pp. 121 033–121 049, 2020.

[33] M. G. Minguell and R. Pandit, "Tracksafe: A comparative study of data-driven techniques for automated railway track fault detection using image datasets," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106622, 2023.

[34] Z. Allal, H. N. Noura, F. Vernier, O. Salman, and K. Chahine, "Wind turbine fault detection and identification using a two-tier machine learning framework," *Intelligent Systems with Applications*, vol. 22, p. 200372, 2024.

[35] M.-C. Flynn, M. Sztykiel, C. E. Jones, P. J. Norman, G. M. Burt, P. Miller *et al.*, "Protection and fault management strategy maps for future electrical propulsion aircraft," *IEEE Transactions on Transportation Electrification*, vol. 5, no. 4, pp. 1458–1469, 2019.

[36] K. Ahlstrom, J. Torin, K. Fersan, and P. Nobrant, "Redundancy management in distributed flight control systems: experience & simulations," in *Proceedings. The 21st Digital Avionics Systems Conference*, vol. 2, 2002, pp. 13C3–13C3.

[37] N. Costa and L. Sánchez, "Variational encoding approach for interpretable assessment of remaining useful life estimation," *Reliability Engineering & System Safety*, vol. 222, p. 108353, 2022.

[38] T. Berghout, L.-H. Mouss, O. Kadri, L. Saïdi, and M. Benbouzid, "Aircraft engines remaining useful life prediction with an adaptive denoising online sequential extreme learning machine," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103936, 2020.

[39] L. Van Eykeren and Q. Chu, "Sensor fault detection and isolation for aircraft control systems by kinematic relations," *Control Engineering Practice*, vol. 31, pp. 200–210, 2014.

[40] M. A. Kassab, H. S. Taha, S. A. Shedied, and A. Maher, "A novel voting algorithm for redundant aircraft sensors," in *Proceeding of the 11th World Congress on Intelligent Control and Automation*, 2014, pp. 3741–3746.

[41] C. E. Jones, P. J. Norman, G. M. Burt, C. Hill, G. Allegri, J. M. Yon *et al.*, "A route to sustainable aviation: A roadmap for the realization of aircraft components with electrical and structural multifunctionality," *IEEE Transactions on Transportation Electrification*, vol. 7, no. 4, pp. 3032–3049, 2021.

[42] K. Kabbabe Poleo, W. J. Crowther, and M. Barnes, "Estimating the impact of drone-based inspection on the levelised cost of electricity for offshore wind farms," *Results in Engineering*, vol. 9, p. 100201, 2021.

[43] D. A. Rodríguez, C. Lozano Tafur, P. F. Melo Daza, J. A. Villalba Vidales, and J. C. Daza Rincón, "Inspection of aircrafts and airports using uas: A review," *Results in Engineering*, vol. 22, p. 102330, 2024.

[44] A. Reddy, V. Indragandhi, L. Ravi, and V. Subramaniyaswamy, "Detection of cracks and damage in wind turbine blades using artificial intelligence-based image analytics," *Measurement*, vol. 147, p. 106823, 2019.

[45] M. Mandirola, C. Casarotti, S. Peloso, I. Lanese, E. Brunesi, and I. Senaldi, "Use of uas for damage inspection and assessment of bridge infrastructures," *International Journal of Disaster Risk Reduction*, vol. 72, p. 102824, 2022.

[46] R. Watson, M. Kamel, D. Zhang, G. Dobie, C. MacLeod, S. G. Pierce *et al.*, "Dry coupled ultrasonic non-destructive evaluation using an over-actuated unmanned aerial vehicle," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 4, pp. 2874–2889, 2022.

[47] T. Zhao, R. Watson, D. Zhang, R. McMillan, W. Galbraith, C. MacLeod *et al.*, "A pulsed eddy current sensor for uav deployed pipe thickness measurement," *IEEE Sensors Journal*, pp. 1–1, 2024.

[48] J. Faiz and M. Soleimani, "Dissolved gas analysis evaluation in electric power transformers using conventional methods a review," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 24, no. 2, pp. 1239–1248, 2017.

[49] Y. Hua, Y. Sun, G. Xu, S. Sun, E. Wang, and Y. Pang, "A fault diagnostic method for oil-immersed transformer based on multiple probabilistic output algorithms and improved ds evidence theory," *International Journal of Electrical Power & Energy Systems*, vol. 137, p. 107828, 2022.

[50] M. S. Travers, "Chapter 14 - reducing collisions with structures," in *Conservation of Marine Birds*, L. Young and E. VanderWerf, Eds. Academic Press, 2023, pp. 379–401.

[51] M. Fonseca-Badillo, L. Negrete-Navarrete, A. González-Parada, and A. Castañeda-Miranda, "Simulation and analysis of underground power cables faults," *Procedia Engineering*, vol. 35, pp. 50–57, 2012, international Meeting of Electrical Engineering Research 2012.

[52] B. Taormina, J. Bald, A. Want, G. Thouzeau, M. Lejart, N. Desroy *et al.*, "A review of potential impacts of submarine power cables on the marine environment: Knowledge gaps, recommendations and future directions," *Renewable and Sustainable Energy Reviews*, vol. 96, pp. 380–391, 2018.

[53] X. Jiang, E. Corr, B. Stephen, and B. G. Stewart, "Impact of increased penetration of low-carbon technologies on cable lifetime estimations," *Electricity*, vol. 3, no. 2, pp. 220–235, 2022.

[54] F. Fan, B. Stephen, K. Bell, D. Infield, and S. McArthur, "Impacts of measurement errors on real-time thermal rating estimation for overhead lines," *IEEE Transactions on Power Delivery*, vol. 38, no. 2, pp. 1086–1096, 2023.

[55] M. F. Goni, M. Nahiduzzaman, M. Anower, M. Rahman, M. Islam, M. Ahsan *et al.*, "Fast and accurate fault detection and classification in transmission lines using extreme learning machine," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 3, p. 100107, 2023.

[56] S. Bandara, P. Rajeev, E. Gad, B. Sriskantharajah, and I. Flatley, "Damage detection of in service timber poles using hilbert-huang transform," *NDT & E International*, vol. 107, p. 102141, 2019.

[57] M. Hosseini, J. Helm, B. Stephen, and S. D. J. McArthur, "Automated feature validation of trip coil analysis in condition monitoring of circuit breakers," in *Proceedings of the European Conference of the PHM Society 2018*, C. S. Kulkarni and T. Tinga, Eds., vol. 4. PHM Society European Conference, 2018.

[58] S. Wang, Y. Zhou, and Z. Ma, "Research on fault identification of high-voltage circuit breakers with characteristics of voiceprint information," *Scientific Reports*, vol. 14, no. 9340, 2024.

