AI-ENABLED SMART FIELD CONTROL OPTIMIZATION FOR BUILDING CENTRAL COOLING SYSTEMS

XIE LINGYUN

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Building Environment and Energy Engineering

AI-Enabled Smart Field Control Optimization for Building Central Cooling Systems

Xie Lingyun

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Aug 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no materials previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

________________________(Signed)

XIE Lingyun (Name of student)

ABSTRACT

Abstract of thesis entitled: AI-enabled smart field control optimization for building central cooling systems

Submitted by : XIE Lingyun

For the degree of : Doctor of Philosophy

at The Hong Kong Polytechnic University in August, 2025

Rapid advances in Artificial Intelligence (AI) are reshaping energy-intensive industries, offering transformative opportunities to optimize complex engineering systems for both sustainability and performance. Given that building HVAC systems account for a large portion of global energy usage, researchers have increasingly explored AI-enabled strategies to enhance intelligent facility management and advance energy conservation goals. However, practical deployment in real-world environments remains constrained by key barriers, including high computational demands that limit real-time feasibility, reliance on centralized architectures vulnerable to communication bottlenecks, insufficient robustness under system uncertainties, limited adaptability to equipment degradation, and scalability challenges in complex multi-zone applications.

This PhD research addresses the limitations of current HVAC optimization practices by proposing a generic AI-empowered framework and a suite of innovative control strategies—including lightweight online optimization, energy-gradient-guided decision-making, incremental online learning, and parallel distributed architectures—thereby enabling efficient, robust, and scalable real-time deployment of intelligent optimization in building central cooling systems.

A comprehensive review is conducted to examine the progression of HVAC control strategies, ranging from traditional rule-based methods to contemporary AI-enabled approaches. The review reveals notable progress in the theoretical development of intelligent control strategies. It also reveals several persistent challenges that limit the real-world applicability of current strategies. A notable lack of autonomous, interpretable, and resilient AI mechanisms underscores the gap between theoretical advances and practical implementation. These insights define the research objectives and guide the design of the proposed strategies.

To enable AI-enabled field-level optimization, a generic framework and associated strategies are developed, integrating AI-based control optimization into BAS by adopting smart control stations at the BAS field level. The proposed framework comprises two core functional modules: (1) an AI operating environment that supports lightweight and real-time execution of AI models, and (2) a comprehensive suite of AI functional boxes that ensure effective and reliable execution of AI algorithms. The framework is validated by integrating two physical smart control stations with a BAS testbed. Control robustness and energy performance are further assessed through hardware-in-the-loop testing with a simulated dynamic HVAC system. Test results show that the proposed framework and associated strategies can maintain stable and robust control under various critical conditions. Building on this foundation, the subsequent optimization strategies are also validated on the same hardware-in-the-loop testbed using smart control stations.

To address the challenge of computational burden in real-time optimization, a simplified deep learning-enabled Genetic Algorithm is proposed, ensuring optimization intervals are short enough for online applications. This algorithm also significantly reduces CPU and memory usage, enabling deployment on a miniaturized control station for field implementation. To further enhance stability and reliability, a robust assurance scheme is incorporated, which switches to expert knowledge-based control under abnormal conditions. Hardware-in-the-loop testing validates the proposed strategy, confirming its computational efficiency, control effectiveness, and operational robustness. Test results show that the optimal control strategy achieves an average of 7.66% energy savings and exhibits strong operational robustness.

To enhance robustness against data uncertainties, a novel local energy gradient-guided AI-empowered online optimization strategy is proposed. Relative energy consumption in neighboring regions is employed to guide the optimization process, thereby improving control stability and enhancing resilience to input data noise. This strategy involves theoretically formulating a physical variable, the “energy gradient”, to guide the AI-empowered online optimization method. A machine learning model is employed to capture the trend of the energy gradient, guiding the progressive adjustment of setpoints. A comprehensive system mechanism analysis is conducted to uncover system dynamics and provide theoretical support for the proposed strategy. Results show that the proposed strategy exhibits superior stability and robustness, with an energy-saving effect comparable to those achieved by conventional online optimization strategies, achieving approximately 5% energy savings.

To improve long-term adaptability under system degradation and dynamics, an incremental online learning strategy using structural augmentation of the optimization model is introduced. A residual neural network branch is added alongside the offline-trained hybrid model and selectively updated online, enhancing adaptability while preserving pre-trained knowledge. A low learning-rate Adam optimizer is adopted to continuously update the model using real-time data, maintaining stability and preventing catastrophic forgetting. Test results show that the incremental online learning approach effectively restores system performance after degradation, reducing energy consumption by 5.77%, while the original optimization algorithm achieves only 1.12% savings under degraded conditions.

To enhance system-level scalability and fault tolerance, a parallel distributed optimal control strategy is developed. Based on a decentralized operational framework, agents spatially distributed across field stations or edge devices can perform optimization tasks independently. By leveraging dynamic task partitioning, peer-to-peer communication, and voting-based consensus mechanisms, these agents collaboratively achieve global system optimization without the need for centralized coordination. Hardware-in-the-loop tests are conducted by deploying the strategy on four real smart control stations. Test results show that the proposed strategy improves scalability, operational reliability, and optimization efficiency, while delivering 9.16% energy savings.

In conclusion, the framework and strategies developed in this PhD research bridge the gap between academic innovation and practical deployment of AI-enabled control optimization in large-scale building cooling systems, offering a scalable, reliable, and energy-efficient pathway toward intelligent building operations.

PUBLICATIONS ARISING FROM THIS THESIS

Journal Papers

[1] **Lingyun Xie**, Kui Shan, Hong Tang, Shengwei Wang. AI-empowered online control optimization for enhanced efficiency and robustness of building central cooling systems. *Advances in Applied Energy, 2025, 18*, 100220.

[2] **Lingyun Xie**, Kui Shan, Shengwei Wang. A generic framework and strategies for integrating AI into building automation systems for field-level optimization of HVAC systems. *Energy*, 2025, *333*, 137405.

[3] **Lingyun Xie**, Kui Shan, Hong Tang, Shengwei Wang. AI-empowered online optimization adopting local energy gradient for robust and energy-efficient control of building cooling systems. *Journal of building engineering, under review*.

[4] **Lingyun Xie**, Kui Shan, Hangxin Li, Shengwei Wang. A parallel distributed approach for AI-enabled online optimization of building cooling systems with enhanced efficiency, reliability and scalability. *Applied Energy, under review*.

Conference Papers

[1] **Lingyun Xie**, Kui Shan, Hong Tang, Shengwei Wang. (2025). A hybrid model-based online learning technology for AI-enabled adaptive optimization control in building central chiller plants. *The 13th International Conference on*

Sustainable Development in the Building and Environment, July 27 to 31, 2025, London, England.

Patents

[1] Shengwei Wang, **Lingyun Xie,** Hangxin Li. An embedded smart control station for the optimal control of building energy systems. *China patent:* 202521375575.1.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Shengwei Wang, for his unwavering guidance, encouragement, and intellectual support throughout the course of my PhD journey. His profound insights and patient mentoring have been instrumental in shaping both the direction and the quality of this research. I am also grateful to Professor Fu Xiao, whose insightful discussions and constructive suggestions have provided me with new perspectives and helped me overcome many research challenges.

I am sincerely thankful to Dr. Kui Shan and Dr. Hangxin Li for their valuable advice and constructive feedback at various stages of this work. Their critical questions and perspectives greatly enhanced the depth and rigor of my study. Special thanks go to Professor Xiuming Li for his kind assistance with the hardware construction of the control station.

My heartfelt thanks also go to my colleagues and friends in the research team for their generous sharing of experiences, stimulating discussions, collaborative spirit, and moral support during challenging times.

Finally, I am deeply grateful to my family, particularly my parents, for their unconditional love, sacrifices, and faith in me. Without their continuous encouragement and emotional support, this journey would never have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

Abbreviations

ACO	Ant Colony Optimization
AI	Artificial Intelligence
ANN	Artificial Neural Network
BAS	Building Automation System
BPNN	Back-Propagation Neural Network
BWO	Beluga whale optimization
CNN	Convolutional Neural Network
COP	Coefficient of Performance
CPU	Central Processing Unit
DDC	Direct Digital Control
DL	Deep Learning
DR	Demand Response
EER	Energy efficiency ratio
ELM	Extreme Learning Machine
FIS	Fuzzy Inference System
FLOPs	Floating-point operations per second
GA	Genetic Algorithm
GPU	Graphics Processing Unit
GWO	Grey Wolf Optimizer
XGBoost	Extreme Gradient Boosting
HIL	Hardware-in-the-loop
HVAC	Heating Ventilation Air Conditioning
IAQ	Indoor Air Quality
IPWOA	Improved Parallel Whale Optimization Algorithm
ISO	International Organization for Standardization
KNN	K-nearest neighbors
LAN	Local Area Network
LMTD	Logarithmic mean temperature difference
ML	Machine Learning

MPC	Model Predictive Control
NTU	Number of transfer units
PID	Proportional-Integral-Derivative
PLR	Partial Load Ratio
PMV	Predicted Mean Vote
PPD	Predicted Percentage Dissatisfied
PSO	Particle Swarm Optimization
PV	Photovoltaics
RF	Random Forest
RL	Reinforcement Learning
SD	Standard Deviation
SMA	Simple Moving Average
SSA	Sparrow search algorithm
SVM	Support Vector Machines
SVR	Support Vector Regression
TPU	Tensor Processing Unit
TR	Tons of refrigeration

Symbols

k	Penalty factor
L	Loss function
M	Volume flow rate, Ls^{-1}
N	Number of operational units
P_{tot}	Total energy consumption
Q	Heat exchange capacity, kJ
R	Regularization terms
T	Temperature, K
W	Power consumption, kWh

Greek letters

λ	Regularization coefficient
η	Learning rate
θ	Model parameters

Subscripts

e	Evaporator
ch	Chillers
cw	Cooling water
$chws$	Return chilled water

$comp$	Compressor
cws	Supply cooling water
cwr	Return cooling water
db	Dry-bulb
nom	Nominal conduction
set	Setpoint
sw	Seawater
wb	Wet-bulb

CHAPTER 1 INTRODUCTION

This chapter provides an overview of the thesis, which is divided into three sections. Section 1.1 introduces the background and motivation behind the study. In Section 1.2, the aim and objectives of the thesis are presented. Section 1.3 outlines the organization of the thesis and provides a brief description of each chapter.

1.1 Background and motivation

Global energy consumption has grown sharply in recent decades, primarily due to economic growth, rapid urbanization, and rising expectations for thermal comfort in residential and commercial spaces (Swagatika Priyadarshini et al., 2025). Within this trend, the building sector stands out as one of the largest contributors, responsible for nearly 40% of global energy use, of which heating, ventilation, and air conditioning (HVAC) systems constitute a substantial portion (International Energy Agency, 2021). In high-density urban regions such as Hong Kong, buildings consume over 80% of total energy, with almost 40% attributed to air conditioning alone (Z. Chen et al., 2023). These figures highlight the urgent need to improve the operational efficiency and sustainability of building cooling systems as a critical pathway toward meeting climate targets and mitigating the environmental impacts of rising energy demand.

Despite the pivotal role that HVAC systems play in building energy consumption, their operation is still predominantly governed by conventional control strategies, such as simple on/off switching and Proportional-Integral-Derivative (PID) control (Aste et al., 2017; Ulpiani et al., 2016). While these approaches are valued for their simplicity and cost-effectiveness, they are inherently limited in addressing the highly nonlinear, dynamic, and uncertain nature of building thermal environments (Borase et al., 2021). These shortcomings are particularly evident under fluctuating occupancy patterns, variable outdoor climatic conditions, and progressive degradation of system components. As a result, HVAC systems frequently operate in suboptimal states, causing unnecessary energy consumption and reduced overall performance (Parvin et al., 2021).

Recent advances in computational capabilities, combined with the increasing availability of large-scale datasets generated by smart meters, Internet of Things (IoT) devices, and advanced sensor networks, have created new opportunities for applying Artificial Intelligence (AI) to building system optimization (Biswas et al., 2024; Halhoul Merabet et al., 2021). AI techniques—including machine learning, deep learning, and heuristic algorithms— have shown considerable potential in tackling the complexity, nonlinearity, and uncertainty inherent in HVAC operations. These methods enable predictive modeling, anomaly detection, and adaptive optimization, thereby enhancing both energy efficiency and system resilience (Wong & Li, 2010).

Despite these advances, the large-scale deployment of AI-based solutions in real-world building environments remains constrained by critical challenges spanning computational, operational, and architectural domains. Overcoming these barriers is essential to achieving reliable, efficient, and scalable AI-enabled optimization control in practice. The main challenges can be summarized as follows:

- *Reliance on centralized architectures and onsite infrastructure limitations.* Many current AI applications depend on cloud-based or centralized servers for computation and decision-making. Although such architectures provide abundant computational resources, they introduce drawbacks including network latency, bandwidth constraints, data privacy concerns, and vulnerability to communication failures (Bi et al., 2015; Wollschlaeger et al., 2017). Deploying optimization control directly at the field level can alleviate transmission bottlenecks; however, most existing field platforms lack the computational capacity, runtime environments, and functional support required for executing advanced algorithms effectively. Furthermore, the absence of a unified onsite framework that integrates both hardware and software support for AI technologies limits their practical adoption and hinders the creation of standardized datasets, thereby exacerbating current issues of data quality and availability.

- *Computational complexity and real-time feasibility.* Advanced AI optimization algorithms often involve significant computational demands. While these methods offer high accuracy and predictive capabilities, their computational intensity poses difficulties for real-time implementation, particularly in large-scale or multi-zone systems. Recent advances in processors and computing hardware have greatly increased available computational power at the server level; however, communication latency and reliability issues still constrain cloud-based solutions. At the same time, field-level platforms must balance computing capability, cost, and compatibility to enable practical AI deployment, yet such platforms remain scarce. Under these circumstances, balancing algorithmic complexity with operational feasibility is crucial for achieving real-time, reliable AI-enabled optimization control.
- *Limited adaptability to system changes and equipment degradation.* AI models trained on historical datasets often exhibit reduced reliability when faced with system changes such as equipment aging, sensor drift, or evolving operational patterns (Veloso & Loschi, 2021). Over time, discrepancies between the

optimization model calculations and actual system behaviour can lead to suboptimal control actions and even destabilize system performance.

- *Scalability and reliability in control architectures*. Distributed control frameworks have been explored to reduce computational burden and improve resilience. However, many existing solutions employ hierarchical architectures that rely on a central coordinator to integrate local optimization outcomes and maintain global system coherence (Su & Wang, 2020). Such reliance introduces single points of failure and can constrain scalability and flexibility (Ding et al., 2024).

These challenges underscore a significant research gap. Despite substantial academic progress, a considerable divide remains between theoretical studies on AI-based optimization for chiller systems and their practical deployment in real-world building environments. There is an urgent need for practical, scalable, and robust operational frameworks that can effectively implement AI-enabled optimization strategies at the field level. At the same time, advances in both optimization algorithms and control methodologies are required to reduce computational complexity, improve optimization efficiency, and ensure operational reliability. Such capabilities are critical for enabling control strategies that dynamically adapt to evolving system conditions while maintaining real-time performance in complex, large-scale building environments. Bridging this gap is essential to translating the theoretical advancements in artificial intelligence into tangible energy savings and measurable operational improvements in practice.

1.2 Aim and objectives

The PhD study is dedicated to advancing the integration of AI-enabled methodologies with the practical operational requirements of building central cooling systems. Its primary aim is to develop innovative solutions that embed AI directly into field-level Building Automation Systems (BAS), thereby overcoming computational and architectural constraints while delivering adaptive, efficient, and robust performance under dynamic, large-scale operating conditions. By tackling both technical and implementation challenges, the study seeks to make substantial theoretical and practical contributions to the development of next-generation intelligent control strategies for building energy systems, enabling significant progress toward sustainable and energy-efficient building operations. The specific objectives of this research are as follows:

i. Conduct a comprehensive literature review to systematically examine the evolution of control optimization strategies for HVAC systems, identify key limitations in efficiency, adaptability, robustness, scalability, and real-world feasibility, and establish the research gaps that inform this study.

ii. Design and implement a generic framework for integrating AI-driven online control optimization into BAS, enabled through smart control stations deployed at the field level. The framework will support lightweight, real-time AI model execution, ensuring both optimization effectiveness and operational reliability.

iii. Develop an AI-empowered online control optimization strategy that combines a simplified deep learning–enhanced Genetic Algorithm with a robust assurance scheme to improve computational efficiency and ensure stable, reliable control in practical deployments.

iv. Incorporate physical variables such as energy gradients into the AI optimization process to guide decision-making based on relative rather than absolute energy changes, thereby enhancing robustness to measurement noise, modeling errors, and data anomalies.

v. Propose an incremental online learning approach with structural model augmentation, enabling continuous adaptation of the optimization model to evolving system dynamics and equipment degradation while preserving long-term energy efficiency and operational reliability.

vi. Establish a scalable, fault-tolerant parallel distributed optimization framework in which multiple agents execute optimization tasks collaboratively without a central coordinator, thereby improving computational efficiency, scalability, and resilience for large-scale building systems.

1.3 Organization of this thesis

This PhD thesis consists of eight chapters, organized as follows.

Chapter 1 outlines the background, motivation, and research objectives, highlighting the need for advanced control strategies in central cooling systems and the opportunities enabled by AI technologies. The chapter also presents the overall structure.

Chapter 2 provides a comprehensive literature review covering the evolution of control strategies for building HVAC systems. It details traditional rule-based and optimization-based approaches, as well as recent advances in AI-based methods, including machine learning, deep learning, and heuristic optimization. The chapter also explores existing frameworks for field-level control, distributed architectures, and adaptive strategies to cope with system changes. Critical research gaps are identified to motivate the proposed solutions.

Chapter 3 introduces the development of a generic framework and strategies for integrating AI into BAS for field-level optimization of central cooling systems. It details both the hardware and software architecture of the proposed smart field control stations, explaining how the integration of AI algorithms into embedded systems enables local decision-making while maintaining real-time responsiveness and reliability.

Chapter 4 describes the design and implementation of AI-empowered online control optimization methods for enhancing the efficiency and robustness of building central cooling systems. It elaborates on the formulation of the optimization problem, discusses innovative algorithmic solutions that combine machine learning and heuristic techniques, and analyzes how these methods reduce computational complexity while maintaining high performance. The chapter also presents simulation and experimental results validating the effectiveness of the proposed strategies under dynamic operational conditions.

Chapter 5 presents the development of an energy gradient-guided AI-empowered online optimization strategy. This chapter introduces the novel concept of using relative energy variations—termed “energy gradients”—to guide optimization processes. It explains how this physical variable helps overcome the challenges associated with relying solely on absolute energy measurements, improving the robustness of the control system in the presence of measurement noise, modeling errors, and operational uncertainties. The proposed method’s theoretical underpinnings, algorithmic design, and experimental validation are detailed in depth.

Chapter 6 focuses on incremental online learning with model structural augmentation for AI-enabled control optimization, specifically addressing the challenges posed by equipment degradation in long-term system operation. The chapter discusses the limitations of static AI models and introduces online learning mechanisms that allow models to evolve in response to changing system dynamics. It elaborates on techniques for updating model parameters incrementally without triggering forgetting, ensuring continued accuracy and operational efficiency. A hardware-in-the-loop test is presented to demonstrate the effectiveness of the proposed adaptive control methods.

Chapter 7 develops a parallel distributed approach for AI-enabled online optimization in building cooling systems. It details the development of a multi-agent architecture that eliminates the need for centralized coordination, thus improving computational efficiency, scalability, and fault tolerance. The chapter discusses the algorithmic mechanisms enabling cooperative optimization among distributed agents, as well as elite migration strategies that enhance global optimization performance. Hardware-in-the-loop testing results are provided to validate the robustness and practical applicability of the proposed distributed framework.

Chapter 8 summarizes the key findings and contributions of this research, highlighting the novel methodologies and technological advancements developed throughout the study. It discusses the implications of the proposed strategies for real-world deployment in building cooling systems and identifies several promising directions for future research. These include further exploration of autonomous decision-making under extreme conditions, integration with broader building energy ecosystems, and potential applications of emerging AI technologies in the field.

CHAPTER 2   LITERATURE REVIEW

The landscape of building energy management is undergoing a rapid transformation, driven by global sustainability goals, the rapid proliferation of Internet of Things (IoT) devices, and the advancing capabilities of artificial intelligence (AI). Although academic research on AI-enabled control optimization has produced promising results, its practical deployment in real-time, large-scale, and complex building environments—particularly for cooling systems—remains a significant challenge. This chapter presents a comprehensive review of related work across three dimensions: (i) conventional and advanced control strategies for HVAC systems, (ii) state-of-the-art AI-enabled optimization approaches and their inherent limitations, and (iii) recent developments in field-level distributed control optimization frameworks. These perspectives collectively establish the foundation for identifying the research gaps that this thesis aims to address.

2.1 Control Optimization Strategies in Building HVAC Systems

2.1.1 Control optimization objectives

Control optimization in building HVAC systems has become a key approach for improving operational efficiency while maintaining indoor environmental quality. The primary objectives are typically grouped into five categories: thermal comfort, indoor air quality (IAQ), energy efficiency, operational cost, and system performance metrics such as the coefficient of performance (COP) and energy efficiency ratio (EER) (Azzi et al., 2024).

Thermal comfort remains a foundational criterion, typically quantified using indices such as Predicted Mean Vote (PMV) or Predicted Percentage Dissatisfied (PPD).

Barone et al.(Barone et al., 2023) developed a physiological thermal comfort model integrated with a building simulation tool for HVAC optimization, enabling dynamic control of air temperature and humidity and eliminating over 3,650 annual hours of thermal discomfort compared with fixed set-point strategies. Similarly, Carli et al.(Carli et al., 2019) applied a model predictive control strategy that optimizes thermal comfort based on the PMV index, allowing online adjustment of indoor conditions while minimizing energy consumption under real-world operating scenarios.

From an operational perspective, energy efficiency and system efficiency are often the dominant control goals, especially under growing pressure for carbon neutrality. Many studies have formulated objectives around minimizing total energy consumption or maximizing COP. For example, Yuan et al.(Yuan et al., 2021) proposed a model-free reinforcement learning control strategy for HVAC optimization, achieving up to 7.7% energy savings over rule-based control, along with notable reductions in discomfort and operational costs after long-term exploration. Asad et al.(Asad et al., 2019) developed an adaptive model-based real-time optimization framework for large-scale HVAC systems, which improves energy efficiency by updating performance models online using real-time measurements, achieving 8% energy savings and a 99% reduction in computational burden. Homod(Homod, 2018) proposed a hybrid control strategy combining fuzzy inference and neural networks for HVAC systems, which enhanced thermal comfort and achieved significant energy savings compared to traditional hybrid PID cascade control.

Operational cost optimization has also gained prominence, particularly in the context of demand response (DR) programs. Under time-of-use pricing and peak demand charges, cost-minimization strategies become essential. Sun et al.(Sun et al., 2024) proposed a hierarchical HVAC control framework that minimizes electricity costs through demand allocation using genetic algorithms, achieving a 6.5% operational cost reduction under dynamic electricity pricing in a real commercial building. Yang et al.(Y. Yang et al., 2020) proposed a decentralized control framework for multizone HVAC systems that incorporates inter-zone thermal coupling and employs an accelerated distributed augmented Lagrangian method to minimize operational costs efficiently.

Across these studies, optimization problems are typically formulated as multi-objective or constrained single-objective problems, where energy, cost, or comfort indices are minimized or maximized under specific operational constraints. However, most existing strategies remain rooted in deterministic formulations focused on static targets. While effective under stable conditions, such approaches often lack adaptability in dynamic, uncertain environments. This limitation motivates part of the present research, which emphasizes relative energy variations as decision-making signals to enhance robustness and practical applicability in real-world engineering contexts.

2.1.2 Control optimization methods

Control optimization methods for HVAC systems have evolved substantially, progressing from traditional rule-based techniques to sophisticated AI-driven methodologies. This section provides a chronological and methodological overview of key control paradigms, categorized into traditional control, deterministic optimization, and artificial intelligence (AI)–enabled approaches.

Traditional control methods

In practical HVAC engineering, traditional control methods—such as rule-based logic, on/off switching, and Proportional–Integral–Derivative (PID) control— remain widely deployed due to their simplicity, reliability, and ease of implementation. These approaches typically rely on fixed thresholds or empirical relationships to regulate equipment operation. While effective in stable conditions, their static logic cannot adapt to dynamic environmental or internal variations, including occupancy changes, fluctuating loads, and shifting outdoor weather (Borase et al., 2021; Dai et al., 2024). As a result, these methods often lead to unnecessary energy consumption, comfort deviations, and inadequate responsiveness under real-world operating conditions.

Deterministic optimization approaches

To address the limitations of static control logic, deterministic optimization methods—such as gradient descent (Hochreiter et al., 2001), Newton method (Mu et al., 2015), and dynamic programming (Deng et al., 2015)—were introduced to provide more flexible and predictive control frameworks. These methods depend on explicit mathematical models representing the thermodynamic and control behavior of HVAC systems. For example, Pombeiro et al.(Pombeiro et al., 2017) employed a dynamic programming approach to optimize HVAC operations in a PV-integrated building, minimizing operational costs while maintaining thermal comfort under time-varying electricity prices and battery constraints. Vakiloroaya et al.(Vakiloroaya et al., 2013) developed a gradient projection–based optimization strategy for air-cooled HVAC systems using empirical simulation models, enabling optimal setpoint determination under transient conditions and achieving notable energy savings in real-world deployments.

While deterministic optimization provides theoretical rigor and controllable outcomes, it faces practical challenges. These include high computational demands, limited scalability, sensitivity to initial conditions, and the risk of convergence to local optima.

Such limitations reduce their suitability for complex, nonlinear, and real-time HVAC applications.

Artificial Intelligence-enabled strategies

To overcome these limitations, AI–based methods have gained increasing attention, offering data-driven flexibility and strong adaptability to complex, nonlinear, and uncertain environments. AI techniques for HVAC optimization can be broadly classified into machine learning (ML), deep learning (DL), heuristic algorithms, and fuzzy control, each with distinct advantages. These methodologies are examined in detail in 2.2.

2.1.3 Control optimization architectures

The architecture of control optimization in building HVAC systems significantly influences system scalability, responsiveness, fault tolerance, and computational efficiency. Based on the degree of centralization and agent autonomy, these architectures are generally classified into four categories: centralized, hierarchical centralized, hierarchical distributed, and fully distributed systems. Each reflects a distinct trade-off between global coordination and local autonomy, with varying implications for real-world deployment.

Centralized control architecture

Centralized architecture is widely adopted in current HVAC optimization frameworks, particularly those involving advanced control algorithms such as Model Predictive Control (MPC) or heuristic optimization techniques. In this structure, a single supervisory controller— typically located in a central control room or cloud platform—collects real-time data from all subsystems, performs system-wide optimization calculations, and dispatches setpoints to local actuators (Espín-Sarzosa et al., 2020).

While effective in small-medium-scale systems, this approach suffers from scalability constraints, high communication overhead, and single points of failure, making it unsuitable for large-scale or geographically dispersed systems under real-time or unreliable network conditions.

Hierarchical centralized control architecture

To mitigate the limitations of pure centralization, hierarchical centralized architectures divide the system into multiple control layers. The top layer performs global optimization, while lower layers execute localized control tasks, reducing computational burden and improving response times. For example, Zhao et al.(Zhao et al., 2021) proposed a hierarchical centralized control architecture for the integrated battery thermal management and HVAC system in connected HEVs, where an upper-layer planner optimizes temperature trajectories based on look-ahead driving data, and a lower-layer MPC ensures real-time tracking and energy efficiency, achieving up to 10.47% HVAC energy savings. Wang et al. (H. Wang & Wang, 2021) used a hierarchical structure for HVAC demand response, where a regulation bidding controller coordinated with a power-following controller to maintain comfort while optimizing baseline power and regulation capacity.

Despite these improvements, such systems still depend on reliable inter-layer communication and remain vulnerable to supervisory layer failures, limiting adaptability under dynamic conditions.

Hierarchical distributed control architecture

This architecture retains a hierarchical structure but introduces distributed computation at lower layers. Local controllers (agents) perform partial optimization of their subsystems, while a coordinator ensures global coherence (Jendoubi & Bouffard, 2023). The overall optimization task is decomposed into subproblems solved locally and integrated via coordination mechanisms. For instance, Su et al. (Su & Wang, 2020) applied dual decomposition for distributed optimal control of HVAC systems, breaking down the complex optimization task into several simpler ones. Task management is distributed between a coordinator agent and a set of component agents, supporting efficient system operation through decentralized and cooperative control. Similarly, Mtibaa et al.(Mtibaa et al., n.d.) employed a central agent to coordinate local agents dedicated to regional control optimization, successfully achieving energy savings of 8.8%.

While hierarchical distributed approaches improve robustness and local responsiveness, the coordinating layer still represents a partial centralization bottleneck, introducing synchronization overhead and scalability limits in large-scale applications.

Fully distributed control architecture

Fully distributed architectures represent a paradigm shift toward complete decentralization. Here, autonomous agents communicate peer-to-peer, make local decisions based on both local data and exchanged information, and reach system-wide consensus via iterative negotiation or distributed optimization (N. Yang et al., 2020; Zhang et al., 2023).

This architecture offers superior scalability, fault tolerance, and robustness, particularly under asynchronous updates, heterogeneous configurations, or communication delays. For example, Li and Wang (W. Li & Wang, 2022) proposed a fully distributed optimal control strategy for multi-zone HVAC systems, where control agents exchange information locally without a coordinator, achieving higher robustness, lower complexity, and better control performance than hierarchical distributed approaches in IoT-enabled building automation networks. Similar successes have been reported in power systems (Simpson-Porco et al., 2015), vehicle networks (T. Jiang et al., 2019), and industrial automation (Zhu & Ren, 2021).

Despite these advances, the application of fully distributed control architectures in AI-enabled optimization of building cooling systems remains insufficiently explored. In particular, it remains unclear how to decompose AI-enabled control optimization tasks across multiple agents in a parallel and cooperative manner without relying on physical subsystem boundaries. In real-time control optimization scenarios, additional challenges arise in enabling efficient inter-agent communication and consensus-based decision-making under limited computational and communication resources. Furthermore, dynamic and heterogeneous system environments (such as fluctuating agent availability (Fan et al., 2022), asynchronous computation (Assran et al., 2020), and intermittent connectivity (Azemi & Wahid, 2021) pose substantial challenges to achieving robust real-time operation. Effectively addressing these challenges is essential for enabling scalable, reliable, and energy-efficient control optimization in practical cooling system applications.

2.2 Technologies of AI-based strategies in building HVAC systems: State-of-the-art

2.2.1 Definitions and scope of AI for building and engineering applications

Artificial Intelligence (AI) has been defined in diverse ways by scholars and institutions over the decades. In 1955, John McCarthy, widely regarded as the father of AI, described it as *"the science and engineering of making intelligent machines"* More recently, Gillath defined AI as *"simulation of human intelligence processes by machines"* (Gillath et al., 2021). The International Organization for Standardization (ISO) further specifies AI as *"engineered systems that can generate intelligent outputs such as forecasts and recommendations"* (The International Organization for Standardization, 2020). Collectively, these definitions emphasize AI’s central objective: to create systems that can mimic or simulate human cognitive functions such as learning, reasoning, and problem-solving.

The evolution of AI is illustrated in Figure 2.1. Over the decades, AI has progressed through multiple waves: from symbolic logic in the 1950s–1970s (Shortliffe et al., 1979), to expert systems in the 1980s (Kastner & Hong, 1984), and later to data-intensive machine learning in the 1990s. Today, AI has entered a new era of accelerated development. Breakthroughs in big data, high-performance computing (GPU/TPU), and deep learning (DL) have given rise to various AI subfields, such as large language models (e.g., GPT series, DeepSeek)(Naveed et al., 2024) and computer vision frameworks (e.g., YOLO, ViT)(Rajput et al., 2024). AI technologies are widely applied in autonomous driving, medical diagnostics, financial forecasting, and many other domains, continually reshaping industries and pushing the boundaries of intelligent systems.

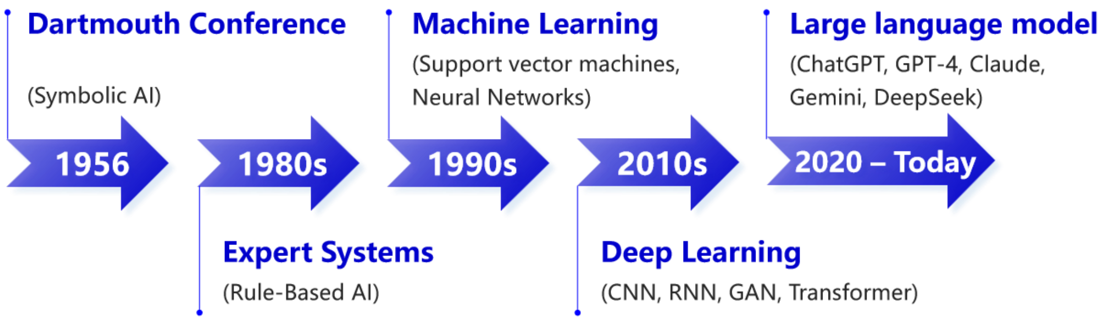


Figure 2.1 The evolution of artificial intelligence

While the historical and technological evolution of AI has significantly expanded its applications, it is essential to distinguish between two fundamental definitions of AI, particularly in the context of building energy applications.

a) Broad definition of AI

This type of definitions encompasses the AI technologies that dominate both historical development and contemporary applications. It includes expert systems, conventional machine learning models, and computerized domain-specific automation. These technologies are useful in improving efficiency, automating repetitive tasks, and providing automated smart decisions without online human involvement. In building energy applications, such systems are widely applied for energy optimization (e.g., supervisory controls), fault detection, and predictive maintenance. However, they lack general reasoning capabilities and human-like autonomy. They typically require human involvement in establishing and updating embedded knowledge bases, and adapting to unforeseen scenarios. As a result, while these tools enable smarter operations, they remain functionally constrained by their design assumptions and operating environments.

b) Rigorous definition of AI

A rigorous definition of AI aligns closely with its foundational ambition: achieving human-like capabilities in learning, reasoning, and autonomous problem-solving. Under this definition, AI is expected to operate with decision-making autonomy comparable to human experts, dynamically adapting to real-world uncertainties without human intervention. Recent advances, such as large language models (e.g., ChatGPT) that showcase capabilities in language comprehension and knowledge synthesis, represent meaningful progress toward this target. Likewise, several academic studies have proposed adaptive control strategies that, under specific assumptions (e.g., reliable data availability or controlled test environments), somehow fulfill the criteria of rigorous AI.

2.2.2 AI-enabled control optimization in building HVAC systems

In recent years, research on AI technologies in the building energy sector has advanced considerably, progressively influencing diverse aspects of energy-efficiency optimization, intelligent management, and sustainable operation. Artificial Intelligence (AI) technologies—such as machine learning (ML) (Ibrahim et al., 2020), deep learning (DL) (Aslam et al., 2021), heuristic algorithms (Pombeiro et al., 2017), adaptive control (Y. Jiang et al., 2023), and fuzzy control (Abuhussain et al., 2023) — enable dynamic adjustment of system operations to enhance efficiency and stability (Himeur et al., 2021). These approaches support key functionalities such as load forecasting, adaptive parameter tuning, and global optimization across multiple subsystems. Table 2.1 provides a summary of representative AI techniques applied in building energy systems.

- Machine learning (ML)

  ML techniques are widely applied in building energy consumption prediction, leveraging historical and real-time operational data to forecast demand patterns and guide energy management decisions (Zhou et al., 2021). Algorithms such as Random Forest (RF) (Z. Gao et al., 2022), Support Vector Regression (SVR) (Qin & Wang, 2022), and K-means clustering (K. Li et al., 2022) can capture nonlinear energy-use patterns and provide accurate short-term cooling and heating load predictions.

  A particularly promising subfield is Reinforcement Learning (RL) (Perera & Kamalaruban, 2021), which learns optimal control policies through continuous interaction with dynamic environments. RL has been applied to autonomously adjust chiller sequencing, valve positions, and pump speeds. For example, Dai et al.(Dai et al., 2023a) proposed an RL-enabled iterative learning control strategy for air-conditioning systems, reducing morning start-up time by improving cooling uniformity across zones and achieving up to 17.8% energy savings.

- Deep learning (DL)

  DL models, especially Artificial Neural Networks (ANNs), can capture complex, high-dimensional relationships between environmental conditions and system responses. Such models are particularly effective for black-box modeling of HVAC systems, allowing accurate prediction and control of indoor conditions based on external climate and internal occupancy data (Y. Gao & Ruan, 2021). DL can also be integrated into broader building control systems to optimize lighting and ventilation schedules based on occupancy patterns, thereby reducing unnecessary energy use (Choi et al., 2021).

- Heuristic and evolutionary algorithms

  Heuristic optimization algorithms (Ifaei et al., 2023), such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO), are well-suited for solving multi-objective optimization problems in HVAC control, balancing energy efficiency, occupant comfort, and operational costs (Z. Li et al., 2024) (L. Li et al., 2021). These algorithms can be coupled with Model Predictive Control (MPC) to dynamically identify optimal control parameters in complex, nonlinear system landscapes. For instance, Mtibaa et al. (Mtibaa et al., 2021) developed an MPC framework solved via GA for multi-zone HVAC optimization, achieving over 50% energy savings and significant reductions in occupant discomfort.

