DEVELOPMENT OF STOCHASTIC OCCUPANCY MODELLING METHODS AND OCCUPANCY-INTEGRATED MPC FOR SMART BUILT ENVIRONMENT CONTROL AND ITS IMPLEMENTATION

ZHANG HANBEI

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Building Environment and Energy Engineering

Development of stochastic occupancy modelling methods and occupancy-integrated MPC for smart built environment control and its implementation

ZHANG Hanbei

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

June 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 material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

_____ (Signed)

ZHANG Hanbei (Name of student)

ABSTRACT

Abstract of thesis entitled: Development of stochastic occupancy modelling methods and occupancy-integrated MPC for smart built environment control and its implementation

Submitted by: ZHANG Hanbei

For the degree of: Doctor of Philosophy

at The Hong Kong Polytechnic University in August 2025

Modern HVAC systems, which dominate building energy use, face the critical challenge of balancing the conflicting objectives of thermal comfort, indoor air quality (IAQ), and stringent energy efficiency targets. Among the various factors, occupant behavior has emerged as one of the most significant drivers of building energy use. However, traditional building automation systems often rely on oversimplified assumptions about occupancy patterns, resulting in substantial energy wastage due to oversupplying unoccupied spaces, while also risking compromised occupant comfort and satisfaction.

With the advancements of Internet of Things (IoT) devices and information communication technologies in smart buildings, real-time occupancy information—such as presence, number of occupants, and personalized feedback—is increasingly accessible. Additionally, IoT-enabled built environment data is being extensively collected, providing unprecedented opportunities for data-driven building management. However, there remains a significant gap in the development of data-driven modeling and occupancy-integrated control systems capable of effectively leveraging IoT data to match building energy services with actual occupancy demand. Traditional data-driven modeling methods often suffer from high model complexity and poor generalization ability, limiting their scalability and applicability across diverse building spaces. At the same time, the suboptimal performance of simple, reactive control strategies highlights

the need for the advanced stochastic occupancy-integrated model predictive control (MPC) strategy. Furthermore, there is a lack of real-world implementation and validation of MPC systems under living conditions with active occupant participation and feedback.

This research aims to address the aforementioned challenges by developing advanced stochastic occupancy modeling methods to accurately identify diverse occupancy patterns, integrating stochastic occupancy prediction into the MPC strategy to align building energy system control with real occupancy demand, and implementing the MPC strategy in an IoT-enabled real-world living lab environment to evaluate its practical impact.

This study first proposes the Adaptive B-Spline-based inhomogeneous Markov chains (IMC) method for stochastic occupancy modelling. This method introduces a dynamic knot adjustment mechanism to better capture occupancy variations across different types of spaces. The proposed method achieves significantly improved prediction accuracy while reducing model complexity. (Chapter 3)

To address the poor generalization issue, this study further develops a band structure-integrated continuous-time inhomogeneous Markov chain (CTIMC) modelling method. The band structure constrains transitions between states to physically meaningful neighboring states, reflecting realistic occupant movement and behavior. Compared to traditional discrete-time IMC models, the band structure-integrated CTIMC method demonstrates superior generalization ability and computational efficiency. (Chapter 4)

This research further integrates the stochastic occupancy prediction into the MPC strategy for multi-objective optimal built environment control. The stochasticity and time-inhomogeneity of occupancy heat gains and CO_2 generations are embedded in the prediction of built environment and energy consumption in MPC. TRNSYS simulation demonstrates the effectiveness of the stochastic occupancy-integrated MPC in achieving significant energy savings while improving thermal comfort and IAQ.

(Chapter 5)

This study further develops an IoT-enabled architecture for intelligent built environment management, integrating stochastic occupancy modeling and real-time data into an Occupant-in-the-loop MPC strategy and validates through field implementations. Deployed in a living lab, the system optimizes HVAC operations, balancing thermal comfort, air quality, and energy efficiency. A one-month experimental evaluation showed 54.9% energy savings, improved comfort, and sufficient ventilation, advancing real-world MPC implementation and future IoT-based building management enhancements. (Chapter 6)

Finally, to enhance the scalability and cost-effectiveness of strategy deployment, this study develops an IoT retrofit scheme that considers the trade-off between costs and model accuracy. A data-driven modeling method using low-cost data is proposed, with systematic analysis conducted on model accuracy and the associated costs. The results indicate that the proposed modeling method based on low-cost data can achieve prediction accuracy comparable to modeling methods using high-cost, high-accuracy data. The proposed method facilitates large-scale, cost-effective deployment of the intelligent built environment management system. (Chapter 7)

This thesis provides a comprehensive framework for intelligent built environment management, combining cutting-edge stochastic modeling, predictive optimization, and real-time occupant interaction. By addressing the challenges posed by dynamic and uncertain occupancy patterns, this research contributes to more sustainable, scalable, and occupant-centric building operations. The findings have broad implications for a range of building types and operational scenarios, offering new insights into the integration of data-driven modeling, IoT, and advanced control technologies to optimize the built environment.

PUBLICATIONS

Journal Papers:

- **Zhang H**, Thilker CA, Madsen H, Li R, Xiao F, Ma T, et al. Stochastic occupancy modeling for spaces with irregular occupancy patterns using adaptive B-Spline-based inhomogeneous Markov Chains. Building and Environment 2024;261:111721. https://doi.org/10.1016/j.buildenv.2024.111721. (Chapter 3)
- **Zhang H**, Thilker CA, Xiao F, Madsen H, Li R, Ma T et al. Physics-informed band structure-integrated Continuous-time inhomogeneous Markov chains for stochastic occupancy modelling. Advanced Engineering Informatics(under review) (Chapter 4)
- **Zhang H**, Thilker CA, Xiao F, Madsen H, Li R, Ma T, et al. Stochastic occupancy-integrated MPC for multi-objective optimal built environment control. Build Simul 2025. https://doi.org/10.1007/s12273-025-1300-4. (Chapter 5)
- **Zhang H**, Zhang J., Xiao F, Ma T, Xu K. Thilker CA, Madsen H. IoT-enabled Occupant-in-the-loop MPC for intelligent built environment management: Field implementation in an office living lab. Journal of Building Engineering (under review) (Chapter 6)
- **Zhang H**, Thilker CA, Madsen H, Li R, Xiao F, Ma T, et al. Development of IoT retrofit schemes for the scaled deployment of the building thermal dynamics concerning costs (In preparation) (Chapter 7)
- **Zhang H**, Xiao F, Zhang C, Li R. A multi-agent system based coordinated multi-objective optimal load scheduling strategy using marginal emission factors for building cluster demand response. Energy and Buildings 2023;281:112765. https://doi.org/10.1016/j.enbuild.2022.112765.
- Zhang J, Xiao F, Li A, Ma T, Xu K, **Zhang H**, et al. Graph neural network-based spatio-temporal indoor environment prediction and optimal control for central air-conditioning systems. Building and Environment 2023;242:110600. https://doi.org/10.1016/j.buildenv.2023.110600.
- Ma T, Zhang J, Xiao F, Xu K, **Zhang H**, et al. Digital twin for 3D interactive building operations: Integrating BIM, IoT-enabled building automation systems, AI, and mixed reality. Automation in Construction. 2025;176:106277. https://doi.org/10.1016/j.autcon.2025.106277

- Liao W, Xiao F, Li Y, **Zhang H**, Peng J. A comparative study of demand-side energy management strategies for building integrated photovoltaics-battery and electric vehicles (EVs) in diversified building communities. Applied Energy 2024;361:122881. https://doi.org/10.1016/j.apenergy.2024.122881.
- Xu K, Chen Z, Xiao F, Zhang J, **Zhang H**, Ma T. Semantic model-based large-scale deployment of AI-driven building management applications. Automation in Construction 2024;165:105579. https://doi.org/10.1016/j.autcon.2024.105579.
- Yu H, Xiao F, **Zhang H**, Liao W, Li Y. A Two-stage coordinated restoration scheme of hybrid AC/DC distribution grid considering cold load pickup and resilience enhancement. Sustainable Cities and Society 2024:105959. https://doi.org/10.1016/j.scs.2024.105959.

Conference Papers:

- **Zhang H.** & Xiao, F. (2022). Model-based multi-objective optimal energy scheduling for energy management of distributed energy system considering different utility functions. **CUE2022.**
- **Zhang H.** & Xiao, F. (2023). Model-based energy scheduling strategy for distributed energy system in joint electricity energy and carbon trading markets considering marginal emission factors. **SuDBE2023: Sustainable Urban Renewal.**
- **Zhang H.** & Xiao, F. (2025). Data-Driven Modeling of Building Thermal Dynamics and Indoor CO_2 Concentration Using Stochastic Differential Equations. **Nexus 2025.**

ACKNOWLEDGMENTS

I would like to express my deepest and most heartfelt gratitude to my supervisor, Dr. Linda Fu Xiao, for her unwavering guidance, invaluable insights, and steadfast encouragement throughout the journey of this research. Her readily available supervision, thoughtful suggestions, and warm patience have been a cornerstone of my academic progress. I am truly privileged to have had the opportunity to learn and grow under her mentorship, and I will always cherish the profound impact she has had on my academic and personal development.

I would also like to extend my heartfelt thanks to Professor Shengwei Wang for his invaluable suggestions, consistent support, and constructive feedback throughout the entire course of my Ph.D. study. His wisdom and encouragement have been a source of inspiration, and his guidance has greatly shaped the direction and depth of my research. I am deeply honored to have had his support during this transformative journey.

My sincere appreciation also goes to my fellow colleagues, friends and family, whose unwavering support, and kind assistance have been a constant source of motivation. Their encouragement has been instrumental in helping me overcome challenges and achieve milestones throughout this journey. The bonds we have formed during these unforgettable years will always hold a special place in my heart, and I will treasure the memories of our shared experiences.

To all of you, thank you from the bottom of my heart.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Background and motivations

The built environment management is undergoing a transformative shift driven by digitalization and intelligentization, with the integration of Internet of Things (IoT) technologies emerging as a cornerstone of modern building automation systems [1]. It was estimated that the number of connected IoT devices installed in commercial smart buildings was over 1.5 billion in 2022, and projected to be 3.25 billion by 2028 [2]. IoT-enabled solutions, characterized by interconnected sensors, cloud- and edge-based data analytics platforms, and advanced control systems, are redefining how buildings are managed, optimized, and sustained [3]. These advancements facilitate the monitoring and control of Heating, Ventilation, and Air Conditioning (HVAC) systems, lighting, and other critical infrastructure, enabling unprecedented operational efficiency [4].

However, this technological evolution occurs against a backdrop of escalating energy demands and environmental urgency. Buildings account for approximately 30–40% of global energy consumption, nearly 28% of carbon emissions [5]. In Hong Kong, buildings consume over 90% electricity [6]. The great share of building energy consumption underscores their pivotal role in achieving the United Nations' Sustainable Development Goals (SDGs) for 2030, particularly SDG 7 (affordable and clean energy) and SDG 11 (sustainable cities and communities) [7]. Effective management of the built environment is not only about ensuring occupant comfort but is also imperative for energy conservation, carbon reduction, and climate resilience.

The necessity to optimize built environment management is amplified by the dual challenge of ensuring occupant well-being while minimizing carbon footprints. Modern HVAC systems, which dominate building energy use (accounting for 40–60% in commercial buildings [8]), must balance the conflicting objectives of thermal comfort and air quality with stringent energy efficiency targets. Among various factors, occupant behavior has been identified as one of the most significant influences on

building energy consumption [9]. However, traditional building automation systems often fail to integrate occupancy-related data, relying instead on fixed schedules or oversimplified assumptions about occupancy patterns [9]. This leads to energy waste from overcooling, overheating, or over-ventilating unoccupied spaces. A report [10] indicated that approximately 40% of building energy was wasted due to inappropriate temperature setpoints and mismatched occupancy schedules. This inefficiency highlights a critical gap in conventional automation systems: the neglect of occupant behavior as a dynamic variable [11]. These behaviors, influenced by cultural, psychological, and contextual factors, introduce uncertainty that rigid automation systems struggle to accommodate [9]. Consequently, the lack of occupant-centric adaptability in automation systems not only undermines energy-saving potential but also risks compromising occupant satisfaction, creating a paradox where technological advancements fail to align with human needs [12].

1.1.1 Building energy sufficiency

To achieve the goal of carbon neutrality by 2050 and ensure the sustainable development of humanity, Building Energy Sufficiency(BES) has emerged as a primary strategy for significantly reducing global carbon emissions [13]. BES refers to the provision of building-related energy services in an equitable, reasonable, and ecological manner [14]. The level of BES can be classified as insufficient, sufficient, or excessive [15,16]. When building services are insufficient, both energy use and emissions remain low; however, inadequate services can lead to serious issues, such as health risks during extreme weather, increased indoor air pollution, and other negative outcomes [17]. Conversely, when the building services are excessive, energy consumption and emissions will exceed sustainable limits, aggravating inequality and social injustice [16].

Achieving BES requires that occupants' needs are met by building energy systems, while limiting energy consumption to sustainable levels [18]. Figure 1-1[18] illustrates the three levels of BES. In subplot (a), the actual occupant demand for air conditioning

is depicted in terms of time, space, and indoor temperature. If the service provided falls short of this demand, or fails to meet personalized needs, the result is inadequate service and occupant dissatisfaction, as shown in subplot (c). Conversely, if energy services are provided to meet the highest possible demand at all times and spaces, as in subplot (d), occupants may be satisfied, but this leads to oversupply, substantial energy waste, and unequal energy use. True sufficiency is achieved only when the timing and spatial distribution of services align with actual demand, and personalized requirements are met, as illustrated in subplot (b).

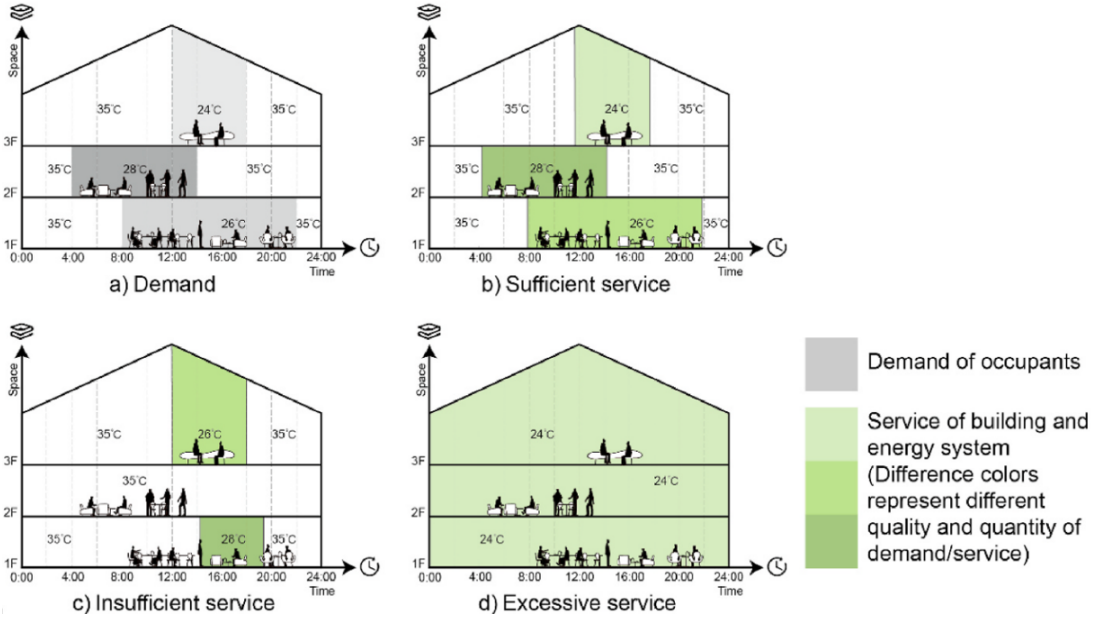


Figure 1-1. Matching the demand of occupants with building service [18]

The characteristics of occupancy behaviors in buildings can be summarized as (1) partial occupancy in space and time [19–23]; (2) diversified requirements [24,25]; (3) preference to control and adjust [26]. It is crucial to explore and address the requirements of occupants and align with their diverse requirements as closely as possible [18]. The goal is to achieve a balance between occupant demands and the system's supply.

1.1.2 Potentials of IoT-enabled advanced built environment control to achieve building energy sufficiency

The HVAC system plays a significant role in regulating indoor temperature and humidity, providing fresh air to meet the working and living needs of occupants. The conventional control system of HVAC often lack of occupant-centric adaptability, leading to lowered energy efficiency, as the loads on unoccupied spaces are still considered part of the actual demand of the HVAC system. Matching HVAC system controls to actual occupancy demand to achieve the BES is an effective way to reduce energy consumption without sacrificing occupant comfort and system functionality.