[59] Q. Yang, J. Ruan, Z. Zhuang, and D. Huang, "Condition evaluation for opening damper of spring operated high-voltage circuit breaker using vibration time-frequency image," *IEEE Sensors Journal*, vol. 19, no. 18, pp. 8116–8126, 2019.

[60] A. A. Razi-Kazemi, M. Vakilian, K. Niayesh, and M. Lehtonen, "Circuit-breaker automated failure tracking based on coil current signature," *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 283–290, 2014.

[61] O. T. Bindingsbø, M. Singh, K. Øvsthus, and A. Keprate, "Fault detection of a wind turbine generator bearing using interpretable machine learning," *Frontiers in Energy Research*, vol. 11, 2023.

[62] R. Melício, V. Mendes, and J. Catalão, "Power converter topologies for wind energy conversion systems: Integrated modeling, control strategy and performance simulation," *Renewable Energy*, vol. 35, no. 10, pp. 2165–2174, 2010.

[63] J. Liang, K. Zhang, A. Al-Durra, S. Muyeen, and D. Zhou, "A state-of-the-art review on wind power converter fault diagnosis," *Energy Reports*, vol. 8, pp. 5341–5369, 2022.

[64] V. L. Jantara and M. Papaelias, "Chapter 5 - wind turbine gearboxes: Failures, surface treatments and condition monitoring," in *Non-Destructive Testing and Condition Monitoring Techniques for Renewable Energy Industrial Assets*, M. Papaelias, F. P. G. Márquez, and A. Karyotakis, Eds. Boston: Butterworth-Heinemann, 2020, pp. 69–90.

[65] J. C. Lopez and A. Kolios, "Risk-based maintenance strategy selection for wind turbine composite blades," *Energy Reports*, vol. 8, pp. 5541–5561, 2022.

[66] D. Zhou, D. Huang, J. Hao, H. Wu, C. Chang, and H. Zhang, "Fault diagnosis of gas turbines with thermodynamic analysis restraining the interference of bound-

ary conditions based on stn," *International Journal of Mechanical Sciences*, vol. 191, p. 106053, 2021.

[67] M. A. Chao, D. S. Lilley, P. Mathé, and V. Schloßhauer, "Calibration and Uncertainty Quantification of Gas Turbine Performance Models," in *Turbo Expo: Power for Land, Sea, and Air*, vol. Volume 7A: Structures and Dynamics, June 2015.

[68] A. Gofuku, "Integrated diagnostic technique for nuclear power plants," *Nuclear Engineering and Technology*, vol. 46, no. 6, pp. 725–736, 2014.

[69] X. Zhong and H. Ban, "Crack fault diagnosis of rotating machine in nuclear power plant based on ensemble learning," *Annals of Nuclear Energy*, vol. 168, p. 108909, 2022.

[70] P. Kundu, A. K. Darpe, and M. S. Kulkarni, "A review on diagnostic and prognostic approaches for gears," *Structural Health Monitoring*, vol. 20, no. 5, pp. 2853–2893, 2021.

[71] H. M. Hashemian and W. C. Bean, "State-of-the-art predictive maintenance techniques*," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 10, pp. 3480–3492, 2011.

[72] Z. Dong, Y. Pan, and X. Huang, "Parameter identifiability of boolean networks with application to fault diagnosis of nuclear plants," *Nuclear Engineering and Technology*, vol. 50, no. 4, pp. 599–605, 2018, international Symposium on Future I&C for Nuclear Power Plants (ISOFIC2017).

[73] J. Coble, P. Ramuhalli, R. Meyer, and H. Hashemian, "Online sensor calibration assessment in nuclear power systems," *IEEE Instrumentation & Measurement Magazine*, vol. 16, no. 3, pp. 32–37, 2013.

[74] H. Hashemian, "On-line monitoring and calibration techniques in nuclear power plants (iaea-cn–164)," International Atomic Energy Agency (IAEA), Tech. Rep., 2009.

[75] H. M. Hashemian, *Cross-Calibration Technique*, 1st ed., ser. Power Systems. Springer Berlin, Heidelberg, 2006.

[76] S.-Y. Chu and C.-J. Kang, "Development of the structural health record of containment building in nuclear power plant," *Nuclear Engineering and Technology*, vol. 53, no. 6, pp. 2038–2045, 2021.

[77] G. West, P. Murray, S. Marshall, and S. McArthur, "Improved visual inspection of advanced gas-cooled reactor fuel channels," *International Journal of Prognostics and Health Management*, vol. 6, no. Special Issue - Nuclear Energy, jul 2015.

[78] M. G. Devereux, P. Murray, and G. M. West, "A new approach for crack detection and sizing in nuclear reactor cores," *Nuclear Engineering and Design*, vol. 359, p. 110464, 2020.

[79] X. Sun, B. Lin, J. Bao, V. Giurgiutiu, T. Knight, P.-S. Lam *et al.*, "Developing a structural health monitoring system for nuclear dry cask storage canister," in *Smart Materials and Nondestructive Evaluation for Energy Systems 2015*, N. G. Meyendorf, Ed., vol. 9439, International Society for Optics and Photonics. SPIE, 2015, p. 94390N.

[80] C. Nash, P. Karve, and D. Adams, "Diagnosing nuclear power plant pipe wall thinning due to flow accelerated corrosion using a passive, thermal nondestructive evaluation method: Feasibility assessment via numerical experiments," *Nuclear Engineering and Design*, vol. 386, p. 111542, 2022.

[81] R. Gomasa, V. Talakokula, S. Kalyana Rama Jyosyula, and T. Bansal, "A review on health monitoring of concrete structures using embedded piezoelectric sensor," *Construction and Building Materials*, vol. 405, p. 133179, 2023.

[82] The Rt Hon Anne-Marie Trevelyan MP and The Rt Hon Alok Sharma MP, "End to coal power brought forward to october 2024," Department for Business, Energy & Industrial Strategy, Tech. Rep., 2021, accessed: 30/09/2021. [Online]. Available: https://www.gov.uk/government/news/end-to-coal-power-brought-forward-to-october-2024

[83] World Nuclear Association, "How can nuclear combat climate change?" World Nuclear Association, Tech. Rep., Accessed: 30/09/2021. [Online]. Available: https://world-nuclear.org/nuclear-essentials/how-can-nuclear-combat-climate-change

[84] World Nuclear Association, "Outline history of nuclear energy," World Nuclear Association, Tech. Rep., Updated May 2024, accessed: 23/04/2021. [Online]. Available: https://world-nuclear.org/information-library/current-and-future-generation/outline-history-of-nuclear-energy

[85] L. J. Bond, P. Ramuhalli, M. S. Tawfik, and N. J. Lybeck, "Prognostics and life beyond 60 years for nuclear power plants," in *2011 IEEE Conference on Prognostics and Health Management*, 2011, pp. 1–7.