- Adaptive control and fuzzy control

  Adaptive control strategies allow HVAC systems to self-tune parameters in response to real-time performance feedback, thereby maintaining efficiency under changing operational conditions (Sun et al., 2022). Fuzzy control, in contrast, is particularly effective in environments with high uncertainty or incomplete data, as it relies on linguistic rules rather than precise mathematical models. This makes it well-suited for control scenarios involving human comfort or qualitative assessments (Martinez-Molina & Alamaniotis, 2020).

In summary, AI-driven control optimization has been studied as a way to move beyond static, rule-based operations toward more adaptive and predictive strategies. By integrating methods ranging from machine learning and deep learning to heuristic and fuzzy approaches, research indicates potential to balance comfort, efficiency, and cost-effectiveness, although large-scale real-world deployment remains limited.

Table 2.1 AI technologies on system control optimization

Refs.	AI techniques	Reference variables	Contributions
(Z. Gao et al., 2022)	Random Forest (RF) + Improved Parallel Whale Optimization Algorithm (IPWOA) + Extreme Learning Machine (ELM)	Previous cooling load, indoor personnel flow, outdoor dry-bulb temperature, indoor flow, window-to-wall ratio, and solar irradiance	• RF is used for feature selection. • IPWOA is used to optimize the neural network. • ELM serves as the predictive model, offering higher learning efficiency. • Compared with benchmark models, this method achieves higher prediction accuracy.
(Qin & Wang, 2022)	Support Vector Regression (SVR) + Reinforcement Learning (RL)	Temperature setpoint, wind speed, wind direction, temperature, humidity, level of CO_2, amount of people	• SVR is used to predict indoor conditions. • Reinforcement learning is used to train a control agent optimizing comfort and energy. • The integrated system achieved over 90% prediction accuracy.
(K. Li et al., 2022)	K-means + K-nearest neighbors (KNN)	Temperature, solar radiation, occupancy, historical load, time features	• K-means is used for clustering training data. • KNN is used to classify unknown samples. • Achieving a mean absolute percentage error of 1.03%.
(Y. Gao & Ruan, 2021)	Long Short-Term Memory (LSTM) + Attention mechanism	Weather (temp, humidity, wind), pressure, holiday	• LSTM is used to encode temporal dependencies. • Attention is used for interpretability at both the hidden state and feature levels. • Model achieved <1.3% MAPE deviation from best-case baseline.
(Choi et al., 2021)	Convolutional Neural Network (CNN)	Real-time occupant count from camera input	• CNN is used for real-time detection and counting of occupants. • Applied occupancy count to HVAC and lighting control.

(Z. Li et al., 2024)	Beluga Whale Optimization (BWO) + Sparrow Search Algorithm (SSA)	Part Load Ratio (PLR) of each chiller unit	• Achieved up to 10.2% annual energy savings. • SSA producer operator is used to accelerate convergence. • IBWO achieved lower power consumption and faster convergence • Saving up to 205 kW in real-world chiller systems.
(L. Li et al., 2021)	Back-Propagation Neural Network (BPNN) + Adaptive Grey Wolf Optimizer (GWO)	Inlet velocity, inlet temp, number and location of heat sources, airflow config	• BPNN is used to predict indoor temperature and thermal comfort index. • Adaptive GWO is used to optimize input parameters. • The proposed method improves time efficiency by 71–77%.
(Sun et al., 2022)	Adaptive least squares	Supply water temperature	• An adaptive control model is used to calculate optimal water temperature dynamically. • The proposed method cut energy use by 34.24%.
(Martinez-Molina & Alamaniotis, 2020)	Fuzzy Inference System (FIS)	Indoor temp, humidity, and occupancy	• FIS is used to control HVAC operation time. • Compared to manual control, FIS reduced the HVAC operation time by 5.7%.

2.2.3 Limitations and challenges of AI-enabled control optimization

Despite the increasing attention and substantial research efforts of AI technologies in building energy applications, practical deployment remains limited due to various technical, practical and commercial challenges. These challenges are associated to algorithmic design, data adequacy and quality, as well as real-world operating conditions. Table 2.2 summarizes the ten categories of AI techniques concerned and their challenges when applied in the building lifecycle.

- *Machine Learning (ML).* Widely applied for load forecasting, classification, and fault detection. ML models, such as SVM, RF and XGBoost, learn statistical patterns from labeled data. Their effectiveness has been demonstrated in well-instrumented environments, but the performance still depends heavily on data availability/adequacy, quality, and representativeness (Mahesh, 2020). A major barrier is the lack of cross-building data standardization, which undermines portability and scalability (Naidu et al., 2023). Reinforcement Learning (RL), a subset of ML, has shown promise in adaptive control (e.g., HVAC scheduling). However, its trial-and-error learning process is risky in real world applications, where suboptimal actions may compromise comfort or safety (Gu et al., 2024). Reinforcement Learning also requires large datasets and computational resources to converge, limiting its suitability for mission-critical facilities such as hospitals or data centers (Nguyen et al., 2020).
- *Deep Learning (DL).* Techniques, such as ANN, LSTM, and Transformer architectures, have been widely applied to modeling complex, nonlinear, and high-dimensional energy systems, particularly for sequence prediction and sensor fusion tasks. However, their computational intensity challenges deployment on

resource-constrained hardware (C. Chen et al., 2020). Moreover, their black-box nature hinders explainability, raising concerns about transparency and trust in high-stakes applications such as HVAC control or energy dispatch (ŞAHiN et al., 2025).

- *Heuristic and Evolutionary Algorithms.* Algorithms, such as GA, PSO, and ACO, are useful in offline optimization (e.g., HVAC sizing, design trade-offs), due to their ability to effectively explore nonlinear and high-dimensional solution spaces. However, their reliance on iterative, population-based computation reduces responsiveness, making them unsuitable for online adaptive control. Furthermore, once optimized, they often lack generalizability across new conditions without manual recalibration (Ghavifekr, 2021).

- *Fuzzy Logic-based control.* Fuzzy control systems approximate human reasoning under uncertainty, and have been applied when mathematical models are unavailable but expert knowledge is abundant. Nevertheless, building effective fuzzy rules requires substantial domain expertise, and updating them under changing conditions is labor-intensive (Michailidis et al., 2023). Their rigid structure and high maintenance costs limit scalability in dynamic or complex environments.

- *Adaptive Control.* These methods dynamically adjust control parameters in response to real-time feedback, providing robust performance under variability. However, their dependence on dense sensor networks introduces vulnerabilities: sensor degradation, communication delays, or missing data can impair effectiveness (Gholamzadehmir et al., 2020). Moreover, deployment costs remain prohibitive, especially in legacy buildings or retrofit projects (Mobaraki et al., 2021).

Table 2.2 Summary of functions and challenges of AI techniques in building

AI Category	Suitable Scenarios	Technical Challenges	Implementation Challenges
Machine Learning	Prediction, classification with historical data	Data quality dependency, generalization issues	Lack of standardization across diverse buildings and systems
Deep Learning	Complex nonlinear modeling, large-scale multivariate systems	High computational cost, limited interpretability	Safety and regulation concerns in critical infrastructure
Heuristic Algorithms	Multi-objective optimization, periodic scheduling	Poor real-time adaptability, long iteration times	Limited to offline optimization, lacks flexibility in dynamic control
Fuzzy Control	Systems with well-defined expert rules	Rule definition relies on expert experience, hard to scale	Costly and labor-intensive to maintain and expand rule sets
Adaptive Control	Dynamic environments with load/weather variability	Requires dense, real-time sensor feedback	High hardware and integration costs

A critical observation is that most current applications fall within the broad definition of AI, relying on pattern recognition, predictive analytics, and optimization under predefined objectives. These approaches enhance efficiency and reduce manual workload but remain constrained by design assumptions and operating contexts. At the same time, a growing body of research has begun to exhibit features consistent with the rigorous definition of AI. For example, certain adaptive control strategies or advanced learning frameworks demonstrate autonomous decision-making under controlled conditions, showing that rigorous AI is attainable in principle. However, their practical effectiveness in professional engineering contexts remains limited. Challenges such as data representativeness and deployment feasibility often restrict their practical effectiveness, meaning that only a limited number of rigorous AI approaches currently outperform traditional methods in practice.

Looking ahead, the transition from broad AI toward more rigorous AI in building energy applications is already underway, but the ultimate vision of ambitious AI—

truly autonomous systems capable of human-like reasoning, self-learning, and continuous adaptation—remains a long-term goal. Achieving this evolution will require not only algorithmic breakthroughs but also robust data ecosystems, trustworthy and explainable intelligence, resilient adaptability under uncertainty, and socio-technical integration within policy and industry frameworks.

2.3 Field-Level Control Optimization for Building HVAC Systems

The increasing complexity of building energy systems, combined with the inherent limitations of centralized and cloud-based control architectures, has driven a paradigm shift toward field-level control. In this approach, control intelligence is embedded directly within local devices and subsystems, enabling real-time decision-making without dependence on remote servers. By hosting optimization models within programmable controllers or smart stations at the field level, it becomes possible to eliminate transmission bottlenecks, safeguard data privacy, and enhance system responsiveness through direct local data interaction (Navon & Goldschmidt, 1999).

Recent research has demonstrated the potential of such field-level intelligence. For instance, Li et al.(W. Li & Wang, 2020) implemented a multi-objective optimization framework distributed across on-site sensors to enable optimal, real-time control of multi-zone ventilation systems. Dai et al. (Dai et al., 2023b) proposed a multi-agent-based distributed cooperative control strategy deployed on site-level controllers to schedule cooling supply in air-conditioning systems, achieving 19.7% electricity savings during fast demand response events. Similarly, Wang et al.(S. Wang et al., 2019) deployed a decentralized swarm intelligence optimization algorithm into distributed microprocessor units within HVAC systems, enabling collaborative global optimization and reducing energy costs through self-organized control.

While these approaches highlight the feasibility of deploying simplified or decomposed control strategies at the field level, they also reveal key limitations. Lightweight deployment often requires extensive pre-configuration efforts, which constrain scalability and adaptability. More critically, existing field-level platforms typically lack the computational capacity, runtime environments, and integrated functionalities necessary for robust execution of advanced AI algorithms (Sahni et al., 2022). As a result, most implementations rely on offline-trained models or static rule sets, restricting their long-term adaptability and falling short of the robustness required for complex, large-scale building systems operating under dynamic conditions.

2.4 Summary of Key Findings and Research Gaps

The review presented in the preceding sections confirms that AI-enabled control optimization has emerged as a powerful paradigm in building energy systems. Nevertheless, the transition from promising theoretical models to practical, scalable, and robust implementations remains limited—particularly in the context of cooling systems, which account for a substantial share of energy consumption in commercial buildings.

Despite notable progress, persistent gaps between academic research and real-world deployment can be traced to several critical limitations:

- Challenges in real-time, field-level implementation of AI-enabled strategies.

  Although advanced AI techniques for HVAC optimization have been demonstrated in simulations and pilot projects, their practical deployment on local controllers is still rare. This is largely due to hardware constraints (e.g., limited CPU power, memory, and runtime environments), the absence of standardized frameworks for embedding AI into edge computing platforms, and the complexity of integration with existing BAS infrastructures. Overcoming these barriers requires the development of lightweight AI solutions tailored to embedded hardware with real-time processing capabilities, along with flexible frameworks that enable seamless integration of AI-enabled strategies at the field level.

- Reliability challenges under uncertainty.

  Current model-based optimization strategies often depend on static system models, which limit their adaptability to equipment degradation, measurement noise, and unforeseen operational changes. Such sensitivity to modeling inaccuracies can result in performance degradation, suboptimal control, or even instability. There is a pressing need for AI-enabled methods with noise tolerance, adaptive recalibration, and self-correcting capabilities to ensure reliable long-term performance across a building system's lifecycle.

- Architectural bottlenecks from centralized and hierarchical control.

  Many AI-based optimization frameworks rely on centralized or hierarchically distributed architectures. While these structures facilitate global coordination, they introduce network latency, bandwidth limitations, and single points of failure, which hinder real-time applicability—especially in mission-critical systems. Addressing these issues requires architectures that are fully decentralized, scalable, and fault-tolerant.

- Limited progress toward autonomous, transparent, and interpretable AI control.

  Most current implementations align with the broad, functional definition of AI and operate as task-specific decision-support tools. They lack the capacity for autonomous decision-making, interpretable reasoning in safety-critical contexts, and dynamic multi-objective optimization under uncertainty. Achieving the vision of general-purpose intelligent control will require AI systems that can reason under uncertainty, balance competing objectives (e.g., comfort vs. cost), and evolve operational policies over extended time horizons without human intervention.

In summary, while AI and optimization technologies have advanced rapidly, a substantial research-to-practice gap persists in the field of building central cooling systems. This gap is rooted in computational complexity, data quality limitations, model adaptability, and architectural constraints inherent to current building automation infrastructures. Closing this gap is essential to fully realize the potential of AI-enabled smart field control, enabling buildings to achieve both high energy efficiency and operational resilience in increasingly dynamic and complex energy environments.

The subsequent chapters of this thesis introduce novel strategies designed to address these challenges, proposing deployable AI-based solutions for real-time, field-level optimization in building central chiller systems.

CHAPTER 3 DEVELOPMENT OF A GENERIC FRAMEWORK AND STRATEGIES FOR INTEGRATING AI INTO BAS FOR FIELD-LEVEL OPTIMIZATION

This chapter presents the development of a generic framework and practical strategies for integrating artificial intelligence (AI) into Building Automation Systems (BAS) to enable field-level optimization of central chiller systems. Motivated by the growing need for energy-efficient and intelligent building operations, the proposed framework bridges the gap between advanced AI algorithms and real-world BAS deployments by establishing a unified architecture that supports seamless integration, reliable execution, and scalable adaptation of AI-driven control strategies. This framework encompasses both hardware and software innovations, including the design of AI-enabled smart control stations that operate at the field level and a modular software architecture capable of hosting diverse AI models for real-time decision-making. Through systematic development and experimental validation, this chapter demonstrates how the proposed framework can effectively overcome key challenges associated with computational constraints, system integration, and operational robustness in AI-based optimization of large-scale cooling systems. The following sections detail the conceptual architecture, technical design, implementation details, and testing methodologies underpinning this generic AI integration framework.

3.1 Proposed generic framework and strategies

3.1.1 Overview of the proposed framework

The proposed framework is built around the deployment of AI-enabling smart control stations to enable field-level integration of AI control within BAS environments. As shown in Figure 3.1, these stations operate alongside existing BACnet/IP controllers and network infrastructure, forming a distributed control system that supports real-time data exchange and AI-enabled decision-making at the edge. Each control station collects operational data from HVAC equipment, executes AI-based optimization algorithms, and issues control commands in a closed-loop manner. The framework adopts a software architecture comprising two synergistic modules. The on-device execution environment enables lightweight, real-time inference of AI models, facilitating the development and deployment of AI algorithms at the edge. The functional support module ensures the orderly and stable execution of optimization tasks.

By combining embedded intelligence with protocol-level integration, the proposed framework bridges the gap between advanced AI algorithms and practical BAS implementation, enabling scalable and reliable deployment of AI-enabled optimization control across a wide range of building environments.

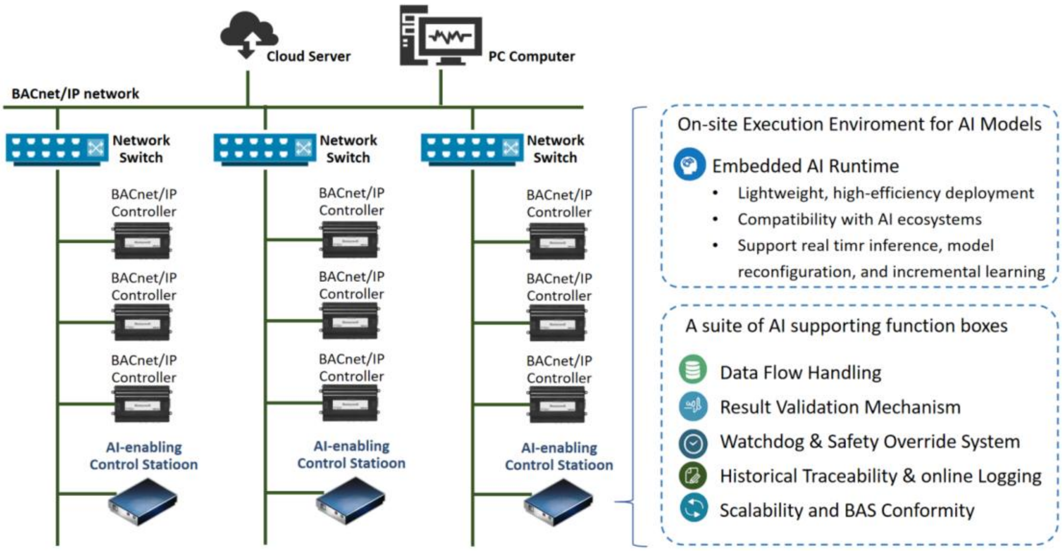


Figure 3.1 Generic AI field integration architecture and major AI supporting functions

3.1.2 Detailed design of the framework and strategies

3.1.2.1 On-site operation environment for AI models

Each smart control station is equipped with a dedicated AI-supportive runtime environment. This environment forms the computational backbone of the proposed framework, enabling real-time, edge-level AI-driven decision-making without relying on cloud or centralized computing resources. The control stations are deployed on embedded edge computing platforms, such as Raspberry Pi and NVIDIA Jetson, which offer an optimal balance of computational performance, power efficiency, and integration flexibility.

To ensure broad compatibility and future extensibility, the embedded runtime environment is designed to integrate with mainstream AI ecosystems seamlessly. It supports the deployment of models developed in high-level programming languages, particularly Python, which provides access to robust scientific computing libraries including NumPy, SciPy, and PyTorch. This setup enables flexible development and rapid prototyping of complex optimization algorithms directly at the field level, significantly accelerating model iteration and validation cycles. In addition to static inference, the edge operating environment supports a suite of software tools for real-time reasoning, adaptive model tuning, and incremental learning, allowing the system to dynamically adapt to changing HVAC loads, occupancy conditions, and operational uncertainties. This architectural design not only mitigates latency and bandwidth limitations associated with centralized computing but also enhances system robustness and data privacy by localizing both data processing and decision-making.

During the optimization process, the control strategy enabled by the AI algorithms can generally be formulated as shown in Eq. (3.1). The system state variables are collected in real time by the smart control station, providing the necessary input for optimization. Based on a trade-off between predicted energy consumption and operational requirements, the algorithm computes the optimal control variables to guide system behavior.

$$u_t = \arg\min_{u} J(x_t, u) = f(x_t, u) + \lambda \cdot C(u) \tag{3.1}$$

where, x_t represents the system state variables at time t (e.g., water temperature, flow rate). u denotes the control variables (e.g., temperature setpoints, differential pressure setpoints). $f(x_t,u)$ is the energy consumption prediction function, $C(u)$ is the constraint term to prevent extreme or unsafe control outputs, and λ is the regularization coefficient.

3.1.2.2 Supporting functions for online AI applications

Beyond the operating environment for AI tools, the proposed framework provides a comprehensive suite of AI functional boxes that bridge AI-driven decision-making with the operational logic of traditional BAS. This module supports the full control lifecycle, ensuring intelligent, stable, and advanced real-time optimization within existing BAS infrastructures. It is divided into three functional components:

Data transmission and processing

The AI-enabled control stations communicate with field-level devices and the BMS via widely adopted open protocols, including BACnet/IP, Modbus TCP, and MQTT. This allows seamless bi-directional communication and facilitates both polling-based and event-driven data acquisition strategies. The default data polling interval is set to one minute, though it can be adjusted depending on system requirements. As illustrated in Figure 3.2, a typical AI control cycle consists of four stages:

- Data acquisition: Real-time operational variables are collected from the system.
- Optimization calculations: The optimization algorithm is executed to obtain the optimal setpoints.
- Results transmission: The optimal setpoints are transmitted to the DDC controller for execution.
- Waiting for the next optimization cycle: The system enters an idle or standby state to synchronize processes and prevent potential conflicts.

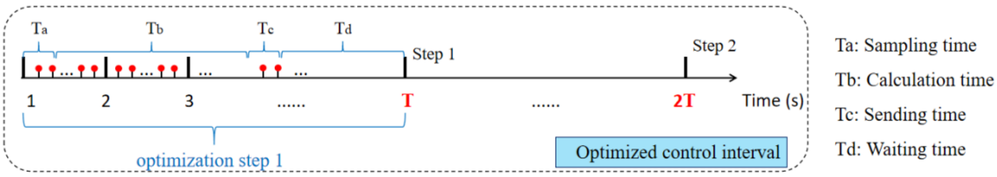


Figure 3.2 Optimization process and timing rules

To ensure the reliability and integrity of control inputs, a data preprocessing mechanism is implemented. This mechanism includes outlier detection (using statistical thresholding or range checks), missing data imputation (via linear or spline interpolation), and signal smoothing (using moving average filters). These preprocessing routines enhance data quality, enabling stable and trustworthy optimization outcomes. All sensor readings, optimization decisions, and contextual metadata are stored in a lightweight, SQL-based database (e.g., SQLite), which ensures full traceability, supports offline diagnostics, and facilitates long-term behavioral learning and auditability.

Stability and robustness assurance mechanisms

A watchdog mechanism is introduced to ensure the stability of AI execution, as illustrated in Figure 3.3. This watchdog monitors the optimization latency within each control interval and enforces a strict time budget (e.g., ≤5 minutes). If the algorithm fails to produce results within the allowable latency window, the watchdog automatically activates a fallback protocol that restores predefined default setpoints to maintain continuous system operation and prevent disruption. This predefined setpoint scheme serves as a backup strategy and can be flexibly configured based on system requirements and operational constraints. The fallback setpoints may consist of conventional fixed baseline values, the last valid optimized setpoints, or values derived from empirical formulas. This approach ensures system stability and reliability even during computational delays or failures, preventing operational disruptions and minimizing efficiency losses.

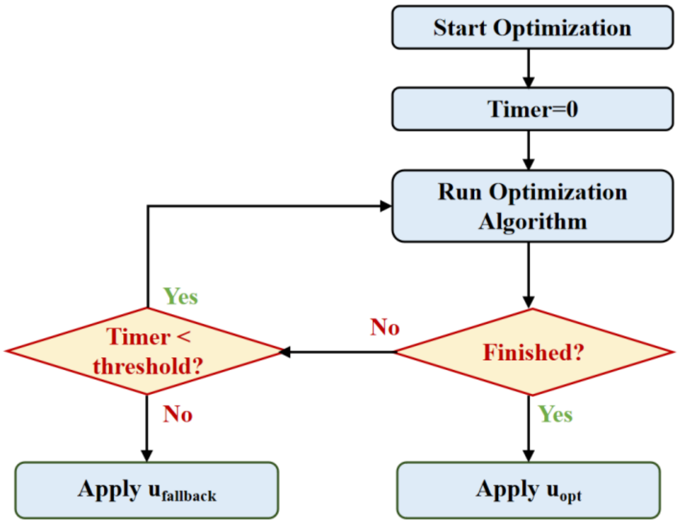


Figure 3.3 Flowchart of the watchdog mechanism for real-time optimization control

In addition, a result validation mechanism is employed to minimize unnecessary control fluctuations and abnormal control decisions. When a new control setpoint, u_t, is generated, it is first compared against the previously implemented setpoint u_{t-1}. The update is only activated if the change exceeds a predefined threshold Δu_{min}. This logic is formalized as follows:

$$\Delta u = |u_t - u_{t-1}| \tag{3.2}$$

$$\mathrm{u}_{opt} = \begin{cases} u_t, & if\ \Delta u \geq \Delta u_{min} \\ u_{t-1}, & if\ \Delta u \leq \Delta u_{min} \end{cases} \tag{3.3}$$

This mechanism prevents frequent or insignificant adjustments that could lead to actuator wear or system instability. Next, this new control setpoint u_t will be subject to anomaly detection to determine whether it can be applied to the optimized operation of the chiller. As highlighted in Eqs. (3.4)- (3.6), a simple moving average (SMA) and corresponding standard deviation (SD) are computed over a rolling window of previously applied setpoints. The new optimized setpoint u_t is only accepted if it lies within a dynamically determined confidence interval, defined by $\pm K{\cdot}SD_t$ around the SMA:

$$SMA_t = \frac{1}{N}\sum\nolimits_{i=t-N+1}^{t} x_i \tag{3.4}$$

$$SD_t = \sqrt{\frac{1}{N}\sum\nolimits_{i=t-N+1}^{t}(x_i - SMA_t)^2} \tag{3.5}$$

$$\mathrm{u}_{opt} = \begin{cases} u_t, & if\ \Delta u \leq k \cdot SD_t \\ u_{t-1}, & if\ \Delta u \geq k \cdot SD_t \end{cases} \tag{3.6}$$

where, SMA_t represents the simple moving average at time t, N is the size of the moving average window, and x_i is the data point at position i. The SD_t indicates the standard deviation within the window, and k is the threshold width coefficient.

Stability and robustness assurance mechanisms are integrated between the smart control station and the Direct Digital Control (DDC) layer to ensure reliable operation of the AI-enhanced optimization algorithms. Specifically, the watchdog mechanism and the result validation mechanism are designed to monitor the optimization process and validate the results, respectively. If the optimization process fails to produce results within a predefined five-minute window, or if the generated setpoints exhibit abnormal deviations, the system automatically triggers a fallback protocol. Once the optimization process returns to normal operation, the station seamlessly transitions back to the AI-enhanced control strategy, ensuring stable and uninterrupted system operation.

Extended functionalities

To support the scalability and adaptability of optimization strategies in practical implementation, the proposed framework integrates a set of extended functionalities, with a particular emphasis on online learning and distributed multi-agent coordination—both of which are essential for ensuring long-term effectiveness and robustness in complex and dynamic building environments.

For online learning, the framework adopts a hybrid model-based architecture designed to balance generalization and adaptability. The model structure comprises two distinct parameter groups: a set of pre-trained shared layers, which encode general system knowledge obtained from offline training across various operating conditions, and a trainable update layer, which is specifically designed to be incrementally updated using real-time operational data from the deployed system. This structure allows the model to retain its general reasoning capability while continuously adapting to evolving local conditions, such as changes in occupancy patterns, equipment aging, or seasonal variations. The online learning process is governed by a generalized optimization rule:

$$\theta_N \leftarrow \theta_N - \eta \cdot u\left(\nabla_{\theta_N}\left[\frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i; \theta_S, \theta_N)\big) + R_{\theta_N}\right]\right) \tag{3.7}$$

where, θ_S and θ_N denote the shared and trainable parameters. L is the loss function. R represents regularization terms (e.g., L1/L2). u (·) defines the optimizer (e.g., SGD, Adam). The symbol η denotes the learning rate and the ∇_{θ_N} represents the direction and magnitude of the steepest increase in the loss landscape.

In addition, the framework supports distributed multi-agent coordination, enabling multiple smart control stations to manage large-scale or zoned building systems collaboratively. Each control station operates as an independent agent with its local sensing, optimization, and actuation capabilities, facilitating spatially distributed deployment. The proposed framework supports fully distributed optimization. Since each control station can independently execute AI algorithms, optimization tasks can be decomposed at the algorithmic level into subtasks processed in parallel by different agents. For example, in a genetic algorithm–based optimization, each agent can evaluate the fitness of different subsets of candidate solutions. This parallelization significantly reduces computation time and enables the system to scale efficiently as additional agents are integrated. To enable coordination without a central controller, the framework employs peer-to-peer communication via standard protocols such as UDP/IP or TCP/IP. Instead of transmitting raw data, agents exchange summarized status indicators or optimization intents—for instance, local operational states or estimated energy-saving benefits of current solutions. These exchanged signals accelerate the optimization convergence and enable the use of consensus algorithms to achieve globally coherent solutions. All AI enhancements are implemented in a non-disruptive manner. The control stations interface seamlessly with existing BAS infrastructure, requiring no modifications to legacy devices or network configurations. This design philosophy ensures backward compatibility and supports incremental integration, allowing building operators to gradually adopt AI-based control without disrupting ongoing operations or compromising system reliability.

3.2 Description of the AI-enabled smart control stations

The proposed AI-enabled control station can be flexibly constructed using embedded devices or industrial-grade edge computing platforms, depending on the application scale and computational requirements. In this study, two types of smart control stations are deployed to verify communication, control, and real-time optimization functionalities.

The first control station is developed based on a Raspberry Pi, which is selected for integration with a BMS testbed to validate its communication and control capabilities. Commonly used in education and prototyping, the Raspberry Pi features a quad-core ARM Cortex-A72 processor and supports up to 8GB of LPDDR4 RAM, which is sufficient to meet the testing requirements of this study. As illustrated in Figure 3.4, the station is assembled with a touchscreen interface, cooling fans, and a protective enclosure, ensuring reliable operation and portability in experimental test environments.

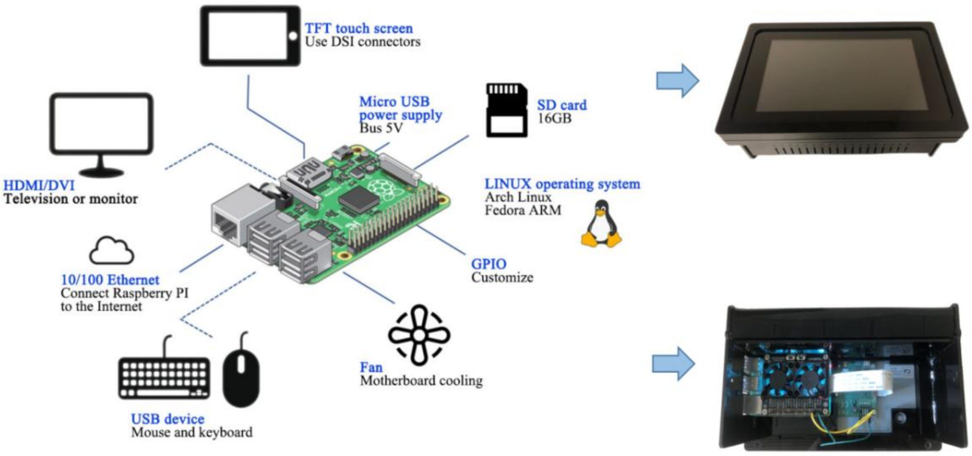


Figure 3.4 Function composition and appearance of Raspberry Pi-based control station

The second control station adopts the NVIDIA Jetson Orin™ edge computing platform for hardware-in-the-loop (HIL) simulation with a virtual dynamic cooling system. While the Raspberry Pi—initially designed for education, prototyping, and personal projects—offers cost-effectiveness and flexibility, it may encounter processing limitations when applied to large-scale or computationally intensive systems. In contrast, the Jetson Orin™ platform, as shown in Figure 3.5, features a 1024-core GPU based on the NVIDIA Ampere architecture and a 6-core Arm® Cortex®-A78AE CPU, delivering up to 40 TOPS of AI computing power. This industrial-grade configuration provides superior computational efficiency and robustness, making it well-suited for real-time optimization tasks in complex, large-scale HVAC systems.

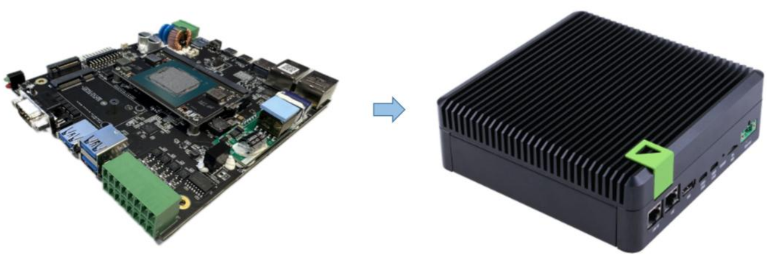


Figure 3.5 Appearance of the control station based on the Jetson Orin™ platform

3.3 Test scenarios and arrangement

3.3.1 Basic function test on BMS testbed

A testbed is constructed using selected IoT components to verify the basic communication and control functions of the proposed control station. As shown in Figure 3.6, the setup includes a global network controller (BR50), a programmable logic controller (BCU-843), a valve driver (A9116), and a temperature and humidity transmitter (RH-5I-R-N10). The specifications of each device are summarized in Table 3.1.

In this test, the control station generates random control commands, which are transmitted to the BR50 controller via the BACnet/IP protocol. The BR50 then relays the commands to the BCU for execution. The valve driver acts as the actuator, while the transmitter provides real-time indoor temperature readings. These sensor readings are displayed on the control station interface and stored in a local database for post-analysis. This experiment replicates a simplified building automation scenario, validating the feasibility of deploying the proposed AI-enabled control framework within a live environment.

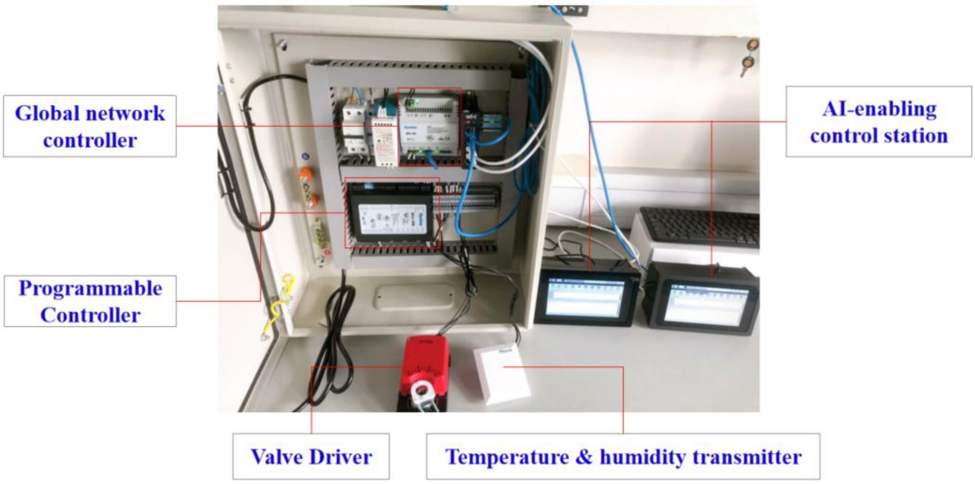


Figure 3.6 Image of the smart station integrated within the experimental platform

Table 3.1 The parameters of each component in the BMS testbed

Equipment	Model	Parameter
Global network controller	BR-50	10M Ethernet and MS/TP bus
Programmable Controller	BCU-843	Eight general inputs, four digital outputs, three analogue outputs
Valve Driver	A9116	Control mode: 0-10V / 4-20mA analog control
Temperature transmitter	RH-5I-R-N10	Accuracy: 0.5°C; Range: 0-50°C

3.3.2 Performance test on a hardware-in-the-loop test rig

To evaluate real-time optimization performance, a hardware-in-the-loop test rig is constructed, as illustrated in Figure 3.7. This platform connects an NVIDIA Jetson-based control station to a high-fidelity virtual cooling system developed in Dymola. The model, packaged as a Functional Mock-up Interface (FMI), runs in real time on a host computer and simulates the dynamic behavior of a large-scale, seawater-cooled cooling system.

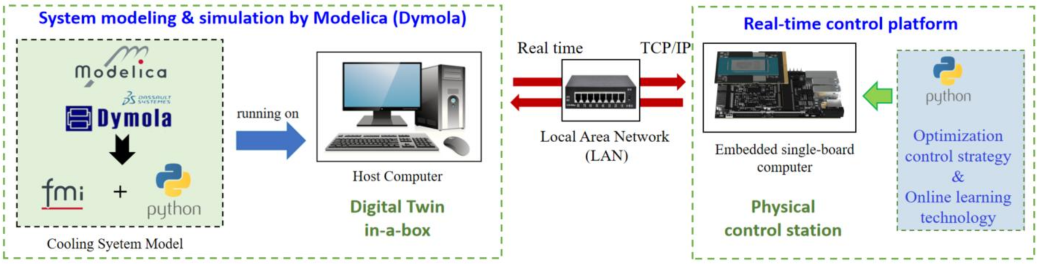


Figure 3.7 Outline of the hardware-in-the-loop test

The cooling system model is based on the actual configuration of a commercial building located in West Kowloon, Hong Kong. The cooling system includes a total of 21 chillers (15 × 3517 kW and 6 × 1758 kW), designed to meet a total cooling load of 55,331 kW. The “Buildings” library of Dymola is used to model all system components, including heat exchangers, chillers, and pumps. Outdoor weather conditions and building cooling load profiles are used as dynamic inputs to simulate realistic operational scenarios.

For the optimization of the seawater-cooled cooling system, the specific objective function is formulated by Eq. (3.1). In this context, the overall optimization objective *J(x)* aims to minimize the total power consumption of the system while incorporating a penalty term related to thermal comfort. The control variables u u include the chilled water supply temperature (T_{cws}), the chilled water return temperature (T_{cwr}), and the chilled water supply temperature at the heat exchanger outlet (T_{chws}). As shown in Eq. (3.8), the first part of the objective function represents the sum of power consumption, whereas the second part penalizes deviations of the chilled water supply temperature from the desired setpoint to ensure occupant comfort.

$$\min_{T_{cws},T_{cwr},T_{chws}} J(x) = (W_{sw} + W_{cw} + W_{ch} + W_{chw}) + k \cdot \left(T_{chws} - T_{chws,set}\right)^2 \quad (3.8)$$

where, W_{sw}, W_{cw}, W_{ch}, and W_{chw} represent the power consumptions of the seawater pump, condenser water pump, chiller unit, and chilled water pump, respectively. The k is the penalty factor.

