HYBRID SURROGATE MODEL FOR PRESSURE AND TEMPERATURE PREDICTION IN A DATA CENTER AND ITS APPLICATION

HYBRID SURROGATE MODEL FOR PRESSURE AND TEMPERATURE PREDICTION IN A DATA CENTER AND ITS APPLICATION

BY

Sahar Asgari

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

TITLE**:** Hybrid Surrogate Model for Pressure and Temperature Prediction in a Data Center and Its Application

AUTHOR: Sahar Asgari,

M.A.Sc. (University of Tehran, Iran),

B.Sc. (Sahand University of Technology, Iran).

SUPERVISORS:

Professor Ishwar K. Puri

Professor Rong Zheng

NUMBER OF PAGES: 164

Abstract

One of the crucial challenges for Data Center (DC) operation is inefficient thermal management which leads to excessive energy waste. The information technology (IT) equipment and cooling systems of a DC are major contributors to power consumption. Additionally, failure of a DC cooling system leads to higher operating temperatures, causing critical electronic devices, such as servers, to fail which leads to significant economic loss. Improvements can be made in two ways, through (1) better design of a DC architecture and (2) optimization of the system for better heat transfer from hot servers.

Row-based cooling is a suitable DC configuration that reduces energy costs by improving airflow distribution. Here, the IT equipment is contained within an enclosure that includes a cooling unit which separates cold and back chambers to eliminate hot air recirculation and cold air bypass, both of which produce undesirable airflow distributions. Besides, due to scalability, ease of implementation, and operational cost, row-based systems have gained in popularity for DC computing applications. However, a general thermal model is required to predict spatiotemporal temperature changes inside the DC and properly apply appropriate strategies. As yet, only primitive tools have been developed that

are time-consuming and provide unacceptable errors during extrapolative predictions. We address these deficiencies by developing a rapid, adaptive, and accurate hybrid model by combining a DDM and the thermofluid transport relations to predict temperatures in a DC. Our hybrid model has low interpolative prediction errors below 0.7 ^{o}C and extrapolative errors less than one half of black-box models. Additionally, by changing the studied DC configuration such as cooling unit fans and severs locations, there are a few zones with prediction error more than 2 ^{o}C.

Existing methods for cooling unit fault detection and diagnosis (FDD) are designed to successfully overcome individually occurring faults but have difficulty handling simultaneous faults. We apply a gray-box model involves a case study to detect and diagnose cooling unit fan and pump failure in a row-based DC cooling system. Fast detection of anomalous behavior saves energy and reduces operational costs by initiating remedial actions. Cooling unit fans and pumps are relatively low-reliability components, where the failure of one or more components can cause the entire system to overheat. Therefore, appropriate energy-saving strategies depend largely on the accuracy and timeliness of temperature prediction models. We used our gray-box model to produce thermal maps of the DC airspace for single as well as simultaneous failure conditions, which are fed as inputs for two different data-driven classifiers, CNN and RNN, to rapidly predict multiple simultaneous failures. Our FDD strategy can detect and diagnose multiple faults with accuracy as high as 100% while requiring relatively few simultaneous fault training data samples.

Our gray-box model exhibits superior performance compared with black-box models, such as ANN and NARX models. An application of the gray-box model involves a case study to detect single and simultaneous cooling unit failures in a row-based cooled DC.

Key words: Data center, row-based, temperature prediction, data-driven, ANN, CNN, RNN, machine learning, anomaly detection and diagnosis.

Acknowledgments

First of all, I would like to express my deepest gratitude and thanks to my Ph.D. supervisor Dr. Ishwar Puri, for his great support and warm-hearted advice towards the accomplishment of my research. I have been extremely lucky to have a supervisor who cared so much about my work. The door to Dr. Puri's office was always open whenever I ran into a trouble spot or had a question about my research or writing. It was a real privilege and an honor for me to share his exceptional scientific knowledge but also of his extraordinary human qualities.

I also like to extend my sincere appreciation to Dr. Rong Zheng for her constant support, invaluable guidance, and constructive suggestions, which were determinant for the accomplishment of the work presented in this thesis. She constantly encouraged and challenged me to explore beyond my set criteria. At many stages in the course of this research project, I benefited from her advice, particularly so when exploring new ideas.

I want to thank my advisory committee, Dr. Rong Zheng and Dr. James Cotton for their excellent constructive insights and invaluable feedback that greatly helped to clarify issues and to improve the quality of the thesis.

I also wish to express my profound gratitude and love to my parents for their endless love, wise counsel, and sympathetic ear throughout my life. This accomplishment would not have been possible without them. I would also like to thank my beloved husband, Hamidreza, for his love and also unfailing support and continuous encouragement through the process of researching and writing this thesis

Last, but not least, I want to thank all my lab mates, especially Dr. Hosein Moazamigoodarzi, Mr. Rohit Gupta, and Dr. Peiying Jennifer Tsai, for their continual support, discussions, and debates which helped me to promote my knowledge in my field. Thank you.

Declaration of Academic Achievements

This dissertation was used to fulfill the requirements of Ph.D. degree. All the research projects were conducted from September 2017 to August 2021. During the period of this study, we developed a gray-box model that combines machine learning with the thermofluid transport equations relevant for a row-based cooled DC to predict steady-state and transient temperatures in server CPUs and cold air inlet to the servers. An artificial neural network (ANN) embedded in the gray-box model predicts pressures, which provide inputs for the thermofluid transport equations that predict the spatio-temporal temperature distributions. The model is validated with experimental measurements for different (1) server workload distributions, (2) cooling unit set-point temperatures, and (3) the airflow of the cooling units. This gray-box model exhibits superior performance compared to a conventional zonal temperature prediction model and an advanced black-box model that is based on a nonlinear autoregressive exogenous model. An application of the gray-box model involves a case study to detect cooling unit fan and pump failure in a row-based DC cooling system. The major contribution of this thesis work was from me.

This thesis has resulted in three manuscripts, and I am the first author for all of them. Two of these manuscripts have been published in Elsevier journals and one is under review. I took the primary role in the design and conduction of the experiments, simulations, data analyses, and modeling reported in these manuscripts. Dr. Puri and Dr. Zheng provided helpful suggestions and guidance to flourish the initial idea. I wrote the first draft of the manuscripts. Dr. Puri, Dr. Zheng, and other authors provided editorial and technical input to generate the final draft of the papers. The papers are listed below:

1. **Sahar Asgari**, Hosein Moazamigoodarzi, Peiying Jennifer Tsai, Souvik Pal, Rong Zheng, Ghada Badawy and Ishwar K. Puri, *"Hybrid Surrogate Model for Online Temperature and Pressure Predictions in Data Centers"*, **Published** in *Future Generation Computer Systems,* 2021.
2. **Sahar Asgari**, SeyedMorteza MirhoseiniNejad, Hosein Moazamigoodarzi, Rohit Gupta, Rong Zheng, and Ishwar K. Puri, *"A Gray-Box Model for Real-Time Transient Temperature"*, **Published** in *Applied Thermal Engineering*, 2020.
3. **Sahar Asgari**, Rohit Gupta, Ishwar K, Puri, and Rong Zheng, *"A Data-Driven Approach to simultaneous Fault Detection and Diagnosis in Data Centers"*, **Under review** in *Applied Soft Computing*.

Table of Contents

Table of Figures

List of Tables

List of Abbreviations and Symbols

ANN	Artificial Neural Network
BPTT	Backpropagation Through Time
CFD	Computational Fluid Dynamic
CNN	Convolutional Neural Network
CRAC	Computer Room Air Conditioner
CRAH	Computer Room Air Handler
DC	Data Center
DDM	Data-Driven Model
FDD	Fault Detection and Diagnosis
GPR	Gaussian Process Regression
IT	Information Technology
LSTM	Long Short-Term Memory
NARX	Nonlinear Autoregressive Exogenous
OCSVM	One-Class Support Vector Machine
PCA	Principal Component Analysis

POD	Proper Orthogonal Decomposition
RBF	Radial Basis Functions
RNN	Recurrent Neural Network
SVM	Support Vector Machine
SVR	Support Vector Regression

Chapter 1

Introduction

In recent years, the unprecedented growing demand for cloud computing, online applications, and internet services has led to tremendous growth in size, number, and power consumption of data centers (DCs). It is estimated that by 2025 DC energy usage will account for 20% of worldwide consumption [1-3]. This growth generates many concerns regarding the electricity demand and environmental impacts from the DCs.

The energy density in DCs is very high and can be 10-100 times more than for conventional office buildings [4]. DC devices draw in raw electric power, produce some useful work, but more than 98% of the electricity is transformed to low-grade heat which the thermal management system must remove from the DC and release into the ambient air [5]. Maintaining a suitable environment for information technology (IT) infrastructure is the first priority in DC. Based on the ASHRAE guideline, the allowable rack inlet air temperature is maintained between 20 ^{o}C and 24 ^{o}C. At high temperatures, local hot spots can emerge which may lead to IT equipment failure, performance imbalance, excessive

power consumption, and are a threat to reliability. Thus, cooling systems have a critical role in continuously maintaining the safe, consistent, and reliable operation of DCs

The cooling required to maintain the IT equipment within a safe operating condition is one of the major contributors to DC power consumption. Depending on the specific IT equipment, cooling systems consume 24-60% of the total energy consumed by a DC [6, 7]. Therefore, an inefficient cooling structure leads to significant energy waste. Despite the liquid cooling technology such as heat reuse and efficient transferring heat, air cooling is the preferred method employed in DCs, which will remain for the foreseeable future due to its reliability, simplicity of air handling, lower capital and maintenance costs, and the uncertainties associated with liquid cooling systems [8-12].

1.1 Air Cooling

Although liquid cooling technology has undergone many improvements over the years, a large number of DCs still use air cooling systems to maintain the environmental conditions suitable for IT equipment operation. If a server becomes too hot, onboard logic will turn it off to prevent damage to the server. Therefore, the heat generated in a server should be extracted immediately. The cooling occurs in three steps: (1) Server cooling, where the IT equipment generates heat as the electronic components within them use electricity. Then, the IT equipment fans draw cold air across the internal components to remove the heat from the CPU and transfers it to the air flow passing over it. (2) Space cooling, the computer room air conditioner or handler (CRAC/CRAH) unit provides cold air to the IT equipment. (3) Finally, at the facility level, heat is rejected outside of the DC.

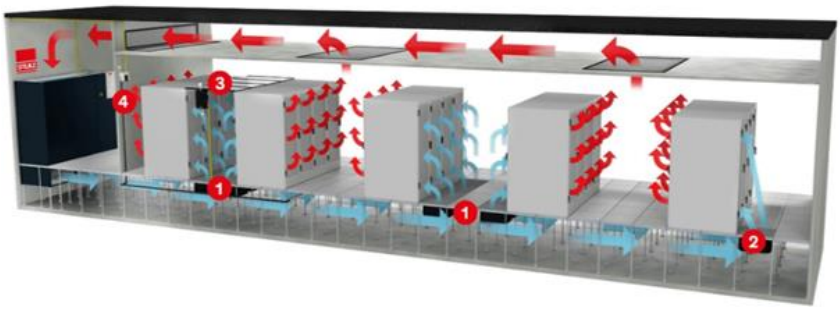
The hot aisle/cold aisle configuration is used in DCs to conserve energy and lower cooling costs by managing airflow. In this design, server racks are lined up in alternating rows in which the cold air intake side of the servers faces one way and the exhausted hot air side the other way. Cold aisles face CRAC/CRAH output ducts, while hot aisles face return ducts. The hot aisle/cold aisle configuration minimizes two major air distribution problems identified in DCs, i.e, bypass and recirculation, and optimize the thermal performance of DCs. If the cold air supplied to the IT equipment is insufficient, the hot air exhausted from servers is recirculated to the IT equipment inlets by the fans inside the servers, increasing the overall inlet air temperature. Bypass occurs when part of the cold airflow returns to the cooling unit without contributing at all to server cooling [13, 14].

1.2 DC cooling Architecture

Air cooling technology has improved significantly over the years. For decades, DCs used raised floor cooling systems to deliver cold air to servers. In a raised floor DC, the cold air from the CRAC/CRAH pressurizes the space below the raised floor and forces air through the perforated tiles that lie in fronts of servers. After passing through the servers, the exhaust air is returned to the CRAC/CRAH to be cooled. One of the major disadvantages of using this kind of DCs is cold and hot air mixing, which in turn increases the server inlet temperature and decreases the efficiency of the cooling system.

To remedy this problem, row-based cooling systems are used, where the cooling units are located between IT racks or mounted above them in DC cabinets. Thus, delivering cold air to a row of racks is easier and results in energy- and cost-efficiency. Figure 1-1 shows these two architectures.

a) Raised floor DC



b) Row-based cooling DC

Figure 1-1: Two different locations to put a cooling unit (architecture) that supplies cool air directly to the IT equipment [15].

1.3 Data-driven models

The advent of data-driven models (DDMs) has created a major development in the DC industry where they are being increasingly used as a replacement for CFD simulations and physics-based models. As a result of the nonlinearity and complexity of the airflow in a DC, sophisticated models are called to represent the complex relationships among the system-state (input, internal, and output) variables. The most widely established techniques which are appropriate for DC study are support vector regression (SVR), Gaussian process regression (GPR), artificial neural network (ANN), and nonlinear autoregressive exogenous (NARX) model.

The SVR algorithm is a regression algorithm that is suitable for both linear and nonlinear datasets. It attempts to set the error within a certain threshold and minimize the generalization error bound. Kernel functions in SVR such as linear, poly, and radial basis functions (RBF) help to find a higher dimensional space for a nonlinear dataset problem in which turns the dataset related to linear regression in that space.

GPR is a nonparametric kernel-based probabilistic model with a finite collection of random variables. It is also a powerful predictive tool for data that is highly non-linear. Several different kernel functions such as rational quadratic, Matérn, squared exponential, and periodic kernels, each with unique properties and characteristics, can be used when fitting the model [16].

The most widely established machine learning-based technique is ANN and is a very effective method for complex and nonlinear systems. It is a highly robust and sophisticated technique to capture the general trend of the complex input and output variables. Typically, ANN includes an input layer, some hidden layers, and an output layer [17]. Each layer consists of a number of neurons and the hidden and output-layer neurons are each linked to the neurons in the previous layer.

The NARX model has been used in DC for different applications [18, 19]. It is a neural network with connections from both system inputs and feedbacks from outputs to model the nonlinearity in a DC. NARX is advantageous in modeling time-series data since the model (1) is better at discovering long time dependences, (2) is more effective at learning, (3) has faster convergence, (4) has negligible computational complexity, and (4) has scalability, making it applicable for large DCs [20-23].

In this study, ANN, SVR, and GPR techniques are used for pressure forecasting. It was found that SVR and GPR algorithms are less accurate than ANN. Because SVR and GPR typically need separate models for each point prediction and then combine them to obtain a complete profile which is costly in comparison with using a single ANN model [48]. Besides, SVR and GPR algorithms require some prior knowledge about the system and input-output relationship to have precise prediction which is difficult to get the relationship due to the nonlinearity and complexity of the airflow distribution in a DC. While ANN finds the relationship of the system and makes more accurate predictions.

For the transient case study, NARX is an appropriate model. However, it ignores important facets of the flow physics and heat transfer that can lead to large prediction errors in extrapolative predictions. To address this deficiency, a gray-box model is introduced that combines machine learning with the thermofluid transport equations relevant to predict transient temperatures. Table 1-1 provides a qualitative comparison of some of the distinctive features of the four machine learning techniques used in this study.

1.4 System fault detection

Common failures in electronic products can be traced back to thermal-related issues. The lifespan of electronic components in a DC depends on the environment temperature and is shortened significantly at high temperatures. Pumps and fans in the cooling systems are widely used to create cold airflow in electronic products for cooling purposes. Since they are critical for thermal management in electronic products, the reliability of electronic products is heavily dependent on the pump and fan reliability [24-27].

Table 1-1. Summary of DDM methods tested.

Feature	ANN	SVR	GPR	NARX
Input data	Non-time series data	Non-time series data	Non-time series data	Time series data
Output prediction space	Handle multiple output points in a single model	Cannot[a] handle multiple output points in a single model	Cannot handle multiple output points in a single model	Handle multiple output points in a single model
	Non-kernel based method	kernel-based method[b]	Kernel-based method	Non-kernel based method
Uncertainty	Deterministic	Deterministic	Stochastic	Deterministic

[a] SVR and GPR typically require separate models for each point prediction.
[b] Kernel based methods need some prior knowledge about the system and input-output relationship.

Anomalies are often the result of exceptional system conditions and do not describe the common functioning of the underlying system. Fast anomaly detection is one of the key requirements for economical and optimal process operation management. Many neural network models have been developed to detect faults in a system and shown to be highly successful. One-class support vector machine (OCSVM) and Nonlinear AutoRegressive Exogenous (NARX) techniques are two different fault detection techniques that have been used for different applications [28]. OCSVM is a special variant of the general support vector machine (SVM) and only uses the normal operation data for training. It constructs the tightest decision boundary that encloses all data with minimal slacks. If a new sample locates within the boundary, it is classified as a normal operation point; otherwise, it is labeled as an abnormality. Since no faulty data is needed for training, the OCSVM can be trained easily and has been applied widely for fault detection [29].

NARX is a popular machine learning algorithm that characterizes complex nonlinear mappings between the input and output time-series data. A NARX network with

embedded memory (tapped delay line) can be utilized to detect faults in a system [30, 31]. First, a NARX network is trained and is used to predict target features given past inputs. If the distance between the predicted and actual values exceeds a threshold over several consecutive data samples, an anomaly is detected.

1.5 System fault diagnosis

The fault type can be determined using different DDMs. The data-driven functional models can classify the value of the response variable, with respect to the different failures. Three common DDMs for fault diagnosis are ANN, convolutional neural network (CNN), and recurrent neural network (RNN).

CNN is a feedforward neural network with a set of non-linear transformation functions that has presented excellent success and high performance to solve many classification problems [32, 33]. Convolutional networks sometimes offer some significant advantages over conventional neural networks, especially when it comes to classification problems. CNN transforms the input into a form that is easier to process, while still retaining the essential features. The crucial features are extracted by applying the convolutional filter on the initial inputs where the redundant data are eliminated. This, in turn, decreases the execution time of the algorithm. Next, a pooling layer is applied where the spatial size of the convoluted features will be attempted to be reduced. Finally, the output of the pooling layer needs to be flattened to be used in a regular neural network.

The third type of data-driven methodology for fault diagnosis is RNN [34]. RNNs possess connections that have loops, feedback, and memory to the networks over time. This memory allows this type of network to learn and generalize across sequences of inputs

rather than individual patterns. A state of the art RNN is Long Short-Term Memory (LSTM) which has shown a better performance than a vanilla RNN [35]. LSTM is trained using Backpropagation Through Time (BPTT) and overcomes the vanishing gradient problem.

1.6 Performance metrics

There is a need to evaluate the performance of the different prediction models using criteria such as accuracy rate, error rate, precision, and recall. The accuracy rate is the percentage of correct classifications while the error rate is the percentage of incorrect classifications. Precision and recall metrics are two important metrics to assess the performance of the classifiers. Precision represents the portion of positive samples that were correctly classified to the total number of positive predicted samples and recall determines the positive correctly classified samples to the total number of positive samples. By combining these two terms, a new metric can be obtained to evaluate the performance of the classifiers which is called F1-Score. F1-Score is the harmonic mean of precision and recall.

Chapter 2

Literature review

2.1 Thermal modeling and temperature prediction of DCs

DC designers require an accurate temperature prediction model for energy-efficient real-time management of computing infrastructure. It is important to examine and evaluate DC thermal models before their implementation in high power density computer rooms [36]. There are three main methods for predicting temperatures in a DC, including (1) white-box [37-42], (2) black-box [43-48], and (3) gray-box models [49-54].

White-box, or physics-based, models are based on an understanding of physical laws and the underlying engineering principles. A white-box model based on computational fluid dynamics (CFD) simulations is time-consuming and very expensive [55, 56]. In this technique, numerical methods are used to solve the differential equations which extract the thermal dynamics of the DC environment. A comprehensive number of boundary conditions and parameters need to be considered for both the servers and DC room, such as servers and cooling units airflow rates, DC room dimensions, etc. A series of DC thermal simulations that used CFD are reviewed by Rambo and Joshi [57]. This method is not flexible to changes in the DC, i.e. servers location and statue (on-off) and models cannot be re-run for each change due to high computational cost.

Black-box models are the mathematical functional relationship between system inputs and outputs to predict system operations but without an understanding of the underlying physical and thermodynamics principles. They have enough accuracy if training data are numerous. Black-box models are used to obtain fast interpolative temperature predictions in DCs, however, they have poor accuracy for extrapolative datasets.

A gray-box method is a combination of white-box and black-box models which includes some aspects of the system physics. Therefore, extrapolative prediction errors are reduced below those of black-box models. A 2D hybrid thermal model is proposed in [51] to predict the temperatures around the servers in a DC. Here, the authors considered the first law of thermodynamics, as well as sensor observations with the auto-regression model. Such an approach can be trained using airflow measurements at the front, or cold ends, of servers. However, it is not practical in a DC due to the complexities associated with measurements and the negligence of hot air recirculation. Another airflow and temperature prediction tool has implemented 3D zonal modeling in [41] but utilizes the zones that are too large to accurately predict temperatures at server inlets. The model also requires airflows that must be obtained for each prediction through computationally expensive CFD simulations.

Available temperature prediction methods suffer from at least one of the following limitations. (1) They are not generic models applicable for several configurations. (2) These prediction algorithms are usually inappropriate for transient operation. (3) The computational time they require can be of the order of several minutes or even hours,

making the models unsuitable for real-time applications. (4) Temperature predictions are only available over short durations and not until steady-state conditions are reached. (5) Comprehensive effects of all important operating conditions, such as cooling unit set-point, airflow, and server workload, are not included. (6) Finally, the methods generally ignore important aspects of flow physics and heat transfer.

2.2 Fault detection and diagnosis of a DC

Failure in the cooling system reduces IT equipment lifetime, the reliability of the DC and increases economic losses. DC designers try to increase system reliability by adding cooling units which leads to extra cost. Smart system monitoring for fault detection and diagnosis (FDD) can be used to detect and diagnose upcoming failures and schedule maintenance actions.

FDD algorithms can be classified into two categories, (1) independent and (2) simultaneous FDD [58, 59]. An independent FDD analyses only one fault type at a time, while the simultaneous FDD can detect and diagnose two or more mutually exclusive faults occurring concurrently.

Studies on independent and simultaneous FDD methodology of cooling systems can be separated into two categories, namely, model-based and data-driven [60, 61]. In a model-based FDD, a physics-based model or semi-empirical mechanistic representation of the cooling system is established to simulate the dynamic behavior of the system under the normal operation condition. Next, the distance between the system and the mentioned model is calculated. Finally, a residual analysis of the distance is conducted to determine

if any fault exists in the system. The literature includes numerous instances that use mechanistic physical models to diagnose some common faults. Some popular model-based FDD techniques include symbolic time-series analysis, interacting multi-model, smooth variable state space, cross wavelet transform, and multi-modal decomposition [62]. Despite the existence of several model-based FDD strategies, it is often challenging to establish accurate physics-based representation to simulate the anomalous behavior of dynamic systems in real-time. Furthermore, these methodologies are often prohibitively computation-intensive, limiting their implementation in control systems [63]. Therefore, data-driven FDD methods have attracted increasing attention.

Data-driven approaches that currently dominate air-conditioning FDD literature do not require considerable model knowledge. By collecting a certain amount of data identifying the essential features of the dynamic system, it can learn fault patterns from historical information, thereby demonstrating the ability to predict faults [64]. Several types of time-series signals such as (a) acoustic, (b) vibration, and (c) electrical signals have been used for FDD of an air-conditioning system. However, these methodologies have salient drawbacks such as (1) low signal to noise ratio, (2) costly data acquisition system for high-frequency mechanical or electrical measurement, (3) availability of single-point contact measurement for each component, and (4) high computational requirement for transforming large time-domain signals to the frequency domain in real-time [65, 66]. These drawbacks of the aforementioned techniques are overcome by obtaining real-time spatial thermal measurements using temperature probes due to their ease of installation in DCs, cost-effectiveness, and low computational post-processing requirements. This

dynamic thermal information is used in popular data driven FDD algorithms such as principal component analysis (PCA), artificial neural networks (ANN), support vector machines (SVM), and combinations of these techniques to identify cooling system faults in DC [67, 68].

There have been several attempts in the FDD literature to detect a single fault of the air-conditioning systems. However, literature on developing algorithms accurately to detect two or more simultaneously occurring faults is relatively sparse. Different faults can occur simultaneously in many real applications, and cooling units in DCs are no exception [69]. The main challenge in simultaneous FDD for the cooling cycle in DCs is that the number of combinations of multiple independent faults is large, thereby resulting in numerous possible categories of simultaneous fault training patterns. Therefore, the acquisition of large-scale datasets for simultaneous faults is difficult and expensive. To the best of our knowledge, only a few previous investigations aimed to study simultaneous FDD in the cooling systems [70, 71].

Existing FDD methods suffer from one or more limitations such as, (1) the need for historical time series of simultaneous fault data for model training, (2) requirement of an experimental set up for data generation, which may affect the health and productivity of the system, especially when simultaneous faults occur, (3) establishing an accurate systematic physical model is difficult because the real systems become increasingly complex which challenges the implementation of an FDD method based on the physical method, and (4) the computational time they require can be of the order of several minutes or even hours, making the models unsuitable for real-time applications.

2.3 References

[1] Y. Li, X. Wang, P. Luo, and Q. Pan, "Thermal-aware hybrid workload management in a green datacenter towards renewable energy utilization," Energies, vol. 12, no. 8, p. 1494, 2019.

[2] S. MirhoseiniNejad, H. Moazamigoodarzi, G. Badawy, and D. G. Down, "Joint data center cooling and workload management: A thermal-aware approach," Future Generation Computer Systems, vol. 104, pp. 174-186, 2020.

[3] D. Andrews and B. Whitehead, "Data Centres in 2030: Comparative Case Studies that Illustrate the Potential of the Design for the Circular Economy as an Enabler of Sustainability," in Sustainable Innovation 2019: 22nd International Conference Road to 2030: Sustainability, Business Models, Innovation and Design, 2019.

[4] Shehabi A, Smith SJ, Masanet E, Koomey J. Data center growth in the United States: decoupling the demand for services from electricity use. Environmental Research Letters. 2018 Dec 18;13(12):124030.

[5] Ebrahimi K, Jones GF, Fleischer AS. A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. Renewable and Sustainable Energy Reviews. 2014 Mar 1;31:622-38.

[6] M. Salim and R. Tozer, "Data Centers' Energy Auditing and Benchmarking-Progress Update," ASHRAE transactions, vol. 116, no. 1, 2010.

[7] H. Lu, Z. Zhang, and L. Yang, "A review on airflow distribution and management in data center," Energy and Buildings, vol. 179, pp. 264-277, 2018.

[8] A. Carbó, E. Oró, J. Salom, M. Canuto, M. Macías, and J. Guitart, "Experimental and numerical analysis for potential heat reuse in liquid cooled data centres," Energy Conversion and Management, vol. 112, pp. 135-145, 2016.

[9] T. Gao, M. David, J. Geer, R. Schmidt, and B. Sammakia, "Experimental and numerical dynamic investigation of an energy efficient liquid cooled chiller-less data center test facility," Energy and buildings, vol. 91, pp. 83-96, 2015.

[10] T. Gao, M. David, J. Geer, R. Schmidt, and B. Sammakia, "A dynamic model of failure scenarios of the dry cooler in a liquid cooled chiller-less data center," in 2015 31st Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), 2015, pp. 113-119: IEEE.

[11] J. Dai, M. M. Ohadi, D. Das, and M. G. Pecht, OPTIMUM COOLING OF DATA CENTERS. Springer, 2016.

[12] H. Moazamigoodarzi, R. Gupta, S. Pal, P. J. Tsai, S. Ghosh, and I. K. Puri, "Modeling temperature distribution and power consumption in IT server enclosures with row-based cooling architectures," Applied Energy, vol. 261, p. 114355, 2020.

[13] K. Dunlap and N. Rasmussen, "Choosing between room, row, and rack-based cooling for data centers," APC White Paper, vol. 130, 2012.

[14] T. Evans, "The different types of air conditioning equipment for IT environments," White Paper, vol. 59, pp. 2004-0, 2004.

[15] Moazamigoodarzi H. DISTRIBUTED COOLING FOR DATA CENTERS: BENEFITS, PERFORMANCE EVALUATION AND PREDICTION TOOLS (Doctoral dissertation).

[16] Asgari S, Moazamigoodarzi H, Tsai PJ, Pal S, Zheng R, Badawy G, Puri IK. Hybrid surrogate model for online temperature and pressure predictions in data centers. Future Generation Computer Systems. 2020;114:531-47.

[17] K. Mehrotra, C. K. Mohan, and S. Ranka, Elements of artificial neural networks. MIT press, 1997.

[18] S. MirhoseiniNejad, F. M. García, G. Badawy, and D. G. Down, "ALTM: Adaptive learning-based thermal model for temperature predictions in data centers," in 2019 IEEE Sustainability through ICT Summit (StICT), 2019, pp. 1-6: IEEE.

[19] A. Di Piazza, M. C. Di Piazza, and G. Vitale, "Solar and wind forecasting by NARX neural networks," Renewable Energy and Environmental Sustainability, vol. 1, p. 39, 2016.

[20] E. Diaconescu, "The use of NARX neural networks to predict chaotic time series," Wseas Transactions on computer research, vol. 3, no. 3, pp. 182-191, 2008.

[21] H. Xie, H. Tang, and Y.-H. Liao, "Time series prediction based on NARX neural networks: An advanced approach," in 2009 International conference on machine learning and cybernetics, 2009, vol. 3, pp. 1275-1279: IEEE.

[22] S. M. Guzman, J. O. Paz, and M. L. M. Tagert, "The use of NARX neural networks to forecast daily groundwater levels," Water resources management, vol. 31, no. 5, pp. 1591-1603, 2017.

[23] J. M. P. Menezes Jr and G. A. Barreto, "Long-term time series prediction with the NARX network: An empirical evaluation," Neurocomputing, vol. 71, no. 16-18, pp. 3335-3343, 2008.

[24] X. Tian, "Cooling fan reliability: failure criteria, accelerated life testing, modeling and qualification," in RAMS'06. Annual Reliability and Maintainability Symposium, 2006., 2006, pp. 380-384: IEEE.

[25] X. Jin, E. W. Ma, T. W. Chow, and M. Pecht, "An investigation into fan reliability," in Proceedings of the IEEE 2012 Prognostics and System Health Management Conference (PHM-2012 Beijing), 2012, pp. 1-7: IEEE.

[26] X.-q. Wen and L.-r. You, "A residual lifetime prediction method of cooling fan based on the operating point offset distance," in 2016 Chinese Control and Decision Conference (CCDC), 2016, pp. 2972-2976: IEEE.

[27] R. Fezai, K. Abodayeh, M. Mansouri, H. Nounou, and M. Nounou, "Fault diagnosis of biological systems using improved machine learning technique," International Journal of Machine Learning and Cybernetics, pp. 1-14, 2020.

[28] Asgari S, MirhoseiniNejad S, Moazamigoodarzi H, Gupta R, Zheng R, Puri IK. A gray-box model for real-time transient temperature predictions in data centers. Applied Thermal Engineering. 2020 Nov 13:116319.

[29] S. Yin, X. Zhu, and C. Jing, "Fault detection based on a robust one class support vector machine," Neurocomputing, vol. 145, pp. 263-268, 2014.

[30] Bangalore P, Letzgus S, Karlsson D, Patriksson M. An artificial neural network-based condition monitoring method for wind turbines, with application to the monitoring of the gearbox. Wind Energy. 2017 Aug;20(8):1421-38.

[31] Li Y, Wang X, Luo P, Pan Q. Thermal-aware hybrid workload management in a green datacenter towards renewable energy utilization. Energies. 2019 Jan;12(8):1494.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436-444, 2015.

[33] S. Li, G. Liu, X. Tang, J. Lu, and J. Hu, "An ensemble deep convolutional neural network model with improved DS evidence fusion for bearing fault diagnosis," Sensors, vol. 17, no. 8, p. 1729, 2017.

[34] J. T. Connor, R. D. Martin, and L. E. Atlas, "Recurrent neural networks and robust time series prediction," IEEE transactions on neural networks, vol. 5, no. 2, pp. 240-254, 1994.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[36] C. D. Patel, R. Sharma, C. E. Bash, and A. Beitelmal, "Thermal considerations in cooling large scale high compute density data centers," in 8th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), pp. 767–776, 2002.

[37] R. K. Sharma, C. E. Bash, C. D. Patel, R. J. Friedrich, and J. S. Chase, "Balance of power: Dynamic thermal management for internet data centers," IEEE Internet Computing, vol. 9, no. 1, pp. 42-49, 2005.

[38] Q. Tang, T. Mukherjee, S. K. Gupta, and P. Cayton, "Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters," in 2006 Fourth International Conference on Intelligent Sensing and Information Processing, 2006, pp. 203-208: IEEE.

[39] H. S. Erden, H. E. Khalifa, and R. R. Schmidt, "A hybrid lumped capacitance-CFD model for the simulation of data center transients," Hvac&R Research, vol. 20, no. 6, pp. 688-702, 2014.

[40] H. Moazamigoodarzi, S. Pal, S. Ghosh, and I. K. Puri, "Real-time temperature predictions in it server enclosures," International Journal of Heat and Mass Transfer, vol. 127, pp. 890-900, 2018.

[41] Z. Song, B. T. Murray, and B. Sammakia, "A compact thermal model for data center analysis using the zonal method," Numerical Heat Transfer, Part A: Applications, vol. 64, no. 5, pp. 361-377, 2013.

[42] R. Zhou, Z. Wang, C. E. Bash, and A. McReynolds, "Data center cooling management and analysis-a model based approach," in 2012 28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2012, pp. 98-103: IEEE.

[43] J. Athavale, Y. Joshi, and M. Yoda, "Artificial neural network based prediction of temperature and flow profile in data centers," in 2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2018, pp. 871-880: IEEE.

[44] J. Moore, J. S. Chase, and P. Ranganathan, "Weatherman: Automated, online and predictive thermal mapping and management for data centers," in 2006 IEEE international conference on Autonomic Computing, 2006, pp. 155-164: IEEE.

[45] M. Zapater, J. L. Risco-Martín, P. Arroba, J. L. Ayala, J. M. Moya, and R. Hermida, "Runtime data center temperature prediction using Grammatical Evolution techniques," Applied Soft Computing, vol. 49, pp. 94-107, 2016.

[46] L. Wang, G. von Laszewski, F. Huang, J. Dayal, T. Frulani, and G. Fox, "Task scheduling with ANN-based temperature prediction in a data center: a simulation-based study," Engineering with Computers, vol. 27, no. 4, pp. 381-391, 2011.

[47] R. Lloyd and M. Rebow, "Data driven prediction model (ddpm) for server inlet temperature prediction in raised-floor data centers," in 2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2018, pp. 716-725: IEEE.

[48] J. Athavale, M. Yoda, and Y. Joshi, "Comparison of data driven modeling approaches for temperature prediction in data centers," International Journal of Heat and Mass Transfer, vol. 135, pp. 1039-1052, 2019.

[49] Z. Song, B. T. Murray, and B. Sammakia, "A dynamic compact thermal model for data center analysis and control using the zonal method and artificial neural networks," Applied thermal engineering, vol. 62, no. 1, pp. 48-57, 2014.

[50] L. Li, C.-J. M. Liang, J. Liu, S. Nath, A. Terzis, and C. Faloutsos, "Thermocast: a cyber-physical forecasting model for datacenters," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 1370-1378.

[51] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, "A cyber–physical systems approach to data center modeling and control for energy efficiency," Proceedings of the IEEE, vol. 100, no. 1, pp. 254-268, 2011

[52] Chen J, Tan R, Wang Y, Xing G, Wang X, Wang X, Punch B, Colbry D. A high-fidelity temperature distribution forecasting system for data centers. In2012 IEEE 33rd Real-Time Systems Symposium 2012 Dec 4 (pp. 215-224). IEEE.

[53] L. Parolini, B. Sinopoli, and B. H. Krogh, "Model predictive control of data centers in the smart grid scenario," IFAC Proceedings Volumes, vol. 44, no. 1, pp. 10505-10510, 2011.

[54] E. Pakbaznia, M. Ghasemazar, and M. Pedram, "Temperature-aware dynamic resource provisioning in a power-optimized datacenter," in 2010 Design,

Automation & Test in Europe Conference & Exhibition (DATE 2010), 2010, pp. 124-129: IEEE.

[55] Nada, S., M. Said, and M. Rady, Numerical investigation and parametric study for thermal and energy management enhancements in data centers' buildings. Applied Thermal Engineering, 2016. **98**: p. 110-128.

[56] Chen, J., et al. A high-fidelity temperature distribution forecasting system for data centers. in 2012 IEEE 33rd Real-Time Systems Symposium. 2012. IEEE.

[57] J. Rambo and Y. Joshi, "Reduced-order modeling of turbulent forced convection with parametric conditions," International Journal of Heat and Mass Transfer, vol. 50, no. 3-4, pp. 539–551, 2007. Bibliography 107.

[58] Z. Zhang, S. Li, Y. Xiao, and Y. Yang, "Intelligent simultaneous fault diagnosis for solid oxide fuel cell system based on deep learning," Applied Energy, vol. 233, pp. 930-942, 2019.

[59] H. Li and J. E. Braun, "A methodology for diagnosing multiple simultaneous faults in vapor-compression air conditioners," HVAC&R Research, vol. 13, no. 2, pp. 369-395, 2007.

[60] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods," Computers & chemical engineering, vol. 27, no. 3, pp. 293-311, 2003.

[61] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies," Computers & chemical engineering, vol. 27, no. 3, pp. 313-326, 2003.

[62] Y. Yu, D. Woradechjumroen, and D. Yu, "A review of fault detection and diagnosis methodologies on air-handling units," Energy and Buildings, vol. 82, pp. 550-562, 2014.

[63] A. Beghi, R. Brignoli, L. Cecchinato, G. Menegazzo, M. Rampazzo, and F. Simmini, "Data-driven fault detection and diagnosis for HVAC water chillers," Control Engineering Practice, vol. 53, pp. 79-91, 2016.

[64] K. Yan, W. Shen, T. Mulumba, and A. Afshari, "ARX model based fault detection and diagnosis for chillers using support vector machines," Energy and Buildings, vol. 81, pp. 287-295, 2014.

[65] Gong CS, Su CH, Chuang YC, Tseng KH, Li TH, Chang CH, Huang LH. Feature extraction of rotating apparatus using acoustic sensing technology. In2019 Eleventh

International Conference on Ubiquitous and Future Networks (ICUFN) 2019 Jul 2 (pp. 254-256). IEEE.

[66] N. Baydar and A. Ball, "Detection of gear failures via vibration and acoustic signals using wavelet transform," Mechanical Systems and Signal Processing, vol. 17, no. 4, pp. 787-804, 2003.

[67] D. Li, Y. Zhou, G. Hu, and C. J. Spanos, "Fault detection and diagnosis for building cooling system with a tree-structured learning method," Energy and Buildings, vol. 127, pp. 540-551, 2016.

[68] S. Wang and F. Xiao, "AHU sensor fault diagnosis using principal component analysis method," Energy and Buildings, vol. 36, no. 2, pp. 147-160, 2004.

[69] I. Velibeyoglu, H. Y. Noh, and M. Pozzi, "A graphical approach to assess the detectability of multiple simultaneous faults in air handling units," Energy and Buildings, vol. 184, pp. 275-288, 2019.

[70] I. Velibeyoglu, H. Y. Noh, and M. Pozzi, "A graphical approach to assess the detectability of multiple simultaneous faults in air handling units," Energy and Buildings, vol. 184, pp. 275-288, 2019.

[71] A. Piacentino and M. Talamo, "Critical analysis of conventional thermoeconomic approaches to the diagnosis of multiple faults in air conditioning units: Capabilities, drawbacks and improvement directions. A case study for an air-cooled system with 120 kW capacity," International journal of refrigeration, vol. 36, no. 1, pp. 24-44, 2013.

Chapter 3

Problem statement and research objectives

Temperature predictions offer a way to optimize server inlet air temperatures and reduce energy waste from over-cooling. Row-based cooling architectures for DCs have been made available only recently, especially for high-density DCs. Improvement in the thermal performance of DCs, e.g., thermal aware workload management, employing model-based control methods, fault detection, and testing "what if" scenarios to characterize the influence of operating conditions on temperature distribution, require a real-time temperature prediction tool. Developing temperature prediction tools for enclosed DCs with row-based cooling architectures is another undiscovered area for the thermal management of DCs.

The available temperature prediction methods have six main limitations: (1) They are not generic models applicable for several configurations, (2) their prediction algorithms are usually inappropriate for transient operation, (3) the computational time they require can be of the order of several minutes or even hours, making the models unsuitable for

real-time applications, (4) temperature predictions are only available over short durations and not until steady-state conditions are reached, (5) comprehensive effects of all important operating conditions, such as cooling unit set-point, airflow, and server workload, are not included, and (6) the methods generally ignore important aspects of flow physics and heat transfer. Therefore, the first and second objectives are to propose steady-state and transient gray-box zonal models to obtain real-time temperature distributions inside the DC that are confined within an enclosure cooled by row-based cooling units with separated cold and hot chambers.

Studying an application of our three-dimensional gray-box temperature prediction model is essential for demonstrating its applicability. The operation of cooling systems is of the critical importance to maintain a secure, reliable, and stable environment while ensuring energy efficiency and adhering to safety guidelines of computing infrastructures. A survey of over 55,000 air conditioning units revealed that more than 90% had experienced one or more faults. Cooling units operated under faulty conditions in a DC exacerbates its energy consumption and cost, damages the IT equipment while diminishing the computing efficiency. Therefore, fast detection of abnormal behavior of cooling units in a DC is of great significance. Different faults can occur simultaneously in many real applications, and cooling units in DCs are no exception. The main challenge in simultaneous FDD for the cooling cycle in DCs is that the number of combinations of multiple independent faults is large, thereby resulting in numerous possible categories of simultaneous fault training patterns. Therefore, the third objective of this research is to use

the gray-box temperature prediction model for studying simultaneous FDD in the cooling systems.