The expansion of IoT devices has enabled real-time occupancy information to become increasingly accessible through diverse IoT sensors in smart buildings, such as motion detectors and cameras. And the actual occupant needs can be more accurately identified and predicted through advanced data-driven modelling methods by exploiting the occupancy pattern. If the HVAC systems are controlled according to the actual occupants' needs, there is a high potential for achieving significant improvements in indoor thermal comfort, air quality and energy efficiency.

1.2 Aim and objectives

The aim of this PhD study is to 1) develop stochastic occupancy modeling methods to identify occupancy patterns and predict actual occupant needs, while accounting for occupancy uncertainties, 2) develop a stochastic occupancy-integrated MPC strategy to align HVAC system control with actual occupancy demand, 3) design and implement an IoT-enabled architecture that integrates advanced data-driven modeling methods and predictive optimization coupled with real-time occupant feedback to examine the effectiveness of the MPC in a real-world condition, and 4) develop the IoT retrofit scheme considering costs to boost the scalability and applicability for strategy deployment.

To achieve the aim, the following specific objectives are set and attained in this study:

(1) To develop stochastic occupancy modelling methods for predicting the actual occupant demands accounting for the periodicity and stochasticity characteristics of occupancy. The model parameters can be learnt by making effective use of the data available in the today's IoT occupancy sensors. The developed stochastic occupancy model is suitable for the applications of online occupancy prediction and occupancy-integrated MPC.

(2) To develop a stochastic occupancy-integrated MPC strategy that integrate stochastic occupant number prediction for multi-objective optimal built environment control of spaces. The stochastic differential equation (SDE)-based building dynamic modelling method is developed integrating occupancy disturbances prediction including occupancy heat gains and CO_2 generations.

(3) To develop a TRNSYS-python co-simulation platform for performance evaluation of the stochastic occupancy-integrated MPC strategy considering discrepancies of models adopted by MPC and the actual process of the target system.

(4) To design and deploy an IoT-enabled architecture that seamlessly integrate data acquisition, transmission, data-driven modeling, and model predictive optimization for advanced built environment management.

(5) To implement the Occupant-in-the-loop (OITL) MPC control strategy that integrates historical occupancy patterns, real-time occupancy data, and real-time personalized feedback through the IoT infrastructure in an office living lab to evaluate the practical effectiveness of the MPC in the real-world condition with active occupant participation.

(6) To develop the IoT retrofit scheme considering the trade-off between costs and model accuracy for the scalable and cost-effective strategy deployment. A data-driven modeling method using low-cost data is developed. The model accuracy and the associated costs are systematically analyzed.

1.3 Research scope and framework

The whole thesis is divided into 8 chapters. Figure 1-2 outlines the research framework and highlights the novel contributions of this study. The main content of each chapter is summarized as follows:

Chapter 1: This chapter outlines the background and motivation of the study, which aims to develop stochastic occupancy modeling methods and model-based control strategies for building energy systems. The goal is to align energy usage with actual occupant needs in terms of timing and quality of service, thereby achieving building energy sufficiency. This is accomplished with the support of IoT technologies in smart buildings. The chapter also presents the research aims, objectives, and overall thesis organization.

Chapter 2: A comprehensive literature review is conducted, focusing on stochastic occupancy modeling methods, occupancy-centric control strategies for the built environment, MPC strategies, and IoT architectures for building management systems. Based on the review, research gaps are identified and summarized.

Chapter 3: This chapter details the proposed Adaptive B-Spline-based inhomogeneous Markov chains (IMC) method for stochastic occupancy modeling to quantify the periodicity and stochasticity characteristics of occupant number. This method dynamically adjusts knot density of B-splines to capture variations in occupancy patterns, such as sharp transitions in different types of spaces. This method achieves significantly improved prediction accuracy with fewer model parameters compared to conventional methods. By optimally adjusting knot density, this method provides a both tailored and automated solution for occupancy modelling in diverse occupancy scenarios.

Chapter 4: This chapter elaborates on the band structure-integrated continuous-time inhomogeneous Markov chain method for stochastic occupancy modeling, which enhances model interpretability and generalization ability. The incorporated band structure within the transition rate matrix ensures both physical interpretability and computational efficiency of the model.

Chapter 5: This chapter describes the development of the MPC strategy integrating the stochastic occupancy models developed in Chapter 3 for building energy management. The MPC strategy adopts the forward Kolmogorov equations to generate future occupancy trajectories, which are then embedded into the cost function and dynamic equations of the MPC. This integration enables the optimization of HVAC control to meet the actual occupancy demand in partial occupancy condition, balancing energy efficiency with occupant thermal comfort and IAQ. Simulations conducted in TRNSYS demonstrate the energy savings, enhanced comfort and air quality improvements achieved by the stochastic occupancy-integrated MPC compared to baseline strategies.

Chapter 6: This chapter presents the implementation of the IoT-enabled OITL MPC strategy based on the control scheme developed in Chapter 5 in a real-world office living lab with active occupancy participation. Real-time occupancy data from camera-based IoT sensors and personalized feedback from QR code-based voting systems are integrated in the OITL MPC to dynamically adjust control setpoints, which are applied to the HVAC systems through the IoT infrastructure. This enhances system adaptability and satisfies occupants' diversified and personalized requirements. Field implementation verifies the effectiveness of integrating real-time occupancy feedback into predictive control systems.

Chapter 7: This chapter details the development of an IoT retrofit scheme that considers the trade-off between costs and model accuracy for the scalable and cost-effective deployment of strategy developed in Chapter 6. A data-driven modeling method using low-cost data is proposed, with systematic analysis conducted on model accuracy and the associated costs. The results indicate that the prediction accuracy of the proposed modeling method based on low-cost data is comparable to that of the modeling methods adopted in Chapter 6, which uses high-cost, high-accuracy data. The proposed method facilitates large-scale, cost-effective deployment of the intelligent built environment management system.

Chapter 8: Main conclusions and contributions are summarized. Limitations and recommendations for the future study are presented.

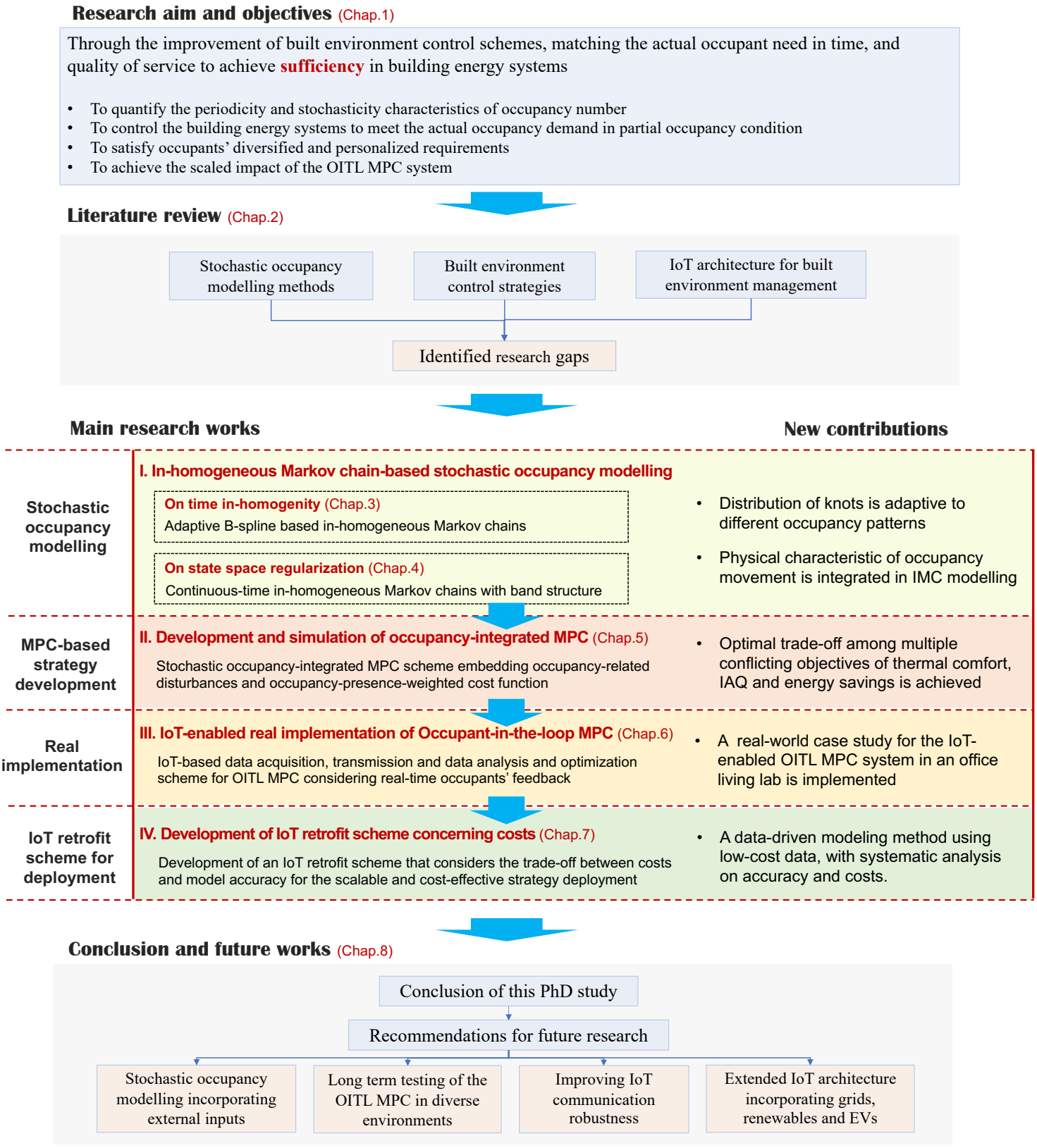


Figure 1-2. Research framework

2 Literature review

2.1 Stochastic occupancy modelling

Periodicity and stochasticity of occupancy are nature characteristics exhibited in occupancy patterns [27]; which is where stochastic models have an inherent advantage [28]. Stochastic models can consider the random characteristics of occupant number and provide probabilistic predictions[29]. Markov Chains (MC) based models have been commonly adopted for stochastic occupancy prediction in the literature [30]. A Markov chain is a stochastic process that describes transitions between states across a timespan [31]. In MC-based stochastic occupancy modeling, states can represent occupancy behavior [9] (such as adjusting the shading [32], opening/closing the window [33], etc.), occupancy presence [34,35] or different levels of occupant number [28], and transition probabilities represent the probabilities of transitioning between different states.

Markov Chain is characterized by a transition matrix describing the probabilities of particular transitions, and an initial state (or initial distribution) across the state space [36]. Based on whether the transition matrix varies with time, the MC can be classified as (time) homogeneous or (time) in-homogeneous. The in-homogenous Markov chain (IMC) is defined by a time-varying transition matrix, which is constructed using transition probabilities that change over time [37]. One challenging problem of modeling an IMC is to estimate a high number of parameters [38] that capture the time-inhomogeneity characteristic of occupancy patterns.

2.1.1 Time-inhomogeneity in Markov chains modelling

A commonly used method for estimating the parameters of an IMC for occupancy modeling, as described in the literature [27,28,39–43], is known as the "Counting Method" [44]. This method involves counting two types of occurrences: (1) transitions from one state to another and (2) transitions from one state to other states. By dividing the count of the first type of occurrence by the second type for each state at each time

slot, the transition probabilities of each state at each time slot can be estimated. Richardson et al. [40] applied the counting method to construct an in-homogeneous transition probability matrix for predicting the occupant numbers in domestic households with 10-minute intervals. Similarly, Xie et al. [42]used the counting method to estimate the parameters of an IMC for predicting hourly occupancy presence states in a hotel room, resulting in $2 \times 2 \times 24$ parameters. In another study, Salimi et al. [43] employed the counting method to estimate the parameters of an IMC for predicting the occupancy working states (a total of 5 states) in an open-plan office. The resulting matrix had dimension of $5 \times 5 \times 288 \times 7$, representing a 5-min time-step for one day (i.e., 288 time slots) and each day for a week (i.e., 7 days). Other literature also utilized the similar counting method with slight variations, such as incorporating a forgetting factor [39], integrating a moving-window [27,28,45], and employing frame-scene analysis [41].

While this simple counting process allows for the estimation of the time-varying transition matrix of the IMC, the counting method exhibits high model complexity. Since the matrix is built separately at each time slot, the parameters are not tied together or regularized in time, and there is no temporal-correlation between each matrix at nearby time slots. As a result, a large number of model parameters are needed to estimate the inhomogeneous transition matrix for every time slots. The high model complexity poses challenges in computational efficiency for large-scale deployment of stochastic occupancy prediction models in building clusters or large complex buildings with numerous rooms.

To reduce the number of model parameters, several assumptions can be made. First, it is presumed that the occupant number in office exhibits periodic variation on weekdays. Second, it is assumed that the transition probabilities are not discontinuous over time but rather show temporal-correlation with nearby time slots and continuously vary over time [38]. Based on these assumptions, smoothing functions are suitable for estimating the parameters of the time-varying transition matrix of the IMC. Splines, which are piece-wise polynomials connected by a vector of knots [46], are commonly used as

smoothing functions in various applications, such as digital image correction [47], precipitation prediction [48], and more. Splines provide a flexible approach for function approximation and interpolation.

In the field of building environments, there are also studies that have adopted spline-based methods to estimate dynamic solar gains for indoor thermal dynamic modeling [49–51]and to interpolate the transition probability matrix of the IMC for occupancy modeling [34,38]. Wolf et al. [38] employed the B-Splines in a hidden Markov-Switching model to identify five occupancy activity levels in a school classroom and a Danish summerhouse. Similarly, Anderersen et al. [34] adopted the B-Splines for dynamic modeling of occupancy presence in single-person and multi-persons offices using in-homogeneous Markov Chains.

However, previous studies have adopted a uniform distribution of knots, where the distances between knots are equal, to construct the B-Splines [49]. This approach does not guarantee optimal interpolating performance for capturing occupancy patterns in terms of both model accuracy and model complexity. Increasing the number of knots is necessary to improve forecasting accuracy, but it also results in increased model complexity. This becomes particularly challenging for rooms with significant occupancy fluctuations within short time intervals, such as meeting rooms. Additionally, knots placed in periods where rapid fluctuations of occupancy do not occur, such as early morning and late-night periods, are redundant and not effectively utilized.

2.1.2 State space regularization in Markov chains modelling

Discrete-time IMC (DTIMC) has naturally become a standard approach over the past few decades to model the occupancy pattern [27,31,34,35,38]. Despite their wide utility, DTIMC-based modelling methods face several limitations.

First, the number of model parameters increases rapidly with the number of states, making it challenging to incorporate additional states for finer modeling. Since the transition probabilities are established in discrete-time, the DTIMC modelling method exhibit high model complexity in the state space. Between two discrete time steps, there

is no restriction or constraint on changes in the number of occupants in DTIMC model. Therefore, it is necessary to model almost each element within the state transition probability matrix individually. Due to the row-sum constraint of the transition probability matrix, only one element in each row of the matrix is non-free. Therefore, modelling a DTIMC with s states and smoothing function containing N parameters requires estimating $N \times s \times (s-1)$ free parameters in total, which grows quadratically with the number of states. Second, this model complexity introduces a significant problem of overfitting to training data. Spikes and non-smoothness are common in the identified time-varying transition probabilities [52]. The DTIMC model fits the training data too closely and is likely to learn the noise and random fluctuations of occupancy sensor, potentially resulting in poor generalization to unseen data. Another significant limitation of DTIMC models is their rigid applicability in fixed time interval. For example, the parameters identified on data sampled at 10-minute intervals cannot be directly used to make predictions at finer intervals, such as 5-minute or 1-minute resolutions. Such cases require resampling the training data and re-identifying a new model. Therefore, while DTIMC methods have been widely adopted for a long time, their inherent limitations highlight the need for more flexible and efficient approaches to model occupancy patterns.

The continuous-time Markov chains (CTMC) model transition rate between states in continuous time. In a CTMC, the state transitions are governed by a transition rate matrix, where each entry indicates the rate at which the system transitions between states in a small (infinitesimal) time increment. Over the decades, CTMC has been applied in modelling dynamics of some nature processes, such as chemical reaction kinetics [53], species biological diversity [54,55], and cloud cover transition [56]. In the research of modelling cloud cover transition [56], the cloud cover is measured as an integer between 0 and 9, where 0 corresponds to completely clear sky, 8 to completely overcast sky, and 9 indicates that the cloud cover is unobservable. In the CTMC-based modelling of the transition among the ten states of different levels of cloud cover, some kind of band structure in the transition rate matrix was adopted. This band structure

indicates that the process of cloud variation can only jump to its immediate neighbors in a small time increment. For example, slim cloud (state 1) can only jump to completely clear sky (state 0) or to slightly more cloud (state 2), but not to completely overcast sky (state 8) in a small time increment.