[86] G. J. Toman and A. Mantey, "Cable system aging management for nuclear power plants," in *2012 IEEE International Symposium on Electrical Insulation*, 2012, pp. 315–318.

[87] M. Muhlheim, R. Belles, and R. Hardin, "Criteria for determining the safety of wireless technologies at nuclear power plants," *18th International Probabilistic Safety Assessment and Analysis*, 7 2023.

[88] F. Zhang and J. B. Coble, "Robust localized cyber-attack detection for key equipment in nuclear power plants," *Progress in Nuclear Energy*, vol. 128, p. 103446, 2020.

[89] F. Zhang, H. A. D. E. Kodituwakku, J. W. Hines, and J. Coble, "Multilayer data-driven cyber-attack detection system for industrial control systems based on network, system, and process data," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4362–4369, 2019.

[90] Nuclear Threat Initiative, "Nti nuclear security index: Falling short in a dangerous world," Nuclear Threat Initiative, Tech. Rep., 2023.

[91] X. Zhao, J. Kim, K. Warns, X. Wang, P. Ramuhalli, S. Cetiner *et al.*, "Prognostics and health management in nuclear power plants: An updated method-centric review with special focus on data-driven methods," *Frontiers in Energy Research*, vol. 9, 2021.

[92] N. Nakagawa, F. Inanc, A. Frishman, R. B. Thompson, W. Junker, F. Ruddy *et al.*, "On-line nde and structural health monitoring for advanced reactors," *Key Engineering Materials*, vol. 321-323, pp. 234–239, 2006.

[93] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *Journal of Big Data*, vol. 7, no. 41, 2020.

[94] S. Dutta, B. Basu, and F. A. Talukdar, "Classification of motor faults based on transmission coefficient and reflection coefficient of omni-directional antenna using dcnn," *Expert Systems with Applications*, vol. 198, p. 116832, 2022.

[95] F. Pallonetto, C. Jin, and E. Mangina, "Forecast electricity demand in commercial building with machine learning models to enable demand response programs," *Energy and AI*, vol. 7, p. 100121, 2022.

[96] J. Guo, J.-L. Wan, Y. Yang, L. Dai, A. Tang, B. Huang *et al.*, "A deep feature learning method for remaining useful life prediction of drilling pumps," *Energy*, vol. 282, p. 128442, 2023.

[97] J. Wang and F. Biljecki, "Unsupervised machine learning in urban studies: A systematic review of applications," *Cities*, vol. 129, p. 103925, 2022.

[98] S. Ochella, M. Shafiee, and C. Sansom, "Adopting machine learning and condition monitoring p-f curves in determining and prioritizing high-value assets for life extension," *Expert Systems with Applications*, vol. 176, p. 114897, 2021.

[99] S. Messaoud, A. Bradai, S. H. R. Bukhari, P. T. A. Quang, O. B. Ahmed, and M. Atri, "A survey on machine learning in internet of things: Algorithms, strategies, and applications," *Internet of Things*, vol. 12, p. 100314, 2020.

[100] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 160, 2021.

[101] R. He, Z. Tian, and M. Zuo, "Machine prognostics under varying operating conditions based on state-space and neural network modeling," *Mechanical Systems and Signal Processing*, vol. 182, p. 109598, 2023.

[102] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: Bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006. [Online]. Available: https://www.jstor.org/stable/25470329

[103] R. Nisbet, J. Elder, and G. Miner, "Chapter 8 - advanced algorithms for data mining," in *Handbook of Statistical Analysis and Data Mining Applications*, R. Nisbet, J. Elder, and G. Miner, Eds. Boston: Academic Press, 2009, pp. 151–172.

[104] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed., ser. Springer Series in Statistics. Springer, 2009.

[105] A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, *Data Mining in Agriculture*, 1st ed., ser. Springer Optimization and Its Applications. Springer, 2009.

[106] A. Thompson, "Uncertainty evaluation for machine learning: metrology requirements and open challenges," in *Mathematical and Statistical Methods for Metrology Conference (MSMM)*, 2021. [Online]. Available: http://www.msmm2021.polito.it/programme

[107] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to rul prediction," *Mechanical Systems and Signal Processing*, vol. 104, pp. 799–834, 2018.

[108] E. Hart, "Wind turbine dynamics identification using gaussian process machine learning," Ph.D. dissertation, University of Strathclyde, 2018.

[109] M. Baur, P. Albertelli, and M. Monno, "A review of prognostics and health management of machine tools," in *The International Journal of Advanced Manufacturing Technology*, 03 2020.

[110] K. Goebel, N. Eklund, and P. Bonanni, "Fusing competing prediction algorithms for prognostics," in *IEEE Aerospace Conference Proceedings*, vol. 2006, 01 2006, p. 10 pp.

[111] M. Arias Chao, C. Kulkarni, K. Goebel, and O. Fink, "Fusing physics-based and deep learning models for prognostics," *Reliability Engineering & System Safety*, vol. 217, p. 107961, 2022.

[112] P. Kundu, A. K. Darpe, and M. S. Kulkarni, "Development of data-driven, physics-based, and hybrid prognosis frameworks: a case study for gear remaining useful life prediction," *Journal of Intelligent Manufacturing*, 2024.

[113] P. Ramuhalli, C. Walker, V. Agarwal, and N. J. Lybeck, "Development of prognostic models using plant asset data," Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States); Idaho National ..., Tech. Rep., 2020.

[114] L. Liao and F. Köttig, "Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction," *IEEE Transactions on Reliability*, vol. 63, no. 1, pp. 191–207, 2014.

[115] T. von Hahn and C. Mechefske, "Knowledge informed machine learning using a weibull-based loss function," *Journal of Prognostics and Health Management*, vol. 2, no. 1, p. 9–44, Jul. 2022, code available: https://github.com/tvhahn/weibull-knowledge-informed-ml.

[116] E. Hart, "Wind turbine dynamics identification using gaussian process machine learning," PhD thesis, University of Strathclyde, Glasgow, UK, 2018, available at https://stax.strath.ac.uk/concern/theses/ht24wj46b.

[117] J. Görtler, R. Kehlbeck, and O. Deussen, "A visual exploration of gaussian processes," *Distill*, 2019, https://distill.pub/2019/visual-exploration-gaussian-processes.