The smart control station runs an AI-enhanced Genetic Algorithm (GA) in Python and communicates with the virtual cooling system via TCP/IP over a local area network. In this AI-enhanced optimization algorithm, a hybrid model is incorporated as the fitness evaluation function, in which sub-models exhibiting complex and highly nonlinear behaviors are represented by machine learning models. This approach reduces the high computational cost of evaluating individual fitness and accelerates the overall optimization process. By adjusting setpoints in real time and receiving system feedback, this setup enables comprehensive testing of the proposed framework and strategies under practical operating conditions. To evaluate optimization performance, two control strategies are implemented and compared:

- Baseline control strategy: A widely used approach in practice that operates the system with fixed setpoints.
- Real-time optimization deployed using the proposed framework: The smart control station executes an AI-enhanced Genetic Algorithm (GA) to dynamically optimize the cooling water supply and return temperature setpoints based on real-time measurements.

In addition to performance evaluation, the stability and robustness assurance mechanisms adopted in the proposed framework are also tested to validate their practical effectiveness. Specifically, these mechanisms were individually assessed under controlled disturbance scenarios: (1) a watchdog mechanism to enforce latency constraints, and (2) a result validation mechanism to ensure reliable setpoint updates. The watchdog mechanism is tested by intentionally inducing delays and failures in the optimization process, including forced computation stalls and exception triggers. These scenarios simulate real-world conditions such as algorithmic overload or system interruptions. The objective is to verify whether the system could detect timeout events and autonomously activate a fallback protocol to maintain uninterrupted operation using previously applied setpoints. The result validation mechanism is evaluated by injecting fluctuating and abnormal setpoints into the optimization process. This includes scenarios involving frequent minor updates and artificially inserted outliers. The effectiveness of the context-aware filtering strategy is confirmed by monitoring whether unnecessary or extreme control updates are successfully identified and discarded.

3.4 Test results and discussion

3.4.1 Functionality validation

The functionality of the proposed framework and strategies is first verified using a simplified BMS testbed comprising essential field-level components. During the test period from January 18th to February 27th, the control station successfully generated and transmitted control signals to the global network controller (BR50) via the BACnet/IP protocol. These commands were subsequently executed by the downstream BCU controller, triggering appropriate actuation of the valve driver. Real-time status updates from the actuator were accurately relayed back to the control station, completing the feedback loop. Simultaneously, continuous temperature measurements were collected from the transmitter and displayed on the station interface, while the data was logged in a local SQL-based database for post-processing and traceability.

As shown in Figure 3.8, the temperature data exhibit stable temporal continuity, with no data loss or communication failures observed over the six-week operation window. The success of this test confirms that the proposed control station can reliably perform core BAS functions, including bidirectional communication, actuator control, sensor data acquisition, and local data storage.

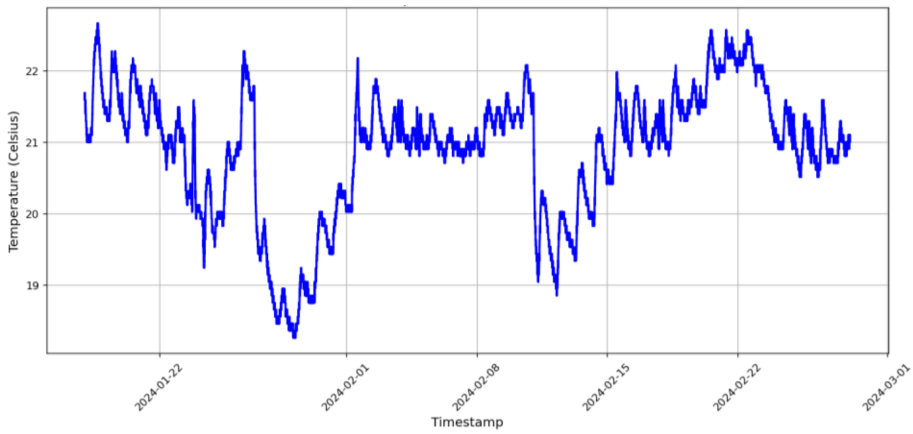


Figure 3.8 Time-series temperature data from sensor measurements

3.4.2 Energy performance and robustness of online optimization

The energy-saving performance of the proposed framework and strategies is evaluated through hardware-in-the-loop (HIL) tests. As illustrated in Figure 3.9, the cooling water setpoint trajectories under two control strategies are compared: one using the proposed AI-enabled optimization strategy and the other using a fixed setpoint baseline strategy. Compared with the baseline, the average return temperature of the cooling water and the supply temperature under the optimized control strategy were 2.30 K higher and 1.93 K lower, respectively. These adjustments resulted in a reduction in the cooling water flow rate but a corresponding increase in the seawater flow rate.

This optimization can be attributed to the system's specific characteristics: the cooling water pumps, due to long-distance piping, consume significantly more energy than the seawater pumps. Therefore, the optimization algorithm prioritizes reducing the energy consumption of the cooling water pumps, even at the expense of a slight increase in the energy usage of the seawater pumps.

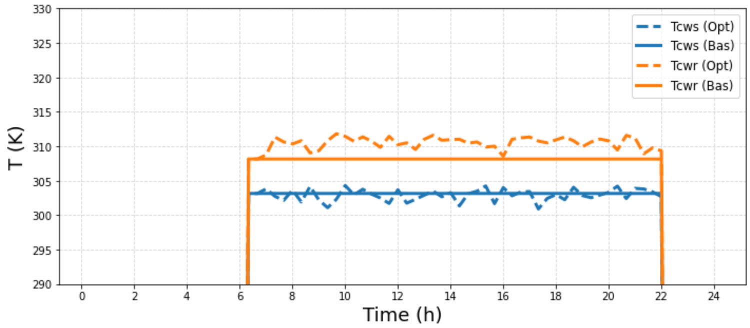


Figure 3.9 Temperature setpoints under two control strategies in a typical day

Figure 3.10 shows the total energy consumption recorded in the tests. By continuously balancing the energy usage of the seawater pumps, cooling water pumps, and chillers, the AI-enabled optimization strategy achieved a total energy saving of 7.66% compared to the baseline control strategy. This verifies the effectiveness of the proposed framework and strategies in improving system-level efficiency through adaptive and data-driven control.

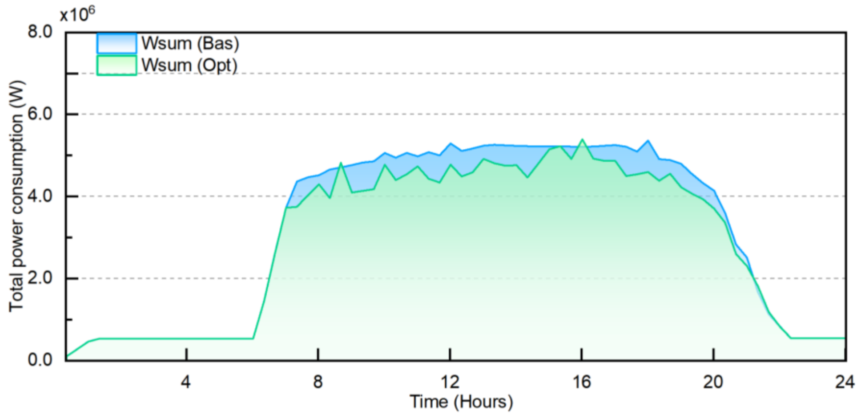
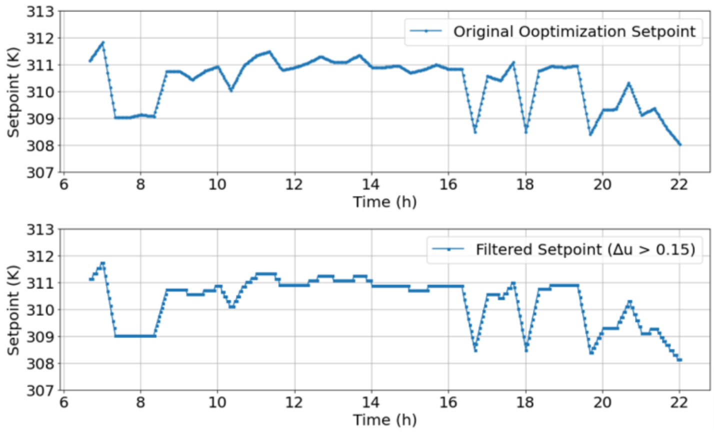
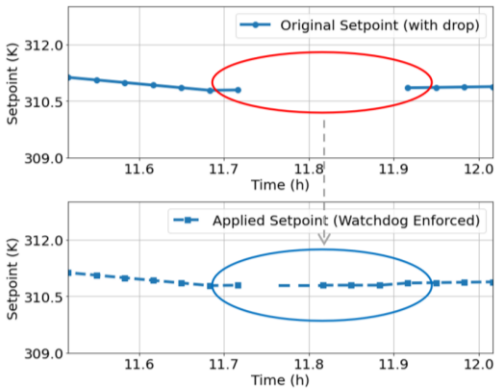


Figure 3.10 Total power consumption under two control strategies in a typical day

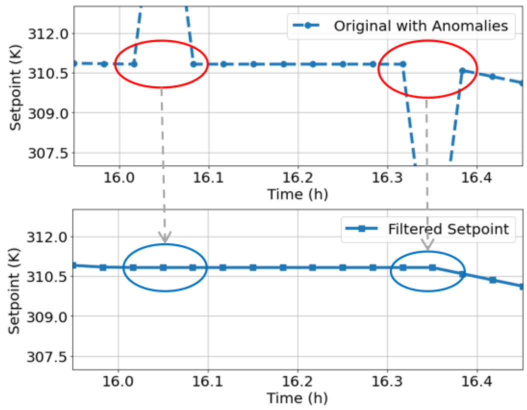
Figure 3.11 presents the test results of the stability and robustness mechanisms integrated into the proposed framework. Figure 3.11 (a) shows the effectiveness of the result validation mechanism in filtering out minor setpoint fluctuations. A threshold of $\Delta u > 0.15$ K was applied to suppress frequent but insignificant control updates that could otherwise trigger unnecessary actuator actions and lead to system instability. As a result, the refined setpoints exhibit a smoother profile with timely updates, indicating enhanced stability in control dispatching without compromising responsiveness. Figure 3.11 (b) shows the response of the watchdog to optimization interruptions. In this test, intentional computation stalls were introduced at specific intervals, preventing the generation of new setpoints. When the delay exceeded the predefined timeout of 300 seconds, the watchdog mechanism triggered the fallback protocol, reverting the system to the last valid setpoint and ensuring continuous operation. Figure 3.11 (c) demonstrates the result validation mechanism under more aggressive anomaly conditions. Sharp spikes and drops were artificially injected into the setpoint sequence to simulate abnormal control outputs. By applying a filter based on moving averages and standard deviations, the system effectively identified and rejected these outliers, maintaining previously validated setpoints and ensuring stable control behavior.



(a) Filtering setpoints via result validation mechanism



(b) Fallback enforcement via watchdog mechanism



(c) Anomaly detection and filtering via result validation mechanism

Figure 3.11 Verification of stability and robustness mechanisms in the framework

3.5 Summary

This paper presents a generic framework for embedding AI-based online optimization into BAS by deploying AI-enabling smart control stations at the field level. The framework comprises two core functional modules: an AI runtime environment for AI model inference and learning, and a comprehensive suite of AI functional components for data exchange, control robustness, and extended functionalities. Based on the tests and experimental validation, the following conclusions are drawn:

- Tests on both the BMS testbed and a hardware-in-the-loop test rig verify the proposed framework. Results confirm that the field smart control stations can reliably host and execute AI-enhanced optimization algorithms at the edge, while maintaining compatibility with the BAS environment.
- Compared with the baseline control strategy, the proposed framework adopts smart control stations for real-time optimization and achieves a significant energy saving of 7.66% through the deployment of an AI-enhanced Genetic Algorithm.

- The integrated watchdog and result validation mechanisms show strong capabilities in ensuring system robustness and stability under various disturbance scenarios. The watchdog module can effectively monitor execution latency and trigger fallback control upon detecting computation failures. The result validation mechanism can perform output filtering and outlier suppression efficiently, thereby maintaining stable and reliable control performance.

This framework provides a practical, scalable, and non-intrusive pathway for embedding AI-enhanced optimization into real-world applications at the field level. Future research will focus on advancing multi-agent coordination, enabling large-scale distributed control, and enhancing online learning adaptability, thereby supporting intelligent and resilient energy management by adopting AI at the edge in increasingly complex built environments.

CHAPTER 4 AI-EMPOWERED ONLINE CONTROL OPTIMIZATION FOR ENHANCED EFFICIENCY AND ROBUSTNESS

This chapter presents the development and implementation of an AI-empowered online control optimization method aimed at enhancing both the efficiency and operational robustness of building central cooling systems. Motivated by the increasing demand for intelligent energy management and the challenges associated with applying advanced optimization algorithms in real-world scenarios, this study introduces a deep learning-enabled genetic algorithm designed to achieve rapid optimization with reduced computational complexity. Furthermore, a robust assurance scheme is integrated to ensure stable operation under abnormal conditions by enabling seamless transitions between AI-driven control and an expert knowledge-based fallback scheme. Through systematic algorithm design, model development, and hardware-in-the-loop testing, this chapter illustrates how the proposed approach can significantly improve energy efficiency while maintaining reliable and practical online control performance. The following sections elaborate on the conceptual framework, technical formulations, system integration, and experimental validations supporting the proposed AI-empowered optimization methodology.

4.1 Proposed AI-empowered control optimization method

4.1.1 Overview of the control method

Figure 4.1 illustrates the generation and control process of the proposed AI-empowered optimization method, designed to reduce computational complexity and enhance application robustness. The core concept involves simplifying the AI-based optimization algorithm to enhance its suitability for field applications on miniaturized stations and integrating an anomaly detection mechanism to maintain the reliability of the optimization process.

The optimization algorithm is a deep learning-enabled genetic algorithm, in which the fitness function's optimization model is trained as a hybrid model to achieve a balance between efficiency and reliability. A physical model developed in Dymola is utilized for comprehensive simulation, generating a complete dataset of system operation. Following data processing, sensitivity analysis, and feature selection, the hybrid model is established and integrated into the genetic algorithm (GA) for optimization calculations. To further enhance system robustness, an anomaly detection mechanism based on a dynamic threshold detection method is incorporated, forming a robust assurance scheme. This mechanism enables automatic switching to an expert knowledge-based scheme whenever anomalies are detected, ensuring continuous and stable operation.

Under conventional control, the Direct Digital Control (DDC) controller follows a fixed setpoint method, continuously acquiring cooling system parameters and adjusting pump and chiller operations via PID feedback control. In the optimal control mode, the AI-empowered optimization algorithm is deployed on a control station, where it processes real-time operational data to generate optimized setpoints and transmits them to the DDC controller. The anomaly detection mechanism is embedded within the controller to validate the received setpoints. If no anomalies are detected, the controller updates the setpoints accordingly; otherwise, it switches to a predefined robust assurance scheme to maintain system stability.

The detailed working mechanism of the deep learning-enabled genetic algorithm and the robust assurance scheme is elaborated in Section 4.1.2.

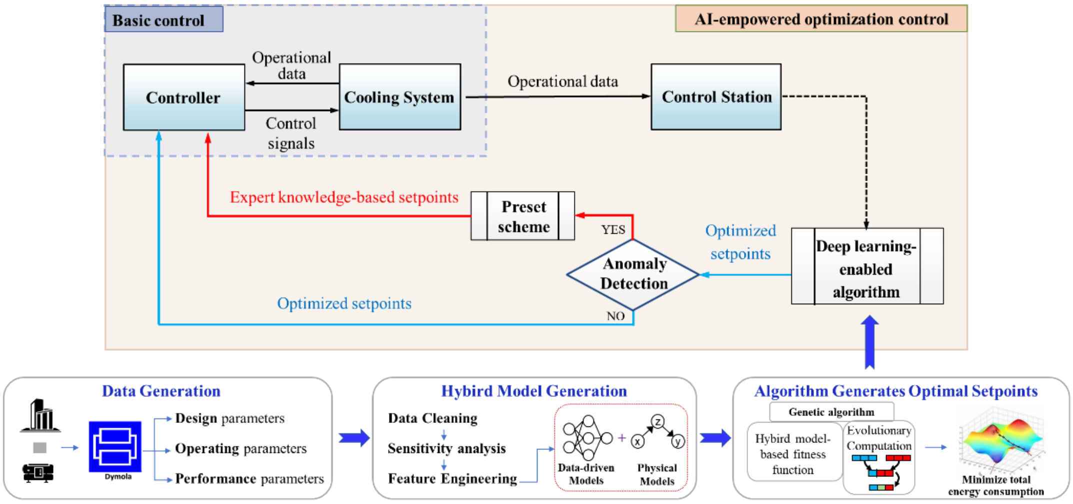


Figure 4.1 Generation and control process of the AI-empowered optimization method

4.1.2 Formulation of the AI-empowered optimization method

4.1.2.1 Deep learning-enabled genetic algorithm for online control

One of the main challenges in applying advanced optimization algorithms in the field is meeting computational demands. Traditional optimization algorithms, such as genetic algorithms, typically find optimal results by evaluating the effectiveness of various schemes through repeated assessments and selections. This process consumes significant computing resources and time, making it less practical for real-time applications.

The deep learning-enabled genetic algorithm proposed in this paper is a hybrid model-based approach in where key nonlinear models of the cooling system are trained as Artificial Neural Network (ANN) models. While pure black-box models can significantly enhance computational efficiency, their lack of interpretability can limit the reliability of applications. The hybrid model, which retains core physical laws, effectively improves adaptability during implementation. The flowchart of the deep learning-enabled genetic algorithm for optimal control is depicted in Figure 4.2. Initially, the control station receives real-time operational parameters from the cooling system and generates an initial population, where each individual represents a different setpoint combination. These initial solutions are then fed into the hybrid model, which evaluates the fitness score for each offspring based on energy consumption. The best candidates are selected for evolutionary operations—selection, crossover, and mutation. Finally, the algorithm determines the optimal setpoint combination, enabling the cooling system to operate efficiently. Additionally, the genetic algorithm used in this study combines roulette wheel selection and elitism retention methods to balance exploration and exploitation, ensuring faster convergence and stable optimization results. Through extensive testing, the following key parameters were selected to ensure efficient convergence of the algorithm: (1) The initial population size is 20; (2) The number of generations is 16; (3) The crossover rate is 0.75; (4) The mutation rate is 0.05.

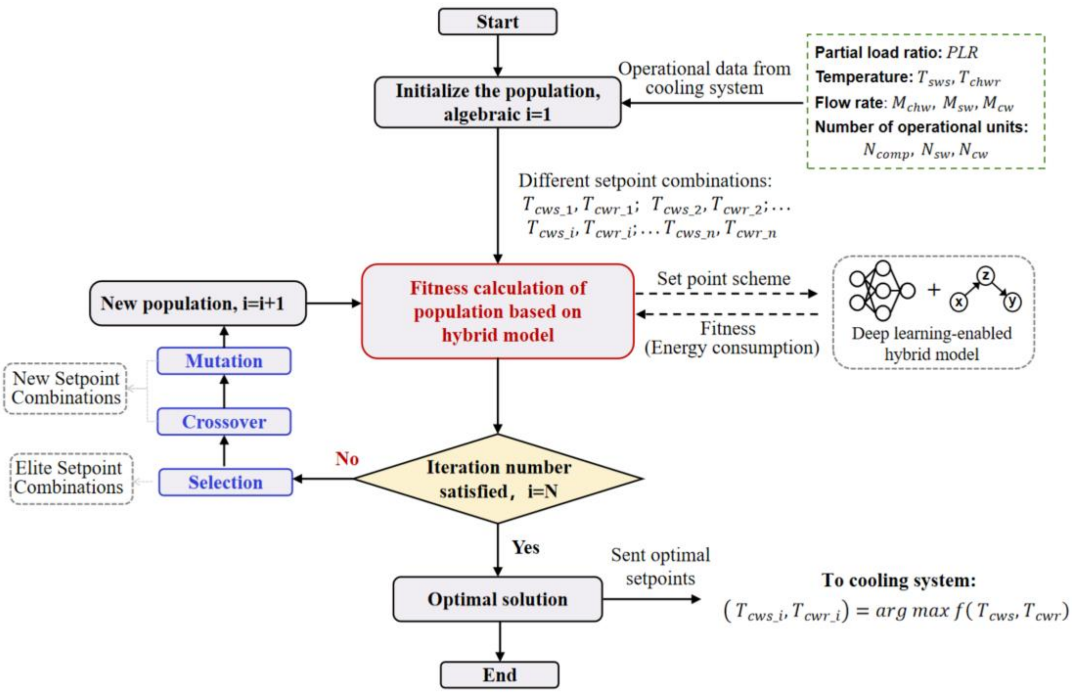


Figure 4.2 Flowchart of the Deep Learning-Enabled Genetic Algorithm

Through sensitivity analysis, it was found that the power consumption of the chiller is highly correlated with the flow rates of seawater and cooling water, while the corresponding control temperature significantly influences the power consumption of the pump. Based on these findings, this study selects the cooling water supply temperature (T_{cws}) and the cooling water return temperature (T_{cwr}) as the decision variables for the genetic algorithm. The fitness calculation formula is defined in Eq. (4.2), where ϵ is a small constant to prevent division by zero. The total system energy consumption, denoted as W_{sum}, primarily consists of the energy consumption of the seawater pump (W_{sw}), cooling water pump (W_{cw}), and chiller unit (W_{comp}). To drive the optimization process toward energy efficiency, the reciprocal of W_{sw} is incorporated into the fitness function, ensuring that higher fitness values correspond to lower energy consumption. During the evolution process of the genetic algorithm, selection, crossover, and mutation operations are iteratively performed, evaluating the fitness of each individual (setpoint combination).

Additionally, a penalty function for the chilled water supply temperature is introduced into Eq. (4.2) to ensure that the optimization results align with the building's operational requirements. In this function, k indicates the penalty factor, T is the temperature, and the subscripts chw and set represent the chilled water and set value. In this study, the identification of the penalty factor k is based on repeated tuning through simulation tests. The initial range of k was discretized into a grid, and tests were conducted under various load and external environmental conditions. The effect of different k values on the chilled water supply temperature (T_{chws}) was carefully observed. When the setpoint $T_{chws,set}$ was set to 280.15 K, a value of k =5×10^{-8} was chosen to ensure the chilled water supply temperature does not exceed 281.15 K.

$$W_{sum} = W_{sw} + W_{cw} + W_{comp} \tag{4.1}$$

$$Fitness = \frac{1}{W_{sum} + \epsilon} - k \cdot \left| T_{chws} - T_{chws,set} \right| \tag{4.2}$$

4.1.2.2 Robust control scheme integrating expert knowledge

The robust assurance scheme is designed between the control station and the Direct Digital Control (DDC) controller to enhance system reliability. As depicted in Figure 4.1, the optimal setpoints generated by the AI-empowered algorithm are sent to a detection node at the controller, where their validity is assessed. A dynamic threshold detection method, based on a simple moving average (SMA), is introduced to evaluate the reliability of the control station's outputs.

As highlighted in Eqs. (4.3)- (4.6), the detection method involves the calculation of the moving average and standard deviation (SD) of a set of temperature points over a specified historical data window. The system identifies abnormal deviations in the optimization setpoints by establishing dynamic thresholds at two SDs above and below the moving average. Such deviations may indicate potential algorithmic failures or inefficiencies, triggering an automatic switch to a preset scheme based on expert knowledge. In addition, if the DDC controller does not receive an updated setpoint from the control station within 20 minutes, a system alert is triggered, indicating a possible failure in communication. In such cases, the system automatically switches to a robust assurance scheme to maintain stable operation. Once the controller resumes receiving valid setpoints, the system switches back to the optimal control scheme. This robust assurance mechanism ensures proper system operation and prevents potential operational failures.

$$SMA_t = \frac{1}{N} \sum_{i=t-N+1}^{t} x_i \tag{4.3}$$

$$Upper\ Threshold_t = SMA_t + k \times SD_t \tag{4.4}$$

$$Lower\ Threshold_t = SMA_t - k \times SD_t \tag{4.5}$$

$$SD_t = \sqrt{\frac{\sum_{i=t-N+1}^{t}(x_i - \bar{x})^2}{N-1}} \tag{4.6}$$

where, SMA_t represents the simple moving average at time t, N is the size of the moving average window, and x_i is the data point at position i. The SD_t indicates the standard deviation within the window, k is the threshold width coefficient, and $\bar{x}$ is the mean of the data points within the window.

In addition, an empirical fitting formula based on expert knowledge is introduced to provide set points under abnormal conditions. According to theoretical research and experiments conducted by a chiller manufacturer, the cooling water temperature optimization set point is influenced by several factors, including the outdoor wet-bulb temperature (T_{wb}) and its design value, the partial load ratio of the chiller system (PLR), and the flow rate of the cooling water (m_{cw}):

$$T_{cwr,set} = A \times T_{wb} + B \times PLR - C \times T_{wb,set} - D \times m_{cw}/Load + 37 \tag{4.7}$$

4.2 Model development and description

4.2.1 Reference building and cooling system

A commercial building located in West Kowloon, Hong Kong, is selected as the reference building for optimization tests. As shown in Figure 4.3**,** the building comprises four towers (T1A, T1B, T2A, T2B) and a podium, with a total area of 288,010 m^2. The total design cooling load is 55,331 kW. To satisfy this cooling demand, the building is equipped with 21 seawater-cooled chillers, including 15 chillers rated at 3517 kW and 6 chillers rated at 1758 kW, as summarized in Table 4.1.

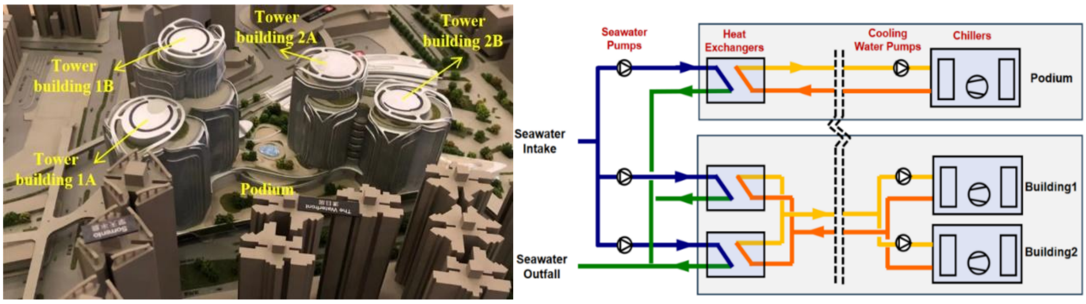


Figure 4.3 Overview of the buildings and the central cooling system layout

Table 4.1 Key building parameters and cooling system specifications

Building —	Tower 1A	Tower 1B	Tower 2A	Tower 2B	Podium
Gross Floor Area (m^2)	58770	77000	52490	43750	56000
Air Conditioning Area (m^2)	48528	60351	43751	36694	13156 (Retail) 11744 (F&B) 13687 (Arcade)
Block load (kW)	10544	12960	9415	7906	4526, 4846, 5135
Load ratio	0.19	0.23	0.17	0.14	0.26
Chiller plants	3*3517kW +1*1758kW	3*3517kW +1*1758kW	2*3517kW +2*1758kW	2*3517kW +1*1758kW	5*3517kW +1*1758kW

The central cooling systems described above are simulated in Dymola, a software platform based on the Modelica language (Fritzson & Engelson, 1998), which is capable of simulating and analyzing complex dynamic systems with high accuracy and flexibility. As shown in Figure 4.4, the system structure is organized from left to right, representing the seawater side, cooling water circulation, chiller units, and load side, respectively. Additionally, from top to bottom, the structure follows the hierarchy of the Podium, Building 1, and Building 2. The models for components, such as pumps, heat exchangers, and chillers, are developed using the "Buildings" library. Basic control logic, including PID feedback control and device start-stop control, is incorporated into the model to simulate fundamental operational behaviors. During the simulation, external environmental parameters, including seawater temperature and building load variations, are input into the seawater side and load side of the

Modelica model for dynamic execution. The load profiles are collected from a large existing building by the same developer nearby. The seawater temperature is based on the Hong Kong Environmental Protection Department (EPD, 2023).

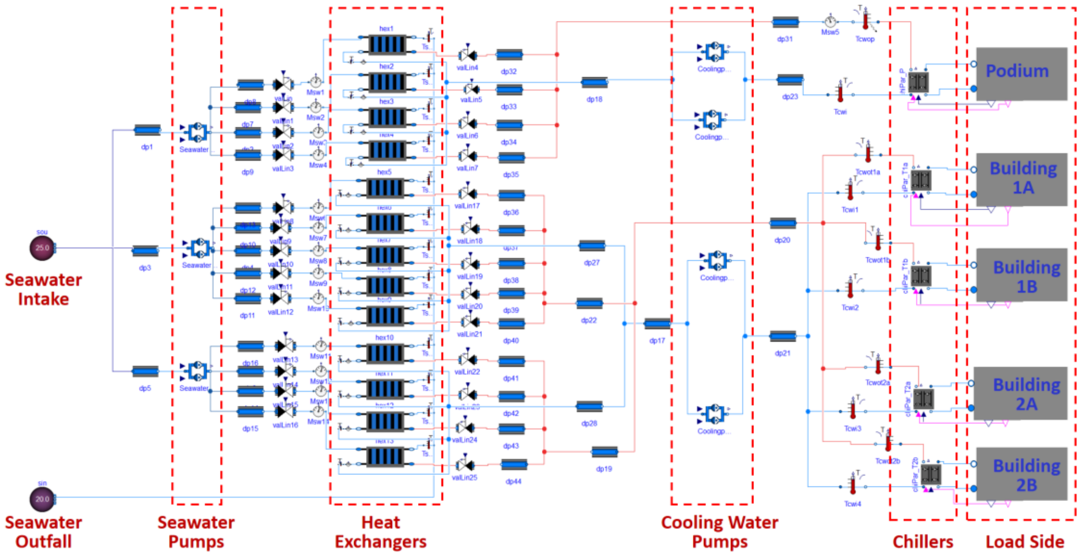


Figure 4.4 Seawater-cooled central cooling system in Dymola

4.2.2 *Hybrid energy models of cooling systems for online optimization*

A hybrid energy model, serving as the fitness function for the genetic algorithm (GA) in online optimization, enables efficient computation of the total power consumption of a central cooling system. Figure 4.5 illustrates the structure of this hybrid model and its function in the optimization process. This model is primarily responsible for fitness evaluation in the genetic algorithm optimization, generating fitness scores correlated with total system energy consumption to determine the optimal offspring. In this hybrid model, Artificial Neural Network (ANN) models are used for chillers and heat exchangers due to their significant nonlinearity and complex parameter settings. The cooling water pump and seawater pump models adopt polynomial fitting derived from their design parameters, as described in Eqs. (4.8) and (4.9). Each component in the models is interconnected according to its physical connections and physical laws, ensuring reliability and interpretability during application.

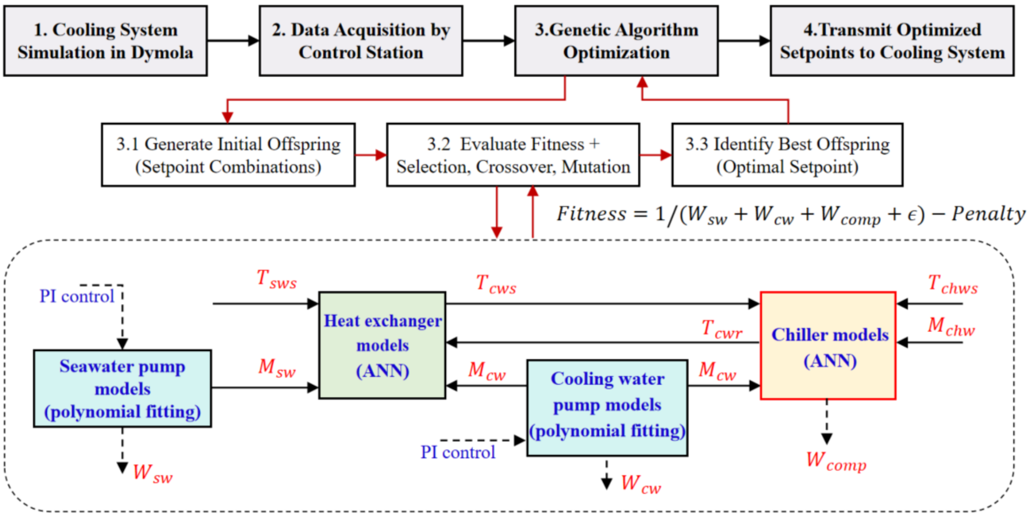


Figure 4.5 Hybrid model structure and its function in the optimization process

$$W_{sw} = \pi_1 + \pi_2\left(\frac{M_{sw}}{M_{sw,nom}}\right) + \pi_3\left(\frac{M_{sw}}{M_{sw,nom}}\right)^2 + \pi_4\left(\frac{M_{sw}}{M_{sw,nom}}\right)^3 \tag{4.8}$$

$$W_{cw} = \mu_1 + \mu_2\left(\frac{M_{cw}}{M_{cw,nom}}\right) + \mu_3\left(\frac{M_{cw}}{M_{cw,nom}}\right)^2 + \mu_4\left(\frac{M_{cw}}{M_{cw,nom}}\right)^3 \tag{4.9}$$

where, W and M are the energy consumption and flow rate, respectively. The subscript *nom* represents the nominal condition. The parameters π_1~π_4 and μ_1~μ_4 can be obtained from manufacturers or by curve-fitting using operational data.

The details of the ANN models for the heat exchangers and chillers are illustrated in Figure 4.6. Both models utilize a two-hidden-layer structure, with each hidden layer consisting of ten nodes. For the heat exchanger model, there are four input parameters: seawater flow rate (M_{sw}), seawater supply temperature (T_{sws}), cooling water flow rate (M_{cw}), and cooling water return temperature (T_{cwr}). The output of this model is the cooling water supply temperature (T_{cws}), which is crucial for coupling with the chiller unit. The chiller model uses five input parameters: cooling water supply temperature (T_{cws}), cooling water flow (M_{cw}), chilled water supply temperature (T_{chws}), chilled water flow (M_{chw}), and the partial load ratio (*PLR*). The outputs of the chiller model include the coefficient of performance (*COP*) of the unit and the cooling capacity (Q_e), which can be further used to calculate the chiller's power consumption. The data for building the ANN models is derived from comprehensive full-operation simulations of the Modelica model, covering all reasonable setpoint combinations and environmental parameters under various system operating conditions. A total of 60,000 valid data points are generated for machine learning. This simulation approach aims to provide the ANN model with a diverse and representative dataset, enabling it to accurately simulate the behavior of the heat exchanger and chiller under various conditions.

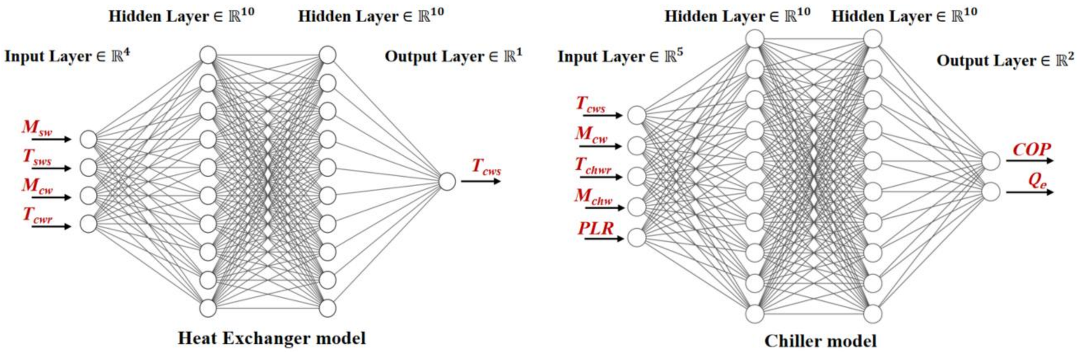


Figure 4.6 Overview of ANN model configurations, inputs, and outputs

For model validation, 4,000 sets of data, which are different from the data used for model training, were randomly generated by the Modelica models (shown in Figure 4.4). These validation datasets were used to assess the performance of the trained ANN models. The validation results are shown in Figure 4.7. The output of the heat exchanger model is a temperature value, which exhibits a relatively small fluctuation range and a strong logical correlation with the input parameters. As a result, the average relative error of the T_{cws} is 0.01%. For the chiller model, the average relative errors of the output values for the *COP* and Q_e are 0.29% and 0.18%, respectively. These results demonstrate that the ANN models trained in this study meet the accuracy requirements.