Inspired by the cloud cover modelling, in fact, such a band structure also exists in the physical process of changes in the occupant number in a room, that is, from 1 people to empty or to 2 people but not to a large amount of people, say 30, in a small time increment. The band structure imposed by the continuous-time model reflects the continuous-time dynamics of the system. In physical systems, this is equivalent to imposing physical constraints. Note there is no physical constraint in discrete-time MC modelling method. In the occupant number modelling, it may be logical to constrain or limit the number of persons entering the room in a small time interval. This can be thought of as a kind of prior information that is put into the model and therefore constraining it to act in the imposed way and hence introducing regularization in state space.

As mentioned earlier, discrete-time MC models require estimating $O(s^2)$ parameters. However, by imposing the band structure in the CTMC modelling, it can reduce the number of parameters to $O(s)$, which is statistically more robust and reduces the risk of overfitting. Physical models are characterised by their interpretability. And since continuous-time models incorporate physical dynamics, as a consequence, the parameters are often easy to interpret. This enables the modeller to relate and physically validate the model with the real-life system in a natural way.

To the best of the author's knowledge, no previous study has developed the band structure that reflects the characteristics of occupant number variation and incorporated the band structure in CTMC-based stochastic occupancy modelling method. Besides, the band structure previously used for cloud cover modeling may have a limitation that it only allows transitions between immediate adjacent states. Within this structure, it might struggle to capture the characteristics of rapid cloud cover changes. Similarly, in occupancy modeling, such a band structure, which only allows transitions between

adjacent number of occupants, may be inadequate for identifying occupancy characteristics in rooms with drastic occupancy changes. It is necessary to extend this band structure to enhance the applicability of the model to the physical process with the characteristic of drastic changes. In addition, no prior research has conducted a comprehensive evaluation comparing the DTIMC and the CTIMC with band structure stochastic occupancy modeling methods, particularly in terms of the time-variation of different state transition probabilities, the number of model parameters (model complexity), and the prediction accuracy on both the training and testing datasets (model generalization ability).

2.2 Built environment control strategies

2.2.1 Conventional control strategies

The conventional control strategy commonly used in modern BAS allows users to manually adjust the temperature setpoint via local thermostats, with no direct intervention from the BAS. Meanwhile, the fresh air flow rate setpoint remains fixed, predetermined by operators based on the room's designed occupancy capacity. Under this framework, HVAC systems operate continuously 24/7, maintaining the set temperature and ventilation rates regardless of actual occupancy conditions.

While this setup is relatively simple and grants users full control over temperature adjustments, allowing them to modify the setpoint whenever they feel too hot or cold, it comes with significant drawbacks. A common issue arises when users leave the room without resetting the thermostat, resulting in substantial energy waste during unoccupied periods. Additionally, the fixed fresh air flow rate often leads to over-ventilation, making it difficult for operators to implement energy-saving strategies. Furthermore, if a previous occupant sets the temperature too high or too low, incoming users may experience immediate discomfort upon entering the room.

Overall, these limitations highlight the inefficiencies of conventional control strategy, underscoring the need for smarter, more adaptive solutions that optimize both energy

management and occupant comfort.

2.2.2 Occupancy-based control strategies

With the developments of Internet of Things (IoT) devices and information communication technologies in smart buildings, real-time occupancy information, such as occupancy presence and number, is increasingly accessible through a variety of sensors like motion sensor [39] and camera [52]. As a result, "occupancy-based control" (OBC) strategy [29,57], as an easy and straightforward way of integrating realistic occupancy in built environment control, is gaining research and application attention. OBC strategy typically provides control setpoints in response to real-time occupancy information with simple logic rules [58]. For example, OBC strategy for temperature control [29,59,60] adjusts indoor temperature setpoints once change in indoor occupancy presence is detected. In addition, the OBC strategy for ventilation control [61] determines the required fresh air flow rate for a building zone based on real-time occupant numbers, which is also known as the occupancy-based demand controlled ventilation (Occupancy-based DCV)[62].

A lot of literature agree that OBC strategy can achieve significant energy savings during unoccupied periods compared to fixed-schedule strategy[29,59,60]. However, OBC strategy with the simple control logic rules that solely reactive to occupancy may be difficult to account for numerous complex factors influencing built environment, such as outdoor weather conditions, thermal mass, HVAC system capacity, potentially leading to suboptimal thermal comfort[12] and IAQ [63]. It is reported that rooms employing OBC strategy may remain uncomfortable for a period after occupants enter [29], even though the OBC strategy adjusts temperature setpoint in time. Occupancy-based DCV, which adopts fixed per-person fresh air flow rates based on designed space type, may fail to account for the variability of human activities [64].

2.2.3 Model predictive control strategies

To address the challenges associated with OBC strategy, control systems can leverage

MPC strategy that integrate occupancy prediction for proactive management rather than reactive responses [12]. MPC allows for the incorporation of disturbance prediction models alongside real-time measurements to effectively handle complex endogenous and exogenous disturbances. This approach formulates a numerical optimization problem to achieve multiple control objectives [65]. In the past decade, the development of MPC for built environment control has attracted increasing research interest [27,66–69].

2.2.3.1 Simulation study

Most studies developed the MPC integrating occupancy prediction for built environment control through simulations [27,39,65,67,70].Some studies [67,71] formulated MPC in simulation assuming the future occupancy in prediction horizon is fixed and equal to the real-time value in occupancy dataset. However, this approach remains fundamentally a reactive control to real-time occupancy rather than constituting proactive control based on occupancy pattern.

Occupancy presence prediction [39,65–70,72,73] has been integrated in MPC for built environment control. However, in rooms with highly stochastic occupancy patterns, only consider the occupancy presence may be insufficient to capture the impact of occupancy-related disturbances, including occupancy heat gains and CO_2 generation, which are critically important factors influencing thermal comfort and IAQ.

Recently, a study [27] formulated an MPC for indoor temperature control and ventilation control integrating occupant number prediction. Various occupancy prediction methods, such as Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short-Term Memory network (LSTM), IMC were compared. The predicted occupant number was used to determine the fresh air flow rate in pandemic situation. However, this study employed static calculation of fresh air flow rate using simple product of the number of people while not characterizing the dynamic cumulative effect of pollutants emitted by occupants over time. In addition, the influence of heat disturbances on the built environment associated with occupant

number was also omitted.

Furthermore, since the model errors, or discrepancies between the target system and the prediction models are inevitable, using a different model for performance evaluation of the MPC is critically important to test the realistic performance. However, the model discrepancy issue was omitted in the most past simulation researches [27,39,67,69,72,73].

2.2.3.2 Experiment study

Most studies were conducted and tested through simulations, with only a few implemented and tested in real buildings [74–77]. Many of the real implementation studies [68,76,78] assumed fixed occupancy schedules, overlooking realistic occupancy data and the uncertainties of occupancy behaviors.

Goyal [68] assumed fixed occupancy schedule in the formulation of MPC and tested under this assumption. Ham [78] tested the offset-free MPC assuming occupancy heat gains a constant disturbance. However, the time-inhomogeneity of occupancy disturbances is not considered, and the estimated constant value has unsatisfied prediction abilities for the time-varying disturbances [79]. De Coninck et al. [76] estimated occupancy levels using plug load data and directly applied the occupancy levels from the prior week as predictions for the current week in the formulation of MPC. This method may be suitable for space types with stable and consistent occupancy patterns but may not well-suited for spaces with variable and volatile occupancy patterns.

Furthermore, existing studies primarily emphasized technical optimization rather than occupant-centric adaptability. None have integrated real-time occupant feedback mechanisms during experimental tests, leading to rigid automation systems where users are unable to promptly report discomfort or directly influence automated decisions. Testing MPC systems in real-world buildings with active occupant participation would be highly valuable, as it would account for the uncertainty of occupancy behaviors and enhance responsiveness to real-time occupant feedback.

2.3 IoT architecture for built environment management system

In the research field of IoT-enabled built environment management system, some articles have proposed the concept of a three-layer architecture for IoT-enabled built environment management systems, comprising the perception layer, network layer, and application layer [3]. The perception layer is responsible for sensing and data acquisition. It typically comprises various IoT devices and sensors, such as temperature, humidity, CO_2 sensors, occupancy detectors, and smart meters, which continuously monitor the physical environment and human activities within buildings. The network layer ensures reliable transmission of sensor data from the perception layer to higher-level platforms. Serving as the communication backbone, it utilizes diverse communication protocols and technologies—such as Wi-Fi, Zigbee, LoRaWAN, and 4G—to facilitate both local and cloud-based data exchange, supporting scalability and interoperability across heterogeneous devices. At the top of the architecture, the application layer hosts a suite of functional modules that enable data-driven building management. Some articles [1-3] have highlighted this layer's holistic capability, encompassing data storage and management, real-time data visualization, advanced analytics, and automated decision-making. Specifically, this layer supports functions such as energy management [80], occupant safety [81] and comfort enhancement [82]. However, in the real implementation of the IoT architecture, the majority focus on real-time monitoring and data visualization [10-13]. Some studies have explored remote manual control [86,87] or rule-based control by employing rule engines [88,89]. Kelly et al.[82] implemented an IoT system for indoor environmental condition monitoring and utility usage for residential buildings. This system used ZigBee wireless sensor network and a gateway to bridge information transformation. Dragulinescu et al. [83] developed a collaborative IoT platform that monitors home environment and alerts the inhabitants when a danger occurs. Mataloto et al.[87] implemented a LoRa-based building energy management system, where users could browse data and sent control

signal from the front-end application. Bashir and Gill [90] proposed an integrated framework of IoT big data analytics for monitoring and controlling the building in real time. Their simulation encompassed environmental parameters such as oxygen levels, smoke or hazardous gases, and luminosity across multiple zones, with data processed in a Cloudera Hadoop Distributed File System to trigger control actions when measurements deviated from predefined comfort thresholds. Despite these advancements, several limitations persist. First, the control strategies implemented are primarily comfort-oriented and generally neglect energy consumption. Second is the lack of validation in real-world IoT environments, being limited to data analysis and visualization rather than closed-loop control.

Crucially, there remains a notable gap in the literature regarding data-driven modeling and control methodologies that fully exploit IoT data, particularly those capable of integrating stochastic occupant behavior modeling and building dynamics. Even fewer studies report on the deployment of optimized, model-based control strategies that leverage these data-driven models to achieve optimal setpoints in building environment management.

2.4 Summary of research gaps

A comprehensive literature review is conducted in this chapter, focusing on stochastic occupancy modeling methods, built environment control strategies, and IoT architectures for building management systems. Based on the review, research gaps are identified and summarized as follows.

(1) High model complexity in stochastic occupancy modeling

While IMC have been widely adopted for stochastic occupancy modeling, existing methods result in high model complexity due to the large number of parameters required for time-varying transition matrices. This issue is particularly pronounced in large-scale buildings with numerous rooms, where computational efficiency becomes a challenge. Despite attempts to reduce complexity using smoothing

functions like splines, the uniform distribution of knots often fails to optimize the balance between model accuracy and complexity. Further improvements are needed to dynamically adjust knot density based on occupancy fluctuations.

(2) Overfitting and poor generalization in occupancy models

Discrete-Time Inhomogeneous Markov Chain (DTIMC) models often face overfitting issues, especially during periods with sparse data transitions (e.g., early morning and late-night). These models tend to learn noise and random fluctuations, which negatively impact generalization to unseen data. Additionally, the fixed time-step nature of DTIMCs limits their flexibility, making them unsuitable for predictions at finer or coarser time intervals without retraining. A more generalizable approach is required to overcome these limitations.

(3) Suboptimal performances of occupancy-based control strategies

Although OBC strategies utilizing real-time occupancy data have shown significant energy-saving potential compared to fixed-schedule strategies, their reactive nature often leads to suboptimal thermal comfort and indoor air quality (IAQ), particularly during the periods of people getting in the room or in rooms with highly stochastic occupancy patterns.

(4) Limited integration of occupancy prediction in MPC strategies

Although some studies have integrated occupancy prediction into MPC for built environment control, several limitations remain. Most approaches rely on occupancy presence predictions or fixed assumptions of real-time occupancy, which fail to account for highly stochastic patterns and the cumulative effects of occupancy-related disturbances, such as heat gains and CO2 generation. Additionally, while advanced prediction methods have been explored, their integration often simplifies critical dynamics, such as pollutant accumulation and heat disturbances. Moreover, the lack of consideration for model discrepancies during performance evaluations limits the reliability of simulation-based MPC studies in real-world applications. Addressing these gaps is essential for developing robust and practical MPC systems.

(5) Lack of real-world implementation of MPC with occupant personalized feedback

While MPC strategies integrating occupancy prediction have been studied in simulations, real-world implementations remain scarce. Existing implementations often rely on fixed occupancy schedules, neglecting realistic and uncertain occupancy behaviors. Furthermore, most studies adopt rigid automation systems without incorporating real-time occupant feedback mechanisms. This limits adaptability and responsiveness, as occupants are unable to promptly report discomfort or influence automated decisions. Experimental testing of MPC systems with active occupant participation is crucial to address these gaps.

(6) Limited integration of data-driven models and advanced control in IoT-enabled built environment management systems

IoT-enabled built environment management systems are primarily focused on data monitoring, visualization, and basic rule-based control. There is still a lack of data-driven modeling and control methods for effective utilization of IoT data, particularly those that incorporate occupant pattern modeling and building dynamics modeling. Even rarer are implementations that utilize these models for optimized automatic control in real-world scenarios. The lack of such comprehensive systems limits the potential for achieving energy sufficiency and occupant-centric building management.

3 Adaptive B-spline-based inhomogeneous Markov chains for stochastic occupancy modelling

This section introduces a novel method called the *Periodic B-Spline with Adaptive Knots* for estimating the parameters of transition matrix of the IMC in stochastic occupancy modeling. This method optimizes the distribution of knots based on specific characteristics of occupancy exhibited in different types of room. For example, residential rooms typically have occupants indoors during the evening and vacant during the daytime, while office rooms have occupants during the daytime and are empty at night. Meeting rooms, on the other hand, are often unoccupied for most of the time with occasional periods of high occupancy when a large number of people enter and exit within a short duration. The proposed Adaptive B-Spline method automatically adjusts the density of knots, reducing it in areas with less variation in occupancy and increasing it in areas with significant changes in occupancy. This allows for better capturing the characteristics of occupancy patterns. The effectiveness of the proposed method is demonstrated using six months of occupant number data collected within a meeting room. A comprehensive comparison is conducted between the proposed adaptive B-Spline method and other methods, including counting method, and uniform B-Spline method. The comparison is based on both model accuracy and model complexity.

Figure 3-1 shows the methodology. It includes three parts, (1) In-homogeneous Markov Chain modeling, (2) parameter estimation, and (3) model evaluation. The In-homogeneous Markov Chain modeling methods in the literature (counting method and uniform B-Spline method) are first provided, followed by the proposed adaptive B-Spline method. The second part introduces how the parameters in transition matrix are estimated using the maximum log-likelihood method. The third part introduces the model evaluation method. The Akaike information criterion (AIC) and Bayesian information criterion (BIC), which balance the model accuracy and model complexity

are used to compare the overall performance of methods.