Chapter 4

Hybrid Surrogate Model for Online Temperature and Pressure Predictions in Data Centers

This chapter is reproduced from *"Hybrid Surrogate Model for Online Temperature and Pressure Predictions in Data Centers"*, ***Sahar Asgari***, *Hosein Moazamigoodarzi, Peiying Jennifer Tsai, Souvik Pal, Rong Zheng, Ghada Badawy and Ishwar K. Puri, Published in Future Generation Computer Systems, 2021.*

The author of this thesis is the first author and the main contributor of this publication. Her main contributions to this work consist of introducing the idea of gray-box model in a row-based cooling DC, writing the manuscript, formulating the problem, conducting the experiments, running CFD simulations, implementing the framework, and generating the numerical results.

4.1 Abstract

The increase in cloud computing and big data storage has led to significant growth in data center (DC) infrastructure that is now estimated to consume more than 1.5% of the world's electricity. Due to suboptimal DC design and operation, a significant fraction of this energy is wasted because of the cooling systems inability to effectively distribute cold air to servers. Consequently, additional cooling air must be circulated inside a DC to prevent local hot spots, which leads to undercooling at other locations. Row-based cooling is an emerging architecture that provides more effective airflow distribution, which lowers energy consumption. Since available methods are unsuitable for accurate online predictions, a general thermal model is required to predict spatiotemporal temperature changes inside a DC and hence optimize airflow distribution for this architecture. Typical approaches include physical models, computational fluid dynamics (CFD) simulations, and black-box data-driven models (DDMs). All three approaches are limited because they do not encapsulate the entirety of relevant operational parameters, are time-consuming and can provide unacceptable errors during extrapolative predictions. We address these deficiencies by developing a fast, adaptive, and accurate hybrid surrogate model by combining a DDM and the thermofluid transport relations to predict temperatures in a DC. Training data for the DDM is obtained from CFD simulations. An artificial neural network (ANN) with the Rectified Linear Unit (ReLU) activation function is shown to predict pressure distributions accurately in a row-based cooling DC. These predicted pressures are inputs for thermofluid transport equations to determine the temperature distribution. The applicability of the model is demonstrated by comparing predictions with experimental

measurements that characterize the influence of varying server workload distribution and cooling unit operational conditions, i.e., temperature set-point, airflow rate, and fan locations, on the temperature distribution. The model can be used to (1) improve cooling configuration design, (2) facilitate thermally aware workload management, and (3) test "what if" scenarios to characterize the influence of operating conditions on the temperature distribution.

Key words: Data center, data-driven models, row-based cooling architecture, temperature prediction, ANN, SVR.

4.2 Introduction

Data centers (DCs) play a critical role in facilitating the digital processes that support our daily lives and economic productivity. Their rapid growth has led to a significant expansion in DC services and facilities [1-3]. The DC industry consumes more than 1.5% of the world's electricity which is estimated to increase by 15-20% annually [4-6]. Although liquids provide considerably higher heat transfer than air, most DCs employ air cooling due to the simplicity of air handling [7-9].

Air cooling systems face two major distribution problems, namely, hot air recirculation and cold air bypass, both of which produce undesirable flow distributions [10-12]. When the cold air supplied to IT equipment is insufficient, hot air exhausted from servers recirculates to the cold chamber where it mixes with cold air, thus raising the rack inlet temperature. Bypass occurs when a portion of the cold airflow returns to the cooling unit without contributing to server cooling. Poor cooling system design and operating

conditions result in inadequate air distribution, thus requiring additional cold air to maintain the IT equipment safely [13-17]. Ineffective cooling is estimated to be responsible for a third of total DC power consumption [5, 8].

To remedy this problem, row-based cooling is an emerging architecture that minimizes hot and cold air mixing, providing better cold air distribution [18]. Optimizing the thermal performance of this architecture requires a model that can predict spatiotemporal temperature variations.

The literature contains several approaches to predict the temperature in a DC [19-23]:

1) Simplified physics-based models that are fast computationally, but insufficiently adaptive to physical changes within a DC. Due to simplifying assumptions, they have relatively poor accuracy [24-28].
2) Computational fluid dynamics (CFD) simulations that provide temperature and airflow distributions with high precision, but are computationally expensive, particularly for large DCs [29-33].
3) Data-driven modeling (DDM) methods that provide fast predictions are simple to implement and capable of approximating complex functional relationships [34-38].

Training data for a DDM is usually obtained from either CFD simulations or experiments. The model is then trained to represent the relations among system state variables (input, internal, and output) [39, 40]. DDMs are classified as either black-box (e.g., with no knowledge of thermodynamics laws) or gray-box (built with partial

knowledge of thermodynamics laws) [34, 41]. Black-box DDMs are extensively used to provide fast temperature predictions in DCs [42, 43]. They are computationally inexpensive. However, since black-box DDMs ignore flow physics and thermodynamics, their accuracy for the extrapolative prediction can be inadequate when there are minor changes in cooling configurations or IT equipment [24, 44].

Gray-box DDMs are built with partially understood physical processes and combined with data-driven approximations to predict air temperatures at discrete locations, such as server inlets and outlets [45-47]. Even though existing gray-box DDMs for DC temperature predictions include some physics, they failed to characterize important phenomena, such as hot air recirculation, which can lead to significant prediction error. Furthermore, most gray-box DDMs reported in the literature employ regression, which is inappropriate for a DC due to the complexity and nonlinearity of the governing equations [48].

While available methods can predict temperatures, (1) the computational time that they require is on the order of several minutes to hours, which is unsuitable for real-time applications, (2) they do not include the effects of important operating conditions, such as cooling unit set-point, airflow, and server workload and location, (3) they cannot adapt reasonably to configuration changes, such as the locations of the cooling unit fans, and (4) they usually ignore important facets of flow physics and heat transfer.

Here, we propose a more general, accurate, and fast surrogate model which combines fundamental thermofluid relations with data-driven solutions to make on-the-fly

predictions of the steady-state pressures, airflow, and temperature distributions in an enclosed DC that utilizes a row-based cooling architecture. CFD simulations are used to generate the training dataset. Three machine learning algorithms, Artificial Neural Network (ANN), Support Vector Regression (SVR), and Gaussian Process Regression (GPR) are employed in conjunction with a 3D zonal model. The applicability of the proposed method is demonstrated by investigating the effect of (1) workload distribution, (2) operating parameters of the cooling units, (3) server placements, and (4) locations of cooling unit fans. The results show that the maximum temperature prediction error is 2.7 ^{o}C and computation time is less than 4 seconds. In summary, the major contributions of this study are:

- Integrating DDMs and physics-based relations to predict DC temperatures.
- Introducing a very computationally efficient and accurate 3D temperature prediction model.
- Investigating the effects of different server workloads and cooling conditions on temperatures.
- Providing a surrogate model that is adaptive to changes in the locations and status of cooling unit fans and server utilization.

The remainder of this study is organized as follows. Section 2 discusses some of the similar researches done in the past. Section3 introduces the proposed surrogate model and its framework. Section 4 provides temperature profiles and evaluates the predictions. Finally, section 5 summarizes the findings.

4.3 Related work

The literature contains numerous methods to predict DC temperatures that can be broadly divided into three categories, i.e., (1) physics-based models, (2) CFD simulations, and (3) data-driven models.

With judicious simplifications of rigorous physical laws, physics-based models are able to predict air temperatures at discrete locations in a DC, such as server inlets and outlets. An example is the lumped-capacitance mathematical model for predicting server inlet and outlet air temperatures [49]. In [49], the server thermal capacitance and effectiveness are determined from air temperature measurements at server inlets and outlets. A thermodynamics-based lumped capacitance model can provide very rapid predictions but with limited spatial information. Such a model is therefore unable to provide the fine-grained local data required to ensure the reliable operation of every server. Such a limitation can be overcome by a zonal method that is an intermediate approach between full CFD simulations and a multi-node lumped model [45]. In zonal methods, a DC is partitioned into a number of characteristic zones to which fundamental conservation laws are applied to predict zonal airflows and temperatures. A physics-based zonal model based on mass and energy conservation relations for each zone within the enclosure can predict real-time temperatures inside a DC [24]. Although these physics-based models are computationally fast, their accuracy is limited due to simplifying assumptions.

CFD simulations of DCs can predict local temperatures, airflows and pressures, characterize the influence of power density and Computer Room Air Conditioning (CRAC)

location on DC performance [29, 50]. However, because of its long execution time, a CFD simulation has limited utility for simulating a medium to large size DC that contains hundreds of racks and thousands of servers. Since thermal outages require immediate actions for safe DC operations, faster ways of improving thermal management must be identified.

The literature includes several works that use DDMs, classified as either black-box or gray-box models, to predict the temperature distribution in a DC. Black box approaches relate outputs, e.g., temperatures, to inputs through equations that ignore the flow physics, where training data can be obtained from an experimentally validated CFD model. While the interpolative prediction errors from various DDMs are typically low, extrapolative prediction errors tend to be much larger [34]. An alternative is offered by an adaptive learning-based thermal model that employs a black-box to predict the temperatures of critical zones using DC operation variables as inputs [42].

In contrast, a gray-box method includes some aspects of the system physics to predict temperatures so that extrapolative prediction errors are reduced below those of black-box models. A 2D hybrid approach that considers the first law of thermodynamics, as well as sensor observations, can be used with auto-regression to predict DC temperatures [46]. Such a model can be trained using airflow measurements at the front, or cold ends, of servers. But it is not practical in a DC due to the complexities associated with measurements and the negligence of hot air recirculation. Another airflow and temperature prediction tool has implemented 3D zonal modeling in [45] but utilizes the zones that were too large to accurately predict temperatures at server inlets. The associated model also

requires airflow rates for each prediction that must be obtained through computationally expensive CFD simulations.

To overcome these predictive challenges, for the first time for a DC environment, we develop a 3D gray-box zonal model that predicts pressures, airflows, and temperatures, which accounts for air recirculation and is computationally efficient.

4.4 Methodology

A surrogate thermal model is created using the airflow momentum, mass, and energy balance equations, where unknown parameters are estimated using DDMs, as depicted in Figure 4-1. First, CFD simulations are validated using experimental measurements for different cooling airflows. Then, three DDMs (ANN, SVR, and GPR) are compared to determine the more appropriate algorithm, which is then trained to predict pressures. Next, the predicted pressures are included in the momentum, mass, and energy relations to predict the airflow and temperatures. Finally, the results from the surrogate model are compared with experiments.

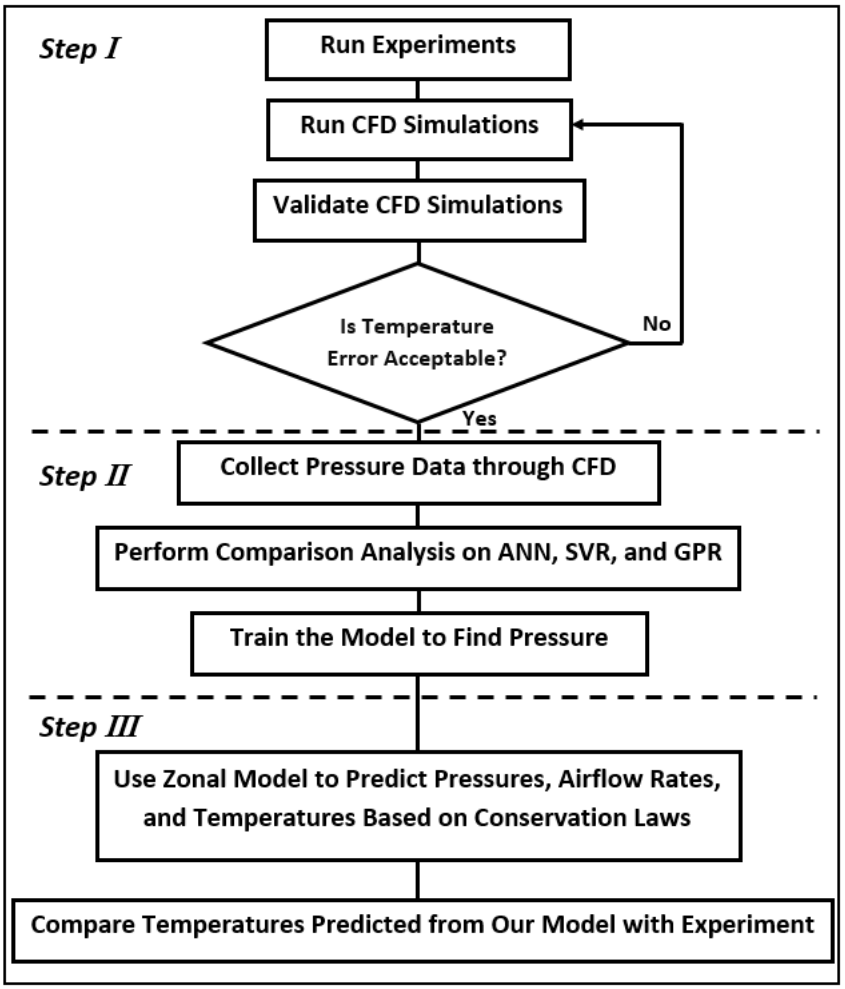


Figure 4-1: Flowchart of the model development for predicting airflow, pressure, and temperature.

4.4.1 In-row cooling architecture and experimental setup

We instrument an in-row cooling modular DC with thermocouples for temperature measurements. Their locations and an airflow schematic are illustrated in Figure 4-2. The DC houses five racks and two in-row cooling units that are placed at the left and right ends of the enclosure. Each cooling unit contains 3 sets of fans. The enclosure is 3.2 m long, 1.4 m wide and 2.05 m high. Cold air from the cooling units into the cold chamber is drawn to

the servers to cool them, where warm air is generated and expelled into the hot chamber and from where it is returned to the cooling units. The racks are partially populated with scattered servers and the empty spaces are blocked with blanking panels. Hot air recirculation and cold air bypass airflows may occur in the cold and hot chambers due to the local pressure differences between these two chambers.

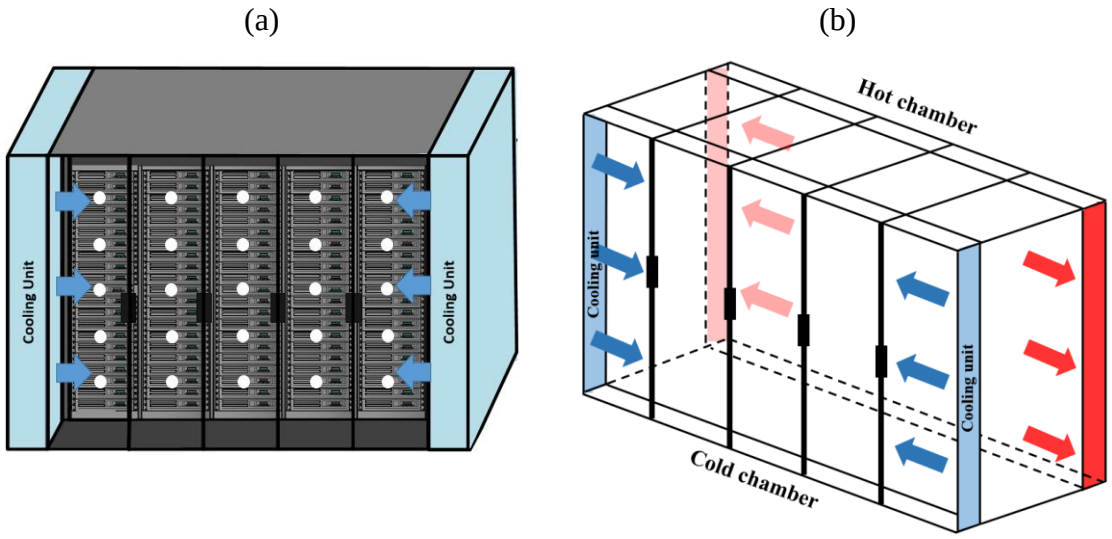


Figure 4-2. Illustration of the experimental row-based cooling DC with 5 racks. (a) Thermocouple locations and (b) airflow schematic. The enclosure is 3.2 m long, 1.4 m wide and 2.05 m high.

4.4.2 Computational fluid dynamics (CFD)

CFD simulations of a row-based cooling DC are performed using ANSYS Fluent 18.0. The flow is simulated using the Reynolds Averaged Navier-Stokes (RANS) model in combination with the standard k-ε turbulent model [51]. For steady-state analysis, the second-order upwind scheme is adapted for the convection term and the semi-implicit method used for the pressure-linked equation (SIMPLE) algorithm. Mesh sensitivity is determined based on the grid convergence index (GCI) for coarse, medium and fine meshes

with 2.6 million, 3.3 million, and 4.4 million nodes, respectively. Based on the GCI, the intermediate mesh is selected for all simulations for which details are provided in the appendix.

A flow field is characterized by mass, momentum, and total energy balances that are described by the continuity, momentum, and energy conservation equations. Solutions to the corresponding mathematical equations provide the local velocities, pressures and temperatures of the fluid in the modeled domain.

The conservation of mass for fluid flow is the continuity equation [52],

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \bar{v}) = 0, \tag{4-1}$$

where ρ is the fluid density, $\bar{v}$ is the fluid velocity, and t is time. Newton's second law of motion is applied to a fluid element that provides the conservation of momentum. When it is applied to fluid flow, the momentum equation [52],

$$\frac{\partial (\rho \bar{v})}{\partial t} + \nabla.(\rho \bar{v} \bar{v}) = -\nabla p + \nabla.(\bar{\bar{\tau}}) + \rho \bar{g} + \bar{F}, \text{ and} \tag{4-2}$$

where p denotes the static pressure, $\bar{\bar{\tau}}$ the stress tensor, $\rho \bar{g}$ and $\bar{F}$ the gravitational body force and external body force. The conservation of energy represents the first law of thermodynamics for a control volume and provides the energy equation [53],

$$\frac{\partial}{\partial t}(\rho E) + \nabla.\left(\bar{v}(\rho E + p)\right) = \nabla.\left(k_{eff} \nabla T - \sum_n h_n \bar{J}_n + \left(\bar{\bar{\tau}}_{eff}.\bar{v}\right)\right) + S_h, \tag{4-3}$$

where E is the total energy, k_{eff} the effective conductivity, h_n the enthalpy of species *n*, $\bar{J}_n$ the diffusion flux of species *n*, and S_h the heat of chemical reaction that is assumed to be zero. Because of its nonlinearity, additional terms arise in the momentum conservation equation that corresponds to turbulent stresses. These additional terms must be related to

the averaged flow variables using a turbulence model. We use the standard k–ε turbulence model, which is a two-equation model that provides a general description of turbulence [54, 55]. The relations for the turbulent kinetic energy k and energy dissipation rate ε are [56],

$$\frac{\partial(\rho k)}{\partial t} + \frac{\partial(\rho k v_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[\frac{\mu_t}{\sigma_k}\frac{\partial k}{\partial x_j}\right] + 2\mu_t E_{ij}E_{ij} - \rho\varepsilon, \text{ and} \quad (4\text{-}4)$$

$$\frac{\partial(\rho \varepsilon)}{\partial t} + \frac{\partial(\rho \varepsilon u_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[\frac{\mu_t}{\sigma_\varepsilon}\frac{\partial \varepsilon}{\partial x_j}\right] + C_{1\varepsilon}\frac{\varepsilon}{K}2\mu_t E_{ij}E_{ij} - C_{2\varepsilon}\rho\frac{\varepsilon^2}{k}, \quad (4\text{-}5)$$

where v_i represents the velocity component in the corresponding direction i, E_{ij} the component of the rate of deformation, μ_t the eddy viscosity, and σ_k, σ_ε, $C_{1\varepsilon}$, and $C_{2\varepsilon}$ are constants. Eq. (4-4) determines the scale of the turbulence, whereas the Eq. (4-5) determines the energy in the turbulence.

The racks of Figure 4-2 are modeled as recirculation boundaries, and the cooling units as mass flow inlets and pressure outlets for the cold air supply and the return air, respectively. The gaps between the racks that can cause air recirculation if not properly sealed, are modeled as porous media using a power-law model [24, 57].

4.4.3 CFD validation

Since a row-based DC is sensitive to cooling unit airflow, three different cases are considered to validate the CFD simulations, i.e., with (1) high ($\dot{m}_{CU} \gg \sum \dot{m}_s$), (2) sufficient ($\dot{m}_{CU} \cong \sum \dot{m}_s$), and (3) low ($\dot{m}_{CU} \ll \sum \dot{m}_s$) flow rates. Due to the emergence of hot spots inside the enclosure, the last of these three cases is the most challenging to maintain equipment integrity during implementation, and therefore it is used to test model robustness.

The temperature differences between CFD predictions and experimental measurements are characterized through $Err = |T_{Exp} - T_{CFD}|$. Figure 4-3 presents *Err* values for the cold chamber at different sensor locations. Values of *Err* are smaller across the cold chamber for the high (< 0.6 ^{o}C) and sufficient (< 1.1 ^{o}C) cooling unit airflows, but there is a larger 1.8 ^{o}C difference in a single zone for the low airflow, which occurs due to hot air recirculation in this region when the pressure is higher in the hot than in the cold chamber. Overall, the values of *Err* are relatively small.

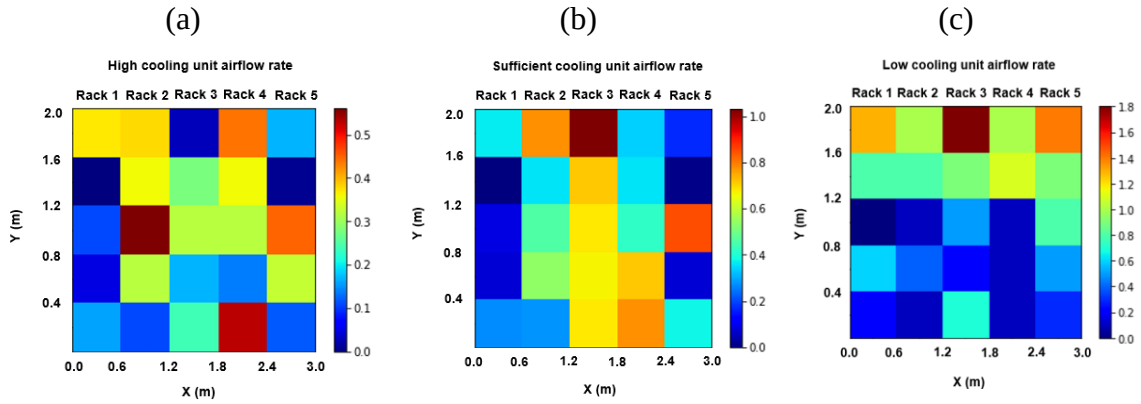


Figure 4-3. Temperature differences between CFD predictions and experimental measurements at the various experimental sensor locations for a) high, b) sufficient, and c) low cooling unit airflows.

4.4.4 Thermal model

CFD simulations of DCs are complex and computationally expensive. Zonal models offer a faster and reasonably accurate alternative. In a zonal model, the DC environment is divided into a coarse grid of zones with the assumption that the conditions inside each zone are spatially uniform. A set of non-linear coupled equations consisting of mass, momentum, and energy conservation equations is applied for each uniform zonal volume [58-60].

$$\sum_j \dot{m}_{j\to i} = 0, \tag{4-6}$$

$$\sum F = \sum_j \left(\dot{m}_{j\to i} v_{j\to i}\right)_{out} - \sum_j \left(\dot{m}_{j\to i} v_{j\to i}\right)_{in}, \text{ and} \tag{4-7}$$

$$\sum_j Q_{j\to i} + Q_{source} = \rho_i V_i c_p \frac{\partial T_i}{\partial t}, \tag{4-8}$$

where $\dot{m}$ denotes the interfacial mass flow rate transferred from cell *j* to cell *i*, F body force, v velocity, ρ density, Q heat flux, Q_{source} internal heat source, c_p specific heat capacity, V_i cell volume, and T_i air temperature. The mass conservation equation (Eq. (4-6)) illustrates that the amount of mass within the control volume remains constant, i.e., it is neither created nor destroyed. The momentum conservation equation (Eq.(4-7)) captures that the momentum can only change through the actions of forces, as described by Newton's laws of motion. The energy conservation equation (Eq. (4-8)) indicates that while energy can be converted from one form to another, the total energy within a control volume remains fixed. Integrated forms of conservation laws (Eqs. (4-6) to (4-8)) predict pressures, temperatures, and mass flowrates.

A schematic of the 3D zonal model for a single rack within an enclosure is represented for a row-based cooling architecture DC in Figure 4-4. A total of 50 zones are created within the cold and hot chambers, where mass flowrates for each zone are obtained by applying Eqs. (4-6), (4-7) and (4-8) for temperature (using PYTHON 3.7).

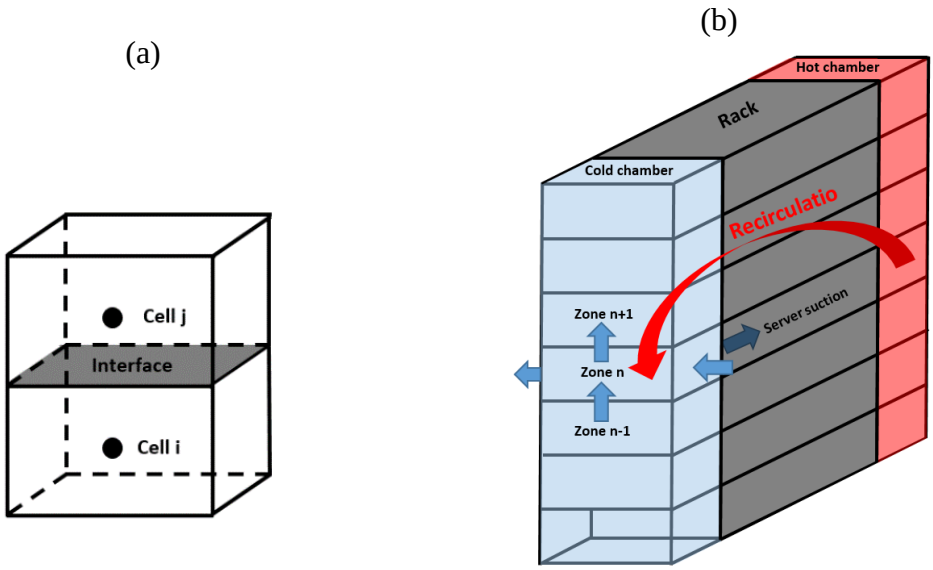


Figure 4-4. Zonal model for (a) cells and interfaces, and (b) the 3D zonal model around a rack.

4.4.5 Data-driven models

Due to the nonlinearity and complexity of the airflow distribution in a DC, DDMs are suitable for representing the multifaceted relationships among the system-state (input, internal, and output) variables. These models can replace CFD simulations and physics-based models, where the most widely established techniques appropriate for DC investigations are SVR, GPR, and ANN [61].

SVR is a regression algorithm suitable for both linear and nonlinear functions that minimizes the generalization error bound subject to error tolerance. Kernel functions in SVR, such as linear, polynomial, and radial basis functions (RBF), help find a higher dimensional representation for the input data, which transform non-linear relationships into linear ones in that space.

GPR is a nonparametric kernel-based probabilistic model with a finite collection of random variables. It is also a powerful predictive tool for data that is highly non-linear. Several different kernel functions such as rational quadratic, Matérn, squared exponential, and periodic kernels, each with unique properties and characteristics, can be used when fitting the model.

The most widely established machine learning-based technique for complex and nonlinear systems is ANN. The technique is highly robust and sophisticated, being able to reproduce the complex general trends for input and output variables. Typically, ANN includes an input layer, some hidden layers and an output layer [62].

Each layer consists of a number of neurons, where the hidden and output-layer neurons are each linked to the neurons in the previous layer. The main challenge with ANN is the choice of model complexity. When the number of parameters is far larger than the available training data, overfitting may happen. Else, unfitting may occur.

Since the DC problem includes nonlinear statistical data, SVR and GPR with non-linear kernel functions, and ANN-based models are compared to determine the more appropriate algorithm. It is worth mentioning that the dataset is generated using experimentally validated CFD simulations, where many realistic scenarios are simulated to provide the input parameter, i.e., cooling unit airflow, and specify the output parameter, which is the static pressure at the different interfaces of a zone.

Table 4-1 lists the independent and dependent variables. The dataset is divided into two portions, 80% of which is used for training and 20% for testing and validation.

Table 4-1. Independent and dependent variables for DDMs.

Independent variable	Range	Dependent variable
Cooling unit airflow rate	0.40 – 2.4 (kg/s)	Static pressures
Cooling unit set-point	18 – 22 (^{o}C)	
Servers workload	0 – 100%	

4.5 Results and discussion

4.5.1 Comparison of the data-driven models

The ANN, GPR, and SVR comparison allows us to select a better algorithm to develop the surrogate model. In order to find the optimum combination of DDM parameters, 5-fold cross-validation is used. The performance of the model is evaluated by comparing the predictions with a set of test data using the root mean square error,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{4-9}$$

where y_i is the observed value, $\hat{y}_i$ predicted value, and *n* the number of samples.

The hyper-parameters of the DDMs are varied. For SVR, three types of kernel functions, linear, polynomial, and RBF are explored, for GPR, four types of kernel functions, rational quadratic, Matérn, squared exponential, and periodic kernels are investigated and for ANN, three different activation functions, Rectified Linear Unit (ReLU), tanh, and logistic are considered. The performance of the RBF kernel function in SVR depends primarily on two important parameters, penalty (C) and Gaussian kernel function (γ), implementation of GPR requires the choice of a suitable kernel function, and

ANN is most sensitive to the number of hidden layers and the number of neurons in each layer.

The comparative analysis for the SVR is presented in Figure 4-5. The minimum RMSE value of the test dataset for the polynomial and RBF kernel functions are 1.9 Pa and 1.5 Pa, respectively, which in the polynomial kernel function occur at 5 degrees and in RBF kernel function at $C = 974$ and $\gamma = 6$.

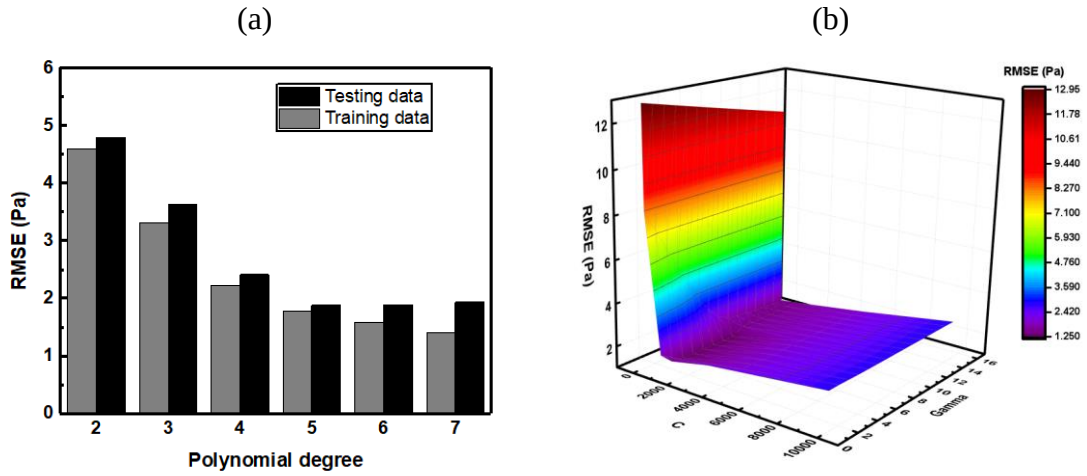


Figure 4-5. Comparison of different kernels of the SVR algorithm. (a) RMSE vs polynomial degree and (b) RBF kernel with a 3D view of RMSE vs C and γ.

Figure 4-6 shows the RMSE values in training and testing using each of the four kernels, i.e., rational quadratic, Matérn, squared exponential, and periodic kernels. We find that the rational quadratic and Matérn kernels reproduce the pressure data more accurately than the squared exponential and periodic kernels. The rational quadratic kernel fits to the data with 1.45 and 1.78 Pa training and testing RMSE, respectively, while the periodic kernel is worse in training as well as testing with 3.14 and 3.49 Pa train and test RMSE, respectively.

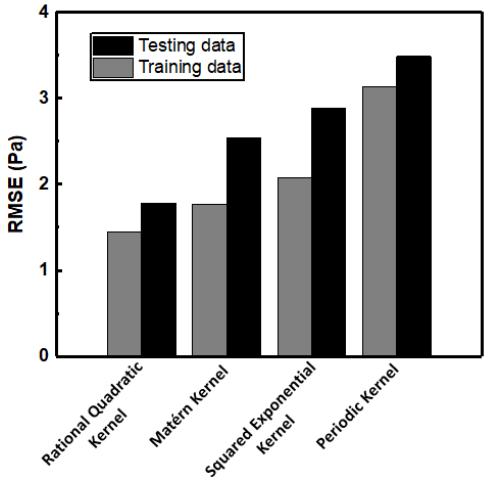


Figure 4-6. GPR algorithm with different kernels: rational quadratic, Matérn, squared exponential, and periodic kernels. All samples have length-scale parameter ℓ=1 which controls how close two points have to be in order to be considered near and thus be highly correlated.

Figure 4-7 provides results for the ANN algorithm with the testing and training datasets. Different activation functions are investigated, where the minimum value of $RMSE_{test}$ is illustrated by the hollow shape. Figure 4-7 (a) and (b) demonstrate the RMSE of the testing and training dataset using the ReLU activation function. The minimum $RMSE_{test}$ value occurs with three hidden layers and 16 neurons. When the tanh activation function is applied, as shown in Figure 4-7 (c) and (d), $RMSE_{test}$ first decreases and then increases due to overfitting. Here, two hidden layers with 9 neurons each yield the minimum $RMSE_{test}$. Similarly, in Figure 4-7 (e) and (f) for the logistic activation function, the minimum $RMSE_{test}$ is obtained with one hidden layer with 6 neurons each.

Table 4-2 summarizes the results of the comparative analysis for GPR, SVR, and ANN. The minimum $RMSE_{test}$ = 0.52 with ANN using the ReLU activation function. Therefore, the ANN with 3 hidden layers and 16 number of neurons is considered to be

more appropriate for our model and thereafter used to predict the pressures. The data flow for the surrogate model used to predict pressures and temperatures is depicted in Figure 4-8.

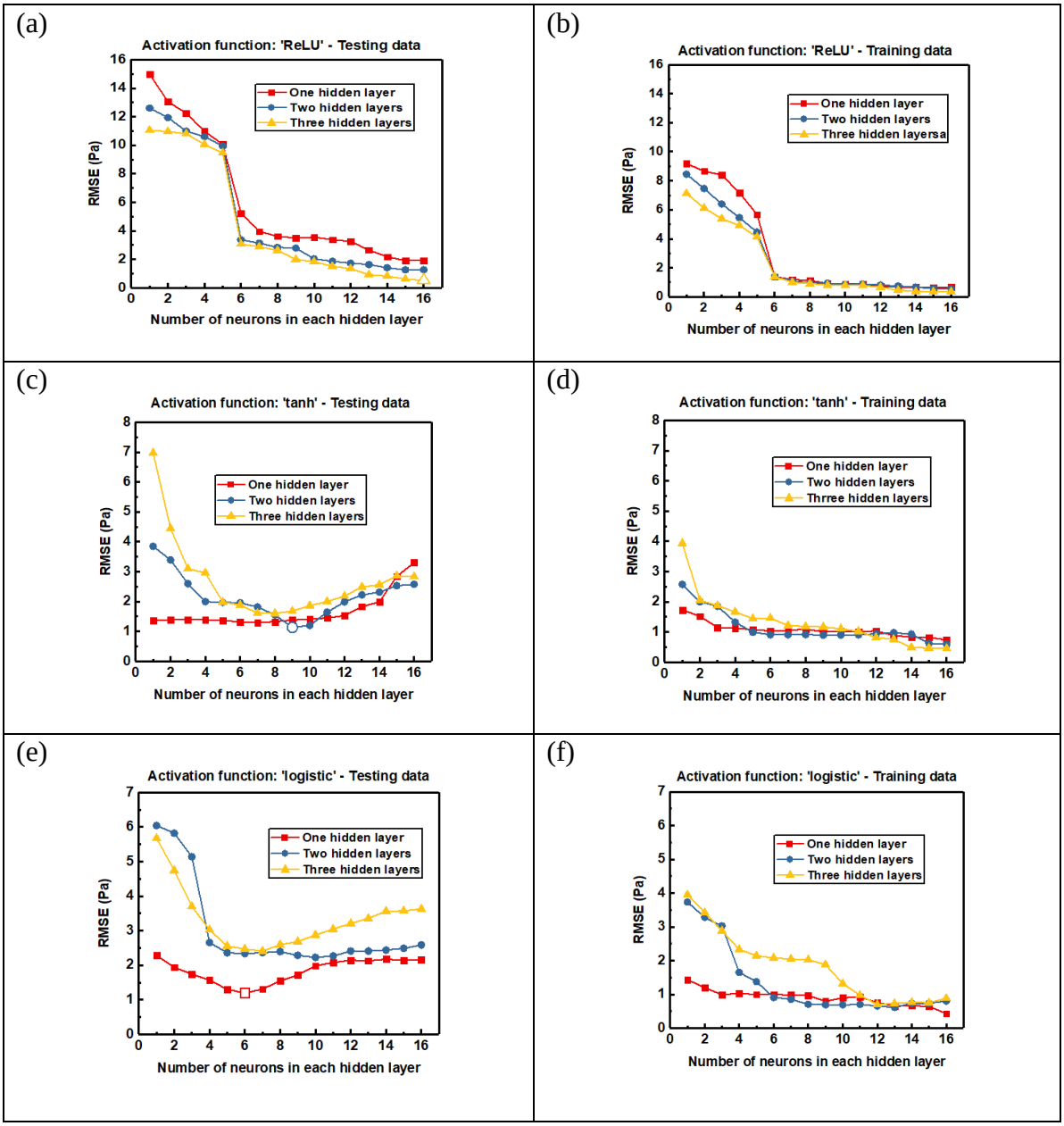


Figure 4-7. Comparison of different activation functions of the ANN algorithm based on RMSE vs. the number of neurons in each hidden layer for testing and training data. (a-b) ReLU, (c-d) tanh, and (e-f) logistic activation functions.

Table 4-2. Results of the comparative analysis of DDMs.

Algorithm	Kernel function	Min test RMSE	Algorithm	Kernel function	Min test RMSE	Algorithm	Activation function	Min test RMSE
GPR	Rational Quadratic	1.78	SVR	Linear	4.27	ANN	ReLU	**0.52**
	Matérn	2.54		RBF	1.50		tanh	1.13
	Squared Exponential	2.89		Poly	1.23		logistic	1.19
	Periodic	3.49						

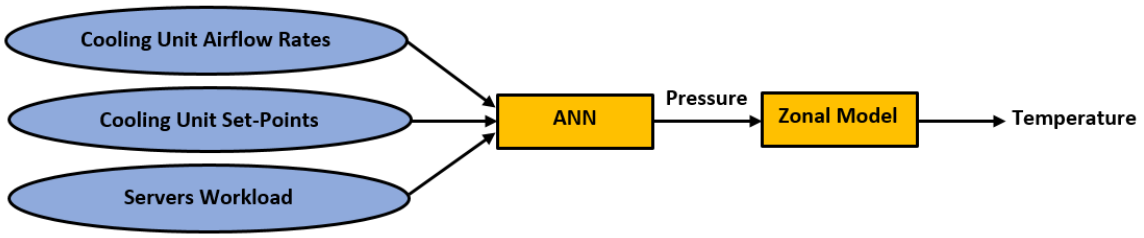


Figure 4-8. The data flow within the surrogate model for prediction used to predict pressures and temperatures.

4.5.2 Sample size required to train ANN

We produce a labeled dataset from the CFD simulations for cooling unit airflow changes in a row-based cooling DC. Each instance in the dataset consists of a collection of pressure data at 60 locations for 52 cooling unit airflows. Therefore, the dataset vector contains a 52×60 matrix. We train the model using different dataset sizes to determine an optimum sample size that provides the best trade-off between accuracy and training time. Table 4-3 shows the average prediction errors as the input data sizes are varied from 24×60 to 52×60. The error decreases with increasing numbers of samples until 44×60 beyond which the decrease in prediction errors is negligible. Therefore, the inputs for further investigations with the ANN algorithm correspond to a 44×60 matrix.

Table 4-3. Average train and test prediction error as the sample size changes.

Number of samples in the dataset	Average train RMSE (Pa)	Average test RMSE (Pa)
24×60	3.50	5.41
28×60	2.87	3.72
32×60	1.71	2.54
36×60	0.97	1.47
40×60	0.68	1.10
44×60	0.46	0.52
48×60	0.45	0.51
52×60	0.45	0.51

4.5.3 Surrogate model prediction and validation

Several realistic scenarios are now considered to determine the temperature distributions in the cold chamber, where model accuracy is evaluated by comparing the predicted results with those obtained from experiments.

4.5.3.1 Influence of cooling system operating conditions

Since the cooling unit airflow plays an important role in the temperature distribution of a DC, we again consider the high ($\dot{m}_{CU} \gg \sum \dot{m}_s$), sufficient ($\dot{m}_{CU} \cong \sum \dot{m}_s$), and low ($\dot{m}_{CU} \ll \sum \dot{m}_s$) airflows for workload and set-point temperature, specified as 100% and 18 ^{o}C, respectively. Figure 4-9 presents the temperature predictions using the surrogate model and temperature differences between model predictions and experiment ($\Delta = \left|T_{\text{Exp}} - T_{Model}\right|$) for these scenarios.

The temperature contours are mostly spatially uniform for the high airflow and values of Δ are lower than 0.8 ^{o}C (Figure 4-9 (b)). For sufficient airflows, the cold chamber pressure is slightly lower than in the hot chamber, leading to hot air recirculation through

the gaps between the racks, producing hot zones (Figure 4-9 (c) and (e)). Figure 4-9 (d) presents values of Δ, where all zones have differences lower than 1°C.

For the lowest airflow, Figure 4-9 (f) shows that 22 out of 25 zones have differences lower than 2 °C and only three zones have Δ > 2 °C. Even for this restrictive case, there is a good agreement between the predicted and experiment temperatures. However, due to hot air recirculation and the uncertainties of the gap resistances that should be viewed as porous media, the predicted temperatures for very few zones are quite different from those obtained from the experiments. Both methods reasonably represent how cold air enters the cold chamber (from the right and left sides), leads to a temperature profile with higher temperature along the middle rack (Rack 3), while side racks (Rack 1 and 5) maintain lower temperatures.