[118] J. Donlevy, K. Jagan, and A. Thompson, "A gaussian process approach to uncertainty evaluation for machine learning," in *Mathematical and Statistical Methods for Metrology Conference (MSMM)*, 2021. [Online]. Available: http://www.msmm2021.polito.it/programme

[119] K. Jagan and A. Forbes, "Approximating gaussian process regression models using banded matrices," in *Mathematical and Statistical Methods for Metrology Conference (MSMM)*, 2021. [Online]. Available: http://www.msmm2021.polito.it/programme

[120] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, no. 12, pp. 371–421, 2008. [Online]. Available: http://jmlr.org/papers/v9/shafer08a.html

[121] H. Olsson, K. Kartasalo, N. Mulliqi, M. Capuccini, P. Ruusuvuori, H. Samaratunga *et al.*, "Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction," *Nature Communications*, vol. 13, no. 7761, 2022.

[122] V. N. Balasubramanian, S.-S. Ho, and V. Vovk, "Foreword," in *Conformal Prediction for Reliable Machine Learning*. Boston: Morgan Kaufmann, 2014, pp. xv–xvi.

[123] R. Foygel Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "The limits of distribution-free conditional predictive inference," *Information and Inference: A Journal of the IMA*, vol. 10, no. 2, pp. 455–482, 08 2020.

[124] N. Kudryashova, T. Amvrosiadis, N. Dupuy, N. Rochefort, and A. Onken, "Parametric copula-gp model for analyzing multidimensional neuronal and behavioral relationships," *PLoS Comput Biol*, vol. 18, no. 1, 2021.

[125] A. Moller, A. Lenkoski, and T. Thorarinsdottir, "Multivariate probabilistic forecasting using ensemble bayesian model averaging and copulas," *Quarterly Journal of the Royal Meteorological Society*, vol. 139, p. 982 – 991, 2013.

[126] J. Hernández-Lobato, J. R. Lloyd, and D. Hernández-Lobato, "Gaussian process conditional copulas with applications to financial time series," *Advances in Neural Information Processing Systems*, vol. 26, 2013. [Online]. Available: https://proceedings.neurips.cc/paper/2013

[127] M. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Annales de l'ISUP*, vol. VIII (3), pp. pp.229–231, 1959.

[128] J. Singer and D. Andrade, "Large-sample statistical methods," *International Encyclopedia of Education (Third Edition)*, pp. pp.232–237, 2010.

[129] M. Käärik, A. Selart, and E. Käärik, "The use of copulas to model conditional expectation for multivariate data," in *Proceedings of the 2011 World Statistics Congress*. Dublin: International Statistical Institute, 2011. [Online]. Available: https://2011.isiproceedings.org/papers/950771.pdf

[130] A. AghaKouchak, A. Bardossy, and E. Habib, "Copula-based uncertainty modeling: Application to multi-sensor precipitation estimates," *Hydrological Processes*, vol. 24, no. 15, pp. 2111–2124, 2010.

[131] A. Possolo, "Copulas for uncertainty analysis," *Metrologia*, vol. 47, no. 3, p. 262, apr 2010.

[132] E. S. Simpson, J. L. Wadsworth, and J. A. Tawn, "A geometric investigation into the tail dependence of vine copulas," *Journal of Multivariate Analysis*, vol. 184, p. 104736, 2021.

[133] D. Lopez-Paz, J. M. Hernández-Lobato, and G. Zoubin, "Gaussian process vine copulas for multivariate dependence," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 2. Atlanta,

Georgia, USA: PMLR, 17–19 Jun 2013, pp. 10–18. [Online]. Available: https://proceedings.mlr.press/v28/lopez-paz13.html

[134] Z. Liu, P. Zhou, X. Chen, and Y. Guan, "A multivariate conditional model for streamflow prediction and spatial precipitation refinement," *Journal of Geophysical Research: Atmospheres*, vol. 120, no. 19, pp. 10,116–10,129, 2015.

[135] L. A. Bull, D. Di Francesco, M. Dhada, O. Steinert, T. Lindgren, A. K. Parlikad *et al.*, "Hierarchical bayesian modeling for knowledge transfer across engineering fleets via multitask learning," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 7, pp. 821–848, 2023.

[136] B. Schulz, M. El Ayari, S. Lerch, and S. Baran, "Post-processing numerical weather prediction ensembles for probabilistic solar irradiance forecasting," *Solar Energy*, vol. 220, pp. 1016–1031, 2021.

[137] B. Stephen, R. Telford, and S. Galloway, "Non-gaussian residual based short term load forecast adjustment for distribution feeders," *IEEE Access*, vol. 8, pp. 10 731–10 741, 2020.

[138] High-level expert group on artificial intelligence (AI HLEG), "Ethics guidelines for trustworthy ai," European Commission, Tech. Rep., 2018. [Online]. Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

[139] O. for Artificial Intelligence, "National ai strategy," HM Government, United Kingdom, Tech. Rep., 2021. [Online]. Available: https://lordslibrary.parliament.uk/artificial-intelligence-development-risks-and-regulation/

[140] White House Office of Science and Technology Policy, "Blueprint for an ai bill of rights: Making automated systems work for the american people," United States of America, Tech. Rep., 2022. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[141] I. Saifa and B. Ammanath, "'trustworthy ai' is a framework to help manage unique risk," MIT Technology Review, Tech. Rep., 2020.

[Online]. Available: https://www.technologyreview.com/2020/03/25/950291/trustworthy-ai-is-a-framework-to-help-manage-unique-risk/

[142] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum *et al.*, "Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations," *Minds & Machines*, vol. 28, pp. 689–707, 2018.

[143] M. D. R. AI_, "Montreal declaration for a responsible development of artificial intelligence," Universite de Montreal, Tech. Rep., 2018. [Online]. Available: https://montrealdeclaration-responsibleai.com/

[144] W. Wu, T. Huang, and K. Gong, "Ethical principles and governance technology development of ai in china," *Engineering*, vol. 6, no. 3, pp. 302–309, 2020.

[145] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electron Markets*, vol. 31, pp. 447–464, 2021.

[146] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[147] D. H. Mcknight, M. Carter, J. B. Thatcher, and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 2, jul 2011.

[148] J. B. Thatcher, D. H. McKnight, E. W. Baker, R. E. Arsal, and N. H. Roberts, "The role of trust in postadoption it exploration: An empirical examination of knowledge management systems," *IEEE Transactions on Engineering Management*, vol. 58, no. 1, pp. 56–70, 2011.

[149] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 1995.

[150] D. H. McKnight, V. Choudhury, and Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," *Information Systems Research*, vol. 13, no. 3, pp. 334–359, 2002.