However, in real-world applications, challenges such as sensor errors, noise, and other system biases are likely to arise. In such cases, the ANN, as a data-driven model, can easily adapt and correct itself through online learning and self-updating, thereby enhancing its robustness to noise and system deviations. Furthermore, in this study, the combination of physical models and ANN models significantly improves the system's interpretability and reliability. By integrating the physical principles of the system with the data-driven learning capabilities of the ANN, the hybrid model-based approach offers both efficient computation and stable operation.

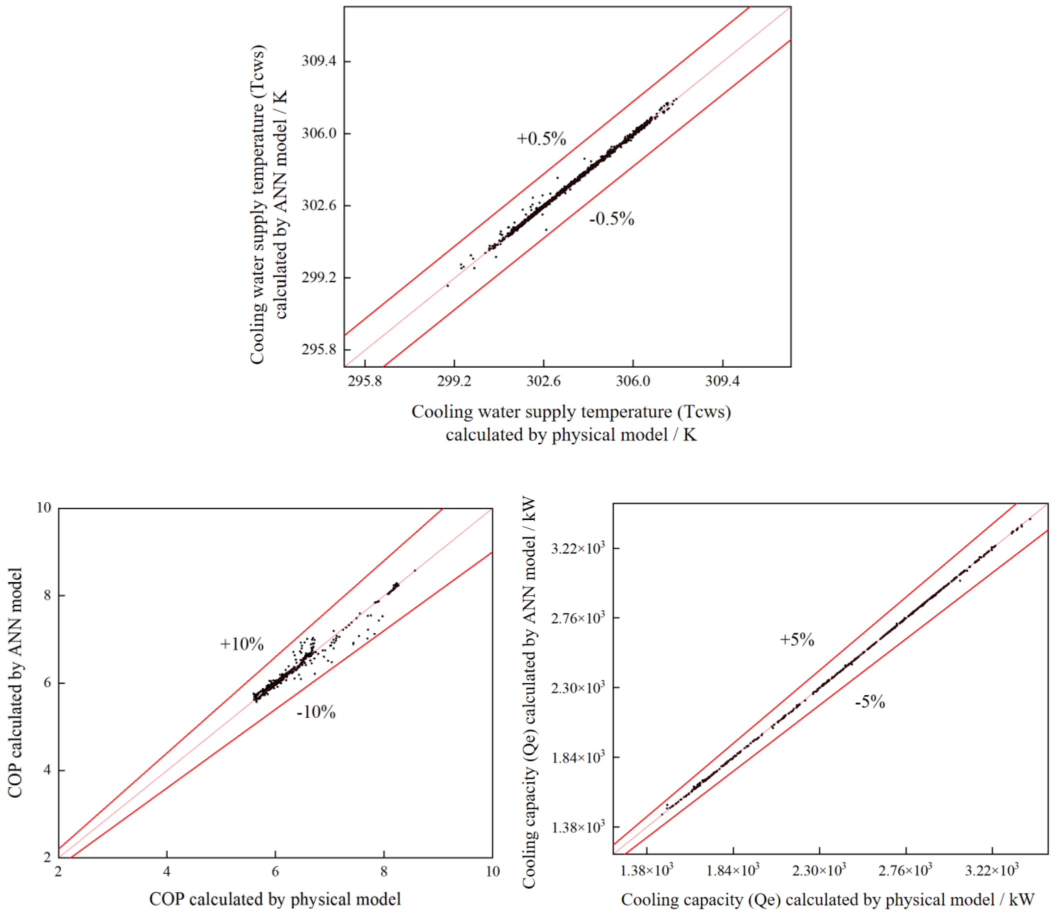


Figure 4.7 Comparison between the data and the model prediction

4.3 Test platform and validation arrangement

To rigorously evaluate the proposed optimization method under realistic dynamic conditions, a hardware-in-the-loop (HIL) simulation platform is employed. This approach integrates physical control hardware with a high-fidelity virtual environment, enabling real-time assessment of algorithm performance, communication protocols, and control responsiveness. The following section outlines the experimental setup, building upon the control stations and test platforms introduced in Sections 3.2 and 3.3.

4.3.1 Description of the smart station and test platform

The Raspberry Pi-based smart control station, previously described in Section 3.2, is utilized here for low-cost and flexible deployment in the HIL setup. Figure 4.8 presents both the schematic diagram and the implementation diagram of the hardware-in-the-loop test platform. The physical smart station, equipped with the deep learning-enabled genetic algorithm, controls a simulated, real-time virtual dynamic cooling system. The cooling system model was developed in Dymola using the Modelica language and packaged as a Functional Mock-up Interface (FMI) file, serving as a "digital twin in a box" hosted on the computer. Within the communication framework, the TCP/IP protocol is employed over the local area network, simulating the real-time data transmission process during optimization.

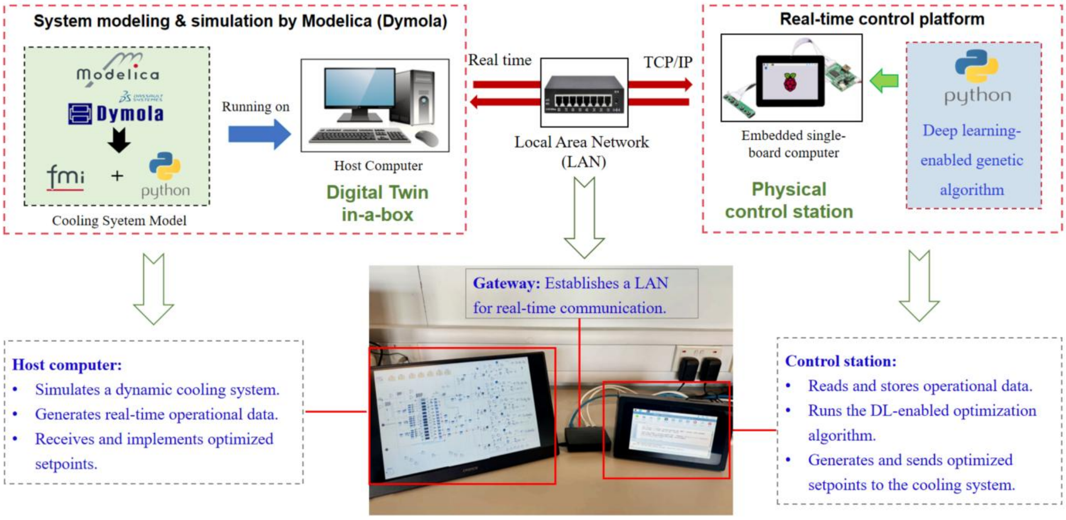


Figure 4.8 Hardware-in-the-loop test platform: schematic and implementation

4.3.2 The test arrangement

The proposed AI-empowered optimization method is deployed on the smart station and evaluated using the hardware-in-the-loop test platform. Real-time validation tests are conducted over a day using weather data representative of a typical summer day in Hong Kong. The analysis begins with an evaluation of computation time per iteration, comparing the results before and after integrating the hybrid model to quantify improvements in computational efficiency.

Subsequently, comparative studies are conducted to evaluate the performance of the proposed AI-empowered optimization method against the traditional control method.

- *Baseline control method:* This is a commonly used method for chiller systems that does not involve optimization. Fixed setpoints for the cooling water supply temperature and return water temperature are maintained throughout the operation.
- *AI-empowered optimization control method:* A deep learning-enabled genetic algorithm is used to optimize the cooling water supply and return temperatures at fixed intervals, improving the system's efficiency.

During the tests, an optimization process within each control interval (5 minutes in this study) is divided into four distinct steps:

- Sampling: collecting real-time system data,
- Calculation: performing the optimization computation,
- Sending: sending optimized setpoints to the system,
- Waiting: waiting until the subsequent optimization interval.

To validate the robustness assurance scheme, three hours during a typical day are selected for the validation test. The controller is programmed to send a set of optimal set points every optimization interval, allowing for the evaluation of the Direct Digital Control (DDC) in responding to normal reception, data anomalies, and interruptions.

4.4 Test results and discussion

4.4.1 Optimization efficiency

The number of floating-point operations per second (FLOPs), serving as a standard measure to evaluate the processing power of a computing device, is used to quantify the performance of the physical control station. The theoretical value of FLOPs can be calculated using: FLOPs ≈ Clock Frequency (GHz) × Floating-point Operations per Cycle.

The smart station used in the hardware-in-the-loop test is equipped with a quad-core ARM Cortex-A72 processor, with each core capable of a maximum clock frequency of 1.5 GHz and supporting up to three floating-point operations per cycle. Thus, the theoretical value of FLOPs can reach 18 GFLOPs (1.5 GHz×4 cores×3). An Intel Core i7 920 processor typically achieves around 63 GFLOPs, suggesting that the station is only suitable for lightweight computing tasks.

Figure 4.9 shows a comparison between the computation times for 200 iterations without and with the hybrid model. In the former, the genetic algorithm uses a fitness function based on a purely physical model. Specifically, the heat exchanger model is established using the logarithmic mean temperature difference (LMTD) method, while the chiller model consists of the compressor, condenser, expansion valve, and evaporator, also modeled using the heat and mass transfer principles. The detailed modeling methods have been thoroughly discussed in previous research(Xie et al., 2021). The results show that the average computation time per iteration using the physical model is about 3.22 seconds, while the hybrid model reduces this to 0.27 seconds. When applied to a genetic algorithm with an initial population of 16 and 8 generations, the physical model requires approximately 6.9 minutes to complete a single optimization computation. By replacing key models with ANN models, the optimization time is reduced to 0.6 minutes under the same conditions, enabling practical real-time applications of AI-empowered optimization control.

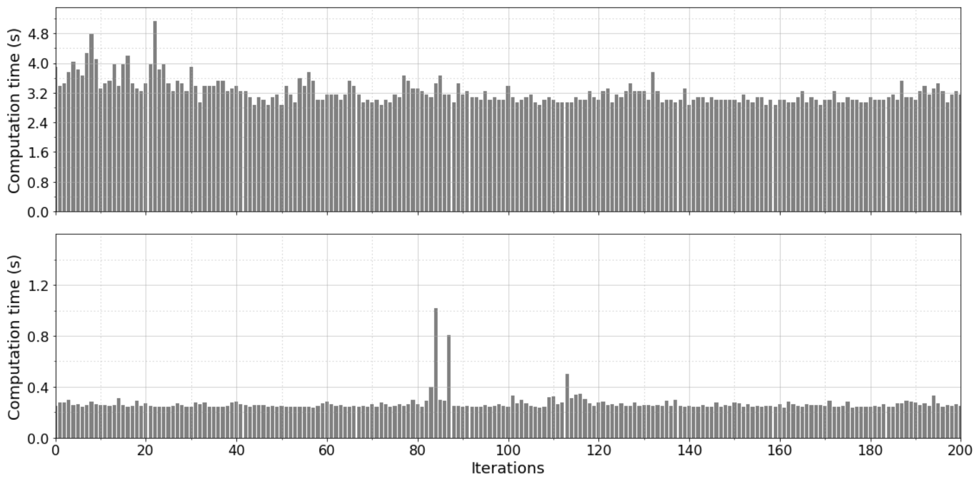


Figure 4.9 Iterative computation time: comparison before (top) and after (bottom) introducing the hybrid model

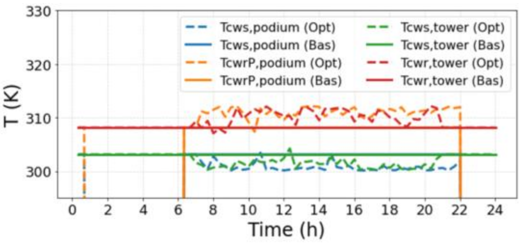
4.4.2 Control performance

Figure 4.10 presents the profiles of major system operation variables controlled by the proposed AI-empowered optimization method compared to those under the control of the baseline method with fixed setpoints. Figure 4.10 (a) shows that, compared to the baseline, the optimized cooling water return temperature increases by 2.30 K on average, while the supply temperature decreases by 1.93 K. These adjustments reduce cooling water flow rates but increase seawater flow rates, as illustrated in Figure 4.10 (b).

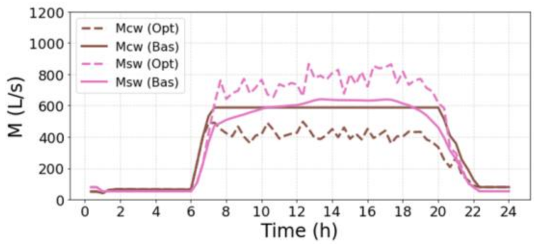
The operational numbers of chillers and pumps follow a rule-based mechanism, ensuring that each unit operates within an acceptable efficiency range. As a result, the number of seawater pumps increases to handle the higher flow rate, while the total number of chillers and cooling water pumps remains unchanged, as shown in Figure 4.10 (c).

Figure 4.10 (d) demonstrates the impact of this method on power consumption. While seawater pump power consumption increases, cooling water pump power consumption decreases significantly. Furthermore, although the higher condenser temperature difference increases compressor energy consumption, this effect is mitigated by a rise in chilled water temperature, as shown in Figure 4.10 (e). Optimization constraints and penalty functions effectively limit the increase in chilled water temperature, ensuring that comfort requirements are not compromised.

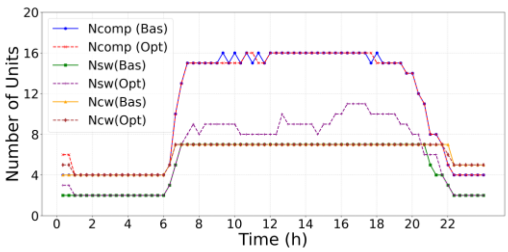
Finally, Figure 4.10 (f) shows the overall reduction in total energy consumption. By properly balancing the energy use of seawater pumps, cooling water pumps, and chillers, the AI-empowered optimization algorithm achieves an energy saving of 7.66% compared to the baseline case.



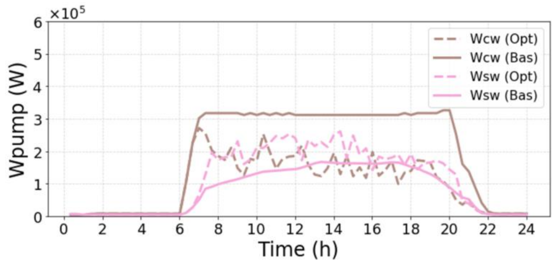
(a) Cooling water temperature set points



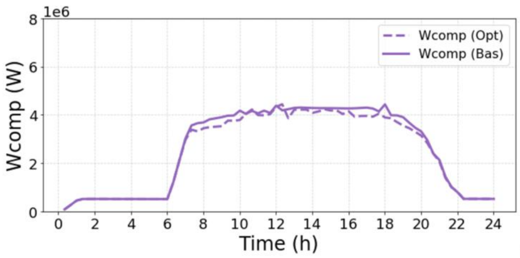
(b) Flow rate of pumps



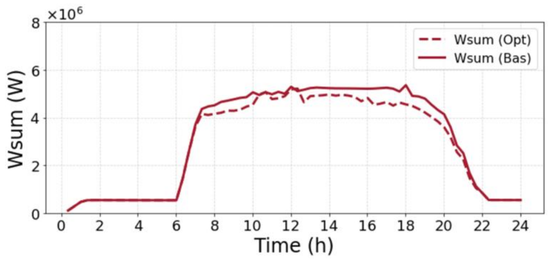
(c) Number of units in operation



(d) Power consumption of pumps



(e) Power consumption of chillers



(f) Total power consumption

Figure 4.10 Main operation variables under two control methods in a typical day

Figure 4.11 illustrates the detailed breakdown of hourly energy consumption and the profile in the Energy Efficiency Ratio (EER). The hourly energy consumption trends provide a clear representation of the instantaneous energy usage fluctuations under different control methods. The EER is defined as the ratio of building load to total system energy consumption, serving as an indicator of overall system efficiency. In the baseline control method, both the water pumps and the chillers operate based on fixed setpoints, resulting in relatively constant energy consumption ratios for each component. In contrast, the optimized control scheme utilizes global optimization through a genetic algorithm, which leads to a more efficient distribution of energy.

The results show that the energy savings of the cooling pumps and chillers significantly outweigh the increased energy consumption of the seawater pumps, leading to a higher EER. As a result, the red line representing the optimized EER in the figure shows a higher value compared to the black line of the basic control method.

Detailed energy savings are presented in Table 4.2. With the global optimization control, chiller energy consumption was reduced by 4.45%, saving approximately 2,605.7 kW. The cooling water pumps achieved substantial energy savings of 47.9% (4,340.1 kW), due to the reduced cooling water flow rate. Since seawater does not require long-distance transport as cooling water, the seawater pump's power consumption is about half that of the cooling water pump. As a result, under the global optimization, seawater pump energy consumption increased by 38.1% (1,469.3 kW). Eventually, the total energy consumption for the basic control scheme was 71,496.6 kW, while the optimized control method achieved a total energy consumption of 66,020.1 kW and an overall energy savings of 7.66% (5,476.5 kW).

Given the increasing importance of the sustainability of HVAC systems, it is essential to analyze the potential reduction in carbon footprint associated with energy conservation. In this paper, the average carbon emission factor of electricity in Hong Kong (0.529 kg CO_2/kWh) is used to calculate the energy savings. The results show that, under typical daily summer conditions, the proposed method can reduce CO_2 emissions by approximately 2.9 tons per day.

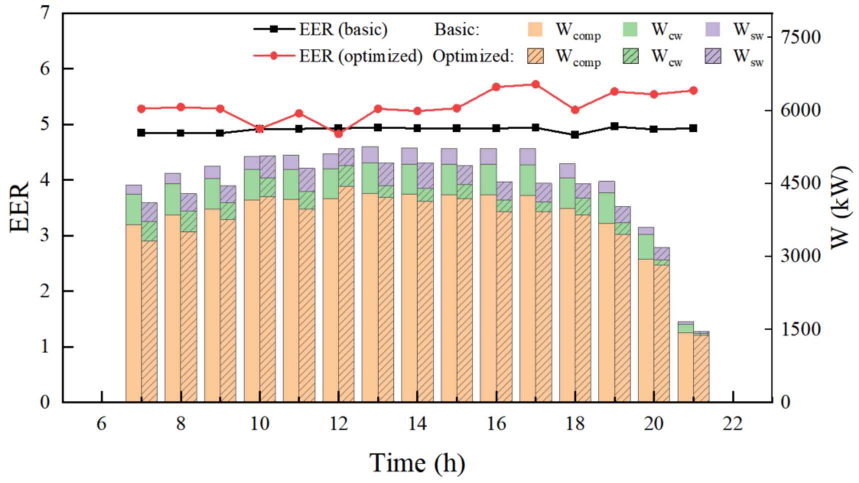


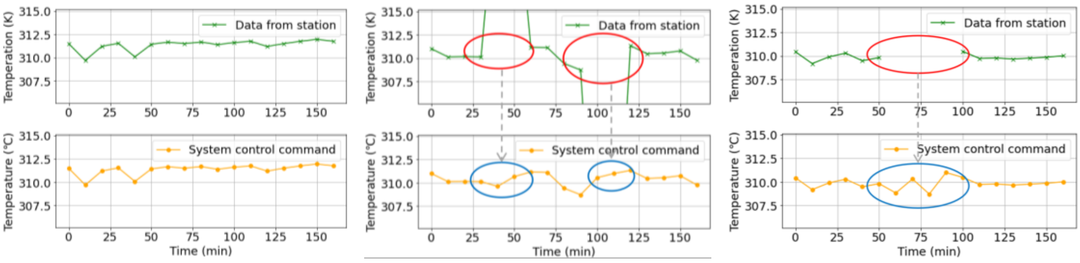
Figure 4.11 Comparison between hourly energy consumption and Energy Efficiency Ratio under baseline and AI-empowered optimization methods

Table 4.2 Comparison of energy consumption throughout the day

Energy consumption (kW)	Basic operation	Optimized operation	Comparison
Compressors	58572.1	55966.4	4.45%
Cooling water pumps	9066.6	4726.5	47.9%
Seawater pumps	3857.9	5327.2	-38.1%
Total	71496.6	66020.1	7.66%

4.4.3 Robust assurance in practical operation

Various abnormal disturbances are also introduced in the optimization tests. The operation of the DDC controller is depicted in Figure 4.12. For regular optimization control processes, the controller precisely achieves the optimal setpoint received from the smart station. When the data from the optimization algorithm overshoots due to external noise or other transmission problems, the controller automatically switches to the expert knowledge-based scheme, as illustrated in Figure 4.12 (b). Figure 4.12 (c) shows the control reaction when it is unable to receive information from the smart station for an extended period, whether due to a physical layer failure (such as disconnection, poor connection, or port damage) or a network issue (such as packet loss). The test results demonstrate that the robust assurance scheme can mitigate the impact of these abnormal situations by generating set-points based on the current operational state and previous optimization experience. Although it exhibits some fluctuations and cannot achieve global optimization accurately, this scheme effectively ensures the stability and reliability of AI-empowered optimization in real-time operation.



(a) Normal reception (b) Data anomalies (c) Data interruption

Figure 4.12 Controller response in different situations

4.5 Summary

This paper presents an AI-empowered optimization method for practical implementation at the field level. A hardware-in-the-loop test is conducted to validate the computational efficiency, control performance, and operational robustness of the proposed method. The results show that the proposed AI-empowered optimization method accelerates the optimization process by a factor of 6.9 and achieves an energy saving of 7.66%. The robust assurance scheme effectively mitigates the impacts of abnormal disturbances through an expert-based switching mechanism, ensuring stable and reliable system operation. This research advances the boundaries of AI applications in the building and energy system fields, however, there is still much to explore regarding efficiency and adaptability under dynamic working conditions in practical applications. Future research should further address practical application issues, such as developing generalized AI models and transitioning toward adaptive control.

CHAPTER 5 AI-EMPOWERED ONLINE OPTIMAL CONTROL STRATEGY ADOPTING ENERGY GRADIENT FOR ENHANCED ROBUSTNESS AND ENERGY-EFFICIENCY

This chapter introduces an AI-empowered online optimal control strategy that adopts the concept of energy gradient to achieve robust and energy-efficient control of building central cooling systems. Recognizing the limitations of conventional optimization methods that rely on absolute energy consumption calculations—often vulnerable to measurement errors and system uncertainties—this study proposes a novel approach that leverages the relative variations in energy consumption, encapsulated in a physical variable termed the “energy gradient.” By integrating this physical insight with machine learning models, the proposed strategy guides the progressive adjustment of control setpoints, enabling more stable and reliable operation under dynamic conditions. This chapter details the theoretical formulation of the energy gradient concept, the development of hybrid modeling techniques, and the practical implementation of the optimization algorithm on embedded control hardware. Validation through hardware-in-the-loop testing demonstrates that the proposed method not only enhances control robustness against noise and uncertainties but also achieves comparable energy savings compared to conventional optimization strategies. The subsequent sections elaborate on the methodological framework, system modeling, algorithmic design, and experimental results that collectively substantiate the effectiveness of this innovative optimization approach.

5.1 Conventional optimization methods for building cooling systems

Conventional optimization methods for building cooling systems predominantly rely on calculations of absolute energy consumption to identify optimal operational setpoints. In these approaches, the total energy use of system components—including chillers, cooling water pumps, and seawater pumps—is directly computed under various control configurations, and optimization algorithms are then applied to search for the setpoints that minimize this absolute consumption. The role of AI techniques in such conventional frameworks has primarily been to simplify complex physical models or accelerate the computational search process through machine learning-based predictions or heuristic optimization algorithms. However, despite their theoretical efficacy, these methods often face significant challenges in practical applications. The inherent complexity of the optimization algorithms, coupled with the demand for real-time response in dynamic operational environments, results in high computational burdens and potential delays.

Furthermore, absolute energy consumption calculations are highly sensitive to measurement errors and modeling inaccuracies, making them less reliable under noisy or uncertain conditions. As a result, while many AI-empowered methods have demonstrated promising results in simulations, their practical deployment for real-time optimization of large-scale cooling systems remains limited. These challenges highlight the need for alternative approaches that can enhance both the robustness and computational efficiency of online optimal control. In this context, the concept of utilizing relative energy variations—introduced via the notion of an "energy gradient"—offers a novel pathway for overcoming the limitations associated with absolute energy-based optimization methods.

5.2 Proposed energy gradient-guided AI-empowered online optimal control strategy

This section begins by introducing the core concept of the proposed energy gradient-guided online optimal control strategy for cooling systems. It then provides a detailed explanation of the optimization control workflow and the strategy development process.

5.2.1 Concept of "energy gradient" and its application for control optimization

In the online optimal control of a cooling system, identifying the optimal setpoint is essentially a trade-off process that balances the energy consumption of the seawater pumps, cooling water pumps, and chillers to minimize the total energy consumption. The conventional strategy typically employs an absolute total energy consumption as the objective function in the optimization algorithm to identify the setpoints that minimize total energy consumption. Such a conventional strategy is referred to as the "absolute total energy consumption-based optimization strategy" in this paper. Due to the complexity of deriving analytical solutions, advanced algorithms are often employed to determine the optimal combination of setpoints (T_{set1}, T_{set2}), as described in Eq. (5.1).

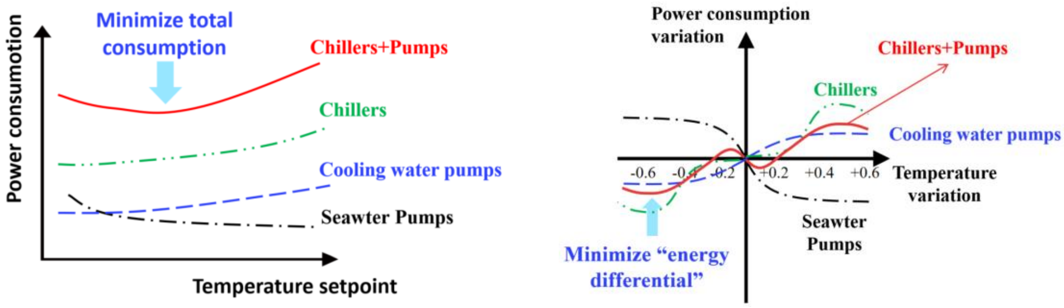
$$Min\ W_{sum}(T_{set1}, T_{set2}) = Min\ [W_{sw}(T_{set1}, T_{set2}) + W_{cw}(T_{set1}, T_{set2}) + W_{ch}(T_{set1}, T_{set2})] \tag{5.1}$$

where, W_{sum} is the total energy consumption. The subscripts *sw*, *cw*, and *ch* represent the seawater pump, cooling water pump, and chiller, respectively.

Unlike this strategy, which directly calculates the energy consumption for various setpoint combinations, the concept of "energy gradient" ($\frac{dW_{sum}}{dT_{set}}$) is defined to quantify the relative change in energy consumption among various sub-systems when varying their setpoints. It is represented by a first-order derivative of energy consumption (W) with respect to the temperature of setpoint (T). The introduction of this variable helps identify the direction of the setpoint changes that lead to a reduction in global energy consumption. The proposed energy gradient-guided AI-empowered online optimal control strategy utilizes this relative correlation to identify the resetting directions and quantities for the setpoints ($\Delta T_{set1}, \Delta T_{set2}$), as detailed in Eq. (5.2)

$$Min\ \Delta W_{sum}(\Delta T_{set1}, \Delta T_{set2}) \approx Min\ \left\{\frac{\partial W_{sum}}{\partial T_{set1}}\Delta T_{set1} + \frac{\partial W_{sum}}{\partial T_{set2}}\Delta T_{set2} + \frac{1}{2}\left[\frac{\partial^2 W_{sum}}{\partial T_{set1}^2}(\Delta T_{set1})^2 + \frac{\partial^2 W_{sum}}{\partial T_{set2}^2}(\Delta T_{set2})^2 + \frac{\partial^2 W_{sum}}{\partial T_{set1}\,\partial T_{set2}}\Delta T_{set1}\Delta T_{set2}\right]\right\} \tag{5.2}$$

Figure 5.1 illustrates the alternative processes for searching optimization solutions, adopting the absolute total energy consumption-based optimization strategy and the energy gradient-guided optimization strategy using Eqs. (5.1) and (5.2) respectively. As the temperature setpoints vary, the energy consumption of pumps and chillers either increases or decreases. The conventional online optimal control strategy compares the total energy consumption across the entire variation spectrum of setpoints to identify the absolute minimum. In contrast, the proposed strategy examines relative energy consumption and guides continuous and gradual adjustment towards the direction of reducing total energy consumption. For instance, when a temperature setpoint decreases by 0.4°C, the energy consumption of seawater pumps increases by 5 kW, while the energy consumption of chillers and cooling pumps decreases by 3 kW and 4 kW, respectively, resulting in a net energy saving of 2 kW. Under these conditions, reducing this setpoint is a viable optimization adjustment direction.



Absolute energy consumption Relative energy consumption

Figure 5.1 Process of searching for the optimal solution using two control functions

In the process of energy gradient-guided online optimal control, the algorithm continuously updates optimization adjustments under a dynamic system environment, gradually approaching the real-time optimal setpoint of the cooling system. This progressive, dynamic search process is illustrated in Figure 5.2. At time θ_1, the theoretical global optimal setpoint is $T_{set,opt(\theta_1)}$, while the current system operating setpoint is $T_{set(\theta_1)}$. Guided by the optimization algorithm, the setpoint incrementally moves toward $T_{set,opt(\theta_1)}$ at each optimization interval t, adapting to real-time operational conditions. However, as the system reaches time θ_2, the global optimal setpoint may shift to $T_{set,opt(\theta_2)}$ due to changes in system dynamics. In response, the optimization algorithm continues to track the updated optimal direction, dynamically adjusting the setpoints to maintain energy-efficient performance under evolving conditions.

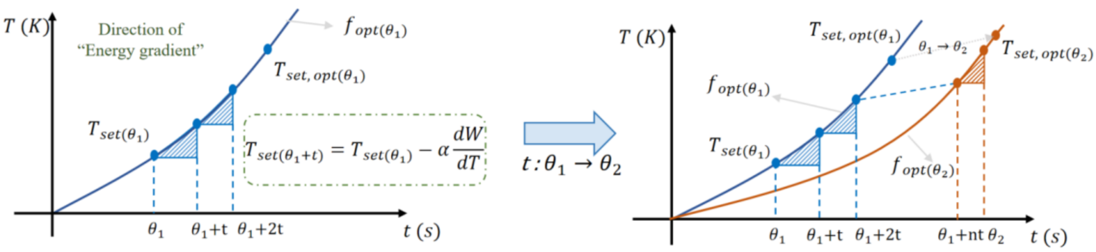


Figure 5.2 Progressive dynamic optimization process based on "energy gradient"

5.2.2 *Outline of proposed online optimal control strategy*

Figure 5.3 illustrates the proposed energy gradient-guided AI-empowered online optimal control strategy. The optimization algorithm can be deployed on either edge devices or a central computer (server). Each device functions as a smart control station, capable of acquiring operational data, executing optimization computation, and applying control decisions to the cooling systems in real-time. Each optimization interval can be divided into three steps: ***1)*** Real-time data collection, ***2)*** Genetic Algorithm-based search, and ***3)*** Execution of control decisions.

The process begins with the collection of real-time measurements by a smart control station from field digital controllers via the IP network. Key system operational variables, including temperature, pressure, flow rate, number of running devices, and building load, are continuously collected and stored in the station's internal database. Then, a Genetic Algorithm (GA)-based optimization scheme is employed to explore candidate combinations of setpoints. Each candidate solution is evaluated using an ANN model pretrained by machine learning, which estimates the energy gradients with respect to changes in setpoints. This allows the optimization scheme to prioritize candidate solutions that achieve minimum total energy use, according to the objective function defined in Eq. (5.3).

$$Min\ \Delta W_{sum}(\Delta T_{set1}, \Delta T_{set2}) = Min\ [F_{ANN_{sw}}(\Delta T_{set1}, \Delta T_{set2}) + F_{ANN_{cw}}(\Delta T_{set1}, \Delta T_{set2}) + F_{ANN_{ch}}(\Delta T_{set1}, \Delta T_{set2})] \tag{5.3}$$

where, F_{ANN} represents the artificial neural network model-based fitness function, which estimates the change in energy use resulting from the adjustments to the two temperature setpoints $(\Delta T_{set1}, \Delta T_{set2})$.

Finally, the optimized setpoints are sent to the control system for execution.

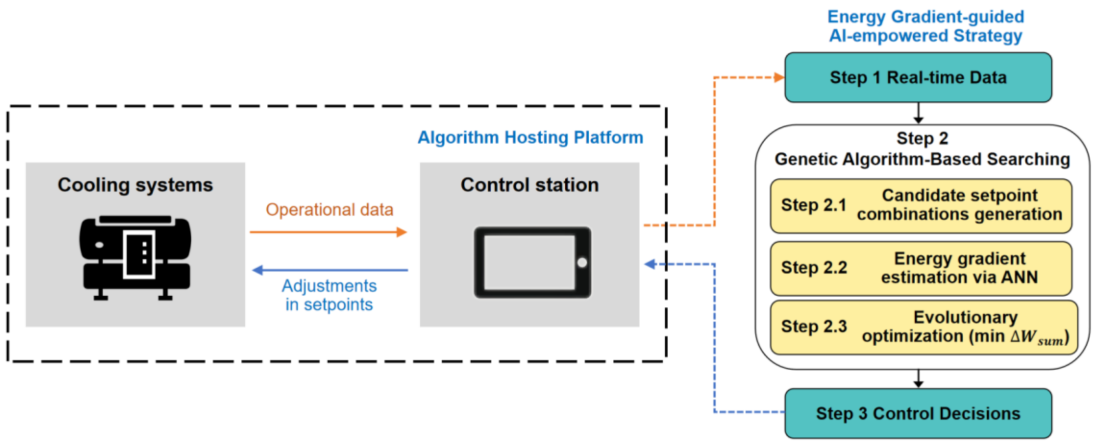


Figure 5.3 Outline of the proposed online optimal control strategy

5.3 System overview and model construction

5.3.1 Description of the test system

The test system adopted in this chapter is based on the reference building and cooling infrastructure previously described in Section 4.3.1. To briefly reiterate, the selected facility is a large-scale commercial building complex located in West Kowloon, Hong Kong, with a gross floor area of 288,010 m² and a total design cooling load of 55,331 kW. The cooling demand is met by a seawater-cooled central chiller system comprising 21 chiller units, distributed across five building zones.

Building upon the configuration outlined earlier, this section provides additional technical specifications of the cooling system components to support detailed model development and subsequent optimization tasks. The composition of the cooling system is shown in Figure 5.4, which incorporates:

- 13 seawater pumps (260 L/s each) to transfer heat to the seawater intake system,
- 15 cooling water pumps (195 L/s) and 6 auxiliary pumps (100 L/s) for flexible operation under variable load conditions,

- 15 chillers rated at 3517 kW and 6 chillers rated at 1758 kW, with the lower-capacity units dedicated to supporting 24-hour commercial zones requiring uninterrupted cooling.

The equipment configuration is further detailed in Table 5.1, including specifications of chiller compressors, heat exchangers, and associated pump systems. The refrigerant used is R134a, and all chillers are centrifugal type, ensuring high efficiency and stable operation under partial load conditions.

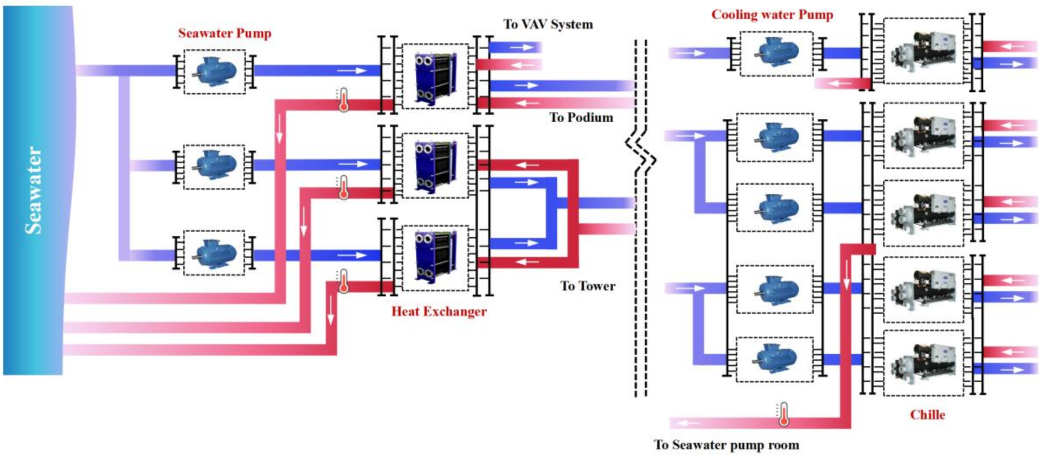


Figure 5.4 Cooling System Composition

Table 5.1 System Equipment Specifications

Equipment	Specifications	
Chiller	Manufacturer: Carrier 19XR7172, Carrier 23XRV5959, Type: Comply water cooled, Cooling capacity (kW): 3517, 1758, Power input (kW): 529, 266.2, Full load COP: 6.2, Refrigerant: R134a	
	Compressors	Type: Centrifugal, Motor speed (rpm): 3000
	Evaporator	Type: Shell & Tube, Chilled water in/out temp: 12.3/6.8°C, Flow (l/s): 152.8, 76.4, Pressure drop (kPa): 73.7, 37
	Condenser	Type: Shell & Tube, Cooling water in/out temp: 29.5/34.5°C, Flow (l/s): 194.5, 97.2, Pressure drop (kPa): 58.1, 37.1
Seawater pump	Type: Horizontal Split Casing, Flow rate (l/s): 260, Operation head (m): 30, Motor Rating (kW): 110, Full Load Speed (r/min): 1480	
Cooling water pump	Type: Horizontal Split Casing, Flow rate (l/s): 195, 100, Operation head (m): 39, Motor Rating (kW): 110/ 55, Full Load Speed (r/min): 1450	

Chilled water pump	Type: Horizontal Split Casing, Flow rate (l/s): 152.3, 76.1, Operation head (m): 33, Motor Rating (kW): 75/ 37, Full Load Speed (r/min): 1450

The simulation model is developed in Dymola using the Modelica language, leveraging the “Buildings” library to model all major system components. The system architecture is modularly organized into distinct subsystems, including the seawater circuit, heat exchanger network, pumping stations, chillers, and terminal load side. Environmental boundary conditions, such as outdoor temperature and dynamic cooling loads, are input from empirical datasets, as previously described in Section 4.3.1.