Figure 4-10 summarizes the average values of Δ at the rack inlets for the three cooling unit airflow rates. The middle rack has the largest difference due to hot air recirculation and, Δ increases as airflow decreases.

Next, the cooling system set-point temperature is increased while the airflow rate ($\dot{m}_{CU} \cong \sum \dot{m}_s$) and the server workloads (100%) are held constant. Figure 4-11 (a-b) presents the temperature profiles and Δ values for a set-point temperature of 18 °C. In Figure 4-11 (c-d) this setpoint is increased to 22 °C. The 4 °C increase in the set-point increases the local temperatures in the entire cold chamber and the average value of Δ also increases by 48%. At the higher setpoint temperature, 20 out of 25 zones have Δ values lower than 1 °C and 5 zones have one between 1 °C and 1.5 °C. The mean difference for both cases is lower

than 0.8 °C and in both cases, the middle rack has the highest difference. Overall, the results from the model are consistent with those from the experiments.

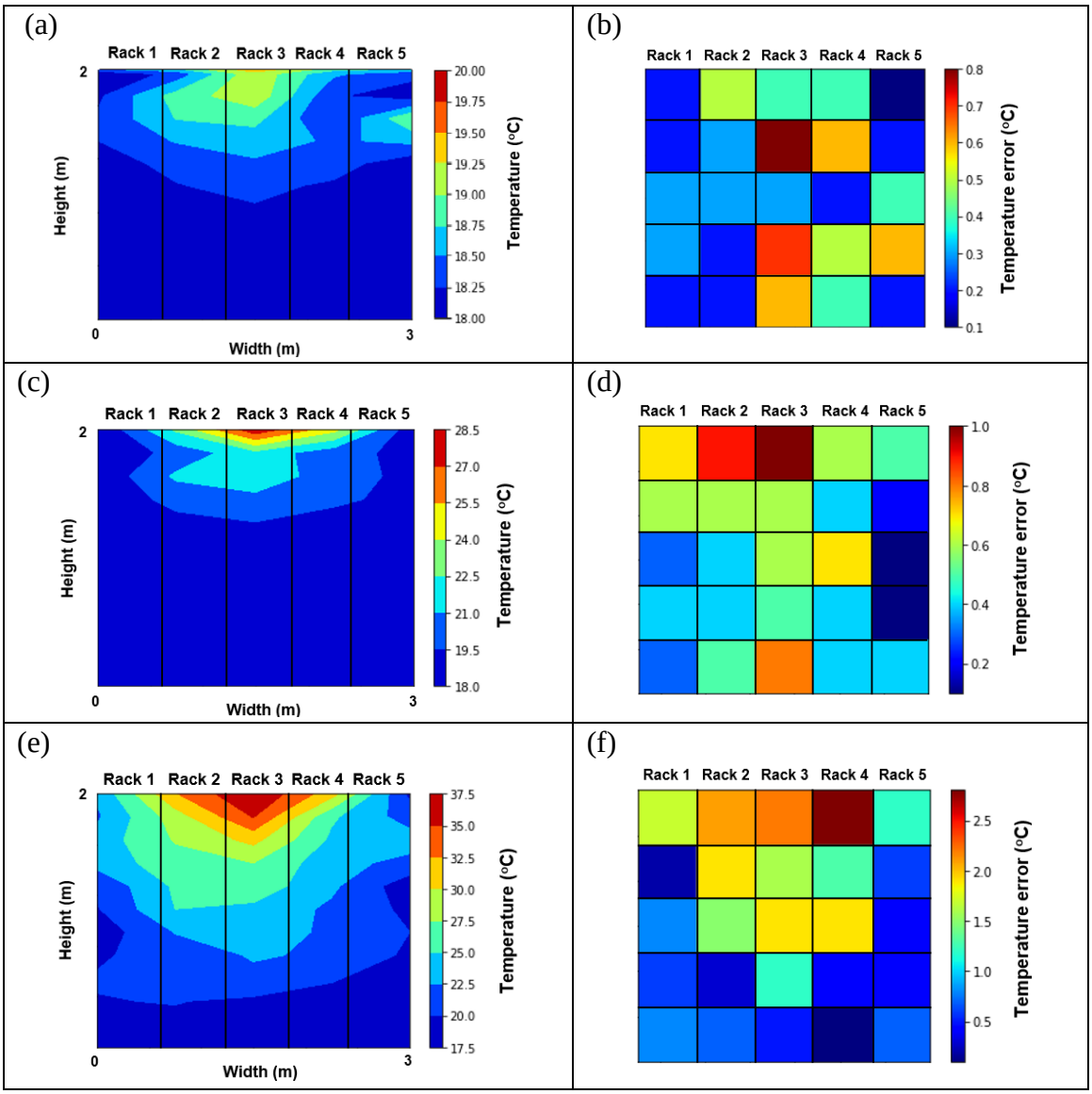


Figure 4-9. Effect of cooling unit airflow on the temperature distribution and temperature differences between model predictions and experimental results ($\Delta = |T_{Exp} - T_{Model}|$) for (a-b) high, (c-d) sufficient, and (e-f) low airflow rates.

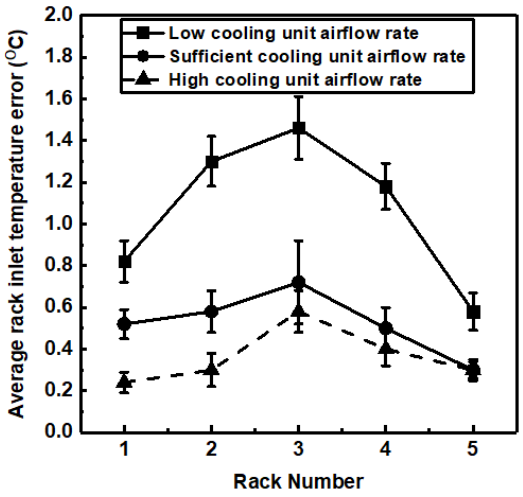


Figure 4-10. Average temperature differences between the model predictions and experiments ($\Delta = |T_{Exp} - T_{Model}|$) at the rack inlets for the three cooling unit airflow rates.

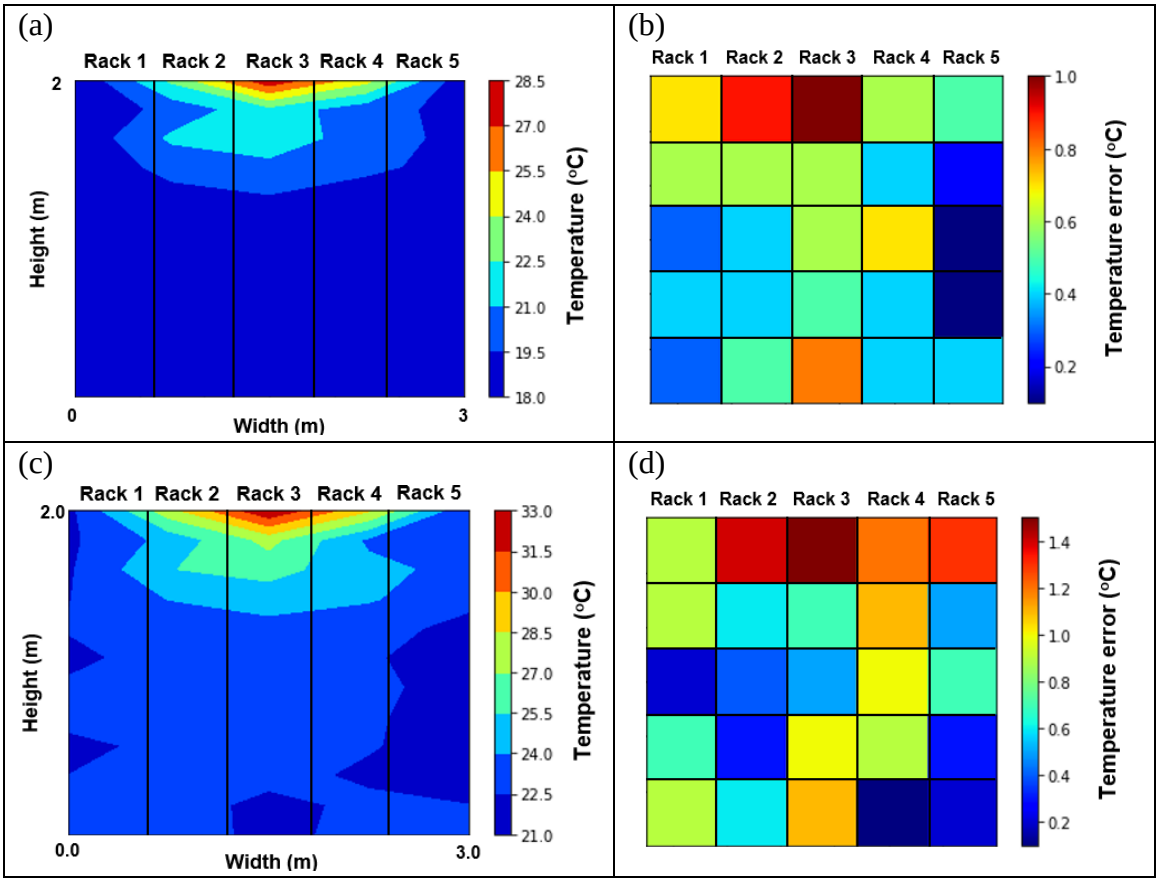


Figure 4-11. Effect of set-point temperature of the cooling unit on the temperature distribution and temperature differences between the model predictions and experiments ($\Delta = |T_{Exp} - T_{Model}|$) for (a-b) set-point at 18 °C and (c-d) at 22 °C.

4.5.3.2 Varying server workload

Figure 4-12 (a-b) presents the cold chamber temperature distribution and Δ values when all servers are operating at a 100% workload (18.9 kW in total), while Figure 4-12 (c-d) presents distributions and values of Δ for 50% workloads (12.6 kW in total). Both cases have cooling unit airflow rates of 1.2 kg/s and set-points of 18 ^{o}C. Reducing server workload decreases the maximum cold chamber temperature from 28.5 ^{o}C to 25 ^{o}C and the average value of Δ changes 19%. Again, the model is applicable and only a few zones have Δ > 1 ^{o}C.

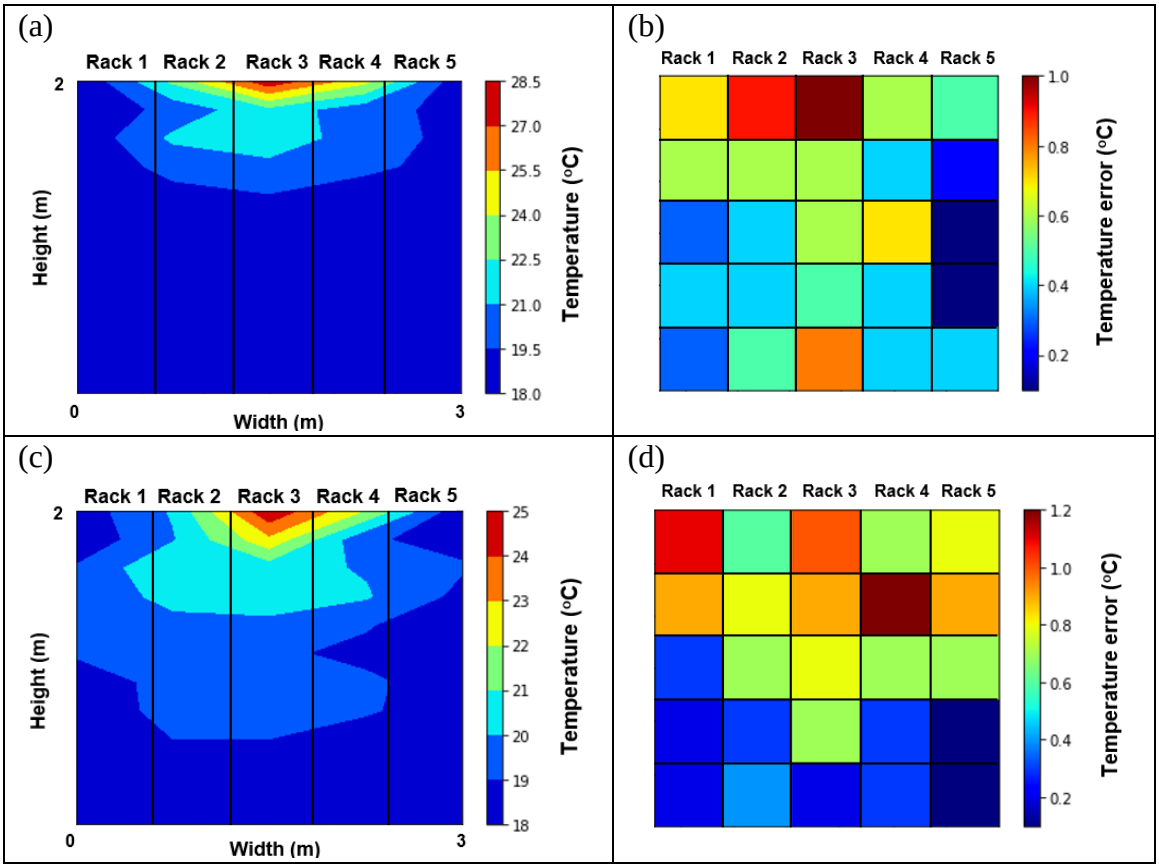


Figure 4-12. Effect of server workload on temperature distribution and temperature differences between the model predictions and experiments ($\Delta = |T_{Exp} - T_{Model}|$) for (a-b) server workloads set at 100% and (c-d) at 50%.

4.5.4 Robustness of surrogate model

Typical DCs contain a dynamic environment in which it is virtually impossible to collect data for all scenarios from either experiments or CFD simulations. A simpler predictive model must be robust enough to adapt to small changes. To investigate the adaptability of the surrogate model to changes, two complex scenarios different from the previous scenarios are investigated based on the same model trained in the previous sections. The first scenario changes the cooling system configuration and the second one alters the server locations. Here, CFD simulation results are compared with those from the surrogate model since it is not feasible to conduct experiments for these scenarios. Additionally, a total of 120 zones are created within the cold and hot chambers to accurately predict temperatures at server inlets.

4.5.4.1 Changing the locations and status of the fans

There are two row-based cooling units in the DC, at the right and left ends of the cold chamber, respectively. Each unit has 3 sets of fans, as shown in Fig. 4-13. Altering their locations can significantly change the temperature profile because the airflow distribution is changed. Figure 4-13 (a) shows the original configuration of the DC as reported in the previous sections where there is sufficient cooling unit airflow rate. In Figure 4-13 (b) and (c), the fans in either the right or left cooling units are moved downward, and in Figure 4-13 (d) the middle fans of the right cooling unit are turned off.

Figure 4-14 demonstrates that when the fan locations and their status is altered, the temperature distribution also changes. In Figure 4-14 (c), changing the locations of right

cooling unit fans downward produces hot areas at the tops of racks 3, 4, and 5. Similarly, in Figure 4-14 (e), these hot areas appear at the tops of racks 1, 2, and 3. Figure 4-14 (g) shows a band of warmer temperatures on the top of rack 3 due to the off-duty middle fan. All of these changes increase the temperatures in some areas, particularly where the fans are turned off.

Values of $D = |T_{\mathrm{CFD}} - T_{Model}|$ when the right-hand side fans are moved downwards are provided in Figure 4-14 (d). Here, 7 out of 60 zones have $D > 2$ ^{o}C, 2 zones have 1.5 ^{o}C < D < 2 ^{o}C, and the remainder have D < 1.5 ^{o}C. Figure 4-14 (f) shows only 1 zone with $D > 2$ ^{o}C, 7 zones with 1.5 ^{o}C < D <1.8 ^{o}C and the rest with D < 1.5 ^{o}C. For the case shown in Figure 4-14 (h), only 2 zones have $D > 2$ ^{o}C. These results indicate that the surrogate model generalizes satisfactorily to the changes in the locations of the cooling system fans and their status.

4.5.4.2 Changes in server location

The locations of servers and blanking panels in a rack also impact the airflow and, consequently, temperature distribution. Blanking panels in a DC fill the empty spaces in the racks and represent solid obstacles to prevent cold air bypass or hot air recirculation. The server locations are changed by sparsely distributed (Figure 4-15 (a)) or concentrating (Figure 4-15 (b)) to investigate the robustness of the surrogate model.

Figure 4-16 (a-b) and (c-d) shows that the temperature profiles and hot zones change as the server locations are changed from the original sparse locations. With concentrated servers, the hot zones spread towards the tops of racks 2, 3, and 4, while the

sparsely distributed servers produce a hot zone on top of the rack 3. When aggregated, the maximum temperature in the hot areas (25 ^{o}C) is lower than when the servers are sparsely placed (28.5 ^{o}C). Figure 4-16 (d) shows that for only 3 of 60 zones D > 2 ^{o}C, for 7 zones 1 ^{o}C < D < 1.5 ^{o}C, and for 50 zones D < 1 ^{o}C. Thus, the surrogate model is robust to the changes in server locations.

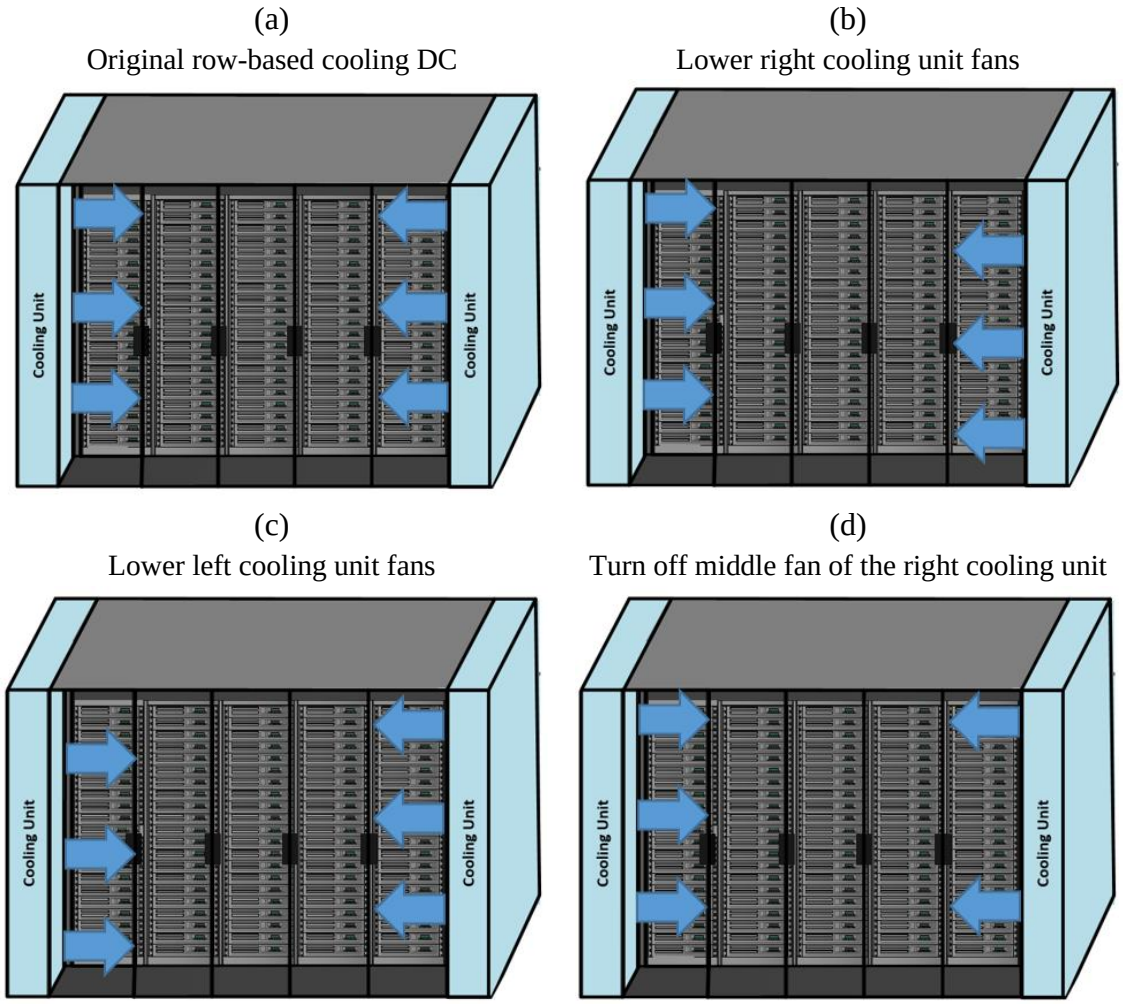


Figure 4-13. Four configurations of the cooling unit fans for the in-row DC cooling architecture.

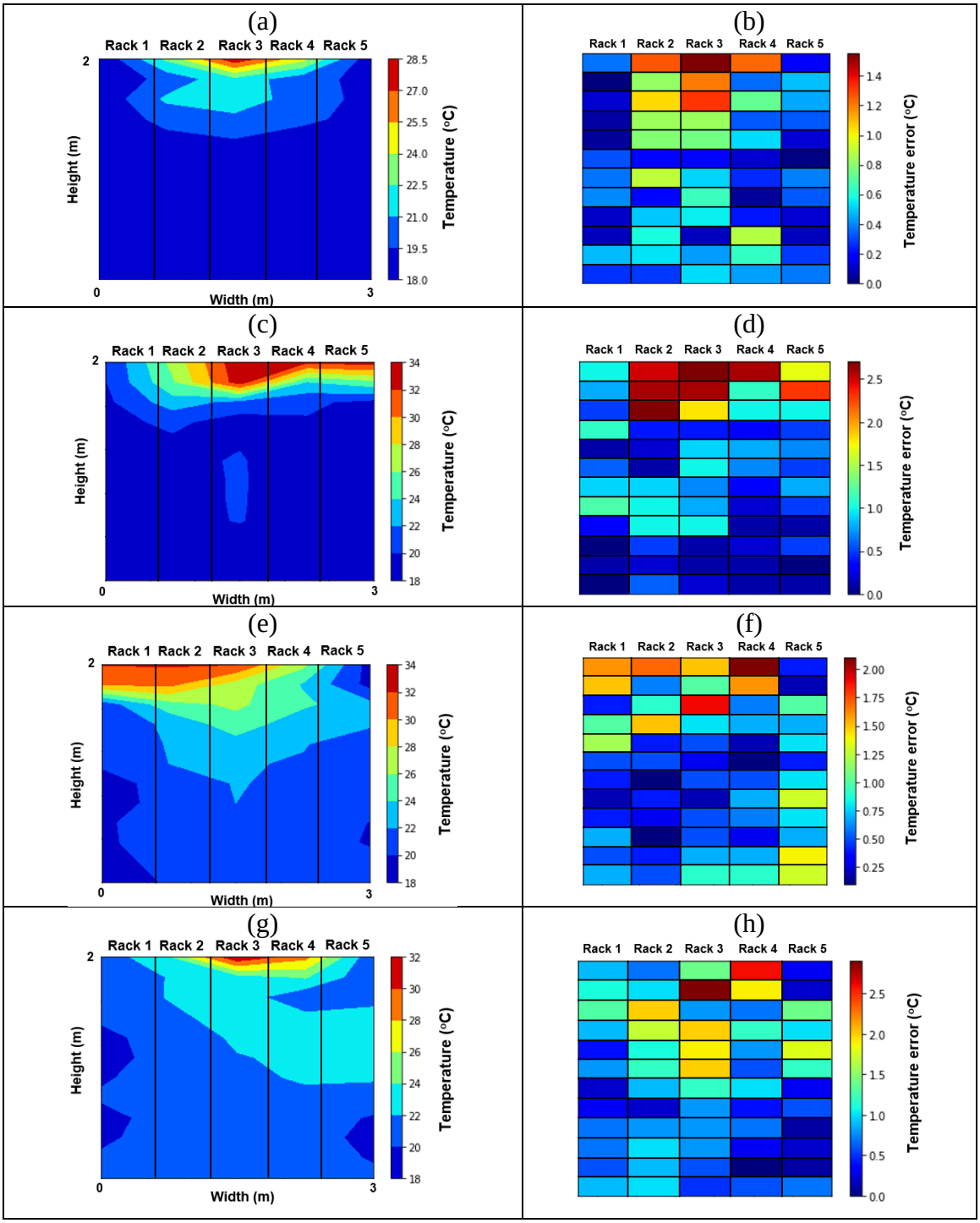


Figure 4-14. The temperature contours and temperature differences between the model predictions and CFD simulations ($D = |T_{CFD} - T_{Model}|$) for (a-b) the original row-based cooling DC configuration, (c-d) moving the fans of the right cooling unit downwards, (e-f) moving the fans of the left cooling unit downwards, and (g-h) turning off the middle fan of the right cooling unit.

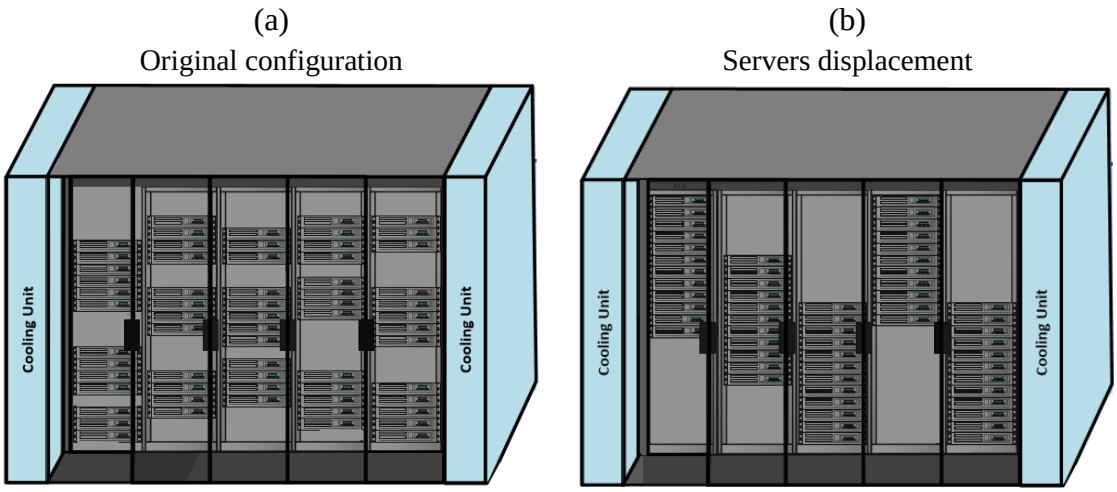


Figure 4-15. The two configurations for server locations reconfiguration. (a) The original scattered configuration and (b) servers aggregated around specific locations.

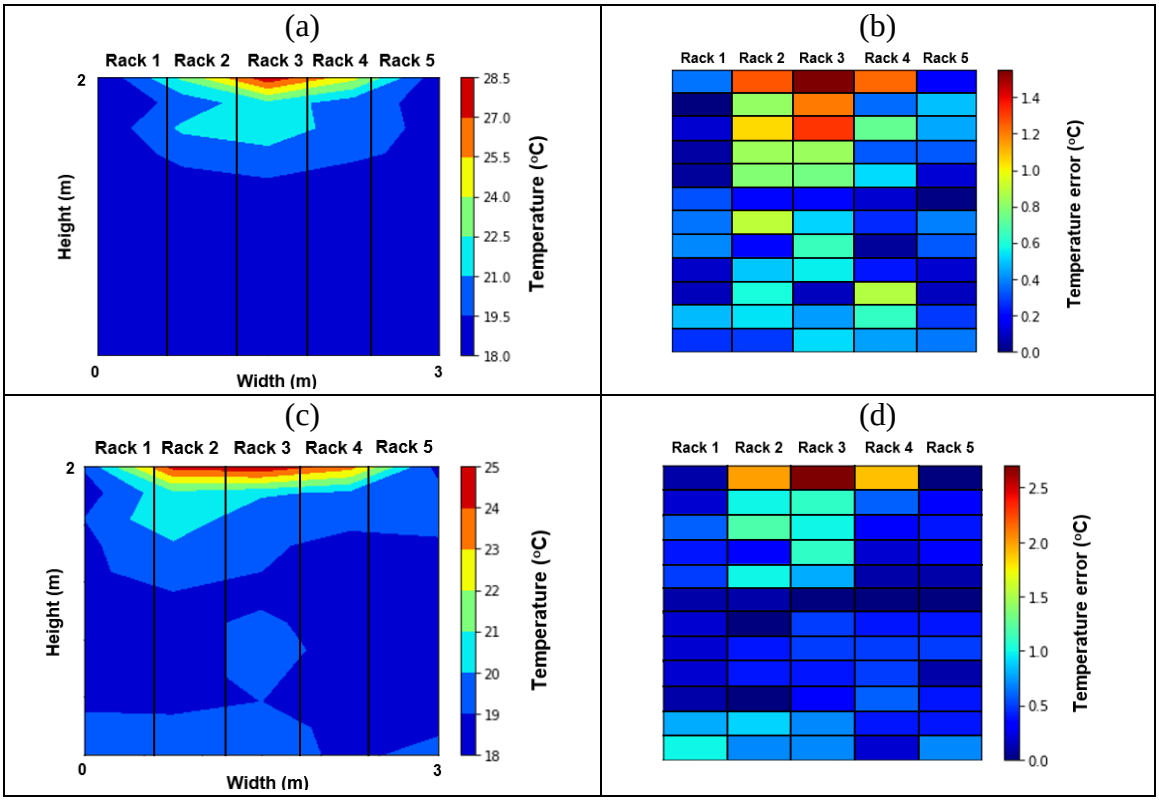


Figure 4-16. Temperature contours and temperature differences between the model predictions and CFD simulations ($D = |T_{CFD} - T_{Model}|$) when server locations are changed from scattered to aggregated in a row-based cooling DC. (a-b) Original scattered configuration and (c-d) aggregated servers.

4.6 Computation time

The computing cost is an important aspect of CFD simulations and surrogate model predictions. Table 4-4 summarizes the computing time on a personal computer with a Core i7-8700 CPU at 3.20 GHz,16 GB memory, and Windows 10 with a 64-bit operating system for the case with 100% server utilization, cooling unit temperature 18 ^{o}C and airflow rate of 1.2 Kg/s. For a typical single steady-state case, CFD simulations and experiments require roughly 28,800 and 234,400 seconds, respectively, whereas, in contrast, the surrogate zonal model requires only 3.6 seconds.

Table 4-4. Time to make experimental measurements of pressures and temperatures for a typical steady-state case, and the corresponding computational times required to obtain predictions from the CFD simulation and the surrogate model.

Method	Time to obtain results for a typical steady-state scenario (second)
CFD	~ 28,800
Experiment	~ 234,00
Surrogate model	~ 4

4.7 Conclusion

We present a machine learning-based surrogate model to predict the pressure, airflow rate and temperature distribution in a modular DC with a row-based cooling architecture. The surrogate model is an inexpensive tool that provides predictions at comparable accuracy as those from more detailed and computationally expensive CFD simulations. This model can be used to (1) improve cooling system design, (2) facilitate thermally aware workload management, and (3) test “what if” scenarios to characterize the influence of operational

conditions on the temperature distribution. The model applies the mass, momentum, and energy conservation equations to each zone for which unknown parameters in the conservation equations are obtained from DDMs. The model is developed by (1) collecting data from experimentally validated CFD simulations, (2) applying ANN to the data to predict pressure, and (3) applying the mass, momentum, and energy conservation relations to each zone to determine the zonal temperature.

Comparing the accuracy of the DDM predictions shows that the ANN algorithm with the ReLU activation function is more appropriate for the particular DC configuration. The cooling unit operation and server workload are varied to characterize their influence on the thermal performance of the DC using the ReLU-based ANN DDM. We determine:

- The cooling unit airflow rate has the most significant influence on the temperature distribution in the cold chamber and DMM accuracy. Increasing the cold air supply lowers the average predicted temperature in the cold chamber and improves the model accuracy.
- A 4 ^{o}C increment in the set-point temperature of the cooling unit results in an average of 4.2 ^{o}C rise in the predicted temperature in the front chamber, while the average prediction error increases by 48%.
- A 50% increase in server workload results in a 0.8 ^{o}C increment in the average temperature in the front chamber and the average temperature prediction error changes 19%.

- The middle rack (rack 3) has a higher temperature due to hot air recirculation and the end racks (racks 1 and 5) have lower temperatures since these racks are adjacent to the cooling units.

The robustness of the surrogate model is demonstrated by investigating the effect of server spatial distribution, cooling unit fan locations and on/off status. Our results from the adaptability examinations show the following:

- As the cooling unit fans are moved downwards, hot zones emerge at the tops of the racks and the prediction error for a few zones is larger than 2 ^{o}C.
- When the fan in the middle is turned off, a band of warmer temperatures emanates from the off-duty middle fan, but the prediction error is larger than 2 ^{o}C for only 2 zones.
- By changing the server locations from a scattered to an aggregated distribution, the spatial locations of hot zones change, but 95% of these zones still have a prediction error lower than 1.5 ^{o}C.

Hence, we demonstrate that the surrogate machine learning model can predict temperatures rapidly and accurately while adapting to the changes in operating conditions. Implementation of this model is promising for understanding DC configurations and their operation in order to enhance energy savings.

4.8 Acknowledgment

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under a collaborative research and development (CRD) project,

Computationally efficient Surrogate Models. We thank colleagues from CINNOS Mission Critical Incorporated who provided insight and expertise.

4.9 References

1. Uzaman, S.K., et al., *A systems overview of commercial data centers: Initial energy and cost analysis.* International Journal of Information Technology and Web Engineering (IJITWE), 2019. **14**(1): p. 42-65.

2. Joshi, Y. and P. Kumar, *Energy efficient thermal management of data centers*. 2012: Springer Science & Business Media.

3. Rong, H., et al., *Optimizing energy consumption for data centers.* Renewable and Sustainable Energy Reviews, 2016. **58**: p. 674-691.

4. Shuja, J., et al., *Sustainable cloud data centers: a survey of enabling techniques and technologies.* Renewable and Sustainable Energy Reviews, 2016. **62**: p. 195-214.

5. Ebrahimi, K., G.F. Jones, and A.S. Fleischer, *A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities.* Renewable and Sustainable Energy Reviews, 2014. **31**: p. 622-638.

6. Brunschwiler, T., et al., *Toward zero-emission data centers through direct reuse of thermal energy.* IBM Journal of Research and Development, 2009. **53**(3): p. 11: 1-11: 13.

7. Dai, J., et al., *OPTIMUM COOLING OF DATA CENTERS*. 2016: Springer.

8. Daraghmeh, H.M. and C.-C. Wang, *A review of current status of free cooling in datacenters.* Applied Thermal Engineering, 2017. **114**: p. 1224-1239.

9. El-Sayed, N., et al. *Temperature management in data centers: why some (might) like it hot*. in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*. 2012.

10. Cho, J., J. Yang, and W. Park, *Evaluation of air distribution system's airflow performance for cooling energy savings in high-density data centers.* Energy and buildings, 2014. **68**: p. 270-279.

11. Patankar, S.V., *Airflow and cooling in a data center.* Journal of Heat transfer, 2010. **132**(7).

12. Cho, J. and B.S. Kim, *Evaluation of air management system's thermal performance for superior cooling efficiency in high-density data centers.* Energy and buildings, 2011. **43**(9): p. 2145-2155.

13. Moazamigoodarzi, H., et al., *Influence of cooling architecture on data center power consumption.* Energy, 2019. **183**: p. 525-535.

14. Huang, Z., et al., *Numerical simulation and comparative analysis of different airflow distributions in data centers.* Procedia Engineering, 2017. **205**: p. 2378-2385.

15. Lyu, C., et al., *Enclosed aisle effect on cooling efficiency in small scale data center.* Procedia Engineering, 2017. **205**: p. 3789-3796.

16. Schiavon, S., et al., *Simplified calculation method for design cooling loads in underfloor air distribution (UFAD) systems.* Energy and Buildings, 2011. **43**(2-3): p. 517-528.

17. Zhang, K., et al., *Recent advancements on thermal management and evaluation for data centers.* Applied Thermal Engineering, 2018. **142**: p. 215-231.

18. Dunlap, K. and N. Rasmussen, *Choosing between room, row, and rack-based cooling for data centers.* APC White Paper, 2012. **130**.

19. Sun, H., P. Stolf, and J.-M. Pierson, *Spatio-temporal thermal-aware scheduling for homogeneous high-performance computing datacenters.* Future Generation Computer Systems, 2017. **71**: p. 157-170.

20. Zhao, X., et al., *A smart coordinated temperature feedback controller for energy-efficient data centers.* Future Generation Computer Systems, 2019. **93**: p. 506-514.

21. Samadiani, E., et al., *Reduced order thermal modeling of data centers via distributed sensor data.* Journal of heat transfer, 2012. **134**(4).

22. Zapater, M., et al., *Runtime data center temperature prediction using Grammatical Evolution techniques.* Applied Soft Computing, 2016. **49**: p. 94-107.

23. Choi, J., et al., *A CFD-based tool for studying temperature in rack-mounted servers.* IEEE transactions on computers, 2008. **57**(8): p. 1129-1142.

24. Moazamigoodarzi, H., et al., *Real-time temperature predictions in it server enclosures.* International Journal of Heat and Mass Transfer, 2018. **127**: p. 890-900.

25. Tang, Q., et al. *Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters*. in *2006 Fourth International Conference on Intelligent Sensing and Information Processing*. 2006. IEEE.

26. Erden, H.S., H.E. Khalifa, and R.R. Schmidt, *A hybrid lumped capacitance-CFD model for the simulation of data center transients.* Hvac&R Research, 2014. **20**(6): p. 688-702.

27. Taneja, S., Y. Zhou, and X. Qin, *Thermal benchmarking and modeling for HPC using big data applications.* Future Generation Computer Systems, 2018. **87**: p. 372-381.

28. Li, X., et al., *Holistic energy and failure aware workload scheduling in Cloud datacenters.* Future Generation Computer Systems, 2018. **78**: p. 887-900.

29. Nada, S., M. Said, and M. Rady, *Numerical investigation and parametric study for thermal and energy management enhancements in data centers' buildings.* Applied Thermal Engineering, 2016. **98**: p. 110-128.

30. Moore, J.D., et al. *Making Scheduling" Cool": Temperature-Aware Workload Placement in Data Centers*. in *USENIX annual technical conference, General Track*. 2005.

31. Chen, J., et al. *A high-fidelity temperature distribution forecasting system for data centers*. in *2012 IEEE 33rd Real-Time Systems Symposium*. 2012. IEEE.

32. Nada, S., M. Said, and M. Rady, *CFD investigations of data centers' thermal performance for different configurations of CRACs units and aisles separation.* Alexandria engineering journal, 2016. **55**(2): p. 959-971.

33. Macedo, D.G., et al. *Improving Airflow and Thermal Distribution in a Real Data Centre Room Through Computational Fluid Dynamics Modeling*. in *2019 8th International Conference on Industrial Technology and Management (ICITM)*. 2019. IEEE.

34. Athavale, J., Y. Joshi, and M. Yoda. *Artificial neural network based prediction of temperature and flow profile in data centers*. in *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*. 2018. IEEE.

35. Moore, J., J.S. Chase, and P. Ranganathan. *Weatherman: Automated, online and predictive thermal mapping and management for data centers*. in *2006 IEEE international conference on Autonomic Computing*. 2006. IEEE.

36. Shrivastava, S.K., J.W. VanGilder, and B.G. Sammakia. *Data center cooling prediction using artificial neural network*. in *ASME 2007 InterPACK Conference collocated with the ASME/JSME 2007 Thermal Engineering Heat Transfer Summer Conference*. 2007. American Society of Mechanical Engineers Digital Collection.

37. Wang, L., et al., *Task scheduling with ANN-based temperature prediction in a data center: a simulation-based study.* Engineering with Computers, 2011. **27**(4): p. 381-391.

38. Song, Z., B.T. Murray, and B. Sammakia, *Airflow and temperature distribution optimization in data centers using artificial neural networks.* International Journal of Heat and Mass Transfer, 2013. **64**: p. 80-90.

39. Solomatine, D.P. and A. Ostfeld, *Data-driven modelling: some past experiences and new approaches.* Journal of hydroinformatics, 2008. **10**(1): p. 3-22.

40. De Lorenzi, F. and C. Vömel, *Neural network-based prediction and control of air flow in a data center.* Journal of Thermal Science and Engineering Applications, 2012. **4**(2).

41. Song, Z., B.T. Murray, and B. Sammakia, *A dynamic compact thermal model for data center analysis and control using the zonal method and artificial neural networks.* Applied thermal engineering, 2014. **62**(1): p. 48-57.

42. MirhoseiniNejad, S., et al. *ALTM: Adaptive learning-based thermal model for temperature predictions in data centers*. in *2019 IEEE Sustainability through ICT Summit (StICT)*. 2019. IEEE.

43. Zhang, K., et al., *Machine learning-based temperature prediction for runtime thermal management across system components.* IEEE Transactions on parallel and distributed systems, 2017. **29**(2): p. 405-419.

44. Varsamopoulos, G., et al., *Using transient thermal models to predict cyberphysical phenomena in data centers.* Sustainable Computing: Informatics and Systems, 2013. **3**(3): p. 132-147.

45. Song, Z., B.T. Murray, and B. Sammakia, *A compact thermal model for data center analysis using the zonal method.* Numerical Heat Transfer, Part A: Applications, 2013. **64**(5): p. 361-377.

46. Li, L., et al. *Thermocast: a cyber-physical forecasting model for datacenters*. in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011.

47. Tang, Q., S.K.S. Gupta, and G. Varsamopoulos, *Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach.* IEEE Transactions on Parallel and Distributed Systems, 2008. **19**(11): p. 1458-1472.

48. Jiang, Z., et al. *Data-driven Thermal Model Inference with ARMAX, in Smart Environments, based on Normalized Mutual Information*. in *2018 Annual American Control Conference (ACC)*. 2018. IEEE.

49. Salih Erden, H., H. Ezzat Khalifa, and R.R. Schmidt, *Determination of the lumped-capacitance parameters of air-cooled servers through air temperature measurements.* Journal of Electronic Packaging, 2014. **136**(3).

50. Hassan, N., M.M.K. Khan, and M. Rasul, *Temperature monitoring and CFD analysis of data centre.* Procedia Engineering, 2013. **56**: p. 551-559.

51. Fluent, A., *12.0 User's guide.* Ansys Inc, 2009. **6**.

52. Molchanov, A.M., *NUMERICAL METHODS FOR SOLVING THE NAVIER-STOKES EQUATIONS.* 2019.

53. Temam, R., *Navier-Stokes equations: Theory and numerical analysis(Book).* Amsterdam, North-Holland Publishing Co.(Studies in Mathematics and Its Applications, 1977. **2**: p. 510.