[151] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao *et al.*, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[152] F. Tramèr, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert *et al.*, "Fairtest: Discovering unwarranted associations in data-driven applications," in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2017, pp. 401–416.

[153] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. PMLR, 23–24 Feb 2018, pp. 107–118. [Online]. Available: https://proceedings.mlr.press/v81/menon18a.html

[154] P. Biecek and T. Burzykowski, *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. [Online]. Available: https://pbiecek.github.io/ema/

[155] L. Deng and Y. Liu, *Deep Learning in Natural Language Processing*. Springer Singapore, 2018, ch. 1.5.2 Limitations of Current Deep Learning Technology.

[156] L. Deng and Y. Liu, *Deep Learning in Natural Language Processing*. Springer Singapore, 2018, ch. 8.6 Summary.

[157] Y. Wang, L. Lin, and J. Chen, "Communication-efficient adaptive federated learning," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162. PMLR, Jul 2022, pp. 22 802–22 838.

[158] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, "A survey on federated learning: challenges and applications," *International Journal of Machine Learning and Cybernetics*, vol. 14, p. 513–535, 2023.

[159] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Medical Image Analysis*, vol. 71, p. 102062, 2021.

[160] M. Minkkinen, J. Laine, and M. Mäntymäki, "Continuous auditing of artificial intelligence: a conceptualization and assessment of tools and frameworks," *Digital Society*, vol. 1, no. 21, 2022.

[161] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *Journal of Information Security and Applications*, vol. 58, p. 102717, 2021.

[162] BIPM, IEC, IFCC, ILAC, ISO, IUPAC *et al.*, *"Type A evaluation of standard uncertainty"*. Joint Committee for Guides in Metrology, JCGM 100:2008, 2008, ch. 4.2, pp. 10–11. [Online]. Available: https://doi.org/10.59161/JCGM100-2008E

[163] BIPM, IEC, IFCC, ILAC, ISO, IUPAC *et al.*, *"Type B evaluation of standard uncertainty"*. Joint Committee for Guides in Metrology, JCGM 100:2008, 2008, ch. 4.3, pp. 11–14. [Online]. Available: https://doi.org/10.59161/JCGM100-2008E

[164] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009, risk Acceptance and Risk Communication.

[165] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Machine Learning*, vol. 110, p. 457–506, 2021.

[166] L. Breiman, "Random forests," *Machine Learning*, vol. 45, p. 5–32, 2001.

[167] J. Li and M. Lin, "Ensemble learning with diversified base models for fault diagnosis in nuclear power plants," *Annals of Nuclear Energy*, vol. 158, p. 108265, 2021.

[168] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[169] A. Forbes, K. Jagan, J. Donlevy, and J. Alves e Sousa, "Optimization of sensor distribution using gaussian processes," *Measurement: Sensors*, vol. 18, p. 100128, 2021.

[170] Y. Koucha, A. Forbes, and Q. Yang, "A bayesian conformity and risk assessment adapted to a form error model," *Measurement: Sensors*, vol. 18, p. 100330, 2021.

[171] Canada Nuclear Safety Commission, "Maintenance programs for nuclear power plants," Regulatory Document, RD/GD-210, November 2012, online: nuclearsafety.gc.ca.

[172] C. Yung and A. Bonnett, "Repair or replace?" *IEEE Industry Applications Magazine*, vol. 10, no. 5, pp. 48–58, 2004.

[173] N. Jammu and P. Kankar, "A review on prognosis of rolling element bearings," *International Journal of Engineering Science and Technology*, vol. 3, no. 10, 2011.

[174] Rexnord Industries, LLC, Gear Group, *FAILURE ANALYSIS GEARS-SHAFTS-BEARINGS-SEALS*, Rexnord Industries.

[175] S. Kumar, D. Mukherjee, P. Guchhait, M. R. Banerjee, A. K. Srivastava, D. Vishwakarma *et al.*, "A comprehensive review of condition based prognostic maintenance (cbpm) for induction motor," *IEEE Access*, vol. 7, pp. 90 690–90 704, 07 2019.

[176] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," *Mechanical Systems and Signal Processing*, vol. 163, p. 108105, 2022.

[177] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022.

[178] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Christoph Molnar, 2022, ch. Interpretable Models. [Online]. Available: christophm.github.io/interpretable-ml-book/

[179] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifier," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017

[180] H. Tan and H. Kotthaus, "Surrogate model-based explainability methods for point cloud nns," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 2927–2936.

[181] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan *et al.*, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017

[182] L. S. Shapley, *17. A Value for n-Person Games.* Princeton: Princeton University Press, 1953, pp. 307–318.

[183] G. V. D. Broeck, A. Lykov, M. Schleich, and D. Suciu, ""on the tractability of shap explanations"," *Journal of Artificial Intelligence Research*, vol. 74, pp. 851–886, 2022.

[184] O. Amin, B. Brown, B. Stephen, and S. McArthur, "A case-study led investigation of explainable ai (xai) to support deployment of prognostics in the industry," in *PHM Society European Conference*, vol. 7, no. 1, 2022, p. 9–20.

[185] I.-S. F. B. Shukla and I. Jennions, "Opportunities for explainable artificial intelligence in aerospace predictive maintenance," *Proceedings of the European Conference of the PHM Society*, vol. 5, p. 11, 2020.

[186] G. Liu, W. Shen, L. Gao, and A. Kusiak, "Knowledge transfer in fault diagnosis of rotary machines," *IET Collaborative Intelligent Manufacturing*, vol. 4, no. 1, pp. 17–34, 2022.

[187] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[188] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 2, pp. 151–166, 2020.

[189] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning," *Journal of Big Data*, vol. 4, Dec 2017.

[190] B. Maschler and M. Weyrich, "Deep transfer learning for industrial automation: A review and discussion of new techniques for data-driven machine learning," *IEEE Industrial Electronics Magazine*, vol. 15, no. 2, pp. 65–75, 2021.

[191] A. Shamsi, H. Asgharnezhad, S. S. Jokandan, A. Khosravi, P. M. Kebria, D. Nahavandi *et al.*, "An uncertainty-aware transfer learning-based framework for covid-19 diagnosis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1408–1417, 2021.

[192] Q. Xu, S. Ali, T. Yue, and M. Arratibel, "Uncertainty-aware transfer learning to evolve digital twins for industrial elevators," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ser. ESEC/FSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 1257–1268.

[193] G. Colella, V. Lange, and F. Duddec, "Transfer learning for metamodel construction to enable uncertainty quantifications in crash design based on scarce data availability," in *ISMA-USD 2022*. Zenodo, Jan 2023.