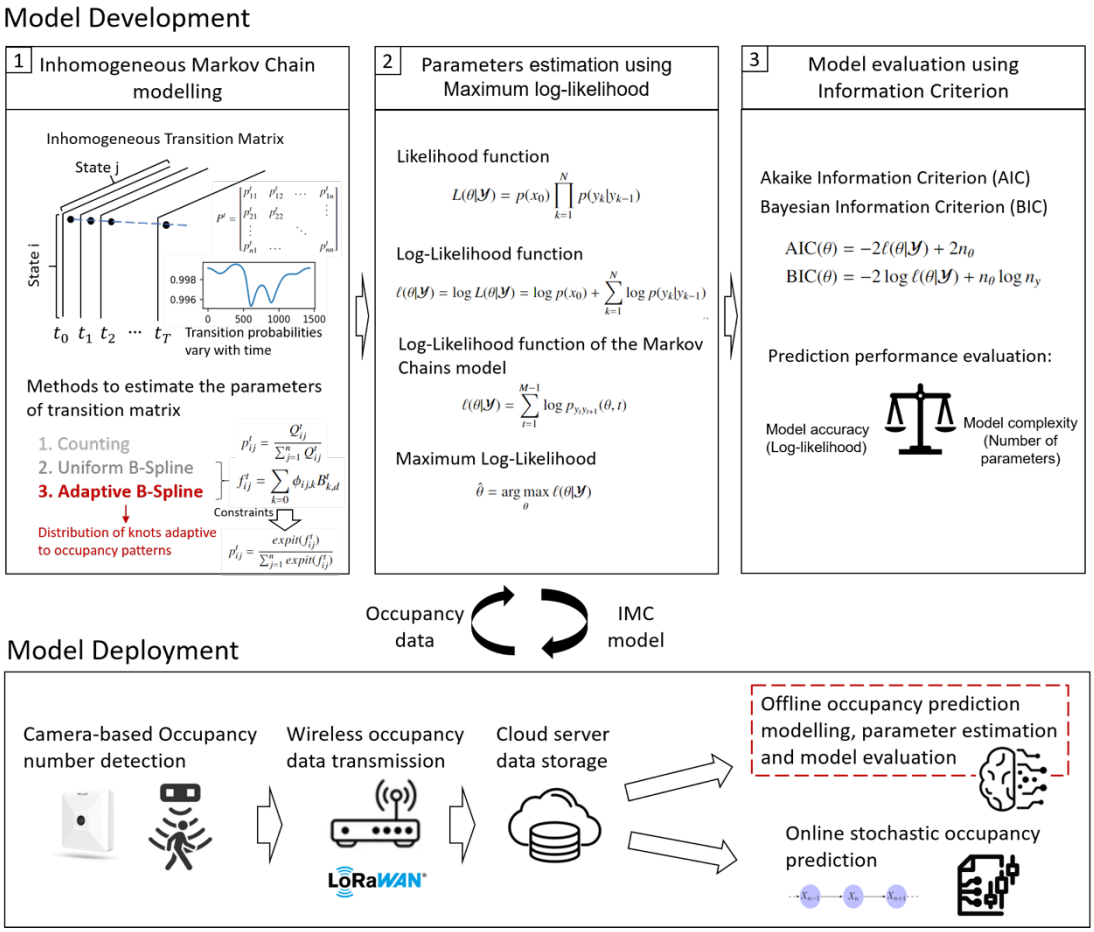


Figure 3-1. Methodology of the proposed adaptive B-spline-based IMC occupancy modelling method

3.1 Brief introduction of Markov chains

The method in this chapter is of the following kind; Markov chains. The Markov property can be simply stated popularly by: Given the present, the future is independent of the past. To more formally state this, let $\mathcal{Y}_k = \{y_1, y_2, \dots, y_k\}$be a collection of observations of an underlying stochastic process $\{X_t; t \geq 0\}$ (in discrete or continuous time) at times $t_1 < t_2 < \cdots < t_k$. Then the Markov property can be stated as:

$$\mathbf{E}\{X_t|\mathcal{Y}_k\} = \mathbf{E}\{X_t|y_k\} \tag{3–1}$$

for $t \geq t_k \geq 0$.

This model class includes a vast number of models and is particularly popular [30,91]. A homogeneous, discrete-time Markov chain can be written in the following way:

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & & \vdots \\ \vdots & & \ddots & \\ p_{s1} & \cdots & & p_{ss} \end{bmatrix} \tag{3–2}$$

where s is the number of states in the Markov chain, $\boldsymbol{P}$ is the so-called transition probability matrix. The matrix has the properties that each row is a margin distribution, i.e. each row sums to 1 and describes the probabilities of the process jumping to each state. The i'th row is thus the probabilities of the process jumping from the i'th state to each other state.

Note also, that none of the transition probabilities, p_{ij}, are time dependent; hence *homogeneous* in time. This is, however, often not a sufficient assumption, e.g. in this example, where the meeting room is much less likely to go from unoccupied to occupied during the night compared to day.

3.2 Inhomogeneous Markov chains

In-homogeneous Markov chains (IMC), on the other hand, allow for time-dependent transitions, where the transition probabilities between states change over time. An IMC can be written as:

$$\boldsymbol{P_t} = \begin{bmatrix} p_{t,11} & p_{t,12} & \cdots & p_{t,1s} \\ p_{t,21} & p_{t,22} & & \vdots \\ \vdots & & \ddots & \\ p_{t,s1} & \cdots & & p_{t,ss} \end{bmatrix} \tag{3–3}$$

It should be noted that the properties of the matrix still hold. The elements in the matrix change over time meaning that the parameters to be estimated in IMC are significantly higher than the homogeneous Markov Chains. The counting method is commonly used in past literature [27,28,39–41,45] to estimate the time-varying transition matrix, which is elaborated in the following section.

3.2.1 Counting method

The counting method calculates the occurrence from one state jump to another state.

The transition probability of $p_{t,ij}$ is estimated as follows,

$$p_{t,ij} = \frac{Q_{t,ij}}{\sum_{j=1}^{n} Q_{t,ij}} \tag{3–4}$$

where $Q_{t,ij}$ is the occurrence from state *i* to state *j* at time step *t*; $\sum_{j=1}^{n} Q_{t,ij}$ is the sum of occurrence from state *i* to other states at time step *t*. Estimation of transition probability matrices for every time step $\mathrm{t} = 1, \ldots, \mathrm{T}$ of the entire available data would result in $\mathrm{T} \times \mathrm{s} \times (\mathrm{s} - 1)$ free parameters, which means that the model complexity is very high. We separate the data into weekdays and weekends, and hence use all data from (e.g.) 10 o'clock to form the transition matrix for that transition time. For example, if a time step of one hour within a day is considered, it would need 24 matrices; if a time step of 5 minutes is used, it would require 288 matrices. We consider time steps of 1 minutes, which results in 1440 matrices. The counting method estimates the parameters of the transition probability matrix separately at each time step, resulting in discrete values rather than continuous and lacking temporal correlation between the matrix.

To reduce the number of model parameters, smoothing functions, such as Splines are suitable for estimating the parameters of the time-varying transition matrix of the IMC. This study proposes a method using periodic B-Spline with adaptive knots for constructing the spatial-temporal correlation of the transition probability matrix, which improves the Uniform B-Spline (knots are equal-distant). The details of the B-Spline methods are elaborated in the next subsections.

3.2.2 B-Splines method

Splines are piece-wise polynomials that have continuous differentiability up to a certain order. The points at which these piece-wise polynomials connect are called knots. Splines are in practice commonly used for interpolation, smoothing, and function approximation [92], due to several advantageous properties. They exhibit smoothness up to a certain order, possess compact support, and have explicit and relatively simple formulas [49].

This chapter aims to estimate the In-homogeneous transition probabilities from measured occupant number data using Splines as smooth functions of time of day. The proposed method is based on a specific type of spline function: the B-spline. By defining a knot sequence, a set of B-splines that form a basis for all splines with this same knot sequence can be constructed. The detailed introduction of B-Splines can be found in [92]. The estimated functions are obtained by summing the scaled B-splines, which is defined by:

$$f_{t,ij} = \sum_{n=1}^{N} \phi_{ij,n} B_{n,D}(t) \tag{3–5}$$

where $f_{t,ij}$ is the unconstrainted form of parameters, which need to be transformed to meet the constraints as the probabilistic variables $p_{t,ij}$. The parameter transformation is detailed in Section 3.2.3. $\phi_{ij,n}$ is the scaling coefficient, which is the model parameter need to be estimated. The number of scaling coefficients to be estimated for interpolating the functions is *N*. $B_{n,D}(t)$ is called the blending functions corresponding to the n^{th} scaling coefficient, which is piece-wise polynomials connected by a vector of knots $(h_1, \ldots, h_{\mathrm{K}})$. *D* is the order of blending function, which is one greater than the degree of the polynomials that define the curve segments [92]. Typically linear, quadratic, or cubic splines are used. For example, the order of a spline (*D*) based on cubic polynomials is 4. Vector of knots is a monotonically increasing list of floats, $h_{\mathrm{k}} \leq h_{\mathrm{k+1}}, 1 \leq \mathrm{k} < \mathrm{K}$. The number of knots is $\mathrm{K} = \mathrm{D} + \mathrm{N}$ (i.e., the order plus the number of scaling coefficients) [92]. $\mathrm{B_{k,d}}$ is defined by the recursive Cox-de Boor equations [92] using the vector of knots:

$$B_{k,1}(t) = \begin{cases} 1, if\ h_k \leq t \leq h_{k+1} \\ 0, \quad otherwise \end{cases} \tag{3–6}$$

$$B_{k,d}(t) = \frac{t - h_k}{h_{k+d-1} - h_k} B_{k,d-1}(t) + \frac{h_{k+d} - t}{h_{k+d} - h_{k+1}} B_{k+1,d-1}(t) \tag{3–7}$$

Eq.(3–6) defines the blending function when $\mathrm{d} = 1$, Eq.(3–7) recursively defines the blending function when $\mathrm{d} > 1$ using the last order of blending functions until it reaches

D. Figure 3-2 shows an example of defining the cubic B-Splines ($\mathrm{D} = 4$) with 8 knots ($\mathrm{K} = 8$) and 4 scaling coefficients ($\mathrm{N} = 4$) using the recursive Cox-de Boor equation.

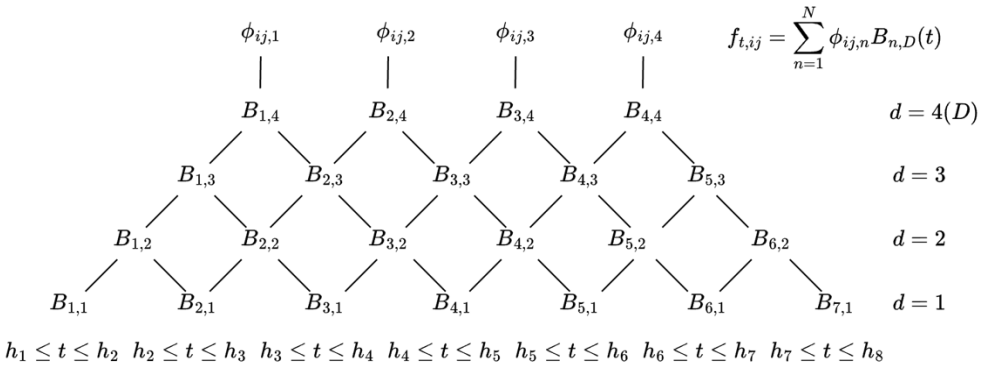


Figure 3-2. Illustration of recursive Cox-de Boor equation for defining the cubic B-Splines with 8 knots and 4 scaling coefficients

3.2.2.1 B-Splines with uniform knots

Uniform distribution of knots, where the distances between knots are equal, has been adopted to construct the B-Splines in previous studies [34,38] for IMC modeling. For example, a normalized uniform knot vector with 6 knots gives (0, 0.2, 0.4, 0.6, 0.8, 1). In many use cases, one is only interested in a series of B-splines that covers a finite range of the x-axis. However, the sum of a finite set of cardinal B-splines will go to zero at the boundary regions. Therefore, the cardinal B-splines near the boundaries can be modified to add up to unity for the entire finite domain [49] using *open uniform knots*.

The standard form of *open uniform knots* begins with *D* zeros (where *D* is the order) then increases uniformly, until ending with *D* copies of the last value. For example, for a quadratic (order 3) B-spline with six scaling coefficients, the standard normalized open uniform knot vector is (0,0,0,0.25,0.5,0.75,1,1,1).The Cox-de Boor recursion formula can be used to define B-splines with non-uniform knot distances and to shift the B-spline along the x-axis by positioning the knots at equal locations. In order to still use Eq.(3–7) in this generalized case, division by zero is defined as zero [49,92].

Figure 3-3 (a-d) illustrates the variation of B-Splines with number of knots using *open uniform knots*. The knot vector of Figure 3-3 (a) and (b) is (0,0,0,0,1,1,1,1) and (0,0,0,0,0.5,1,1,1,1), respectively. By utilizing *open uniform knots*, it is possible to estimate non-zero values at the left and right side of the spline domain for function approximation. Figure 3-3(d-f) shows the variation of B-Splines by scaling the individual basis splines ($B_{n,4}, n = 1..7$) using different scaling coefficients. When using B-Splines for functions approximation, it is necessary to estimate the optimal scaling coefficients from the measurement data. Increasing the number of knots results in finer approximation of functions, thereby improving prediction accuracy.

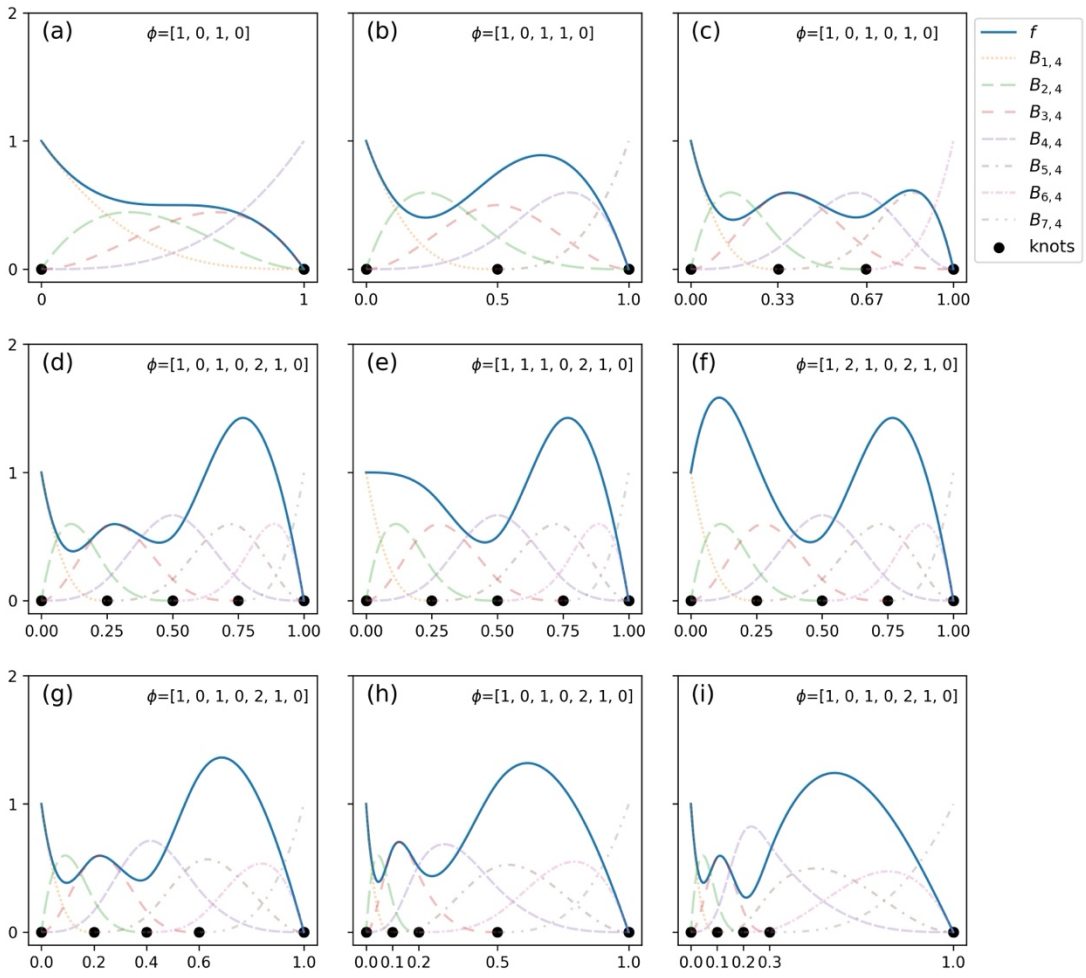


Figure 3-3. The variation of cubic B-Splines with number of knots (a-d), scaling coefficients (d-f), and distribution of knots (d,g-i)

Adopting Spline method for modeling the IMC can significantly reduce parameters and model complexity compared with the counting method. Estimation of transition probability matrix of the entire available data would result in $N \times s \times (s-1)$ free parameter, where the number of scaling coefficients (N) is significantly lower than the time steps (T).

However, this approach does not guarantee optimal interpolation performance in terms of both model accuracy and model complexity for capturing occupancy patterns. For rooms with significant occupancy fluctuations within short time intervals, such as meeting rooms, a finer granularity of knots is required, but this also increases the complexity of model. In addition, placing knots in periods where rapid occupancy fluctuations do not occur, such as early morning and late-night periods, is redundant and not effectively utilized.

3.2.2.2 B-Splines with adaptive knots

This study introduces a novel method called the Periodic B-Spline with Adaptive Knots for estimating the parameters of transition matrix of the IMC in stochastic occupancy modeling. Figure 3-3(d,g-i) illustrates the variation of B-Splines with distribution of knots. By adjusting the distribution of knots, splines can be positioned closer together in parts of their domain where more fluctuation occur, while being spaced further apart in parts of their domain where there is less fluctuation. The parameters $\Delta = (\Delta_1 \dots \Delta_{\mathrm{N-D-1}})$ represent the distances between interior knots, which are optimized along with the scaling coefficients based on the specific characteristics of occupancy exhibited in different types of rooms. It is important to note that knots positioned at the start and end boundaries do not require optimization. As described in Section 3.2.2.1, the number of knots at the endpoints in a B-spline is 2D, which are not optimized. While the positions of the N-D interior knots (non-endpoint knots) are optimized. The parameters to be optimized are the distances between these N-D interior knots, resulting in a total of N-D-1 parameters.