54. Phan, L., B. Hu, and C.-X. Lin, *An evaluation of turbulence and tile models at server rack level for data centers.* Building and Environment, 2019. **155**: p. 421-435.

55. Fulpagare, Y. and A. Bhargav, *Advances in data center thermal management.* Renewable and Sustainable Energy Reviews, 2015. **43**: p. 981-996.

56. Wilcox, D.C., *Turbulence modeling for CFD*. Vol. 2. 1998: DCW industries La Canada, CA.

57. Wan, J., et al., *Air flow measurement and management for improving cooling and energy efficiency in raised-floor data centers: A survey.* IEEE Access, 2018. **6**: p. 48867-48901.

58. Megri, A.C. and F. Haghighat, *Zonal modeling for simulating indoor environment of buildings: Review, recent developments, and applications.* Hvac&R Research, 2007. **13**(6): p. 887-905.

59. Wurtz, E., L. Mora, and C. Inard, *An equation-based simulation environment to investigate fast building simulation.* Building and Environment, 2006. **41**(11): p. 1571-1583.

60. White, F.M., *Viscous flow in ducts.* Fluid mechanics, 1999. **3**.

61. Athavale, J., M. Yoda, and Y. Joshi, *Comparison of data driven modeling approaches for temperature prediction in data centers.* International Journal of Heat and Mass Transfer, 2019. **135**: p. 1039-1052.

62. Mehrotra, K., C.K. Mohan, and S. Ranka, *Elements of artificial neural networks*. 1997: MIT press.

63. Roache, P.J., *Perspective: a method for uniform reporting of grid refinement studies.* 1994.

64. Celik, I.B., et al., *Procedure for estimation and reporting of uncertainty due to discretization in {CFD} applications.* 2008.

65. KARIMI, M., et al. Quantification of Numerical and Model Uncertainties in the CFD Simulation of the Gas Holdup and Flow Dynamics in a Laboratory Scale Rushton-Turbine Flotation Tank. in the 9th International Conference on CFD in the Minerals and Process Industries. 2012.

4.10 Appendix

4.10.1 Mesh independence study

The accuracy of the results from CFD simulation relies on mesh quality. In order to evaluate uncertainties and discretization errors in the simulation, we use the matrix called the grid convergence index (GCI) based on Richardson Extrapolation that creates bounds for discretization error [63].

To apply the GCI method, three meshes with different grid spacings h_1, h_2, and h_3 that represent coarse, medium, and fine meshes, respectively, are built. Each grid spacing yields three solutions f_1, f_2, and f_3. The grid refinement factor,

$$r_{k,k+1} = \frac{h_{k+1}}{h_k} \tag{A. 4-1}$$

is calculated, where k denotes the mesh level. Based on experience, the desired value for r is greater than 1.3 [64]. The order of convergence,

$$p = \ln\left(\frac{f_3 - f_2}{f_2 - f_1}\right) \Big/ \ln r \tag{A. 4-2}$$

The GCI for the fine mesh,

$$GCI_{fine} = \frac{F_S|\varepsilon|}{(r^p - 1)} \tag{A. 4-3}$$

where ε denotes the relative error, F_S the safety factor, where the range $1.25 \leq F_S \leq 3$ is recommended [65].

The mesh independence study is performed based on the GCI for a row-based cooling architecture DC in which the grid is more refined around critical boundaries. Table

A. 4-1 provides details for calculating GCIs from Eqs. (A. 4-1) to (A. 4-3) for three meshes with 2.6 million, 3.3 million, and 4.4 million nodes, respectively. According to Table A. 4-1, the numerical uncertainty in the coarse and fine-grid solutions for cold and hot chambers reveals that there are no significant differences between these two GCI_{12} and GCI_{23}. Both 3.3 million and 4.4 million nodes will lead to similar results while the calculations with 4.4 million nodes will result in more computational efforts.

Figure A. 4-1 and Figure A. 4-2 present the temperature and pressure profiles for the three grids with more details, where the numerical uncertainty is indicated by error bars. The small values of GCI (GCI < 0.03% for temperature and GCI < 0.04% for pressure) reveal that the results of the simulation cannot be improved by refining the mesh. Indeed, the average deviations of temperature and pressure within 60 zones are less than 2% between the medium mesh and fine mesh. Therefore, the medium mesh is selected as the optimum mesh for all simulations to reduce computational time.

Table A. 4-1. Calculation of discretization error.

Case mesh	Cells number	Refinement factor, r	GCI-Cold chamber temperature	GCI-Hot chamber temperature	GCI-Cold chamber pressure	GCI-Cold chamber pressure
Coarse (1)	2,556,883	1.40	0.04%	0.06%	0.10%	0.02%
Medium (2)	3,322,654					
Fine (3)	4,397,640	1.40	0.02%	0.03%	0.04%	0.01%

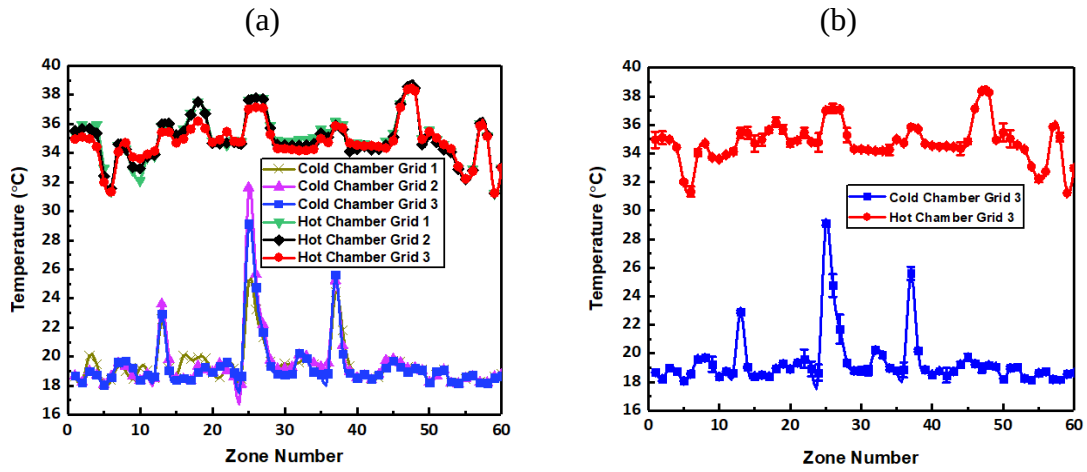


Figure A. 4-1. GCI results for the (a) temperature profile within 60 zones and (b) discretization error for the fine-grid solution.

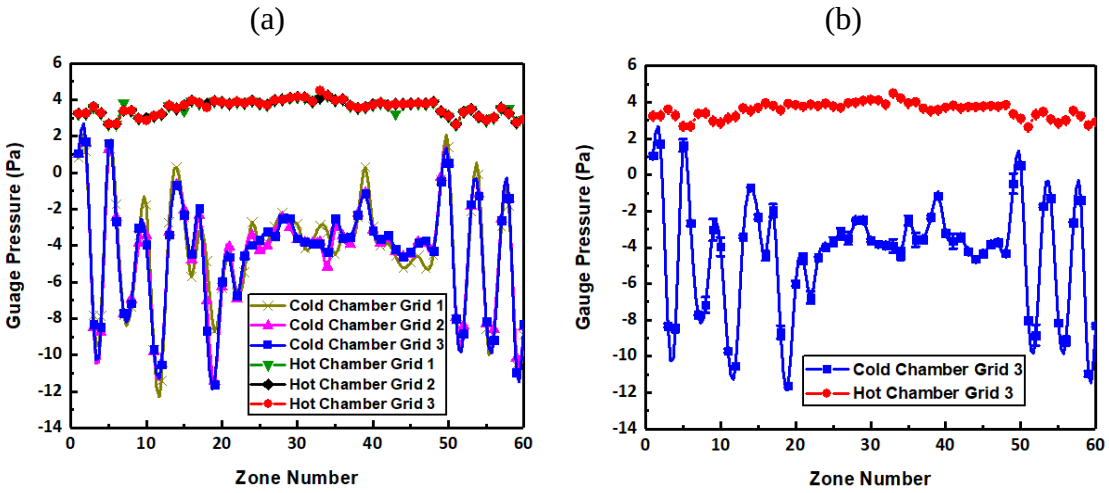


Figure A. 4-2. GCI results for the (a) pressure profile within 60 zones, and (b) discretization error for the fine-grid solution.

Chapter 5

A Gray-Box Model for Real-Time Transient Temperature

This chapter is reproduced from *"A Gray-Box Model for Real-Time Transient Temperature",* ***Sahar Asgari****, SeyedMorteza MirhoseiniNejad, Hosein Moazamigoodarzi, Rohit Gupta, Rong Zheng and Ishwar K. Puri, Published in Applied Thermal Engineering, 2020.*

The author of this thesis is the first author and the main contributor of this publication. Her main contributions to this work consist of introducing the idea of using gray-box model for transient study, writing the manuscript, formulating the problem, conducting the experiments, running CFD simulations, implementing the framework, constructing the algorithms, and generating the numerical results.

5.1 Abstract

In response to the need to improve the energy efficiency of data centers (DCs), system designers now incorporate solutions such as continuous performance monitoring, automated diagnostics, and optimal control. While these solutions must ideally be able to predict transient conditions, in particular real time DC temperatures, existing forecasting methods are inadequate because they (1) make restrictive assumptions about system configurations, (2) are extremely time-consuming for real time applications, (3) are accurate only over limited time horizons, (4) fail to accurately model the effects of operating conditions, such as cooling unit operation conditions and server workloads, or (5) ignore important facets of the flow physics and heat transfer that can lead to large prediction errors in extrapolative predictions. To address these deficiencies, we develop a gray-box model that combines machine learning with the thermofluid transport equations relevant for a row-based cooled DC to predict transient temperatures in server CPUs and cold air inlet to the servers. An artificial neural network (ANN) embedded in the gray-box model predicts pressures, which provide inputs for the thermofluid transport equations that predict the spatio-temporal temperature distributions. The model is validated with experimental measurements for different (1) server workload distributions, (2) cooling unit set-point temperatures and (3) the airflow of the cooling units. This gray-box model exhibits superior performance compared to a conventional zonal temperature prediction model and an advanced black-box model that is based on a nonlinear autoregressive exogenous model. An application of the gray-box model involves a case study to detect cooling unit fan failure in a row-based DC cooling system.

Key words: Datacenter, real-time temperature prediction, fault detection, ANN, NARX.

Nomenclature

Uppercase letters

C_p	Specific heat capacity (kJ kg^{-1} K^{-1})
CU	Cooling unit
E	Total energy (kJ)
E_{ij}	Rate of deformation
F	Body force (N)
$\bar{J}_n$	diffusion flux of species n
MAX_{AE}	Maximum absolute error
MSE	Mean squared error
$\dot{P}_k$	Power consumption of server k (kW)
Q	Heat flux (kW)
Q_{source}	Internal heat source (kW)
S_h	Heat of chemical reaction (kJ)
STD_{AE}	Standard deviation of absolute error
T	Absolute temperature (K)
T_{out}	Server exhaust temperature (K)
T_{in}	Server inlet temperature (K)
V	Volume (m^3)
X	Thermal mass of server (kJ K^{-1})
$\hat{y}$	Predicted value

Lowercase letters

g	Gravitational acceleration (m s^{-2})
h	Enthalpy (kJ)
k	Turbulent kinetic energy
k_{eff}	Effective conductivity
$\dot{m}$	Mass flow rate (kg s^{-1})
n	Server number
dp	Pressure drop (Pa)
p	Static pressure (Pa)
t	Time (s)
v	Velocity (m s^{-1})

Subscripts and superscripts (uppercase)

AE	Absolute error
EXP	Experiment
$Pred$	Prediction
$Pres$	Pressure
S	

Subscripts and superscripts (lowercase)

i	Index of zone in x-direction
j	Index of zone in y-direction
n	Number of species
s	Server

Greek letters

ρ	Density (kg m^{-3})
$\bar{\bar{\tau}}$	Stress tensor (Pa)
ε	Energy dissipation rate
μ_t	Eddy viscosity (m^2 s^{-1})

5.2 Introduction

Cloud computing has driven significant growth in data centers (DCs) and consequently global energy consumption by DCs accounts for about 3.5 % of worldwide electricity use. By 2025 DC energy use is anticipated to account for 20% of worldwide consumption [1-3]. Depending on the specific IT equipment, cooling units account for 24-60% of the total energy consumed by a DC [4, 5] so that ineffective cooling leads to significant energy waste [6-8]. While liquid cooling is promising for its effectiveness and offers the possibility of heat reuse [9, 10], air cooling is the preferred method employed in DCs, which will remain for the foreseeable future due to its reliability, simplicity of air handling, lower capital and maintenance costs, and the uncertainties associated with liquid cooling systems [11-13]. To decrease the energy consumption of air-cooling systems, the designer must consider (1) improving the airflow distribution in a DC and (2) optimizing the system for effective heat transfer.

Improvements in airflow distribution reduce energy costs by favorably influencing server CPU temperatures. This is accomplished by using a suitable DC cooling configuration, such as a row-based cooling within an enclosure that separates the chilled and hot air to eliminate hot air recirculation and cold air bypass, both of which produce undesirable airflow distributions [14-18]. Minimizing the total airflow and maximizing supply air temperatures improve the efficiency of the air-handler. This optimization requires that temperatures must be accurately predicted to apply appropriate strategies properly [19]. For example, controllers can be programmed to take actions that minimize the airflow and maximize supply air temperatures, while complying with the ASHRAE

guideline on the maximum allowable rack inlet air temperature [20, 21]. Fast detection of anomalous behavior also saves energy and reduces operational costs by initiating remedial actions. Cooling unit fans are relatively low-reliability components, where the failure of one or more fans can cause the entire system to overheat. Therefore, appropriate energy-saving strategies depend largely on the accuracy and timeliness of temperature prediction models.

Several methods are available to predict the temperatures in a DC, including white-box [22-27], black-box [28-33], and gray-box models [34-39]. White-box, or physics-based, models are based on an understanding of physical laws and the underlying engineering principles. While some white-box models are computationally fast, they generally adapt insufficiently to rapid operational changes within a DC. Furthermore, due to simplifying assumptions, such models have poor accuracy. In black-box models, system inputs and outputs are correlated through a mathematical function to predict system operations, but without an understanding of the underlying physical and thermodynamics principles. They are accurate if training data are abundant. Black-box models are used to obtain fast interpolative temperature predictions in DCs, e.g., steady-state and transient air temperatures, but their accuracy in making extrapolative predictions is limited [28, 40].

Hybrid or gray-box models combine physics-based white-box models with data obtained from experiments or simulations to develop approximate model parameter values. Thus, gray-box models are more general than black-box models and can provide extrapolative predictions with higher accuracy than white-box models. Although existing gray-box models for DC temperature predictions include some aspects of physical laws,

they fail to characterize important phenomena, such as hot air recirculation, making their predictions unreliable. Furthermore, most gray-box models in the literature employ linear regression, which is inappropriate for a DC due to the complexity and nonlinearity of the governing equations [41]. One model utilizes autoregression to predict transient temperatures in a DC with a 2D hybrid approach that represents the first law of thermodynamics and also includes sensor observations [35]. It is trained using airflow measurements at the front, or cold ends, of servers, but this is not practical in all DCs due to measurement complexities and the model also ignores hot air recirculation [40].

In summary, existing forecasting methods suffer from one or more of the following limitations. (1) They are not generic models applicable for several configurations, (2) their prediction algorithms are usually inappropriate for transient operation, (3) the computational time they require can be of the order of several minutes or even hours, making the models unsuitable for real-time applications, (4) temperature predictions are only available over short durations and not until steady-state conditions are reached, (5) comprehensive effects of all important operating conditions, such as cooling unit set-point, airflow, and server workload, are not included, and (6) the methods generally ignore important aspects of flow physics and heat transfer.

We present a gray-box model for thermal anomaly detection that predicts the transient CPU and inlet air temperatures in an enclosed DC by combining fundamental thermofluid relations with a data-driven solution. The model employs an artificial neural network (ANN) in conjunction with a 3D zonal model to find unknown parameters, and it is trained with data obtained from CFD simulations. We compare it with a black-box model

based on a nonlinear autoregressive exogenous model (NARX) and a conventional zonal model developed in [13], where the airflow within each zone is determined using a mechanical resistance circuit analysis. These flowrates contribute to a zonal energy balance to predict the temperature of each zone in an in-row cooling unit DC. To demonstrate the utility of the gray-box model in data center monitoring, we consider the problem of detecting fan failures in a modular data center, using a classifier trained from the predictions of the gray-box model.

All three models perform well for interpolative predictions, but our gray-box model outperforms for extrapolative predictions under different scenarios. To the best of our knowledge, this is the first study to compare a 3D gray-box model with black-box and conventional zonal models for transient temperature predictions in a DC.

Below, Section 2 introduces the details of the model and its framework. Section 3 compares the transient CPU and inlet air temperatures using the gray-box and black-box models and provides an application of our gray-box model. Finally, Section 4 summarizes our conclusions.

5.3 Methodology

We develop a gray-box thermal model to predict transient server CPU and inlet air temperatures. As depicted in Figure 5-1, pressure data are first collected from experimentally validated CFD simulations. These data are used to train an ANN to predict pressure in different zones. Next, the predicted pressures are applied in the momentum, mass, and energy relations to predict the temperatures.

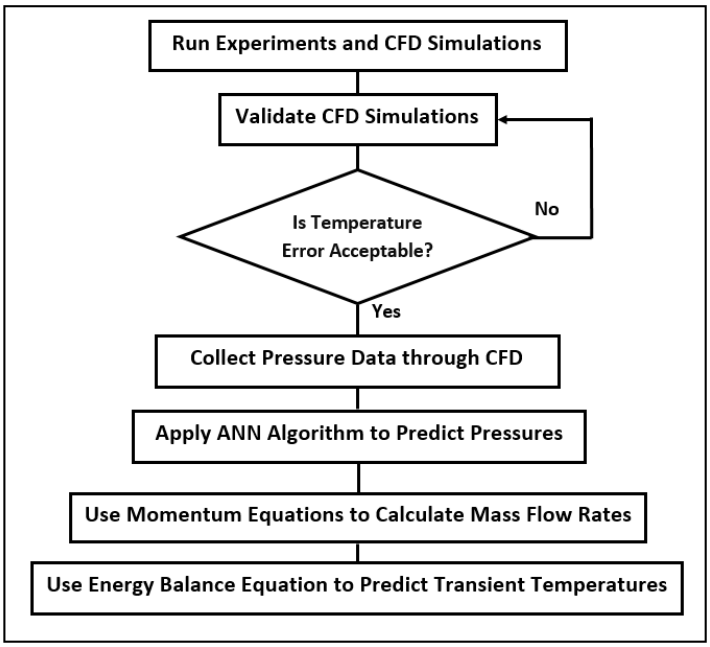


Figure 5-1. Block diagram of the gray-box model for transient temperature predictions.

5.3.1 System description

Figure 5-2 presents the configuration of an in-row cooling modular DC that is instrumented with 25 thermocouples to obtain air temperature measurements at the front of the racks in the cold chamber. Also shown is a schematic of the airflows within the enclosure. The DC houses two in-row cooling units that are placed at the left and right ends of the enclosure and five racks with 64 servers inside them. The servers are sparsely distributed and their CPU temperatures are measured by temperature sensors integrated into the core of the servers using the MobaXterm interface software. This software reports and records all on-board sensors measurements.

The cooling units draw warm air from the back (hot) chamber, extract heat from it and release cold air into the front (cold) chamber. Servers take in the cold air from the cold chamber and expel warm air to the hot chamber from where it is returned to the cooling

units. The racks are partially populated with servers and the empty spaces are blocked with blanking panels. There may be airflow leakage either from the hot to the cold chamber or vice versa due to the local pressure differences between these two chambers.

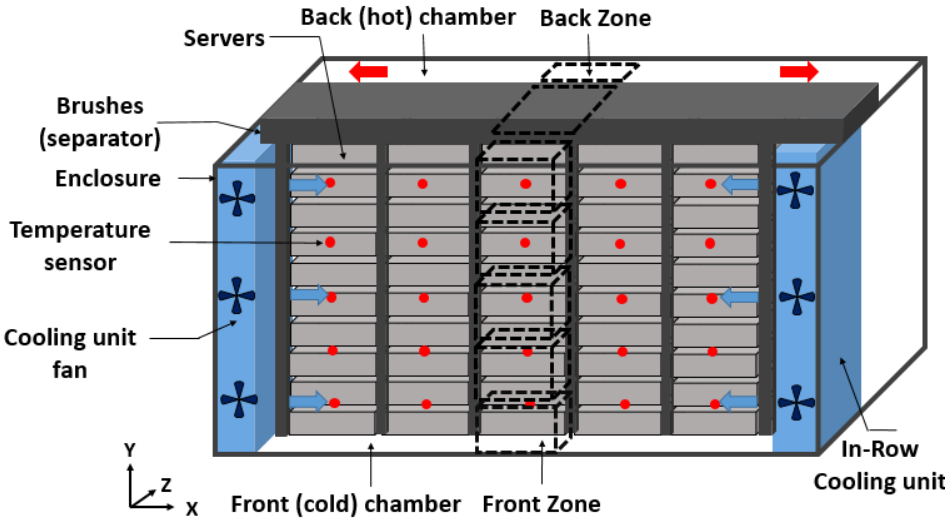
5.3.2 Computational fluid dynamics (CFD)

The CFD simulations for a row-based cooling DC are performed using ANSYS Fluent with the temperatures and turbulent flow field modeled using energy equations and a realizable k-ε model [42-44]. A mesh independence analysis is performed based on the grid convergence index (GCI) for coarse, medium, and fine meshes with 2.6 million, 3.3 million, and 4.4 million nodes, respectively. Based on the GCI, an intermediate mesh is selected for all simulations. For the transient analysis, the second-order upwind scheme is adopted for the convection term and the semi-implicit method used for the pressure-linked equation (SIMPLE) algorithm. The racks in Figure 5-2 are modeled as recirculation boundaries, the cooling units as mass flow inlets and pressure outlets for the cold air supply and the return air, respectively. The gaps between the racks, which can cause air recirculation if not properly sealed, are modeled as porous media using a power-law model to account for their resistance,

$$dp = -C_0|v|^{C_1}, \tag{5-1}$$

where dp denotes the pressure drop across the porous zone, $|v|$ the velocity magnitude, and C_0, and C_1 are empirical coefficients determined from experiments [45].

a)



b)

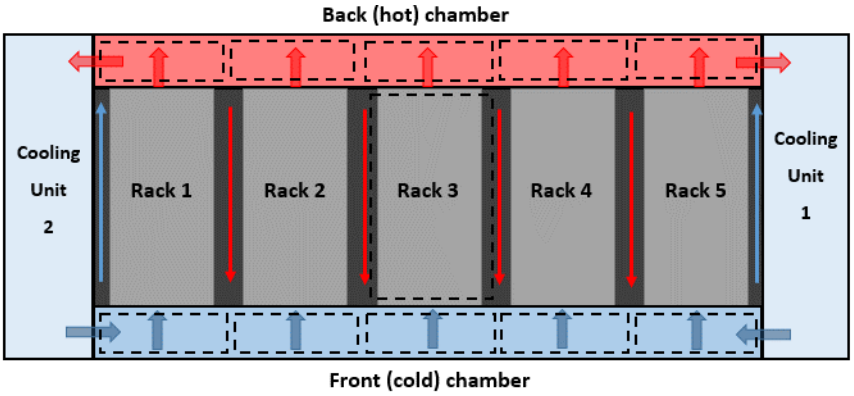


Figure 5-2. Schematic of the DC enclosure with five racks and two in-row cooling units. (a) Thermocouple locations and (b) top view of the airflow distribution. The enclosure is 3.2 m long, 1.4 m wide, and 2.05 m high.

The governing equations for mass, momentum, and energy conservation are [46],

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho \bar{v}) = 0, \tag{5-2}$$

$$\frac{\partial (\rho \bar{v})}{\partial t} + \nabla.(\rho \bar{v} \bar{v}) = -\nabla p + \nabla.(\bar{\bar{\tau}}) + \rho \bar{g} + \bar{F}, \text{ and} \tag{5-3}$$

$$\frac{\partial}{\partial t}(\rho E) + \nabla.\left(\bar{v}(\rho E + p)\right) = \nabla.\left(k_{eff}\nabla T - \sum_n h_n \bar{J}_n + \left(\bar{\bar{\tau}}_{eff}.\bar{v}\right)\right) + S_h, \quad (5\text{-}4)$$

where p denotes the static pressure, $\bar{\bar{\tau}}$ the stress tensor, $\rho\bar{g}$ and $\bar{F}$ the gravitational body force and external body force, E the total energy, k_{eff} the effective conductivity, h_n the enthalpy of species n, $\bar{J}_n$ the diffusion flux of species n, and S_h the heat of chemical reaction that is assumed to be zero.

The relations for the turbulent kinetic energy k and energy dissipation rate ε are [47],

$$\frac{\partial(\rho k)}{\partial t} + \frac{\partial(\rho k v_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[\frac{\mu_t}{\sigma_k}\frac{\partial k}{\partial x_j}\right] + 2\mu_t E_{ij}E_{ij} - \rho\varepsilon, \text{ and} \quad (5\text{-}5)$$

$$\frac{\partial(\rho \varepsilon)}{\partial t} + \frac{\partial(\rho \varepsilon u_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[\frac{\mu_t}{\sigma_\varepsilon}\frac{\partial \varepsilon}{\partial x_j}\right] + C_{1\varepsilon}\frac{\varepsilon}{K}2\mu_t E_{ij}E_{ij} - C_{2\varepsilon}\rho\frac{\varepsilon^2}{k}, \quad (5\text{-}6)$$

where v_i represents the velocity component in the corresponding direction i, E_{ij} the component of the rate of deformation, μ_t the eddy viscosity, and σ_k, σ_ε, $C_{1\varepsilon}$, and $C_{2\varepsilon}$ are constants.

5.3.3 Gray-box model

Even though CFD simulations can predict DC temperatures through white-box models, these approaches are computationally very expensive. Zonal models are a faster and reasonably accurate alternative, where the DC environment is partitioned into a grid of coarse zones with the assumption that the physical quantities inside each zone are spatially uniform. A set of nonlinear coupled equations consisting of the mass, momentum, and energy conservation relations is applied for each uniform zonal volume [48-50]. Figure 5-3 depicts the 3D zones inside the enclosure for a row-based cooling architecture DC. A total

of 75 zones are created by considering the (1) fronts of servers, (2) backs of servers, and (3) servers themselves. Figure 5-2.a presents a schematic of the zones for the middle rack. There are 5 zones in front of the rack, 5 zones at the back of that rack, and another 5 zones within the rack itself, i.e., there are 15 zones overall. This scheme is followed for all other racks. Since the system contains 5 racks, the total number of zones is $15 \times 5 = 75$. Servers are scattered in the racks and empty spaces in the racks are blocked by blanking panels in the DC configuration that we have investigated. Thus, each zone may contain either a single server or more than one server.

If the inlet and exit airflows are known for each zone, the energy balance equations can be applied to determine temperatures. To predict airflows in each zone, data is first collected using CFD simulations for a range of variables (Table 5-1). Next, an ANN is trained to characterize the relation between the zonal pressures and cooling configuration [51].

Table 5-1. Independent and dependent variables for DDMs.

Independent variable	Range	Dependent variable
Cooling unit airflow rate	0.40 – 2.4 (kg/s)	
Cooling unit set-point	16 – 22 (^{o}C)	Static pressures
Servers workload	0 – 100%	

We select ANNs due to their high capacity to model behaviors of complex and nonlinear systems. They are able to reproduce the complex general trends for input and output variables. Typically, an ANN consists of an input layer, some hidden layers, and an

output layer [52]. Each layer contains a number of neurons, where the hidden and output-layer neurons are each linked to the neurons in the previous layer. The main challenge with ANN is the choice of model complexity. When the number of parameters is far larger than the available training data, overfitting may happen, else unfitting may occur. The relevant parameters of the ANN are provided in Table 5-2. To train the network, we use the ReLu (rectified linear unit) activation function in the intermediate layers and the Levenberg-Marquardt back-propagation algorithm (LMA) that minimizes nonlinear functions.

Table 5-2. Parameters of the ANN model.

Model attributes	Details
Number of Layers	5
Number of Neurons in Layer 1 - 5	66 – 4 – 16 – 5 – 60
Activation function	Relu (rectified linear unit)
Training algorithm	Levenberg-marquardt back-propagation algorithm (LMA)

Given the predicted pressure for each zone by the trained ANN, the inlet and exit airflows of each zone can be determined from the mass and momentum conservation equations as follow,

$$\sum_j \dot{m}_{j\to i} = 0, \text{ and} \tag{5-7}$$

$$\sum F = \sum F_{Press} + \sum F_{Body} = \sum(\dot{m}v)_{out} - \sum(\dot{m}v)_{in}, \tag{5-8}$$

where $\dot{m}_{j\to i}$ denotes the interfacial mass flow rate transferred from cell j to cell i, F pressure and body forces in x, y, and z direction, v velocity, and ρ density. In Eq. (5-7), the mass within the control volume is constant. Eq. (5-8) shows that the momentum

changes only through the action of forces described by Newton's laws of motion. Table 5-3 contains expressions for force terms in Eq. (5-8).

Table 5-3. Expressions for the terms in Eq. (5-8).

$\sum F = \sum F_x + \sum F_y + \sum F_z$	
$\sum F_x = \sum F_{x,Press} + \sum F_{x,Body}$ (Force in *x* direction)	$(PA)_{x,out} - (PA)_{x,in}$
$\sum F_y = \sum F_{y,Press} + \sum F_{y,Body}$ (Force in *y* direction)	$(PA)_{y,out} - (PA)_{y,in} + \rho g$
$\sum F_z = \sum F_{z,Press} + \sum F_{z,Body}$ (Force in *z* direction)	$(PA)_{z,out} - (PA)_{z,in}$

The energy balance for two different types of zones must be considered. For an active server, it is,

$$X\dot{P}_n - \dot{m}_{s,n}c_P\left(T_{out,n} - T_{in,n}\right) = Y\frac{\partial T_{CPU,n}}{\partial t}, \tag{5-9}$$

where *n* denotes the server number, $\dot{m}_{s,n}$ is server mass flow rate, $\dot{P}$ denotes the total power consumption of the corresponding server, *X* is a coefficient that determines the power usage by CPUs, c_P denotes specific heat capacity, $T_{out,n}$ is server exhaust temperature, $T_{in,n}$ is the temperature of the corresponding cold chamber zone, *t* is time, and *Y* is the empirical coefficient for the thermal mass of a server available from the literature [13]. The other energy balance is for the airside within the in-row cooling unit,

$$\sum_j Q_{j\to i} + Q_{source} = \rho_i V_i c_p \frac{\partial T_i}{\partial t}, \tag{5-10}$$

where Q indicates heat flux, Q_{source} the internal heat source, V_i cell volume, ρ_i the air density, and T_i the air temperature at the inlet of a server. The inputs and outputs of the gray-box model are depicted in Figure 5-4. Eq. (5-9) shows that while energy can be converted from one form to another, the total energy within the control volume is constant. Eq. (5-10) provides the temperature change as server heat is added to the system. Integrated

forms of the mass, momentum and energy conservation laws (Eqs. (5-7)-(5-10)) are used to predict pressure, temperature, and mass flowrates.

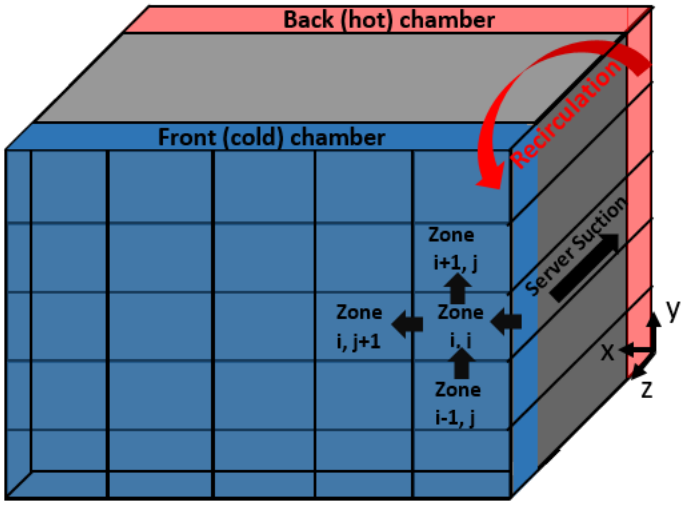


Figure 5-3. 3D zones inside the enclosure of a row-based cooling DC.

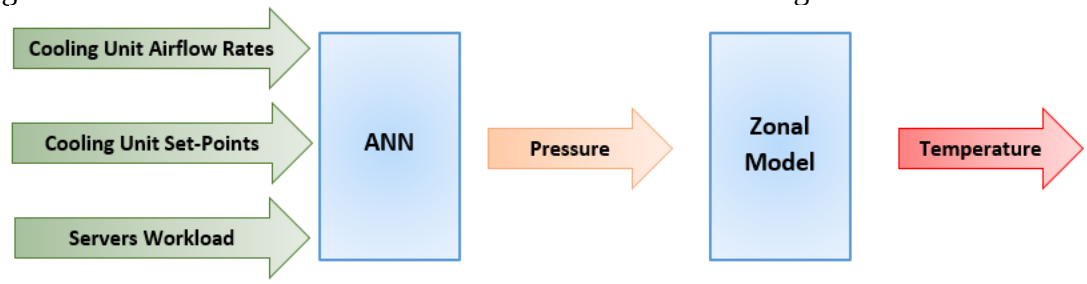


Figure 5-4. The data flow within the gray-box model for temperature predictions.

5.3.4 Failure detection

Next, we consider a use case of the gray-box model for transient temperature predictions during system failure. Failure detection is an essential aspect of highly reliable systems. To detect failures, a classification approach can be employed based on the different system behaviors during normal and failure conditions. In the classification problem, an instance associated with a set of attributes (features) is taken as input and the goal is to assign a class label (e.g., normal or abnormal) to that instance. Machine learning models such as

ANN can be applied to detect anomalies by learning the spatial and temporal characteristics of the temperature distributions for different conditions.

To gain an understanding of the kinds of features that can be used to distinguish different failure scenarios, we use the gray-box model to predict the temperature distributions in the cold chamber for normal operating condition and 10 minutes after various fan failures. Figure 5-5 shows the corresponding temperature values at different locations. Clearly, each type of fan failure, including the case of no failure, has a unique signature that can be used to infer which fan has stopped working.

We generate labeled training data for the gray-box model under normal and abnormal (e.g., fan failure) conditions in a row-based cooling DC using CFD simulations. Each instance in the training set consists of a collection of temperature readings at 25 locations every minute during a 10-minute interval and the corresponding label. Therefore, each observation is a 25x10 matrix. Furthermore, we use feature selection to determine a subset of 13 (out of the 25) sensor locations, which are the most informative for failure detection. The input to the classifier is reshaped to a 130x1 vector. We train an ANN with one input layer, several hidden layers, and one output layer. At run time, the ANN classifier takes real measurements from thermal sensors at target locations as inputs and predicts whether a fan failure has occurred and if so which fan it is. In this case study, we assume at any time, only one fan can fail.

5.4 Results and discussion

We evaluate the fidelity of CFD simulations and the proposed gray-box model, as well as the failure prediction algorithm using a row-based modular DC that is depicted in Figure

5-2. Twenty-five thermocouples are mounted along the cold chamber to collect temperature measurements.

5.4.1 CFD validation

To validate the CFD simulations, the airflow through the cooling unit is altered from a high value of 1.9 kg/s ($\dot{m}_{CU} \gg \sum \dot{m}_s$) to a lower value of 0.9 kg/s ($\dot{m}_{CU} \ll \sum \dot{m}_s$), and its set-point temperature from 18ºC to 16ºC after 300 seconds for 100% IT load at 20 kW. The transient temperatures from CFD simulations and experimental measurements over a period of 1200 seconds are shown in Figure 5-6, and the CFD simulations evaluated in Table 5-4 based on the following performance comparison metrics [53],

Maximum absolute error:

$$MAX_{AE} = MAX\left|T_{CFD} - T_{Exp}\right|. \tag{5-11}$$

Mean squared error:

$$MSE = \frac{1}{n}\sum\left(\left|T_{CFD} - T_{Exp}\right|\right)^2. \tag{5-12}$$

Standard deviation of absolute error:

$$STD_{AE} = \sqrt{\frac{1}{n-1}\sum\left(\left|T_{CFD} - T_{Exp}\right| - \frac{1}{n}\sum\left|T_{CFD} - T_{Exp}\right|\right)^2}. \tag{5-13}$$

Before any change in the cooling unit operating conditions, the maximum temperature difference between the CFD simulations and the experimental measurements at any location is smaller than 1.5% (0.3 °C), indicating relatively small errors. Hot spot formation is unlikely to occur for this case due to an oversupply of cold air, which produces a more uniform temperature distribution in the front chamber. After 600 seconds, MAX_{AE}, MSE and STD_{AE} increase slightly, where racks 3 and 4 show greater deviations of the

predictions from measurements due to hot air recirculation. At 1200 seconds, only a single zone has an 8% deviation from experiments with $MAX_{AE} \approx 1.83$ ºC, values for which are lower than 4.5 % for the remaining 20 zones. Therefore, we conclude that the CFD simulations provide reasonably accurate predictions of transient temperatures.

Table 5-4. Performance of CFD simulation.

	t = 0 Second					**t = 300 Seconds**				
	Rack 1	Rack 2	Rack 3	Rack 4	Rack 5	Rack 1	Rack 2	Rack 3	Rack 4	Rack 5
MAX_{AE}	0.19	0.29	0.29	0.22	0.18	0.18	0.24	0.19	0.20	0.18
MSE	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.01
Std_{AE}	0.04	0.12	0.09	0.06	0.05	0.04	0.10	0.07	0.06	0.07
	t = 600 Seconds					**t = 1200 Seconds**				
	Rack 1	Rack 2	Rack 3	Rack 4	Rack 5	Rack 1	Rack 2	Rack 3	Rack 4	Rack 5
MAX_{AE}	0.25	0.51	1.07	1.06	0.58	0.34	0.57	1.83	1.46	0.74
MSE	0.03	0.15	0.30	0.28	0.09	0.05	0.20	0.86	0.61	0.13
Std_{AE}	0.09	0.11	0.35	0.40	0.23	0.11	0.17	0.62	0.50	0.25

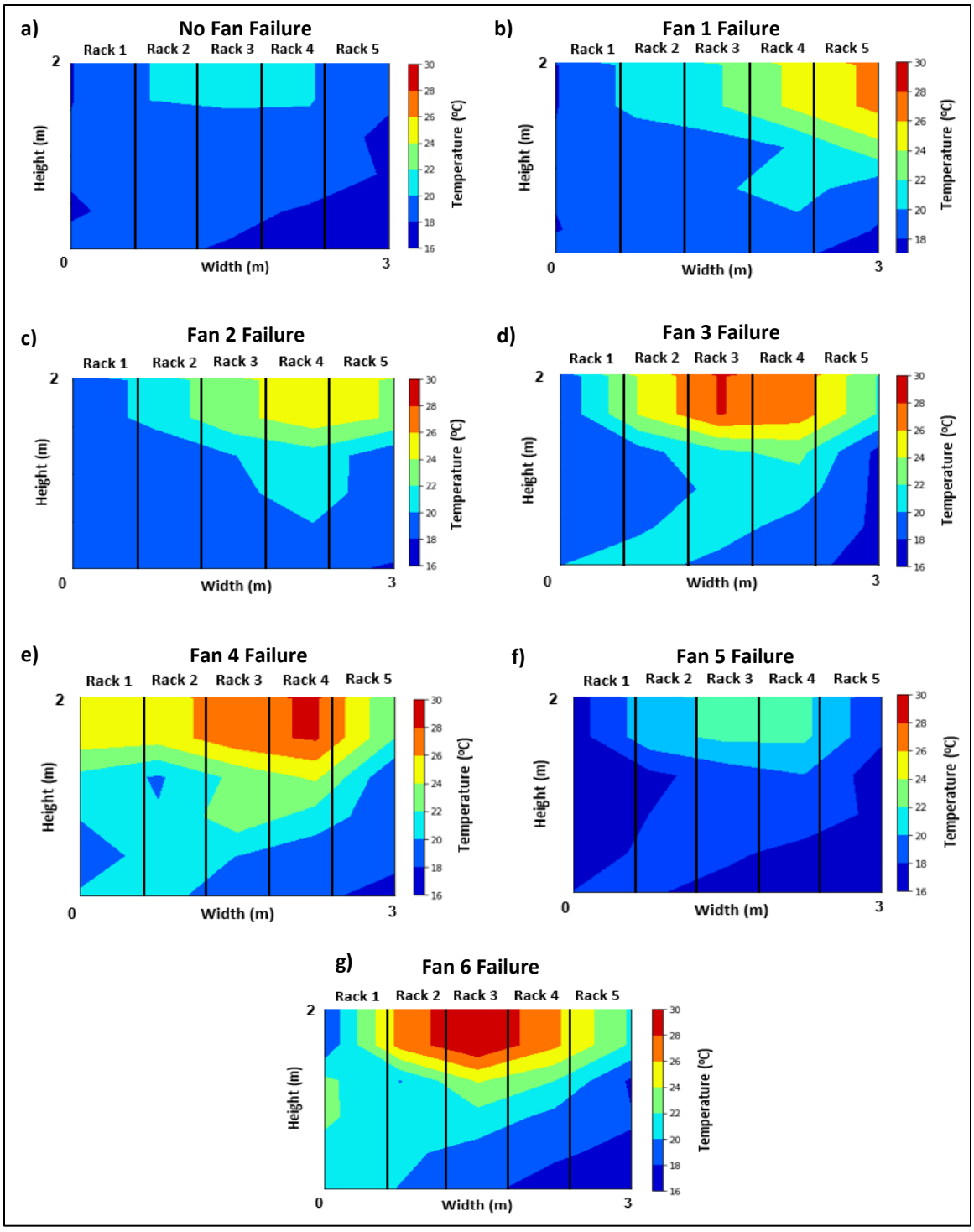


Figure 5-5. The temperature profiles for the normal thermal state and six thermal fault states induced by the cooling unit fans when the set-point temperature and server workloads are set to 17 °C and 100%, respectively.

