Reliability analysis of a CRAC system under uncertainties

By

Ahsan Javed

(Registration No: 00000335017)

Supervisor: Dr. Tayyab Zafar

Department of Mechatronics Engineering

College of Electrical & Mechanical Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis by Ahsan Javed (Registration No. 00000330517) of Electrical and Mechanical Engineering College has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature of Supervisor: _____

Dr. Tayyab Zafar

Dated: 26-July-24

Signature of HOD: _____

Dr. Hamid Jabbar

Date: 26-07-2024

Signature of Dean: _____

Brig Dr. Nasir Rashid

Date: 26 JUL 2024

DEDICATION

This thesis is dedicated to Dr. Tayyab Zafar for his invaluable guidance and mentorship, to my mother for her unwavering support and encouragement, and to my wife for her endless patience, understanding, and motivation throughout this journey. Thank you all for your immeasurable contributions to my life and work.

ACKNOWLEDGEMENTS

I extend my deepest gratitude to my supervisor Dr. Tayyab Zafar for his exceptional insight and scholarly guidance throughout the course of completing my thesis. His expertise, dedication, and unwavering support have been instrumental in shaping both the direction of my research and my personal growth as a scholar.

I am also thankful to my Co-Supervisor Dr. Muhammad Mubashir Saleem for his understanding, core expertise, and constructive criticism, which proved crucial in navigating through this research endeavor.

I am equally grateful to my Guidance and Evaluation Committee Member Dr. Anas Bin Aqeel for his continuous feedback, encouragement, and guidelines, which motivated and propelled me forward.

Special thanks are also due to Dr. Ameer Hamza for the guidelines and supervision provided, which served as a guiding light throughout this thesis journey.

ABSTRACT

The rapid increase in data center size and number, driven by escalating internet and cloud computing demands, has led to high energy consumption and public concern. Densely packed high-powered systems within data centers generate significant radiant heat, necessitating effective and Reliable cooling solutions for maintaining uptime. This research presents a novel method to evaluate Computer Room Air Conditioning (CRAC) system performance and efficiency. Firstly, a rack-level heat transfer probabilistic constraint is introduced, integrating environmental conditions such as ambient temperature, humidity, and airflow patterns, which significantly impact heat transfer processes and are accurately incorporated to reflect real-world scenarios. Additionally, the model accounts for specific configurations and thermal properties of data server racks, enabling precise simulation of heat generation and dissipation patterns. The probabilistic variables undergone training including the layout of servers, types of cooling mechanisms employed, and the material properties of the racks. Secondly, modelling the CRAC system's heat transfer rate as random distribution facilitates effective thermal load management and balances computational demands with accuracy. Based on the output from two probabilistic performance functions, a multi-response Gaussian process (AMRGP) model is developed using an adaptive sampling technique, enhancing predictive accuracy and efficiency by training the predicted responses with a learning U-function to calculate the probability of failure and reliability of the model. The proposed method also improves risk assessment by predicting the likelihood of failure events, aiding in the development of a powerful tool for designing and evaluating CRAC system reliability in complex and uncertain environments. This research thus represents a significant advancement in the field of data center engineering, providing a robust framework for future development in thermal management and reliability assessment.

Keywords: Data center, CRAC system, Radiation heat transfer, MCS, Reliability predictions, Adaptive sampling, Thermal management tools.

TABLE OF CONTENTS

LIST OF TABLES

Page No.

LIST OF FIGURES

Page No.

CHAPTER 1: INTRODUCTION

1.1 Background and Motivation

Data centers have become the nerve centers of modern digital operations, playing a pivotal role in storing, processing, and delivering vast amounts of data essential for businesses, organizations, and individuals. Within these facilities, the efficient management of temperature and humidity levels is vital to ensure the reliable operation of critical equipment. Computer Room Air Conditioning (CRAC) systems emerge as key components in maintaining optimal environmental conditions within data centers. These systems are specifically designed to regulate temperature and humidity levels, effectively dissipating the heat generated by densely packed servers and networking equipment.

The background of the CRAC system stems from the escalating demands of data center operations, with a surge in data traffic from 2010 to 2018 due to the rise of cloud-based services and applications provided by major commercial cloud services providers such as Google, Facebook, and Amazon [1]. With rising concerns about energy consumption and environmental sustainability, there has been a growing emphasis on making data center operations more energy-efficient and reliable. CRAC systems, which typically account for a significant portion of energy usage within data centers, have been a focal point for energy efficiency initiatives. Manufacturers have introduced more energy-efficient designs and technologies, such as variable-speed compressors and economizer modes, to reduce energy consumption and improve the coefficient of performance.

The reliability analysis of CRAC systems in data centers is driven by the critical importance of cooling systems in maintaining the reliability, availability, and energy efficiency of data center infrastructure. According to [2], by conducting quantifiable reliability analysis on component's inherent characteristics, data center operators can identify vulnerabilities, optimize performance, and implement proactive measures to ensure continuous and uninterrupted operations, thereby guaranteeing the delivery of reliable and robust data center services. The main motivation of this study comes from the

need for the reliability analysis of data center's spatial data, intensive radiation heat uncertainty quantifications, and surrogate modelling using the Kriging method. A classification derived from the synopsis of data center energy consumption and reliability modelling is offered in Figure 1.

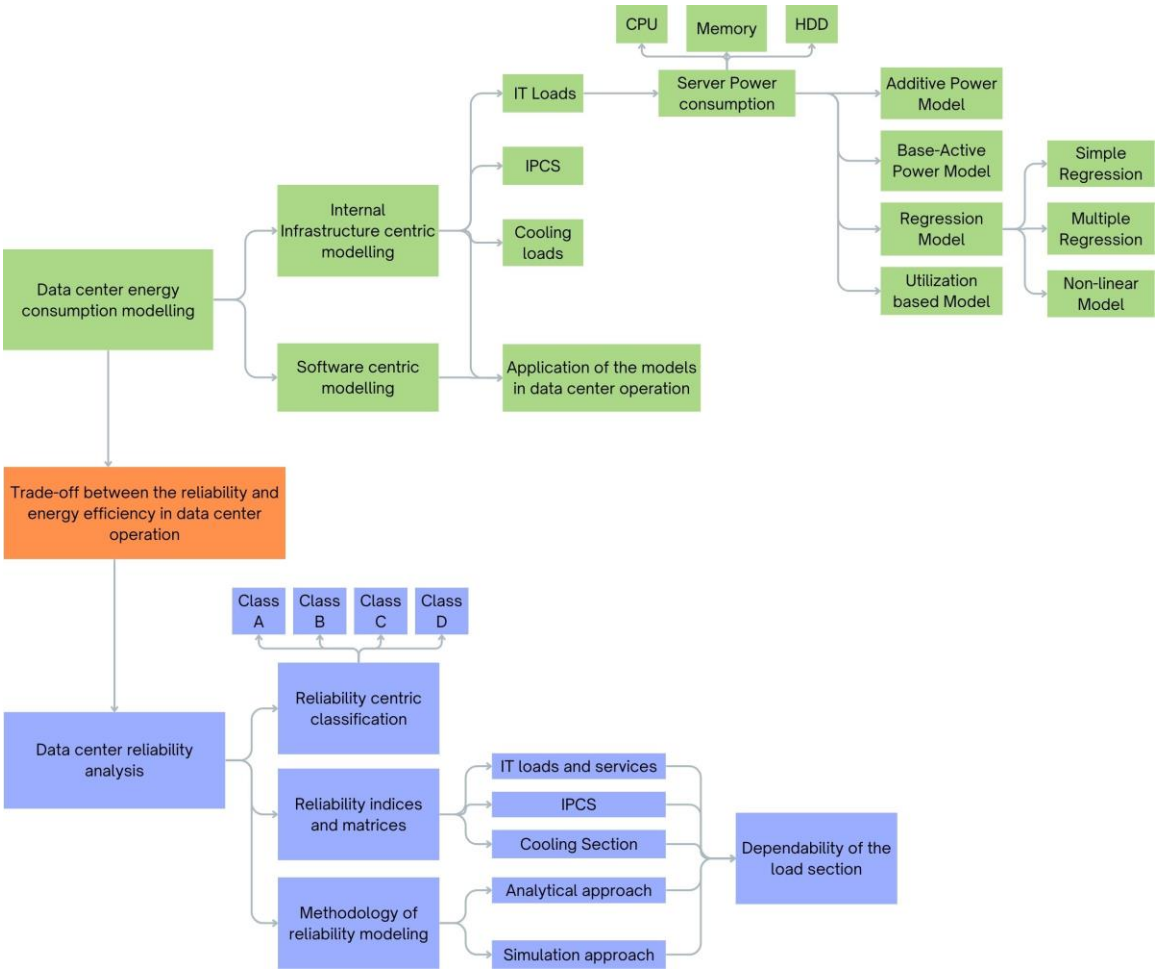

Figure 1.1: Spectrum of data center reliability analysis and energy consumption modelling

1.2 Environmental Classes of Data Centers

The CRAC system of data centers serves a multifaceted role critical to their efficient operations. Its foremost importance lies in temperature regulations, as the CRAC system diligently maintains optimal temperatures within the data center. By effectively managing heat dissipated from servers and networking equipment, it prevents overheating, thus

protecting from hardware failures and costly downtime. As data centers become more sophisticated and capable of handling big data, there arises a critical need for a standardized framework to categorize varying levels of data centers capacities. The American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) has risen to this challenge by introducing ASHRAE 2008, and 2011 standards, which were further restructured into the 2015 ASHRAE environmental classes for data center applications [3]. These classes provide guidelines for CRAC system design and implementation to enterprise engineers. This section discusses three data center environmental classes, each representing allowable environmental ranges and thresholds for critical equipment and the density of the data center.

ENVIRONMENTAL CLASSES OF DATA CENTER

2015 CLASSES	APPLICATIONS	INFORMATION TECHNOLOGY EQUIPMENT	ENVIRONMENTAL CONTROL	RECOMMNEDED TEMPEREATURE (°C) & HUMIDITY RANGES (%RH)
A1	Data Center	Enterprise Server Storage Products	Tightly Controlled	15°C to 32°C & 60% RH
A2 A3 A4	Data Center	Volume Servers, Storage Products, Super Computers, Workstations, SCADA Servers	Optimum Control for Class A2 Some Control, Use of free Cooling Technique for Class A3 Some, Control, Near full-time usage of free cooling for Class A4	10°C to 35°C & 8% to 80% RH for Class A2 5°C to 40°C & 8% to 85% RH for Class A3 5°C to 45°C & 8% to 90% RH for Class A4
B	Office, Home, Transportable Environment, Etc.	Workstations, Printers, Mobile Data Server Modules, Personal Computers	Minimal Control	5°C to 35°C & 8% to 80% RH
C	Point-of-sale, Industrial, Factory, etc.	Point-of-sale equipment, Ruggedized Controllers, or computer and PDAs	No Control	5°C to 40°C & 8% to 80% RH

Figure 1.2: Environmental classifications of data center infrastructure

1.2.1 Class A: Tightly Controlled Environments

Class A represents mission-critical data centers where maintaining tight temperature and humidity tolerances is imperative for optimal equipment performance and reliability. It allows only $2.4\,min/year$ downtime, with fully redundant equipment commissioning. The recommended temperature range is $64.4°F$ to $80.6°F$ with dew point

and humidity ranges of $10.4°F$ DP & 20% RH to $59°F$ DP and 80% RH. Class A environments comprise of enterprise servers, storage products, and volume servers. These data centers are equipped with advanced cooling technologies, monitoring mechanisms, and robust equipment to withstand continuous heat flux and changing ambient conditions.