5.3.2 Model construction of the proposed online optimal control strategy

For the development of the energy gradient-guided AI-empowered online optimal control strategy, a critical step is to construct an ANN model capable of capturing the trend of the energy gradient in response to changes in setpoints. The physical model built in Dymola is utilized to perform comprehensive simulations, generating a complete dataset of system operations by varying load levels, ambient weather conditions, and equipment configurations within the simulation environment.

This dataset is used to train the ANN model, which captures the nonlinear mapping relationship between setpoint adjustments and the physical variable called "energy gradient" proposed in this study. The details of the ANN model are presented in Table 5.2, featuring a relatively complex three-layer hidden architecture with 20:40:40 neurons across the layers. The input features include: Building cooling load (*Load*), Cooling water supply and return temperatures (T_{cws}, T_{cwr}), Number of operating seawater and cooling water pumps (Num_{sw}, Num_{cw}), Chilled water flow rate and supply temperature (M_{chw}, T_{chws}), and intended setpoint adjustments for cooling water supply and return temperatures ($\Delta T_{cws,set}$, $\Delta T_{cwr,set}$). The outputs are the predicted variations in energy consumption for chillers (ΔW_{ch}), cooling water pumps (ΔW_{cw}), and seawater pumps (ΔW_{sw}). The LeakyReLU activation function is applied to ensure stable learning and to prevent the "dying ReLU" problem, enabling smooth gradient propagation throughout the network and enhancing the robustness of the model.

Table 5.2 System equipment specifications

Inputs	Ten features: *Load*, T_{cws}, T_{cwr}, Num_{sw}, Num_{cw}, M_{chw}, T_{chws}, $\Delta T_{cws,set}$, $\Delta T_{cwr,set}$
Hidden Layer	Three hidden layers
Neurons	20:40:40
Activation Function	LeakyReLU activation function
Optimizer	adam optimizer
Loss function	Mean squared error
Outputs	Three neurons: ΔWch, ΔWcw, ΔWsw

5.4 Test arrangement

A comprehensive test has been arranged to thoroughly evaluate the proposed energy gradient-guided AI-empowered online optimal control strategy, including mechanism analysis and comparative analysis.

5.4.1 System mechanism analysis

System mechanistic analysis focuses on discovering the internal dynamics of the cooling system through parameter correlation and energy-saving potential assessment.

- *Parameter Correlation Analysis:* This analysis identifies causal relationships among operational and performance parameters of the cooling system. Statistical strategies are employed to quantify the influence weights between parameters, providing a deeper understanding of the energy-saving mechanisms and helping the selection of key input variables.

- *Energy-Saving Potential Analysis:* This analysis constructs three-dimensional total energy consumption maps across different load conditions and setpoint combinations. These visualizations provide an intuitive understanding of the energy landscape, assist in defining optimization boundaries, and highlight regions with high energy-saving potential.

5.4.2 Comparative analysis

The comparative analysis focuses on operational robustness, setpoint adjustment stability, and energy-saving performance. It compares a baseline control method, a conventional optimization strategy, and the proposed optimization strategy, as listed below.

- *Baseline control:* A widely used approach in practice that operates the system with fixed setpoints and does not involve optimization.
- *Conventional GA-based online optimal control:* Utilizes a Genetic Algorithm (GA) to optimize T_{cws} and T_{cwr}, using a fitness function that directly minimizes the absolute total energy consumption of the cooling system.
- *Proposed energy gradient-guided online optimal control:* Also adopting GA, this strategy uses a fitness function to estimate relative changes in energy consumption for individual components (e.g., chiller, cooling pump, seawater pump), rather than calculating absolute energy consumption values.

Table 5.3 summarizes the optimization model configurations for the two GA-based strategies. In the proposed energy gradient-guided strategy, a smaller population size and fewer generations are selected to match the reduced search space. The optimization interval is shortened to 2 minutes due to the lower computational complexity. Furthermore, a higher mutation rate is adopted to mitigate the risk of premature convergence and avoid falling into local optima.

Table 5.3 Optimization model configurations in two online optimal control strategies

Types	Conventional online optimal control	Energy gradient-guided online optimal control
Strategy	Global optimization $W = f(T_{cws,set}, T_{cwr,set})$	Approaching optimization $\Delta W = f(\Delta T_{cws,set}, \Delta T_{cwr,set})$
Fitness functions	Physical formula-based	ANN model-based
Population size	20	10
Generational number	16	8
Optimization interval	10 mins	2 mins
Crossover rate	0.75	0.75
Mutation rate	0.05	0.12
Outputs	W_{ch}, W_{cw}, W_{sw}	ΔW_{ch}, ΔW_{cw}, ΔW_{sw}

To evaluate the robustness of the online optimal control strategies, two types of input perturbations—Gaussian random noise and systematic bias—were independently introduced to the water temperature and flow rate inputs of the model. The perturbation ranges were set to 0–0.5 K for temperature and 0–2% for flow rate, covering typical sensor uncertainty levels observed in HVAC systems. For each error type, a series of simulations was conducted to compare the relative errors in estimated energy consumption between the conventional online optimal control strategy and the proposed energy gradient-guided online optimal control strategy.

Beyond evaluating robustness, the stability and energy-saving performance of both control strategies were further validated through a hardware-in-the-loop test on a physical control station. The hardware-in-the-loop test platform used in this chapter follows the same foundational setup introduced in Section 4.3.2, where a real-time digital twin of the cooling system, developed and simulated in Dymola, is integrated with a Raspberry Pi-based control station over a TCP/IP-based local area network. This architecture enables closed-loop validation of control strategies in a near-realistic environment.

5.5 Test results and discussions

5.5.1 System mechanism analysis: parameter dynamic correlations and energy gradient trends

5.5.1.1 Parameter correlation analysis based on dynamic operation data

The correlation of parameters in the system operation process is categorized into operational parameters and performance parameters, as illustrated in Figure 5.5. First, the building load (*Load*) demonstrates a strong correlation with the chilled water flow (M_{chw}) and the number of chillers in operation (Num_{ch}), with correlation coefficients of 0.82 and 0.83, respectively. It also exhibits moderate correlations with the cooling water flow rate (M_{cw}) and seawater flow rate (M_{sw}), with coefficients of 0.53 and 0.7, respectively.

Under a certain building load, the total power consumption, which includes the power consumption of the chillers (W_{ch}), cooling water pumps (W_{cw}), and seawater pumps (W_{sw}), needs to be minimized. Among these, the chiller power consumption (W_{ch}) is highly correlated with the water flow rate variables (M_{sw}, M_{cw}, and M_{chw}), while the pump power consumption is directly related to their flow rates. Furthermore, the pump states show relatively high correlations with the water temperatures they control. For example, the seawater flow rate (M_{sw}) and the number of seawater pumps (Num_{sw}) are closely related to the cooling water supply temperature (T_{cws}), while the cooling water pump flow rate (M_{cw}) and number (Num_{cw}) are highly correlated with the return cooling water temperature (T_{cwr}). In this context, the W_{sw} and W_{cw} also exhibit a certain degree of negative correlation with T_{cws} and T_{cwr}, with correlation coefficients ranging from -0.4 to -0.6. In this study, the control strategies optimize the cooling water supply and return temperatures (T_{cws} and T_{cwr}) setpoints to maximize system efficiency.

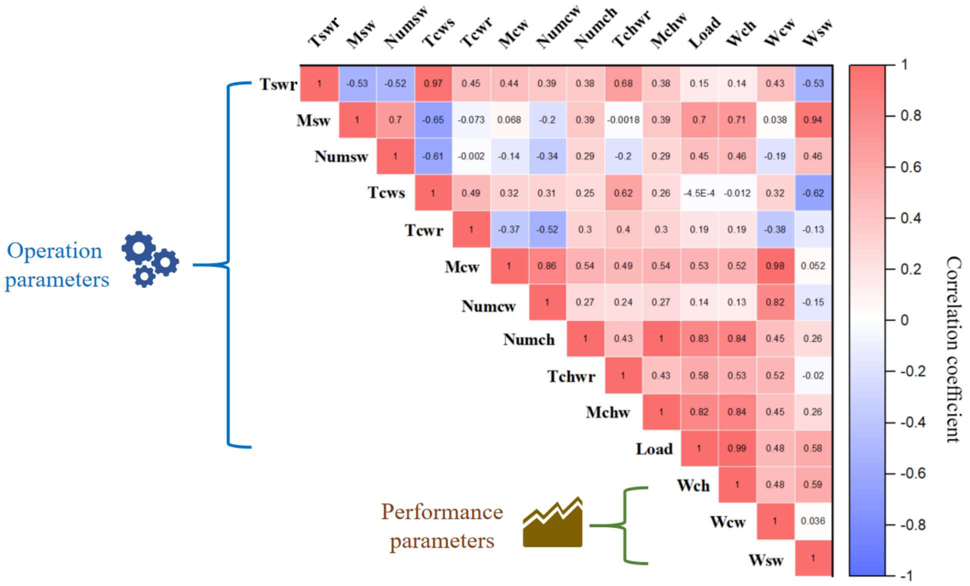
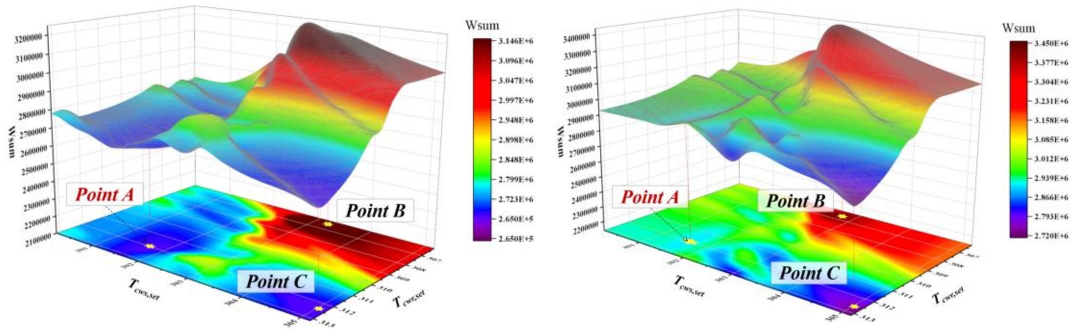


Figure 5.5 Correlation coefficient matrix table for the studied system

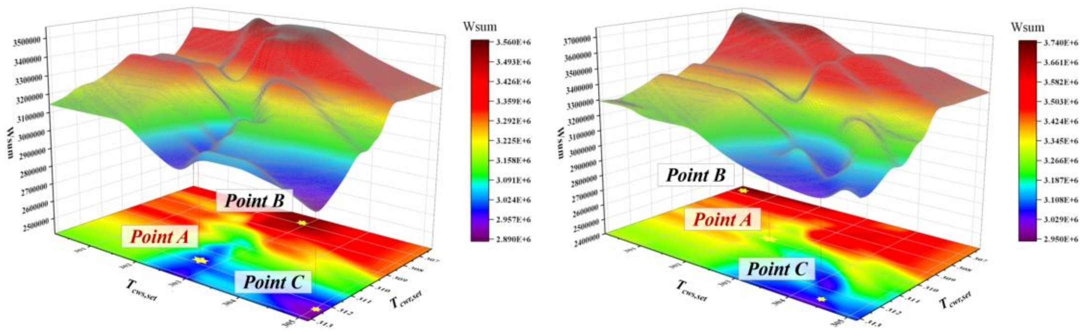
5.5.1.2 Energy saving potential analysis

A 3D energy consumption map is constructed to visualize how energy usage varies with Tcws and Tcwr under different load conditions, as shown in Figure 5.6. The results show that the system exhibits greater energy-saving potential under partial load conditions. Therefore, the number of operating chillers should be controlled to keep them running within the 60%–80% load range.

Under a specific load, a lower Tcwr often results in peak energy consumption at point B, while increasing both Tcws and Tcwr can lead to a global minimum in total energy consumption at point C. However, the condenser entering water temperature is too high to ensure reliable chiller operation. To address this issue, an optimization space is defined to constrain a feasible range for setpoints. Within this constrained region, point A is identified as the optimal trade-off between energy savings and system stability.



(a) Load at 65% (b) Load at 69%



(c) Load at 74% (d) Load at 80%

Figure 5.6 3D energy consumption maps (energy usage vs. supply and return cooling water temperature) under different cooling load conditions

5.5.2 Comparative analysis of two online optimal control strategies

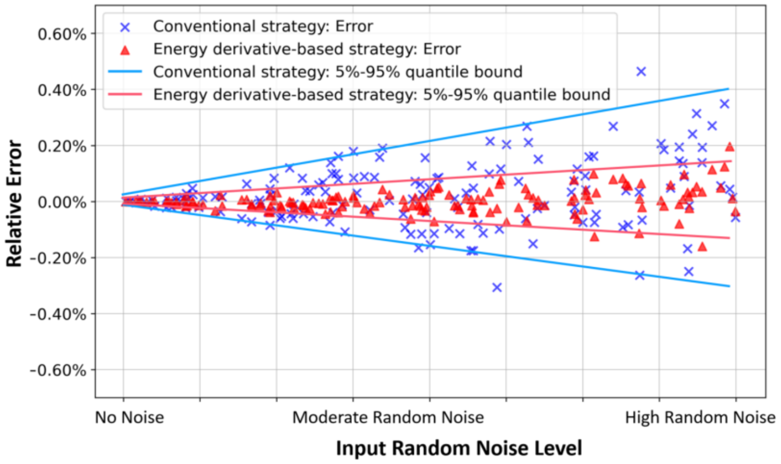
5.5.2.1 Robustness analysis

Two types of input perturbations were introduced to comprehensively evaluate the robustness of the online optimal control strategies under measurement uncertainty and bias: random noise and systematic bias. These perturbations were independently applied to the water temperature (0–0.5 K) and flow rate (0–2%) input variables. The resulting variations in the estimated energy consumption were analyzed for both the conventional online optimal control strategy and the proposed energy gradient-guided strategy, as illustrated in Figure 5.7.

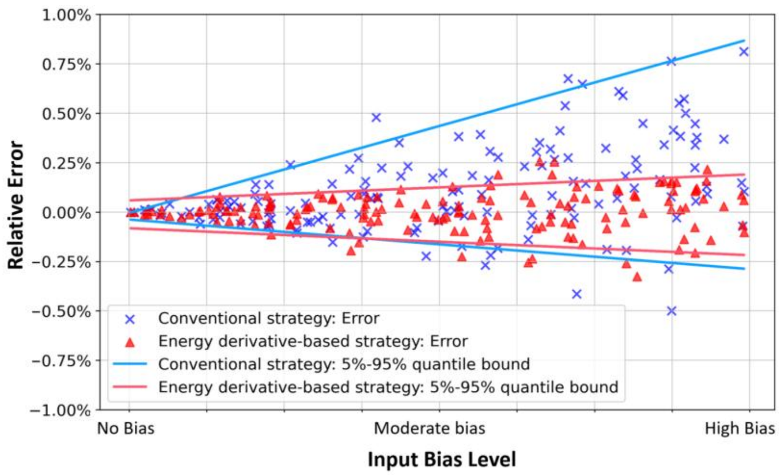
Figure 5.7 (a) illustrates the relative error of the model output under random noise. As the noise level increases, both strategies exhibit increasing output fluctuations, with the conventional strategy displaying a wider dispersion and greater variability in relative error. In contrast, the proposed energy gradient-guided strategy maintains a more centralized error distribution, with narrower bounds. Even at high noise levels, the proposed strategy constrains the majority of estimation errors within ±0.2%, whereas the conventional strategy shows error bounds approaching ±0.4%. This indicates the greater robustness of the proposed strategy in resisting stochastic fluctuations and maintaining stability in energy estimation.

Figure 5.7 (b) presents the relative error of the model output under systematic input bias, where persistent sensor drift introduces a random positive or negative offset in measurements. The error distribution of the conventional online optimal control strategy expands asymmetrically as the bias increases, reaching error bounds exceeding ±0.75% at high bias levels. Conversely, the proposed energy gradient-guided strategy maintains a significantly tighter and more symmetrical error band, with the output error remaining constrained within approximately ±0.25%. This demonstrates that the proposed strategy is less sensitive to sensor drift and offset errors.

These findings demonstrate that by determining control adjustments based on energy gradients rather than raw sensor values, the proposed strategy exhibits inherent resilience to sensor noise and drift, thereby improving the robustness of control performance.



(a) Under random input noise

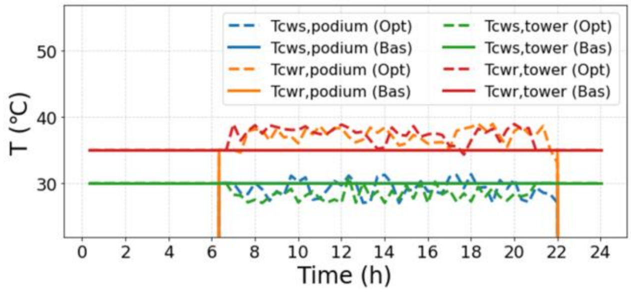


(b) Under systematic input bias

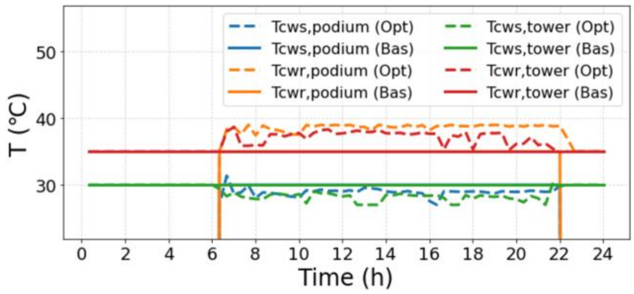
Figure 5.7 Relative error in estimated energy consumption under varying error levels

5.5.2.2 Stability analysis

The optimization setpoints of two online optimal control strategies on a typical day are illustrated in Figure 5.8**.** Each control strategy is evaluated against the baseline control, in which fixed set points are adopted. The building is divided into the podium and tower sections according to their functions, and the results representing the podium (subscript P) and the tower (subscript T) are plotted separately. Overall, it can be observed that both online optimal control strategies dynamically adjust the cooling water supply temperature setpoint ($T_{cws,set}$) and cooling water return temperature setpoint ($T_{cwr,set}$) throughout the day. On average, the T_{cws} of the conventional online optimal control strategy is reduced by 1.17K, while T_{cwr} is increased by 2.14K. In the proposed strategy, T_{cws} is decreased by 1.45K and T_{cwr} is increased by 2.86K. This difference is attributed to the conventional online optimal control strategy directly determining absolute setpoint values, which tends to result in larger fluctuations and overshoots—as observed around 14:00 and 18:00 in Figure 5.8 (a).



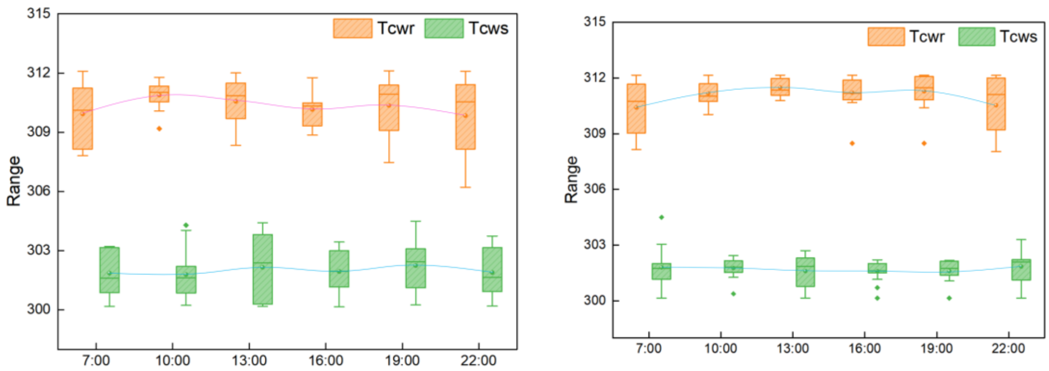
(a) Conventional online optimal control strategy VS. baseline control



(b) Proposed online optimal control strategy VS. baseline control

Figure 5.8 Optimization setpoints of two optimization strategies on a typical day

Figure 5.9 presents the box plot of setpoints generated by the two online optimal control strategies. In the conventional online optimal control strategy (Figure 5.9a), the variations in both T_{cws} and T_{cwr} are more significant, indicating a relatively unstable control process. By contrast, the proposed online optimal control strategy (Figure 5.9b) exhibits a much narrower temperature variation range, particularly during high cooling load conditions. These results demonstrate that the proposed strategy enhances the stability of the control process by adopting the physical variable - “energy gradient”.



(a) Conventional online optimal control strategy

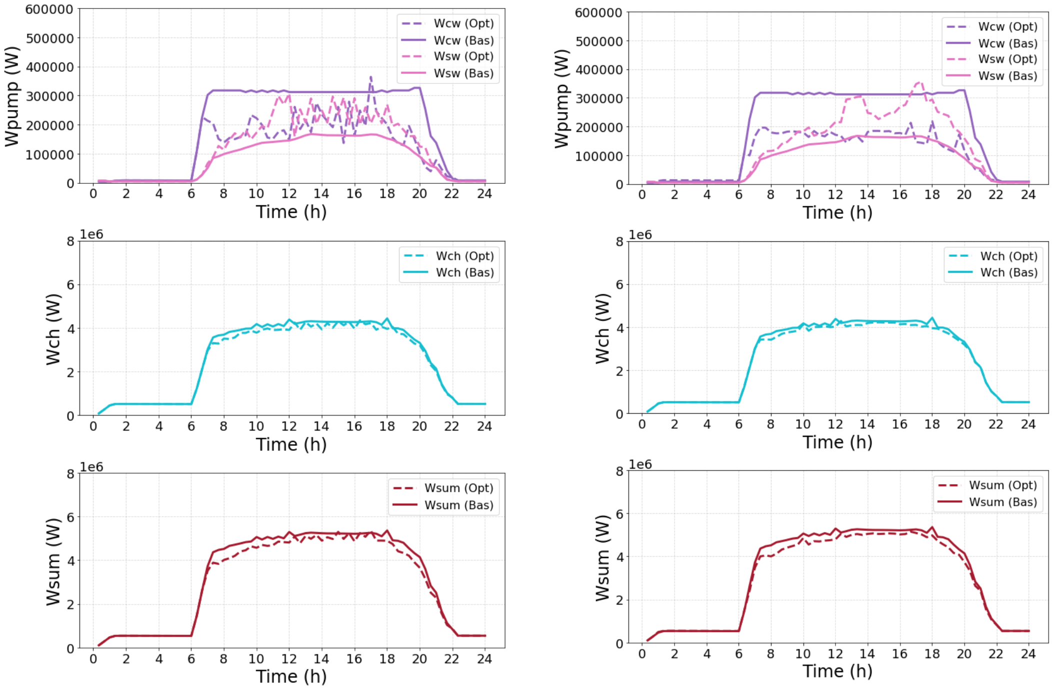
(b) Proposed online optimal control strategy

Figure 5.9 Box plot of the setpoints in two optimization strategies

5.5.2.3 Energy-saving performance

Figure 5.10 presents the energy consumption of the seawater pumps (W_{sw}), cooling water pumps (W_{cw}), and chillers (W_{ch}) over a typical working day, comparing the conventional and proposed online optimal control strategies with the baseline control. The results show that under both online optimal control strategies, W_{cw} is significantly reduced, while W_{sw} is slightly increased. This is because the rise in T_{cwr} reduces the cooling water flow requirement, which leads to a reduction in cooling water pump power. T_{cws} is decreased to ensure adequate heat dissipation, which increases the seawater flow rate and W_{sw}. However, since the cooling water pump has a higher energy consumption due to the long-distance pipeline, the optimization algorithm prefers to reduce W_{cw}, even if it results in a marginal increase in W_{sw}.

Ultimately, the proposed energy gradient online optimal control strategy achieves a 5.3% total energy saving compared with the baseline control, while the energy saving rate is 5.7% for the conventional online optimal control strategy. These results indicate that, while the energy gradient-guided online optimal control strategy enhances robustness and operational stability, it sacrifices a small portion of energy-saving potential in favor of greater control reliability.



(a) Conventional online optimal control strategy VS. baseline control

(b) Proposed online optimal control strategy VS. baseline control

Figure 5.10 Energy consumption of two online optimal control strategies compared with baseline control on a typical day

5.6 Summary

This study proposed an energy gradient-guided AI-empowered online optimal control strategy aimed at enhancing the stability and robustness of the online optimal control process in cooling systems. This strategy introduces a physical variable, the "energy gradient", as physical guidance for iteratively adjusting control setpoints in real time. Based on mechanism analysis and comparative analysis through hardware-in-the-loop testing, the following conclusions can be drawn:

- The energy gradient-based online optimal control strategy effectively captures local sensitivity trends in individual equipment energy consumption. By leveraging parameter correlation analysis and energy-saving potential analysis, the strategy identifies optimal setpoints that achieve a balance between system efficiency and stability under dynamic load conditions.
- Compared to conventional online optimal control strategies, the proposed energy gradient-guided strategy significantly enhances control robustness under both random noise and systematic sensor bias. In robustness tests, Gaussian random noise and systematic bias were independently introduced to flow rate (0–2%) and temperature (0–0.5 K) inputs. The proposed strategy consistently maintained lower relative output errors, reducing estimation error by 50% under high random noise. Under substantial systematic bias, it constrained error growth within ±0.25%, whereas the conventional strategy exhibited errors exceeding ±0.75%. These results demonstrate the ability of the proposed strategy to mitigate the effects of random noise and maintain low sensitivity to systematic offsets.
- The proposed strategy enhances control stability by estimating relative energy variation rather than the absolute value. Hardware-in-the-loop validation tests show that it reduces temperature fluctuation amplitudes during high cooling load conditions. The strategy avoids overshoots and maintains setpoint distributions within narrower ranges than the conventional online optimal control strategy.
- In terms of energy saving, the proposed energy gradient-guided online optimal control strategy achieves a total energy reduction of 5.3% compared with the baseline control, which is comparable to the 5.7% achieved by the conventional online optimal control strategy. The slight compromise in energy efficiency is justified by increased operational reliability and a smoother control process.

This work demonstrates the potential of energy gradient-informed, progressive adjustment of online optimal control strategies in real-world control systems. It provides a practical alternative to traditional black-box AI strategies by combining physics knowledge with AI-empowered algorithms. Future studies will focus on extending this strategy to hybrid energy systems, improving model adaptability, and addressing long-term degradation and generalization issues.

CHAPTER 6 INCREMENTAL ONLINE LEARNING WITH MODEL STRUCTURAL AUGMENTATION FOR AI-ENABLED OPTIMAL CONTROL CONSIDERING EQUIPMENT DEGRADATION

This chapter presents an incremental online learning strategy with model structural augmentation, designed to enable AI-based optimal control of building cooling systems under conditions of equipment degradation. While AI-empowered optimization methods have demonstrated remarkable capabilities in improving energy efficiency, their effectiveness can deteriorate over time as system components age or drift from their original operating conditions. To overcome this challenge, this study introduces a hybrid modeling approach that combines offline-trained models with an adaptive residual learning branch, which incrementally updates in response to real-time operational data. This strategy maintains the generalization capability of the offline model while dynamically correcting prediction errors caused by system degradation. Implemented on an embedded smart control station and validated through hardware-in-the-loop testing, the proposed approach demonstrates significant improvements in maintaining optimization performance, achieving notable energy savings even under degraded conditions. The following sections detail the theoretical framework, algorithm design, hardware implementation, and experimental validation of this innovative online learning strategy.

6.1 Challenges in AI-enabled optimization under equipment degradation

In building cooling systems, conventional optimization and control strategies are predominantly model-based, relying on mathematical or data-driven models to characterize the system's energy consumption and operational performance. These methods establish relationships between system inputs, such as setpoints or environmental conditions, and outputs like energy usage or thermal comfort, and they employ various optimization algorithms to search for control strategies that minimize energy consumption while maintaining operational constraints. Under design conditions and stable system performance, these model-based approaches have proven highly effective, achieving significant energy savings and operational improvements in both simulations and real-world applications.

However, as cooling systems operate over extended periods, inevitable phenomena such as equipment degradation—including heat exchanger fouling, declining chiller efficiency, sensor drift, and component wear—gradually alter the system's actual behavior. Such degradation leads to discrepancies between the model's predictions and the real system response, undermining the accuracy of optimization outcomes. Consequently, the optimization strategies derived from static or offline-trained models may become increasingly inaccurate, resulting in suboptimal control actions, reduced energy efficiency, and potential operational instability. These challenges underscore the limitations of existing AI-enabled optimization methods, which frequently assume static system conditions and fail to adapt to gradual system changes over time. Addressing these issues necessitates the development of adaptive learning mechanisms capable of continuously updating system models to reflect evolving operational states, thereby ensuring sustained optimization performance despite equipment aging and degradation.

6.2 Proposed incremental online learning strategy

6.2.1 Overview of the incremental online learning

The proposed strategy leverages an offline-trained hybrid model to capture the fundamental dynamics of the system and mitigate the risk of overfitting when real-time data are limited. A low learning-rate Adam optimizer is employed to facilitate stable and gradual online updates, effectively preventing catastrophic forgetting and preserving previously acquired knowledge. As illustrated in Figure 6.1, the proposed strategy comprises the following key steps:

1) *Initial model construction and offline pretraining:* A comprehensive simulation is performed to generate extensive datasets encompassing a wide range of design parameters, operating conditions, and system performance metrics. Based on these data, a hybrid model is constructed to capture the fundamental system behavior while maintaining flexibility for subsequent adaptation.

2) *Model structural augmentation:* To prepare the model for online adaptation, the proposed strategy augments the network architecture by introducing a residual learning branch. This additional residual neural network processes real-time data and produces corrective signals that are combined with the output of the pre-trained offline model.

3) *Online learning trigger mechanism:* During normal operation, the hybrid model continuously performs optimization based on its current parameters. To ensure model accuracy under changing system conditions or equipment degradation, an

online monitoring mechanism evaluates the deviation between model predictions and real system measurements. When the prediction error exceeds a predefined threshold, the incremental online learning process is activated to refine the model and reduce discrepancies.

4) *Low learning rate adaptive updating:* Incremental learning is performed using real-time operational data. The Adam optimizer, configured with a low learning rate, facilitates gradual updates to the adaptive layers, preventing catastrophic forgetting and maintaining a balance between model adaptability and robustness. Elastic net regularization is also applied to mitigate the impact of noisy data and preserve the model's ability to generalize across diverse operating conditions.

Through this structured incremental online learning strategy, the hybrid model achieves continuous adaptation to evolving operational conditions and equipment performance changes, while maintaining computational efficiency and physical interpretability.

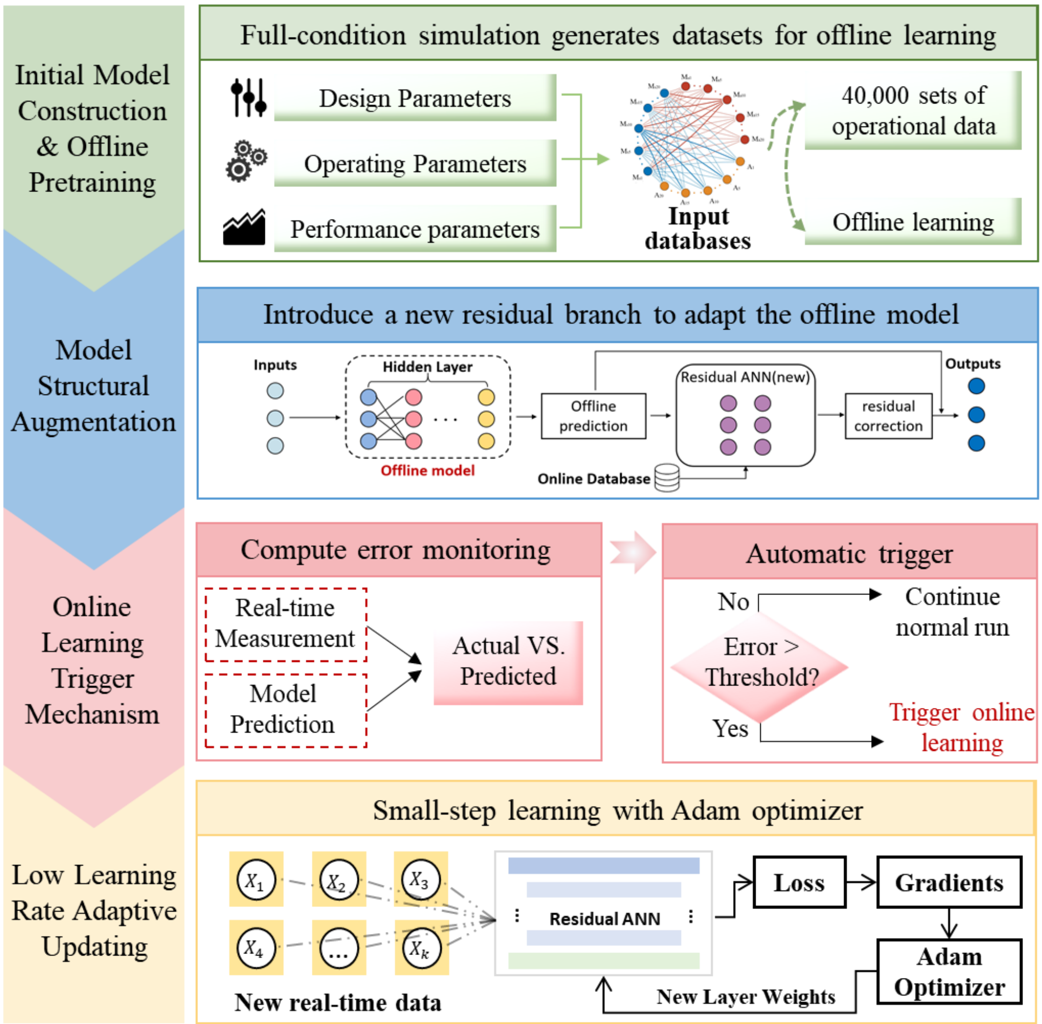


Figure 6.1 Overview of the hybrid model-based online learning strategy

6.2.2 Mechanism of the incremental online learning

The introduction of the Artificial Neural Network (ANN) model significantly lowers the threshold for online training, enabling efficient updates of model weights and biases using real-time operational data. A typical ANN model consists of an input layer, an output layer, and multiple hidden layers. The following formula can describe the relationship between the input and output of a neuron:

$$Y = f\left(\sum_{i=1}^{n} Z_i x_i + b\right) = f(x;\theta) \tag{6.1}$$

where, Z_i represents the weights associated with each input feature x_i, b is the bias term, and f is the activation function. The parameter set θ encompasses all weights and biases in the network.

To mitigate the impact of local data noise and maintain the reliability of physical knowledge, a hybrid model-based online learning strategy is adopted. In this approach, the offline-trained ANN model $f_{offline}(x;\theta_s)$ is preserved, and an additional residual branch $f_{resdual}(x;\theta_R)$ is introduced. The final output of the progressive network is computed as:

$$\hat{Y} = f_{offline}(x;\theta_s) + f_{resdual}(x;\theta_R) \tag{6.2}$$

During online learning, the goal is to minimize the loss function L that measures the discrepancy between predicted values $\hat{Y}$ and the actual observations Y:

$$\mathrm{L}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i,\hat{y}_i) \tag{6.3}$$

where n is the number of training samples.

In conventional gradient descent, only the parameters of the residual branch θ_R are updated:

$$\theta_R \leftarrow \theta_R - \eta\nabla_{\theta_R}\mathrm{L}(Y,\hat{Y}) \tag{6.4}$$

where η is the learning rate.

To avoid overfitting, L1 and L2 regularization terms are added, resulting in the regularized loss:

$$L_{reg}\big(Y,\hat{Y}\big) = \frac{1}{n}\sum\nolimits_{i=1}^{n} L(y_i,\hat{y}_i) + \lambda_1\|\theta_R\|_1 + \lambda_2\|\theta_R\|_2^2 \tag{6.5}$$

where λ_1 and λ_2 is the regularization strength for the L1 norm and L2 norm, respectively.