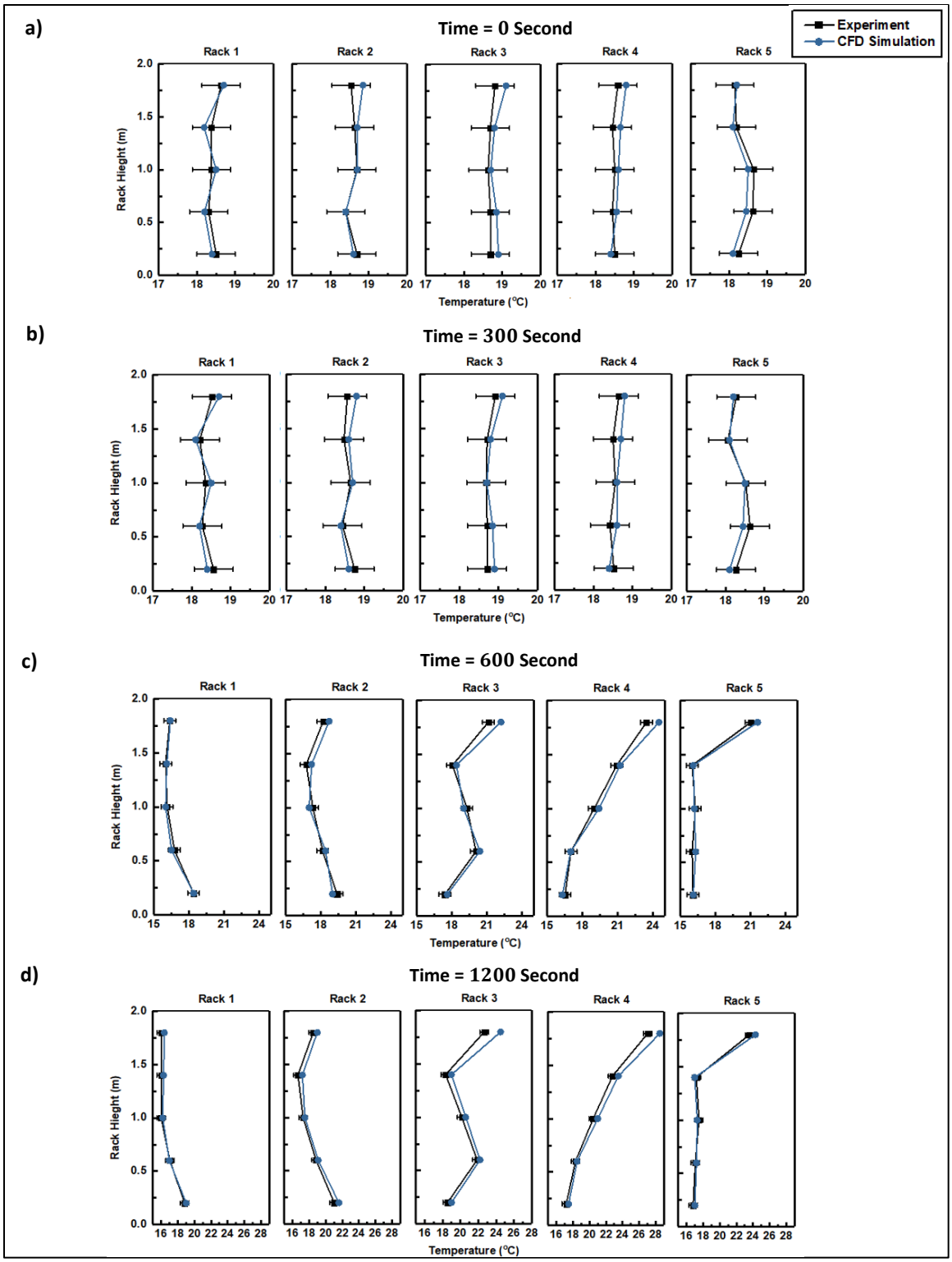


Figure 5-6. Temperature distributions provided by the CFD simulations and experimental measurements for 25 locations in the front chamber shown in Figure 5-2 (a).

5.4.2 Effects of training set size in the gray-box model

Using the CFD simulations, we produce a labeled dataset for cooling unit airflow changes in a row-based cooling DC. Each instance in the dataset consists of a collection of pressures at 60 locations for 56 airflows every second during a 10-seconds interval, i.e., a 56x60x10 tensor. We train the ANN model using different number of airflows to determine the optimum sample size that provides the best trade-off between accuracy and training time. Table 5-5 shows average prediction error as the input data sizes are varied from 24x60x10 to 56x60x10. The error decreases with increasing numbers of samples until a 48x60x10, beyond which the decrease in prediction error is negligible. Therefore, 48 out of a total 56 airflows are chosen.

5.4.3 Baseline black-box model

In this section, we introduce a state-of-the-art black-box thermal model that can predict inlet air and CPU temperatures at successive time steps in an in-row cooling unit DC [40, 54]. Denoting the input and output vectors at time *t* by x_t and y_t, respectively, x_t consists of the cooling unit operational parameters and server workloads, while y_t includes inlet air and CPU temperatures. Given inputs from time *t-m* to *t* and outputs (or measurements) from time *t* to *t-n*, the model predicts the output at time $t + 1$ using the function,

$$\hat{y}_{t+1} = f(x_t, x_{t-1}, \dots, x_{t-m}, y_t, y_{t-1}, \dots, y_{t-n}) \qquad \text{m = 1, 2, 3, … and n = 1, 2, 3, … .} \quad (5\text{-}14)$$

Table 5-5. Average train and test prediction errors as the sample size changes.

Training set size	Average train RMSE (Pa)	Average test RMSE (Pa)
24x60x10	3.72	6.02
28x60x10	2.54	4.24
32x60x10	1.62	3.57
36x60x10	1.05	2.19
40x60x10	0.87	1.40
44x60x10	0.51	0.74
48x60x10	0.39	0.48
52x60x10	0.38	0.51
56x60x10	0.37	0.65

The NARX model referred to previously is adopted. Specifically, a neural network with connections from both system inputs and feedbacks from outputs is used to model the nonlinearity, as shown in Figure 5-7. The closed-loop NARX network with embedded memory (tapped delay line) allows multi-step predictions. NARX is advantageous in modeling time-series data since the model (1) is better at discovering long time dependences, (2) is more effective at learning, (3) has faster convergence, (4) has negligible computational complexity, and (4) has scalability, making it applicable for large DCs [55-58].

To train the black-boxed model, labeled training data every minute over the past 720 s is used as shown in Figure 5-2.a. Realistic scenarios are considered to obtain the input parameters, i.e., changing workloads for 64 servers, the cooling unit airflows that have 6 fans, and the set point temperatures to specify the outputs, which are the inlet air temperatures for different zones and server CPU temperatures. Therefore, the input to the

neural network contains a 700x71 matrix. We train an open loop NARX neural network with a 50-neuron hidden layer, and update the weights and bias values according to Levenberg-Marquardt optimization [55].

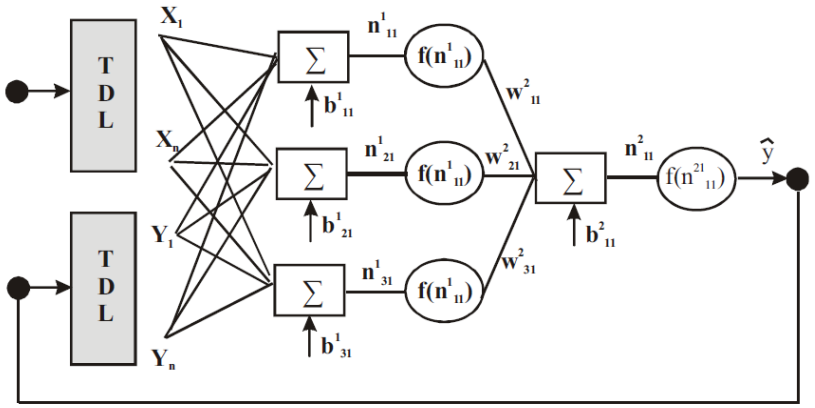


Figure 5-7. NARX neural network with tapped delay line (TDL) at the input (figure taken from [54]).

To develop the black-box model, temperature data is obtained from sensors at 20 second intervals and divided into training and validation sets to determine model parameters for different scenarios, such as changing the cooling unit operating conditions and server workloads. Once the model is trained, it is used to predict these same scenarios. To optimize the performance of the NARX neural network for time-series predictions in nonlinear systems, the hyper-parameters for feedback and the neural network should be carefully chosen. Another important consideration is the amount of training data. While it is expected that with training data obtained over a larger operational duration the model will likely capture the system dynamics and thus enable better predictions, this also compromises its ability to make early predictions due to a longer ramp-up period.

Figure 5-8 shows the average error for the black-box rack inlet temperature predictions in terms of deviations from temperature measurements as the durations over which the training data are obtained are varied from 300 to 700 seconds after an abrupt change in the cooling unit operation and IT load at 60 seconds. The error is large for shorter durations, but the larger data length of 700s provides a far more accurate solution. The errors for racks 1 and 5 are lower than for the other racks due to their proximity to the cooling units where the local temperatures are almost equal to the set-point temperature. For the remaining experiments, we select a training data duration of 700 seconds for the black-box model.

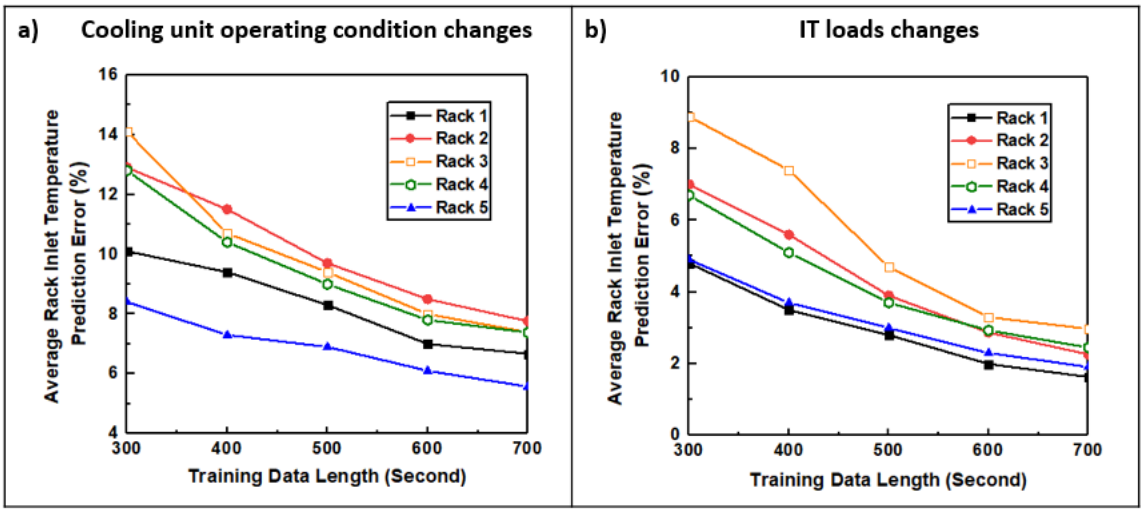


Figure 5-8. Average prediction error for the rack inlet temperature as a function of training data length in the black-box model when operating conditions change abruptly at 60 s. All predictions continue until a steady-state condition is reached.

5.4.4 Transient temperature predictions

The gray-box, conventional zonal, and baseline black-box models are employed to predict transient temperatures of a row-based cooling architecture until a steady-state condition is

reached. The predictions of the models are compared with sensor measurements for several realistic scenarios.

5.4.4.1 Influence of cooling unit operating conditions

The cooling unit operating conditions are critical for DC operation since both airflow paths and set-point temperatures influence the thermal environments within the cold and hot chambers. Reducing airflows or increasing the set-point temperature typically results in higher rack temperatures and thereby higher differences in air temperatures across the racks. For this set of experiments, we consider a high cooling unit airflow (80%) with an 18^{o}C set-point temperature for $t \leq t_0$. At times $t > t_0$, the cooling unit airflow and set-point temperature decrease to 30% and 16 ^{o}C, respectively, but the server workloads remain constant at 100%. Here, since the cold chamber pressure is lower than in the hot chamber, and hot air recirculation occurs through the gaps between the racks producing hot zones.

5.4.4.1.1 Server inlet air temperature prediction

Figure 5-9 compares temperature predictions at the 25 sensor locations in the cold chamber using the three models until a steady-state condition is reached. The temperatures are mostly uniform initially and the models have high accuracy, with lower than 0.3^{o}C differences between the predictions and experiments. In the black-box model, the first 700 s of data are used to train the model and test its interpolative accuracy, while the remainder are used to evaluate its extrapolative accuracy. The interpolative error associated with the model lies below 0.7^{o}C. As the model progresses beyond duration over which the training data are obtained, i.e., after 700 s, the black-box model predicts a sudden increase in temperature that causes larger differences between the predictions and measurements. In

contrast, deviations of the predicted temperatures from measurements are far smaller for the gray-box and conventional zonal models, and the average prediction error is about less than half that of the black-box model.

Figure 5-10 presents the average rack inlet temperature prediction errors for the aforementioned scenario with the three models, defined as $\Delta = \frac{|T_{Exp} - T_{Pred}|}{T_{Exp}}$. Before 700 s, the gray-box model, the conventional zonal model and the black-box model have a maximum Δ of 3.5%, 4.2% and 2.5%, respectively. After 700 s, Δ for the black-box model is larger than for the gray-box and conventional zonal models and for steady-state conditions, all racks show Δ < 4.3% for the gray-box model and Δ < 6.5% for the conventional zonal model while for the black-box model Δ > 20.42%. Therefore, the gray-box model is much more accurate than the conventional zonal model and the black-box model for extrapolative predictions. Figure 5-11 provides transient temperature predictions for six sensors at arbitrarily chosen different locations.

We also observe differences in the prediction accuracies at different rack locations. For example, Δ values are lower for the side racks (racks 1 and 5) since they are close to the cooling units while for the middle racks (racks 2, 3, and 4) Δ is larger due to hot air recirculation, which is not accurately modeled.

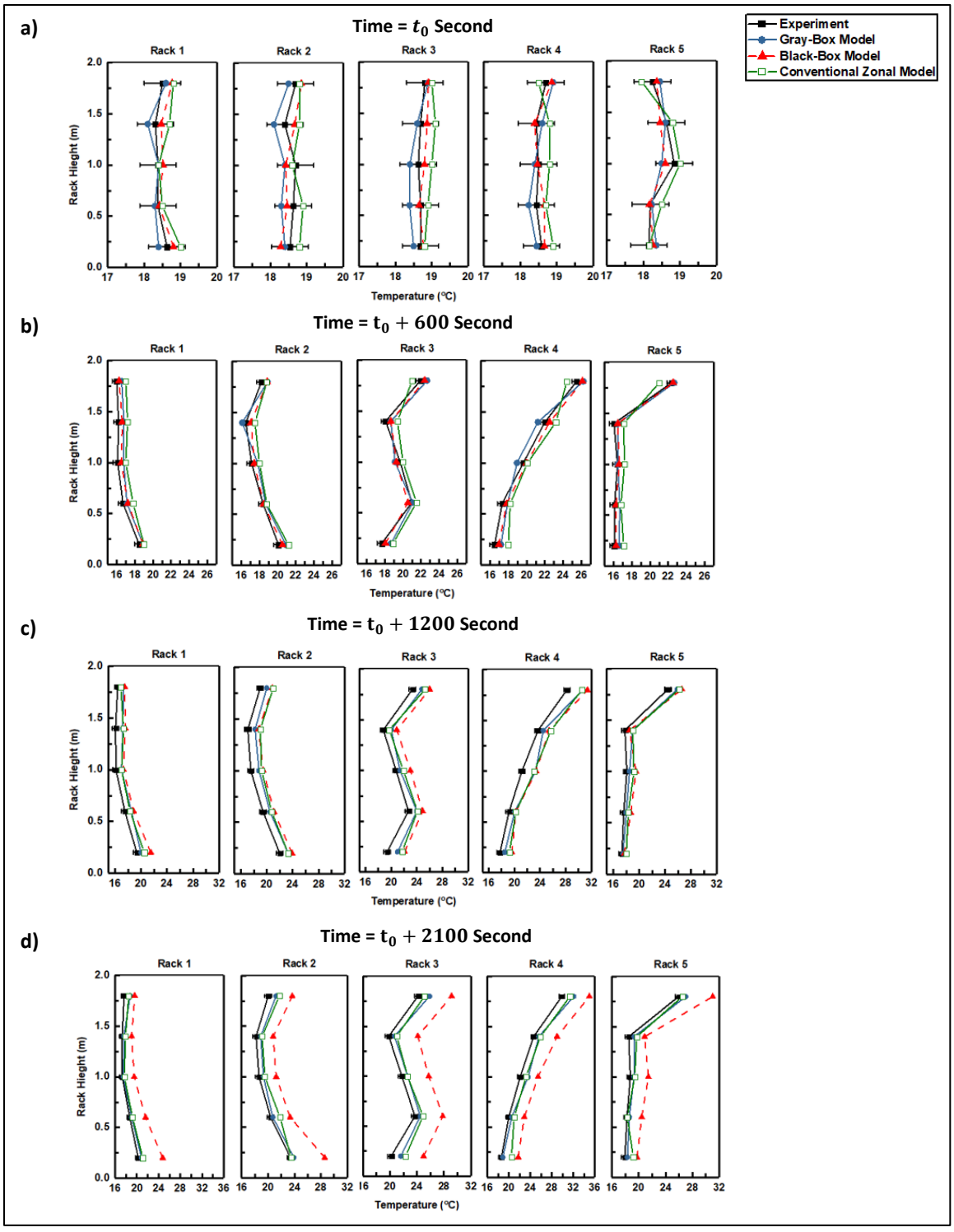


Figure 5-9. Online temperature predictions of the gray-box model (blue solid line), a conventional zonal model (green solid line) and a black-box model (red dash line) versus temperature measurements from experiments (black solid line) in response to an abrupt change in the cooling unit operation at $t = t_0 + 60$ s.

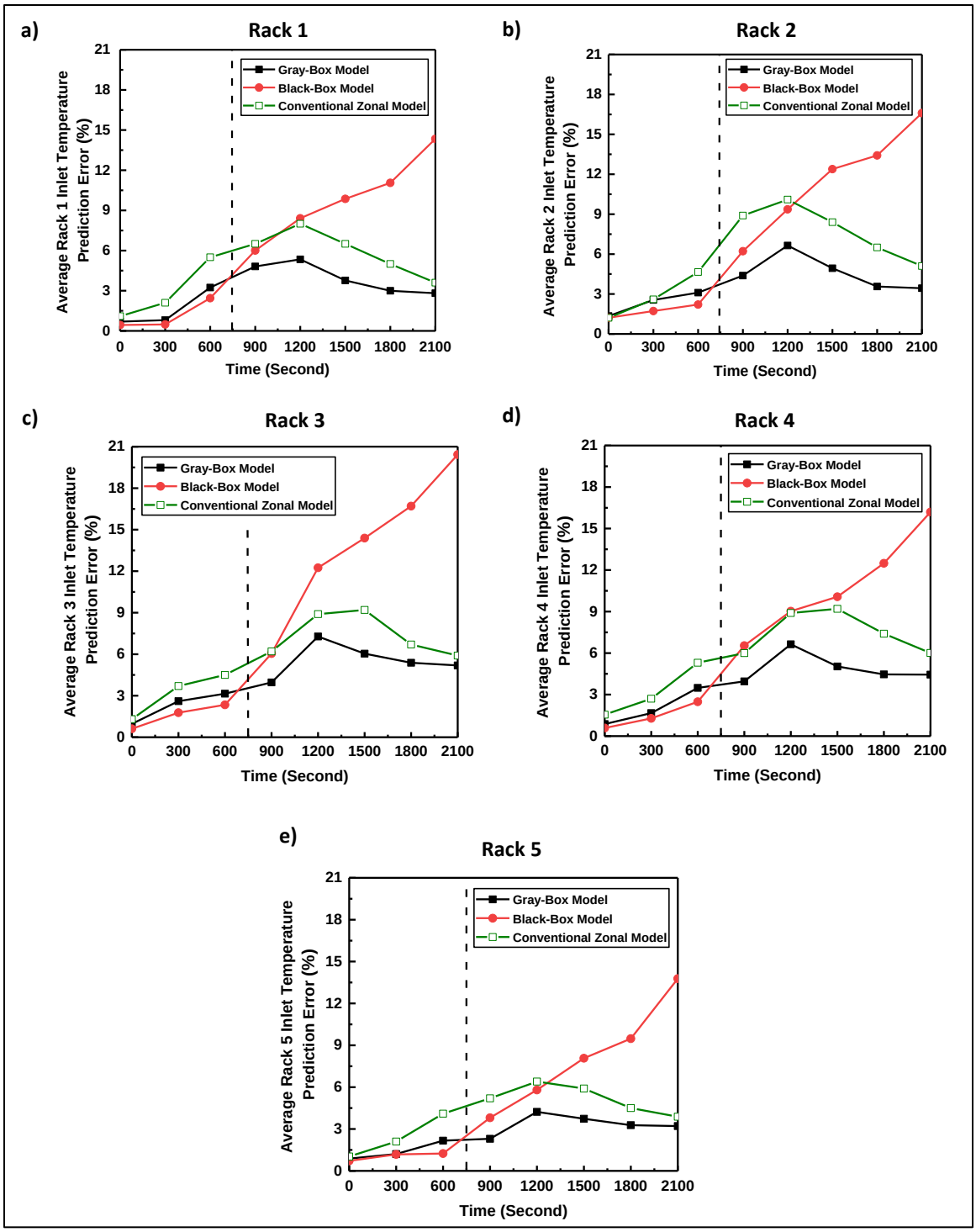


Figure 5-10. Performance comparison: Temperature prediction errors for three models with respect to experimental results, $\Delta = \frac{|T_{Exp} - T_{Pred}|}{T_{Exp}}$, when the cooling unit operation changes abruptly at 60 s.

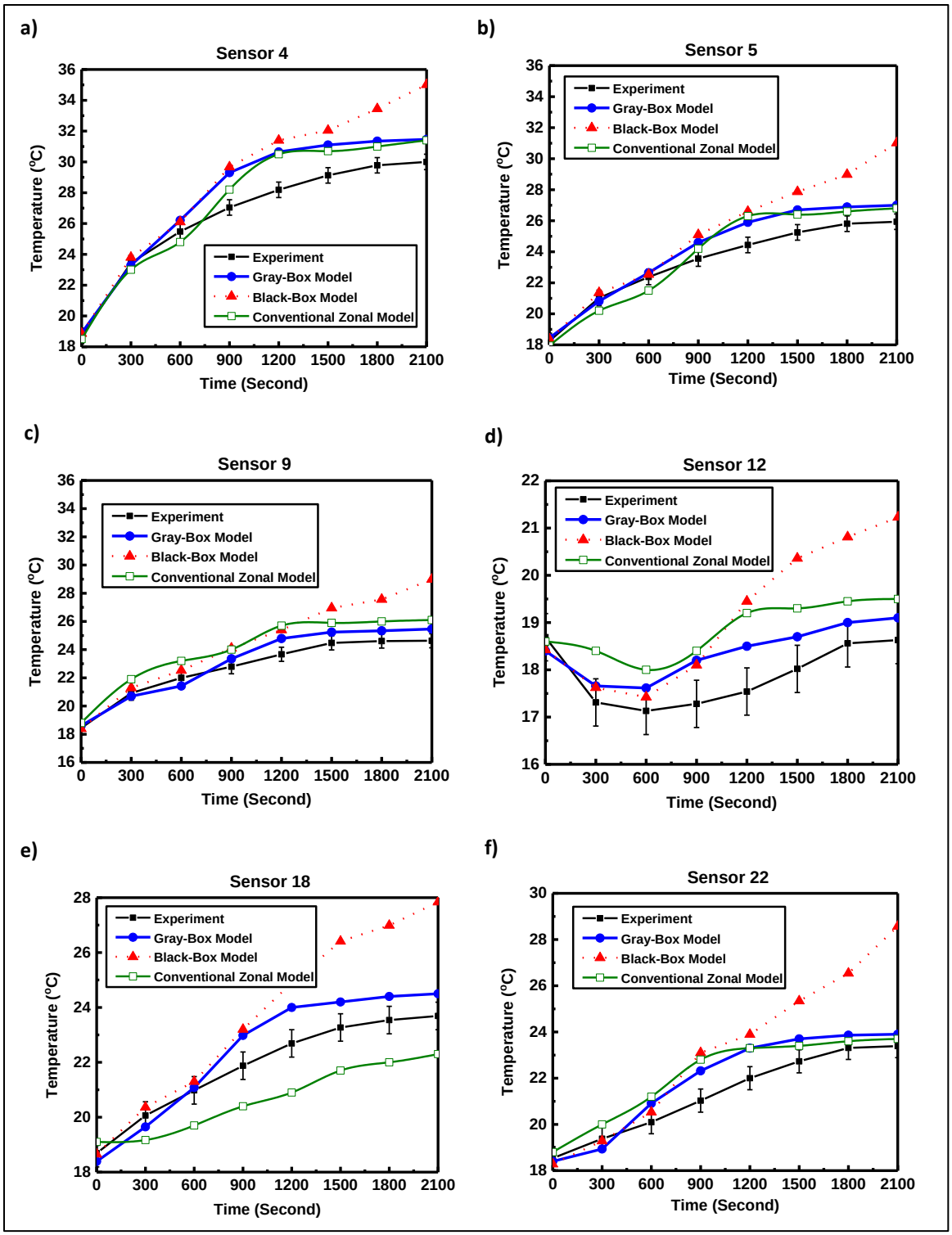


Figure 5-11. Transient temperature predictions from three models at different sensor locations when the cooling unit operation changes abruptly at 60 s.

5.4.4.1.2 Server CPU temperature prediction

Here, the CPU temperatures are predicted by the gray-box and black-box models until a steady-state condition. Since the conventional zonal model developed in [13] does not predict CPU temperatures, its results are excluded in this set of experiments. Figure 5-12 presents CPU temperatures prediction results for two arbitrary chosen servers. Server 23 is located at the top of rack 3 and server 45 in the middle of rack 4. The gray-box model has a higher accuracy than the black-box model for all times. The values of Δ for both models for these two servers are summarized in Table 5-6.

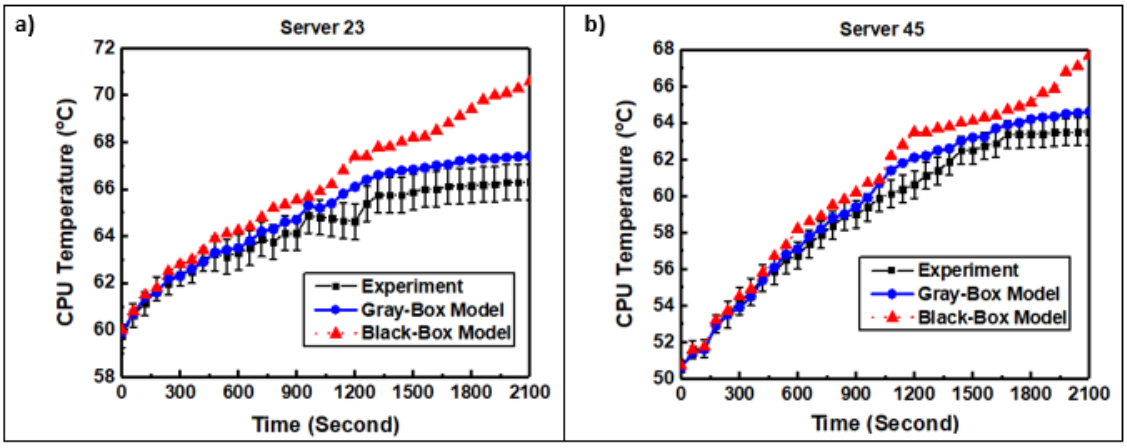


Figure 5-12. CPU temperature predictions from the gray-box and black-box models in response to a change in the cooling unit operation at 60 s until a steady-state condition is reached.

Table 5-6. Relative temperature prediction error between the black-box and gray-box models and experimental measurements for servers 23 and 45 in response to a change in the cooling unit operation at 60 s.

	$\Delta_{\mathbf{Gray-box}}$ (%)	$\Delta_{\mathbf{Black-box}}$ (%)
Server 23	1.02	2.71
Server 45	0.91	2.41

5.4.4.2 Influence of IT load on temperature distribution

When $t \leq t_0$, all servers operate at a 20% workload (8.6 kW in total) while at time $t > t_0$ the server workloads increase to 100% (20 kW in total). Both cases have cooling unit airflows of 1.2 kg/s and a set-point temperature of 18°C.

5.4.4.2.1 Server inlet air temperature prediction

Cold chamber temperature predictions from the gray-box model, the conventional zonal model and the black-box model are shown in Figure 5-13. Before $t_0 + 700$ seconds, all models provide good temperature predictions, but after $t_0 + 700$ the error from the black-box model increases significantly over time, while it increases initially moderately for the gray-box and conventional zonal models but then decreases as the system approaches a steady-state condition. Similar observations can be made from Figure 5-14, which shows the Δ values from the three models. Finally, Figure 5-15 shows the transient temperature predictions for six sensors, which are chosen arbitrarily at different locations.

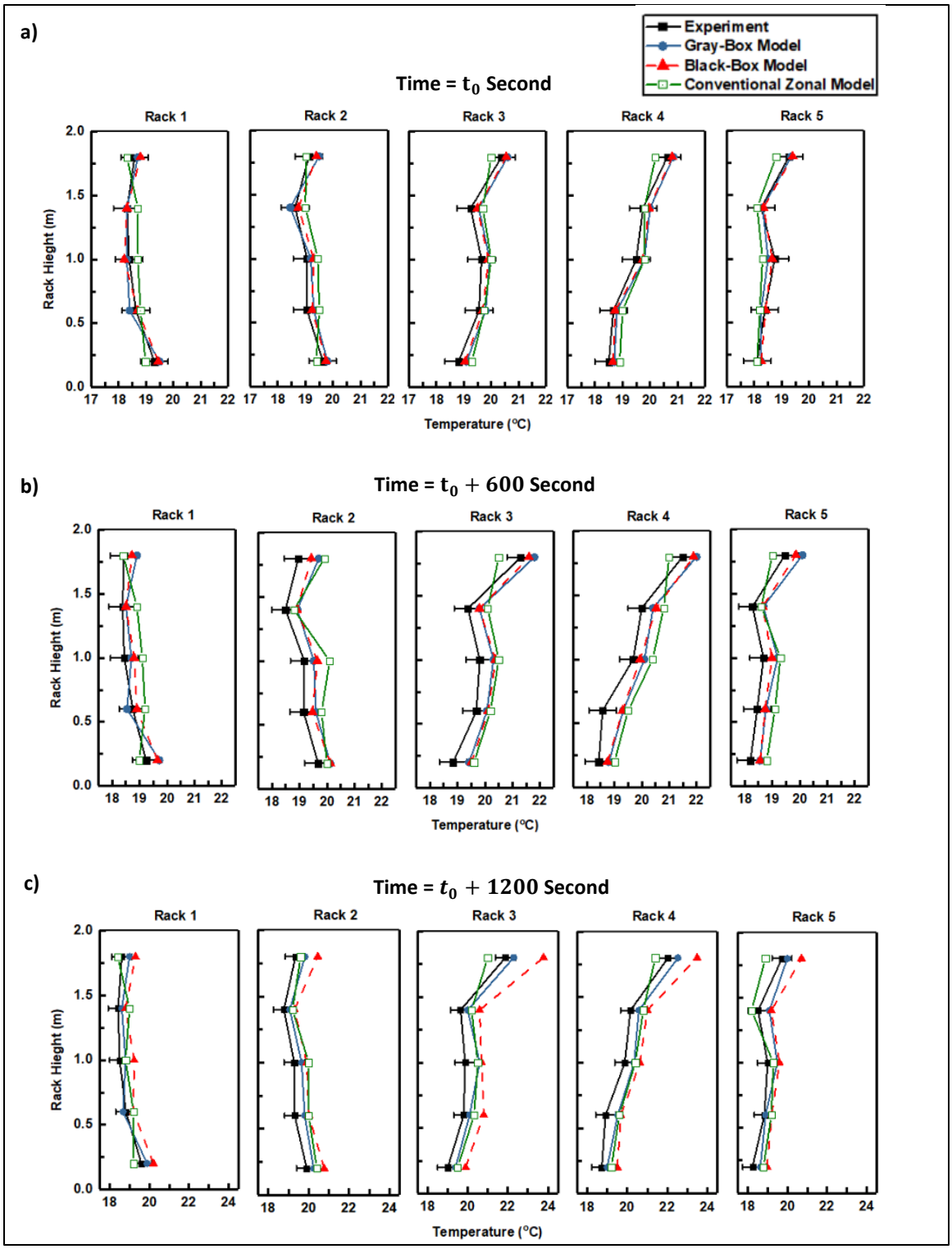


Figure 5-13. Online temperature predictions from the gray-box (blue solid line) and black-box (red dash line) models and experimental temperature measurements (black solid line) in response to a change in the server workloads at $t = t_0 + 60$ s.

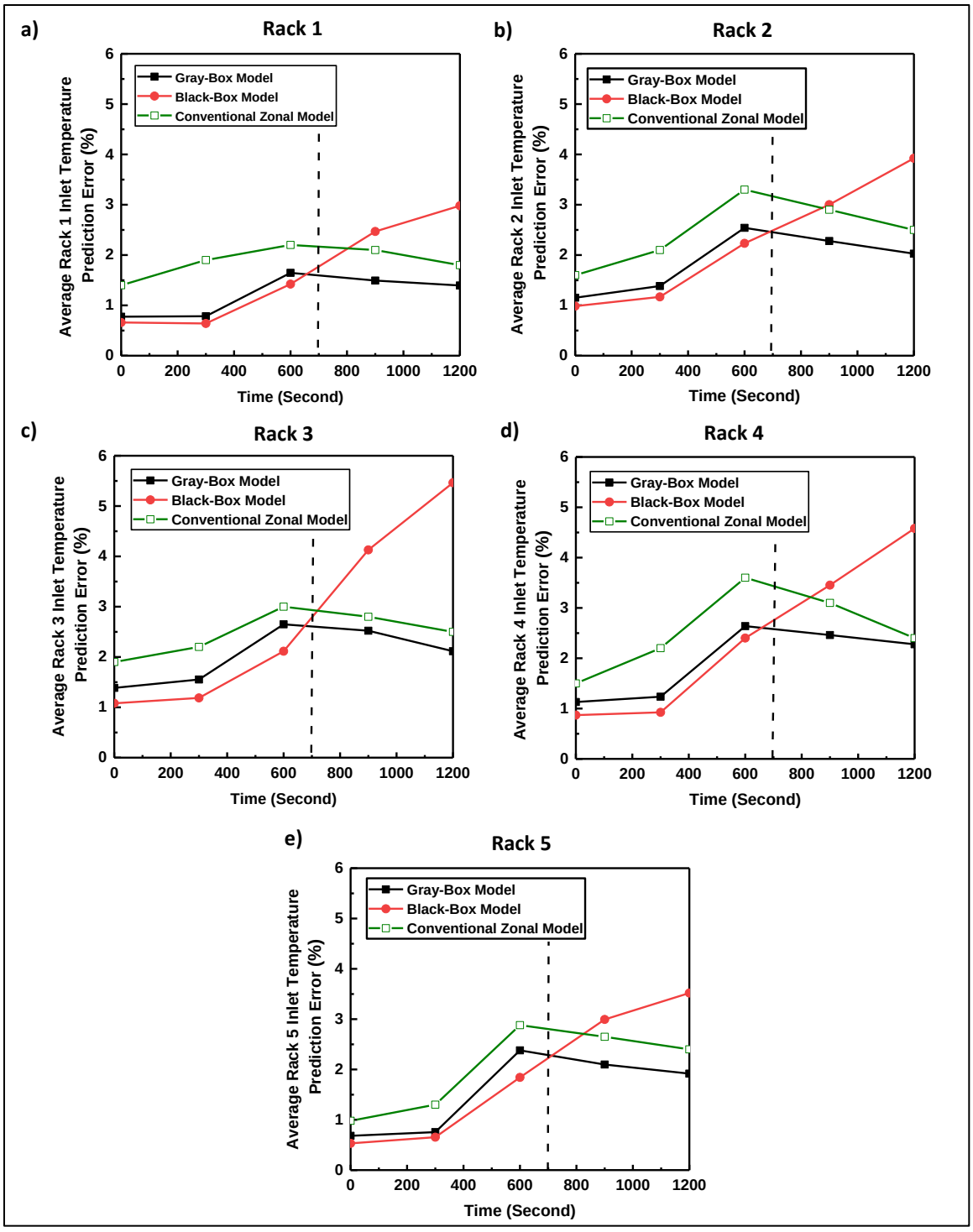


Figure 5-14. Performance comparison: Temperature prediction errors for three models with respect to the experimental result, $\Delta = \frac{|T_{Exp} - T_{Pred}|}{T_{Exp}}$, when the server workloads change at time 60 s.

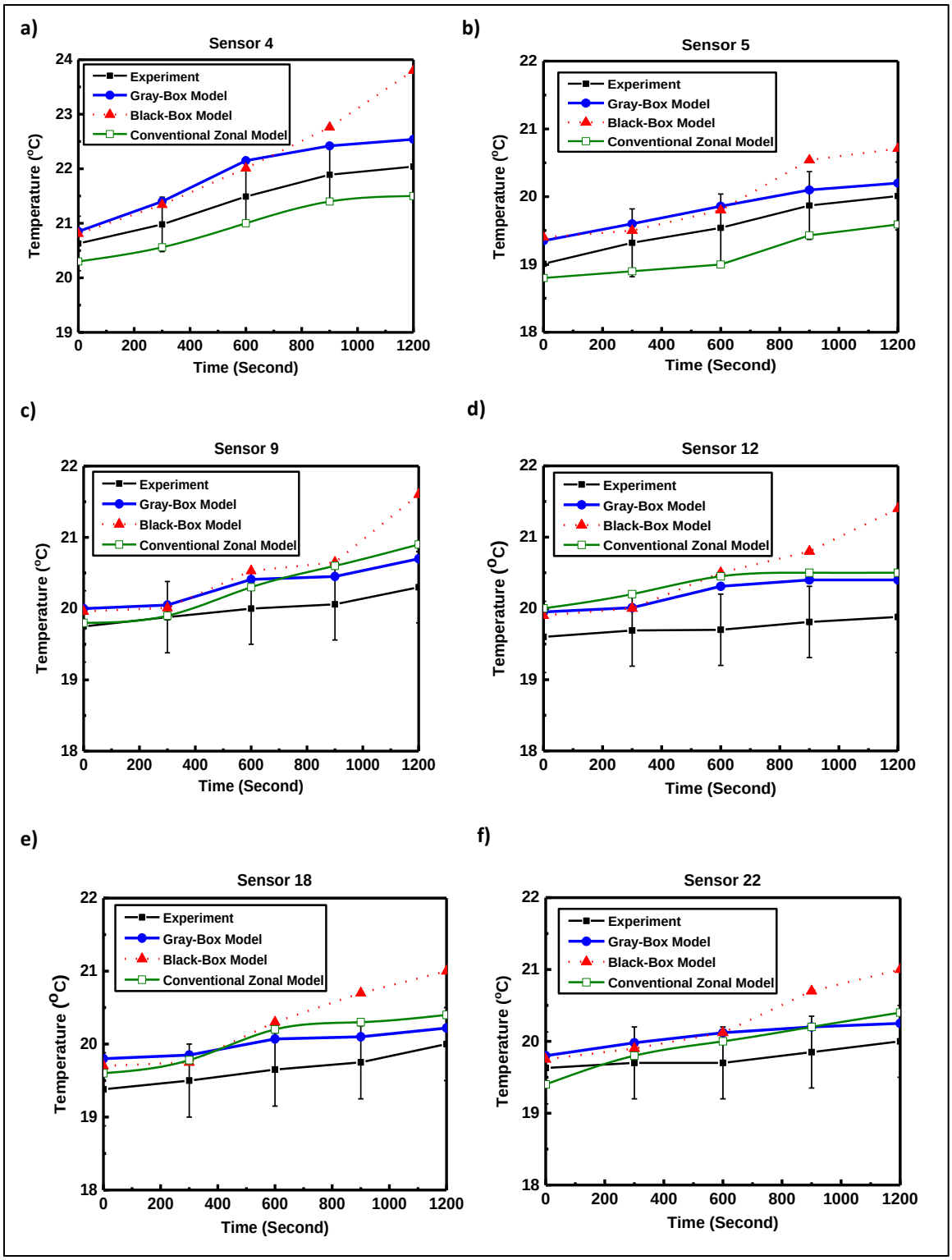


Figure 5-15. Transient temperature predictions from three models at different sensor locations when the server utilization changes at time 60 s.

5.4.4.2.2 Server CPU temperature prediction

We examine the predicted CPU temperatures for two arbitrarily chosen servers (23 and 45) as their workload increases from 20% to 100%. Since the conventional zonal model [13] does not predict CPU temperatures, its results are excluded in this set of experiments. The CPU temperature prediction results are presented in Figure 5-16 and values of Δ are provided in Table 5-7. The gray-box model is in far better agreement with the experiment results.

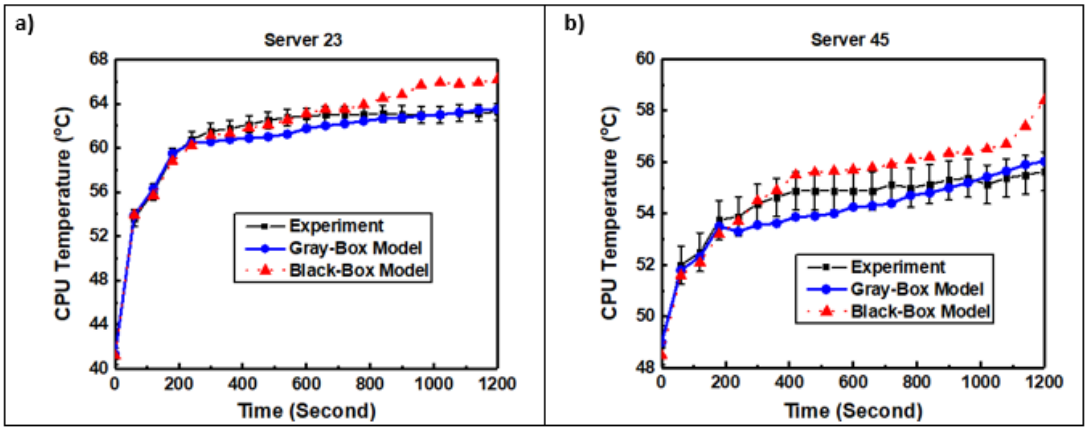


Figure 5-16. CPU temperature predictions in response to a change in the server workload at time 60 s until a steady-state condition is reached from the gray-box and black-box models.