1.2.2 Class B: Controlled Environments

The data centers pertaining to Class B have a wider operating range while still maintaining adequate environmental control. They offer $1.6\ h/year$ downtime with $72\ h$ power failure safety and temperature ranges from $41°F$ to $95°F$ with humidity range of 8% to 80% RH. This class is suitable for data centers where energy efficiency is a priority, as the broader temperature range allows for more flexible cooling strategies without compromising equipment reliability.

1.2.3 Class C: No-Control Environments

Class C data centers expand the permissible temperature ranges further, making them suitable for environments where energy efficiency takes precedence over precise environmental control. They offer $28.8\ h/year$ downtime with partial to no redundancy, and the permissible ambient temperature range is $41°F$ to $104°F$ with humidity ranging from 8% to 80% RH. These data centers can tolerate wider temperature variations, allowing operators to implement economization techniques such as air-side or water-side economizers to reduce cooling costs while maintaining equipment reliability.

1.3 Types of CRAC Systems

Heat removal by the CRAC system in a data center can be conceptualized as the process of transferring heat energy from the data center environment to the outdoors. This transfer can be as straightforward as using an air duct to carry heat energy to an outdoor cooling system. Typically, heat exchanger equipment is commissioned in CRAC systems to facilitate this transfer by moving heat energy from one fluid to another, such as from air to water. Table 1.1 presents seven common CRAC types implemented worldwide with

specific indoor and outdoor heat transfer mechanisms. The fluid medium (liquid or gas) is responsible for transporting the heat energy from the data center.

Table 1.1: Comparison of heat transfer methods of CRAC systems

S. No.	Indoor Heat Exchanger	Transport fluid	Outdoor Heat Exchanger	Advantages	Disadvantages
1	CRAH	Chilled water	Chiller	Low Running Cost. Fewer parts. High heat removal capacity.	High capital cost. Poses threat to IT Equipment.
2	Pumped refrigerant System	Refrigerant	Chiller	High efficiency. Non-conductivity of fluids. Offers chip-level cooling.	High build-up cost. Implementable for closed coupled systems. Required additional pumps and heat exchangers.
3	Air-cooled CRAC	Refrigerant	Condenser	Low overall execution cost. Easy operation and maintenance.	Complicated Piping system. Restriction of long distance in refrigerant piping. Multiple CRAC systems cannot be coupled.
4	Glycol-cooled CRAC	Glycol	Dry cooler	One single factory-sealed equipment housing. Offers long-distance piping network.	Requires additional equipment (pump package, valves). More Capital and installation costs.

				Glycol offers free cooling as an economizer mode.	The volume of refrigerant to be maintained regularly.
5	Water-cooled CRAC	Condenser water	Cooling tower	Low running cost. Highly reliable.	High execution and maintenance costs.
6	Air-cooled self-contained system	Air	Air Duct/Exhaust louvers	Lowest installation costs. High energy savings Suitable for moderate availability requirements.	Less heat removal capacity. Requires ductwork or dropped ceiling in an indoor environment. Demands frequent filter changes
7	Air duct system	Air	Evaporative cooler/RTU	Less space requirements Significant cooling energy savings in mild seasons	Low reliability of data center equipment. Difficult to implement in existing infrastructure.

The fundamentals of each type of CRAC system are as follows:

1.3.1 Computer Room Air Handler (CRAH) System

Computer room air handler (CRAH) systems are also known as chilled water systems, the CRAH integrates the refrigeration components of the CRAC unit with a water chiller. The chiller produces water chiller to around $46 - 59°F$, which is pumped to the CRAC units in the data center. The CRAH system returns warm air to the chiller, which transfers the heat to another stream called condenser water, circulating through a cooling tower. The cooling tower works as a radiator, lowering the water temperature. Key advantages include lower cost, fewer parts, greater heat removal capacity, and enhanced efficiency. However, they require higher capital costs for smaller installations and continuous make-up water. These systems are common in large buildings with data centers

needing moderate-to-high availability, especially for larger installations exceeding $200kW$.

1.3.2 Pumped Refrigerant Heat Exchanger System

This system, also known as a pumper refrigerant system, uses a heat exchanger and pump to isolate the cooling medium from the chilled water, employing refrigerants like $R-134A$ or non-conductive fluids like fluorinert, which are circulated through the system without the need for a compressor. Chilled water from the chiller transfers heat from the pumped refrigerant, which then returns to the cooling unit. The main advantages are preventing water damage to equipment, using oil-less and non-conductive refrigerant, and enhancing efficiency and proximity to servers or direct-to-chip level cooling.

1.3.3 Air-cooled CRAC System

The combination of an air-cooled CRAC unit with a condenser is commonly referred to as an air-cooled CRAC DX system. This system uses refrigerant and a semi-hermetic compressor with a matching evaporative coil, extensively utilized in data centers of various sizes, especially small and medium spaces. The compressor is usually within the CRAC unit, transferring heat from the data centers to outdoor surroundings. Advantages include lower costs and simplified operation and maintenance, although long-distance refrigerant piping is not viable, impacting reliability. Common in wiring closets, computer rooms, and data centers ranging from $7\ to\ 200KW$ with moderate availability requirements.

1.3.4 Glycol-cooled CRAC System

A glycol-cooled CRAC system paired with a dry cooler constitutes what's commonly termed a glycol-cooled system using glycol to extract and transport heat. Glycol pipes are smaller and can cover longer distances, facilitating the servicing of multiple CRAC units. An economizer coil can deactivate the refrigeration cycle, achieving free cooling and reducing operating costs. Regular maintenance is needed to ensure proper

glycol volume and quantity. This system offers reliability with factory-sealed units and is cost-effective for larger data centers.

1.3.5 Water-cooled CRAC System

When a water-cooled CRAC is coupled with a cooling tower, it forms what's commonly known as a water-cooled system. These systems use a water loop called condenser water to gather and transfer heat from the data center. Heat dissipation occurs outdoors via a cooling tower. It is cost-effective for leased data center environments but requires a high initial investment for the cooling tower, pump, and piping system. Reliability concerns arise with non-dedicated cooling towers compared to dedicated ones for CRAC units.

1.3.6 Air-cooled Self-contained System

Indoor self-contained systems are typically limited in capacity (up to $15kW$) due to the space needed for all components and large air ducts. While larger capacities are possible for outdoor systems, they are uncommon for precision cooling. These systems offer the lowest installation cost and are used in wiring closets and computer rooms with moderate availability requirements, occasionally addressing hot spots in the data center.

1.3.7 Direct Fresh Air Evaporative Cooling system

This system combines an air duct with a direct fresh air evaporative cooler. It uses fans and louvers to draw cold outdoor air through filters directly into the data center, regulated by louvers and dampers to maintain environmental set points. The primary cooling mode is economizer or free cooling, with containerized DX air-cooled systems as a backup. Despite filtration, fine particulates like smoke and gases may enter the data center.

There are several types of Computer Room Air Conditioning (CRAC) systems, that plays a crucial role in maintaining the environmental stability of data centers. For this research, DX-type air-cooled CRAC systems have been selected due to their prevalent use

in small to medium-sized data centers and their simple design, lower initial costs, and ease of installation. This simplicity, however, comes at the cost of lower energy efficiency and higher running costs compared to water-cooled systems, particularly in larger installations. To enhance the reliability and efficiency of these systems, metamodelling techniques are employed. By reducing computational demands, facilitating sensitivity analysis, and quantifying uncertainties, metamodelling supports the identification of critical factors affecting CRAC system performance and helps in making informed decisions to enhance reliability. The integration of metamodelling into the design and operation of the CRAC system, provides a robust framework for predicting failures, assessing the risk, and refining the system parameters, thereby ensuring the continuous and efficient operation of data centers.

1.4 Metamodelling for Reliability Analysis

In the domain of Reliability analysis, both analytical and simulation-based methods are integral for comprehensively assessing the performance and efficacy of critical engineering systems. Analytical reliability analysis methods, such as Failure Mode and Effects Analysis (FMEA) and Fault Tree Analysis (FTA), offer systematic approaches to identifying potential failure modes and evaluating their impact on system reliability. However, these methods may be limited by their reliance on static assumptions, potentially overlooking the dynamic and stochastic nature of real-world operating conditions. To supplement these analytical techniques, simulation-based methods, such as Monte Carlo Simulation (MCS), are employed to model the probabilistic behavior of systems across diverse operational scenarios. MCS entails running numerous simulations with randomly generated input variables to estimate the distribution of possible outcomes, effectively incorporating uncertainty and variability in system parameters for a more comprehensive and realistic assessment of reliability and accuracy in performance.

There are various metamodelling techniques employed to approximate the behavior of the probabilistic performance function $g(X)$ of the engineering systems. The basic block diagram of a metamodel reflected as a grey box built on input design variable (X) with associated output $g(X)$ is presented in **Error! Reference source not found.**. This section

outlines the advanced metamodel developed by the researchers and their application within the reliability analysis domain [4].

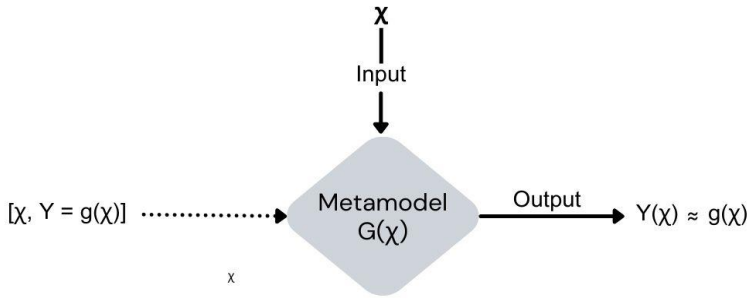

Figure 1.3: A basic functional diagram of interpretable metamodel

1.4.1 Monte Carlo Simulation (MCS)

Monte Carlo Simulations (MCS) build on input variables (x), which are system parameters, while $g(x)$ quantifies the system response. Through random sampling of input variables and evaluation of performance function, MCS determines the probability of failure by a predefined threshold, offering insights into system reliability.

1.4.2 Polynomial Chaos Expansion

Polynomial chaos expansions (PCE) formulate the performance function $g(x)$ using orthogonal multivariate basis functions with respect to joint probability density functions f_x of input variables (x). PCE is typically implanted in its non-intrusive form, where (x) is represented in the standard normal space via variable transformation. However, the efficiency of the PCE method is a trade-off between the dimensionality of input data (x), as larger dimensions lead to an exponential increase in the required experimental design size and costs.