The corresponding gradient update becomes:

$$\theta_R \leftarrow \theta_R - \eta(\nabla_{\theta_R}\frac{1}{n}\sum\nolimits_{i=1}^{n} L(y_i,\hat{y}_i) + \lambda_1 sgn(\theta_R) + 2\lambda_2\theta_R \tag{6.6}$$

In this work, the Adam optimizer with a low learning rate ($\eta = 10^{-5}$) is employed instead of standard gradient descent to enhance learning stability and convergence further. Unlike traditional gradient descent, Adam maintains adaptive estimates of both the first-order moments (mean) and second-order moments (uncentered variance) of the gradients, which allows for dynamic adjustment of learning rates for each parameter. The Adam updates for θ_R are:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t \tag{6.7}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2 \tag{6.8}$$

$$\widehat{m}_t = m_t/(1-\beta_1^t),\ \hat{v}_t = v_t/(1-\beta_2^t) \tag{6.9}$$

$$\theta_R \leftarrow \theta_R - \eta\,\widehat{m}_t/(\sqrt{\hat{v}_t} + \epsilon) \tag{6.10}$$

where g_t is the gradient of the loss with respect to θ_R at iteration t; m_t and v_t represent the exponentially decaying averages of the gradient and its squared value; and β_1 , β_2 , and ϵ are hyperparameters of the Adam algorithm.

By employing a progressive residual learning strategy and Adam optimization with a low learning rate, the online learning process achieves small, stable updates in the residual branch. This mitigates catastrophic forgetting and maintains robust performance under dynamic and uncertain operational conditions.

6.2.3 Description of the online optimal control strategy

In this study, an incremental online learning strategy is integrated into the optimal control process to enhance real-time adaptability and maintain optimization performance under equipment degradation. As illustrated in Figure 6.2, the smart control station collects real-time operational data from the cooling system, which is then stored in a dedicated database. The hybrid model introduced in Section 4.2.2 is employed in this section's optimization algorithm. The AI algorithm continuously reads the updated data and performs online learning to update the ANN models in the hybrid model.

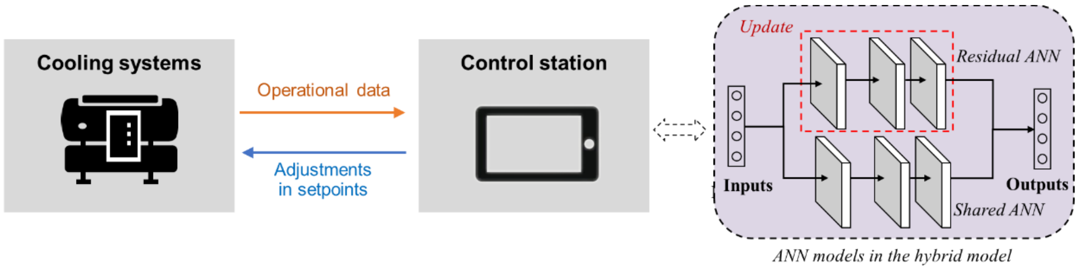


Figure 6.2 Architecture of the residual ANN-based online learning framework

The updated model serves as the foundation for optimization. A hybrid model-based Genetic Algorithm (GA) is employed to determine optimal set points that drive the system towards energy-efficient operation. In particular, the cooling water supply temperature (T_{cws}) and cooling water return temperature (T_{cwr}) are selected as optimization variables. The seawater pump speed is modulated to maintain T_{cws} at its set point using Proportional-Integral (PI) control, while the cooling water pump speed is similarly regulated to achieve the T_{cwr} set point.

The optimization objective is defined as minimizing the total energy consumption of the seawater pumps, cooling water pumps, and chillers, while maintaining appropriate chilled water supply temperatures. The objective function is formulated as follows:

$$min\ W_{sum} = min\left(W_{sw} + W_{cw} + W_{ch} + k \cdot \left(T_{chws} - T_{chws,set}\right)^2\right), \qquad (6.11)$$

where, W_{sum} is the total energy consumption. The subscripts $sw, cw,$ and ch represent the seawater pumps, cooling water pumps, and chillers, respectively. The last term of the equation is the penalty for the chilled water supply temperature. The k indicates the penalty factor, T is the temperature, and the subscripts chw and set represent the chilled water and set value.

6.3 Test platform and arrangement

To comprehensively evaluate the real-time feasibility, responsiveness, and robustness of the proposed AI-enabled optimization strategy, a hardware-in-the-loop (HIL) simulation platform was constructed. This platform establishes a closed-loop testing environment between a high-fidelity virtual cooling system and a physical smart control station. Simulating dynamic operational conditions enables rigorous assessment of the algorithm's behavior in near-real-world scenarios. The overall platform configuration builds upon the system components and architectural details introduced in earlier chapters, with additional implementation-specific arrangements outlined below.

6.3.1 Test procedure

The HIL framework employed here mirrors the setup previously introduced in Section 3.3.2. As shown in Figure 6.3, a high-resolution digital twin of the seawater-cooled chiller system—developed in Dymola and exported as a Functional Mock-up Interface (FMI)—is deployed on a host computer. This model replicates the dynamic behavior of the building energy system under varying external and internal loads.

A physical Jetson Orin™ Nano-based smart control station is integrated into this loop via TCP/IP communication. During testing, the control station continuously receives real-time system state information and responds by computing and dispatching optimized control setpoints every 10 minutes. This iterative closed-loop interaction emulates actual supervisory control workflows and allows for systematic evaluation of control performance under dynamic and uncertain conditions.

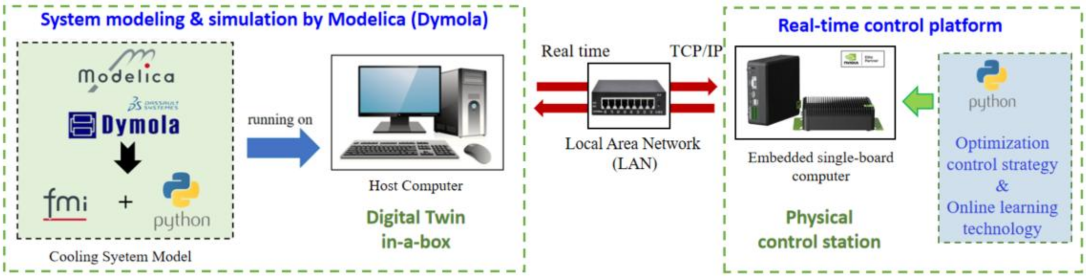


Figure 6.3 Outline of the hardware-in-the-loop test

6.3.2 Description of the reference building and cooling system

The structural and operational configurations of the reference building and its central cooling system—comprising tower zoning, chiller grouping, pump arrangements, and load characteristics—have been comprehensively detailed in Section 5.3.1. For this test, the same building and system architecture are adopted to ensure consistency in simulation fidelity and comparative performance evaluation.

6.3.3 Proposed smart control station

The Jetson Orin™ smart control station used in this section is identical to the one described in Section 3.2. With up to 40 TOPS of AI inference performance, it provides ample computational capacity to support hybrid model-based learning algorithms in real time. Integrated into the HIL environment via a local area network (LAN), the station enables low-latency, bidirectional communication with the virtual chiller plant. Its parallel processing capability and rapid response make it well-suited for field-level deployment, delivering localized intelligent control for complex HVAC systems.

Figure 6.4 Photo of the smart control station

6.3.4 Test scheme

To evaluate the efficacy of a hybrid model-based online learning strategy in refining the optimization algorithm, this study conducts a comparative analysis of the algorithm’s performance before and after modification against a baseline using fixed setpoints, under both design conditions and model degradation conditions, as illustrated in Figure 6.5.

Scheme A involves using the original hybrid model for optimal control and comparing it with the basic control method under design conditions. This comparison establishes a benchmark for assessing the effectiveness of the hybrid model in typical operating scenarios. In Scheme B, the performance of the heat exchanger and chiller in the chiller system model is reduced by 20%, which challenges the accuracy of the previous hybrid model in describing the actual power consumption of the system. The original hybrid model is still used for optimal control to explore the impacts of model inaccuracies on optimization efficiency. Scheme C implements the hybrid model-based online learning strategy to modify the hybrid model using 2 hours of operational data. The system, with deliberately degraded performance, is optimized and controlled again to evaluate the effectiveness of the adaptive model adjustments.

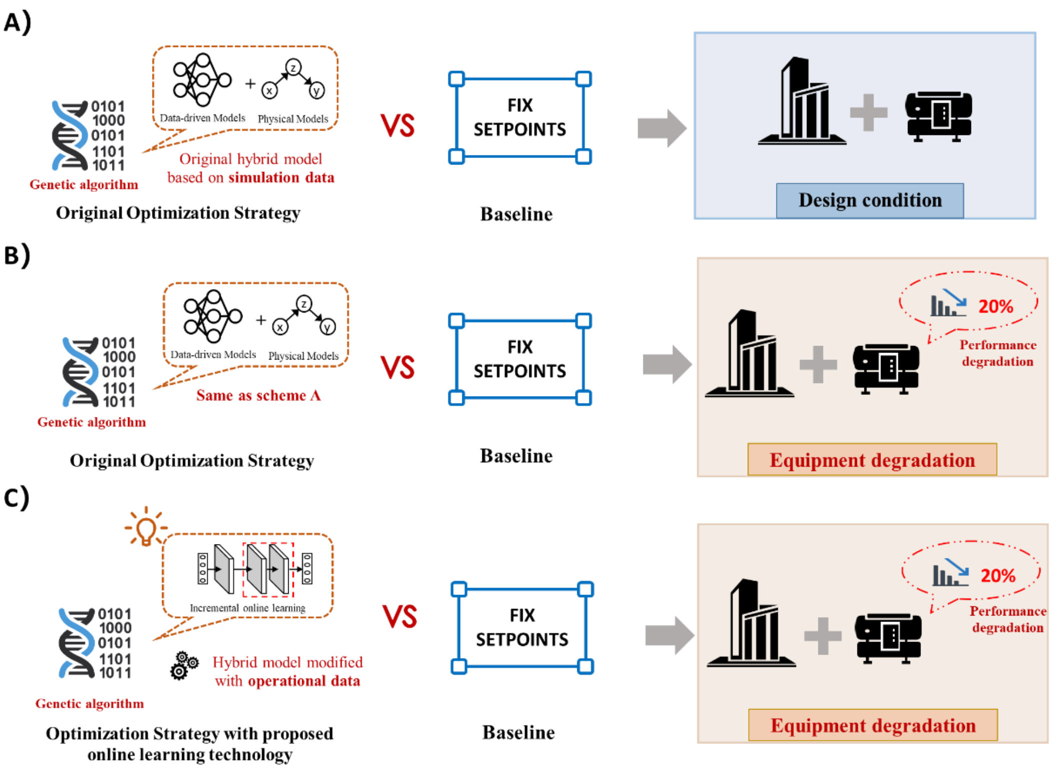


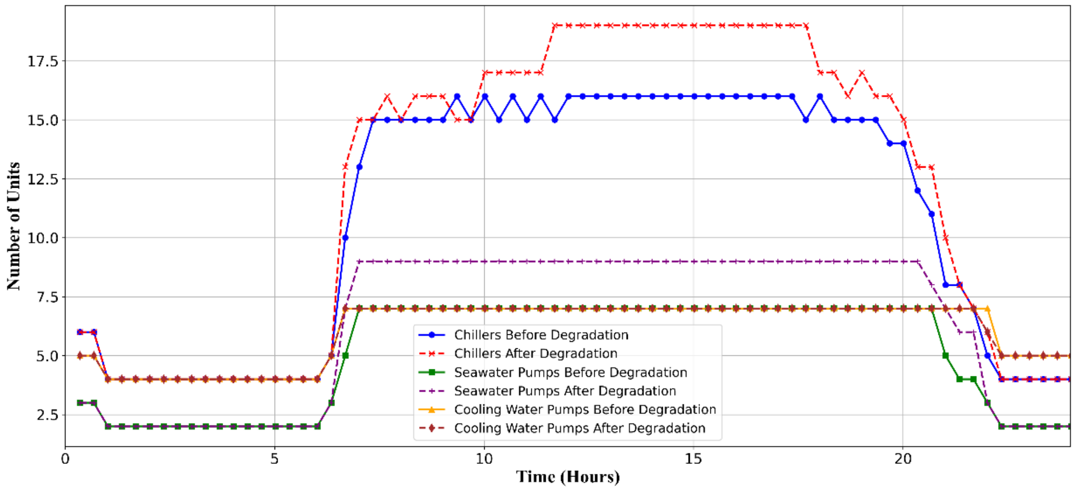
Figure 6.5 Comparison of testing schemes, including: A) the original hybrid model under design conditions, B) the original hybrid model under equipment degradation, and C) the hybrid model modified through incremental online learning under equipment degradation

Furthermore, to rigorously assess the robustness and adaptability of the proposed incremental online learning strategy, additional experiments were conducted on the heat exchanger model. Both the proposed strategy and a conventional online learning strategy—which directly updates the offline-trained model via standard gradient descent—were tested for comparison. Each method underwent a sequence of 15 incremental online learning blocks, with each block using only 100 samples to simulate the limited data availability typical of real-time industrial operations. To systematically examine resilience to data quality issues, controlled random noise ranging from ±0.5% to ±2.5% was deliberately introduced into both input and output data during blocks 5, 6, and 7. This experimental setup was designed to replicate potential anomalies or measurement fluctuations encountered in practical scenarios, thereby evaluating each method's capacity to maintain prediction accuracy and avoid catastrophic forgetting under noisy conditions.

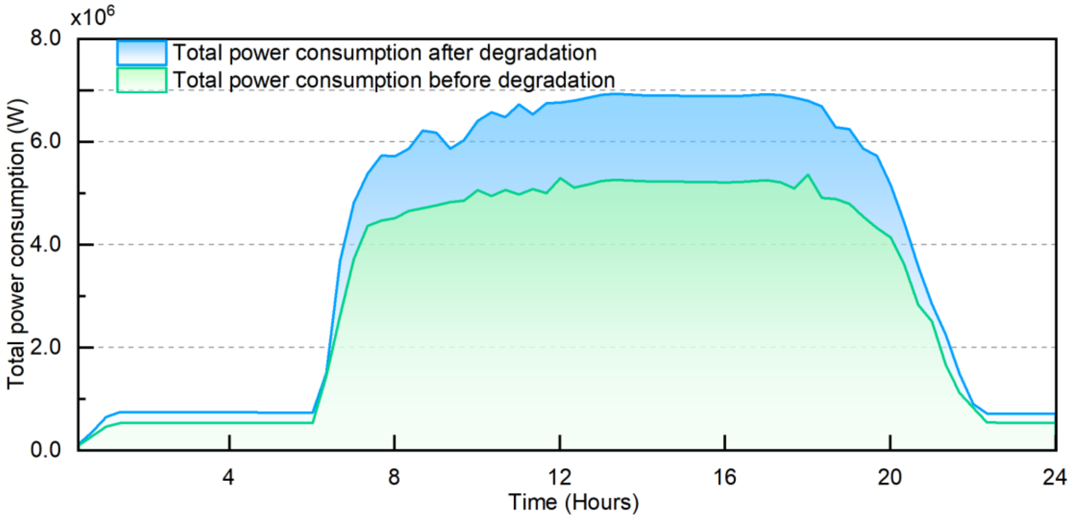
6.4 Test results and comparative analysis

6.4.1 Equipment degradation impact on system operations

This section analyzes the effects of system performance degradation on efficiency via a comparative test that deliberately lowers efficiency by 20%. The operational states before and after the degradation under the basic operational scheme are compared, as illustrated in Figure 6.6. Due to the performance decline of the heat exchanger, an increased volume of seawater was required for cooling, leading to a rise in the number of seawater pumps. And the operational units of chillers also increased due to reduced cooling capacity. Ultimately, the total power consumption, including energy consumption by seawater pumps and chillers, increased by 12%.



(a) Operation number of units



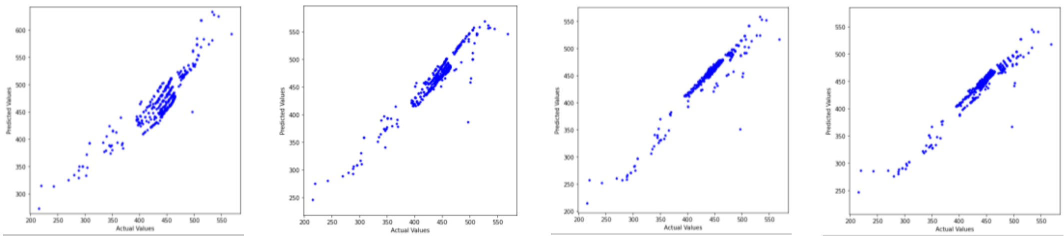
(b) Total power consumption

Figure 6.6 Operational status before and after system degradation

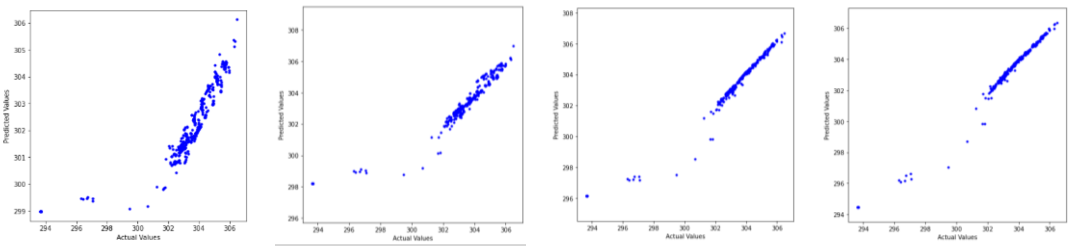
6.4.2 Model deviation and modification effects on optimal control

Figure 6.7 illustrates the progressive evolution of model performance under a 20% equipment performance degradation scenario, demonstrating the effectiveness of the incremental online learning strategy proposed in this study. The figure presents the sequential online learning process for both the chiller and the heat exchanger models. In each scatter plot, the horizontal and vertical axes represent the actual measured values and the model-predicted values, respectively.

As shown in Figure 6.7 (a) and (b), the initial offline-trained models exhibited limited prediction accuracy under degraded conditions, with R^2 values of 0.34 for the chiller's energy consumption and 0.40 for the heat exchanger's temperature output. These results indicate that the original models could not adequately capture the altered system dynamics resulting from equipment deterioration. Through the introduction of a new residual branch into the ANN architecture, coupled with incremental online updates using real-time operational data, the model's regression performance was significantly enhanced. The final R^2 values surpassed 0.9, confirming that the proposed online learning approach successfully compensates for system degradation and preserves modeling accuracy.



(a) Energy consumption of the chillers (unit: kW, R^2 = 0.34, 0.46, 0.89, 0.93)



(b) Temperature output of heat exchanger (unit: K, R^2: 0.40, 0.81, 0.94, 0.98)

Figure 6.7 Model regression performance in the online learning process

In the test arrangements described in Section 6.3.4, the performance of the seawater-source chiller unit under five distinct operational scenarios was assessed:

a. Optimal control under design conditions

b. Basic control under design conditions

c. Basic control with a performance deviation of 20%

d. Optimal control with a performance deviation of 20%

e. Optimal control with model correction

The energy consumption and Energy Efficiency Ratio (EER) under various test conditions are illustrated in Figure 6.8. The results show that the five evaluated schemes can be categorized into two groups: those operating under design conditions and those under application conditions involving a 20% model deviation to simulate performance degradation.

Under design conditions, scheme b (green line), which employs a genetic algorithm for optimization, demonstrates a clear energy-saving advantage compared to the traditional control strategy represented by scheme a (red line). However, when the system performance deteriorates, overall energy consumption increases across all strategies due to shifts in the operating characteristics of the chiller and pump systems. In these degraded conditions, scheme d (dark blue line), which continues to use the original optimization model without adaptation, achieves only a marginal energy saving of 1.12% relative to the traditional scheme c (purple line). This limited improvement highlights that the optimization potential cannot be fully exploited when discrepancies exist between the model and the actual system behavior.

By incorporating the proposed online learning mechanism for model correction, scheme e (light blue line) can adaptively update the optimization model in response to real-time system changes. As a result, it achieves a 5.77% reduction in energy consumption, partially recovering the optimization benefits lost due to degradation. Nevertheless, the energy savings remain lower than the 7.8% observed under design conditions, highlighting the influence of equipment degradation on optimization effectiveness.

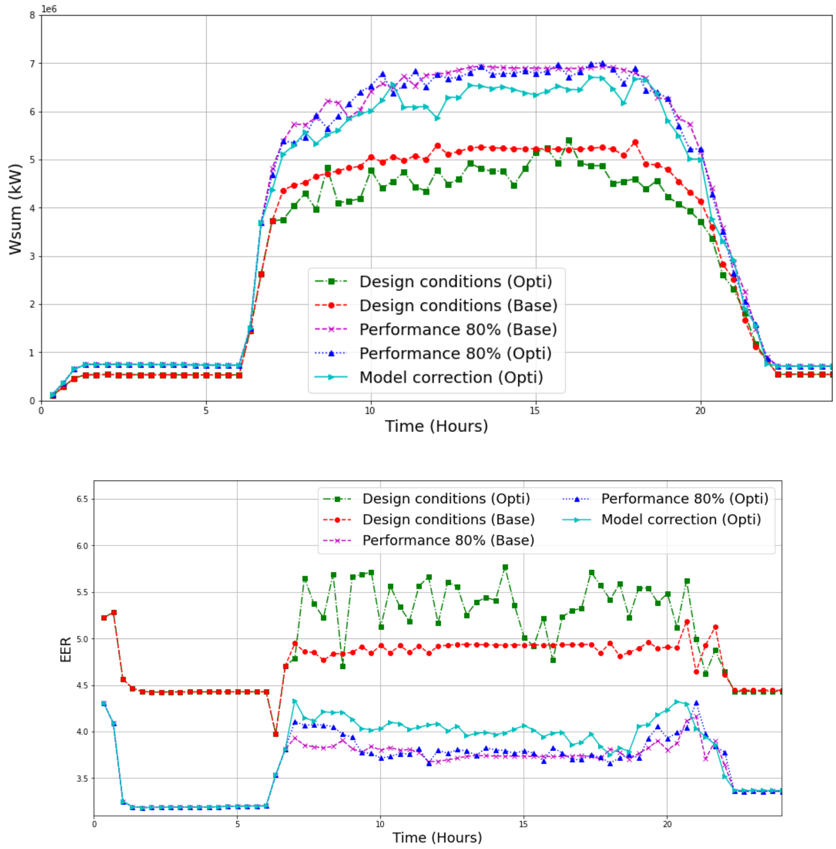


Figure 6.8 Energy consumption and performance under various test conditions

Figure 6.9 presents a time-sliced analysis of power consumption at three key operating hours—10:00, 15:00, and 20:00—offering deeper insights into the system's dynamic behavior under different optimization schemes and conditions.

At 10:00, when overall loads are moderate, performance degradation (scenarios c and d) results in noticeably higher power consumption, particularly in the seawater pump segment (Wswpump). This suggests that model inaccuracies impact components sensitive to flow regulation, such as pumps. By 15:00, during peak load conditions, differences among the schemes become even more pronounced. Under degraded performance, both the baseline and uncorrected optimization strategies (c and d) exhibit elevated total power consumption, mainly due to increased demands from both chillers and cooling water pumps (Wchiller and Wcwpump). In contrast, the corrected optimization (scheme e) demonstrates its effectiveness by significantly reducing pump-related energy consumption. This underscores the capability of online model correction to restore optimal operational patterns during periods of high demand. At 20:00, as system loads decrease, overall differences among the schemes narrow. The corrected optimization strategy (e) continues to maintain lower pump power consumption than the degraded cases (c and d), indicating that even during off-peak hours, online learning can fine-tune operations to preserve efficiency.

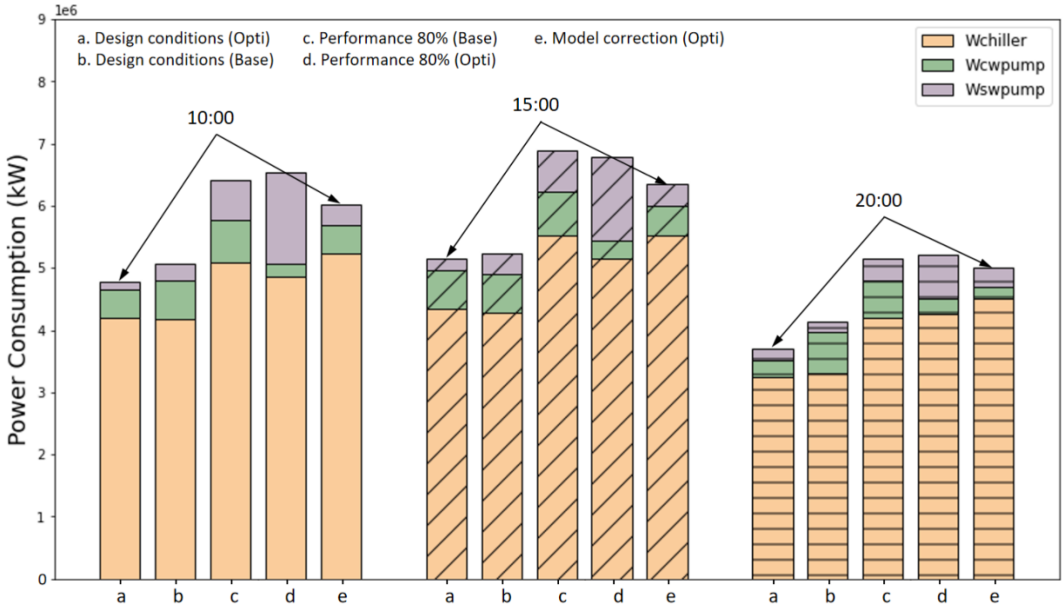


Figure 6.9 Detailed comparison of power consumption across three time points

Figure 6.10 provides a comprehensive visualization of the interplay between operational conditions and system energy consumption throughout a typical day. The figure sequentially arranges five operational scenarios—from optimal design conditions (left) to severe performance degradation and subsequent model correction (right)—thus capturing the full spectrum of system states from ideal to degraded and corrected operation.

In the three-dimensional plot, the bar heights represent total power consumption across time, while color gradients encode the magnitude of energy usage, ranging from lower consumption (blue-green) to higher consumption levels (red). Notably, the darkest red regions concentrate in the scenarios representing performance degradation without correction, clearly illustrating how system aging and model mismatches lead to elevated energy demands. Conversely, under design conditions and after applying model correction strategies, the bars shift toward lighter colors and lower heights, signifying substantial reductions in energy consumption. This visual evidence underscores two critical insights: first, while AI-empowered optimization can achieve significant savings under ideal conditions, its effectiveness diminishes sharply when the system drifts away from its modeled performance envelope. Second—and more importantly—the introduction of online model correction mechanisms successfully restores a significant portion of the lost optimization potential, even under degraded operating conditions. Moreover, the lower subfigure in Figure 6.10 succinctly quantifies these trends by summarizing the total daily energy consumption for each scenario. It shows that the energy-saving curve drops sharply under degradation but climbs back toward optimal levels following model correction. This demonstrates that continuous model adaptation through online learning is essential for sustaining long-term energy efficiency.

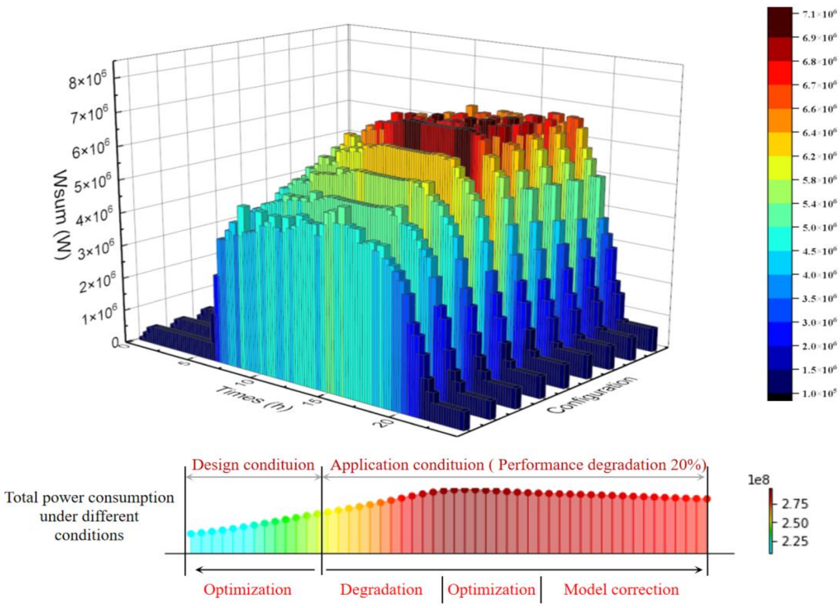
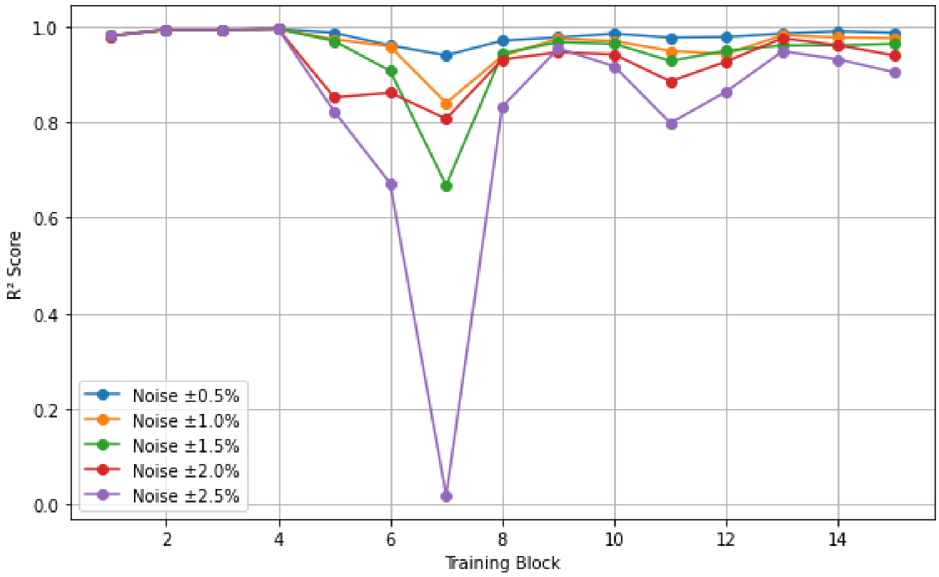


Figure 6.10 Relationship between operational conditions and energy consumption

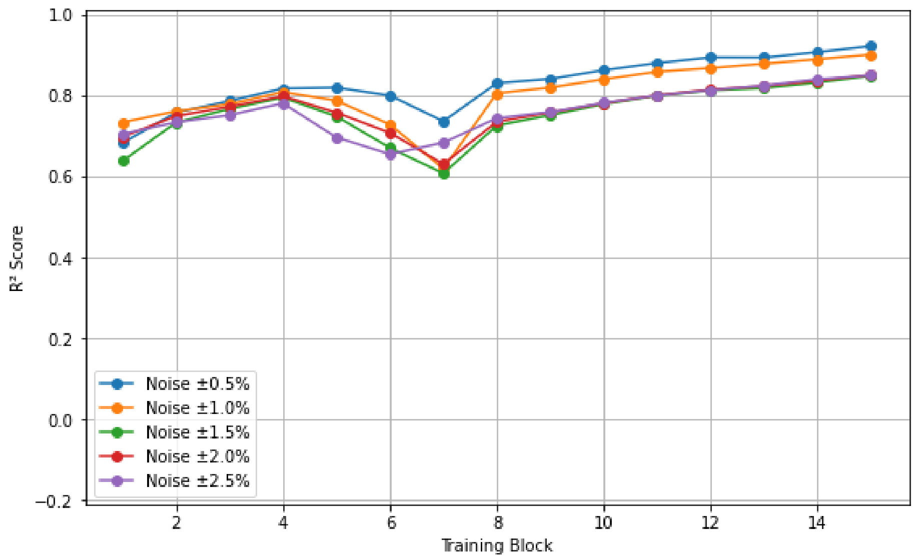
6.4.3 Comparative performance under noisy online learning

Figure 6.11 illustrates the comparative learning performance under random noise disturbances. The conventional method, which employs a higher learning rate and direct updates to the offline model, achieved rapid initial data fitting, with R^2 values nearing 0.98 in the early stages. However, this swift adaptation incurred significant drawbacks: the model exhibited pronounced sensitivity to the injected noise. Specifically, during blocks 5, 6, and 7, severe overfitting occurred, evidenced by a sharp drop in R2. Notably, under the highest noise level (±2.5%), the R^2 value plummeted to as low as 0.02 in block 6. Furthermore, the model experienced catastrophic forgetting, leading to cyclical performance degradation in subsequent learning blocks—for instance, a noticeable decline in R^2 around block 11, highlighting instability in retaining previously learned knowledge.

In contrast, the proposed incremental online learning strategy demonstrated markedly greater robustness. While it experienced a moderate decline in R^2 during the noisy blocks (5 to 7), the degradation was substantially less severe than in the conventional method. Even under the highest noise level of ±2.5%, the proposed strategy maintained R^2 values no lower than 0.60. More importantly, the residual strategy maintained steady performance improvements across subsequent blocks, avoiding the oscillatory patterns observed in the traditional approach. This indicates that the residual framework effectively isolates the influence of noisy data, prevents catastrophic forgetting, and preserves the core knowledge embedded in the offline-trained model. The progressive and conservative nature of the updates enables the model to remain resilient even under challenging, noisy conditions.



(a) Conventional online learning strategy



(b) Proposed online learning strategy

Figure 6.11 Comparison of learning performance under random noise disturbances

6.5 Summary

This study proposed an incremental online learning strategy to mitigate the performance degradation challenges commonly encountered in AI-enabled optimal control of building cooling systems. The strategy was implemented on a real smart control station and validated through hardware-in-the-loop testing. The key findings are summarized as follows:

- Model degradation significantly impacts optimization efficiency. The optimal control strategy based on the initial model achieved significant energy savings of approximately 7.8% under design conditions. However, a 20% degradation in the chiller system's performance led to notable shifts in the operating behavior of both chillers and pumps, resulting in a 12% increase in total energy consumption. Under these degraded conditions, the original optimization algorithm struggled to adapt, causing the optimization efficiency to decline sharply to just 1.12%.

- The proposed incremental online learning strategy successfully mitigated the adverse impacts of equipment degradation. By continuously updating the model with real-time operational data, the strategy improved the predictive accuracy of the optimization model, raising its R^2 value from 0.40 to 0.98. This enhanced modeling capability enabled the system to achieve 5.77% energy savings even under conditions of significant performance degradation.
- The incremental online learning strategy demonstrates superior robustness under noisy operational conditions. Test results indicate that the conventional method is prone to overfitting and catastrophic forgetting, with R^2 values dropping as low as 0.02 under ±2.5% disturbances. In contrast, the proposed strategy maintains stable performance even at the highest noise level, preserving R^2 values above 0.60 and avoiding catastrophic forgetting.

In conclusion, the results demonstrate that the proposed incremental online learning strategy holds substantial promise for maintaining optimization effectiveness in complex, dynamic systems subject to aging and operational drift. This work underscores the importance of integrating adaptive learning mechanisms into AI-driven optimization frameworks to ensure sustained energy efficiency, reliability, and operational resilience in real-world applications.

CHAPTER 7 A PARALLEL DISTRIBUTED AI-ENABLED OPTIMAL CONTROL STRATEGY OF ENHANCED EFFICIENCY, RELIABILITY AND SCALABILITY

This chapter presents a novel parallel distributed approach for AI-enabled online optimization of building cooling systems, aiming to enhance operational efficiency, reliability, and scalability in practical applications. Traditional centralized optimization frameworks, although effective in algorithmic performance, face significant limitations in large-scale systems due to computational bottlenecks, communication latency, and risks of single points of failure. Hierarchical distributed architectures partially alleviate these issues but often depend on central coordinators, thereby retaining vulnerability and limiting true scalability. To address these challenges, this study proposes a fully decentralized framework in which multiple autonomous agents operate in parallel, each equipped with AI-based optimization capabilities and interconnected through peer-to-peer communication and consensus mechanisms. This approach allows distributed computational tasks, dynamic load balancing, and resilient system operation even under agent failures or network disruptions. The following sections detail the theoretical principles, algorithmic architecture, system integration, and experimental validation underpinning this parallel distributed optimization framework.

7.1 Conventional centralized and hierarchical distributed control methods

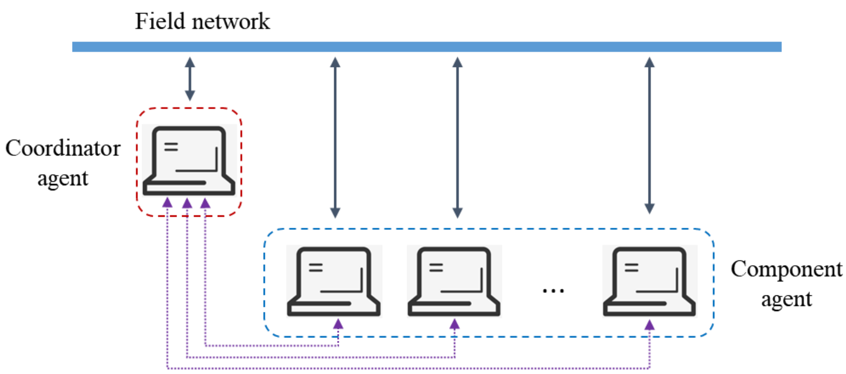
As reviewed in Chapter 2, most AI-enabled optimization strategies for building cooling systems have been implemented within centralized or hierarchical distributed architectures. Centralized approaches leverage powerful computing platforms and global visibility, while hierarchical distributed methods alleviate computational burden by decomposing subproblems and delegating them to local agents. Despite these improvements, both paradigms remain inherently constrained by their reliance on a central coordinating entity. This reliance introduces several critical challenges, including communication bottlenecks, vulnerability to coordinator failures, and limited scalability in large, heterogeneous building environments. These drawbacks become especially problematic in real-time AI optimization scenarios, where rapid responsiveness and resilience to disturbances are essential.