Table 5-7. Relative errors in the temperature predictions between the black-box and gray-box models and experimental results for servers 23 and 45 in response to a change in the server workload at time 60 s.

	$\Delta_{\mathbf{Gray-box}}$ (%)	$\Delta_{\mathbf{Black-box}}$ (%)
Server 23	0.96	1.74
Server 45	0.91	1.61

5.4.5 Thermal anomaly detection and fault classification

Recall from Section 2.4 that a neural network is trained to classify different failure scenarios. Table 5-8 summarizes the train and test accuracy rate of the ANN with different numbers of hidden layers and neurons in each layer. In the experiments, the ReLu activation function is used in the output layer. We find that the ANN with 4 hidden layers and 5 neurons has the best performance with the test accuracy and error rate equal to 95% and 5%, respectively. Further increasing the number of layers or neurons leads to overfitting.

Table 5-8. Parameters of the ANN classification model and its accuracy.

# Hidden layer	# Neurons	Train accuracy	Test Accuracy
1	3	47%	28%
1	5	48%	31%
2	3	54%	42%
2	5	61%	54%
2	7	74%	69%
3	3	88%	81%
3	5	96%	89%
3	10	100%	81%
4	3	97%	90%
4	5	98%	95%
4	10	100%	89%
5	3	100%	91%
5	5	100%	85%

We evaluate the performance of the fault detection and classification model in identifying cooling unit fan failures. Precision and recall metrics can be used to evaluate the classification performance. Table 5-9 provides the precision, recall, and F_{score} of the

resulting ANN classifier. The table indicates the no fan failure condition and the failure of the 1st fan have the highest F_{score} of 0.90, implying that these classes can easily be distinguished from the rest, while the failure of fan 4 has the least F_{score} of 0.74. The overall precision, recall, and F_{score} for the seven conditions is 0.84.

A confusion matrix is provided in Table 5-10 which helps determine misclassified cases, where the columns represent the actual class for the class number and rows indicate the predicted class. The elements of the diagonal contain the total number of correct predictions in each class and the remaining entries summarize the number of misclassifications into other classes.

Table 5-9. Multi-class classification precision, recall, and F_{score} for fans failure using ANN classifier.

Class number	Class definition	Precision	Recall	F_{score}
1	No fan failure	0.90	0.90	0.90
2	Fan 1 failure	0.90	0.90	0.90
3	Fan 2 failure	0.89	0.80	0.84
4	Fan 3 failure	0.80	0.80	0.80
5	Fan 4 failure	0.78	0.70	0.74
6	Fan 5 failure	0.82	0.90	0.86
7	Fan 6 failure	0.82	0.90	0.86
Overall		0.84	0.84	0.84

Table 5-10. Confusion matrix for the studied multi-class classification problem.

		Actual class						
		Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Predicted class	Class 1	9	0	0	0	0	1	0
	Class 2	0	9	1	0	0	0	0
	Class 3	0	1	8	0	0	0	0
	Class 4	0	0	0	8	2	0	0
	Class 5	0	0	0	1	7	0	1
	Class 6	1	0	1	0	0	9	0
	Class 7	0	0	0	1	1	0	9

5.5 Quantitative and quantitative comparison with related works

Table 5-11 presents a summary of key differences between previous work and our proposed method. As previously discussed, computation time and fidelity are two crucial performance metrics for different models in real-time control and fault detection applications. Although CFD simulations have relatively high spatial resolution, they provide reasonable results only after several hours of numerical computation. In contrast, depending on the geometry and projected duration of the prediction, conventional zonal models take a few as tens of seconds to a few minutes to provide useful temperature predictions, although test cases require over an hour of run time to calibrate, optimize and tune the model. An investigation of a conventional zonal model reports that it is possible to include passive server effects [13], although a passive server has a small influence on rates of changes in temperatures and is also not common in a DC because its inclusion reduces reliability. Thus, without significantly influencing the accuracy of the results, we neglect passive servers.

For the tested DC scenario, our black-box and gray-box models can make predictions in less than 0.5 seconds on a personal computer with a Core i7-8700 3.20 GHz CPU and 16 GB memory. Of these two models, the gray-box model has a much higher extrapolative prediction accuracy, is adaptive to changes in DC operation, and can reproduce airflow leakage, such as hot air recirculation and cold air bypass [51]. Since the primary objective of this study is to predict rack inlet temperatures rapidly and accurately, the lower spatial resolution is acceptable for our purposes.

Table 5-11. A comparison of DC temperature prediction models in present and past studies.

Assessment criteria	CFD simulations [44, 59, 60]	Conventional zonal model [13]	Black-box model [28, 32, 33]	Present work
Computation time for one specific scenario	> 1 hours	> 30 seconds	< 10 seconds	< 0.5 seconds
Required time for training/calibrating/setup the model	> 5 hours	> 1 hours	> 1 hour	< 1 hour
Spatial resolution	High (<1 mm)	Low (> 20 cm)	NA	Low (> 20 cm)
Duration of testing/training datasets	NA	NA	> 1 hour	< 1 hour
Interpolative error	NA	NA	< 0.5 ^{o}C	< 0.5 ^{o}C
Extrapolative error	NA	NA	> 1 ^{o}C	< 1.5 ^{o}C
Ability to capture special features				
• Adaptive to changes	No	Yes	No	Yes
• Able to capture airflow leakages through the gaps	No	Yes	No	Yes
• Effect of passive servers	No	Yes	No	No

5.6 Conclusion

We develop a tool for designers and operators to successfully plan, operate, and control the transient behavior of a DC. The hybrid model combines conventional thermodynamics laws with intelligent algorithms to provide real-time temperature predictions of server CPUs and the cold chamber in a modular DC with a row-based cooling architecture. Model performance is compared against a conventional zonal model and an advanced data-driven black-box model for two scenarios, i.e., (1) changes in cooling unit operation and (2) varying server workload. Our findings are summarized below.

1. When the cooling unit fan speeds are changed from 80% to 30% (decreasing the airflow from 1.92 kg/s to 0.96 kg/s) and set-point temperatures decreased from 18^{o}C to 16^{o}C, all models provide accurate predictions with low interpolative errors having values below 0.7^{o}C. However, the black-box model has higher extrapolative errors, while the gray-box and conventional zonal models error decreases by half

as the system reaches an eventual steady state. Hence, the gray-box and conventional zonal models outperforms the black-box model for extrapolative predictions.

2. With the server workload incremented by 80%, all models initially yield a low average error for inlet air temperature predictions. However, this error increases significantly for the black-box model but decreases for the gray-box and conventional zonal model as the system reaches a steady state. Errors in predictions of the server CPU temperatures with the gray-box model are approximately half of those with the black-box model. Overall, the gray-box model again has a better prediction ability than the conventional zonal and black-box models.
3. Prediction errors for the middle rack (rack 3) are higher due to hot air recirculation than for the end racks (racks 1 and 5) since these racks are adjacent to the cooling units.
4. The gray-box model is applied to detect fan failures. Experimental results demonstrate that the classifier trained using predictions from the gray-box model achieve precision, recall, and a F_{score} of 0.84 for one normal and 6 abnormal conditions.

We conclude that the performance of the gray-box model is superior to that of a pure data-driven black-box and conventional zonal models. Our future work will investigate applications of the gray-box model for early fault detection and diagnosis, thermal-aware workload management and tests of what-if scenarios to characterize the

influence of operating conditions on the CPU and inlet air temperature distribution, and model-predictive control for the operation of cooling units.

5.7 Acknowledgment

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under a collaborative research and development (CRD) project, Computationally efficient Surrogate Models. We thank colleagues from CINNOS Mission Critical Incorporated who provided insight and expertise.

5.8 References

[1] Y. Li, X. Wang, P. Luo, and Q. Pan, "Thermal-aware hybrid workload management in a green datacenter towards renewable energy utilization," *Energies*, vol. 12, no. 8, p. 1494, 2019.

[2] S. MirhoseiniNejad, H. Moazamigoodarzi, G. Badawy, and D. G. Down, "Joint data center cooling and workload management: A thermal-aware approach," *Future Generation Computer Systems*, vol. 104, pp. 174-186, 2020.

[3] D. Andrews and B. Whitehead, "Data Centres in 2030: Comparative Case Studies that Illustrate the Potential of the Design for the Circular Economy as an Enabler of Sustainability," in *Sustainable Innovation 2019: 22nd International Conference Road to 2030: Sustainability, Business Models, Innovation and Design*, 2019.

[4] M. Salim and R. Tozer, "Data Centers' Energy Auditing and Benchmarking-Progress Update," *ASHRAE transactions*, vol. 116, no. 1, 2010.

[5] H. Lu, Z. Zhang, and L. Yang, "A review on airflow distribution and management in data center," *Energy and Buildings*, vol. 179, pp. 264-277, 2018.

[6] K. Ebrahimi, G. F. Jones, and A. S. Fleischer, "A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities," *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 622-638, 2014.

[7] H. M. Daraghmeh and C.-C. Wang, "A review of current status of free cooling in datacenters," *Applied Thermal Engineering*, vol. 114, pp. 1224-1239, 2017.

[8] R. Gupta, S. Asgari, H. Moazamigoodarzi, S. Pal, and I. K. Puri, "Cooling architecture selection for air-cooled Data Centers by minimizing exergy destruction," *Energy*, p. 117625, 2020.

[9] A. Carbó, E. Oró, J. Salom, M. Canuto, M. Macías, and J. Guitart, "Experimental and numerical analysis for potential heat reuse in liquid cooled data centres," *Energy Conversion and Management*, vol. 112, pp. 135-145, 2016.

[10] T. Gao, M. David, J. Geer, R. Schmidt, and B. Sammakia, "Experimental and numerical dynamic investigation of an energy efficient liquid cooled chiller-less data center test facility," *Energy and buildings*, vol. 91, pp. 83-96, 2015.

[11] T. Gao, M. David, J. Geer, R. Schmidt, and B. Sammakia, "A dynamic model of failure scenarios of the dry cooler in a liquid cooled chiller-less data center," in *2015 31st Thermal Measurement, Modeling & Management Symposium (SEMI-THERM)*, 2015, pp. 113-119: IEEE.

[12] J. Dai, M. M. Ohadi, D. Das, and M. G. Pecht, *OPTIMUM COOLING OF DATA CENTERS*. Springer, 2016.

[13] H. Moazamigoodarzi, R. Gupta, S. Pal, P. J. Tsai, S. Ghosh, and I. K. Puri, "Modeling temperature distribution and power consumption in IT server enclosures with row-based cooling architectures," *Applied Energy*, vol. 261, p. 114355, 2020.

[14] K. Dunlap and N. Rasmussen, "Choosing between room, row, and rack-based cooling for data centers," *APC White Paper*, vol. 130, 2012.

[15] T. Evans, "The different types of air conditioning equipment for IT environments," *White Paper*, vol. 59, pp. 2004-0, 2004.

[16] J. Cho, J. Yang, and W. Park, "Evaluation of air distribution system's airflow performance for cooling energy savings in high-density data centers," *Energy and buildings*, vol. 68, pp. 270-279, 2014.

[17] I.-N. Wang, Y.-Y. Tsui, and C.-C. Wang, "Improvements of airflow distribution in a container data center," *Energy Procedia*, vol. 75, pp. 1819-1824, 2015.

[18] J. Cho and B. S. Kim, "Evaluation of air management system's thermal performance for superior cooling efficiency in high-density data centers," *Energy and buildings*, vol. 43, no. 9, pp. 2145-2155, 2011.

[19] M. K. Patterson, R. Weidmann, M. Leberecht, M. Mair, and R. M. Libby, "An investigation into cooling system control strategies for data center airflow containment architectures," in *International Electronic Packaging Technical Conference and Exhibition*, 2011, vol. 44625, pp. 479-488.

[20] J. Priyadumkol and C. Kittichaikarn, "A Study of Air Flow through Perforated Tile for Air Conditioning System in Data Center," in *Applied Mechanics and Materials*, 2013, vol. 249, pp. 126-131: Trans Tech Publ.

[21] R. Schmidt and M. Iyengar, "Server rack rear door heat exchanger and the new ASHRAE recommended environmental guidelines," in *International Electronic Packaging Technical Conference and Exhibition*, 2009, vol. 43604, pp. 851-862.

[22] R. K. Sharma, C. E. Bash, C. D. Patel, R. J. Friedrich, and J. S. Chase, "Balance of power: Dynamic thermal management for internet data centers," *IEEE Internet Computing*, vol. 9, no. 1, pp. 42-49, 2005.

[23] Q. Tang, T. Mukherjee, S. K. Gupta, and P. Cayton, "Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters," in *2006 Fourth International Conference on Intelligent Sensing and Information Processing*, 2006, pp. 203-208: IEEE.

[24] H. S. Erden, H. E. Khalifa, and R. R. Schmidt, "A hybrid lumped capacitance-CFD model for the simulation of data center transients," *Hvac&R Research,* vol. 20, no. 6, pp. 688-702, 2014.

[25] H. Moazamigoodarzi, S. Pal, S. Ghosh, and I. K. Puri, "Real-time temperature predictions in it server enclosures," *International Journal of Heat and Mass Transfer,* vol. 127, pp. 890-900, 2018.

[26] Z. Song, B. T. Murray, and B. Sammakia, "A compact thermal model for data center analysis using the zonal method," *Numerical Heat Transfer, Part A: Applications*, vol. 64, no. 5, pp. 361-377, 2013.

[27] R. Zhou, Z. Wang, C. E. Bash, and A. McReynolds, "Data center cooling management and analysis-a model based approach," in *2012 28th Annual IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, 2012, pp. 98-103: IEEE.

[28] J. Athavale, Y. Joshi, and M. Yoda, "Artificial neural network based prediction of temperature and flow profile in data centers," in *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2018, pp. 871-880: IEEE.

[29] J. Moore, J. S. Chase, and P. Ranganathan, "Weatherman: Automated, online and predictive thermal mapping and management for data centers," in *2006 IEEE international conference on Autonomic Computing*, 2006, pp. 155-164: IEEE.

[30] M. Zapater, J. L. Risco-Martín, P. Arroba, J. L. Ayala, J. M. Moya, and R. Hermida, "Runtime data center temperature prediction using Grammatical Evolution techniques," *Applied Soft Computing,* vol. 49, pp. 94-107, 2016.

[31] L. Wang, G. von Laszewski, F. Huang, J. Dayal, T. Frulani, and G. Fox, "Task scheduling with ANN-based temperature prediction in a data center: a simulation-based study," *Engineering with Computers,* vol. 27, no. 4, pp. 381-391, 2011.

[32] R. Lloyd and M. Rebow, "Data driven prediction model (ddpm) for server inlet temperature prediction in raised-floor data centers," in *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2018, pp. 716-725: IEEE.

[33] J. Athavale, M. Yoda, and Y. Joshi, "Comparison of data driven modeling approaches for temperature prediction in data centers," *International Journal of Heat and Mass Transfer,* vol. 135, pp. 1039-1052, 2019.

[34] Z. Song, B. T. Murray, and B. Sammakia, "A dynamic compact thermal model for data center analysis and control using the zonal method and artificial neural networks," *Applied thermal engineering,* vol. 62, no. 1, pp. 48-57, 2014.

[35] L. Li, C.-J. M. Liang, J. Liu, S. Nath, A. Terzis, and C. Faloutsos, "Thermocast: a cyber-physical forecasting model for datacenters," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1370-1378.

[36] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, "A cyber–physical systems approach to data center modeling and control for energy efficiency," *Proceedings of the IEEE,* vol. 100, no. 1, pp. 254-268, 2011.

[37] J. Chen *et al.*, "A high-fidelity temperature distribution forecasting system for data centers," in *2012 IEEE 33rd Real-Time Systems Symposium*, 2012, pp. 215-224: IEEE.

[38] L. Parolini, B. Sinopoli, and B. H. Krogh, "Model predictive control of data centers in the smart grid scenario," *IFAC Proceedings Volumes,* vol. 44, no. 1, pp. 10505-10510, 2011.

[39] E. Pakbaznia, M. Ghasemazar, and M. Pedram, "Temperature-aware dynamic resource provisioning in a power-optimized datacenter," in *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*, 2010, pp. 124-129: IEEE.

[40] S. MirhoseiniNejad, F. M. García, G. Badawy, and D. G. Down, "ALTM: Adaptive learning-based thermal model for temperature predictions in data centers," in *2019 IEEE Sustainability through ICT Summit (StICT)*, 2019, pp. 1-6: IEEE.

[41] Z. Jiang *et al.*, "Data-driven thermal model inference with armax, in smart environments, based on normalized mutual information," in *2018 Annual American Control Conference (ACC)*, 2018, pp. 4634-4639: IEEE.

[42] I. ANSYS, "ANSYS fluent 12.0 User's Guide," *New Hampshire: ANSYS INC*, 2009.

[43] Y. Fulpagare and A. Bhargav, "Advances in data center thermal management," *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 981-996, 2015.

[44] W. A. Abdelmaksoud, H. E. Khalifa, T. Q. Dang, R. R. Schmidt, and M. Iyengar, "Improved CFD modeling of a small data center test cell," in *2010 12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, 2010, pp. 1-9: IEEE.

[45] H. Moazamigoodarzi, P. J. Tsai, S. Pal, S. Ghosh, and I. K. Puri, "Influence of cooling architecture on data center power consumption," *Energy*, vol. 183, pp. 525-535, 2019.

[46] R. Temam, "Navier-Stokes equations: Theory and numerical analysis(Book)," *Amsterdam, North-Holland Publishing Co.(Studies in Mathematics and Its Applications*, vol. 2, p. 510, 1977.

[47] D. C. Wilcox, *Turbulence modeling for CFD*. DCW industries La Canada, CA, 1998.

[48] A. C. Megri and F. Haghighat, "Zonal modeling for simulating indoor environment of buildings: Review, recent developments, and applications," *Hvac&R Research*, vol. 13, no. 6, pp. 887-905, 2007.

[49] F. M. White, "Viscous flow in ducts," *Fluid mechanics*, vol. 3, 1999.

[50] E. Wurtz, L. Mora, and C. Inard, "An equation-based simulation environment to investigate fast building simulation," *Building and Environment*, vol. 41, no. 11, pp. 1571-1583, 2006.

[51] S. Asgari *et al.*, "Hybrid surrogate model for online temperature and pressure predictions in data centers," *Future Generation Computer Systems*, vol. 114, pp. 531-547.

[52] K. Mehrotra, C. K. Mohan, and S. Ranka, *Elements of artificial neural networks*. MIT press, 1997.

[53] A. Afram and F. Janabi-Sharifi, "Gray-box modeling and validation of residential HVAC system for control system design," *Applied Energy,* vol. 137, pp. 134-150, 2015.

[54] A. Di Piazza, M. C. Di Piazza, and G. Vitale, "Solar and wind forecasting by NARX neural networks," *Renewable Energy and Environmental Sustainability,* vol. 1, p. 39, 2016.

[55] E. Diaconescu, "The use of NARX neural networks to predict chaotic time series," *Wseas Transactions on computer research,* vol. 3, no. 3, pp. 182-191, 2008.

[56] H. Xie, H. Tang, and Y.-H. Liao, "Time series prediction based on NARX neural networks: An advanced approach," in *2009 International conference on machine learning and cybernetics*, 2009, vol. 3, pp. 1275-1279: IEEE.

[57] S. M. Guzman, J. O. Paz, and M. L. M. Tagert, "The use of NARX neural networks to forecast daily groundwater levels," *Water resources management,* vol. 31, no. 5, pp. 1591-1603, 2017.

[58] J. M. P. Menezes Jr and G. A. Barreto, "Long-term time series prediction with the NARX network: An empirical evaluation," *Neurocomputing,* vol. 71, no. 16-18, pp. 3335-3343, 2008.

[59] G. Varsamopoulos, M. Jonas, J. Ferguson, J. Banerjee, S. K. Gupta, and I. Lab, "Using transient thermal models to predict cyberphysical phenomena in data centers," *Sustainable Computing: Informatics and Systems,* vol. 3, no. 3, pp. 132-147, 2013.

[60] R. Wang *et al.*, "Toward Automated Calibration of Data Center Digital Twins: A Neural Surrogate Approach," *arXiv preprint arXiv:2001.10681,* 2020.

Chapter 6

A Data-Driven Approach to Simultaneous Fault Detection and Diagnosis in Data Centers

This chapter is reproduced from *"A Data-Driven Approach to simultaneous Fault Detection and Diagnosis in Data Centers", **Sahar Asgari**, Rohit Gupta, Ishwar K, Puri, and Rong Zheng, Published in Applied Soft Computing, 2021*.

Her main contributions to this work consist of introducing the idea of using gray-box model for single and simultaneous fault detection and diagnosis, writing the manuscript, conducting the experiments, implementing the framework, constructing the algorithms, and generating the numerical results.

6.1 Abstract

The failure of cooling systems in data centers (DCs) leads to higher indoor temperatures, causing crucial electronic devices to fail, and produces a significant economic loss. To circumvent this issue, fault detection and diagnosis (FDD) algorithms and associated control strategies can be applied to detect, diagnose, and isolate faults. Existing methods that apply FDD to DC cooling systems are designed to successfully overcome individually occurring faults but have difficulty in handling simultaneous faults. These methods either require expensive measurements or those made over a wide range of conditions to develop training models, which can be time-consuming and costly. We develop a rapid and accurate, single and multiple FDD strategy for a DC with a row-based cooling system using data-driven fault classifiers informed by a gray-box temperature prediction model. The gray-box model provides thermal maps of the DC airspace for single as well as a few simultaneous failure conditions, which are used as inputs for two different data-driven classifiers, CNN and RNN, to rapidly predict multiple simultaneous failures. The model is validated with testing data from an experimental DC. Also, the effect of adding Gaussian white noise to training data is discussed and observed that even with low noisy environment, the FDD strategy can diagnose multiple faults with accuracy as high as 100% while requiring relatively few simultaneous fault training data samples. Finally, the different classifiers are compared in terms of accuracy, confusion matrix, precision, recall and F1-score.

Key words: Data center, Fault diagnosis, Classification, Time-series analysis, Gray-box model.

Nomenclature

ANN	Artificial neural network	IT	Information technology
BPTT	Backpropagation through time	LSTM	Long short-term memory
CFD	Computational fluid dynamics	lpm	Liter per minute
CNN	Convolutional neural network	NARX	Nonlinear AutoRegressive Exogenous
DC	Data center	PCA	Principal component analysis
DDM	Data-driven model	OCSVM	One-class support vector machine
FDD	Fault detection and diagnosis	RBF	Radial basis function
HPC	High-performance computing	RNN	Recurrent neural network
IRC	In-row cooling	SVM	Support vector machine

6.2 Introduction

With the advent of big data, three-dimensional electronic chip stacking, increased interest in tensor processing and deep learning algorithms, worldwide computing loads on the data centers (DCs), and high-performance computing (HPC) clusters are increasing significantly. By 2025, DC energy use is anticipated to account for 20% of worldwide consumption [1-3]. Cooling systems contribute to 24-60% of the total energy consumed by computing infrastructures [4, 5]. The efficient operation of cooling systems is critical for providing a secure, reliable, and stable DC environment while ensuring energy efficiency and compliance with safety guidelines for computing infrastructures.

The primary causes of failures in air-cooled computing equipment include (1) large temperature and humidity fluctuations that decrease equipment lifetime, and (2) the presence of hot (temperature) spots leading to automatic shutdown of the information

technology (IT) equipment [6-9]. Failures of IT equipment can be mitigated by endowing cooling systems with fault detection and isolation (FDI) capacity to ensure self-sustaining DC operation and resilience. A large survey of air conditioning units revealed that more than 90% had experienced one or more faults [10]. Cooling units that operate under faulty conditions in a DC exacerbate the energy consumption and cost, and damage the IT equipment, diminishing computing efficiency [11]. Therefore, there is a pressing need to develop effective fault detection and diagnosis (FDD) solutions for cooling systems in DCs.

There are two classes of FDD algorithms, (1) independent and (2) simultaneous FDD [12, 13]. The first considers only a single fault type at a time, while the latter can detect two or more mutually exclusive faults occurring simultaneously. Studies on FDD can be further separated into two categories, i.e., model-based and data-driven [14-17]. In a model-based FDD [18-20], a semi-empirical mechanistic representation of the system or a physics-based model is established to characterize the dynamic system behaviors under normal operations. Thereafter, characteristic anomalies in the system are detected and diagnosed from the deviations of real-time process outputs from the predicted normal conditions. Popular model-based FDD techniques include symbolic time-series analysis, interacting multi-model, smooth variable state space, cross wavelet transform, and multi-modal decomposition [14, 15, 21]. Despite the existence of several model-based FDD solutions, it is often challenging to establish intricate and accurate physics-based representations for anomalous behaviors of dynamic systems. These methodologies are

also often prohibitively computation-intensive, limiting their implementation in control systems to real-time diagnostics [22, 23].

Several types of time-series signals, such as (a) acoustic, (b) vibration, and (c) electrical signals have been used for FDD of air-conditioning systems. Data-driven approaches such as principal component analysis (PCA), artificial neural networks (ANN), support vector machines (SVM), and combinations of these techniques have been applied to identify cooling system faults in DCs [24-27]. However, these methods suffer from multiple drawbacks, e.g., (1) poor signal to noise ratio, (2) prohibitively expensive data acquisition equipment required for high-frequency mechanical/electrical measurements, (3) lack of single-point contact measurements for each component, and (4) high computational requirement for transforming large time-domain signals to the frequency domain in real-time [13, 28-30]. Issues (2) and (3) can be partly overcome by obtaining real-time spatial thermal measurements since temperature probes can be readily installed in DCs, are cost-effective, and require low computational post-processing.

In contrast to the vast FDD literature about single fault detection in the air-conditioning systems, reports on algorithms that accurately detect two or more simultaneously occurring faults are sparse. Faults can occur simultaneously in many real applications, where cooling units in DCs are no exception [31, 32]. The main challenge for simultaneous FDD in DCs is that the number of combinations of multiple faults is large, resulting in numerous possible fault patterns. Acquiring large-scale datasets for simultaneous faults required for data-driven models is both difficult and expensive. Besides,

data collection during faulty conditions can adversely affect the health and productivity of the system, particularly when simultaneous faults occur.

Here, we develop a novel hybrid FDD framework to detect and diagnose multiple simultaneous faults in DC cooling units. To mitigate the shortage of faulty data in real systems, we extend a previously proposed 3D gray-box transient model by explicitly modeling cooling unit control for multi-rack DCs equipped with in-row cooling (IRC) units. The resulting model allows for the generation of simulated data under normal, fan, or chiller pump failures in the presence of diverse server workloads. Model parameters are calibrated using experimental data obtained from real systems during their normal operation.

In this study, to detect arbitrary faults, two data-driven models are considered, one-class SVM (OCSVM) and a Nonlinear AutoRegressive Exogenous (NARX) neural network model. To diagnose single or simultaneous failures, we train a 2D convolutional neural network (2D-CNN) and a recurrent neural network (RNN) model. The first uses 2D-CNN for feature exaction and the second has long short-term memory (LSTM) to capture long-term temporal dependencies. Additionally, the robustness of the proposed models to noisy environments is investigated by adding Gaussian white noise that is the most common type of noise [33].

We instrument an enclosed modular DC with 25 temperature sensors. The modular DC contains 64 active servers spread across five racks and two IRC units on each side. Experimental data is collected every 30 s under both normal and multiple fault conditions

to validate the proposed models. The OCSVM and NARX models can respectively detect faults with 100% accuracy after 300s and 270s after the onset of faults. For single component failure, both 2D-CNN and RNN models can achieve 100% accuracy when the models are trained and tested with 540s and 420s of data, respectively. In the case of two failures, the RNN and CNN models provide high accuracy up to 100% with 540s and 600s of training data, respectively. To the best of our knowledge, this is the first study to leverage a gray-box model to generate data to train models for individual and simultaneous FDD in a DC cooling system. In summary, the major contributions of this paper lie in:

- Extension of a previous 3D gray-box model [34, 35] by utilizing cooling unit control to provide thermal maps of the DC airspace for single as well as simultaneous failure conditions.
- Introduction of a hybrid FDD framework using a 3D gray-box transient model to detect and diagnose faults.
- Development of a rapid and accurate, single and multiple FDD strategy by using both single failure and a few simultaneous failures in the training data.
- Investigation of the robustness of the models in noisy environments.

The remaining sections are organized as follows. Section 2 introduces in-row cooling unit DCs and the hybrid FDD framework and its implementation. Sections 3 and 4 presents the analysis and discussion of results. Finally, Section 5 summarizes our conclusions.

6.3 Methodology

6.3.1 A modular data center architecture with in-row cooling units

A modular DC contains several server racks, an uninterruptable power supply, power distribution cabinets, power distribution units, air conditioners (in-row coolers in this case), and additional equipment that operates independently and provides network, cabling, and monitoring functions. This class of DCs is gaining in popularity due to improved efficiency, flexibility and expandability, and ease of maintenance.

Figure 6-1 a and b depict an enclosed modular DC geometry equipped with five racks ($2 \times 1 \times 0.6$ m^3) and two IRC units on each side with a fixed speed pump and control valve to regulate the water flow rate. The red dots indicate the positions of the temperature sensing probes across the half-width of the cold and hot chambers. In the front chamber, cold air delivered by the IRC units is drawn into the servers. In the back chamber, warm air exits the servers and is returned to the IRC units. There are 64 servers in total. Rack number 2 contains 12 servers and the rest have 13 servers each. The racks are partially populated with these servers and empty spaces are covered with blanking panels to minimize hot and cold air mixing. However, leakage flows are present inside the enclosure through the air-blocking brushes, either from the back to the front chamber or vice versa across the aisles, depending on the pressure difference across them [36, 37]. Figure 6-1 c shows the energy interactions across different components in the DC. Each of the IRC units includes 3 fans and a fin-tube heat exchanger to transfer the heat from hot air to chilled water stream. The IRC units are equipped with temperature sensors to measure the return air temperature. A setpoint is set for those sensors so that the control system in the units

can adjust the fan speed and the chilled water valve to better match the IT load and server inlet air temperature requirement.

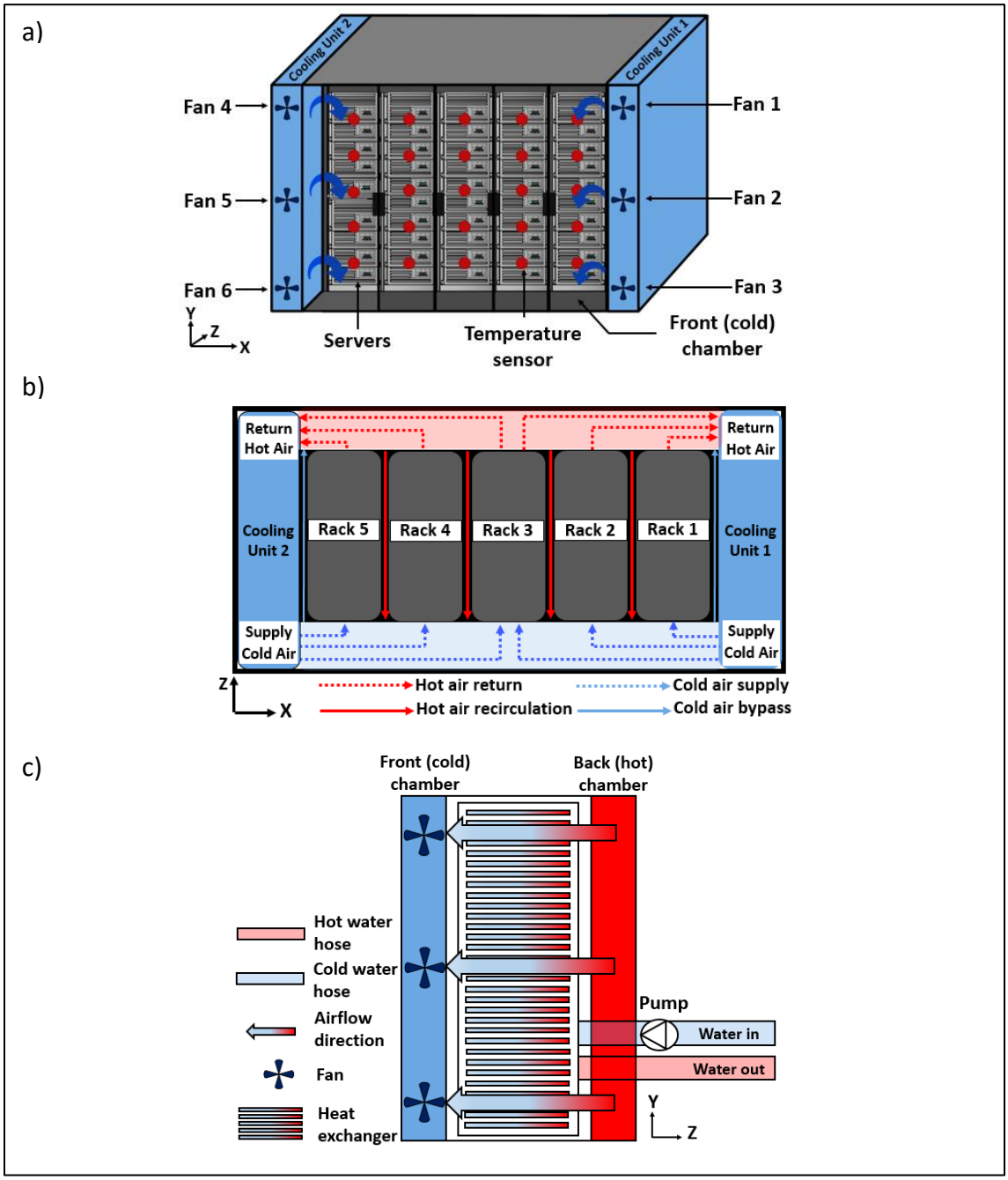


Figure 6-1. (a) Three-dimensional schematic representation of the DC considered for the case study. The red dots indicate positions of the temperature probes across half-width of the cold chamber, (b) top cross-sectional view showing salient airflows inside the enclosure, and (c) IRC schematic. The enclosure is 3.2 m long, 1.4 m wide, and 2.05 m high.

6.3.2 Gray-box temperature prediction

For predicting DC thermal conditions, zonal models represent an intermediate approach between the lumped system and more precise but time-consuming computational fluid dynamics (CFD) models. For a gray-box prediction, the DC environment is partitioned into coarse-grained control volumes, assuming that the physical properties inside each control volume are spatially uniform. A set of nonlinear coupled equations consisting of the mass, momentum, and energy conservation relations is applied for each uniform zonal volume [38]. Figure 6-2 shows the three-dimensional zones inside the enclosure for a row-based cooling architecture DC. A total of 95 zones are created by considering the (1) fronts and (2) backs of servers and cooling units, and the (3) servers themselves.

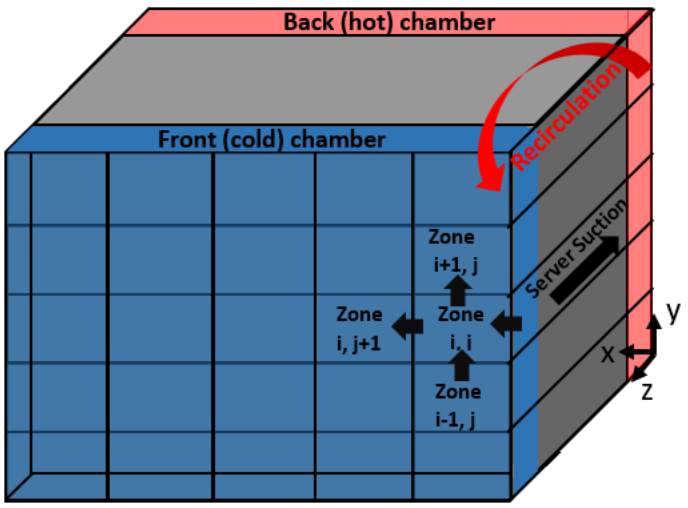


Figure 6-2. Zones considered for temperature prediction in the cold (front) chamber and back (hot) chamber inside the DC enclosure equipped with IRC units.

An artificial neural network (ANN) model has been trained to characterize the relation between the zonal pressures and cooling configuration using data from CFD

simulations [34]. With the predicted pressure for each zone obtained from the trained ANN, the inlet and exit airflows of each zone at time t[1] can be reconstructed by applying mass and momentum balance across these zones and servers for which details are provided in the appendix, which is obtained from [34].

To characterize the effects of the failures of cooling system components in our analysis, we also incorporate transient cooling system control in the gray-box model. The IRC units situated within the DC contain an air-water heat exchanger and fans to extract heat from the DC. The waterside of the heat exchanger is fed with the building chilled water supply using a circulation pump. The gray-box model representing the spatial airside temperature is coupled with the waterside through a transient energy balance for the IRC heat exchanger. Details on the transient energy balance for the airside of the IRC heat exchangers can be found in the appendix, which is obtained from [39]. The data flow within the gray-box model is depicted in Figure 6-3.

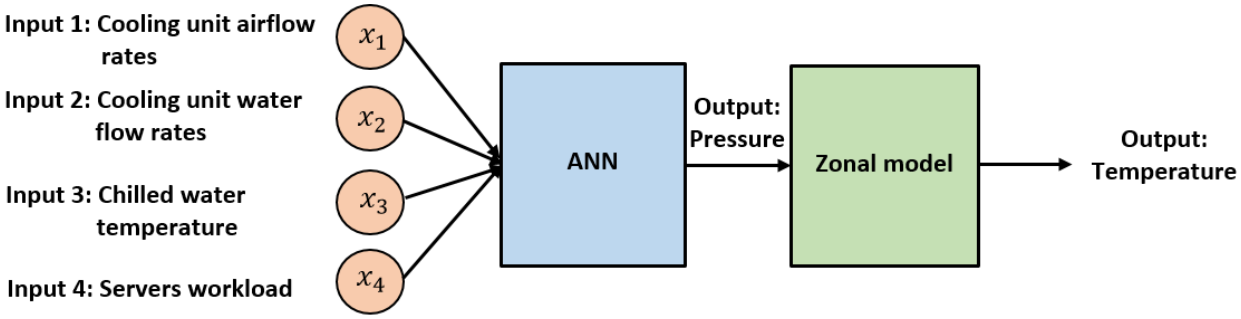


Figure 6-3. The data flow within the gray-box model for temperature predictions.

[1] Unless specified otherwise, time index *t* is omitted from the equations in the appendix.

6.3.3 Fault types

Common failures in electronic products can be traced back to thermal-related issues. The lifespans of electronic components in a DC is significantly shortened when the environment contains high-temperature areas or there are high-temperature fluctuations. Pumps and fans are widely used to maintain cold airflow that cools electronic equipment. Since pump and fan reliability is critical for the proper thermal management and reliability of IT equipment [40-43], we consider pump and fan(s) failures and assess their impact on the system operations. Seven types of faults and their combination are investigated, leading to a total of 21 fault classes and one normal class.

Figure 6-4 shows the temperature distributions under normal conditions, failure of fan 2 and failure of the chiller pump 540 s after the onset of the failure, as predicted by the gray-box model and measured from our experiments in a modular DC. The predictions and measurements are in close agreement. Different failures have different spatial implications, which can be distinguished from the temperature distributions during normal operation. These spatial patterns serve as the basis of the FDD application.

6.3.4 Hybrid simultaneous FDD

6.3.4.1 Solution overview

The development of a simultaneous FDD model consists of four main steps as shown in Figure 6-5. First, time-series temperature data under normal, single, and simultaneous fault conditions are simulated using the gray-box temperature prediction model and experimental set up [35]. Second, fault classification models are trained using faulty data.

We consider two deep neural networks, namely, a 2D-CNN model and an RNN model as described in section 6.3.4.4. Thirdly, an off-line fault detection model is trained with normal data only. During the operational phase, windowed measurements collected from thermal sensors in the modular DC are taken as inputs for the fault detection model. If a fault is detected, the fault classification model is further applied to identify the types of that fault. Alternatively, one can train an online fault detection model with the assumption that the system, at least initially, operates under normal conditions. The online model is then utilized to detect faults in the next window. If no fault is detected, the model is updated with new data. Otherwise, the fault classification model is applied to determine the type of fault. The details of each step follow.

6.3.4.2 Training data generation

The temperature data at 25 sensor locations are simulated at 30-second intervals using the three-dimensional gray-box temperature prediction model described in Section 6.3.2. Experiments show that smaller sampling intervals have only marginal impacts on the accuracy of the resulting model. For robustness, measurements with random initial seeds and server utilizations are collected under normal, single, and simultaneous fault conditions. In total, 200 instances of transient data are generated, each lasting 10 minutes.