1.4.3 Kriging or Gaussian Process Models

Kriging models, also known as Gaussian process models, interpolate $g(x)$ as a Gaussian process indexed by input random variables, with the design variables acting as

training points. Kriging models inherently account for uncertainty and refinement based on the available data. The application of Kriging entails selecting a correlation function, which is considered stationary and in Gaussian form with a polynomial basis. The prediction in the kriging method depends on hypermeters such as α, σ^2, and a correlation factor $\mathcal{R}(x;\theta)$. Where θ, is a hypermeter trained with a common kernel for each dimension of the design point (x) to enhance the accuracy of the model.

1.5 Scope of the Research Work

The dependability of the CRAC systems in data centers on load sections has led to the exploration and development of a novel algorithm for reliability analysis of the complete data center system. The adaptive Multiple Response Gaussian Process (AMRGP) model, designed to handle multi-dimensional output variables as an n-dimensional Gaussian process with a separable covariance function, will be deployed. The research involves generating probabilistic performance functions by integrating radiation heat transfer theory and the sensible heat transfer rate of the CRAC system. A metamodel based on an adaptive sampling method is produced, serving as a spatial representation of the complex and mission-critical model of a data center. The hyperparameters of the metamodel are tuned to predict the response and estimate the joint probability of failure of two highly uncertain subsystems within a data center: the data server rack and the Computer Room Air Conditioning (CRAC) system. The overall accumulative reliability and accuracy of probabilistic constraints are calculated.

1.6 Research Objectives

1. Implementation of a practical engineering problem with uncertainties and development of multiple limit state functions for reliability analysis.
2. Development of a time-independent algorithm using the AMRGP model with adaptive sampling to reduce the computational costs by achieving the predefined targets, which are specified as percentage reduction with absolute values.

3. To identify the effectiveness of the proposed approach and to train the data center model for reliability assessment of already commissioned data centers and their associated CRAC systems.

The thesis is further organized as the literature review which examines the existing body of knowledge, exploring relevant theories, methodologies, and findings in the field. The methodology section outlines the research design, including data collection methods, analytical frameworks, and simulation techniques employed. Subsequently, the results chapter presents the findings of the study, followed by a discussion section that interprets and contextualizes the results within the broader research context. Finally, the conclusion chapter summarizes the key findings, discusses their implications, and suggests avenues for future research.

CHAPTER 2: LITERATURE REVIEW

The inception of reliability analysis in engineering marks a crucial shift towards evaluating system dependability and robustness, driven by the imperative to ensure reliable system function amidst diverse conditions and potential consequences like failure, safety hazards, and financial losses. Reliability analysis involves systematic exploration of methodologies to assess failure likelihood and system performance, including identifying failure modes, quantifying probabilities, and evaluating metrics such as availability, reliability, and maintainability.

2.1: Classical Reliability Analysis Techniques

The need for reliability analysis was incepted by researchers at the end of the 19th century by categorizing its implementation into three broader areas [5];

1. The exploration of system-level reliability analysis gained attention, particularly driven by the need for rigorous treatment of safety aspects in complex systems like nuclear power plants. The researchers highlighted the importance of understanding how various components interact to influence overall system reliability.
2. With the increasing integration of software into various systems, there was a notable rise in the focus on software reliability and functionality of modern systems, prompting efforts made to improve software reliability through systematic testing and refinement processes.
3. Despite the vital role of reliability in ensuring product quality and customer satisfaction, there was a noticeable lack of interest among managers in implementing reliability programs. This indifference towards reliability initiatives prompted the development of incentives to enhance reliability and highlight its significance in achieving long-term success and competitiveness.

Moreover, the reliability analysis method based on mathematical equations was first introduced by [6]. The method consists of calculating different performance metrics

of a repairable component of a system having monotonic behaviour. The main focus was to determine the availability point, the pattern of failures within a specific time frame, and the duration of system downtime. [7],[8] introduces a comprehensive perspective on system failure engineering in which failure was linked with the system's states and was classified in a binary manner based on coherent structural function. However, the representation and modelling of the systems along with its quantification posed the biggest challenge for the researchers. Additionally, a persistent struggle continues to adequately represent, propagate, and quantify the inherited uncertainties in engineering systems. These classical methods were extensively used to evaluate and analyze various engineering components of the systems, but the accuracy in the representation and modelling of multi-state dynamics persists as a challenging task.

2.2: Analytical Methods

Analytical methods for reliability analysis involve mathematical and statistical techniques to assess the reliability of systems, components, or processes. These methods aim to quantify the probability of failure or success, identify potential failure modes, and evaluate the system's performance under various uncertainties [9]. However, analytical methods can be time-consuming for complex systems. Some common analytical methods for reliability analysis developed by researchers include:

Probabilistic modelling involves constructing mathematical models that represent the reliability characteristics of the system or its components. These models often utilize probability distributions to present the likelihood of failure or success over time. Probabilistic models are implemented on many systems by using techniques such as Markov models [10], stochastic processes [11], and Bayesian analysis [12].

The Fault tree analysis method, employed by [13], is a deductive approach used to analyze and visualize the causes of system failures. It starts with a top-level event (system failure) and back propagates through a logical diagram of events and conditions that could eventually lead to the top event. FTA helps identify critical failure paths and assess their probabilities but it is time-consuming and more prone to error.

Failure mode and effects analysis (FMEA) offers a systematic approach to identify potential failure modes and evaluate their effects within a system. [14] & [15] presented a detailed overview of FMEA and its limitations in weighing the risk factors and ranking the failure modes of engineering systems. It involves investigating the severity, occurrence probability, and detectability of each failure mode to prioritize them for mitigation.

A reliability block diagram (RBD) is a graphical method used to model the reliability of complex systems by breaking them down into individual components or blocks and representing their correlations. Unlike previous methods of modelling and constructing fault trees, the RBD method eliminates the need for identifying minimal path sets. However, model mapping is still an open area for research [16]. The unstable and indeterminate performance of analytical methods for multimodal and complex engineering systems has led to the introduction of sampling-based methods.

2.3: Sampling-based Methods

Sampling-based methods of reliability analysis involve using statistical sampling techniques to estimate reliability metrics, such as failure probabilities or system performance when analytical solutions are impractical. These methods are particularly useful for complex systems or situations where explicit mathematical models are difficult to construct. [17] developed a method using sampling-based techniques to perform uncertainty and sensitivity analysis, and the computational costs were optimized using a variance reduction method.

Monte Carlo simulation (MCS) involves generating random samples from probability distributions that represent uncertain input parameters, such as component failure rates or environmental conditions. The performance, including convergence and computational efficiency, of the MCS method was investigated by [18], for series systems with various configurations. Significant reductions in computational cost were achieved using the proposed methodology.

Latin Hypercube Sampling (LHS) is a sampling technique that offers superior efficiency over Monte Carlo simulation by ensuring a more uniform coverage of input

parametric space. It divides each input parameter range into equally probable intervals and samples from each interval once, which reduces the number of samples needed to achieve accurate results [19].

Importance Sampling (IS) is a variance reduction technique that improves the efficiency of a model by biasing the sampling towards regions of the input parametric space that contribute most to the output uncertainty [20].

A significant apprehension in analytical and sampling-based methods revolves around certifying the accuracy of the response surface and the effectiveness of the sampling procedure associated with it. The ultimate goal is to accurately determine the true combination of unknown parameters. The accuracy of the traditional reliability analysis methods is often assessed using a small subset of testing samples in many applications due to the high computational costs of evaluating the entire parametric space [21]. This challenge is further compounded when dealing with high-dimensional unknown model parameters, leading to complexity and error. To address this issue, the Gaussian Process Regression model (GPR) has been established by [22].

2.4: Probabilistic Gaussian Process Regression(GPR) Reliability Modelling

Probabilistic modelling offers the characterization of uncertainties associated with various factors that influence the performance of engineering systems. These uncertainties include material properties, load conditions, environmental conditions, manufacturing variations, and operational parameters. In probabilistic modelling, the uncertainties are represented using probability distributions, such as uniform, normal, and lognormal distributions, to reflect the variability and randomness in the system's behaviour.

In recent years, the Kriging method [23], a spatial interpolation technique, has gained popularity as a surrogate modelling method due to its ability to formulate spatial correlations and handling of high-dimensional data. However, it poses the limitation of modelling a single probabilistic constraint for reliability analysis of engineering systems. To address this problem, Multiple Response Gaussian Process (AMRGP) modelling is offered [24]. AMRGP is a powerful approach for reliability analysis that can handle

multiple correlated responses of a complete engineering system simultaneously. It offers the ability to model complex relationships between response functions, making it suitable for applications where traditional single-response models are inadequate.

Table 2.1: Literature Review of Methods of Reliability Analysis

Classification	Classes	Published Works	Advantages	Limitations
Analytical Methods [9]	Classical Math-based	[5-8]	Provides precise analytical solutions for simple systems.	May not adequately capture complexities the of real-world systems.
	Probabilistic Modelling	[10-12]	Incorporates uncertainty in system parameters, providing a more realistic representation of system behavior.	Requires extensive data and assumptions about probability distributions, which may not always be available.
	Fault Tree Analysis (FTA)	[13]	Offers a structured approach to identifying system failure modes and their causes.	Can become complex and time-consuming for large, interconnected systems.
	Failure Mode and Effect Analysis (FMEA)	[14, 15]	Systematically evaluates potential failure modes and their effects on system performance.	Relies heavily on expert judgment and may not capture all failure modes or their interactions.
	Reliability Block Diagram (RBD)	[16]	Simplifies system representation, facilitating analysis of complex systems.	May oversimplify system interactions, potentially overlooking important failure modes or correlations.
Sampling-based Methods [17]	Monte Carlo Simulations (MCS)	[18]	Provides probabilistic assessment of system reliability,	Can be computationally intensive, especially for complex systems

			accounting for uncertainties and variability.	with numerous parameters.
	Latin Hypercube Sampling (LHS)	[19]	Efficiently samples across the entire parameter space, reducing the number of simulations needed.	Requires careful consideration of the number of samples and distribution of parameters.
	Importance Sampling (IS)	[20]	Focuses simulations on regions of interest, improving efficiency for rare event analysis.	Selection of Importance functions and tuning parameters can be challenging.
GPR Probabilistic Reliability Modelling [22]	Kriging	[23]	Provides a surrogate model to approximate complex system responses, reducing computational burden.	Requires careful selection of correlation function and tuning of hyperparameters.
	Multiple Response Gaussian Process (AMRGP)	[24]	Handles multi-dimensional output variables and provides flexibility in modelling complex system responses.	May require significant computational resources for model training and optimization, especially for large datasets.

2.5: Reliability Analysis of CRAC Systems

The CRAC system within a data center contributes to roughly 40% of its total power usage. Data server equipment produces heat through Joule heating, necessitating effective thermal management to avoid undesirable temperatures. This heat generation is typically quantified in watts, with the power consumed by data servers being largely converted into heat. Hence, the data server's thermal output is directly proportional to its power consumption [25]. Moreover, data server equipment is designed to operate within specific temperature ranges to maintain optimal performance and high reliability.

Temperature fluctuations beyond these ranges can significantly increase the failure rates of semiconductor devices within data servers and telecommunication equipment devices, with failure rates doubling for every $10°C$ rise [26].

The occurrence of uncertainties during operations and variations in environmental temperatures can result in malfunctions and breakdowns within the data center and its systems. To mitigate these uncertainties, it is essential to conduct a detailed reliability analysis during the design phase, well before the construction and deployment of data centers. Researchers have introduced various methodologies and techniques to evaluate the reliability of CRAC systems in data centers.