To overcome these limitations, there is a clear need to advance beyond partially centralized frameworks toward fully distributed control strategies. By eliminating dependence on a supervisory node, fully distributed approaches enable peer-to-peer collaboration among autonomous agents, thereby improving scalability, robustness, and adaptability of optimization in practical building cooling applications.

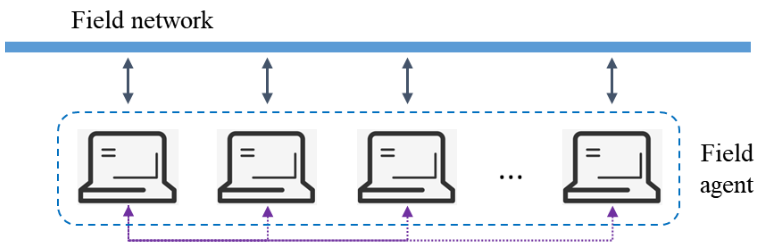
7.2 Proposed parallel distributed AI-enabled optimal control strategy

7.2.1 Overview of the parallel distributed architecture

Figure 7.1 illustrates the frameworks of the conventional hierarchical distributed control strategy and the proposed parallel distributed control strategy. The key difference between the two strategies lies in the decomposition approach. In the conventional strategy, the system is decomposed based on its physical structure, with individual agents responsible for local optimization of specific components. In this case, a coordinator agent is needed for information collection and global optimization, which introduces the risk of single-point failures and limits system scalability.



(a) Hierarchical control strategy (partially distributed)



(b) Proposed parallel distributed control strategy

Figure 7.1 Frameworks of the proposed parallel distributed control strategy and the hierarchical control strategy

To achieve fully distributed parallel computing, the proposed strategy decomposes computational tasks at the algorithmic level rather than based on physical subsystem boundaries, allowing equivalent task deployment across all agents for spatially distributed operation. Specifically, a genetic algorithm (GA) is embedded within each agent to perform local optimization independently. The key parameters of the GA—namely the population size and the number of generations—are adaptively configured based on the number of active agents currently participating in the optimization process. This dynamic task partitioning mechanism ensures balanced computational load distribution and enables real-time scalability across all agents.

During each optimization cycle, agents sequentially broadcast their IP addresses to signal their activation status, allowing all participating nodes to detect peer availability autonomously. Each agent runs an identical optimization algorithm and engages in decentralized coordination via peer-to-peer communication. Throughout the process, agents exchange elite individuals from the GA at designated intervals, facilitating inter-agent knowledge sharing and accelerating overall convergence. Final control decisions are determined through a consensus-based mechanism, in which each agent evaluates and votes on candidate solutions received from its peers. By integrating local computations with shared optimization outcomes, the system can reach a global decision without requiring centralized coordination.

7.2.2 Description of the proposed optimal control strategy

To achieve parallel cooperative optimization across multiple agents, a decentralized operational framework and a structured communication mechanism are established. This section provides a detailed description of the agent configuration and the distributed optimal control process.

7.2.2.1 Configurations of individual agents

For a single agent, the basic functions are acquiring real-time measurement data and sending optimization results to the cooling system. Within the fully decentralized operational framework, agents are further required to broadcast their activation states, synchronize optimization progress, and collaboratively achieve consensus. Figure 7.2 illustrates three communication mechanisms used to achieve the above functions. A single agent adopts three different protocols for different communication purposes. First, the BACnet/IP protocol is employed to establish communication between the agent and the cooling system. The IP addresses of all agents are predefined and stored in a list, allowing each agent to retrieve operational data from the cooling system via its designated port. Secondly, the UDP/IP protocol is used to synchronize activation states among agents through broadcast announcements. Leveraging the connectionless nature of UDP, each agent can broadcast its activation message to all potential peers. By receiving these broadcasts, agents can determine the number and identity of active agents currently participating in the system. Finally, TCP-based connections are established among all active agents based on the known IP addresses. This enables reliable, multi-point point-to-point communications for exchanging optimization data and facilitates the consensus process for selecting the final control decision.

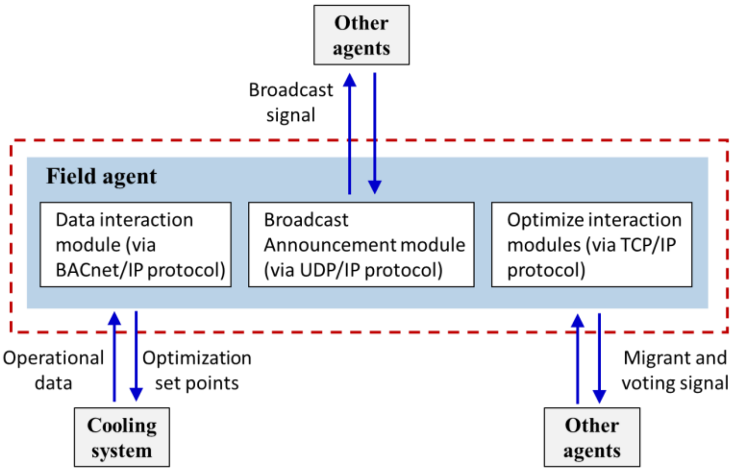


Figure 7.2 Communication frameworks in a single agent

7.2.2.2 Operation procedure of the parallel distributed optimal control strategy

Figure 7.3 illustrates the key steps in the optimization process of the parallel distributed control strategy. This strategy comprises five main steps: broadcast announcement, optimization initialization, genetic algorithm computation, elite migration among agents, and multi-agent consensus. The purpose and process of each step are detailed below.

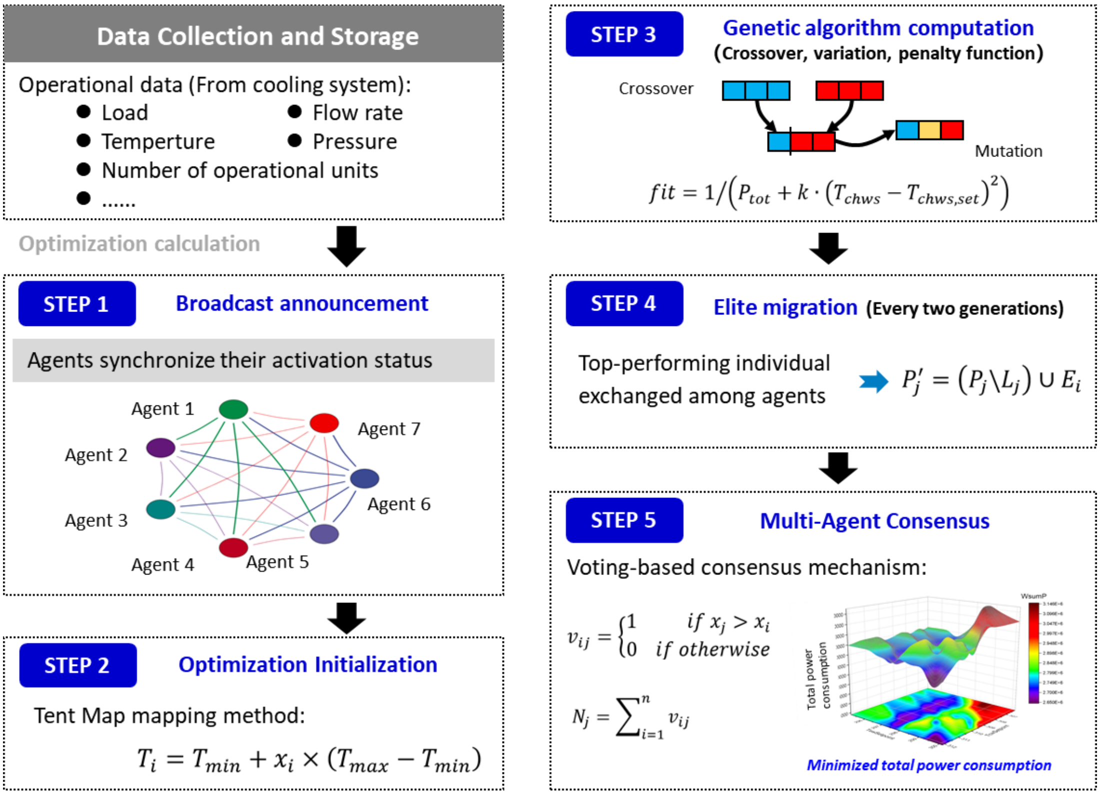


Figure 7.3 Process and major steps of the parallel distributed optimal control strategy

Broadcast announcement

At the start of each optimal control interval, every agent remains inactive while attempting to acquire real-time operational data from the cooling system. After confirming the completeness of the acquired data, the agent transitions into an active state. It then employs a broadcasting protocol (UDP/IP) to announce its activation state to all other agents in the network by sending an activation signal, denoted as S_i. The structure of this signal can be expressed as follows:

$$S_i = \text{"}Agent\text{"}\ i : \{IP_i\}\ \text{"}is\ active\ at\ time\text{"}\ t \tag{7.1}$$

Optimization Initialization

The genetic algorithm (GA) is embedded in each agent to perform optimization independently. After receiving the activation signals S_i from other agents, each agent automatically determines the initial population size p for the genetic algorithm (GA) based on the total number of active agents n. The Tent Map chaotic mapping method is employed to enhance the diversity of the initial population. Specifically, the Tent Map parameter is set to r=2 for each agent, and the initial value x_0 is randomly selected from the interval [0, 1].

Each agent independently runs the Tent Map to generate a sequence $X = \{x_1, x_2, \dots, x_p\}$. Where:

$$x_{k+1} = r \cdot x_k \;\; if \; x_k < 0.5; \;\; x_{k+1} = r \cdot (1 - x_k) \;\; if \; x_k \geq 0.5 \tag{7.2}$$

This sequence X is then mapped into the ranges of the setpoints to determine the initial values for the cooling water supply temperature (T_{cws}) and the cooling water return temperature (T_{cwr}). The mapping is expressed as:

$$T_{cws} = T_{cws,min} + X \cdot \left(T_{cws,max} - T_{cws,min}\right), \tag{7.3}$$

$$T_{cwr} = T_{cwr,min} + X \cdot \left(T_{cwr,max} - T_{cwr,min}\right), \tag{7.4}$$

Genetic algorithm computation

During the GA computation process, each agent functions as an independent "island", evolving its population locally while communicating with other agents at specific intervals to accelerate the optimization process.

The initialization process generates a population of individuals, each representing a candidate solution to the optimization problem. These individuals are evaluated using a fitness function that quantifies how closely each solution approximates the optimal objective. Based on the fitness scores, a selection mechanism is applied to retain the most promising individuals for reproduction. Selected individuals undergo genetic operations—namely, crossover and mutation—to create the next generation. In this study, the crossover and mutation probabilities are set to 0.8 and 0.08, respectively, to maintain an effective balance between exploration of the search space and exploitation of high-quality solutions.

Elite migration among agents

After every two generations of the genetic algorithm, the top-performing individual from each sub-population (i.e., each agent) is selected based on its fitness value and shared with other agents. This elite migration mechanism facilitates inter-agent knowledge exchange and accelerates global convergence by disseminating high-quality solutions throughout the distributed population. More importantly, this mechanism allows an agent to accelerate its optimization process by adopting better solutions from other agents, thereby fully leveraging the computational capacity of all active agents. Through continuous elite migration and peer-to-peer exchange, high-quality candidate solutions are disseminated across the network, accelerating convergence and enhancing overall solution quality. Consequently, the distributed framework not only enables robust cooperative optimization but also significantly improves computational efficiency by parallelizing workloads in a hardware-aware manner. In essence, a greater number of active agents leads to higher computational throughput and faster optimization, highlighting the scalability and efficiency of the proposed architecture.

Let the set of elite individuals from all agents be denoted as E_i. Upon receiving elite individuals from peers, each agent j updates its local population P_j by replacing its least fit individuals with members of E_i. A re-selection process is then performed to ensure the population size remains constant. This strategy maintains genetic diversity while enhancing the overall optimization performance through cooperative evolution across agents. Where L_j is the set of the least fit individuals in P_j that are replaced by incoming elites from E_i.

$$P_j' = \left(P_j \backslash L_j\right) \cup E_i \tag{7.5}$$

<u>*Multi-Agent Consensus*</u>

To achieve a consistent control decision in a fully decentralized environment, this study introduces a voting-based consensus mechanism among multiple agents. In this process, each agent sends its locally computed fitness value to all other active agents in the network. Upon receiving fitness values from peers, each agent compares them with its result and casts a vote of approval or disapproval accordingly. As defined in Eq. (7.6), where agent i receives a candidate result x_j from agent j, it evaluates whether x_j is superior to its computed value x_i, and assigns a binary vote accordingly.

$$v_{ij} = \begin{cases} 1 & if\ x_j > x_i \\ 0 & if\ otherwise \end{cases} \tag{7.6}$$

The resulting voting matrix V=[v_{ij}] represents the voting relationships among all agents, where v_{ij} indicates whether agent i approves the result from agent j. After the voting is complete, the total number of votes N_j received by each result can be calculated as:

$$N_j = \sum_{i=1}^{n} v_{ij} \tag{7.7}$$

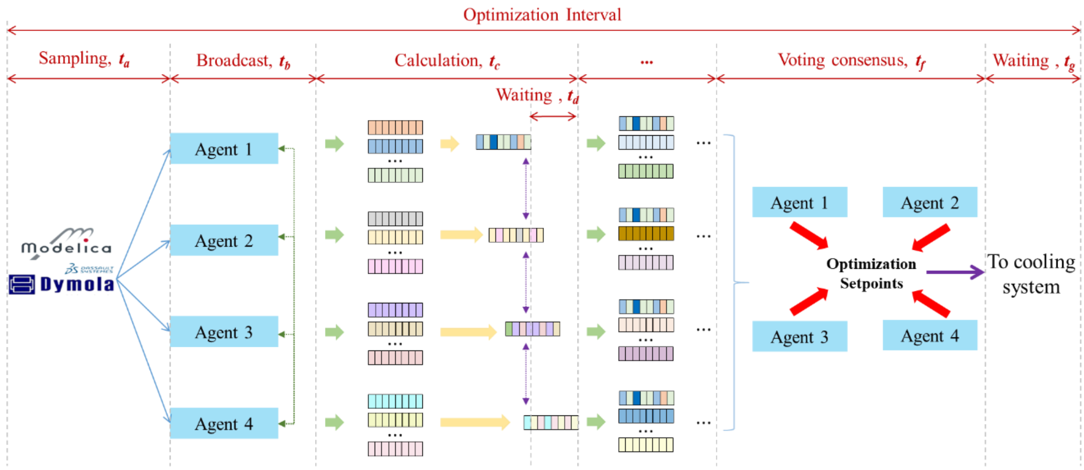
Subsequently, the agent whose solution receives the highest N_j is selected as the consensus result, and its control decision is sent to the cooling system. This voting mechanism ensures the security and consistency of result selection by relying on a majority decision, thereby mitigating the risk of a single agent failure that could cause a decision error.

7.2.2.3 Timing of computation and communication at each optimization interval

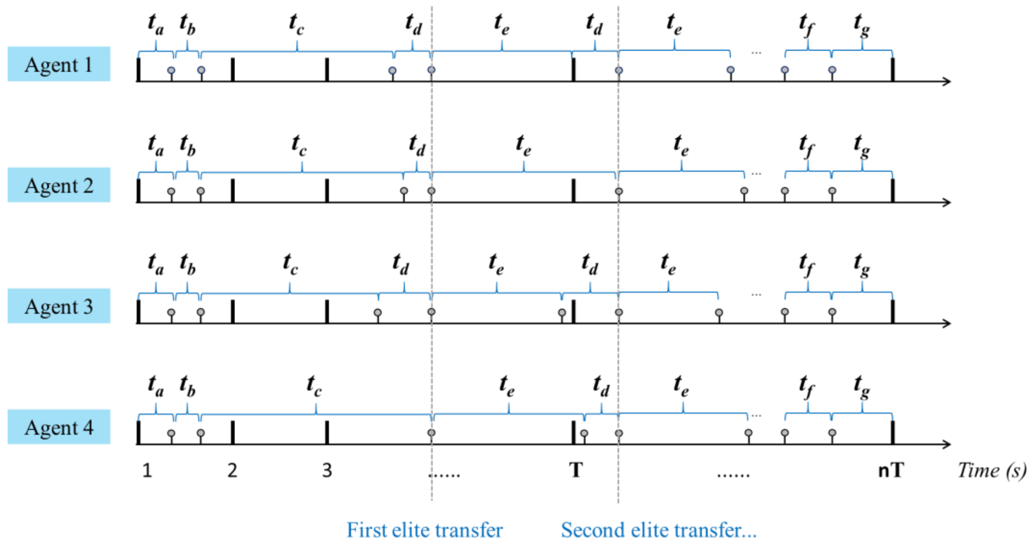
Figure 7.4 illustrates the timing structure of the distributed optimization process among multiple agents. Each agent independently performs its designated tasks while coordinating with others through a waiting mechanism. The process starts with data acquisition (t_a) and broadcast announcement (t_b), during which agents collect operational data from the cooling system and exchange activation signals to identify all active participants.

Subsequently, agents enter the optimization step, where local computation and elite information reception are executed concurrently using a dual-threaded program structure. If an agent completes its computation before others, it enters a waiting state until it receives the elite individuals from all other active agents. The figure depicts two rounds of elite migration, demonstrating how agents synchronize to promote collaborative optimization and accelerate convergence.

Following the optimization and migration phases, the consensus voting mechanism is triggered. In this step, each agent casts votes based on fitness comparisons, and the agent with the highest approval is selected to transmit the final optimization set points to the cooling system. Once the decision is executed, all agents enter a standby mode, awaiting the next optimization cycle.



(a) Distributed optimization process among agents



t_a : Sampling time; t_b : Broadcast time; t_c : Calculation time (Two generation); t_d : Waiting time (Wait for other agents); t_e : Calculation time (With elite); t_f: Voting consensus time; t_g : Waiting time (Wait for the next optimization cycle)

(b) Timing of computation and communication of each agent

Figure 7.4 Timing structure of the distributed optimization process

7.3 Test platform and validation arrangement

7.3.1 Test building and cooling system

The test system employed in this study is based on a large-scale seawater-cooled chiller plant that serves a commercial building complex in West Kowloon, Hong Kong. As detailed previously in Section 4.3.1, the reference building comprises four towers

(T1A, T1B, T2A, T2B) and a podium, with a total gross floor area of 288,010 m² and an estimated peak cooling demand of 55,331 kW. The system includes 21 chillers, comprising 15 high-capacity units (3,517 kW) and 6 medium-capacity units (1,758 kW). The specific chiller distribution across building zones is summarized again in Table 7.1 for ease of reference in the validation context.

Table 7.1 Building chiller system scheme

Building	Chiller scheme
Tower 1A	3×3,517 kW +1×1,758 kW
Tower 1B	3×3,517 kW +1×1,758 kW
Tower 2A	2×3,517 kW +2×1,758 kW
Tower 2B	2×3,517 kW +1×1,758 kW
Podium	5×3,517 kW +1×1,758 kW

To support system-wide optimization, the cooling network is structured into three primary hydraulic loops, each governed by distinct control set-points. The first loop is the seawater loop, which dissipates heat from the cooling water through heat exchangers. In this loop, the cooling water supply temperature (T_{cws}) is maintained at its set point by regulating the speed of the seawater pumps. The second loop is the cooling water loop, which transfers heat from the condenser to the cooling water. The cooling water return temperature (T_{cwr}) is controlled by adjusting the cooling water flow rate to meet its designated set point. The third loop is the chilled water loop, which is responsible for supplying chilled water to the air handling units (AHUs). In this loop, the chilled water supply temperature (T_{chws}) is maintained by the chiller's internal controller and the number of operating chillers. The set-points T_{cws}, T_{cwr}, and T_{chws} significantly affect the power consumption of chillers, chilled water pumps, cooling water pumps, and seawater pumps. It is challenging and necessary to identify the combined optimal set-points of the three parameters to maximize the overall system efficiency in a particular working condition.

To evaluate optimization performance, a fitness function is defined to guide the search for optimal control setpoints. As shown in Eq. (7.8), the fitness function consists of two components: the total power consumption of the cooling system and a penalty term that represents deviations from the desired comfort level.

$$\max_{\{T_{cws}, T_{cwr}, T_{chws}\}} fit = 1/\left(P_{tot} + k \cdot \left(T_{chws} - T_{chws,set}\right)^2\right) \tag{7.8}$$

where, P_{tot} is the total energy consumption, k is the penalty factor.

By maximizing the fitness value, the optimization algorithm effectively balances two objectives: minimizing energy consumption and maintaining thermal comfort. The total energy consumption P_{tot} defined in Eq. (7.9) represents the total power usage of all components of the cooling system:

$$P_{tot} = \sum_{i=1}^{N_{sw}} P_{sw,i} + \sum_{j=1}^{N_{cw}} P_{cw,j} + \sum_{n=1}^{N_{ch}} P_{ch,n} + \sum_{n=1}^{N_{chw}} P_{chw,n} \tag{7.9}$$

where, P_{tot} is the total power consumption. The subscripts $sw, cw, chw,$ and ch represent the seawater pumps, cooling water pumps, chilled water pumps, and chillers, respectively. The N is the number of operation units.

7.3.2 Test platform

To validate the proposed parallel distributed control strategy, the hardware-in-the-loop (HIL) architecture originally introduced in Section 3.3.2 has been extended to support multi-agent coordination. As shown in Figure 7.5, the updated platform retains the core structure of the previously established HIL framework. However, it is enhanced with the deployment of four smart field control stations, enabling distributed optimization across building subsystems.

The virtual cooling system, developed in Dymola using the Modelica language, continues to serve as a high-fidelity digital twin, simulating the full dynamics of a large-scale seawater-cooled HVAC system in real time. Each control station is deployed as an independent optimization agent, responsible for generating local set-points based on its assigned zone's state. All devices communicate over a shared local area network (LAN), ensuring synchronized, low-latency interaction between virtual and physical nodes.

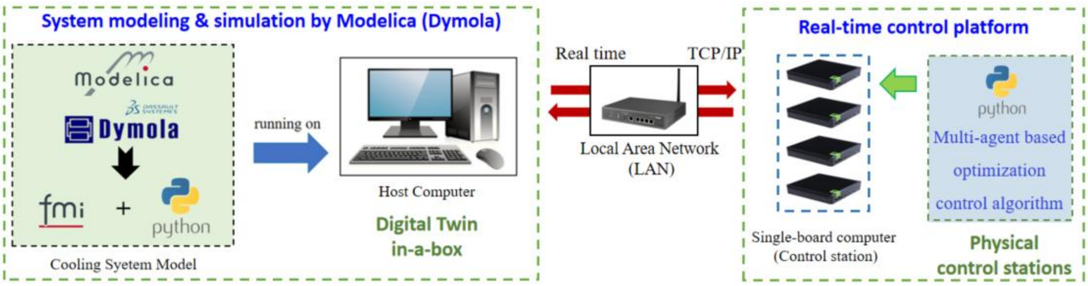


Figure 7.5 Architecture of the hardware-in-the-loop test platform

An overview of the physical setup is shown in Figure 7.6. The host computer runs the virtual plant and distributes real-time state data to the four control agents, each embedded with the previously specified NVIDIA Jetson Orin™ hardware platform (see Section 3.2 for details). The control agents execute parallel optimization routines at regular intervals, returning coordinated set-points to the simulated system for real-time implementation. To facilitate live monitoring, an HDMI screen splitter and portable display are configured to visualize the operation of each agent during testing.

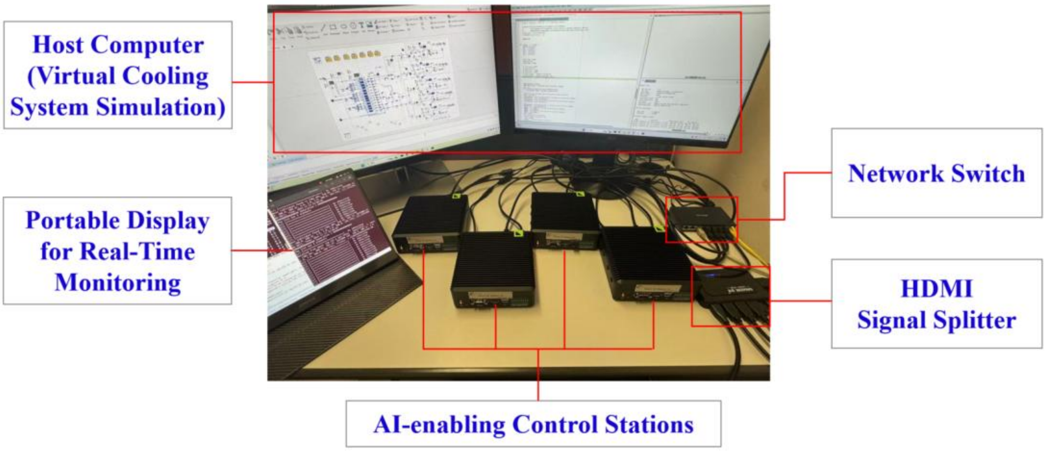


Figure 7.6 Image of the test platform in operation

7.3.3 Validation arrangements

To validate the effectiveness and advantages of the proposed parallel distributed optimal control strategy, three control strategies are tested under the same typical working day for a comparative study:

- *Baseline control strategy:* A conventional control approach commonly applied in practical cooling systems, which does not incorporate any optimization process. Fixed set-points for the cooling water supply temperature and return water temperature are applied throughout system operation.
- *Centralized optimal control strategy:* A benchmark optimal control approach in which the set points of the cooling system are centrally optimized. The optimization algorithm is deployed and executed on the host computer, and all computations are performed in a fully centralized manner.
- *Parallel distributed optimal control strategy:* In this approach, four real smart field control stations are utilized to conduct optimization and control tasks independently. The agents coordinate through peer-to-peer communication and consensus mechanisms to achieve global system optimization without relying on any centralized controller.

In the validation process, a comparative analysis of computation time is first conducted between the centralized optimal control strategy and the proposed parallel distributed optimization control strategy. This comparison aims to evaluate the computational efficiency improvements achieved by the distributed approach. The computation time required for fitness evaluation in the genetic algorithm at each generation is recorded and analyzed.

Subsequently, the energy-saving performance of both optimization-based strategies is evaluated by comparing them with the baseline control strategy. The total energy consumption of the cooling system is measured for each control strategy, enabling a quantitative assessment of the energy efficiency improvements achieved by the parallel distributed optimization approach.

To further evaluate the robustness and scalability of the proposed parallel distributed optimal control strategy, a dedicated fault injection and recovery experiment was conducted. In this test, one control station was intentionally shut down during real-time optimization to simulate an unexpected agent failure. After a predefined interval, the control station was restored, representing the reintegration of a newly available agent into the optimization process. The system's ability to automatically detect agent loss and reintegration, dynamically reallocate computational tasks, and maintain stable optimization performance was assessed to evaluate the self-adaptive capabilities of the distributed framework under dynamic operating conditions.

7.4 Test results and discussion

7.4.1 Capability of the control station

The optimization process of the genetic algorithm involves iterative calculations and repeated evaluations of the fitness values for numerous candidate solutions. The time required for fitness evaluation is directly influenced by the computational capacity of the hardware, making it a key determinant of the overall optimization speed.

Figure 7.7 compares the time consumption for 100 fitness evaluations performed by the optimization algorithms running on the host computer and on a control station. The results indicate that the host computer completes the task in approximately 78.61 ms, while the control station requires about 190.24 ms for the same number of evaluations. This suggests that the host computer offers approximately 2.4 times the computational capability of a single control station.

Given this performance gap, a single control station is insufficient to meet the required optimization interval for real-time implementation. To overcome this limitation, the proposed parallel distributed control strategy utilizes multiple control stations to parallelize and distribute the computational tasks, thereby significantly improving optimization efficiency and enabling real-time control capability. A detailed comparative analysis of the overall optimization efficiency between the centralized optimal control strategy and the parallel distributed optimal control strategy is presented in Section 7.4.2.

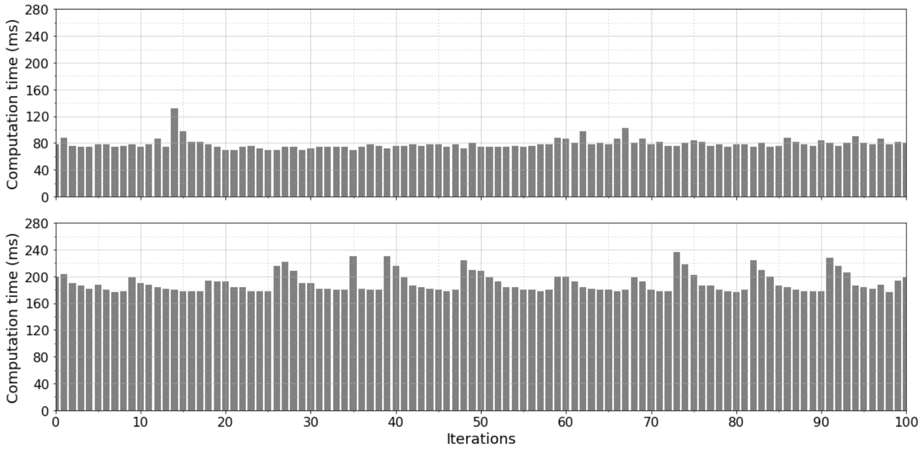


Figure 7.7 Computation time for fitness evaluations

(Top: Host computer, Bottom: Control station)

7.4.2 Optimization efficiency

Figure 7.8 illustrates the evolutionary optimization processes for the parallel distributed optimal control strategy, which operates across multiple control stations, and the centralized optimal control strategy, which is executed on the host computer. The results demonstrate that the proposed strategy achieves optimization efficiency comparable to that of the centralized strategy.

During the evolutionary process, the distributed strategy benefits from faster convergence by combining independent parallel computations with collaborative evolution among agents. Theoretically, with four control stations operating in parallel, the distributed approach could achieve up to a fourfold acceleration in optimization speed, exceeding the 2.4 times computational gap identified in Section 7.4.1. However, this theoretical advantage is partially offset by additional time needed for communication, synchronization, and decision-making processes inherent to the distributed architecture. As shown in Figure 7.8, agent 1 and agent 4 complete two generations of calculations within 11 seconds, but must wait for the other agents to finish their computations before conducting elite migration at 13 seconds. This synchronization mechanism, while essential for maintaining consistency across agents, adds to the overall time consumption. As a result, under the current experimental conditions, both the centralized and distributed optimization strategies require approximately one minute to complete the full optimization cycle. Moreover, the computational efficiency of the distributed strategy can be further scaled by integrating additional agents as needed.

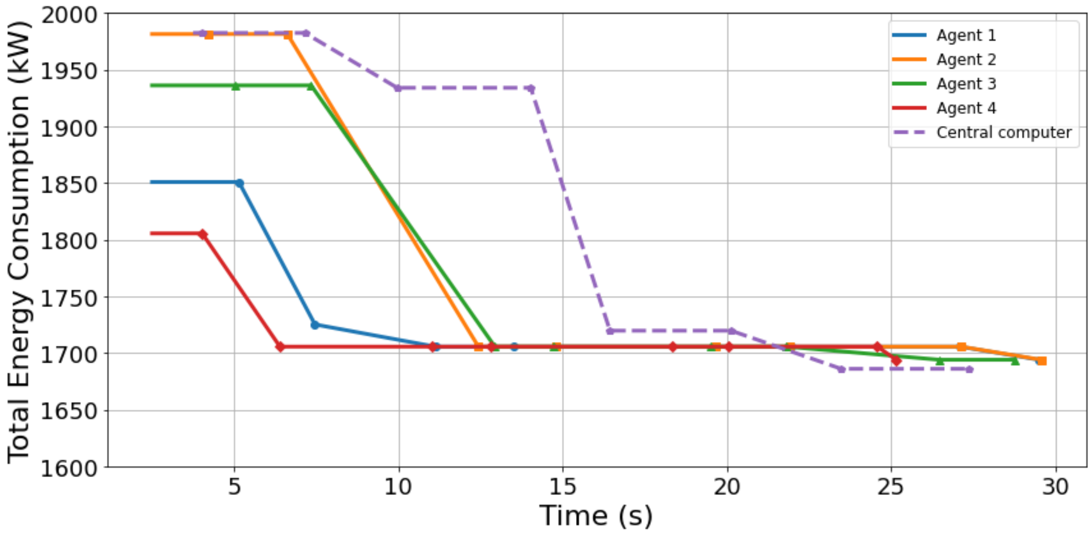


Figure 7.8 Evolutionary process in the parallel distributed control strategy (the four agents) and the centralized control strategy (the central computer)

7.4.3 Energy performance

Figure 7.9 presents the total power consumption profiles of the three control strategies described in Section 7.3.3 on a typical day. During the early stage (0–6 h), the system operates under relatively low load conditions, resulting in minimal differences among the three control strategies. As the system load increases after 6:00 AM, additional chillers and pumps are activated to meet the rising cooling demand, and the advantages of the optimization strategies become increasingly apparent. Both the centralized and parallel distributed optimization strategies effectively reduce power consumption during the peak load period (approximately 7:00 AM to 8:00 PM) compared to the baseline control strategy. Notably, the proposed parallel distributed optimal control strategy achieves slightly higher energy savings. Compared to the baseline, the centralized optimization reduces total energy consumption by 8.90%, while the parallel distributed optimization achieves a reduction of 9.16%.

The superior performance of the parallel distributed strategy is attributed to its ability to leverage multiple agents to collaboratively construct a denser optimization search grid without sacrificing convergence speed. Through parallel population evolution and inter-agent information exchange, the distributed framework maintains efficient global exploration and local exploitation, ensuring convergence performance that remains comparable to centralized optimization despite the decentralized architecture.

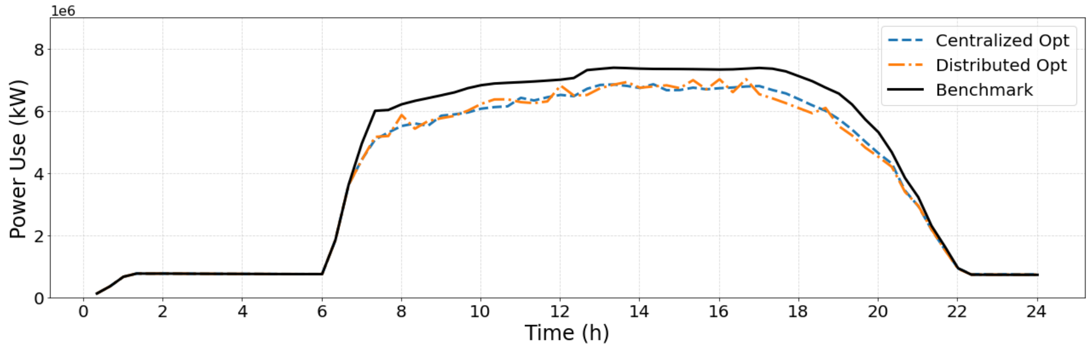


Figure 7.9 Total power consumption of each strategy

7.4.4 Reliability and scalability

Figure 7.10 illustrates the real-time aggregate CPU utilization of the four agents during the fault injection and recovery test. Over the 1000-second simulation period, the system completed eight optimization cycles. After each cycle, once the agents sent the optimized setpoints to the cooling system, all agents entered a waiting status, as reflected by intervals of zero CPU utilization.

The optimization computations were executed using Python with multi-threaded Basic Linear Algebra Subprograms (BLAS) libraries, enabling efficient intra-agent parallelization of linear algebra operations. As can be observed in Figure 7.10, the CPU utilization for each agent remained close to 100% of a single core during most of the computational periods. Occasionally, CPU utilization peaked at approximately 130%, which indicates that certain matrix operations (e.g., matrix multiplication, decomposition, and vectorized calculations) were parallelized by the BLAS library across multiple physical cores, leading to partial multi-core utilization.

At approximately 400 seconds, Agent 3 was intentionally shut down to simulate a fault event. Its CPU utilization dropped immediately to zero. The parallel distributed optimization framework autonomously detected the loss of the agent through the broadcasting protocol. It dynamically reallocated the genetic algorithm's population size and generation count among the remaining three agents. Since the hardware resources of the remaining agents remained unchanged, their CPU utilization patterns showed minimal variation; however, the increased computational task assigned to each agent resulted in an approximate 20-second increase in optimization cycle time.

Subsequently, at approximately 800 seconds, Agent 3 was restored to normal operation. Upon rejoining the system, the agent broadcast its activation states, and the distributed framework automatically updated the number of active agents and redistributed the optimization tasks accordingly. Throughout the entire fault injection and recovery process, the system maintained stable and continuous optimization operations without any interruption or degradation in control performance, thereby demonstrating the reliability and scalability of the proposed parallel distributed optimal control strategy under dynamic operational conditions.