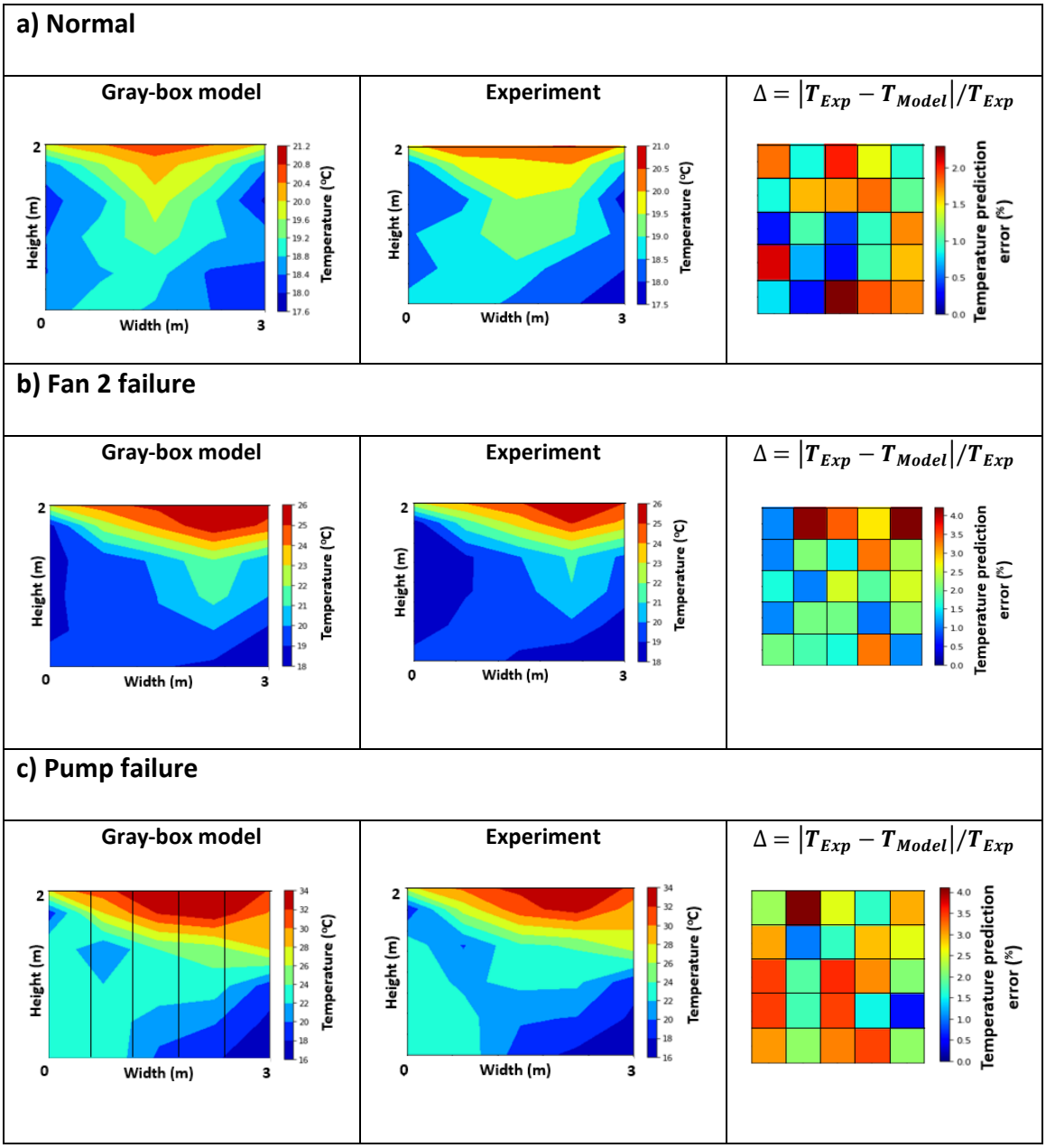


Figure 6-4. Temperature distribution predicted by the gray-box model and measured from our experimental modular DC at t = 540s and prediction error ($\Delta = |T_{Exp} - T_{Model}|/T_{Exp}$) under (a) normal, (b) fan 2 failure and (3) pump failure conditions.

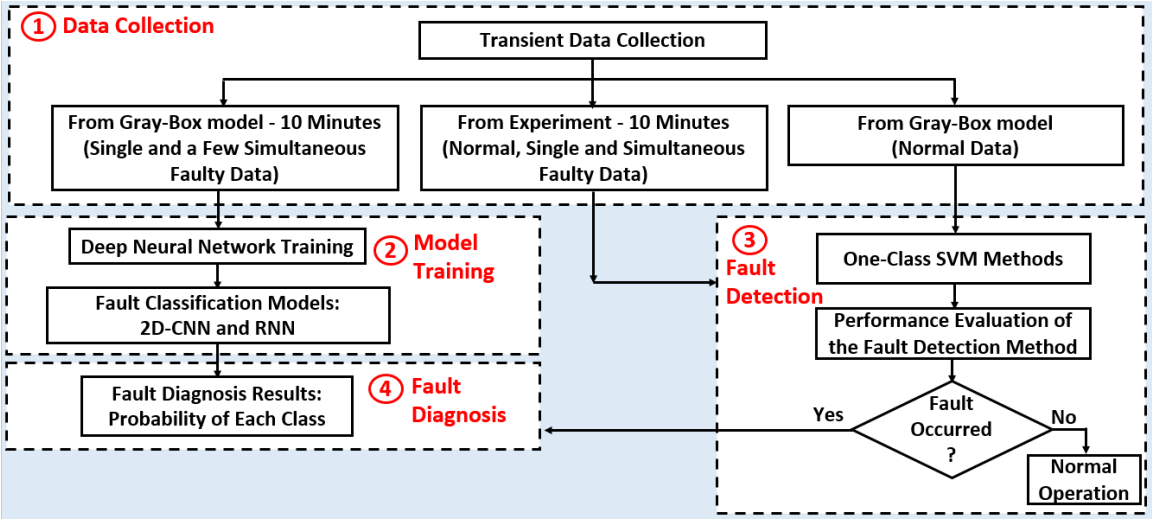


Figure 6-5. Schematic of the FDD strategy.

6.3.4.3 Fault detection

Anomalies are often the result of exceptional system conditions and do not describe the common functioning of the underlying system. Fast anomaly detection is one of the key requirements for economical and optimal process operation management. Many neural network models have been developed to detect faults in a system and shown to be highly successful. We consider OCSVM and NARX techniques. OCSVM is a special variant of the general SVM and only uses the normal operation data for training. It constructs the tightest decision boundary that encloses all data with minimal slacks. If a new sample locates within the boundary, it is classified as a normal operation point; otherwise, it is labeled as an abnormality. Since no faulty data is needed for training, the OCSVM can be trained easily and has been applied widely for fault detection [44].

Formally, let $x_i, i = 1,2, \dots, n$ be normal data points. OCSVM aims to solve the following optimization problem,

$$\min_{r,c,\zeta} r^2 + \frac{1}{\upsilon n}\sum_{i=1}^{n}\zeta_i, \text{ s.t., } \|\Phi(x_i) - c\|^2 \leq r^2 + \zeta_i, \forall i = 1,2,\dots,n, \quad (6\text{-}1)$$

where $\Phi(\cdot)$ is a kernel function and υ is a hyper-parameter that sets an upper bound on the fraction of outliers (i.e., training examples regarded out-of-class) [45, 46]. Here, we adopt a radial basis function (RBF) kernel

NARX is a popular machine learning algorithm that characterizes complex nonlinear mappings between the input and output time-series data. A NARX network with embedded memory (tapped delay line) can be utilized to detect faults in a system [47, 48]. First, a NARX network is trained and is used to predict target features given past inputs. If the distance between the predicted and actual values exceeds a threshold over several consecutive data samples, an anomaly is detected. In this work, we consider a NARX model illustrated in Figure 6-6, which has been previously used in transient temperature prediction for DCs [49].

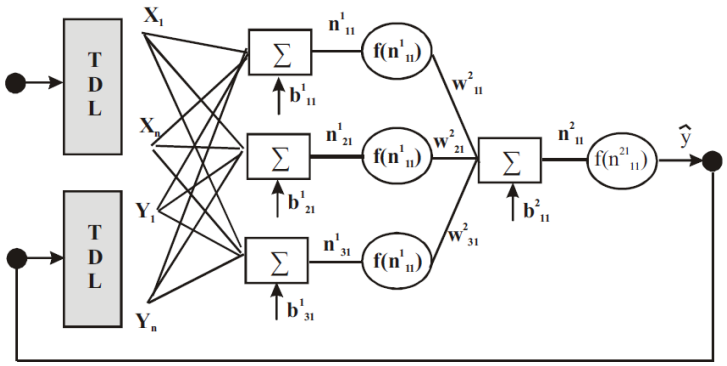


Figure 6-6. NARX neural network with tapped delay line (TDL) at the input (reproduced with permission from [49]).

6.3.4.4 Fault classification

A data-driven model (DDM) is a technique that enables learning from a set of observations. The main concept of a DDM is to find relationships between the system state variables (input, internal, and output) without or with partial knowledge of the physical behavior of the system. Due to the nonlinearity and complexity of the temperature distribution in a DC, DDMs are suitable for representing the multifaceted relationships among the system-state variables. Our goal is to find data-driven functional models that classify the value of the response variables, say temperature, with respect to the pump and fans failures. Due to the spatial and temporal locality of the data, CNN and RNN models are our choices of DDMs for fault classification.

6.3.4.4.1 Convolutional neural network (CNN)

CNN is a feedforward neural network with a set of non-linear transformation functions which has enjoyed much success and demonstrated high performance for solving many classification problems [50, 51]. It transforms the input into a form which is easier to process, while still retaining the essential features. The crucial features are extracted by applying convolutional filters on the initial inputs. Next, a pooling layer is applied to reduce the spatial size of the convoluted features. Finally, the output of the pooling layer must be flattened to be further processed by fully connected layers and an output layer.

The structure of the CNN employed in this study is shown in Figure 6-7a. Input data is of the form (number of samples, height, width, depth). There are 175 training samples, where each sample is a tensor of shape (5, 5, D). In an input sample, the first and second dimensions contain the measurements of sensors arranged in a 5x5 matrix at a time

instance. The depth of the tensors D corresponds to the length of the time window for a 30s sampling interval. Based on hyperparameter tuning, the best results are obtained with four convolutional layers (C1–C4), two max-pooling layers (max-pooling 1 and max-pooling 2), and one fully connected layer. The convolutional layers have kernel sizes of 3×3 and 2×2 with ReLu activation function which is followed by a max-pooling layer which acts as the feature extractor. Finally, one fully connected layer is applied which performs non-linear transformations of the extracted features and acts as the classifier. It also contains a Softmax activation function (or normalized exponential function), which outputs a probability value for each of the classification labels. The network is trained by minimizing the sum cross-entropy loss for all training samples.

6.3.4.4.2 Recurrent neural network (RNN)

Since the input data is inherently time series data, we also provide an RNN model [52]. RNNs possess connections that have loops, feedback, and memory over time. Incorporating memory allows this type of network to learn and generalize across sequences of inputs. A commonly used RNN is Long Short-Term Memory (LSTM), which has shown a better performance than vanilla RNNs [53]. LSTM is trained using Backpropagation Through Time (BPTT) and overcomes the vanishing gradient problem.

Here, we use 2D-CNN for feature extraction from raw temperature measurements of 25 sensors at each time instance. The generated features then serve as inputs to LSTM which outputs the probability of each class. The training set contains a total of 175 samples, where each sample is a 5x5 matrix with D channels, across t timesteps. Therefore, the input shape is (175, t, 5, 5, D). According to hyperparameter tuning, the 2D-CNN contains 2

convolutional layers, one max-pooling layer, and one fully connected layer. The LSTM is trained with 1 hidden layers of 70 neurons and a ReLu activation function.

At run time, all the classifiers take real-time measurements from thermal sensors at target locations as inputs and predict the kind of failure that will occur. Additionally, all steps of the classification problems are implemented in Python 3.8 using Keras with Tensorflow backend, and the model is trained by Adam optimizer as well as the cross-entropy loss function [54].

6.3.4.5 Experimental conditions for FDD validation

Figure 6-1 shows a schematic representation of the experimental DC containing 5 IT racks with 2 IRC units. The experimental operating conditions for racks and cooling units are provided in Table 6-1 and 6-2, respectively. The experimental dataset consists of a collection of temperature readings at 25 locations in the front chamber taken at every 30 seconds over a 10-minute duration.

6.4 Results

6.4.1 Thermal characteristics of faults

We use the gray-box model to predict the temperature distributions in the cold chamber during normal operating conditions and various failures. Several statistical values of temperature measurements over a 10-minute period after the onset of single faults or for normal operations are summarized in Table 6-3. Clearly, the statistics are distinctive so that it is possible to tell apart categories of failures or the no failure condition. For example, in Table 6-3, the mean temperature from the gray-box model and experiment under normal

operating condition is 18.45 °C and 18.41°C, respectively while for any faulty situation it is above 19.45 °C. The standard deviations of temperatures for faulty states are also significantly higher than that during normal conditions. Pump failures are even more prominent with the minimum and maximum temperatures at 24.75 °C and 32.75 °C for the gray-box model and 24.54 °C and 31.99 °C for the experimental results, when the minimum and maximum temperatures deviate from the normal operating condition by 40% and 55%, respectively. As a result of spatial fluctuations in the velocity field, the resulting temperatures vary at specific time instants.

Table 6-4 presents statistics of temperature measurements over a 10-minute period after the onset of two faults and for a normal condition. Again, each type of failure has a unique signature that can be used to infer the category of the failure that has occurred. According to Table 6-4, for multiple failures, the mean of cold chamber temperatures increases significantly beyond 21 °C. Moreover, two simultaneous fan failures result in a higher standard deviation. Comparing Table 6-3 and Table 6-4, we also observe that concurrent faults generally result in higher mean and maximum temperatures and larger standard deviations.

Table 6-1. IT Racks operating conditions.

Rack	Volume flow rate of air ($m^3 s^{-1}$)	IT load (kW)
1	0.22	3.6
2	0.20	4.0
3	0.24	4.1
4	0.20	3.9
5	0.22	3.8

6.4.2 Assessment of fault detection methodologie

We investigate the effectiveness of the OCSVM and NARX in fault detection. In OCSVM, 25 data instances under normal operation condition are used as training data to build the fault detection model. The OCSVM is trained with radial basis function (RBF) kernel function, gamma of 0.04 and $nu = 0.01$. An open-loop NARX neural network with embedded memory (tapped delay line) from our prior work is used to detect faults [35].

Table 6-2. IRC unit operating conditions under normal and faulty scenarios.

Case	Cooling unit	Air flow rate ($m^3 s^{-1}$)	Water flow rate (lpm)
Normal operation	Left	0.51	32
	Right	0.51	
One fan failure	Left	0.51	32
	Right	0.34	
Two fans failure	Left	0.51	32
	Right	0.17	
Pump failure	Left	0.51	0
	Right	0.51	
Pump and one fan failure	Left	0.51	0
	Right	0.34	

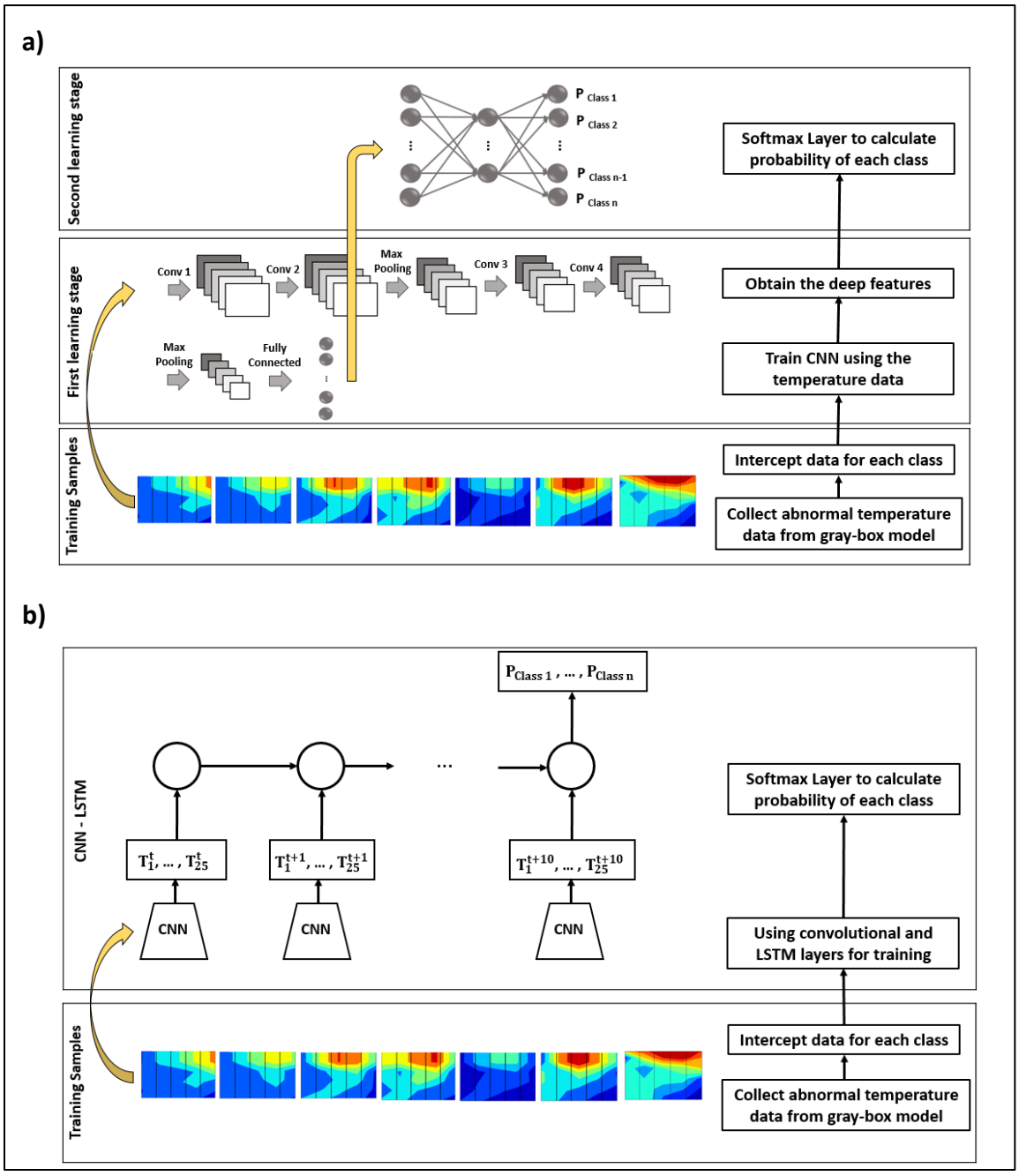


Figure 6-7. The Architecture of neural networks employed in this study. a) CNN and b) RNN.

Table 6-3. Salient thermal features of no-fault single fault states over 10 minutes.

Fault types	Mean (°C)		Standard deviation (°C)		Minimum (°C)		Maximum (°C)	
	Exp.	Model	Exp.	Model	Exp.	Model	Exp.	Model
Fan 1	19.48	19.62	2.17	2.27	17.24	17.36	26.01	26.52
Fan 2	19.51	19.82	1.69	1.88	17.88	17.92	24.71	25.11
Fan 3	20.02	20.34	2.25	2.36	17.68	17.79	26.14	26.81
Fan 4	20.54	20.65	2.74	2.91	17.47	17.50	28.22	28.92
Fan 5	19.66	19.75	1.17	1.22	17.73	17.85	22.71	22.88
Fan 6	20.96	21.06	2.78	2.86	17.71	17.86	28.76	29.02
Pump	27.86	28.02	1.13	1.34	24.54	24.75	31.99	32.75
Normal	18.41	18.45	0.98	0.97	17.61	17.69	21.10	21.19

Table 6-4. Salient thermal features of multiple fault states induced in the cooling unit fans and pump after 10 minutes.

Fault types	Mean (°C)		Standard deviation (°C)		Minimum (°C)		Maximum (°C)	
	Exp.	Model	Exp.	Model	Exp.	Model	Exp.	Model
Fans 1 and 2	20.62	21.07	2.44	2.88	17.55	17.75	27.02	27.52
Fans 1 and 6	22.01	22.68	3.58	3.77	17.65	17.75	31.19	31.94
Fans 2 and 4	21.98	22.91	3.49	3.82	17.72	17.92	31.27	31.73
Pump and Fan 1	27.85	28.50	2.23	2.41	24.21	24.75	35.07	35.88
Pump and Fan 5	28.04	28.45	1.48	1.66	25.09	25.75	31.30	31.38
Normal	18.41	18.45	0.98	0.97	17.61	17.69	21.10	21.19

210 test samples from the testbed are used to evaluate the models, among which 20 instances are from normal conditions, 140 instances contain single failure data (20 per failure) and 50 instances with simultaneous faulty data. Figure 6-8 shows the performance of fault detection based on OCSVM and NARX when the training data length changes from 120 to 360 s. Figure 6-8 represents that the faults are successfully detected by both

models. NARX detects faults with 100% accuracy and F1-Score using 270 s of data, while OCSVM requires data for 30 s more. ROC curves, shown in Figure 6-9, evaluate the trade-off between true and false-positive rates of the two algorithms. With 270 s of data, the AUC of NARX and OCSVM are 0.989 and 0.945, respectively. With 300 s of data, both models achieve an AUC of close to 1.00 which demonstrates the classifiers perfect fault detection ability.

We further evaluate the time to train and make predictions using both algorithms. The experiments are conducted on a desktop PC with a Core i7-8700 CPU at 3.20 GHz,16 GB memory, and Windows 10 with a 64-bit operating system. From Table 6-5, we see that OCSVM is more time-efficient due to lower average training and running times and CPU memory consumption. Since NARX must be trained at run-time, the total time to detect a fault is the sum of training time, inference time and the time to collect sufficient measurement data for prediction. In contrast, OCSVM is pre-trained and thus the total time to detect a fault is only the sum of inference time and the time to collect sufficient measurement data for prediction. Therefore, OCSVM is more favorable despite being slightly less data efficient.

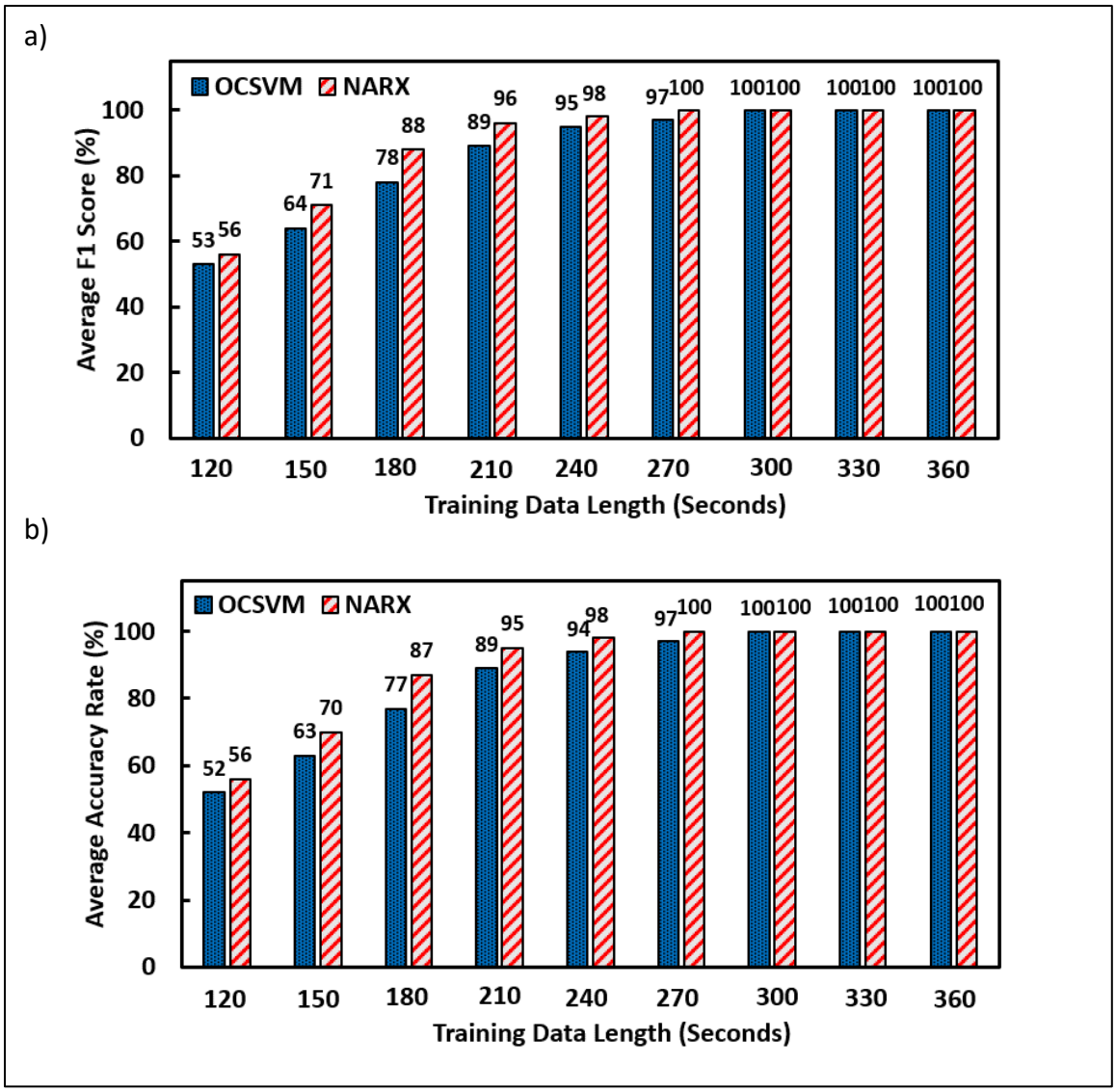


Figure 6-8. Results obtained for fault detection using OCSVM and NARX techniques when the training data length changes from 120 to 360 s.

Table 6-5. Computation time comparison for the OCSVM and NARX neural networks.

Algorithm	Average Training Time (Seconds)	Average Detection Time (Seconds)
OCSVM	~ 3	~ 0.01
NARX	~ 800	~ 0.18

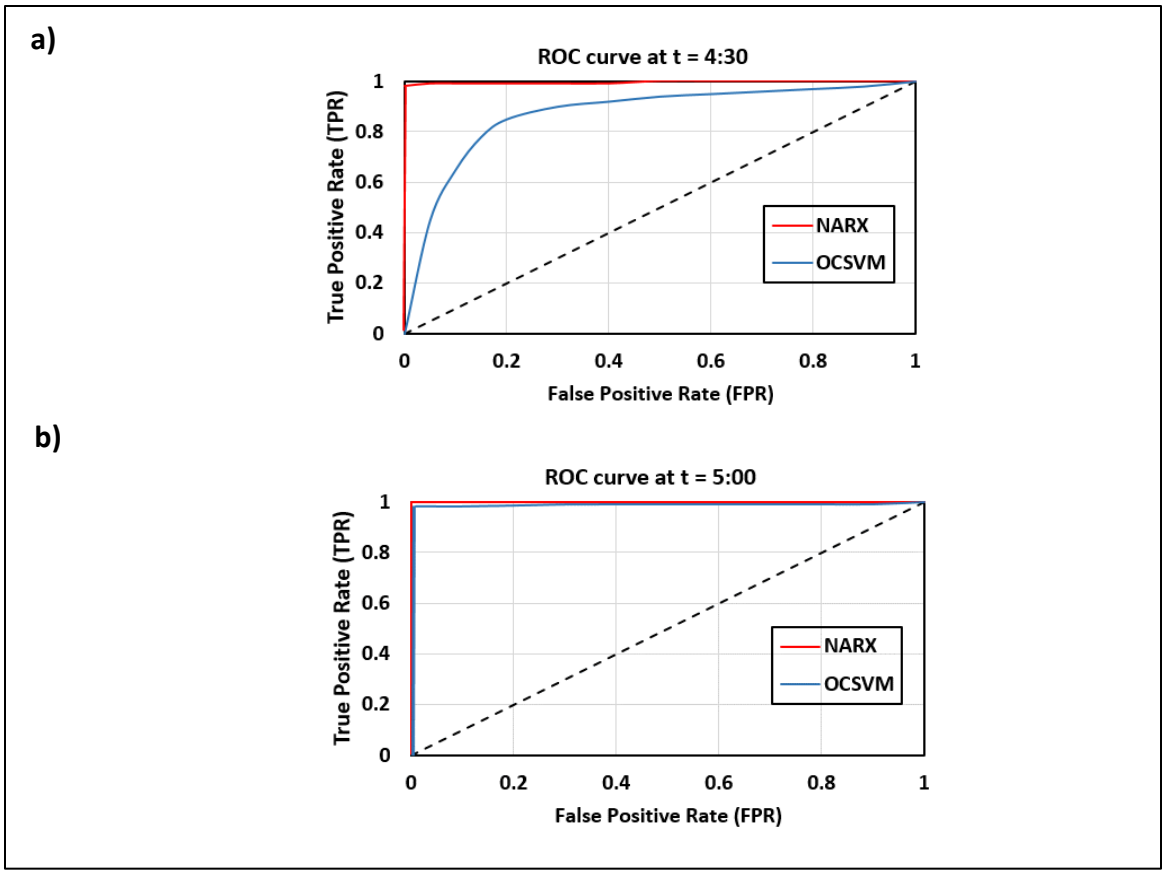


Figure 6-9. Performance of the fault detection algorithms after times (a) 270 s and (b) 300 s.

6.4.3 Comparison of fault diagnosis methodologies: Single-fault

We now investigate independent faults that occur one at a time using experimental data to evaluate the accuracy of the FDD model in a modular DC with an IRC that is depicted in Figure 6-1.

6.4.3.1 Effects of training data duration

To explore the relationship between the duration of training and testing data on independent fault diagnosis performance, a comparative study is performed. Figure 6-10 shows the average F1-Scores and accuracy with different data durations for the CNN and RNN

algorithms. As shown in the figure, the diagnosis performance degrades with decreasing data duration due to insufficient information for classification. RNN is more data-efficient than CNN since the average F1-Scores and accuracy rate of the RNN model are 100% using 420 s of training samples. In comparison, 540 s of data is needed for CNN to reach close to 100% average F1-Scores and accuracy rate.

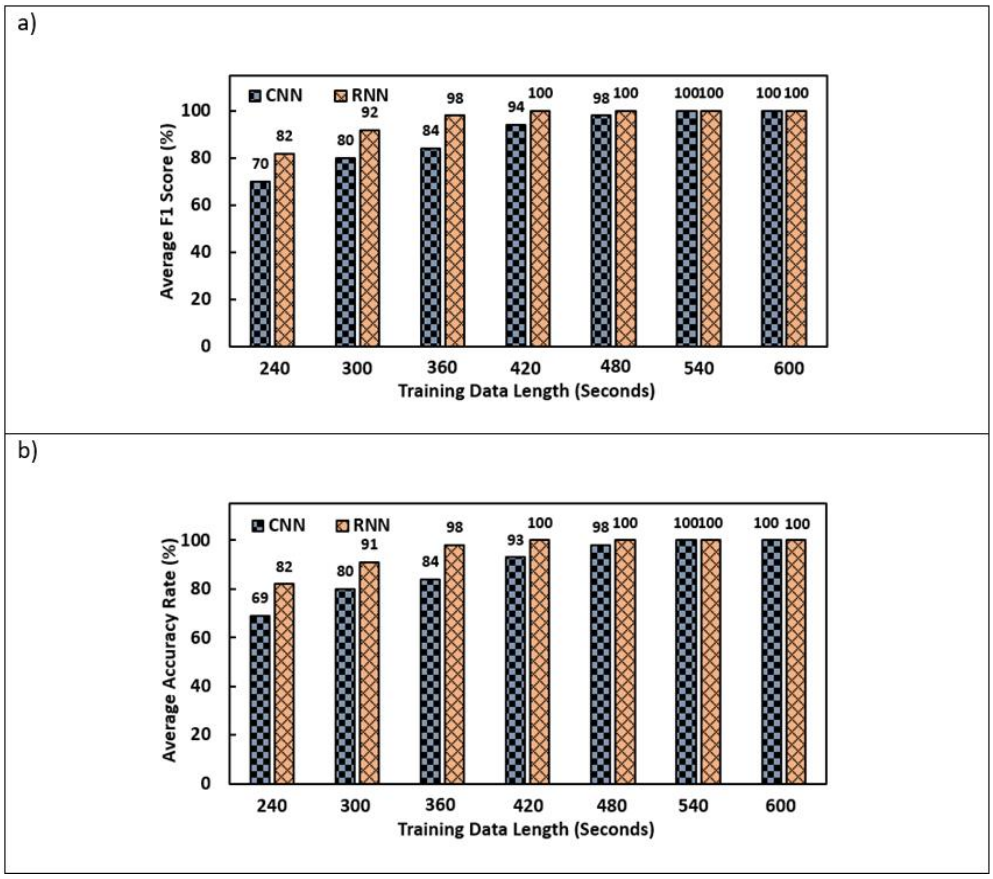


Figure 6-10. Independent fault diagnosis results with different sizes of training samples: (a) Average F1-Scores (%) and (b) Average accuracy rate (%).

6.4.3.2 Influence of initial time of training data

Table 6-6 shows the performance of the two deep learning models (CNN and RNN) as the initial timestamp of the training data is changed while the duration is held constant (i.e.,

300 s). Both models have around 100% average accuracy when the training starts from 200 s and 300 s. However, by shifting the initial time to 0 s, the performance of the CNN and RNN models decreases to 80% and 91%, respectively. Shifting the time window towards zero seconds diminishes the probability of finding a characteristic temporal change in the temperature profile which is used as signatures for detecting component failure.

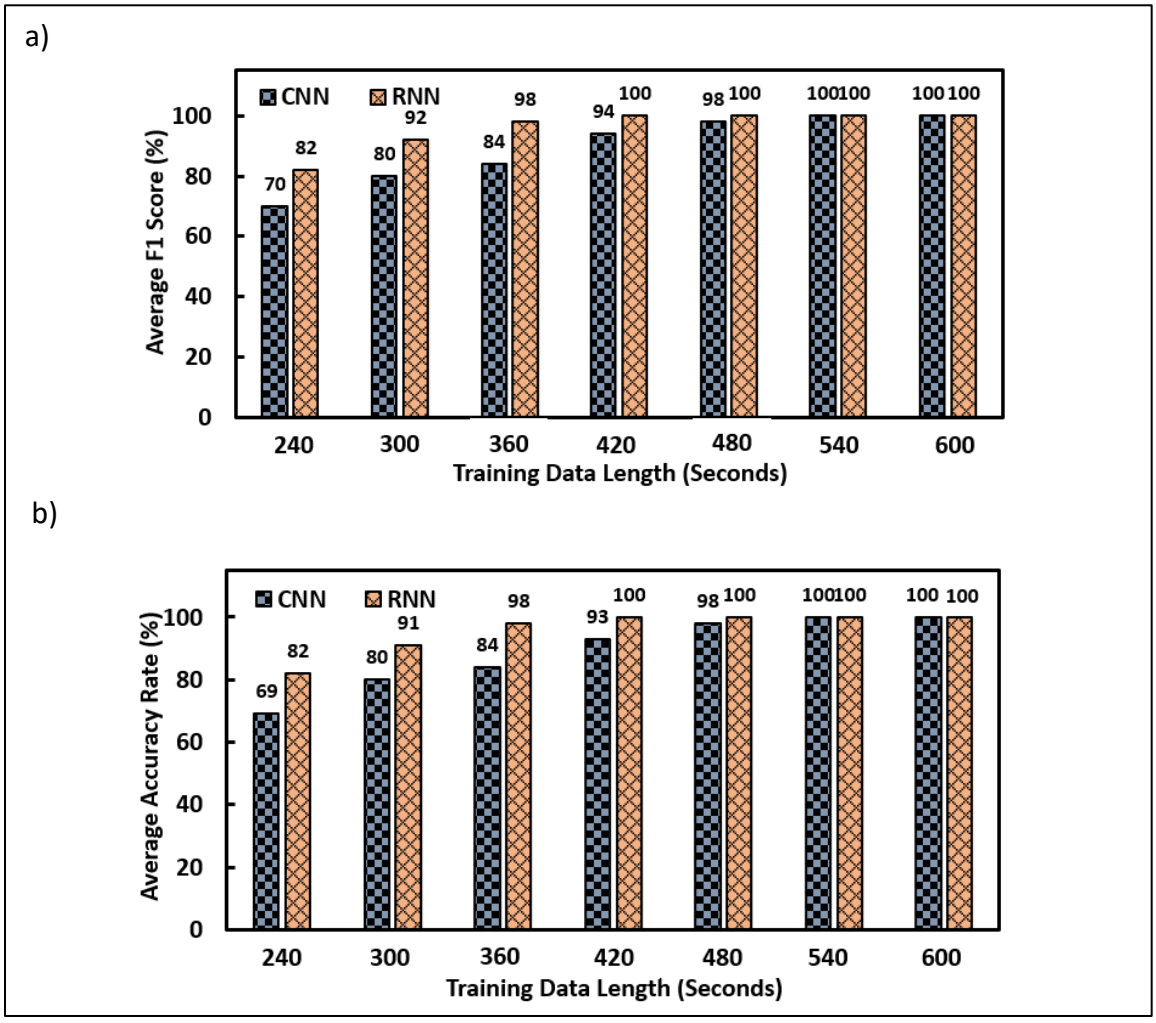


Figure 6-11. Independent fault diagnosis results with different sizes of training samples: (a) Average F1-Scores (%) and (b) Average accuracy rate (%).

Table 6-6. Comparison of the performance of the neural network model for different initial times for the training data (single failure scenario).

	Average accuracy rate (%)	
Training data duration (s)	CNN	RNN
0 - 300	80	91
100 - 400	89	94
200 - 500	98	100
300 - 600	100	100

Figure 6-12 shows the confusion matrix results for the CNN and RNN models for different training data length. The matrix compares the actual target values corresponding to experimentally triggered scenarios with those predicted by the data-driven model. The diagonal of a confusion matrix corresponds to correctly predicted fault classes. As seen in Figure 6-12, when the training data length is short, the number of misclassified testing samples is high. Most misclassifications are among fan faults and confusion between pump and fan failure is rare and non-existent with RNN for data duration no less than 300s. This implies that pump failures lead to distinctive spatial and temporal characteristics from fan failures.

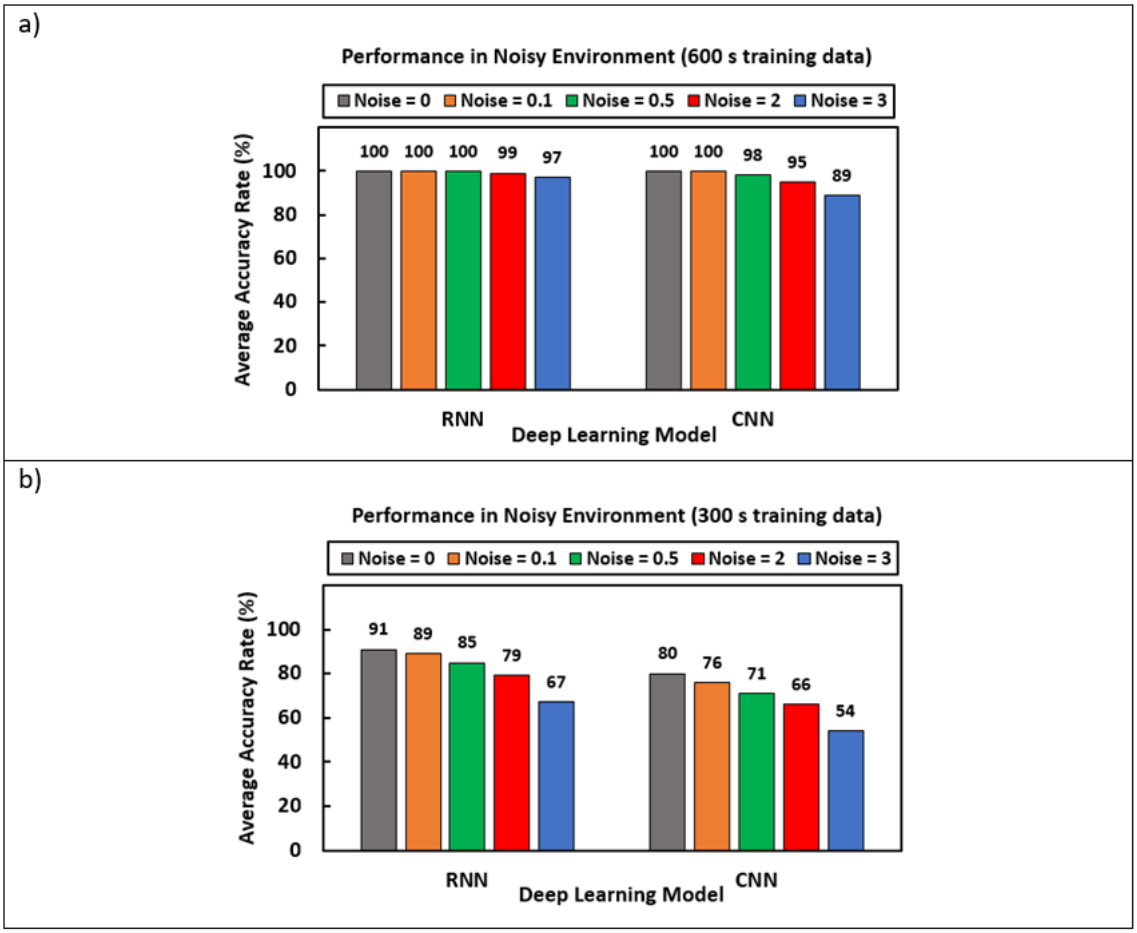


Figure 6-12. Performance of the CNN and RNN with different noise inputs when the training data lengths are a) 600s and b) 300 s.

6.4.3.3 Effect of adding noise in training dataset

In this section, Gaussian noise with standard deviations of 0.1, 0.5, 2, and 3 °C is added to the training dataset only for single fault detection. As seen in Figure 6-12, for a longer training duration (600 s), adding noise does not have a significant influence on the models performance. However, with a shorter duration (300 s), the average accuracy rate degrades by 24% and 26% for the RNN and CNN models, respectively. The degraded performance under large Gaussian noises can be attributed to the different distributions of the training

and testing data (also called domain gap). In practice, such a domain gap can be reduced by selecting a proper noise variance in training data based on the characteristics of the thermocouples used.

6.4.4 Comparison of fault diagnosis methodologies: Multiple faults

6.4.4.1 Effect of training data duration

Table 6-3 reveals seven single failure types, where the possible number of combinations of two simultaneously occurring fault in the system is as high as 128. We evaluate the ability of the proposed single-fault classifiers for detecting multiple faults. If the top-2 softmax output of the classifier is above a pre-defined threshold, two faults are identified. As baseline models, we also train one CNN and one RNN model with similar architecture as those in Section 6.3.4.4 but output 128 classes instead of 7. To distinguish the models, we call the models with 128 output classes CNN_2 and RNN_2, and the original models which use a minimal number of simultaneous fault samples CNN_1 and RNN_1. CNN_2 and RNN_2 are trained with 350 instances of data generated by the gray-box model. While CNN_1 and RNN_1 are trained with 190 instances, among which 175 of them contain single failure data (25 per failure) and 15 contain the simultaneous faulty data of Table 6-4. All four models are evaluated using experimental data from the testbed.