A state-of-the-art Power usage effectiveness (PUE) technique was developed by [27], introducing the Evaporative Cooling Composite |Air Conditioning system (ECCAC). This system integrates the indirect heat transfer method with vapor compression refrigeration techniques to deliver year-round cooling for data centers. Additionally, it utilizes an evaporative condenser to maximize the utilization of natural cooling sources. A refrigerant circuit serves as the cooling carrier instead of air or water, thereby reducing cooling losses. However, achieving a robust coefficient of performance (COP) in the proposed system involves balancing efficiency against cost considerations.

The power consumption modelling technique analyzes the power consumption and losses of data server racks to efficiently design the CRAC system. The impact of power losses, as the number of servers increases, is evaluated by analyzing the percentage of power loss by [28]. The reliability block diagram (RBD) technique is utilized to measure the availability of critical equipment, employing mean-time-to-fail (MTTF) and mean-time-to-repair (MTTR) metrics.

The strategic design selections and control measures in optimizing airflow uniformity, temperature distributions, and overall thermal management efficiency are of paramount importance in data center environments. [29] conducted comprehensive numerical and experimental investigations utilizing a scaled physical model to explore the impact of various factors on data center cooling. These factors include power density, floor tile opening ratio, the lateral spacing between CRAC units, and cold aisles of data server

racks. The study focused on evaluating thermal management performance parameters such as SHI/RHI and Return Temperature Indices (RTI). However, the air balancing of the CRAC system with the complete data center environment remains a challenge.

SCADA is an emerging approach that involves the implementation of two proprietary control strategies: supervisory control and data acquisition (SCADA) and ON/OFF, within a hybrid evaporative and direct expansion CRAC DX model for data center cooling systems. These control schemes can operate independently or synergistically, resulting in energy conservation through methods like airside economization and evaporative cooling [30]. The main objective of the research is to assess how proprietary controls influence the reliability and energy efficiency of the data center. Integrated system-level controls (SCADA) achieve recommended temperature and humidity conditions at all times, with a low cooling power usage effectiveness (PUE) of 1.13 compared to 3.76 with no controls.

Hierarchical modelling techniques investigate reliability and availability metrics and conduct parametric sensitivity analysis of the CRAC system of the data center. The technique revolves around modelling reliability block diagrams (RBD) and utilizing parametric sensitivity analysis on each CRAC component to gauge the system's responsiveness to component failures and repair durations [31]. This technique can segregate the critical components of the CRAC system and those with negligible impact on availability to uphold operational Reliability.

Reliability and availability analysis (RAA) identifies risks and devises potential solutions for CRAC system design and operation. The CRAC systems with hybrid cooling techniques, identical to water-side economizers, require frequent reliability and availability analysis. This is due to its complex configuration and multiple operational modes [32]. The model comprises a Markov chain in conjunction with the reliability block diagram (RBD) method to evaluate the reliability and operational availability of the CRAC system through spatial temperature gradient data.

2.6: Summary of Literature Review

From the review of available literature, it is evident that reliability analysis ensures system dependability and robustness by assessing failure likelihood and performance through classical techniques, analytical methods, and sampling-based methods. Whereas, prevailing reliability analysis techniques encounter difficulties in effectively managing the complexity and uncertainty inherent in modern engineering systems, particularly within the domain of data center cooling systems. For CRAC systems in data centers, reliability analysis manages thermal output, power consumption, and environmental variations using methodologies such as PUE, SCADA controls, and hierarchical modelling. The conventional methodologies struggle to accurately assess the reliability and performance of these systems due to their multi-dimensional nature and dynamic operation conditions, compounded by scalability and efficiency challenges when applied to large-scale data centers, and necessitates real-time reliability analysis. Current methods also cannot often provide continuous, real-time updates, which is essential for timely decision-making in mission-critical data center environments.

The research gap addressed by this work underscores the absence of studies on CRAC systems of data centers and their probabilistic constraints through surrogate modelling, representing a worldwide research gap. This research bridges this gap by introducing a novel algorithm based on the Adaptive Multiple Response Gaussian Process (AMRGP) model and adaptive sampling techniques of reliability analysis for engineering systems, thereby advancing the field. This introduction of a robust framework, provides invaluable insights to practitioners and researchers to optimize airflow, and temperature distribution, and integrate advanced control strategies to enhance the reliability and energy efficiency of data centers.

CHAPTER 3: METHODOLOGY

This study introduces an innovative approach to reliability analysis, accentuating the spatial interpolation of radiation heat transfer and exchange mechanisms. It employs direct sampling techniques to probabilistically evaluate the CRAC system's failure and reliability.

3.1: Reliability Analysis of CRAC System

The CRAC system of the data center utilizes data collected from temperature and humidity sensors positioned at the return air inlet and various strategic spatial points. Operators manually adjust the desired temperature to enhance the CRAC system's performance and heat exchange capability. Effective management of cool and hot airflow within data center aisles is critical, necessitating precise zoning and channelization strategies and arrangement to markedly enhance the efficiency and robust performance of the CRAC system. Consequently, the selection of optimal capacity and layout of data server racks to facilitate active heat exchange by the CRAC system presents a substantial challenge to the industry.

Furthermore, the layout and airflow patterns must strike a balance between the data server's capacity and the cost of execution. This study introduces a novel modelling technique for reliability analysis, highlighting spatial distribution of heat transfer mechanisms from data server racks to the data center environment, and subsequently to the CRAC system. The proposed approach employs adaptive sampling technique to derive probabilities of failure and reliability values. The aim is to model an efficient modelling technique providing high accuracy in the design and equipment selections of the data centers.

Once the data center racks and CRAC model are converged, they can generate temperature distribution patterns and actual heat transfer gradients, offering a benchmarking technique to analyze various load scenarios and assess the effectiveness of

the CRAC system's capacity within the data center. The probability of failure of the proposed method is expressed as:

$$P_f^{DC} = \Pr(f_{\mathrm{X}}(\boldsymbol{x}), \mathrm{X} = 1,2, \dots, n) = \Pr(\min_{\mathrm{X}=1,2,\dots,n} f_{\mathrm{X}}(\boldsymbol{x}) < 0) \quad (1.1)$$

In this context, P_f^{DC} represents the probability that the performance function f_{X}, which is influenced by various input factors denoted as $\boldsymbol{x}$ (such as the area of data server racks, temperature gradients, flow rate of CRAC system, density of air, and air emissivity), yields a value less than 0. The critical performance functions, which define radiation heat transfer and rate of heat transfer by the CRAC system, are symbolized as $f_{\mathrm{X}}(\boldsymbol{x})$. It depends on deterministic design space within $\boldsymbol{x}$ and ensure a reliability level surpassing a predetermined mark.

3.1.1 Radiation Heat Transfer of Data Server Rack

Radiation heat transfer, as a probabilistic constraint, refers to a conditional requirement within a system or process where the transfer of thermal energy through radiation must meet certain probabilistic criteria. This constraint is particularly important where heat transfer plays a crucial role, such as in data centers. It is a mathematical relationship between the temperature of a data center rack and the rate at which it emits thermal radiation. The total radiant heat energy emitted per unit surface area of a rack per unit of time is directly proportional to the fourth power of its absolute temperature (measured in Kelvin). The probabilistic constraint can be expressed mathematically as:

$$R_{rack} = \epsilon \times \sigma \times A \times (T_{surface}^4 - T_{ambient}^4) \quad (1.2)$$

- R_{rack} represents the radiation heat transfer rate of the rack, in watts (W).
- ϵ symbol denotes the emissivity of the rack's surface ranging from 0.1 to 0.9.
- σ is the Stefan-Boltzmann constant, with a value of 5.67×10^8 $(W/(m^2.K^4))$.
- A represents the surface area of the rack in square meters m^2.
- $T_{surface}$ is the temperature of the rack's surface, measured in Kelvin (K).

- $T_{ambient}$ pertains to the ambient temperature of the data center environment.

3.1.2 Heat Transfer Rate of CRAC System

The heat transfer rate of CRAC systems involves the removal of heat from the data center space using airflow, heat exchange, and refrigeration technologies. It requires proper design, sizing, and operation of the CRAC system for efficient heat transfer and maintaining optimal operating conditions of data servers and their equipment in the data center. By multiplying the mass flow rate, density of air, specific heat capacity, and temperature difference, we obtain the total heat transfer rate of the CRAC system. The mathematical equation can be presented as:

$$R_{crac} = \dot{m} \times \rho \times c_p \times (\Delta T) \tag{1.3}$$

- R_{crac} is the heat transfer rate in watts (W).
- $\dot{m}$ represents the mass flow rate of the air in (m^3/h).
- ρ is the density of air in (kg/m^3).
- c_p pertains to the specific heat capacity of air in $(J/kg.K)$.
- ΔT reflects the temperature difference of air medium at CRAC I/O in (K).

3.2: Baseline Method

This work refers to [33] for the baseline method. In contrast to the Kriging model's handling of a single output variable as isotropic Gaussian stochastic process, the AMRGP model is capable of accommodating multidimensional output variables. It establishes a comprehensive approximate relationship between input and output variables. Overall, the proposed approach can simultaneously consider both accuracy and efficiency in reliability estimation, enabling significant time savings while ensuring the correctness of the solution.

Table 3.1: Classification of Kriging and AMRGP method

Classification	Kriging	AMRGP
Modelling approach	Gaussian process based	Extended Gaussian process with separate covariance functions
Handling of output variables	Single output based on spatial correlation	Multiple response variables with correlation and joint predictions
Covariance function	Isotropic or anisotropic	Correlation with input variables and response variables
Application domain	Spatial domain	Engineering, biology, multiple response functions of Mechanical systems
Complexity and scalability	Intensive for large datasets due to big covariance metrics	Scalable interference techniques and advanced Gaussian processes are feasible for response variables.
Predicative performance	Low accuracy for complex and uncertain application	Best accuracy for intricate correlation and complex spatial relationships of variables.

3.3: Adaptive Multiple Response Gaussian Process Model Architecture

The AMRGP method is characterized by the following mathematical formulation:

$$\boldsymbol{\mathcal{Z}}^n(\boldsymbol{\theta}, \boldsymbol{\theta}') \sim MGP(\boldsymbol{\mathcal{T}}^n(\boldsymbol{\theta}, \boldsymbol{\theta}')\boldsymbol{\mathcal{M}}^n, \boldsymbol{\Sigma}^n\boldsymbol{\mathcal{R}}^n((\boldsymbol{\theta}, \boldsymbol{\theta}'), (\boldsymbol{\theta}, \boldsymbol{\theta}')) \quad (1.4)$$

The AMRGP model has the capability of handling multi-dimensional output variables represented as $\boldsymbol{\mathcal{Z}} = [\boldsymbol{\mathcal{Z}_1}, \boldsymbol{\mathcal{Z}_2}, \dots, \boldsymbol{\mathcal{Z}_n}]$, which formulates them as an n-dimensional Gaussian process with separable covariance functions.