[194] J. W. Hines and D. Garvey, "Process and equipment monitoring methodologies applied to sensor calibration monitoring," *Quality and Reliability Engineering International*, vol. 23, no. 1, pp. 123–135, 2007.

[195] H. Harb and A. Makhoul, "Energy-efficient sensor data collection approach for industrial process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 661–672, 2018.

[196] M. Z. A. Bhuiyan, J. Wu, G. Wang, Z. Chen, J. Chen, and T. Wang, "Quality-guaranteed event-sensitive data collection and monitoring in vibration sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 572–583, 2017.

[197] BIPM, IEC, IFCC, ILAC, ISO, IUPAC *et al.*, *"Uncertainty"*. Joint Committee for Guides in Metrology, JCGM 100:2008, 2008, ch. 3.3, pp. 5–7.

[198] Y.-H. Lin and G.-H. Li, "A bayesian deep learning framework for rul prediction incorporating uncertainty quantification and calibration," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7274–7284, 2022.

[199] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf

[200] S. Liu, P. Ram, D. Vijaykeerthy, D. Bouneffouf, G. Bramble, H. Samulowitz *et al.*, "An admm based framework for automl pipeline configuration," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 4892–4899, Apr. 2020.

[201] C. Yang, J. Fan, Z. Wu, and M. Udell, "Automl pipeline selection: Efficiently navigating the combinatorial space," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1446–1456.

[202] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: a survey," *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, 2018.

[203] B. S. P. C. Baker and M. D. Judd, "Compositional modeling of partial discharge pulse spectral characteristics," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 7, pp. 1909–1916, 2013.

[204] M. Z. Ali, M. N. S. K. Shabbir, X. Liang, Y. Zhang, and T. Hu, "Machine learning-based fault diagnosis for single- and multi-faults in induction motors using measured stator currents and vibration signals," *IEEE Transactions on Industry Applications*, vol. 55, no. 3, pp. 2378–2391, 2019.

[205] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," 2020. [Online]. Available: https://arxiv.org/abs/1811.11839

[206] Lee, J. , Qiu, H. , Lin, J. and Rexnord Technical Services, IMS, University of Cincinnati. "Bearing Data Set", NASA Ames Prognostics Data Repository, 2007, http://ti.arc.nasa.gov/project/prognostic-data-repository, NASA Ames Research Center, Moffett Field, CA.

[207] NASA Ames Prognostics Data Repository, "Femto bearing data set," NASA Ames Research Center, Moffett Field, CA, 2012, http://ti.arc.nasa.gov/project/prognostic-data-repository.

[208] R. B. Abernethy, *The New Weibull handbook : reliability and statistical analysis for predicting life, safety, supportability, risk, cost and warranty claims.* Robert B. Abernethy, 2004.

[209] Case Western Reserve University Bearing Data Center, "Seeded fault test data," https://engineering.case.edu/bearingdatacenter, Accessed: Jul. 24, 2024 [Online].

[210] A. Medina-Borja and K. S. Pasupathy, "Uncovering complex relationships in system dynamics modeling: Exploring the use of cart, chaid and sem," in *2007*

International Conference of the System Dynamics Society and 50th Anniversay Celebration, 2007. [Online]. Available: https://proceedings.systemdynamics.org/2007/proceed/

[211] S. K. Gundewar and P. V. Kane, "Bearing fault diagnosis using time segmented fourier synchrosqueezed transform images and convolution neural network," *Measurement*, vol. 203, 2022.

[212] Dr Eric Bechhoefer, Chief Engineer, NRG Systems, "Condition based maintenance fault database for testing of diagnostic and prognostics algorithms," https://www.mfpt.org/fault-data-sets/, data assembled and prepared on behalf of MFPT. Accessed: Jul. 24, 2024 [Online].

[213] W. Liao, R. D. Geest, D. V. Maele, J. C. Poletto, L. P. Selvaraj, T. Ooijevaar *et al.*, "Active learning for gear defect detection in gearboxes," in *Proceedings of the European conference of the PHM Society 2024*, vol. 8, no. 1, 2024, pp. 152–161.

[214] R. Ranjan, S. K. Ghosh, and M. Kumar, "Fault diagnosis of journal bearing in a hydropower plant using wear debris, vibration and temperature analysis: A case study," *Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering*, vol. 234, no. 3, pp. 235–242, 2020.

[215] R. Hughes, T. Haidinger, X. Pei, and C. Vagg, "Real-time temperature prediction of electric machines using machine learning with physically informed features," *Energy and AI*, vol. 14, p. 100288, 2023.

[216] X. Zhou, H. Zhang, X. Hao, X. Liao, and Q. Han, "Investigation on thermal behavior and temperature distribution of bearing inner and outer rings," *Tribology International*, vol. 130, pp. 289–298, 2019.

[217] X. Hao, J. Zhai, J. Liang, Y. Chen, and Q. Han, "Time-varying stiffness characteristics of roller bearing influenced by thermal behavior due to surface frictions and different lubricant oil temperatures," *Tribology International*, vol. 144, p. 106125, 2020.

[218] S. J. Pugh, G. F. Hewitt, and H. Müller-Steinhagen, "Fouling during the use of seawater as coolant - the development of a user guide," *Heat Transfer Engineering*, vol. 26, no. 1, p. 35 – 43, 2005.

[219] H. Hashemian, "Aging management of instrumentation & control sensors in nuclear power plants," *Nuclear Engineering and Design*, vol. 240, no. 11, pp. 3781–3790, 2010.

[220] I. A. E. Agency, ""4.3.1 prognostics" in "4. prognostics and structural material integrity"," in *Advanced Surveillance, Diagnostic and Prognosit Techniques in Monitoring Structures, Systems and Components in Nuclear Power Plants*, vol. NP-T-3.14, Vienna, Austria, 2013, pp. 57–59. [Online]. Available: https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1599_web.pdf

[221] T. Seuaciuc-Osorio, I. Virkkunen, H. Miedl, B. Briquez, H. Abdel-Khalik, C. Lamb *et al.*, ""10.1.4. prediction and prognostics, to better inform maintenance activities" in "chapter 10. nuclear power"," in *ARTIFICIAL INTELLIGENCE FOR ACCELERATING NUCLEAR APPLICATIONS, SCIENCE AND TECHNOLOGY*, Vienna, Austria, 2022, pp. 61–70. [Online]. Available: https://www-pub.iaea.org/MTCD/Publications/PDF/ART-INTweb.pdf

[222] V. Nemani, L. Biggio, X. Huan, Z. Hu, O. Fink, A. Tran *et al.*, "Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial," 2023.