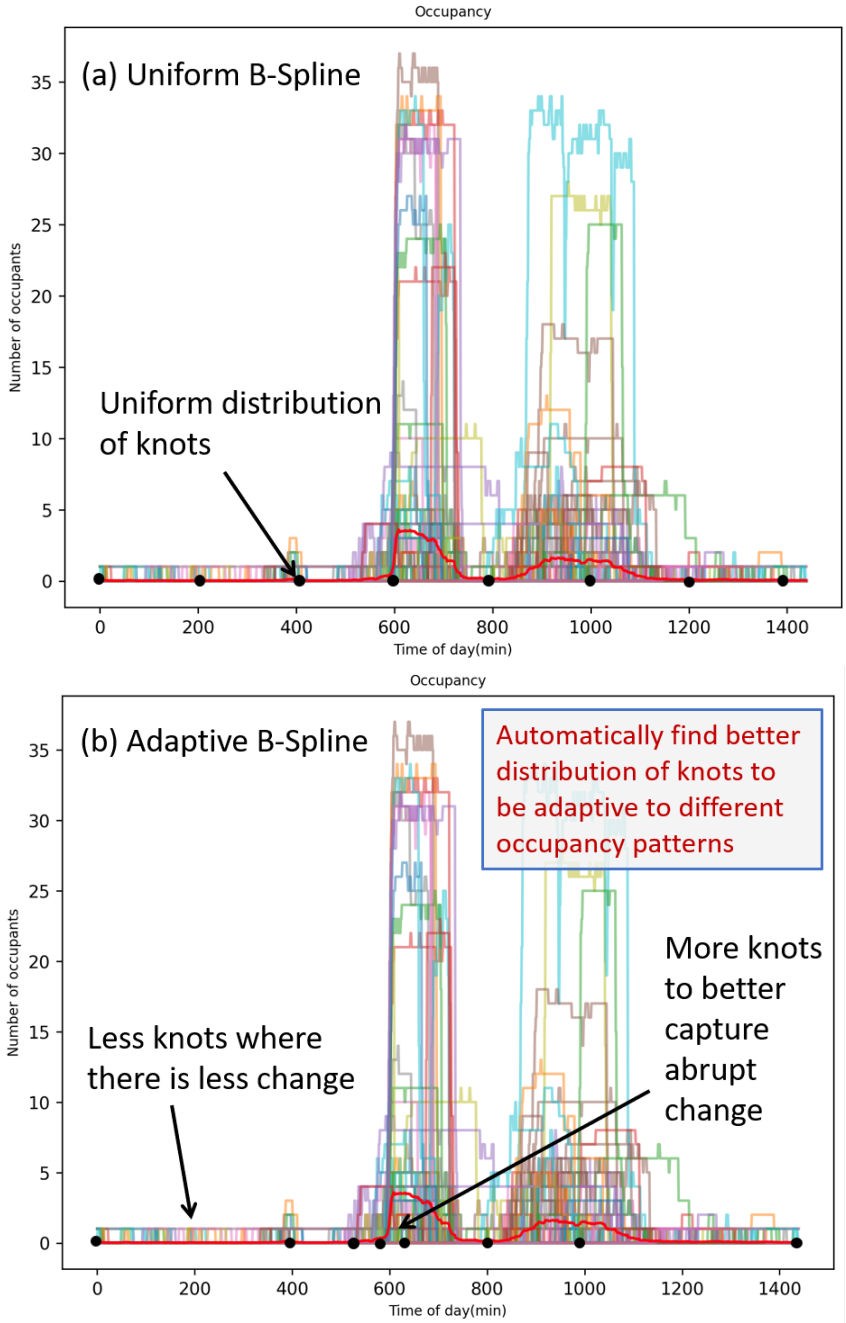


Figure 3-4. Conceptual illustration of the proposed adaptive B-Spline method (a) Uniform B-Spline: the knot sequence is uniformly spaced (b) Adaptive B-Spline: the distribution of knots is adaptive to occupancy patterns

The adaptive B-Spline method has more parameters to be optimized compared to the uniform B-Spline method, and the additional parameters correspond to the number of interior knots. When estimating the transition probability matrix using the adaptive B-Spline method with the entire available data, the number of free parameters would be

$N \times s \times (s-1) + (N - D - 1)$. It is important to note that the increase in the number of model parameters in the adaptive B-Spline method is significantly smaller in magnitude compared to the original number of parameters in the uniform B-Spline model. This is also because we use the same knot placement distribution for all B-splines in the transition probability matrix.

Figure 3-4 shows the conceptual illustration of the proposed adaptive B-Spline method. By automatically adjusting the density of knots, reducing them in areas with less variation in occupancy and increasing them in areas with significant changes in occupancy, the method ensures a better capture of occupancy pattern characteristics.

The most common orders of B-splines in practical applications are quadratic and cubic. This study uses cubic splines only in all the modelling. The reason hereof is its extra degree of freedom compared to quadratic splines, which ensures smoothness to an extra degree since the second order derivatives can be matched between knots [93].

3.2.3 Parameter transformation

A parameter transformation between a constrained and unconstrained domain is necessary to ensure that $p_{t,ij} \in [0,1]$. To apply this, the inverse logit transformation $\text{expit}(\text{x}) = \exp(\text{x}) / (1 + \exp(\text{x}))$ with range $[0,1]$ is used [38]. The row sum constraint (i.e. the sum of probabilities in each row is equal to 1) can be satisfied by dividing each entry by its row sum. Hence, the transition probabilities can be expressed by:

$$p_{t,ij} = \frac{\text{expit}(f_{t,ij})}{\sum_{j=1}^{n} \text{expit}(f_{t,ij})} \tag{3–8}$$

3.2.4 Parameter identification

This study uses the method of maximum likelihood to estimate the parameters in the Markov models, see [44] for an introduction. The maximum likelihood (ML) method is based on maximizing the transition probability density functions (as a function of the parameters) of the observed stochastic process evaluated in those observations. Written

formally, that is

$$\mathrm{L}(\theta|\mathcal{Y}) = \mathrm{p}(\mathrm{x}_0)\prod_{\mathrm{k}=1}^{\mathrm{N}}\mathrm{p}(\mathrm{y}_\mathrm{k}|\mathrm{y}_{\mathrm{k}-1}) \tag{3–9}$$

where L is the likelihood of the set of parameters θ given data $\mathcal{Y}$, x_0 is some prior initial conditions, and p is the transition density of the process. For numerical reasons, it is often advantageous to consider the logarithm of L (which preserves the maximum of L due to its monotonicity), the so-called log-likelihood function

$$\ell(\theta|\mathcal{Y}) = \log \mathrm{L}\,(\theta|\mathcal{Y}) = \log \mathrm{p}\,(\mathrm{x}_0) + \sum_{\mathrm{k}=1}^{\mathrm{N}}\log \mathrm{p}\,(\mathrm{y}_\mathrm{k}|\mathrm{y}_{\mathrm{k}-1}) \tag{3–10}$$

The optimal set of parameters are then computed by the set of parameters that maximizes the log-likelihood

$$\hat{\theta} = \arg\max_{\theta}\ell(\theta|\mathcal{Y}) \tag{3–11}$$

If we let the observation space of the Markov model be the indices of the chosen states, i.e. $\mathrm{y}_\mathrm{t} \in \{1,2,\dots,\mathrm{n}\}$, the log likelihood function in general for a Markov chain process takes the form:

$$\ell(\theta|\mathcal{Y}) = \sum_{\mathrm{t}=1}^{\mathrm{M}-1}\log \mathrm{p}_{\mathrm{y_t y_{t+1}}}(\theta,\mathrm{t}) \tag{3–12}$$

where $\mathrm{p}_{\mathrm{y_t y_{t+1}}}(\theta,\mathrm{t})$ is the $\mathrm{y_t}$th, $\mathrm{y_{t+1}}$th entry in the transition probability matrix at time *t* given θ. It is thus the sum of the logarithm of the transition probabilities.

In case of a homogeneous Markov chain model, the parameters are the transition probabilities themselves, $\theta = p_{ij}$. Here, the ML estimate of the transition probabilities are given in closed form by the counting method, $\hat{p}_{ij} = n_{ij}/N$, where n_{ij} is the total observerd number of transitions from state i to j and N is the number of data points. If the transition probabilities are parameterized somehow (e.g. using a splines expansion), it is necessary to evaluate the sum of the logarithm of the transition probabilities in Eq.(3-12) and solve Eq.(3-11). Therefore, using the form in Eq.(3-8), we need to evaluate the transition probabilities in order to evaluate ℓ and use a numerical solver.

3.2.5 Model evaluation

The models are estimated by means of the maximum likelihood method, and thus should be compared according to how well each model maximizes the likelihood function. The model that has the largest likelihood function is the most accurate, and this can be compared as long the same data is used. However, better model accuracy is not necessarily desirable since over fitting can occur. To account for model complexity and the amount of parameters in the comparisons of models, information criteria like Akaike's or Baye's may be used [44]. These take the form:

$$\mathrm{AIC}(\theta) = -2\ell(\theta|\mathcal{Y}) + 2\mathrm{n}_\theta \tag{3–13}$$

$$\mathrm{BIC}(\theta) = -2\log\ell\,(\theta|\mathcal{Y}) + \mathrm{n}_\theta \log \mathrm{n}_\mathrm{y} \tag{3–14}$$

where AIC and BIC are Akaike's and Baye's information criteria respectively, ℓ is the log-likelihood function, n_θ is the number of estimated parameters, and n_y is the number of data points. The purpose of both of these criteria is to make measures that can compare performance across different models with different model complexity. I.e., these criteria estimate the best model by weighing in the number of parameters into the objective. We say that the best model is the model that minimises the AIC and/or BIC since this model (statistically speaking) has the best trade-off between accuracy and model complexity.

AIC can be derived as an unbiased estimator of the expected Kullbach-Lieber discrepancy between the true and given model [94] (this is where the second term, $2n_\theta$ comes from). The BIC on the other hand can be derived from the Laplace approximation to the Kullbach-Lieber discrepancy, but it generally requires $n_y \gg n_\theta$ to be a sufficient approximation [95].

3.3 Case study

This section provides a brief description of the dataset used in this study. The data originates from an office meeting room at a university in Hong Kong. Figure 3-5 shows the photograph of the room. The meeting room (L × W × H: 4.5 m × 6.5 m × 3 m) has one south-facing exterior wall (4.5 m × 3 m) and one east-facing exterior wall (6.5 m × 3 m). The window-wall-ratios of the south-facing wall and the east-facing wall are 0.3 and 0.35, respectively. The room has a floor area of around 30m^2 and is used for meeting only.

The meeting room is a living lab. A camera-based IoT occupancy sensor was installed at the ceiling of the meeting room. With image detection algorithm embedded, the sensor can provide high accurate regional occupancy counting. Image blurring algorithm is also embedded internally in the device to protect privacy. According to the specification of the sensor, the accuracy of regional occupancy counting is 95% and the detection coverage area is 78m^2. More details of the sensor can be found in [96]. After detection, the total regional occupancy data will be sent to our database through a LoRa gateway. Figure 3-6 shows the detected occupants during a meeting.

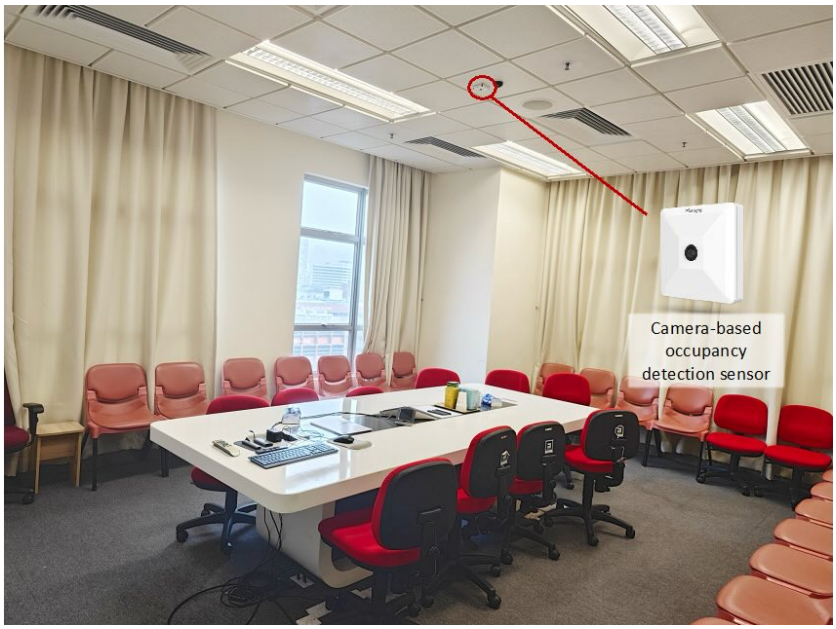


Figure 3-5. Meeting room at a university in Hong Kong

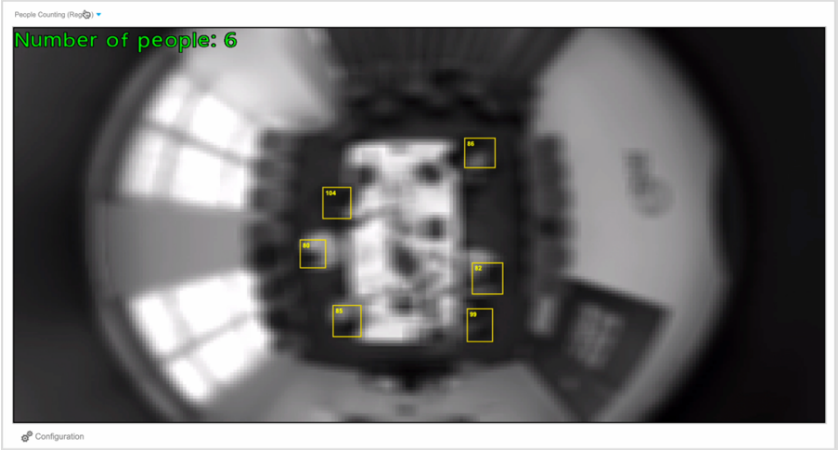


Figure 3-6. Detected occupant number during a meeting

The analysis is conducted on six-month occupancy data from 16th March to 18th September 2023, collected at a one-minute time interval, as shown in Figure 3-7. The analysis specifically focuses on workdays since the meeting room receives fewer visitors on weekends. Note that the data has undergone anomaly detection and replacement procedures. During nighttime hours (between 0 a.m. to 6 a.m.), the data may occasionally show one person detected due to darkness. To address this, a filter has been applied to identify instances where one person appears and disappears within a five-minute interval. While this filter effectively removes most anomalies, some may still remain. In future analyses, an automatic anomaly detection method, combined with multi-sensor fusion technology such as CO_2 and motion sensors, will be adopted to further enhance the data quality. Another idea we plan to carry out is to model the observation noise independently, such that the model becomes a hidden Markov model and is observed through an observation equation with modelled noise.

From the dataset, it is evident that meetings primarily occur between 10 a.m. to 12 p.m. and 2 p.m. to 6 p.m. Additionally, there are instances of people entering the meeting room between 6 a.m. to 7 a.m., primarily comprising cleaning staff. A weekly group meeting involving the entire research group takes place, during which the room can accommodate more than thirty people. However, for the remaining meeting, the number of participants in meetings generally remains below ten.