- In the above equation, $MGP(\boldsymbol{\theta}, \boldsymbol{\theta}')$ denotes the Gaussian process used for constructing the total estimated affiliation between input and output variables.
- $\boldsymbol{\mathcal{T}}^n(\boldsymbol{\theta}, \boldsymbol{\theta}') = [t_1^n(x, x'), t_2^n(x, x'), \dots, t_z^n(x, x')]$ represents the vector of regression models, while $\boldsymbol{\mathcal{M}}^n = [\boldsymbol{m_1^n}, \boldsymbol{m_2^n}, \dots, \boldsymbol{m_z^n},]$ stands for the vector of regression coefficients.

- $\boldsymbol{\mathcal{R}}^n(\boldsymbol{\theta}, \boldsymbol{\theta}')$ denotes a spatial correlation function indicating three-dimensional position, and $\boldsymbol{\Sigma}^n$ represents an unidentified covariance matrix essential for managing multiple failure models.

In this framework, to illustrate the correlation among multi-dimensional output response variables, an isotropic Gaussian model is employed to measure the spatial location and correlation. The isotropic Gaussian model is built from:

$$Y^n(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\sum_{I=1}^{z} \boldsymbol{\omega}_i^n (x_i - x_i')^2\} \tag{1.5}$$

The primary objective of incorporating isotropic Gaussian model is to assess how quickly the correlation among the multiple failure modes decreases to zero, leveraging the roughness parameter denoted as a vector $\boldsymbol{\omega}_i^n = (\omega_1, \omega_2, \dots, \omega_z)$. The lower values of ω_i^n correspond to a well-adjusted AMRGP model for the responses $\boldsymbol{\mathcal{Z}}^n(\boldsymbol{\theta}, \boldsymbol{\theta}')$. The spatial correlation function $\boldsymbol{\mathcal{R}}^n(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is associated with $\boldsymbol{\mathcal{M}}^n, \boldsymbol{\Sigma}^n$ in Equation (1.4) and can be fully expressed by solving the following equations:

$$\widehat{\boldsymbol{\mathcal{M}}^n} = [(\boldsymbol{\mathcal{T}}^n)^T \mathbf{P}^{-1} \boldsymbol{\mathcal{T}}^n]^{-1} (\boldsymbol{\mathcal{T}}^n)^T \mathbf{P}^{-1} \boldsymbol{\mathcal{Z}} \tag{1.6}$$

$$\widehat{\boldsymbol{\Sigma}^n} = \frac{1}{z}[(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{T}}^n \boldsymbol{\mathcal{M}}^n)^T \mathbf{P}^{-1} (\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{T}}^n \boldsymbol{\mathcal{M}}^n)] \tag{1.7}$$

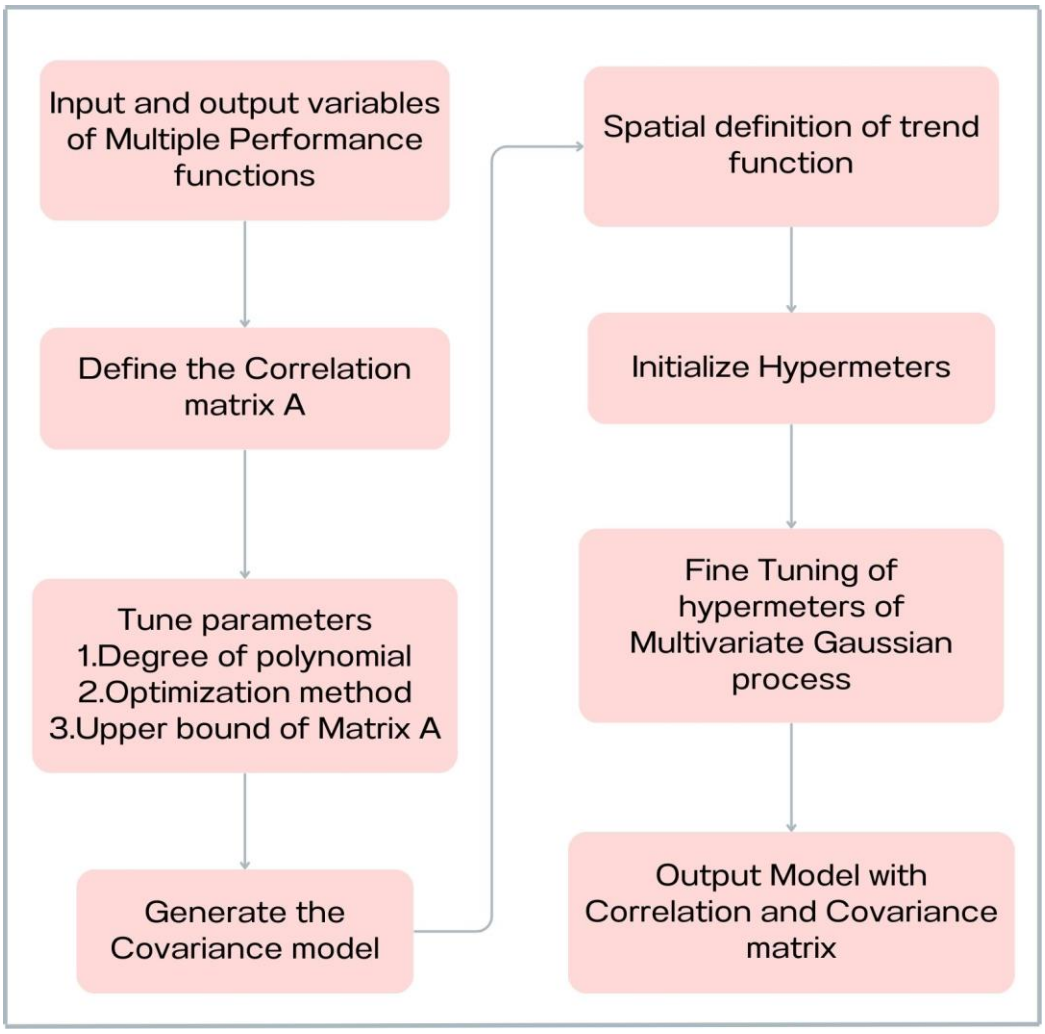

Figure 3.4: Block Diagram of AMRGP Model Architecture

In these equations, **P** is the correlation matrix of $Y^n(\boldsymbol{x}, \boldsymbol{x}')$, $\boldsymbol{\mathcal{Z}}$ is the corresponding matrix of output response values, and z represents the number of sample points to be modeled. The solution for the roughness parameter $\boldsymbol{\omega}_i^n$ is vital for constructing the AMRGP model. To address this, a logarithmic inverse function $\log[z(vect(\boldsymbol{\mathcal{Z}})|\boldsymbol{\mathcal{M}^n}, \boldsymbol{\Sigma}^n, \boldsymbol{\omega}_i^n]$ of the maximum likelihood technique [34] is utilized:

$$\begin{aligned}\text{Log}[z(vect(\boldsymbol{\mathcal{Z}})|\boldsymbol{\mathcal{M}^n}, \boldsymbol{\Sigma}^n, \boldsymbol{\omega}_i^n] = -\frac{nz}{2}\log(2\pi) - \frac{z}{2}\log(|\boldsymbol{\Sigma}^n|) - \\ \frac{n}{2}\log(|\mathbf{P}|) - \frac{1}{2}vect(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{T}}^n\boldsymbol{\mathcal{M}}^n)^T \times (\boldsymbol{\Sigma}^n \otimes \mathbf{P})^{-1} vect(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{T}}^n\boldsymbol{\mathcal{M}}^n)\end{aligned} \tag{1.8}$$

Here, the methodology involves sorting a matrix by a column vector represented by $vect(\boldsymbol{\theta})$ and using the Kronecker product $\otimes$. The static AMRGP model is then constructed using Equation (1.8), facilitating the prediction of multidimensional response values $\widehat{\boldsymbol{\mathcal{Z}}}(x^z)$ at any uncertain point x^z. The mean and variance of these predicted values are subsequently formulated as follows:

$$\mu_{\widehat{\mathcal{Z}}}(x^z) = \boldsymbol{\mathcal{T}}^n(x^z)\boldsymbol{\mathcal{M}}^{\boldsymbol{n}} + c(x^z)^T\mathbf{P}^{-\mathbf{1}}(\boldsymbol{\mathcal{Z}} - \boldsymbol{\mathcal{T}}^n\boldsymbol{\mathcal{M}}^n) \tag{1.9}$$

$$\begin{aligned}\sigma_{\widehat{\mathcal{Z}}}(x^z) = diag_{vec}\Bigg(\sigma \\ \times \Bigg(1 - c(x^z)^T\mathbf{P}^{-1}c(x^z) \\ + \Big(\boldsymbol{\mathcal{T}}^n(x^z)^T - (\boldsymbol{\mathcal{T}}^n)^T\mathbf{P}^{-\mathbf{1}}c(x^Z)\Big)^T \\ \times ((\boldsymbol{\mathcal{T}}^n)^T\mathbf{P}^{-1}\boldsymbol{\mathcal{T}}^n)^{-\mathbf{1}}\Big(\mathcal{T}^n(x^z)^T - (\boldsymbol{\mathcal{T}}^{\boldsymbol{n}})^T\mathbf{P}^{-1}c(x^Z)\Big)\Bigg)\Bigg)\end{aligned} \tag{2.0}$$

The regression model vector is denoted as $\boldsymbol{\mathcal{T}}^n(x^z)$**,** which is formulated at an unknown point x^z, $c(x^z)$ signifies the spatial correlation vector among the unknown point x^z, and the input trial point $N_t = [x_1, x_2, \dots, x_n]$. The $diag_{vec}$ represents the transverse elements of the matrix. The predicted average value of an unforeseen data point can be determined by using the mean of random variables, while the variance offers insights into the level of uncertainty specific to that point.