[223] S. Taghiyeh, D. C. Lengacher, A. H. Sadeghi, A. Sahebi-Fakhrabad, and R. B. Handfield, "A novel multi-phase hierarchical forecasting approach with machine learning in supply chain management," *Supply Chain Analytics*, vol. 3, p. 100032, 2023.

[224] R. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. Melbourne, Australia: OTexts, 2018, vol. 2nd Edition, accessed on: 04/03/2024.

[225] F. Tavazza, B. DeCost, and K. Choudhary, "Uncertainty prediction for machine learning models of material properties," *ACS Omega*, vol. 6(48), pp. 32 431–32 440, 2021.

[226] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, vol. 120, pp. 70–83, 2018.

[227] M. Zamo and P. Naveau, "Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts," *Methematical Geosciences*, pp. 209–234, 2018.

[228] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007. [Online]. Available: https://doi.org/10.1198/016214506000001437

[229] M. B. WILK and R. GNANADESIKAN, "Probability plotting methods for the analysis for the analysis of data," *Biometrika*, vol. 55, no. 1, pp. 1–17, 03 1968.

[230] C. Plumley, "Penmanshiel wind farm data," *Zenodo*, 7 Feb 2022.

[231] SKF, "Condition monitoring," date Accessed: 07/03/2024. [Online]. Available: https://www.skf.com/uk/products/mounted-bearings/bearing-housings/split-pillow-block-housings-snl-2-3-5-6-series/condition-monitoring

[232] L. Huang, F. Da, and S. Gai, "Research on multi-camera calibration and point cloud correction method based on three-dimensional calibration object," *Optics and Lasers in Engineering*, vol. 115, pp. 32–41, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0143816618308273

[233] D. Wu, L. Gao, M. Chinazzi, X. Xiong, A. Vespignani, Y.-A. Ma *et al.*, "Quantifying uncertainty in deep spatiotemporal forecasting," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1841–1851. [Online]. Available: https://doi.org/10.1145/3447548.3467325

[234] A. Hilal, S. A. Bangroo, N. A. Kirmani, J. A. Wani, A. Biswas, M. I. Bhat *et al.*, "Chapter 19 - geostatistical modeling—a tool for predictive soil mapping," in *Remote Sensing in Precision Agriculture*, ser. Earth Observation, S. Lamine, P. K. Srivastava, A. Kayad, F. Muñoz-Arriola, and P. C. Pandey, Eds. Academic Press, 2024, pp. 389–418. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323910682000114

[235] Q. Dai, D. Han, and P. Srivastava, "Chapter 5 - radar rainfall sensitivity analysis using multivariate distributed ensemble generator," in *Sensitivity Analysis in Earth Observation Modelling*, G. P. Petropoulos and P. K. Srivastava, Eds. Elsevier, 2017, pp. 91–102. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128030110000057

[236] B. Lin and X. Dong, "Ship hull inspection: A survey," *Ocean Engineering*, vol. 289, p. 116281, 2023.

[237] G. Dobie, R. Summan, C. MacLeod, and S. Gareth Pierce, "Visual odometry and image mosaicing for nde," *NDT & E International*, vol. 57, pp. 17–25, 2013.

[238] S. Choi, H. Cho, and C. J. Lissenden, "Nondestructive inspection of spent nuclear fuel storage canisters using shear horizontal guided waves," *Nuclear Engineering and Technology*, vol. 50, no. 6, pp. 890–898, 2018.

[239] L. Biondi, M. Perry, C. Vlachakis, and A. Hamilton, "Smart cements: repairs and sensors for concrete assets," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2018*, H. Sohn, Ed., vol. 10598, International Society for Optics and Photonics. SPIE, 2018, p. 105982U.

[240] B. T. Bastian, J. N, S. K. Ranjith, and C. Jiji, "Visual inspection and characterization of external corrosion in pipelines using deep neural network," *NDT & E International*, vol. 107, p. 102134, 2019.

[241] S. Ismail, Z. Salleh, M. Yusop, and F. Fakhruradzi, "Monitoring of barnacle growth on the underwater hull of an frp boat using image processing," *Procedia*

Computer Science, vol. 23, pp. 146–151, 2013, 4th International Conference on Computational Systems-Biology and Bioinformatics, CSBio2013.

[242] Y. D. Yasuda, F. A. Cappabianco, L. E. G. Martins, and J. A. Gripp, "Aircraft visual inspection: A systematic literature review," *Computers in Industry*, vol. 141, p. 103695, 2022.

[243] B. Lin and X. Dong, "A multi-task segmentation and classification network for remote ship hull inspection," *Ocean Engineering*, vol. 301, p. 117608, 2024.

[244] A. Shaukat, Y. Gao, J. A. Kuo, B. A. Bowen, and P. E. Mort, "Visual classification of waste material for nuclear decommissioning," *Robotics and Autonomous Systems*, vol. 75, pp. 365–378, 2016.

[245] J. McAlorum, M. Perry, H. Dow, and S. Pennada, "Robotic concrete inspection with illumination-enhancement," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2023*, Z. Su, B. Glisic, and M. P. Limongelli, Eds., vol. 12486, International Society for Optics and Photonics. SPIE, 2023, p. 124860L.

[246] S. Pennada, M. Perry, J. McAlorum, H. Dow, and G. Dobie, "Performance evaluation of an improved deep CNN-based concrete crack detection algorithm," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2023*, Z. Su, B. Glisic, and M. P. Limongelli, Eds., vol. 12486, International Society for Optics and Photonics. SPIE, 2023, p. 1248615.

[247] R. J. Woodham, "Photometric Stereo: A Reflectance Map Technique For Determining Surface Orientation From Image Intensity," in *Image Understanding Systems and Industrial Applications I*, R. Nevatia, Ed., vol. 0155, International Society for Optics and Photonics. SPIE, 1979, pp. 136 – 143.

[248] J. Wu, "Rotation invariant classification of 3d surface texture using photometric stereo," PhD thesis, Heriot-Watt University, Edinburgh, UK, 2003, available at: https://www.macs.hw.ac.uk/texturelab/ files/publications/phds_mscs/JW.

[249] E. DAVIES, "Chapter 16 - the three-dimensional world," in *Machine Vision (Third Edition)*, 3rd ed., ser. Signal Processing and its Applications, E. DAVIES, Ed. Burlington: Morgan Kaufmann, 2005, pp. 445–485.

[250] S. J. Koppal, *Lambertian Reflectance*. Boston, MA: Springer US, 2014, pp. 441–443.

[251] Q. Zheng, A. Kumar, B. Shi, and G. Pan, "Numerical reflectance compensation for non-lambertian photometric stereo," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3177–3191, 2019.