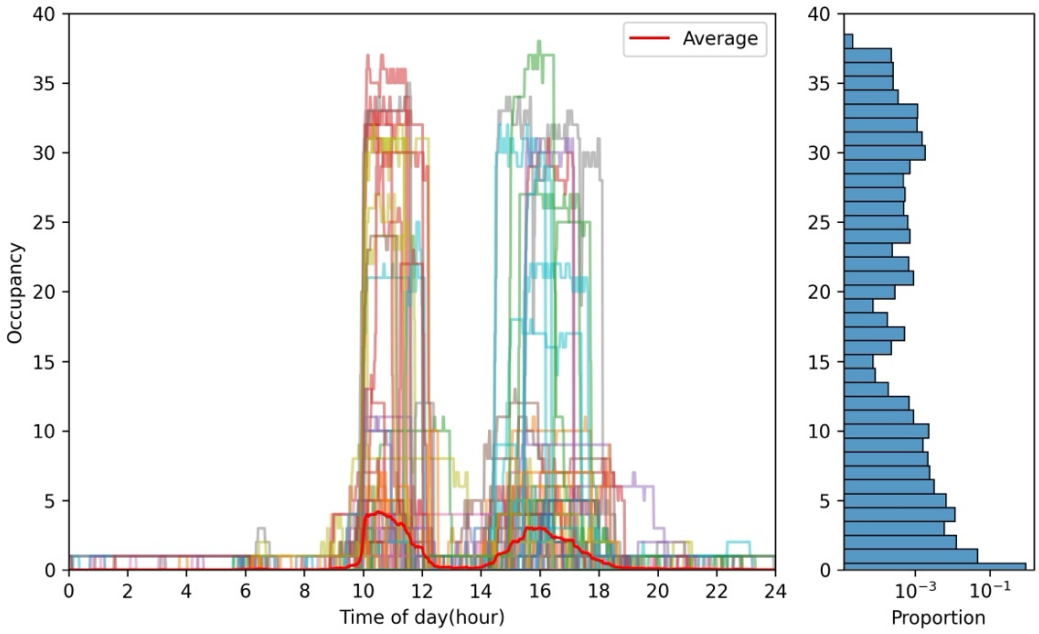


Figure 3-7. Distribution of occupancy Data

When determining the state space dimensions for the IMC model, a trade-off exists between model accuracy and complexity [38]. If the state space is too small, meaning only a few states are used to describe the process of occupancy changes, the model remains relatively simple. In the most extreme case, only two states (e.g., present/absent) are considered. However, this simplification results in low accuracy when predicting occupant numbers. Conversely, as the state space increases, the model's complexity grows rapidly—specifically, complexity increases quadratically with the number of states—making the model more intricate and difficult to identify.

To reduce the number of states while maintaining accuracy, certain ranges of occupant numbers can be aggregated into a single state. In this study, the occupancy data is categorized into nine states based on the distribution of the number of occupants to simplify the model. Table 3-1 shows the correspondence between the occupancy and states. This approach ensures a balance between model accuracy and complexity while effectively capturing the patterns of occupancy variations in the meeting room. For researchers seeking to implement a similar approach, state space discretization can be tailored based on the distribution of occupant numbers in their dataset and the specific

ranges of interest.

Table 3-1 Occupancy states

States	Occupant number	States	Occupant number
0	0	5	5
1	1	6	6-10
2	2	7	11-20
3	3	8	more than 20
4	4		

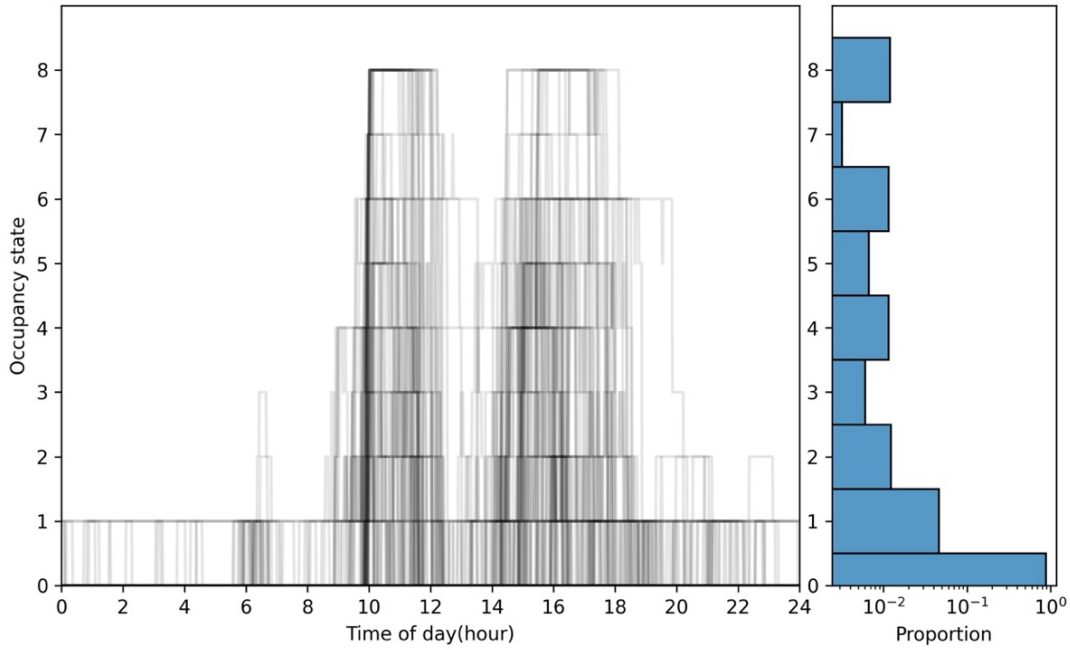


Figure 3-8. Distribution of Occupancy States

Figure 3-8 illustrates the distribution of the different states. The majority of the time, the occupancy is either none (state 0) or single person (state 1), as most observations fall within this range (less than 2 people). Small meetings, which occur frequently, correspond to occupants ranging from 2 to 5 people. Each specific occupancy count within this range is considered a distinct state, i.e., state 2 to 5. Medium-sized meetings, with occupants from 6 to 10 people, are treated as a single state (state 6). The group meeting, which involves more than 20 people and occurs weekly, is also considered a

distinct state (state 8). Occupants ranging from 11 to 20 people, which occur less frequently, are combined into a single state (state 7).

3.4 Results

This section compares the results of occupancy prediction using homogeneous MC and IMC with counting, uniform B-Spline and adaptive B-Spline methods. The evaluation is based on several metrics, including log-likelihood, the number of parameters, AIC, and BIC. Additionally, the temporal evolution of the transition probabilities of the Markov Chains is analyzed. The distribution of knots using the adaptive B-Spline method is compared to the uniform B-Spline method to demonstrate the effectiveness of the proposed adaptive approach.

3.4.1 Homogeneous Markov Chain model

Figure 3-9 shows the transition probabilities of the homogeneous Markov Chain model. It is evident from the figure that the probability of transitioning from state 0 to state 0 is the highest, reaching 0.998. This can be attributed to the typical usage pattern of the meeting room, where it remains unoccupied for the majority of the time.

Figure 3-10 displays the results of simulating the occupancy state 100 times using the homogeneous Markov Chain, along with the distribution of the simulated occupancy states. For each simulation, the initial distribution or state values of occupancy at the start of each day, specifically at 0:00 hours, are used as input for predicting the number of individuals every minute throughout the day using time-varying Markov Chains. This process was repeated 100 times, simulating 100 scenarios. In this study, the input state starting from state 0 was utilized. Noted that the simulations for other methods follow the similar process, only the transition probability matrix are different.

Although the distribution of the occupancy states closely resembles the overall distribution of the measured occupancy states (as shown in Figure 3-8), Figure 3-10 indicates that the changes in occupancy states across different time slots are similar,

which does not align with the actual situation. In reality, it is unlikely for the room to transition from unoccupied to occupied during the nighttime compared to the daytime.

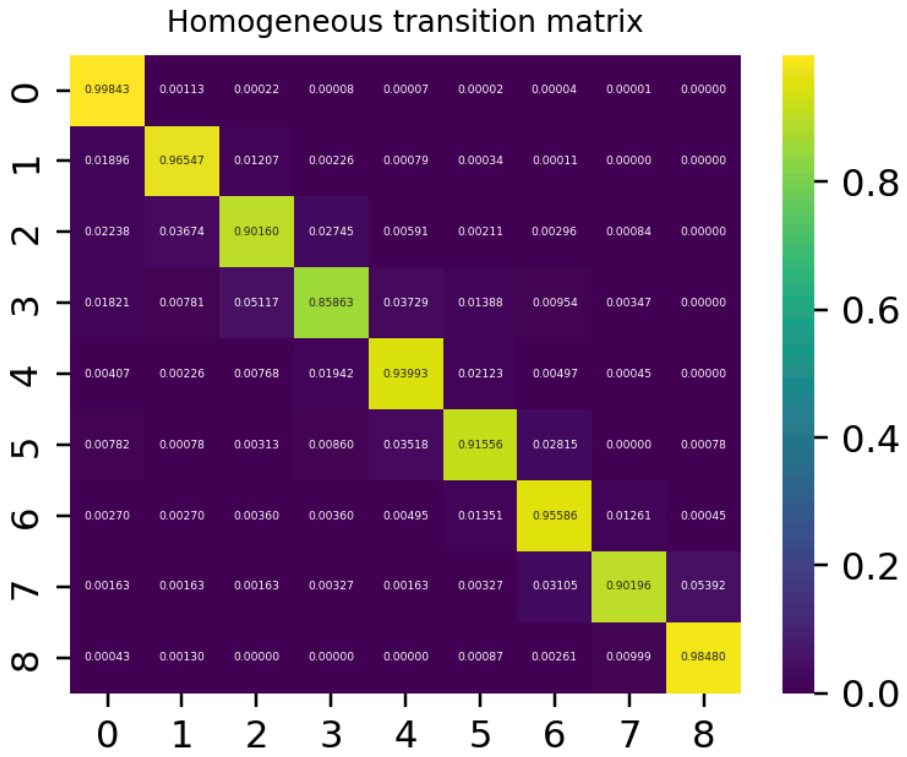


Figure 3-9. Transition matrix of the homogeneous Markov Chain

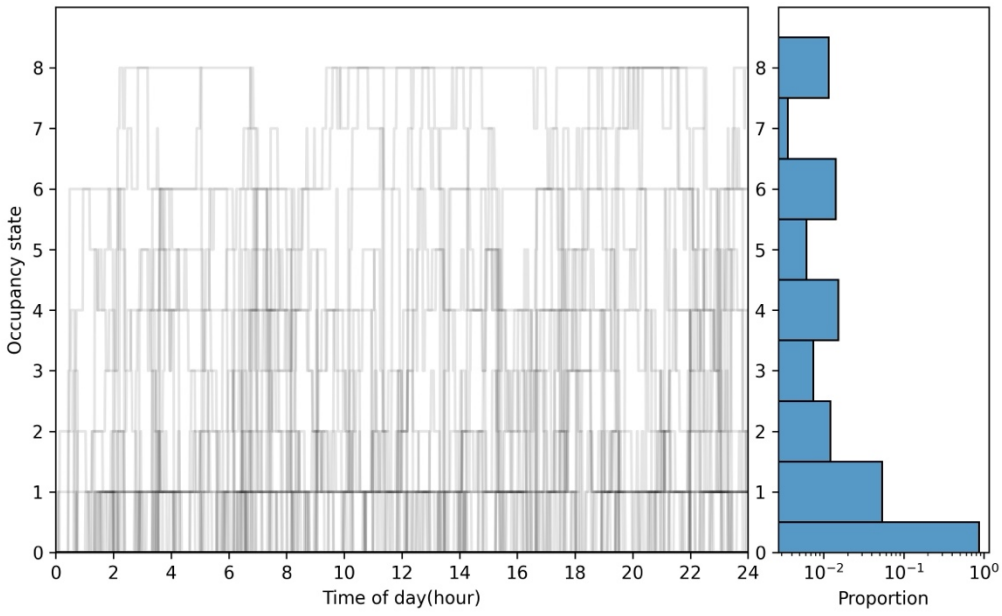


Figure 3-10. 100 simulations using the homogeneous transition matrix

Despite the homogeneous Markov chain model has only 81 parameters, its log-likelihood reaches -8035, with AIC and BIC values of 16231 and 17054, respectively.

3.4.2 Inhomogeneous Markov Chain models

3.4.2.1 Counting method

Figure 3-11 shows the in-homogeneous transition probabilities from states 0,1, and 2 to states 0,1, and 2, respectively using the counting method.

The graph located in the top-left corner (at position (0,0)) represents the transition probabilities from state 0 to state 0 in the meeting room. Around 10 o'clock, there is a significant decrease in the transition probability from 1 to approximately 0.95, indicating that it is less likely for the room to remain unoccupied during this time. This captures the change in the number of people in the room as meetings normally start at 10 o'clock in the morning. Around 12 o'clock, the probability rises again to 0.99, corresponding to the lunchtime break. Around 15 o'clock, it decreases to 0.98, indicating higher probability of afternoon meetings.

The graph at position (0,1) (from state 0 to state 1) also displays this trend of changes in occupancy states. It is worth noting that the transition probability matrix also captures the change in occupancy states from 6 to 7 o'clock when the cleaning staff enters the meeting room to clean. In this period, the transition probability from state 0 to state 0 decreases to 0.985, while the transition probability from state 0 to state 1 increases to 0.015.

Since the counting method simply calculates the frequency of transitions from one state to another, the transition probabilities between adjacent time points are discrete rather than continuous, and there is no temporal-correlation between them. For example, in one minute, the probability may be 0, but in the next minute, it may suddenly become 1. This result does not align well with reality, as in reality, the transition probabilities between adjacent moments typically do not exhibit such large variations.

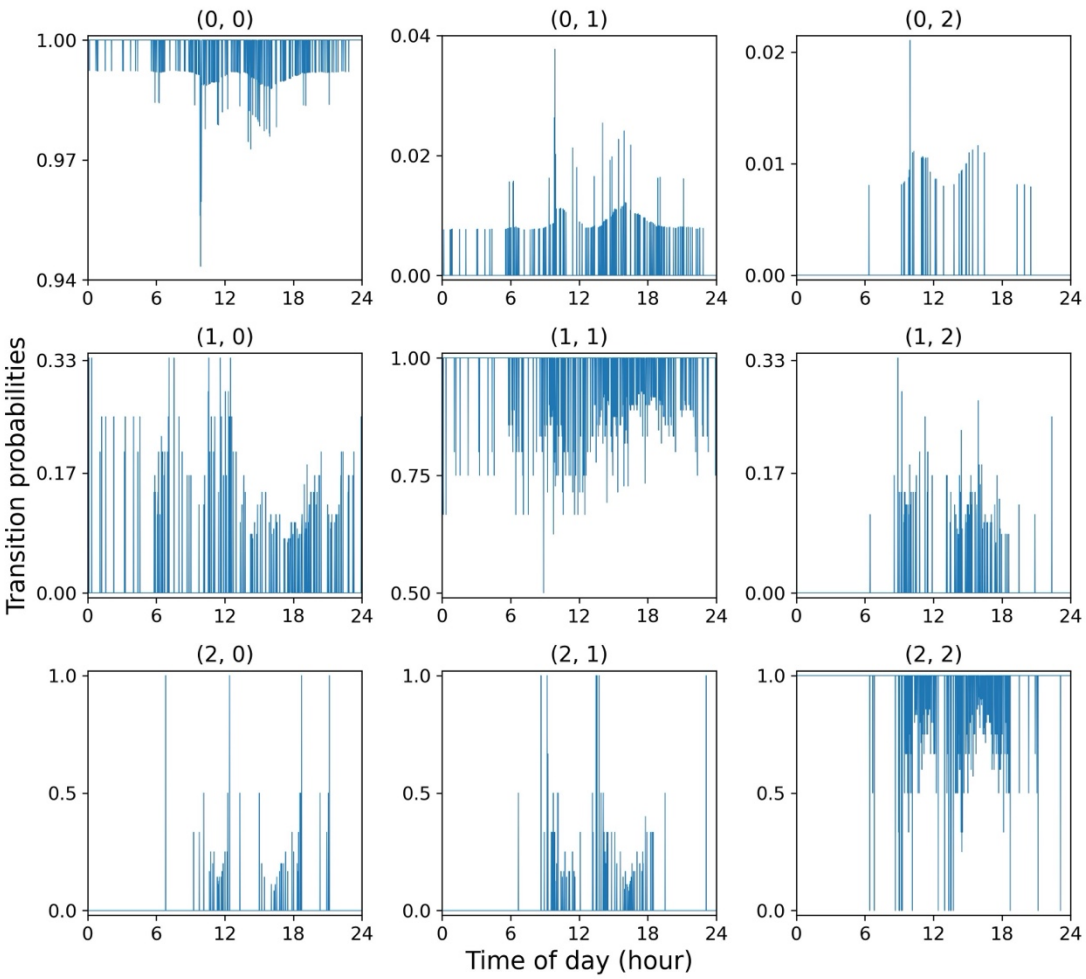


Figure 3-11. In-homogeneous transition probabilities from states 0,1,2 to states 0,1,2 varying with time of day using counting method (the plot at top-left titled (0,0) represents the transition probabilities from state 0 to state 0)

Figure 3-12 shows the results of simulating the occupancy state 100 times using the IMC with counting method. Although the counting method is able to capture the characteristics of the variation in occupancy states throughout the day, as evidenced by Figure 3-12 and a log-likelihood of -4032, it requires a large number of model parameters, reaching 116559 ($1439 \times 9 \times 9$), which results in very high values for the AIC and BIC (241181 and 1425741 respectively).