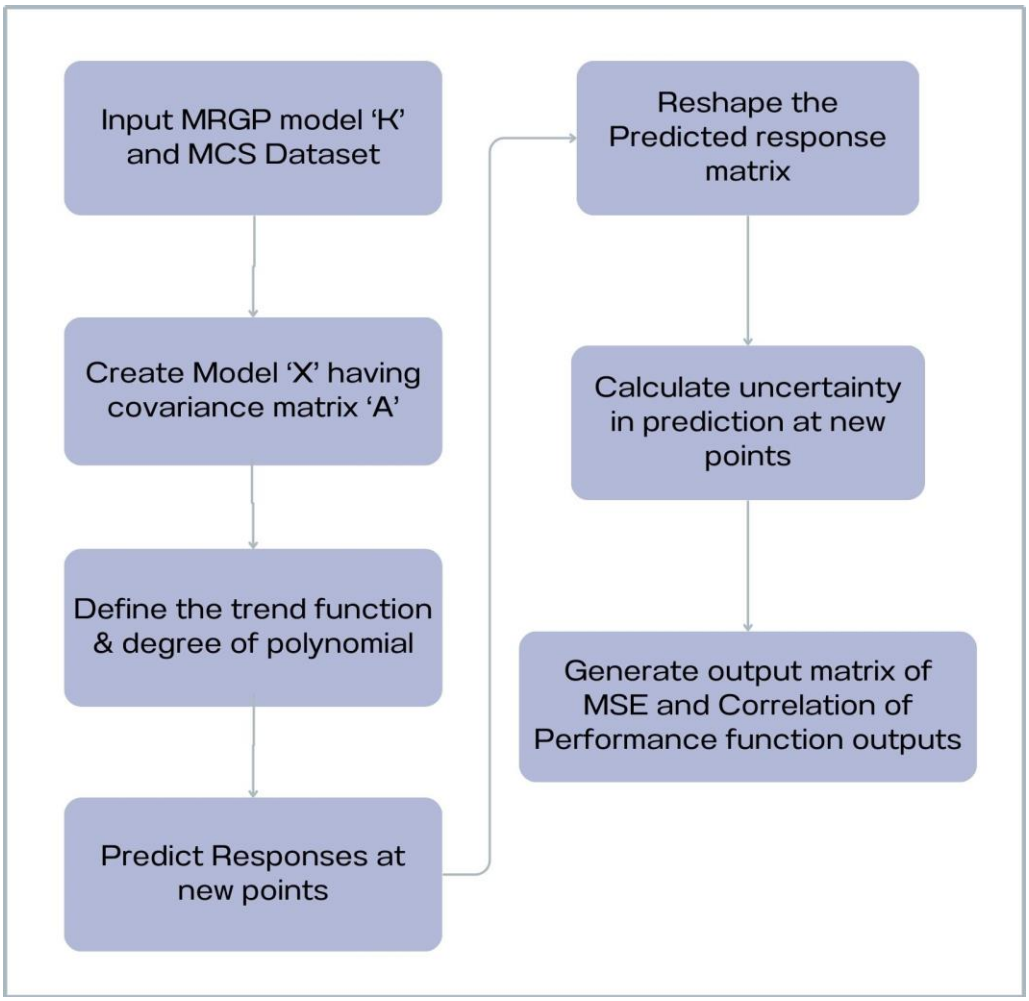

Figure 3.5: Block diagram of AMRGP Prediction Function

The AMRGP model can be utilized to create the limit state surface of a system that accounts for multiple failure modes. Despite the significant improvement in the quality of preliminary sample points generated by the LHS method, the model's accuracy still falls short of requirements due to the limited number of samples. Therefore, there is a need to enhance the adaptability of the AMRGP model to improve its accuracy and efficiency. The initial surrogate model for the extreme value response surface relies on the AMRGP model; therefore, new random variables are initialized by the learning functions. This study instigated advanced learning U-function introduced by [35] to identify the best-updated sample points from the candidate sample pool obtained through MCS and the expression is as follows:

$$\mathcal{U}(x^z) = \left(\left|\frac{\mu_{\widehat{\mathcal{Z}}}(x^z)}{\sigma^2_{\widehat{\mathcal{Z}}}(x^z)}\right|\right) \tag{2.1}$$

Here, $\mu_{\widehat{\mathcal{Z}}}(x^z)$ denotes the predicted mean, and $\sigma_{\widehat{\mathcal{Z}}}(x^z)$ represents the predicted standard deviation of the AMRGP model, respectively. Subsequently, the sample points requiring refinement can be recognized as follows:

$$x^{\Theta} = \min_{x \in x_{mc}} \mathcal{U}(x^z) \geq th_u \tag{2.2}$$

Where, x^{Θ} is the new sample point obtained from the learning function and is reintroduced in N_t, and x_{mc} represents the 10^6 sample repository generated through MCS. Subsequently, the probability of failure comes out to be $P_f^{DC} = P(\mu_{\widehat{\mathcal{Z}}}(x^z) < 0)$ as offered in Equation (1.1).

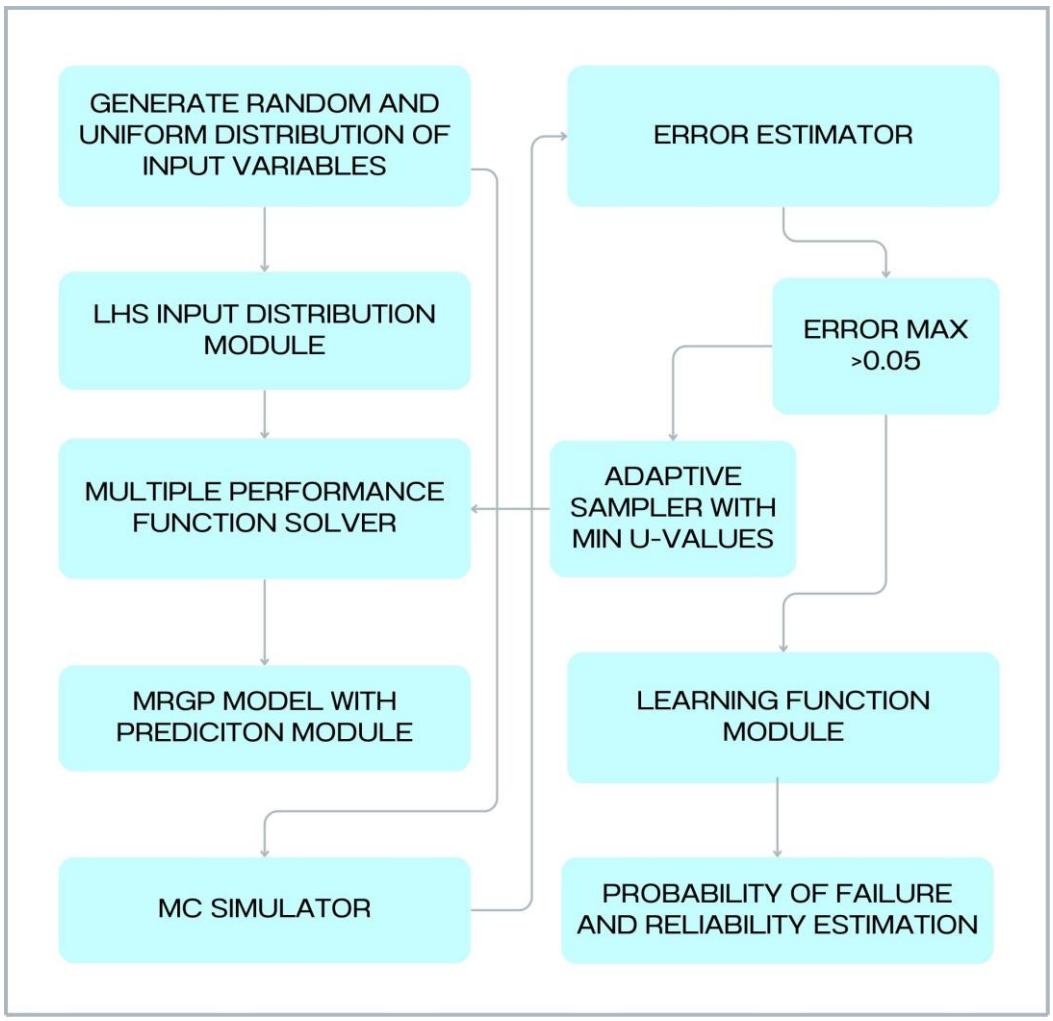

Figure 3.6: Block diagram of Proposed Adaptive AMRGP Algorithm

The coefficient of variation of the MCS method for the estimated probability can be expressed as follows:

$$CV(\mathrm{P}_f) = \sqrt{\frac{1 - \mathrm{P}_f^{\mathrm{DC}}}{(n_{mc} - 1)\mathrm{P}_f^{\mathrm{DC}}}} \tag{2.3}$$

The estimated probability of failure is denoted as $\mathrm{P}_f^{\mathrm{DC}}$ and n_{mc} is the total number of MCS samples. If the coefficient of variance exceeds 5%, the estimation is considered unacceptable, necessitating expansion of the sample repository. On the other hand, if it falls below 5%, the estimation is considered acceptable, indicating that the number of

samples required to achieve accurate and precise results is adequate. Once the new samples have been initialized and the outputs from the performance functions obtained, the erros is calculated to evaluate the model's accuracy. Where, f_{mc} denotes the actual count of failing trajectories from the n_{mc} sample repository, while f_{lhs} is the actual count of failing samples. So, the error expression is formulated as:

$$\zeta = \max_{f_{lhs} \epsilon [0, f_{mc}]} \frac{f_{lhs}}{|f_{mc} - f_{lhs}|} \times 100\% \tag{2.4}$$

3.4: Summary of Proposed Adaptive AMRGP Reliability Analysis Method:

Step1: Generate an initial set of samples denoted as N_t, using Latin hypercube sampling (LHS).

Step2: Utilize the samples generated in Step1 to calculate the corresponding results by training them through the two probabilistic performance functions.

Step3: Initialize a Multiple Response Gaussian Process model using the input samples and the corresponding output.

Step4: Generate new samples by MCS technique through random distributions of input variables.

Step5: Predict the response of the system and formulate the error of each performance function through MCS samples and established AMRGP model.

Step6: Calculate the results by employing the learning function using the MCS samples and the output predictions obtained in Step5.

Step7: Assess the precision of the AMRGP model. If it satisfies the specified threshold of less than 0.05%, then proceed with reliability estimation.

Step8: If the outcomes from the AMRGP model are not precise, identify and reintroduce acquired samples from learning U-function. Incorporate these new samples in N_t and repeat Step7 until accurate results are achieved.

Step9: Estimate the reliability and error of the refined AMRGP model.

CHAPTER 4: DATACENTER MODELLING

4.1: Implementation Details

A practical model of a Class-A data center CRAC system and its layout arrangement is presented in Figure4.7. In this study, a DX-type air-cool CRAC system is selected with an in-row air distribution arrangement which guarantees targeted cooling with exceptional heat dissipation capacity providing high availability and low commissioning cost.

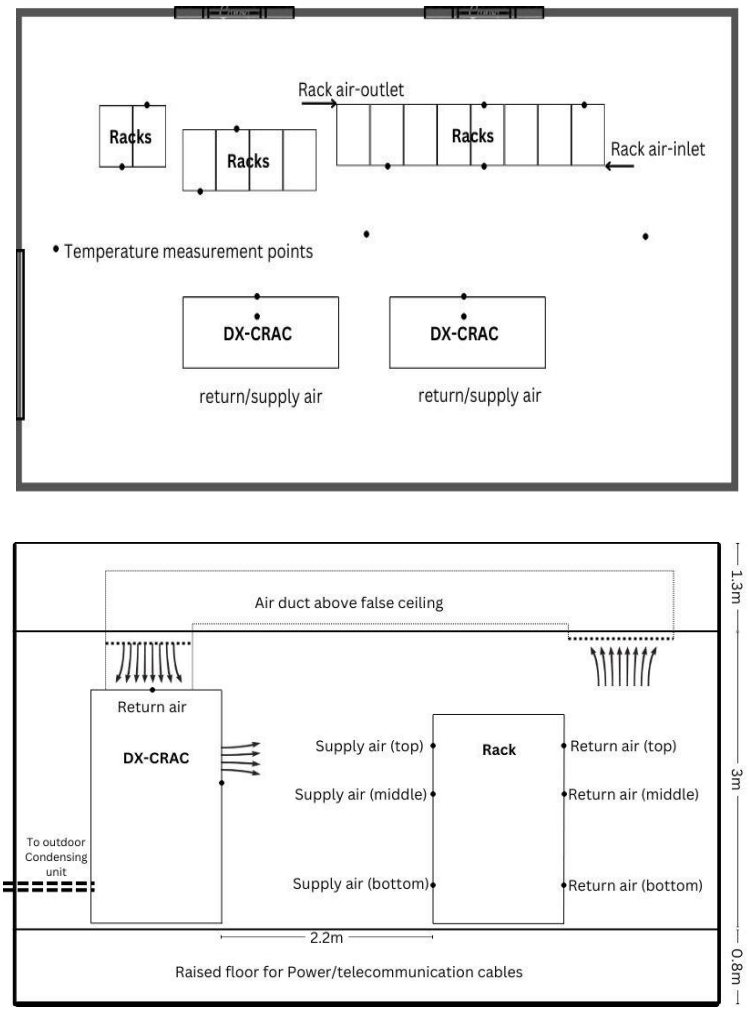

Figure 4.7: (a)Typical Layout Plan of the Data center with inflow air distribution arrangement (b) Cooling System Configuration in Data center

The temperature ranges are determined using commercially accessible computational fluid dynamics (CFD) analysis program Mentor MA FloVENT, the detailed description of this simulation technique is provided in [36]. This packaged software utilizes input as the root assembly with geometrical details of data center enclosure including doors and windows. Fixed flow CRAC is selected for Supply/Return air temperature measurements with monitor points placed as per data center layout plan. The heat transfer power along with the flow rate of the racks of the datacenter is provided to access the gradient distribution of thermal heat transfer, which helps in defining the capacity of the CRAC system.