[252] B. Shi, Z. Mo, Z. Wu, D. Duan, S.-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 271–284, 2019.

[253] F. Pernkopf and P. O'Leary, "Image acquisition techniques for automatic visual inspection of metallic surfaces," *NDT & E International*, vol. 36, no. 8, pp. 609–617, 2003.

[254] R. Dessì, C. Mannu, G. Rodriguez, G. Tanda, and M. Vanzi, "Recent improvements in photometric stereo for rock art 3d imaging," *Digital Applications in Archaeology and Cultural Heritage*, vol. 2, no. 2, pp. 132–139, 2015, digital imaging techniques for the study of prehistoric rock art.

[255] J. Sun, M. Smith, L. Smith, L. Coutts, R. Dabis, C. Harland *et al.*, "Reflectance of human skin using colour photometric stereo: with particular application to pigmented lesion analysis," *Skin Research and Technology*, vol. 14, no. 2, pp. 173–179, 2008.

[256] F. A. Saiz, I. Barandiaran, A. Arbelaiz, and M. Graña, "Photometric stereo-based defect detection system for steel components manufacturing using a deep segmentation network," *Sensors*, vol. 22, no. 3, 2022.

[257] S. Huang, K. Xu, M. Li, and M. Wu, "Improved visual inspection through 3d image reconstruction of defects based on the photometric stereo technique," *Sensors*, vol. 19, no. 22, 2019.

[258] S. Moylan, J. Slotwinski, A. Cooke, K. Jurrens, and M. Donmez, "Proposal for a standardized test artifact for additive manufacturing machines and processes," in *Proceedings of the Solid Freeform Fabrication Symposium, Austin, TX*, 2012-08-15 2012.

[259] J. F. Hughes, A. van Dam, M. McGuire, D. F. Sklar, J. D. Foley, S. K. Feiner *et al.*, *Computer Graphics: Principles and Practice*. Addison-Wesley, Jul 2013, ch. 14.5.2 Implicit Surfaces in Chapter 14: Standard Approximations and Representations, pp. 341–342.

[260] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Computer Graphics (SIGGRAPH 1996 Proceedings)*, 1996.

[261] M. S. Floater and M. Reimers, "Meshless parameterization and surface reconstruction," *Computer Aided Geometric Design*, vol. 18, no. 2, pp. 77–92, 2001.

[262] J. F. O'Brien and J. K. Hodgins, "Animating fracture," in *Communications of the ACM*, vol. 43, no. 7, Jul 2000.

[263] M. A. Kowalski, J. N. Lee Markosian, L. Bourdev, R. Barzel, L. S. Holden, and J. F. Hughes, "Art-based rendering of fur, grass and trees," in *Siggraph 1999*, 1999.

[264] Y. Xiong, A. Chakrabarti, R. Basri, S. J. Gortler, D. W. Jacobs, and T. Zickler, "From shading to local shape," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 67–79, 2015.

[265] J. Ren, F. Wang, J. Zhang, Q. Zheng, M. Ren, and B. Shi, "Diligent102: A photometric stereo benchmark dataset with controlled shape and material variation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 571–12 580.

[266] F. Wang, J. Ren, H. Guo, M. Ren, and B. Shi, "Diligent-$\prod$: Photometric stereo for planar surfaces with rich details – benchmark dataset and beyond," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2023, pp. 9443–9453.

[267] J. McAlorum, H. Dow, S. Pennada, M. Perry, and G. Dobie, "Automated concrete crack inspection with directional lighting platform," *IEEE Sensors Letters*, vol. 7, no. 11, pp. 1–4, 2023.

[268] Blender Online Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Blender Institute, Amsterdam, 2022. [Online]. Available: http://www.blender.org

[269] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "MeshLab: an Open-Source Mesh Processing Tool," in *Eurographics Italian Chapter Conference*, V. Scarano, R. D. Chiara, and U. Erra, Eds. The Eurographics Association, 2008.

[270] J. Park, S. N. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon, "Multiview photometric stereo using planar mesh parameterization," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 1161–1168.

[271] K. Luo, Y. Ju, L. Qi, K. Wang, and J. Dong, "Rmaff-psn: A residual multi-scale attention feature fusion photometric stereo network," *Photonics*, vol. 10, no. 5, 2023.

[272] N. Senin, S. Catalucci, M. Moretti, and R. Leach, "Statistical point cloud model to investigate measurement uncertainty in coordinate metrology," *Precision Engineering*, vol. 70, pp. 44–62, 2021.

[273] R. Yang, Y. Wang, S. Liao, and P. Guo, "Dpps: A deep-learning based point-light photometric stereo method for 3d reconstruction of metallic surfaces," *Measurement*, vol. 210, p. 112543, 2023.

[274] C. Pottier, J. Petzing, F. Eghtedari, N. Lohse, and P. Kinnell, "Developing digital twins of multi-camera metrology systems in blender," *Measurement Science and Technology*, vol. 34, no. 7, p. 075001, mar 2023.

[275] H. Theil, "A rank-invariant method of linear and polynomial regression analysis," *Proceedings of the Royal Netherlands Academy of Sciences*, vol. 53, pp. Part I: 386–392, Part II: 521–525, Part III: 1397–1412, 1950.

[276] P. K. Sen, "Estimates of the regression coefficient based on kendall's tau," *Journal of the American Statistical Association*, vol. 63, no. 324, pp. 1379–1389, 1968.

[277] X. Dang, H. Peng, X. Wang, and H. Zhang, "Theil-sen estimators in a multiple linear regression model," 2008, the University of Mississippi at olemiss.edu.

[278] T. Kärkkäinen and S. Äyrämö, "On computation of spatial median for robust data mining," in *EUROGEN 2005: Evolutionary and Deterministic Methods for Design, Optimization and Control with Applications to Industrial and Societal Problems*, Oct 2005.

[279] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer New York, NY, 2006, ch. 3.3 Bayesian Linear Regression.

[280] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Jun 2001. [Online]. Available: https://www.jmlr.org/papers/v1/

[281] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 05 1992.

[282] A. Genz, "Numerical computation of multivariate normal probabilities," *Journal of Computational and Graphical Statistics*, vol. 1, no. 2, pp. 141–149, 1992.

[283] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 1994.

[284] M. Vilela, N. Halidi, S. Besson, H. Elliott, K. Hahn, J. Tytell *et al.*, "Chapter nine - fluctuation analysis of activity biosensor images for the study of information flow in signaling pathways," in *Fluorescence Fluctuation Spectroscopy (FFS), Part B*, ser. Methods in Enzymology, S. Y. Tetin, Ed. Academic Press, 2013, vol. 519, pp. 253–276. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124055391000099