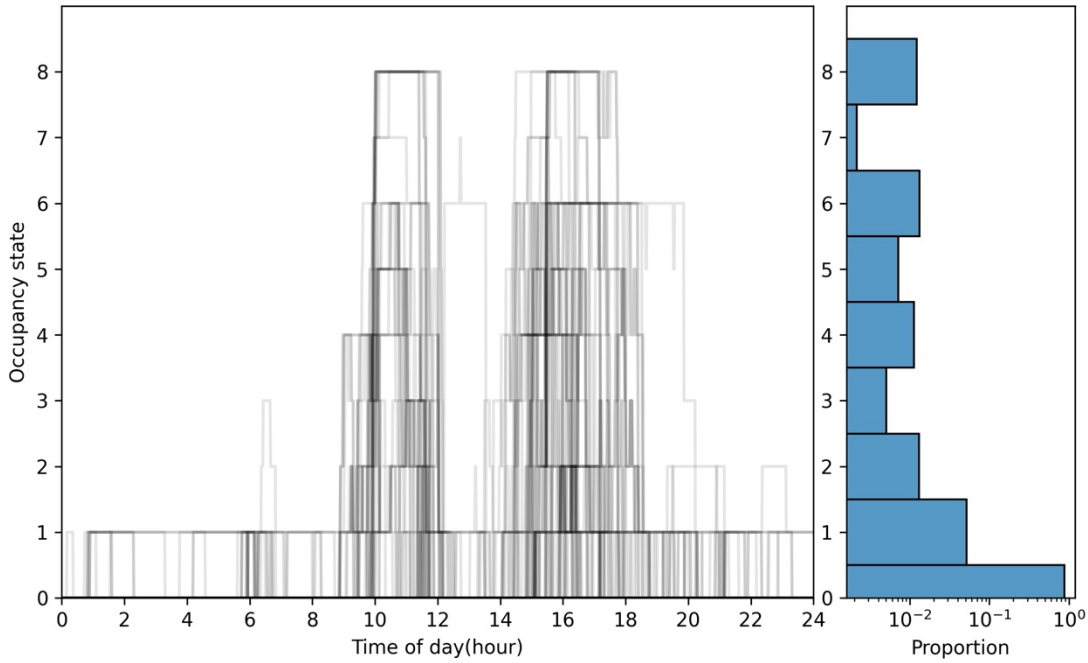


Figure 3-12. 100 simulations of the in-homogeneous transition matrix using counting method

3.4.2.2 Uniform B-spline method

Figure 3-13 illustrates the variation of transition probabilities over time for IMC using the uniform cubic B-Spline method (D = 4).

It compares the transition probabilities identified using different number of scale coefficients (*N*) ranging from 4 to 12 (the number of knots ranges from 8 to 16). *N* refers to the number of scaling coefficients for a single transition probability. The rationale for selecting a range of *N* from 4 to 12 is as follows. When *N* = 4, corresponding to cubic B-splines with *K* = 8, there are no internal knots. This represents the minimum required number of knots. As *N* increases, the number of knots gradually increases. It is observed that the log-likelihood also increases, indicating an improvement in accuracy. However, it is found that the marginal accuracy increment, i.e., the incremental improvement in accuracy gained by adding a knot, generally decreases. When *N* exceeds 12, the marginal accuracy increment becomes few, making it increasingly difficult to further enhance accuracy by adding more nodes. It is also the case for adaptive B-spline method. Therefore, there is no need to display results with a

higher number of knots.

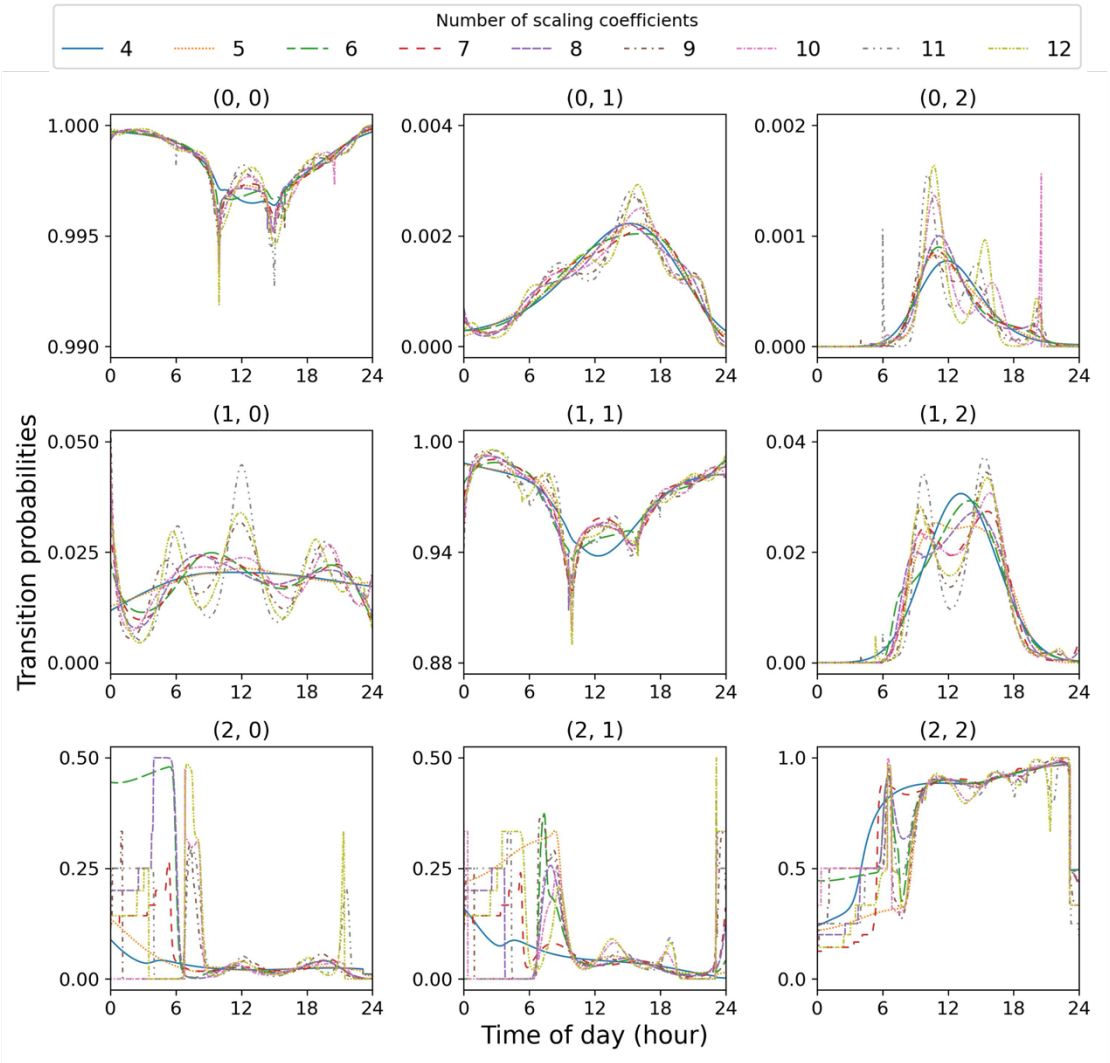


Figure 3-13. In-homogeneous transition probabilities (y axis) from states 0,1,2 to states 0,1,2 as a function of time of day (x axis) identified using the uniform B-Spline method (the plot at top-left titled (0,0) represents the transition probabilities from state 0 to state 0). Each curve is a B-spline with a specific number of scaling coefficients (N)

For the entire transition matrix, the total number of parameters is $N \times 81$. Compared to Figure 3-11, it can be seen that using the B-Spline method allows for smoothing of the transition matrix. Adopting the B-Spline method can significantly reduce the model parameters.

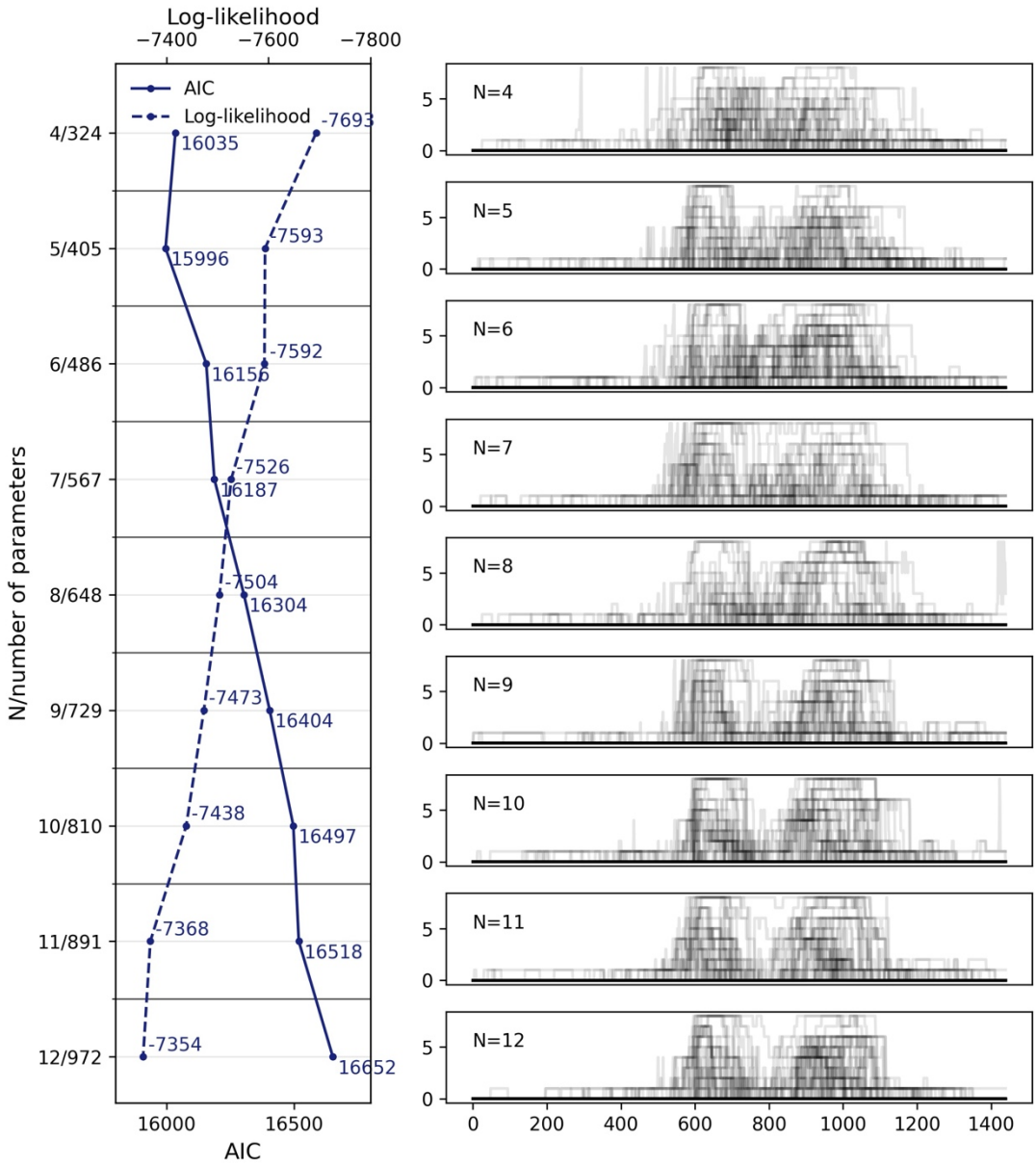


Figure 3-14. Log-likelihood, and AIC varying with number of scaling coefficients (N), and total number of parameters (N×81) and corresponding results of 100 simulations using uniform B-Spline method

In the case of N = 4, the open uniform knot vector used is (0,0,0,0,1,1,1,1), which means there is no interior knots. Figure 3-13 illustrates that this case captures the general trend of the transition probabilities, showing an increase in occupancy during the day and a decrease during the evening. However, it fails to capture finer structures, such as the variation trend during the lunch break at noon. As the number of scale

coefficients (*N*) increases, implying the inclusion of more interior knots and parameters, the transition probabilities gradually capture the bimodal structure more accurately. It is worth noting that in certain time ranges, such as before 6 AM and after 11 PM, where training data samples are limited, the transition probabilities are difficult to converge.

Figure 3-14 compares the log-likelihood, number of parameters and AIC as they vary with the number of scaling coefficients(*N*) in the uniform B-Spline method. The corresponding 100 simulations are shown on the right-hand side. It can be observed that as the number of scaling coefficients increases, the log-likelihood gradually decreases, indicating an improvement in model accuracy. The AIC initially decreases but then starts to increase due to the increase in model parameters. When the number of scaling coefficients is 5, the minimum AIC value is 15,996.

3.4.2.3 Adaptive B-spline method

Figure 3-15 shows the comparison of the log-likelihood and AIC between adaptive B-Spline and uniform B-Spline that varies with the number of scaling coefficients(N). It should be noted that in the adaptive B-Spline method, the distances between knots are also optimized.

Therefore, the adaptive B-Spline method has more parameters compared to the uniform B-Spline method, and the additional parameters correspond to the number of interior knots. Specifically, when N=5, there is 1 interior knots, thus the total parameters of adaptive B-Spline method is 406, 1 higher than the uniform B-Spline method. It can be observed that adopting adaptive B-Spline, both the log-likelihood is increased and AIC is reduced, demonstrated that the adaptive B-Spline method achieved better model performance regarding model accuracy and model complexity.

The right-hand side of Figure 3-15 shows the results of normalized distribution of knots. The upper right sub-figure is the occupant number observation data (same with Figure 3-7), which is used to show how the knots are adaptive to the occupancy patterns. It is worth noting that when there is only one internal knot (N=5), the adaptive method hardly changes the position of the knot.

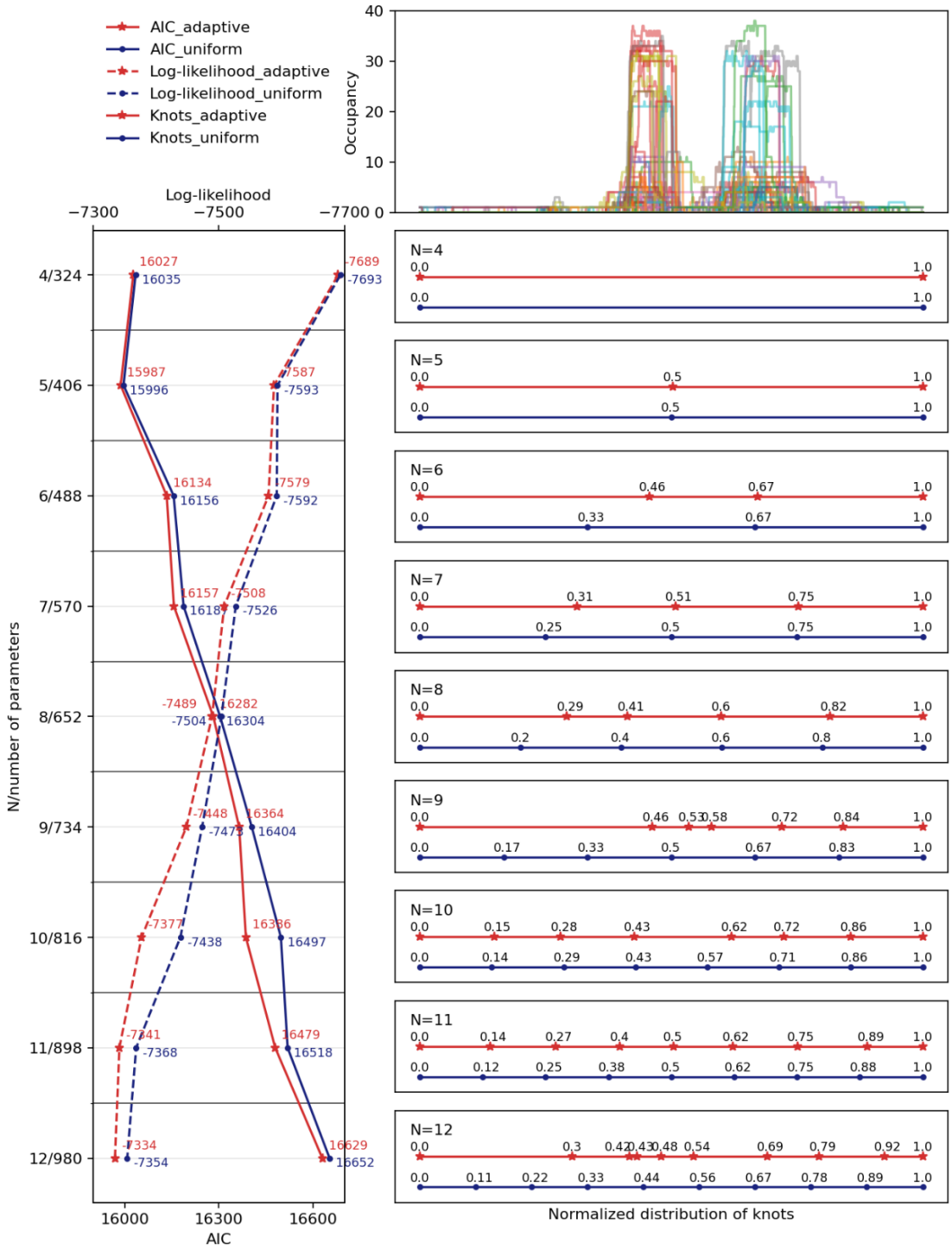


Figure 3-15. Comparison of Log-likelihood, and AIC between adaptive B-Spline and uniform B-Spline method that varies with number of scaling coefficients (N), and total number of parameters and corresponding results of normalized distribution of knots

This is because the normalized position at 0.5 is already the center of the time with significant overall occupancy variation, thus further optimization is practically

unnecessary. When there are knots distributed in position with less occupancy variation, the proposed adaptive B-Spline method can optimize the distribution of interior knots, adjusting them to better position.