Figure 6-13 shows the classifier performance for various data durations. Generally, it is observed that the average accuracy rate for all models increases as data are available over longer durations. As expected, CNN_2 and RNN_2 outperform CNN_1 and RNN_1 in most cases since they are trained with enough multi-fault data. Similar to the single-fault

scenarios, RNN_1 is more data-efficient than CNN_1 and can achieve 100% accuracy with 540 s of data.

The superior performance of CNN_2 and RNN_2 for multi-fault diagnosis comes at the cost of longer data preparation times to generate the needed training data from the gray-box model and longer training time, as summarized in Table 6-7. All computation time is collected from the same desktop PC described in Section 6.3.4.4.

Table 6-7. Data preparation and computation time comparison for multiple FDD models.

Algorithm	Average required time to collect and prepare data from gray-box model (s)	Average running time to train (s)	Average running time to obtain results of a class (s)
CNN_1	~ 105000	~ 18	~ 0.3
RNN_1	~ 105000	~ 10	~ 0.2
CNN_2	~ 2025000	~ 1500	~ 0.3
RNN_2	~ 2025000	~ 100	~ 0.2

6.4.4.2 Influence of initial time of training data

Table 6-8 presents average accuracy of the proposed CNN_1, RNN_1, CNN_2, and RNN_2 models with details of the training duration when two failures occur in the system. The initial time of the training data impacts the model accuracy, with a larger effect on the CNN_1 and RNN_1 models since there are fewer simultaneous failure scenarios in the training data. When the training time window lies between 300 s to 600 s, the average accuracy of both CNN_2 and RNN_2 is 100%, whereas for CNN_1 and RNN_1 it is 88% and 93%, respectively, but which decreases significantly by shifting the initial time for training.

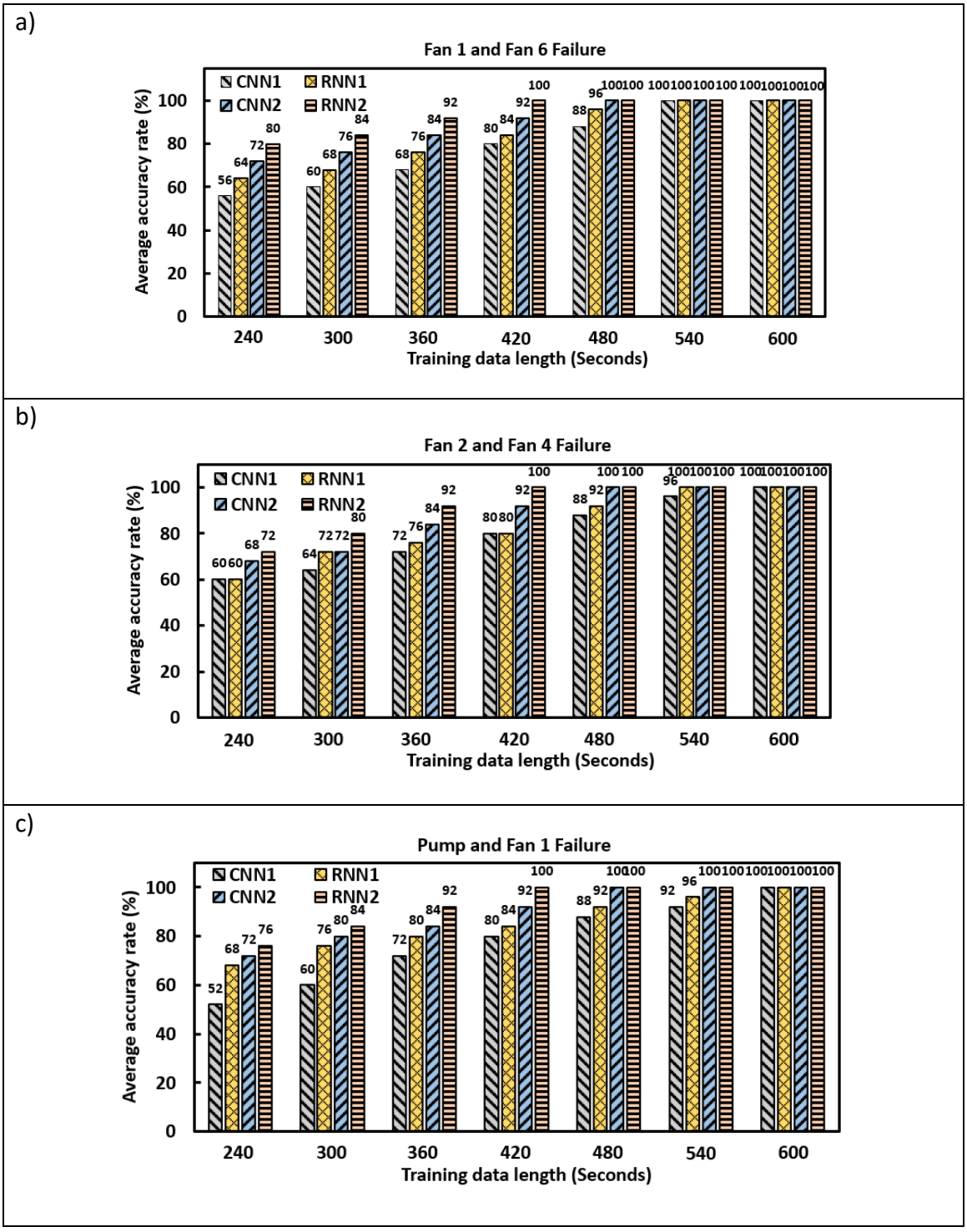


Figure 6-13. Multiple fault diagnosis performance with varying sizes of training samples.

Table 6-8. Comparison of the performance of the neural network model for different initial times of the training data (Simultaneous failure scenarios).

	Average accuracy rate (%)			
Training data duration (s)	CNN_1	RNN_1	CNN_2	RNN_2
0 - 300	61	70	75	82
100 - 400	64	76	80	91
200 - 500	72	87	92	98
300 - 600	88	93	100	100

6.4.4.3 Effect of adding noise in training data

Figure 6-14 shows the performance of the state of the art deep learning-based models when Gaussian noise with standard deviations of 0.1, 0.5, 2, and 3 °C added to the training datasets. When the training duration is 600 s, with small amounts of noise, all algorithms perform well, i.e., with greater than 92% accuracy. However, as the noise increases, the performance is degraded by up to 28%. For a training data length of 300 s, adding noise impacts the performance significantly, where adding 3 °C noise decreases the average accuracy to 40% and 50% for CNN_1 and CNN_2, respectively. Overall, the RNN model outperforms the CNN in the presence of larger amounts of noise

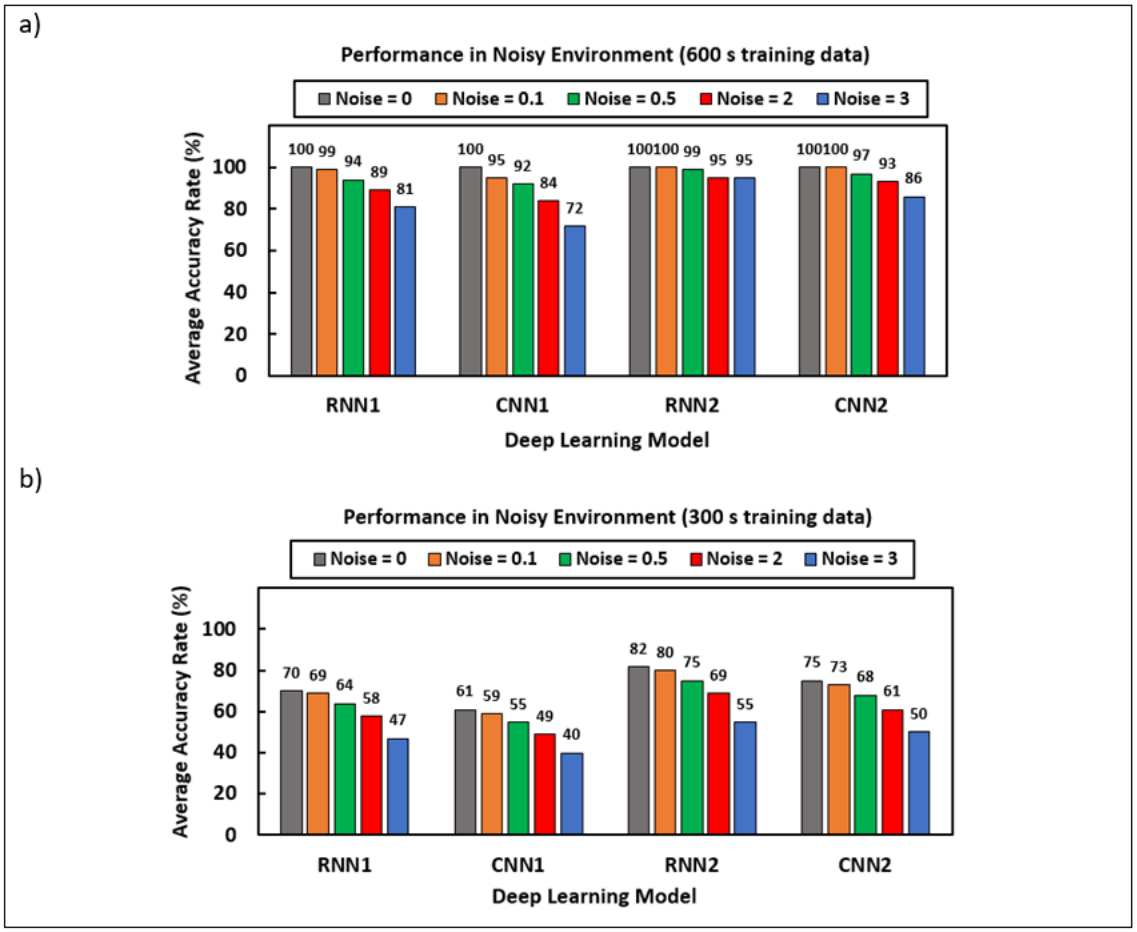


Figure 6-14. Performance of the CNN1, RNN1, CNN2 and RNN2 with different noise inputs when the training data lengths are a) 600s and b) 300 s.

6.5 Discussion

The air temperature at the front inlets of a typical server in a DC must be in the 20 °C – 30 °C range to avoid thermal redlining [55]. Every 10 °C increase over 21 °C decreases the long-term reliability of electronics by 50% [56]. The higher temperatures produce failures that have detrimental effects on hardware performance and reliability. Failure occurrence can potentially cost a company millions of dollars for each minute its network and data are unavailable. Apart from the immediate financial cost, the impact of reduced productivity,

lost opportunities, brand damage, and potential data loss that can affect a business for years to come.

We provide a specific FDD approach that can handle multiple simultaneous faults in DC cooling units. One unique contribution is the rapid diagnosis of multiple faults with high accuracy while requiring relatively few simultaneous fault training data samples. We also incorporate cooling system control into a previously proposed 3D gray-box transient model [34, 35] to gather normal and faulty data for multi-rack DCs equipped with in-row cooling (IRC) units.

The detection and diagnosis of faults requires four steps. First, regular and abnormal temperature data are recorded by using the previous 3D gray-box transient model. Next, two deep neural networks are considered, namely, a 2D-CNN and an CNN-LSTM to identify salient patterns in the normal or faulty time series data. Third, the OCSVM fault detection model is used. Finally, if a fault is detected, the fault classification model is applied further to identify the type of that fault.

We utilized training data from the 3D gray-box model and testing data from the real-world data center presented in Section 2 to validate the FDD model. To avoid overfitting, the k-fold cross validation technique is used for which model evaluation is taken as the average of k model evaluation from each of the k folds of the data. The result depends on the complexity of the fault and the characteristics of the different faults. For instance, the accuracy rate of single fault types is high using short training data lengths, which indicates that the method detects and diagnoses the fault sufficiently rapidly even

with a training set of shorter duration. However, for multiple fault conditions, a longer training data length is required to reach a desirable accuracy. The results of the methods are compared with those obtained previously in Table 6-9.

The methodology is general and can be applied to other types of DCs, such as the rack mountable cooling unit (RMCU) and overhead air delivery (OHAD) DCs [57]. Due to the generality and robustness of the methodology, the method is applicable for many different systems and FDD problems, such as residential cooling units and different industrial HVAC systems [58, 59].

Table 6-9. Comparison between the results of the method with previous results.

Reference	Fault type	Method	Simultaneous faulty data in training dataset	Simultaneous failure	Average accuracy of FDD (%)
[60]	Actuator: Cooling and heating coil, air damper, return fan	Decision tree	NA	No	97
[61]	Actuator: Cooling and heating coil, air damper, return fan	Supervised auto-encoder	NA	No	98
[58]	Actuator and Sensor faults	KPCA and RBF neural network	Yes	Yes	99
[59]	Actuator, Sensor and coil faults	Shallow neural network (SNN)	Yes	Yes	29
Proposed method	Actuator: Cooling and heating coil and return fan	2D-CNN and RNN	A few	Yes	100

6.6 Conclusion

We have developed a low-cost FDD methodology for a gray-box model and machine learning model to detect and diagnose single and multiple faults in the cooling system of a

DC. The methodology utilizes a gray-box temperature prediction model of the system to generate a training dataset for FDD. Salient findings include:

1. Fault detection using OCSVM and NARX is highly accurate. However, OCSVM is advantageous due to its low run-time computation cost.
2. For single failure diagnosis, classifier performance diminishes with decreasing data duration. The average F1-Score and accuracy of the RNN and CNN models get to 100% using 480 s and 540 s of training samples, respectively.
3. For multiple failure detection, the classifiers trained with single fault training data and a few simultaneous fault samples (i.e., CNN_1 and RNN_1) can achieve an average accuracy as high as 100% with 600 s of data. Though the CNN and RNN models trained with data from two simultaneous faults have a higher average accuracy rate of 100%, they incur the cost of significantly higher data preparation and training time.
4. RNN models are generally more data-efficient than CNN models. This may be attributed to parameter sharing and their ability to capture temporal correlation in time series data.

The results show that the FDD strategy can detect multiple failures trained with single and limited number of simultaneous fault data. Therefore, specific fault-tolerant control actions can be applied based on the faults that are diagnosed in the system to reduce operation and maintenance costs.

6.7 Acknowledgment

This research was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under a Collaborative Research and Development (CRD) project, Computationally Efficient Surrogate Models.

6.8 References

[1] Y. Li, X. Wang, P. Luo, and Q. Pan, "Thermal-aware hybrid workload management in a green datacenter towards renewable energy utilization," *Energies,* vol. 12, no. 8, p. 1494, 2019.

[2] S. MirhoseiniNejad, H. Moazamigoodarzi, G. Badawy, and D. G. Down, "Joint data center cooling and workload management: A thermal-aware approach," *Future Generation Computer Systems,* vol. 104, pp. 174-186, 2020.

[3] D. Andrews and B. Whitehead, "Data Centres in 2030: Comparative Case Studies that Illustrate the Potential of the Design for the Circular Economy as an Enabler of Sustainability," in *Sustainable Innovation 2019: 22nd International Conference Road to 2030: Sustainability, Business Models, Innovation and Design*, 2019.

[4] M. Salim and R. Tozer, "Data Centers' Energy Auditing and Benchmarking-Progress Update," *ASHRAE transactions,* vol. 116, no. 1, 2010.

[5] H. Lu, Z. Zhang, and L. Yang, "A review on airflow distribution and management in data center," *Energy and Buildings,* vol. 179, pp. 264-277, 2018.

[6] M. Kheradmandi and D. G. Down, "Data driven fault tolerant thermal management of data centers," in *2020 International Conference on Computing, Networking and Communications (ICNC)*, 2020, pp. 736-740: IEEE.

[7] N. El-Sayed, I. A. Stefanovici, G. Amvrosiadis, A. A. Hwang, and B. Schroeder, "Temperature management in data centers: Why some (might) like it hot," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, 2012, pp. 163-174.

[8] W. Torell, K. Brown, and V. Avelar, "The unexpected impact of raising data center temperatures," *Write paper 221, Revision,* 2015.

[9] H. Moazamigoodarzi, P. J. Tsai, S. Pal, S. Ghosh, and I. K. Puri, "Influence of cooling architecture on data center power consumption," *Energy,* vol. 183, pp. 525-535, 2019.

[10] J. Proctor, "Residential and small commercial central air conditioning; rated efficiency isn't automatic," in *Presentation at the Public Session. ASHRAE Winter Meeting, January*, 2004, vol. 26.

[11] G. Singh, T. C. A. Kumar, and V. Naikan, "Efficiency monitoring as a strategy for cost effective maintenance of induction motors for minimizing carbon emission and energy consumption," *Reliability Engineering & System Safety,* vol. 184, pp. 193-201, 2019.

[12] Z. Zhang, S. Li, Y. Xiao, and Y. Yang, "Intelligent simultaneous fault diagnosis for solid oxide fuel cell system based on deep learning," *Applied Energy,* vol. 233, pp. 930-942, 2019.

[13] H. Li and J. E. Braun, "A methodology for diagnosing multiple simultaneous faults in vapor-compression air conditioners," *HVAC&R Research,* vol. 13, no. 2, pp. 369-395, 2007.

[14] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part I: Quantitative model-based methods," *Computers & chemical engineering,* vol. 27, no. 3, pp. 293-311, 2003.

[15] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies," *Computers & chemical engineering,* vol. 27, no. 3, pp. 313-326, 2003.

[16] R.-E. Precup, P. Angelov, B. S. J. Costa, and M. Sayed-Mouchaweh, "An overview on fault diagnosis and nature-inspired optimal control of industrial process applications," *Computers in Industry,* vol. 74, pp. 75-94, 2015.

[17] A. Albu, R.-E. Precup, and T.-A. Teban, "Results and challenges of artificial neural networks used for decision-making and control in medical applications," *Facta Universitatis, Series: Mechanical Engineering,* vol. 17, no. 3, pp. 285-308, 2019.

[18] J. Schein, S. T. Bushby, N. S. Castro, and J. M. House, "A rule-based fault detection method for air handling units," *Energy and buildings,* vol. 38, no. 12, pp. 1485-1492, 2006.

[19] P. T Agami Reddy PhD, "Development and evaluation of a simple model-based automated fault detection and diagnosis (FDD) method suitable for process faults of large chillers/discussion," *ASHRAE Transactions,* vol. 113, p. 27, 2007.

[20] Y. Zhao, S. Wang, F. Xiao, and Z. Ma, "A simplified physical model-based fault detection and diagnosis strategy and its customized tool for centrifugal chillers," *HVAC&R Research,* vol. 19, no. 3, pp. 283-294, 2013.

[21] Y. Yu, D. Woradechjumroen, and D. Yu, "A review of fault detection and diagnosis methodologies on air-handling units," *Energy and Buildings,* vol. 82, pp. 550-562, 2014.

[22] A. Beghi, R. Brignoli, L. Cecchinato, G. Menegazzo, M. Rampazzo, and F. Simmini, "Data-driven fault detection and diagnosis for HVAC water chillers," *Control Engineering Practice,* vol. 53, pp. 79-91, 2016.

[23] Z. Du, B. Fan, X. Jin, and J. Chi, "Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis," *Building and Environment,* vol. 73, pp. 1-11, 2014.

[24] D. Li, Y. Zhou, G. Hu, and C. J. Spanos, "Fault detection and diagnosis for building cooling system with a tree-structured learning method," *Energy and Buildings,* vol. 127, pp. 540-551, 2016.

[25] S. Wang and F. Xiao, "AHU sensor fault diagnosis using principal component analysis method," *Energy and Buildings,* vol. 36, no. 2, pp. 147-160, 2004.

[26] Z. Du, X. Jin, and Y. Yang, "Wavelet neural network-based fault diagnosis in air-handling units," *Hvac&R Research,* vol. 14, no. 6, pp. 959-973, 2008.

[27] W.-Y. Lee, J. M. House, and N.-H. Kyong, "Subsystem level fault diagnosis of a building's air-handling unit using general regression neural networks," *Applied Energy,* vol. 77, no. 2, pp. 153-170, 2004.

[28] C.-S. A. Gong *et al.*, "Feature extraction of rotating apparatus using acoustic sensing technology," in *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, 2019, pp. 254-256: IEEE.

[29] N. Baydar and A. Ball, "Detection of gear failures via vibration and acoustic signals using wavelet transform," *Mechanical Systems and Signal Processing,* vol. 17, no. 4, pp. 787-804, 2003.

[30] Y. Li, J. X. Gu, D. Zhen, M. Xu, and A. Ball, "An evaluation of gearbox condition monitoring using infrared thermal images applied with convolutional neural networks," *Sensors,* vol. 19, no. 9, p. 2205, 2019.

[31] H. Li and J. E. Braun, "Decoupling features and virtual sensors for diagnosis of faults in vapor compression air conditioners," *International Journal of Refrigeration,* vol. 30, no. 3, pp. 546-564, 2007.

[32] I. Velibeyoglu, H. Y. Noh, and M. Pozzi, "A graphical approach to assess the detectability of multiple simultaneous faults in air handling units," *Energy and Buildings,* vol. 184, pp. 275-288, 2019.

[33] A. Neelakantan *et al.*, "Adding gradient noise improves learning for very deep networks," *arXiv preprint arXiv:1511.06807,* 2015.

[34] S. Asgari *et al.*, "Hybrid surrogate model for online temperature and pressure predictions in data centers," *Future Generation Computer Systems,* vol. 114, pp. 531-547, 2021.

[35] S. Asgari, S. MirhoseiniNejad, H. Moazamigoodarzi, R. Gupta, R. Zheng, and I. K. Puri, "A gray-box model for real-time transient temperature predictions in data centers," *Applied Thermal Engineering,* vol. 185, p. 116319, 2021.

[36] R. Gupta, H. Moazamigoodarzi, S. MirhoseiniNejad, D. G. Down, and I. K. Puri, "Workload management for air-cooled data centers: An energy and exergy based approach," *Energy,* vol. 209, p. 118485, 2020.

[37] H. Moazamigoodarzi, R. Gupta, S. Pal, P. J. Tsai, S. Ghosh, and I. K. Puri, "Modeling temperature distribution and power consumption in IT server enclosures with row-based cooling architectures," *Applied Energy,* vol. 261, p. 114355, 2020.

[38] A. C. Megri and F. Haghighat, "Zonal modeling for simulating indoor environment of buildings: Review, recent developments, and applications," *Hvac&R Research,* vol. 13, no. 6, pp. 887-905, 2007.

[39] H. Moazamigoodarzi, S. Pal, S. Ghosh, and I. K. Puri, "Real-time temperature predictions in it server enclosures," *International Journal of Heat and Mass Transfer,* vol. 127, pp. 890-900, 2018.

[40] X. Tian, "Cooling fan reliability: failure criteria, accelerated life testing, modeling and qualification," in *RAMS'06. Annual Reliability and Maintainability Symposium, 2006.*, 2006, pp. 380-384: IEEE.

[41] X. Jin, E. W. Ma, T. W. Chow, and M. Pecht, "An investigation into fan reliability," in *Proceedings of the IEEE 2012 Prognostics and System Health Management Conference (PHM-2012 Beijing)*, 2012, pp. 1-7: IEEE.

[42] X.-q. Wen and L.-r. You, "A residual lifetime prediction method of cooling fan based on the operating point offset distance," in *2016 Chinese Control and Decision Conference (CCDC)*, 2016, pp. 2972-2976: IEEE.

[43] R. Fezai, K. Abodayeh, M. Mansouri, H. Nounou, and M. Nounou, "Fault diagnosis of biological systems using improved machine learning technique," *International Journal of Machine Learning and Cybernetics*, pp. 1-14, 2020.

[44] S. Yin, X. Zhu, and C. Jing, "Fault detection based on a robust one class support vector machine," *Neurocomputing,* vol. 145, pp. 263-268, 2014.

[45] Y. Xiao, H. Gao, and Y. Yan, "Indirect Gaussian kernel parameter optimization for one-class SVM in fault detection," in *Third International Workshop on Pattern Recognition*, 2018, vol. 10828, p. 108280K: International Society for Optics and Photonics.

[46] J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Technical Report MSR-T R-99–87, Microsoft Research (MSR),* 1999.

[47] P. Bangalore, S. Letzgus, D. Karlsson, and M. Patriksson, "An artificial neural network-based condition monitoring method for wind turbines, with application to the monitoring of the gearbox," *Wind Energy*, vol. 20, no. 8, pp. 1421-1438, 2017.

[48] Y. Cui, P. Bangalore, and L. B. Tjernberg, "An anomaly detection approach based on machine learning and scada data for condition monitoring of wind turbines," in *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 2018, pp. 1-6: IEEE.

[49] A. Di Piazza, M. C. Di Piazza, and G. Vitale, "Solar and wind forecasting by NARX neural networks," *Renewable Energy and Environmental Sustainability*, vol. 1, p. 39, 2016.

[50] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning. nature 521 (7553), 436-444," *Google Scholar Google Scholar Cross Ref Cross Ref*, 2015.

[51] S. Li, G. Liu, X. Tang, J. Lu, and J. Hu, "An ensemble deep convolutional neural network model with improved DS evidence fusion for bearing fault diagnosis," *Sensors*, vol. 17, no. 8, p. 1729, 2017.

[52] J. T. Connor, R. D. Martin, and L. E. Atlas, "Recurrent neural networks and robust time series prediction," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 240-254, 1994.

[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[54] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*: Springer, 2016, pp. 179-187.

[55] F. Robert, "Alternating cold and hot aisles provides more reliable cooling for server farms," *White Paper, Uptime Institute*, 2000.

[56] M. K. Patterson, "The effect of data center temperature on energy efficiency," in *2008 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, 2008, pp. 1167-1174: IEEE.

[57] R. Gupta, S. Asgari, H. Moazamigoodarzi, S. Pal, and I. K. Puri, "Cooling architecture selection for air-cooled Data Centers by minimizing exergy destruction," *Energy*, vol. 201, p. 117625, 2020.

[58] A. Montazeri and S. M. Kargar, "Fault detection and diagnosis in air handling using data-driven methods," *Journal of Building Engineering*, vol. 31, p. 101388, 2020.

[59] U. Ghose and U. Bisht, "Performance Evaluation of Various ANN Architectures Using Proposed Cost Function," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2020, pp. 732-737: IEEE.

[60] R. Yan, Z. Ma, Y. Zhao, and G. Kokogiannakis, "A decision tree based data-driven diagnostic strategy for air handling units," *Energy and Buildings*, vol. 133, pp. 37-45, 2016.

[61] W.-S. Yun, W.-H. Hong, and H. Seo, "A data-driven fault detection and diagnosis scheme for air handling units in building HVAC systems considering undefined states," *Journal of Building Engineering*, vol. 35, p. 102111, 2021.

[62] E. Wurtz, L. Mora, and C. Inard, "An equation-based simulation environment to investigate fast building simulation," *Building and Environment*, vol. 41, no. 11, pp. 1571-1583, 2006.

6.9 Appendix

6.9.1 Transient gray-box thermal model

An artificial neural network (ANN) model has been trained to characterize the relation between the zonal pressures and airflows using data from CFD simulations [34]. With the predicted pressure for each zone obtained from the trained ANN, the inlet and exit airflows of each zone at time t* can be reconstructed by applying mass and momentum balance across these zones [34].

$$\dot{m}^{f}_{(i+1,j)\to(i,j)} + \dot{m}^{f}_{(i-1,j)\to(i,j)} + \dot{m}^{f}_{(i,j+1)\to(i,j)} + \dot{m}^{f}_{(i,j-1)\to(i,j)} + \dot{m}_{l} - \dot{m}_{s} = 0, \quad \text{(6-A.1)}$$

$$\dot{m}^{b}_{(i+1,j)\to(i,j)} + \dot{m}^{b}_{(i-1,j)\to(i,j)} + \dot{m}^{b}_{(i,j+1)\to(i,j)} + \dot{m}^{b}_{(i,j-1)\to(i,j)} + \dot{m}_{l} + \dot{m}_{s} = 0, \text{ and} \quad \text{(6-A.2)}$$

$$\sum F = \sum(\dot{m}v)_{out} - \sum(\dot{m}v)_{in}\ , \quad \text{(6-A.3)}$$

where $\dot{m}_{(i+1,j)\to(i,j)}$ and $\dot{m}_{(i-1,j)\to(i,j)}$ are the mass flows in the *x*-direction, $\dot{m}_{(i,j+1)\to(i,j)}$ and $\dot{m}_{(i,j-1)\to(i,j)}$ represent the mass flows in the *y*-direction, $\dot{m}_{s}$ mass flowrate of a server, $\dot{m}_{l}$ the leakage airflow, F body force, $\dot{m}$ interfacial mass flux, and v velocity across zones.

* Unless specified otherwise, time index *t* is omitted from the equations.

Subsequently, the energy balance for the zones in front and back of servers can be utilized to predict temperatures [39, 62].

$$\rho_a C_{p,a} V_i \left(\frac{T^f_{i,j}(t)-T^f_{i,j}(t-\Delta t)}{\Delta t}\right) = \Phi^f_1(t-\Delta t) + \Phi^f_2(t-\Delta t) + \Phi^f_3(t-\Delta t) + \Phi^f_4(t-\Delta t) + \Phi^f_5(t-\Delta t) + \Phi^f_6(t-\Delta t), \text{ and} \quad (6\text{-A.4})$$

$$\rho_a C_{p,a} V_i \left(\frac{T^b_{i,j}(t)-T^b_{i,j}(t-\Delta t)}{\Delta t}\right) = \Phi^b_1(t-\Delta t) + \Phi^b_2(t-\Delta t) + \Phi^b_3(t-\Delta t) + \Phi^b_4(t-\Delta t) + \Phi^b_5(t-\Delta t) + \Phi^b_6(t-\Delta t)\ , \quad (6\text{-A.5})$$

where ρ_a indicates the density of air, V_i zonal volume, $T^f_{i,j}$ inlet air temperature of a server, $T^b_{i,j}$ the air temperature at the back of a server, and Φ the energy transport term that depends on the relative pressure, mass flow rate and temperature of the respective zone and spatial direction. Table 6- A.1 contains expressions for each term included in Eqs. 6-A.4 and 6-A.5.

The energy balance for an active server, a source of heat, is,

$$X\dot{P}_{(i,j)} - \dot{m}^s_{(i,j)} c_{p,a}\left(T^f_{(i,j)} - T^b_{(i,j)}\right) = Y\frac{dT_{CPU,(i,j)}}{dt}, \quad (6\text{-A.6})$$

where $\dot{m}^s_{(i,j)}$ denotes the server mass flow rate, $\dot{P}_{(i,j)}$ the total power consumption of the corresponding server, X a coefficient that determines the power usage by CPUs, $c_{p,a}$ specific heat capacity, $T^f_{i,j}$ inlet air temperature of a server, $T^b_{i,j}$ the air temperature at the back of a server, t time, and Y an empirical coefficient for the thermal mass of a server that is available from the literature [35].

To characterize the effects of the failures of cooling system components in our analysis, we also incorporate cooling system control in the gray-box model. The IRC units situated within the DC contain an air-water heat exchanger and fans to extract heat from the DC. The waterside of the heat exchanger is fed with the building chilled water supply using a circulation pump. The gray-box model representing the spatial airside temperature is coupled with the waterside through a transient energy balance for the IRC heat exchanger. The transient energy balance for the airside of the IRC heat exchanges reveals [39],

$$\frac{X_a}{2}\left(\frac{T_{h,a}(t)-T_{h,a}(t-\Delta t)}{\Delta t}+\frac{T_{c,a}(t)-T_{c,a}(t-\Delta t)}{\Delta t}\right) \tag{6-A.7}$$
$$=-\frac{UA}{2}\left(T_{h,a}(t-\Delta t)+T_{c,a}(t-\Delta t)-T_{c,w}-T_{h,w}(t-\Delta t)\right)$$
$$+\rho_a C_{p,a}\dot{Q}_a\left(T_{h,a}(t-\Delta t)-T_{c,a}(t-\Delta t)\right),$$

and the energy balance for the waterside leads to

$$\frac{X_w}{2}\left(\frac{T_{h,w}(t)-T_{h,w}(t-\Delta t)}{\Delta t}\right) \tag{6-A.8}$$
$$=\frac{UA}{2}\left(T_{h,a}(t-\Delta t)+T_{c,a}(t-\Delta t)-T_{c,w}-T_{h,w}(t-\Delta t)\right)$$
$$+\rho_w C_{p,w}\dot{Q}_w\left(T_{c,w}-T_{h,w}(t-\Delta t)\right).$$

here $X_w=\rho_w C_{p,w}V_w$ and $X_a=\rho_a C_{p,a}V_a$ are the thermal masses of water and air inside the IRC unit, ρ_w and ρ_a densities of water and air, $C_{p,w}$ and $C_{p,a}$ specific heat capacities of water and air, $\dot{Q}_w$ and $\dot{Q}_a$ the flowrate of water and air prescribed by the DC control system (distributed across two IRC units), $T_{h,w}$ and $T_{c,w}$ the hot and chilled water temperatures, $T_{h,a}$ and $T_{c,a}$ the hot air return and cold air supply temperatures, Δt the time step, UA the

product of universal heat transfer coefficient and the contact area between the two interacting fluid media i.e., air and water. The value of UA as a function of $\dot{Q}_a$ and $\dot{Q}_w$ is obtained from our previous work [37]. In Eqs. (6-A.7) and (6-A.8), $T_{c,a}$ varies with server utilization as the return air temperature to the IRC units is changed.

Table 6-A.2. Expressions for the terms in Eqs. 6-A.4 and 6-A.5.

$\Phi^f_{1\to6}$		$\Phi^b_{1\to6}$	
Φ^f_1 (Horizontal energy transport in the front chamber)		Φ^b_1 (Horizontal energy transport in the back chamber)	
$\left[P^f_{i+1,j}-P^f_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^f_{(i+1,j)\to(i,j)}T^f_{i+1,j}$	$\left[P^b_{i+1,j}-P^b_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^b_{(i+1,j)\to(i,j)}T^b_{i+1,j}$
$\left[P^f_{i+1,j}-P^f_{i,j}\right]< 0$	$C_{p,a}\dot{m}^f_{(i,j)\to(i+1,j)}T^f_{i,j}$	$\left[P^b_{i+1,j}-P^b_{i,j}\right]< 0$	$C_{p,a}\dot{m}^b_{(i,j)\to(i+1,j)}T^b_{i,j}$
Φ^f_2 (Horizontal energy transport in the front chamber)		Φ^b_2 (Horizontal energy transport in the back chamber)	
$\left[P^f_{i-1,j}-P^f_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^f_{(i-1,j)\to(i,j)}T^f_{i-1,j}$	$\left[P^b_{i-1,j}-P^b_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^b_{(i-1,j)\to(i,j)}T^b_{i-1,j}$
$\left[P^f_{i-1,j}-P^f_{i,j}\right]< 0$	$C_{p,a}\dot{m}^f_{(i,j)\to(i-1,j)}T^f_{i,j}$	$\left[P^b_{i-1,j}-P^b_{i,j}\right]< 0$	$C_{p,a}\dot{m}^b_{(i,j)\to(i-1,j)}T^b_{i,j}$
Φ^f_3 (Vertical energy transport in the front chamber)		Φ^b_3 (Vertical energy transport in the back chamber)	
$\left[P^f_{i,j+1}-P^f_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^f_{(i,j+1)\to(i,j)}T^f_{i,j+1}$	$\left[P^b_{i,j+1}-P^b_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^b_{(i,j+1)\to(i,j)}T^b_{i,j+1}$
$\left[P^f_{i,j+1}-P^f_{i,j}\right]< 0$	$C_{p,a}\dot{m}^f_{(i,j)\to(i,j+1)}T^f_{i,j}$	$\left[P^b_{i,j+1}-P^b_{i,j}\right]< 0$	$C_{p,a}\dot{m}^b_{(i,j)\to(i,j+1)}T^b_{i,j}$
Φ^f_4 (Vertical energy transport in the front chamber)		Φ^b_4 (Vertical energy transport in the back chamber)	
$\left[P^f_{i,j-1}-P^f_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^f_{(i,j-1)\to(i,j)}T^f_{i,j-1}$	$\left[P^b_{i,j-1}-P^b_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^b_{(i,j-1)\to(i,j)}T^b_{i,j-1}$
$\left[P^f_{i,j-1}-P^f_{i,j}\right]< 0$	$C_{p,a}\dot{m}^f_{(i,j)\to(i,j-1)}T^f_{i,j}$	$\left[P^b_{i,j-1}-P^b_{i,j}\right]< 0$	$C_{p,a}\dot{m}^b_{(i,j)\to(i,j-1)}T^b_{i,j}$
Φ^f_5 (Energy exchange due to leakage flow in the front chamber)		Φ^b_5 (Energy exchange due to leakage flow in the back chamber)	
$\left[P^b_{i,j}-P^f_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^{b\to f}_{i,j}T^b_{i,j}$	$\left[P^f_{i,j}-P^b_{i,j}\right]\geq 0$	$C_{p,a}\dot{m}^{f\to b}_{i,j}T^f_{i,j}$
$\left[P^b_{i,j}-P^f_{i,j}\right]< 0$	$C_{p,a}\dot{m}^{f\to b}_{i,j}T^f_{i,j}$	$\left[P^f_{i,j}-P^b_{i,j}\right]< 0$	$C_{p,a}\dot{m}^{b\to f}_{i,j}T^b_{i,j}$
Φ^f_6 (Energy exchange due to server suction in the front chamber)		Φ^b_6 (Energy exchange due to server exhaust in the back chamber)	
$-C_{p,a}\dot{m}^s_{i,j}T^f_{i,j}$		$C_{p,a}\dot{m}^s_{i,j}T^f_{i,j}+\dot{P}^s_{i,j}$	

Chapter 7

Conclusions and future directions

7.1 Conclusions

Row-based cooling architecture with enclosure is well established in the DC industry. Therefore, we present a gray-box model to predict the thermal behavior inside a row-based cooled DC running arbitrary workloads and equipped with different distributions of servers. The effect of cooling unit operating conditions and IT loads on the temperature distribution is investigated. The model is validated by comparison with experiments, where the maximum difference between predictions and measurements is less than 7%. The model facilitates real-time control algorithms developed for IT enclosures with row-based cooling architectures. We also compare our methodology with different black-box models as well as a conventional zonal model. The results show that the gray-box model exhibits superior performance in terms of accuracy and computational time.

Additionally, we used our model to automatically detect and diagnose single and multiple failures in the cooling units of DC. The gray-box model provides thermal maps of

the DC airspace for normal and single failure conditions, used as inputs for two different data-driven classifiers, namely, CNN and RNN, to predict multiple failures rapidly. Two fault detection algorithms, OCSVM and NARX, are compared which OCSVM showed superiority over NARX in aspect of computation time and accuracy rate.

In summary, our gray-box model exhibits superior performance compared with black-box models, such as ANN and NARX models. An application of the gray-box model involves a case study to detect single and simultaneous cooling unit failures in a row-based cooled DC. The schematic steps of the study is shown in Figure 7.1.

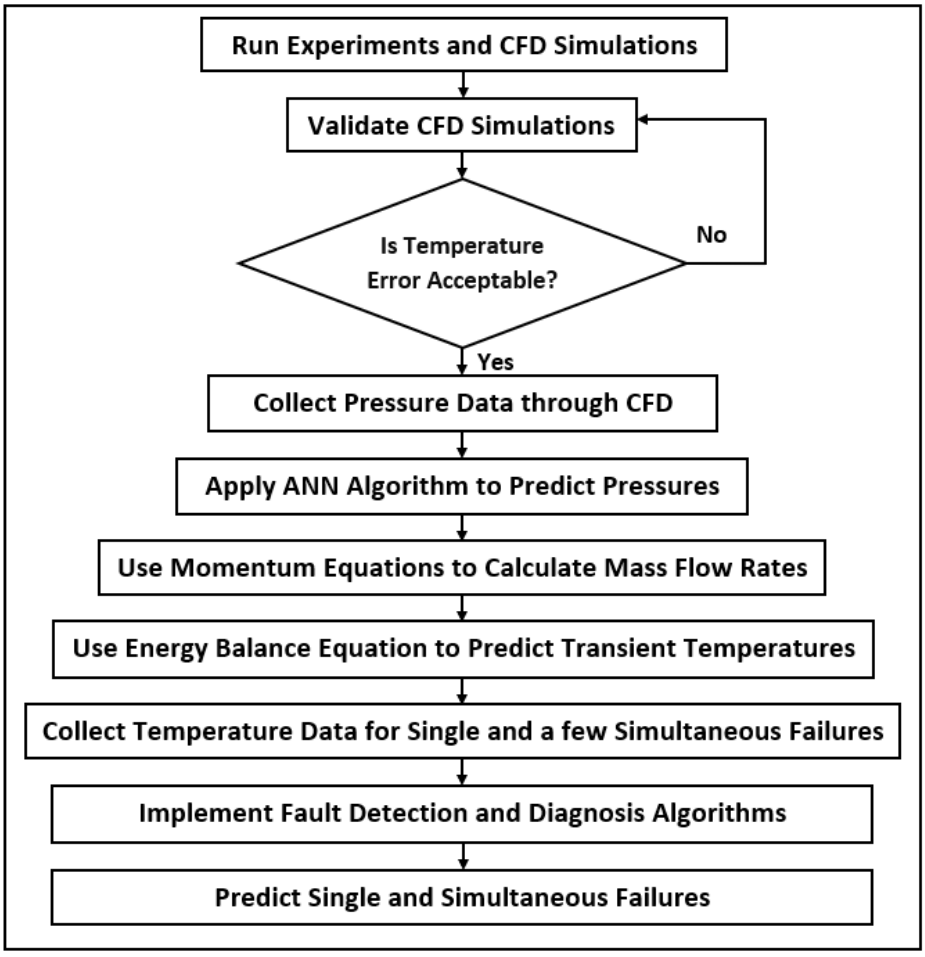


Figure 7-1. Schematic steps of the study.

7.2 Future directions

The results and findings in this work indicate that improving thermal behavior models in DC shows considerable potential for future development. Therefore, the following avenues for future research are recommended based on the results of this research:

- Improving the proposed gray-box model by considering zones for brushes between racks and implement mass and momentum equations.
- CFD model calibration using neural net approaches to improve the model accuracy
- Extending the proposed gray-box model for different kinds of DCs, such as raised-floor and rack-mountable cooling unit DCs.
- Developing a single gray-box model able to predict thermal behavior for different DCs.
- Using the model to estimate exergy destruction of a DC and the potential of waste heat recovery.
- Employing the proposed models for thermal aware workload management in DCs.
- Developing model predictive controllers (MPC) for DCs with row-based cooling architectures using the proposed model.