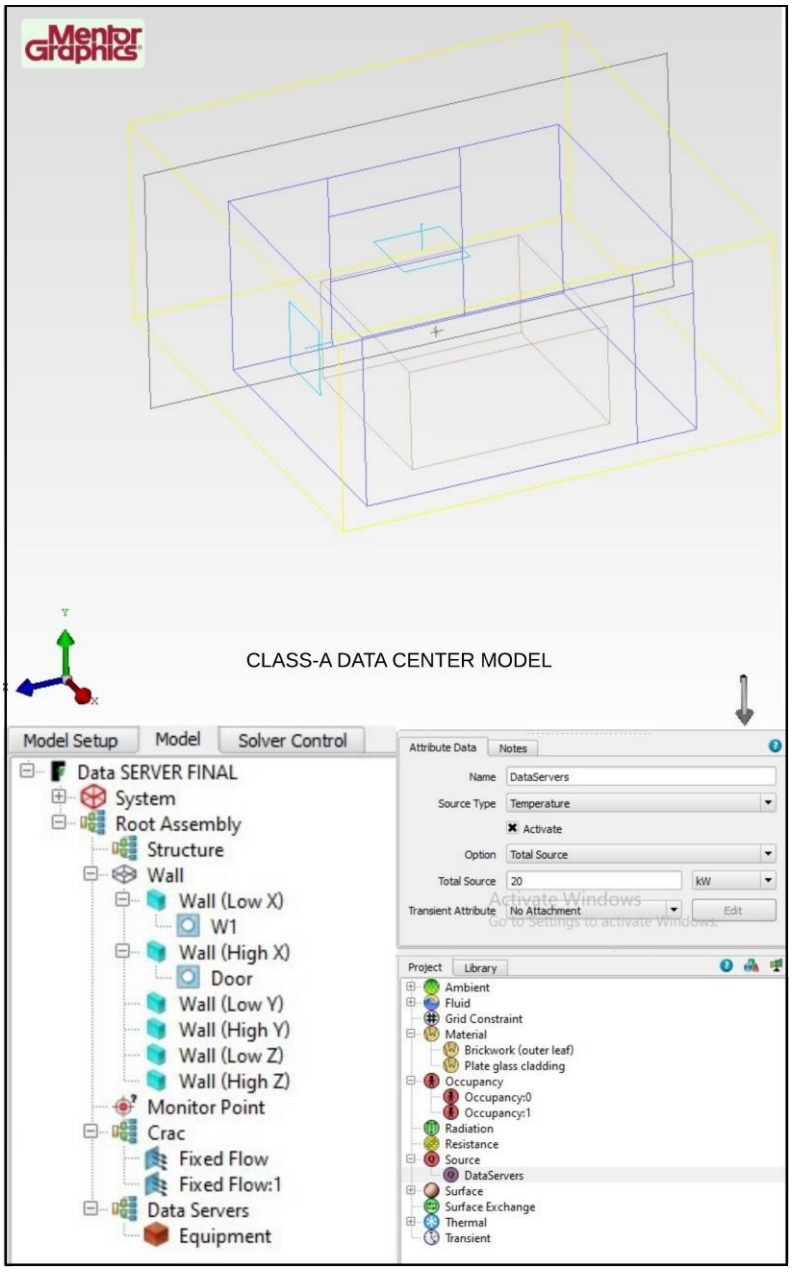

Figure 4.8: FloVENT geometric model of Class-A data center system

The simulation run by the FloVENT generated the transient heat model and the temperature gradient for the datacenter equipment is exhibited as shown in Figure4.9.

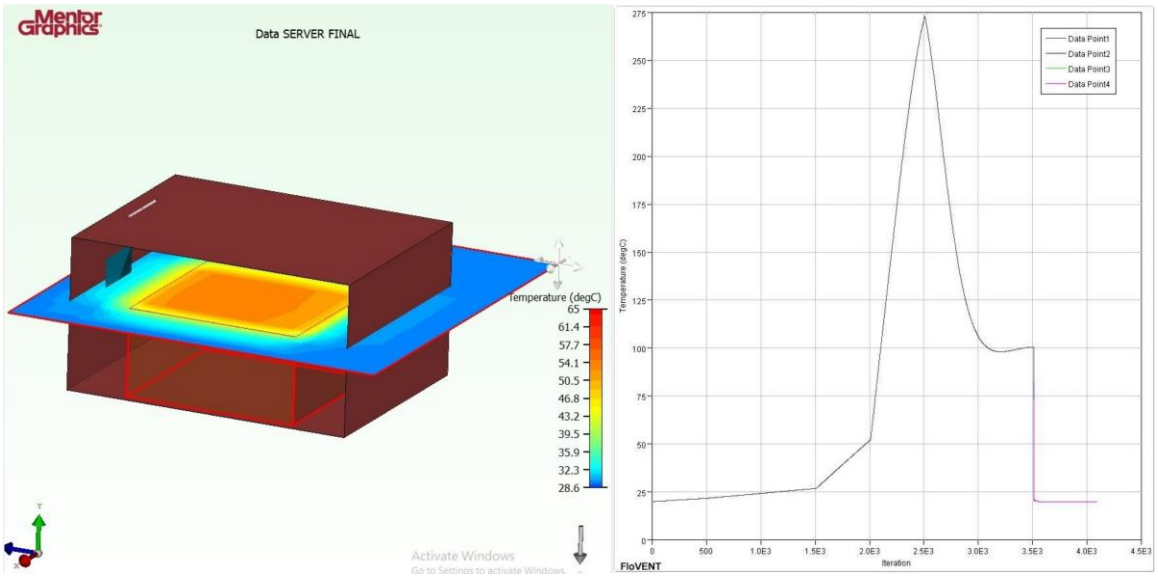

Figure 4.9: (a)Thermal model illustrating the data center environment temperatures after experimental modifications. (b) The graph presents the temperature distribution against the set data points**.**

Once the heat transfer model is converged to the required ambient conditions. The next step involves creating an adaptive surrogate model for the two performance functions as derived in Section 3.1.1 & 3.1.2. The output of the first performance function is a spatial representation of the surface temperature of the data server rack ($T_{surface}$), the ambient temperature of the data center ($T_{ambinet}$) & CRAC air (ΔT) as uniformly distributed gradients. The emissivity of air (ϵ) between the range $0.1 - 0.9$ and the normally distributed area A_x of data center racks are trained as inputs to the model. The uniformly distributed temperature gradients T_1 and T_2 obtained from the FloVENT Model are taken as probabilistic constraints. Flow rate m of the CRAC system is normally distributed and selected using [37], with uniform delta air temperature gradient. A single multi-response surrogate model is then generated to examine the typical distribution of thermal gradient and the effect of the heat transfer rate of the CRAC system from heat-transmitting data server racks. In this method, a crucial task is to identify a single failure point of both performance functions along the entire temperature profile which provides the precise reliability of the system. The parameters describing the distribution of the probabilistic input variables are listed in Table 4.1.

Table 4.1: Dispersal constraints of data server rack and CRAC system

Input parameters	Probabilistic mean	Standard deviation	Type of distribution
m	5	1	Normal distribution
ε	0.8995	1	Normal distribution
A	6	1	Normal distribution
ΔT	$2 \leq T \leq 6$	–	Uniformly distributed
$T_{surface}$	$40 \leq T_1 \leq 58$	–	Uniformly distributed
$T_{ambient}$	$20 \leq T_2 \leq 30$	–	Uniformly distributed

4.2: Machine Setup

Reliability modelling and estimation require special hardware and software equipment. The details of the server machine are as follows:

Table 4.2: Machine setup details (Software + Hardware)

Setup	IT Equipment	Description
Hardware	Motherboard	MSI x570
	CPU	Intel i5-3330
	GPU	NVIDIA Quadro RTX 5000
	RAM	16 GB
	SSD	500 GB
Software	OS	Microsoft Windows 10 Pro
	Language	MATLAB R2018a
	IDE	MATLAB desktop environment
	Library	2GB
	CFD Modelling Tool	FloVENT 11.3

4.3: Fine-tuning Hypermeters of AMRGP Model

Fine-tuning hypermeters of the AMRGP model involves optimizing its configuration to enhance its ability to capture dependencies and make predictions for multiple responses simultaneously. In this study, the upper bound for the correlation matrix is set at 0.6 and the degree of a polynomial function is set as 1.

4.4: Optimization method

A MATLAB optimization function ‘fmincon’ is used for solving constrained nonlinear probabilistic performance functions [38]. This algorithm provides an improved solution by iteratively adjusting the hypermeters as input and returns a scalar value representing the objective to be minimized.

CHAPTER 5: RESULTS AND DISCUSSION

The experimentation aimed to evaluate two distinct probabilistic performance functions of CRAC system and data server racks which are heat-emitting and exchanging sources within the data center environment. Adaptive multiple response Gaussian process model produced notably precise and accurate results. Each performance function underwent scrutiny to establish the critical performance and reliability estimation by formulating the probability of failure.

The built adaptive AMRGP model is based on a combination of variables which are randomly distributed along with performance function data. Initially, 07 samples ($N_t = 7$) are generated to create the surrogate model for both performance functions. The model undergoes iterative refinement till the error margin falls under 5%. Once the model is refined, the reliability values are calculated. The efficiency of the algorithm is evaluated based on the sum of performance function calls, and are lesser than the number of samples used for building the model. The results are then compared to the MCS method which is built on 10^6 samples and is vastly accurate due to its ability to analyze large datasets. The error value is then calculated using the formula $\zeta = \frac{|R_{mc} - R|}{R_{mc}} \times 100$. The proposed adaptive AMRGP method with fewer samples, demonstrates an error of less than 0.09% when compared to the MCS method having 10^6 samples. This indicates accurate and precise results, as decreasing the sample size improves the accuracy.