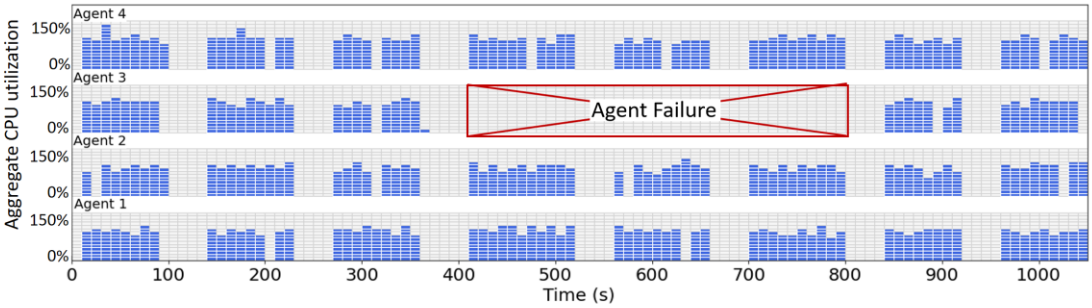


Figure 7.10 Real-time stacked CPU load grid under agent failure and recovery

7.5 Summary

This study proposed a novel parallel distributed optimal control strategy for AI-enabled online optimal control of building cooling systems, addressing key limitations in scalability, reliability, and centralized dependence found in conventional approaches. The proposed strategy was implemented using a parallel multi-agent architecture and validated through hardware-in-the-loop testing. The main findings are summarized as follows:

- The decentralized operational framework enhances optimization efficiency. Although a single control station exhibited lower computing power (190.24 ms per fitness evaluation vs. 78.61 ms on the host computer), the computational tasks were effectively parallelized across four agents. This enabled the proposed strategy to meet real-time optimization demands while supporting system scalability.
- Energy performance was notably improved. Compared to the baseline control strategy, the centralized and parallel distributed optimization strategies achieved energy savings of 8.90% and 9.16%, respectively. These results confirm the effectiveness of the proposed method in enhancing system-level energy efficiency while maintaining robustness.

- Robustness and fault-tolerance were validated through fault injection and recovery tests. The system autonomously identified agent failures, dynamically reallocated optimization tasks among available agents, and maintained uninterrupted optimization operations. Upon agent reintegration, the system adapted in real time and rebalanced computational tasks without compromising performance.

Looking forward, the proposed parallel distributed optimal control strategy presents great potential for broader application in complex and large-scale building energy systems, especially where real-time optimization, system robustness, and deployment flexibility are critical. Future research will focus on further enhancing the distributed optimization strategy to accommodate dynamic system uncertainties, heterogeneous agent capabilities, and more complex multi-objective optimization scenarios. In addition, experimental validations on larger-scale physical systems and long-term field implementations will be pursued to comprehensively assess the practical feasibility, stability, and economic benefits of the proposed strategy in real-world applications.

CHAPTER 8 CONCLUSIONS AND FUTURE WORK

This chapter presents the concluding remarks and outlines the future directions for research based on the findings and contributions of this study. This chapter is divided into three sections, each addressing different aspects of the research conducted. Section 8.1 highlights the main contributions of this study. Section 8.2 presents the conclusions drawn from the analysis and interpretation of the research outcomes. Section 8.3 focuses on recommendations for future work based on the research conducted.

8.1 Main contributions of this study

The main contributions of this PhD study are summarized as follows:

i. A comprehensive review of optimization control strategies for HVAC systems was conducted, outlining the evolution from conventional rule-based and model-driven methods to modern AI-enabled approaches. The review critically identified key gaps in scalability, robustness, and field-level adaptability, motivating the development of novel AI-empowered optimization strategies.

ii. A generic and practical framework was proposed for embedding AI-enabled optimization into Building Automation Systems (BAS) via smart control stations at the field level. This framework supports real-time operation, modular integration, and distributed decision-making, thus providing a scalable alternative to traditional AI-enabled optimization control.

iii. An AI-empowered online optimization strategy was developed, combining deep learning-enhanced evolutionary search with a robust assurance mechanism. This approach balances computational feasibility with reliable operation, supporting practical deployment on resource-constrained field stations.

iv. A physics-guided strategy was introduced by incorporating the concept of energy gradient into online optimization. This method enhances stability and robustness under data uncertainties by guiding decision-making through relative energy variations, ensuring reliable optimization under noisy or imperfect data conditions.

v. An incremental online learning mechanism with structural model augmentation was proposed to maintain long-term adaptability. By introducing a residual learning branch and combining it with hybrid offline–online modeling, the approach ensures sustained performance of AI models under equipment degradation and evolving system dynamics.

vi. A fully distributed, multi-agent control architecture was designed for parallel optimization in large-scale cooling systems. This strategy eliminates reliance on centralized coordination, enables autonomous agent cooperation, and supports scalability and fault-tolerant operation, paving the way for resilient AI-enabled control in real-world building environments.

8.2 Conclusions

Conclusions on the state-of-the-art research and technologies for AI in Building HVAC systems

- AI-enabled control strategies—spanning machine learning, reinforcement learning, and hybrid models—are redefining the optimization landscape of building HVAC systems. These methods demonstrate strong potential for achieving dynamic, data-driven, and multi-objective control, far outperforming conventional rule-based and model-based strategies in adaptability, scalability, and responsiveness.

- Despite advances in AI algorithms, practical deployment at the field level remains limited due to hardware constraints, a lack of standardized AI-integration frameworks, and difficulties in embedding real-time control logic into smart control stations. This limits the application of AI in dynamic, decentralized, or resource-constrained environments, such as on-site chiller control.
- Many AI-based HVAC optimization solutions rely on centralized or hierarchical control architectures, which introduce issues such as single points of failure, latency, and coordination overhead. These structural limitations hinder the robustness, scalability, and real-time deployment of optimization algorithms, especially in complex or large-scale building environments.
- Current applications of AI in building systems primarily fall under the broad definition of AI, acting as decision-support tools. The transition toward truly autonomous, self-evolving, and interpretable AI systems remains in its early stages. Future research must bridge this gap to develop resilient AI agents capable of long-term adaptation, reasoning under uncertainty, and transparent multi-objective decision-making in real-world energy systems.

Conclusions on the generic framework and strategies for integrating AI-based online optimization into BAS

- Tests on both the BMS testbed and a hardware-in-the-loop test rig validated the proposed framework. Results confirm that the field smart control stations can reliably host and execute AI-enhanced optimization algorithms at the edge, while maintaining compatibility with the BAS environment.

- Compared to the baseline control strategy, the proposed framework utilizes smart control stations for real-time optimization and achieves a significant energy saving of 7.66% through the deployment of an AI-enhanced Genetic Algorithm.
- The integrated watchdog and result validation mechanisms demonstrate strong capabilities in ensuring system robustness and stability under various disturbance scenarios. The watchdog module can effectively monitor execution latency and trigger fallback control upon detecting computation failures. The result validation mechanism can efficiently perform output filtering and outlier suppression, thereby maintaining stable and reliable control performance.

Conclusions on the AI-empowered online control optimization strategy for enhanced efficiency and robustness

- An AI-empowered control optimization strategy is developed for real-time implementation in building central cooling systems, featuring a deep learning-enabled genetic algorithm (DL-GA) that balances optimization performance with computational feasibility. The hybrid modeling approach—combining ANN-based data-driven models with physically interpretable formulations—achieves a 91% reduction in computation time compared to traditional physics-only models, thereby enabling field-level deployment on low-power smart stations.
- The proposed DL-GA demonstrates superior energy efficiency through real-time setpoint optimization, achieving an overall system energy saving of 7.66% compared to conventional fixed-setpoint control strategies. The largest improvements are observed in pump energy consumption, with cooling water pump energy usage reduced by 47.9%, outweighing the moderate increase in seawater pump power. These enhancements contribute to an estimated 2.9 tons of

CO_2 emissions reduction per summer day in typical operation scenarios in Hong Kong.

- To ensure operational robustness in practical deployment, a dynamic anomaly detection and fallback mechanism has been incorporated. This robust assurance scheme enables seamless transitions from AI-driven optimization to expert-rule-based control under abnormal conditions such as data corruption or communication loss. Experimental validation via hardware-in-the-loop testing confirms that this approach safeguards control reliability without significantly compromising energy performance.

Conclusions on the AI-empowered online optimization adopting energy gradient for robust and energy-efficient control

- The energy gradient-based online optimization strategy effectively captures local sensitivity trends in individual equipment energy consumption. By leveraging parameter correlation analysis and energy-saving potential analysis, the strategy identifies optimal setpoints that achieve a balance between system efficiency and stability under dynamic load conditions.
- Compared to conventional online optimization strategies, the proposed energy gradient-guided strategy significantly enhances control robustness under both random noise and systematic sensor bias. In robustness tests, Gaussian random noise and systematic bias were independently introduced to flow rate (0–2%) and temperature (0–0.5 K) inputs. The proposed strategy consistently maintained lower relative output errors, reducing estimation error by 50% under high random noise. Under substantial systematic bias, it constrained error growth within ±0.25%, whereas the conventional strategy exhibited errors exceeding ±0.75%.

These results demonstrate the ability of the proposed strategy to mitigate the effects of random noise and maintain low sensitivity to systematic offsets.

- The proposed strategy enhances control stability by estimating relative energy variation rather than the absolute value. Hardware-in-the-loop validation tests demonstrate that it reduces temperature fluctuation amplitudes during high cooling load conditions. The strategy avoids overshoots and maintains setpoint distributions within narrower ranges than the conventional online optimization strategy.
- In terms of energy saving, the proposed energy gradient-guided online optimization strategy achieves a total energy reduction of 5.3% compared with the baseline control, which is comparable to the 5.7% achieved by the conventional online optimization strategy. The slight compromise in energy efficiency is justified by improved operational reliability and a smoother control process.

Conclusions on the incremental online learning with model structural augmentation for AI-enabled control optimization considering equipment degradation

- Model degradation significantly impacts optimization efficiency. The optimized control strategy based on the initial model achieved significant energy savings of approximately 7.8% under design conditions. However, a 20% degradation in the chiller system's performance led to notable shifts in the operating behavior of both chillers and pumps, resulting in a 12% increase in total energy consumption. Under these degraded conditions, the original optimization algorithm failed to adapt effectively, causing the optimization efficiency to decline sharply to just 1.12%.
- The proposed incremental online learning strategy successfully mitigated the adverse impacts of equipment degradation. By continuously updating the model with real-time operational data, the strategy improved the predictive accuracy of

the optimization model, raising its R^2 value from 0.40 to 0.98. This enhanced modeling capability enabled the system to achieve 5.77% energy savings even under conditions of significant performance degradation.

- The incremental online learning strategy demonstrates superior robustness under noisy operational conditions. Test results indicate that the conventional method is prone to overfitting and catastrophic forgetting, with R^2 values dropping as low as 0.02 under ±2.5% disturbances. In contrast, the proposed strategy maintains stable performance even at the highest noise level, preserving R^2 values above 0.60 and avoiding catastrophic forgetting.

Conclusions on the parallel distributed approach for AI-enabled online optimization with enhanced efficiency, reliability and scalability

- The decentralized operational framework enhances optimization efficiency. Although a single control station exhibited lower computing power (190.24 ms per fitness evaluation vs. 78.61 ms on the host computer), the computational tasks were successfully parallelized across four agents. This enabled the proposed strategy to meet real-time optimization demands while supporting system scalability.
- Energy performance was notably improved. Compared to the baseline control strategy, the centralized and parallel distributed optimization strategies achieved energy savings of 8.90% and 9.16%, respectively. These results confirm the effectiveness of the proposed method in enhancing system-level energy efficiency while maintaining robustness.
- Robustness and fault-tolerance were validated through fault injection and recovery tests. The system autonomously identified agent failures, dynamically reallocated optimization tasks among available agents, and maintained uninterrupted

optimization operations. Upon agent reintegration, the system adapted in real time and rebalanced computational tasks without compromising performance.

8.3 Recommendations for future work

This PhD research has made substantial progress in advancing the field of intelligent optimization control for building central cooling systems, particularly through the development of novel control frameworks, optimization algorithms, control strategies, and architectural designs. These contributions collectively enable the practical and reliable deployment of AI-empowered optimization strategies in real-world BAS applications. Nevertheless, several important areas remain open for further exploration and refinement. Accordingly, the following directions are recommended for future work:

- Real-Time Implementation and Validation in Live Operational Environments. Although the proposed control strategies have been rigorously tested using a high-fidelity hardware-in-the-loop (HIL) platform, actual deployment in operational buildings introduces additional layers of complexity and uncertainty. These include heterogeneous equipment configurations, unpredictable system disturbances, and potential limitations in data availability. Future work should focus on the real-world implementation of the proposed methods in the reference building used for simulation studies in this thesis. Such deployment will allow a more thorough evaluation of algorithmic robustness, adaptability, and long-term performance under real operational constraints.
- Extension to Multi-Energy and Multi-Service Systems. While this study concentrates on central cooling systems, the proposed generic framework and

field-level smart control stations possess broad applicability to other supervisory control scenarios. Future research can extend the distributed AI optimization architecture to encompass multi-energy systems that integrate heating, ventilation, power distribution, energy storage, and renewable sources (e.g., photovoltaic or solar thermal systems). Coordinated optimization across multiple energy carriers will enhance overall system efficiency, economic viability, and operational resilience at the community or district scale.

- Development of Self-Diagnosis and Self-Healing Mechanisms in Smart Controllers. In practical applications, smart control stations may encounter unexpected hardware faults, network communication failures, or model drift over time. Therefore, embedding self-diagnostic modules capable of monitoring the health of hardware components and the fidelity of control models is a promising direction. Moreover, the integration of self-healing mechanisms that autonomously reconfigure optimization tasks or leverage redundant agents for recovery will significantly improve system fault tolerance and operational continuity.
- Incorporation of Explainable AI for Enhanced Transparency and Adoption. The deployment of AI in mission-critical building systems demands a high degree of transparency and interpretability. Future work should consider integrating explainable artificial intelligence (XAI) techniques to provide insights into the decision-making processes of deep learning models used in control optimization. Such techniques will improve user trust and facilitate human–machine collaboration, particularly in semi-autonomous or safety-critical BAS applications.

REFERENCES

Abuhussain, M. A., Alotaibi, B. S., Aliero, M. S., Asif, M., Alshenaifi, M. A., & Dodo, Y. A. (2023). Adaptive HVAC System Based on Fuzzy Controller Approach. *Applied Sciences*, *13*(20), Article 20. https://doi.org/10.3390/app132011354

Asad, H. S., Yuen, R. K. K., Liu, J., & Wang, J. (2019). Adaptive modeling for reliability in optimal control of complex HVAC systems. *Building Simulation*, *12*(6), 1095–1106. https://doi.org/10.1007/s12273-019-0558-9

Aslam, S., Herodotou, H., Mohsin, S. M., Javaid, N., Ashraf, N., & Aslam, S. (2021). A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. *Renewable and Sustainable Energy Reviews*, *144*, 110992. https://doi.org/10.1016/j.rser.2021.110992

Assran, B. M., Aytekin, A., Feyzmahdavian, H. R., Johansson, M., & Rabbat, M. G. (2020). Advances in Asynchronous Parallel and Distributed Optimization. *Proceedings of the IEEE*, *108*(11), 2013–2031. https://doi.org/10.1109/JPROC.2020.3026619

Aste, N., Manfren, M., & Marenzi, G. (2017). Building Automation and Control Systems and performance optimization: A framework for analysis. *Renewable & Sustainable Energy Reviews*, *75*, 313–330. https://doi.org/10.1016/j.rser.2016.10.072

Azemi, N. L. M., & Wahid, N. (2021). Uncertainty in internet of things: A review. *International Journal of Advanced Technology and Engineering Exploration*, *8*(75), 422–431. https://doi.org/10.19101/IJATEE.2020.762115

Azzi, A., Tabaa, M., Chegari, B., & Hachimi, H. (2024). Balancing Sustainability and Comfort: A Holistic Study of Building Control Strategies That Meet the Global Standards for Efficiency and Thermal Comfort. *Sustainability*, *16*(5), Article 5. https://doi.org/10.3390/su16052154

Barone, G., Buonomano, A., Forzano, C., Giuzio, G. F., Palombo, A., & Russo, G. (2023). A new thermal comfort model based on physiological parameters for the smart design and control of energy-efficient HVAC systems. *Renewable and Sustainable Energy Reviews*, *173*, 113015. https://doi.org/10.1016/j.rser.2022.113015

Bi, S., Ho, C. K., & Zhang, R. (2015). Wireless powered communication: Opportunities and challenges. *IEEE Communications Magazine*, *53*(4), 117–125. https://doi.org/10.1109/MCOM.2015.7081084

Biswas, P., Rashid, A., Biswas, A., Nasim, M. A. A., Chakraborty, S., Gupta, K. D., & George, R. (2024). AI-driven approaches for optimizing power consumption: A comprehensive survey. *Discover Artificial Intelligence*, *4*(1), 116. https://doi.org/10.1007/s44163-024-00211-7

Borase, R. P., Maghade, D. K., Sondkar, S. Y., & Pawar, S. N. (2021). A review of PID control, tuning methods and applications. *International Journal of Dynamics and Control*, *9*(2), 818–827. https://doi.org/10.1007/s40435-020-00665-4

Carli, R., Cavone, G., Dotoli, M., Epicoco, N., & Scarabaggio, P. (2019). Model predictive control for thermal comfort optimization in building energy management systems. *2019 IEEE International Conference on Systems, Man* *and Cybernetics (SMC)*, 2608–2613. https://doi.org/10.1109/SMC.2019.8914489

Chen, C., Zhang, P., Zhang, H., Dai, J., Yi, Y., Zhang, H., & Zhang, Y. (2020). Deep Learning on Computational-Resource-Limited Platforms: A Survey. *Mobile Information Systems*, *2020*(1), 8454327. https://doi.org/10.1155/2020/8454327

Chen, Z., Xiao, F., Guo, F., & Yan, J. (2023). Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy*, *9*, 100123. https://doi.org/10.1016/j.adapen.2023.100123

Choi, H., Um, C. Y., Kang, K., Kim, H., & Kim, T. (2021). Application of vision-based occupancy counting method using deep learning and performance analysis. *Energy and Buildings*, *252*, 111389. https://doi.org/10.1016/j.enbuild.2021.111389

Dai, M., Li, H., & Wang, S. (2023a). A reinforcement learning-enabled iterative learning control strategy of air-conditioning systems for building energy saving by shortening the morning start period. *Applied Energy*, *334*, 120650. https://doi.org/10.1016/j.apenergy.2023.120650

Dai, M., Li, H., & Wang, S. (2023b). Multi-agent based distributed cooperative control of air-conditioning systems for building fast demand response. *Journal of Building Engineering*, *77*, 107463. https://doi.org/10.1016/j.jobe.2023.107463

Dai, M., Li, H., & Wang, S. (2024). Development and implementation of supply-based feedback controls for energy-efficient and grid-interactive cooling management over entire building daily cycle. *Energy*, *313*, 133858. https://doi.org/10.1016/j.energy.2024.133858

Deng, K., Sun, Y., Li, S., Lu, Y., Brouwer, J., Mehta, P. G., Zhou, M., & Chakraborty, A. (2015). Model Predictive Control of Central Chiller Plant With Thermal Energy Storage Via Dynamic Programming and Mixed-Integer Linear Programming. *IEEE Transactions on Automation Science and Engineering*, *12*(2), 565–579. IEEE Transactions on Automation Science and Engineering. https://doi.org/10.1109/TASE.2014.2352280

Ding, X., Liao, X., Cui, W., Meng, X., Liu, R., Ye, Q., & Li, D. (2024). A Deep Reinforcement Learning Optimization Method Considering Network Node Failures. *Energies*, *17*(17), Article 17. https://doi.org/10.3390/en17174471

EPD, E. P. Department. (2023). (2023). *Marine Water Quality in Hong Kong in 2023*. The Government of the Hong Kong Special Administrative Region. https://www.epd.gov.hk/epd/sc_chi/environmentinhk/water/hkwqrc/waterquality/marine-2.html

Espín-Sarzosa, D., Palma-Behnke, R., & Núñez-Mata, O. (2020). Energy Management Systems for Microgrids: Main Existing Trends in Centralized Control Architectures. *Energies*, *13*(3), Article 3. https://doi.org/10.3390/en13030547

Fan, H., Wang, C., Liu, L., & Li, X. (2022). Review of Uncertainty Modeling for Optimal Operation of Integrated Energy System. *Frontiers in Energy Research*, *9*. https://doi.org/10.3389/fenrg.2021.641337

Fritzson, P., & Engelson, V. (1998). Modelica—A unified object-oriented language for system modeling and simulation. In E. Jul (Ed.), *ECOOP'98—Object-Oriented Programming* (pp. 67–90). Springer. https://doi.org/10.1007/BFb0054087

Gao, Y., & Ruan, Y. (2021). Interpretable deep learning model for building energy consumption prediction based on attention mechanism. *Energy and Buildings*, *252*, 111379. https://doi.org/10.1016/j.enbuild.2021.111379

Gao, Z., Yu, J., Zhao, A., Hu, Q., & Yang, S. (2022). A hybrid method of cooling load forecasting for large commercial building based on extreme learning machine. *Energy*, *238*, 122073. https://doi.org/10.1016/j.energy.2021.122073

Ghavifekr, A. A. (2021). Application of Heuristic Techniques and Evolutionary Algorithms in Microgrids Optimization Problems. In A. Anvari-Moghaddam, H. Abdi, B. Mohammadi-Ivatloo, & N. Hatziargyriou (Eds.), *Microgrids: Advances in Operation, Control, and Protection* (pp. 219–251). Springer International Publishing. https://doi.org/10.1007/978-3-030-59750-4_9

Gholamzadehmir, M., Del Pero, C., Buffa, S., Fedrizzi, R., & Aste, N. (2020). Adaptive-predictive control strategy for HVAC systems in smart buildings – A review. *Sustainable Cities and Society*, *63*, 102480. https://doi.org/10.1016/j.scs.2020.102480

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607. https://doi.org/10.1016/j.chb.2020.106607

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). *A Review of Safe Reinforcement Learning: Methods, Theory and Applications* (arXiv:2205.10330). arXiv. https://doi.org/10.48550/arXiv.2205.10330

Halhoul Merabet, G., Essaaidi, M., Ben Haddou, M., Qolomany, B., Qadir, J., Anan, M., Al-Fuqaha, A., Abid, M. R., & Benhaddou, D. (2021). Intelligent building control systems for thermal comfort and energy-efficiency: A systematic review of artificial intelligence-assisted techniques. *Renewable and Sustainable Energy Reviews*, *144*, 110969. https://doi.org/10.1016/j.rser.2021.110969

Himeur, Y., Alsalemi, A., Al-Kababji, A., Bensaali, F., Amira, A., Sardianos, C., Dimitrakopoulos, G., & Varlamis, I. (2021). A survey of recommender systems for energy efficiency in buildings: Principles, challenges and prospects. *Information Fusion*, *72*, 1–21. https://doi.org/10.1016/j.inffus.2021.02.002

Hochreiter, S., Younger, A. S., & Conwell, P. R. (2001). Learning to Learn Using Gradient Descent. In G. Dorffner, H. Bischof, & K. Hornik (Eds.), *Artificial Neural Networks—ICANN 2001* (pp. 87–94). Springer. https://doi.org/10.1007/3-540-44668-0_13

Homod, R. Z. (2018). Analysis and optimization of HVAC control systems based on energy and performance considerations for smart buildings. *Renewable Energy*, *126*, 49–64. https://doi.org/10.1016/j.renene.2018.03.022

Ibrahim, M. S., Dong, W., & Yang, Q. (2020). Machine learning driven smart electric power systems: Current trends and new perspectives. *Applied Energy*, *272*, 115237. https://doi.org/10.1016/j.apenergy.2020.115237

Ifaei, P., Esfehankalateh, A. T., Ghobadi, F., Mohammadi-Ivatloo, B., & Yoo, C. (2023). Systematic review and cutting-edge applications of prominent heuristic optimizers in sustainable energies. *Journal of Cleaner Production*, *414*, 137632. https://doi.org/10.1016/j.jclepro.2023.137632

International Energy Agency. (2021). *Net zero by 2050—A roadmap for the global energy sector*.

Jendoubi, I., & Bouffard, F. (2023). Multi-agent hierarchical reinforcement learning for energy management. *Applied Energy*, *332*, 120500. https://doi.org/10.1016/j.apenergy.2022.120500

Jiang, T., Fang, H., & Wang, H. (2019). Blockchain-Based Internet of Vehicles: Distributed Network Architecture and Performance Analysis. *IEEE Internet of Things Journal*, *6*(3), 4640–4649. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2018.2874398

Jiang, Y., Zhu, S., Xu, Q., Yang, B., & Guan, X. (2023). Hybrid modeling-based temperature and humidity adaptive control for a multi-zone HVAC system. *Applied Energy*, *334*, 120622. https://doi.org/10.1016/j.apenergy.2022.120622

Kastner, J. K., & Hong, S. J. (1984). A review of expert systems. *European Journal of Operational Research*, *18*(3), 285–292. https://doi.org/10.1016/0377-2217(84)90150-4

Li, K., Zhang, J., Chen, X., & Xue, W. (2022). Building's hourly electrical load prediction based on data clustering and ensemble learning strategy. *Energy and Buildings*, *261*, 111943. https://doi.org/10.1016/j.enbuild.2022.111943

Li, L., Fu, Y., Fung, J. C. H., Qu, H., & Lau, A. K. H. (2021). Development of a back-propagation neural network and adaptive grey wolf optimizer algorithm for thermal comfort and energy consumption prediction and optimization. *Energy and Buildings*, *253*, 111439. https://doi.org/10.1016/j.enbuild.2021.111439

Li, W., & Wang, S. (2020). A multi-agent based distributed approach for optimal control of multi-zone ventilation systems considering indoor air quality and energy use. *Applied Energy*, *275*, 115371. https://doi.org/10.1016/j.apenergy.2020.115371

Li, W., & Wang, S. (2022). A fully distributed optimal control approach for multi-zone dedicated outdoor air systems to be implemented in IoT-enabled building automation networks. *Applied Energy*, *308*, 118408. https://doi.org/10.1016/j.apenergy.2021.118408

Li, Z., Gao, J., Guo, J., Xie, Y., Yang, X., & Li, M.-J. (2024). Optimal loading distribution of chillers based on an improved beluga whale optimization for reducing energy consumption. *Energy and Buildings*, *307*, 113942. https://doi.org/10.1016/j.enbuild.2024.113942

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386.

Martinez-Molina, A., & Alamaniotis, M. (2020). Enhancing Historic Building Performance with the Use of Fuzzy Inference System to Control the Electric Cooling System. *Sustainability*, *12*(14), Article 14. https://doi.org/10.3390/su12145848

Michailidis, P., Michailidis, I., Vamvakas, D., & Kosmatopoulos, E. (2023). Model-Free HVAC Control in Buildings: A Review. *Energies*, *16*(20), Article 20. https://doi.org/10.3390/en16207124

Mobaraki, B., Lozano-Galant, F., Soriano, R. P., & Castilla Pascual, F. J. (2021). Application of Low-Cost Sensors for Building Monitoring: A Systematic Literature Review. *Buildings*, *11*(8), Article 8. https://doi.org/10.3390/buildings11080336

Mtibaa, F., Nguyen, K.-K., Dermardiros, V., & Cheriet, M. (2021). Context-aware Model Predictive Control framework for multi-zone buildings. *Journal of Building Engineering*, *42*, 102340. https://doi.org/10.1016/j.jobe.2021.102340

Mtibaa, F., Nguyen, K.-K., Dermardiros, V., McDonald, S., Venne, J.-S., & Cheriet, M. (n.d.). *Hierarchical Multi-Agent Control Framework for Energy Efficiency and Carbon Emission Reduction in Multi-Zone Buildings*.

Mu, B., Li, Y., Seem, J. E., & Hu, B. (2015). A Multivariable Newton-Based Extremum Seeking Control for Condenser Water Loop Optimization of Chilled-Water Plant. *Journal of Dynamic Systems, Measurement, and Control*, *137*(111011). https://doi.org/10.1115/1.4031051

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. In R. Silhavy & P. Silhavy (Eds.), *Artificial Intelligence Application in Networks and Systems* (pp. 15–25). Springer International Publishing. https://doi.org/10.1007/978-3-031-35314-7_2

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A Comprehensive Overview of Large Language Models* (arXiv:2307.06435). arXiv. https://doi.org/10.48550/arXiv.2307.06435

Navon, R., & Goldschmidt, E. (1999). Automated data acquisition for on-site control. *Proc., 16th IAARC/IFAC/IEEE International Symp. on Automation and Robotics in Construction*, 681–686. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ced5a154415a88a91d8950dda3e29087b3ef0314

Nguyen, T. T., Nguyen, N. D., & Nahavandi, S. (2020). Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications. *IEEE Transactions on Cybernetics*, *50*(9), 3826–3839. https://doi.org/10.1109/TCYB.2020.2977374

Parvin, K., Lipu, M. S. H., Hannan, M. A., Hannan, M. A., Ker, P. J., Abdullah, M. A., Jern, K. P., Rawshan Ara Begum, Begum, R. A., Begum, R., Mansur, M., Muhamad Mansur, Muttaqi, K. M., Mahlia, T. M. I., Zhang, R., & Dong, Z. Y. (2021). Intelligent Controllers and Optimization Algorithms for Building Energy Management Towards Achieving Sustainable Development: Challenges and Prospects. *IEEE Access*, *9*, 41577–41602. https://doi.org/10.1109/access.2021.3065087

Perera, A. T. D., & Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, *137*, 110618. https://doi.org/10.1016/j.rser.2020.110618

Pombeiro, H., Machado, M. J., & Silva, C. (2017). Dynamic programming and genetic algorithms to control an HVAC system: Maximizing thermal comfort and minimizing cost with PV production and storage. *Sustainable Cities and Society*, *34*, 228–238. https://doi.org/10.1016/j.scs.2017.05.021

Qin, H., & Wang, X. (2022). A multi-discipline predictive intelligent control method for maintaining the thermal comfort on indoor environment. *Applied Soft Computing*, *116*, 108299. https://doi.org/10.1016/j.asoc.2021.108299

Rajput, L., Tyagi, N., Tyagi, S., & Tyagi, D. K. (2024). State of the art object detection: A comparative study of YOLO and ViT. *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, 01–06. https://doi.org/10.1109/ISCS61804.2024.10581202

ŞAHiN, E., Arslan, N. N., & Özdemir, D. (2025). Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning.

Neural Computing and Applications, *37*(2), 859–965. https://doi.org/10.1007/s00521-024-10437-2

Sahni, Y., Cao, J., Yang, L., & Wang, S. (2022). Distributed resource scheduling in edge computing: Problems, solutions, and opportunities. *Computer Networks*, *219*, 109430. https://doi.org/10.1016/j.comnet.2022.109430

Shortliffe, E. H., Buchanan, B. G., & Feigenbaum, E. A. (1979). Knowledge engineering for medical decision making: A review of computer-based clinical decision aids. *Proceedings of the IEEE*, *67*(9), 1207–1224. Proceedings of the IEEE. https://doi.org/10.1109/PROC.1979.11436

Simpson-Porco, J. W., Shafiee, Q., Dörfler, F., Vasquez, J. C., Guerrero, J. M., & Bullo, F. (2015). Secondary Frequency and Voltage Control of Islanded Microgrids via Distributed Averaging. *IEEE Transactions on Industrial Electronics*, *62*(11), 7025–7038. IEEE Transactions on Industrial Electronics. https://doi.org/10.1109/TIE.2015.2436879

Su, B., & Wang, S. (2020). An agent-based distributed real-time optimal control strategy for building HVAC systems for applications in the context of future IoT-based smart sensor networks. *Applied Energy*, *274*, 115322. https://doi.org/10.1016/j.apenergy.2020.115322

Sun, Y., Chen, X., Wu, S., Wei, W., Wang, W., & Deng, S. (2022). Performance analysis of air source heat pump space heating system with an adaptive control for supply water temperature. *Applied Thermal Engineering*, *211*, 118401. https://doi.org/10.1016/j.applthermaleng.2022.118401

Sun, Y., Shen, Q., Liu, X., Zhang, L., Li, X., Zhang, T., & Lin, M. (2024). Optimal control strategy for building HVAC systems: Satisfying flexible demand response with different value-based selection. *Energy and Buildings*, *323*, 114823. https://doi.org/10.1016/j.enbuild.2024.114823

Swagatika Priyadarshini, L. A., Kaushik, V., & Kataria, R. (2025). The Role of Anthropogenic Activities, Energy Demand, and Global Climate Change. In P. Singh & B. Ao (Eds.), *The Intersection of Global Energy Politics and Climate Change: A Comprehensive Analysis of Energy Markets and Economics* (pp. 35–61). Springer Nature. https://doi.org/10.1007/978-981-96-0535-4_2

The International Organization for Standardization. (2020). *Information technology—Artificial intelligence—Artificial intelligence concepts and terminology* (Version 22989:2022). https://www.iso.org/standard/74296.html

Ulpiani, G., Borgognoni, M., Romagnoli, A., & Di Perna, C. (2016). Comparing the performance of on/off, PID and fuzzy controllers applied to the heating system of an energy-efficient building. *Energy and Buildings*, *116*, 1–17. https://doi.org/10.1016/j.enbuild.2015.12.027

Vakiloroaya, V., Ha, Q. P., & Samali, B. (2013). Energy-efficient HVAC systems: Simulation–empirical modelling and gradient optimization. *Automation in Construction*, *31*, 176–185. https://doi.org/10.1016/j.autcon.2012.12.006

Veloso, G. A., & Loschi, R. H. (2021). Dynamic linear degradation model: Dealing with heterogeneity in degradation paths. *Reliability Engineering & System Safety*, *210*, 107446. https://doi.org/10.1016/j.ress.2021.107446

Wang, H., & Wang, S. (2021). A hierarchical optimal control strategy for continuous demand response of building HVAC systems to provide frequency regulation service to smart power grids. *Energy*, *230*, 120741. https://doi.org/10.1016/j.energy.2021.120741

Wang, S., Xing, J., Jiang, Z., & Dai, Y. (2019). A Decentralized Swarm Intelligence Algorithm for Global Optimization of HVAC System. *IEEE Access*, *7*, 64744–64757. https://doi.org/10.1109/ACCESS.2019.2913359

Wollschlaeger, M., Sauter, T., & Jasperneite, J. (2017). The Future of Industrial Communication: Automation Networks in the Era of the Internet of Things and Industry 4.0. *IEEE Industrial Electronics Magazine*, *11*(1), 17–27. https://doi.org/10.1109/MIE.2017.2649104

Wong, J. K. W., & Li, H. (2010). Construction, application and validation of selection evaluation model (SEM) for intelligent HVAC control system. *Automation in Construction*, *19*(2), 261–269. https://doi.org/10.1016/j.autcon.2009.10.002

Xie, L., Huang, S., Ye, Y., & Zhang, X. (2021). Regional suitability of heating tower heat pumps based on technical and economic performance maps. *Energy and Buildings*, *245*, 111048. https://doi.org/10.1016/j.enbuild.2021.111048

Yang, N., Li, J., & Chen, J. (2020). Fully distributed hybrid adaptive learning consensus protocols for a class of non-linearly parameterized multi-agent systems. *Applied Mathematics and Computation*, *375*, 125074. https://doi.org/10.1016/j.amc.2020.125074

Yang, Y., Hu, G., & Spanos, C. J. (2020). HVAC Energy Cost Optimization for a Multizone Building via a Decentralized Approach. *IEEE Transactions on Automation Science and Engineering*, *17*(4), 1950–1960. https://doi.org/10.1109/TASE.2020.2983486

Yuan, X., Pan, Y., Yang, J., Wang, W., & Huang, Z. (2021). Study on the application of reinforcement learning in the operation optimization of HVAC system.

Building Simulation, *14*(1), 75–87. https://doi.org/10.1007/s12273-020-0602-9

Zhang, S., Zhang, Z., Cui, R., Li, H., & Yan, W. (2023). Distributed optimal consensus protocol for high-order integrator-type multi-agent systems. *Journal of the Franklin Institute*, *360*(10), 6862–6879. https://doi.org/10.1016/j.jfranklin.2023.05.001

Zhao, S., Amini, M. R., Sun, J., & Mi, C. C. (2021). A Two-Layer Real-Time Optimization Control Strategy for Integrated Battery Thermal Management and HVAC System in Connected and Automated HEVs. *IEEE Transactions on Vehicular Technology*, *70*(7), 6567–6576. https://doi.org/10.1109/TVT.2021.3085938

Zhou, Y., Liu, Y., Wang, D., & Liu, X. (2021). Comparison of machine-learning models for predicting short-term building heating load using operational parameters. *Energy and Buildings*, *253*, 111505. https://doi.org/10.1016/j.enbuild.2021.111505

Zhu, P., & Ren, W. (2021). Fully Distributed Joint Localization and Target Tracking With Mobile Robot Networks. *IEEE Transactions on Control Systems Technology*, *29*(4), 1519–1532. IEEE Transactions on Control Systems Technology. https://doi.org/10.1109/TCST.2020.2991126