Table 5.1: Results of Proposed Reliability analysis methods in comparison with other methods

Methods used	**Number of samples**	$\boldsymbol{P_f^{DC}}$	**Error**
Proposed Adaptive AMRGP	7	0.0012	0.13%
MRGP	20	0.0019	0.16%
Adaptive Kriging	20	0.0037	0.16%
Polynomial Chaos Expansion	1×10^3	0.049	5.10%

MCS	1×10^6	0.0029	-

For Reliability analysis of the Computer Room Air Conditioning (CRAC) system, conducting physical experiments is often impractical due to the high risk and cost associated due to uncertainties. Consequently, computational fluid dynamics (CFD) simulations are frequently employed for Reliability and Availability analysis. Furthermore, building a Gaussian Process model based on a full CFD dataset is very computationally intensive. To overcome this issue, Latin Hypercube Sampling (LHS) is employed to determine the highest levels of variability in input variables ΔT, $T_{surface}$, and $T\_ambient$. The fitted adaptive AMRGP model provides estimated mean function coefficients and the correlation parameters based on the exponential covariance function. The results achieved are highly accurate with low error values. Moreover, a complete heat transfer model of the CRAC system of the datacenter is developed which is the main contribution of this research. The analysis conducted in this study aligns with the general principles of heat transfer and thermodynamics, validating the model's predictions to tackle the computational costs.

This research does not address the energy management techniques developed to enhance the reliability and efficiency of data centers. Instead, the focus is on the heat transfer models of the data center's major components. Additionally, the integration of sustainable and green energy resources presents novel challenges in data center operations. The impact of green technologies, such as renewable energy generation and free cooling techniques, on energy and heat transfer modelling approaches is not covered and can be explored further.

CHAPTER 6: CONCLUSIONS AND FUTURE RECOMMENDATION

In conclusion, the proposed methods offer several advantages over existing analytical approaches. (1) A novel AMRGP Reliability analysis technique is developed for dynamic and very uncertain data center environments, an area that is rarely explored in literature. (2) The proposed method efficiently identified optimal design variables having uniform and random distributions throughout the sampling space. This approach effectively prevents issues related to sample points clustering along the critical design space boundary. (3) The algorithm and model simultaneously consider multiple performance functions, making it suitable for system-level design analysis. (4) The model provides correlation among the multiple performance functions of the CRAC system and data server racks, which can facilitate better equipment selections in the future. (5) The developed algorithm predicts the Reliability and error with notable accuracy and reduced computational cost and energy.

The reliability assessment of the complete data center's infrastructure across various failure scenarios validates the practicality of the proposed approach. Findings reveal a system reliability of 0.9988 under dual failure modes, demonstrating a mere 0.13% deviation from MCS method. Importantly, this method notably enhances computational efficiency.

Future research perspectives are the adaption of sustainable energy resources in data centers and their impact on the reliability of CRAC systems. On the other hand, integrating various robust optimization algorithms with the proposed adaptive AMRGP method presents another promising research direction for Reliability analysis and design optimization of CRAC systems.

REFERENCES

[1] S. K. Uzaman, J. Shuja, T. Maqsood, F. Rehman, and S. Mustafa, "A systems overview of commercial data centers: initial energy and cost analysis," *International Journal of Information Technology and Web Engineering (IJITWE),* vol. 14, no. 1, pp. 42-65, 2019.

[2] M. Wiboonrat, "An empirical study on data center system failure diagnosis," in *2008 The Third International Conference on Internet Monitoring and Protection*, 2008: IEEE, pp. 103-108.

[3] A. TC, "Data center power equipment thermal guidelines and Best practices," *ASHRAE TC 9.9, ASHRAE, USA,* 2016.

[4] R. Teixeira, M. Nogal, and A. O'Connor, "Adaptive approaches in metamodel-based reliability analysis: A review," *Structural Safety,* vol. 89, p. 102019, 2021.

[5] W. Denson, "The history of reliability prediction," *IEEE Transactions on reliability,* vol. 47, no. 3, pp. SP321-SP328, 1998.

[6] T. Aven and U. Jensen, *Stochastic models in reliability*. Springer, 1999.

[7] K.-Y. Cai, "System failure engineering and fuzzy methodology an introductory overview," *Fuzzy sets and systems,* vol. 83, no. 2, pp. 113-133, 1996.

[8] K.-Y. Cai, *Introduction to fuzzy reliability*. Springer Science & Business Media, 2012.

[9] A. Mettas and M. Savva, "System reliability analysis: the advantages of using analytical methods to analyze non-repairable systems," in *Annual Reliability and Maintainability Symposium. 2001 Proceedings. International Symposium on Product Quality and Integrity (Cat. No. 01CH37179)*, 2001: IEEE, pp. 80-85.

[10] M. H. Davis, *Markov models & optimization*. Routledge, 2018.

[11] Z. Zhang, W. Li, and J. Yang, "Analysis of stochastic process to model safety risk in construction industry," *Journal of Civil Engineering and Management,* vol. 27, no. 2, pp. 87-99, 2021.

[12] O. Kammouh, P. Gardoni, and G. P. Cimellaro, "Probabilistic framework to evaluate the resilience of engineering systems using Bayesian and dynamic Bayesian networks," *Reliability Engineering & System Safety,* vol. 198, p. 106813, 2020.

[13] L. A. Jimenez-Roa, T. Heskes, T. Tinga, and M. Stoelinga, "Automatic inference of fault tree models via multi-objective evolutionary algorithms," *IEEE*

transactions on dependable and secure computing, vol. 20, no. 4, pp. 3317-3327, 2022.

[14] D. Panchal, U. Jamwal, P. Srivastava, K. Kamboj, and R. Sharma, "Fuzzy methodology application for failure analysis of transmission system," *International Journal of Mathematics in Operational Research,* vol. 12, no. 2, pp. 220-237, 2018.

[15] J. Huang, J.-X. You, H.-C. Liu, and M.-S. Song, "Failure mode and effect analysis improvement: A systematic literature review and future research agenda," *Reliability Engineering & System Safety,* vol. 199, p. 106885, 2020.

[16] M. C. Kim, "Reliability block diagram with general gates and its application to system reliability analysis," *Annals of Nuclear Energy,* vol. 38, no. 11, pp. 2456-2461, 2011.

[17] J. C. Helton, J. D. Johnson, C. J. Sallaberry, and C. B. Storlie, "Survey of sampling-based methods for uncertainty and sensitivity analysis," *Reliability Engineering & System Safety,* vol. 91, no. 10-11, pp. 1175-1209, 2006.

[18] X. Liu, Z.-J. Cao, D.-Q. Li, and Y. Wang, "Adaptive Monte Carlo simulation method and its applications to reliability analysis of series systems with a large number of components," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering,* vol. 8, no. 1, p. 04021075, 2022.

[19] M. Abyani and M. R. Bahaari, "A comparative reliability study of corroded pipelines based on Monte Carlo Simulation and Latin Hypercube Sampling methods," *International Journal of Pressure Vessels and Piping,* vol. 181, p. 104079, 2020.

[20] W. Yun and Y. Wang, "An efficient Kriging model-based importance sampling method for estimating the failure probability-based parameter global sensitivity index with uncertain distribution parameters," *Aerospace Science and Technology,* vol. 130, p. 107861, 2022.

[21] Z. Hu and S. Mahadevan, "A single-loop kriging surrogate modelling for time-dependent reliability analysis," *Journal of Mechanical Design,* vol. 138, no. 6, p. 061406, 2016.

[22] L. He and Y. Hung, "Gaussian process prediction using design-based subsampling," *Statistica Sinica,* vol. 32, no. 2, pp. 1165-1186, 2022.

[23] T. Zafar, Y. Zhang, and Z. Wang, "An efficient Kriging based method for time-dependent reliability based robust design optimization via evolutionary algorithm," *Computer Methods in Applied Mechanics and Engineering,* vol. 372, p. 113386, 2020.

[24] P. Wei, F. Liu, and C. Tang, "Reliability and reliability-based importance analysis of structural systems using multiple response Gaussian process model," *Reliability Engineering & System Safety,* vol. 175, pp. 183-195, 2018.

[25] ASHRAE, *Thermal guidelines for data processing environments*. ASHRAE, 2021.

[26] N. Rasmussen, "Calculating total cooling requirements for data centers," *White paper,* vol. 25, pp. 1-8, 2007.

[27] Z. Han, X. Sun, H. Wei, Q. Ji, and D. Xue, "Energy saving analysis of evaporative cooling composite air conditioning system for data centers," *Applied Thermal Engineering,* vol. 186, p. 116506, 2021.

[28] K. M. U. Ahmed, M. Alvarez, and M. H. Bollen, "Reliability analysis of internal power supply architecture of data centers in terms of power losses," *Electric Power Systems Research,* vol. 193, p. 107025, 2021.

[29] S. Nada and M. Said, "Effect of CRAC units layout on thermal management of data center," *Applied thermal engineering,* vol. 118, pp. 339-344, 2017.

[30] R. Khalid and A. P. Wemhoff, "Thermal control strategies for reliable and energy-efficient data centers," *Journal of Electronic Packaging,* vol. 141, no. 4, p. 041004, 2019.

[31] L. Souza, K. Camboim, J. Araujo, F. Alencar, P. Maciel, and J. Ferreira, "Dependability evaluation and sensitivity analysis of data center cooling systems," *The Journal of Supercomputing,* vol. 79, no. 17, pp. 19607-19635, 2023.

[32] J. Wang, Q. Zhang, S. Yoon, and Y. Yu, "Reliability and availability analysis of a hybrid cooling system with water-side economizer in data center," *Building and Environment,* vol. 148, pp. 405-416, 2019.

[33] M. Su, G. Xue, D. Wang, Y. Zhang, and Y. Zhu, "A novel active learning reliability method combining adaptive Kriging and spherical decomposition-MCS (AK-SDMCS) for small failure probabilities," *Structural and Multidisciplinary Optimization,* vol. 62, pp. 3165-3187, 2020.

[34] T. Lee and D. Shi, "A comparison of full information maximum likelihood and multiple imputation in structural equation modelling with missing data," *Psychological Methods,* vol. 26, no. 4, p. 466, 2021.

[35] H.-M. Qian, Y.-F. Li, and H.-Z. Huang, "Time-variant system reliability analysis method for a small failure probability problem," *Reliability Engineering & System Safety,* vol. 205, p. 107261, 2021.

[36] T. Mikjaniec, A. Manning, D. Small, and J. VanGilder, "Data center design using improved CFD modelling and cost reduction analysis," in *2011 27th Annual IEEE Semiconductor Thermal Measurement and Management Symposium*, 2011: IEEE, pp. 97-103.

[37] *Uniflair and EcoBreeze Selection Catalog*, 2024. [Online]. Available: http://www.uniflair.co.uk/pdfs/CRAC%20brochure%202015.pdf.

[38] A. M. Albaghdadi, M. B. Baharom, and S. A. bin Sulaiman, "Parameter design optimization of the crank-rocker engine using the FMINCON function in MATLAB," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1088, no. 1: IOP Publishing, p. 012072.