It is clearly show that the density of interior knots in areas with less variation, i.e. early morning, has been reduced, while increased in areas with significant changes in occupancy, i.e. morning and afternoon. For example, when N=9, in the uniform B-Spline method, the interior knots are positioned at (0.17,0.33,0.5,0.67,0.83). The knots at 0.17 and 0.33 correspond to the early morning period when the number of people is low. When using the adaptive B-Spline method, the optimized knot positions are at (0.45,0.53,0.58,0.72,0.84). These positions align with periods of significant fluctuations in the number of people inside the meeting room.

Figure 3-16 illustrates the in-homogeneous transition probabilities identified using adaptive B-Spline and uniform B-Spline when N=9. From the subplots (0,2),(2,0), (2,1), and (2,2), it can be observed that the adaptive B-Spline method may exhibits a horizontal stretching and compressing effect on the transition probabilities along the time axis. This effect is different from the vertical scaling effect of scaling coefficients. The adaptive B-Spline method allows for the stretching of transition probabilities in time intervals with few changes in occupancy, while compressing transition probabilities in time intervals with significant changes. In the compressed zones, where there are higher number of scaling coefficients, the method can obtain better scaling coefficients to more accurately capture occupancy patterns.

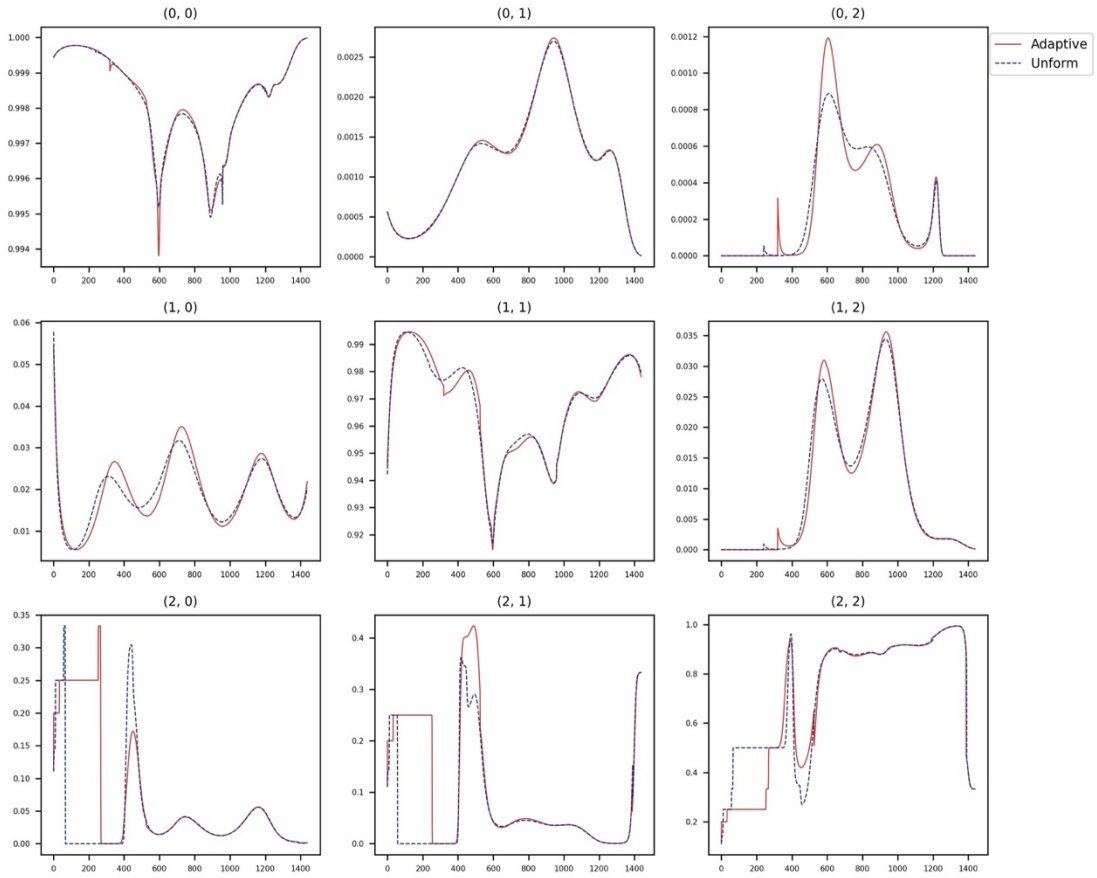


Figure 3-16. Comparison of in-homogeneous transition probabilities identified using adaptive B-Spline and uniform B-Spline when the number of scaling coefficients N=9

Note that during some of the less occupied times (for subplots (2,0), (2,1), and (2,2)), the transition probabilities look non-smooth at times. This is due to the inverse logit-transformation and that there is less-to-no data at these times and therefore, the parameter values here have no influence on the likelihood. The optimized distribution of knot positions noticeably enhances the accuracy of the model. When using different values of N, the adaptive B-Spline method can also effectively distribute knot positions that closely match periods of significant fluctuations in occupancy, demonstrating the effectiveness of the proposed method.

3.4.3 Comparison among models

Table 3-2 lists the AIC, BIC and log-likelihood for all investigated models in this

research. By comparing the log-likelihood, it can be observed that the adaptive B-Spline method is more accurate than the homogeneous Markov Chain model and in-homogenous Markov Chain using uniform B-Spline method under the same number of scaling coefficients.

When N=5, the accuracy of the adaptive B-Spline method even surpasses that of the uniform B-Spline method with N=6. Comparing the AICs and BICs, it can be seen that the adaptive B-Spline method consistently yields lower values compared to the homogeneous Markov Chain model and in-homogenous Markov Chain using uniform B-Spline method under the same number of scaling coefficients. This indicates that the adaptive B-Spline method has better overall performance. When N=5 the adaptive B-Spline method achieves the best overall performance

Table 3-2 Comparison among models

No.	Method	N	AIC	BIC	Log-like.	Params.
0	Homo	-	16231.02	17054.2	-8034.51	81
1	Counting	-	241181.27	1425740.96	-4031.63	116559
2		4	16035.01	17226.5	-7693.5	324
3		5	15996.62	17485.99	-7593.31	405
4		6	16156.23	17943.47	-7592.12	486
5		7	16187.87	18272.99	-7526.94	567
6	Uniform	8	16304.08	18687.07	-7504.04	648
7		9	16404.85	19085.71	-7473.42	729
8		10	16497.54	19476.27	-7438.77	810
9		11	16518.9	19795.51	-7368.45	891
10		12	16652.5	20226.97	-7354.25	972
11	Adaptive	4	16027.77	17219.26	-7689.88	324

12	5	15987.69	17480.73	-7587.84	406
13	6	16134.08	17928.67	-7579.04	488
14	7	16157.04	18253.18	-7508.52	570
15	8	16282.38	18680.07	-7489.19	652
16	9	16364.65	19063.9	-7448.33	734
17	10	16385.98	19386.78	-7376.99	816
18	11	16479.78	19782.13	-7341.89	898
19	12	16629.21	20233.1	-7334.6	980

3.5 Discussion

3.5.1 Adaptability in diverse room types with different occupancy patterns

Although the proposed Adaptive B-spline-based method is theoretically universal and capable of dynamically adjusting the density of knots to varying occupancy patterns across different room types, further empirical validation is necessary to substantiate this generalizability. In future work, we plan to collect occupancy data from a variety of room types—such as residential rooms, large-plan office spaces, classrooms and lecture halls—and apply the proposed method to identify the adaptive knots for each case. Through this empirical analysis, we aim to demonstrate the method's adaptability and effectiveness in capturing diverse occupancy characteristics across different environments.

3.5.2 Disadvantages of discrete-time formulation of Markov chains

The method presented in this chapter estimates each transition probability in the transition probability matrix given the transition time. This discrete-time formulation has two disadvantages. (1) It is not possible to use the estimated model for other transition times. Instead, one needs to estimate a similar model anew. (2) The amount of parameters scales quadratically with the number of states. In certain situations, both

of these issues can be overcome by using a continuous-time description of the transition probabilities. Such a situation is, when the system moves, it moves to neighboring states. In the situation of this meeting room, this would be the case if only one person entered the room at a time. This would result in a continuous-time transition rate matrix having a non-zero diagonal structure and zeros in the off-diagonal elements. This results an amount of parameters that is linear in the number of states and thus significantly reduces the amount of parameters and acts as regularization

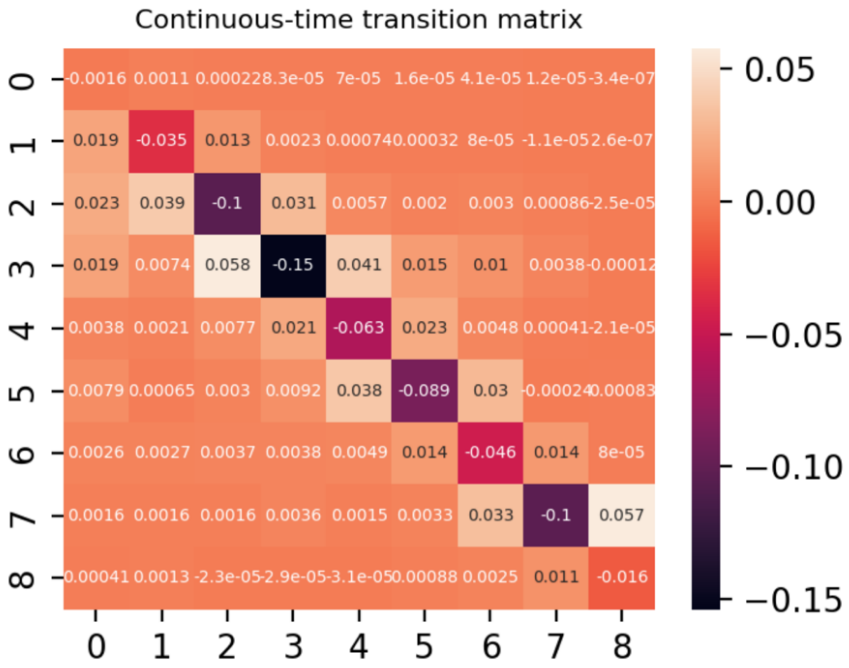


Figure 3-17. Continuous-time homogeneous transition rate matrix

We investigated this hypothesis by taking the matrix logarithm of the transition probability matrix to see if this diagonal-like structure is present. The matrix logarithm operation can convert the discrete-time transition probability matrix to the continuous-time transition rate matrix. Figure 3-17 shows a plot of the continuous-time homogeneous transition rate matrix. It clearly displays this diagonal pattern with a band structure. Modeling this in continuous time with IMC is illustrated in the next section.

3.6 Summary

This section presented discrete time in-homogeneous Markov chain models to model the occupancy levels in a meeting room of an office in Hong Kong. The models used various parametric expansions in time to model the time dependence of each day of the week, where the most advanced model used periodic B-splines where the optimal knots distribution was estimated. The models were tested and compared against each other and to a time-homogeneous model, and results suggest that it was advantageous to also optimize the knots distribution to better describe the different variations in occupancy data.

4 Continuous-time in-homogeneous Markov chains with band structure for improving occupancy model interpretability and generalization ability

At the end of Section 3, we identified the disadvantages of discrete-time formulation of Markov chains. By taking the matrix logarithm of the discrete-time transition probability matrix, a diagonal-like structure is clearly presented. In this section, we delve deeper into the band structure by integrating it into the continuous-time IMC (CTIMC) for stochastic occupancy modeling to capture the intrinsic physical properties of occupancy dynamics.

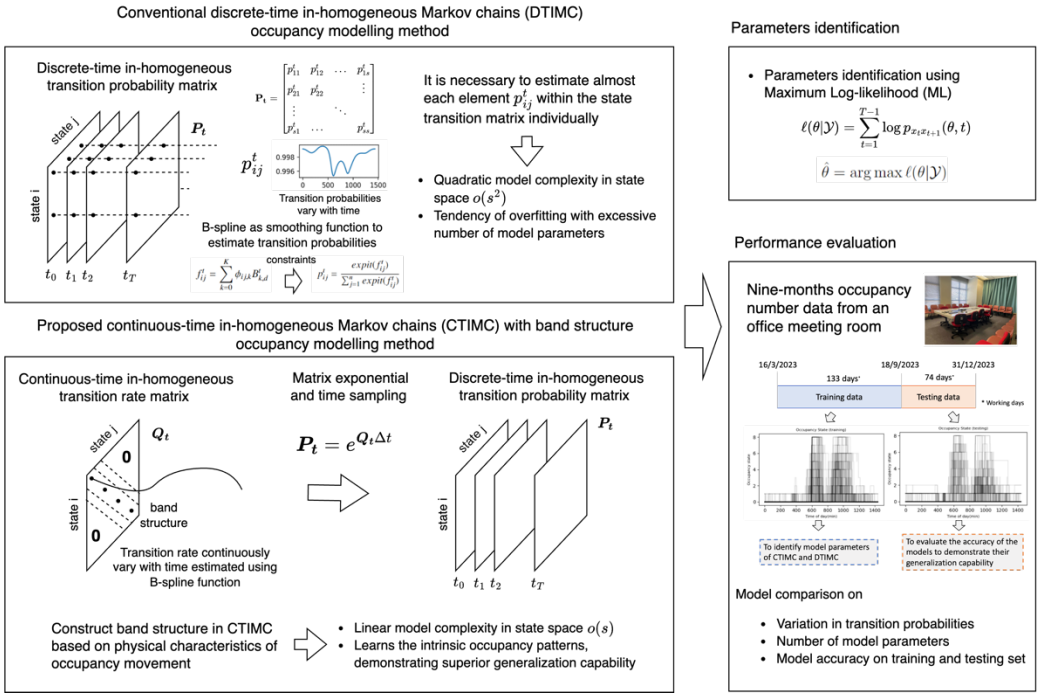


Figure 4-1. Methodology framework

Figure 4-1 illustrates the methodological framework. The traditional DTIMC modelling method from the literature is introduced in Section 3.4.2. The proposed CTIMC with band structure modelling method is elaborated in Section 4.1. The parameter identification method using the maximum log-likelihood technique, which is the same

as outlined in Section 3.2.4. Finally, Section 4.2 and 4.3 present the performance evaluation and comparison between the proposed CTIMC method and the DTIMC method.

4.1 Continuous-time Markov chain models

A continuous-time Markov model can be formulated in the following way. Let X(t) be the stochastic process on the states 0, 1, 2, ... and satisfy

$$p_{ij} = Pr(X(t+u) = j | X(t) = i), \forall u \geq 0 \quad (4\text{–}1)$$

We shall furthermore assume that p_{ij} satisfy the following conditions

$$p_{ij}(h) = \lambda_{ij} h + o(h) \; as \; h^{+} \to 0 \; for \; i \neq j \quad (4\text{–}2)$$

$$p_{ij}(h) = 1 - (\sum_{j} \lambda_{ij}) h^{+} + o(h) \; as \; h \to 0 \; for \; i \neq j \quad (4\text{–}3)$$

$$p_{ij}(0) = \delta_{ij} \quad (4\text{–}4)$$

where h is a small time increment, δ_{ij} is a Dirac delta function equal to 1 if $i=j$ and otherwise 0, and $\lambda_{ij} \geq 0$.

This definition of the stochastic process in continuous time gives rise to formulating the process using the so-called infinitesimal generator of the process; a matrix where each element is the rate at which the system transitions between states in continuous time:

$$\boldsymbol{Q} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1s} \\ q_{21} & q_{22} & & \vdots \\ \vdots & & \ddots & \\ q_{s1} & \cdots & & q_{ss} \end{bmatrix} \quad (4\text{–}5)$$

A CTMC describes the process of continuous rate of state changes, as opposed to changes of probability at discrete time points, which is described by DTMC. Unlike the constraint of a DTMC where each row sums to 1, the constraint for the transition rate matrix in CTMC is that the sum of each row equals 0 (condition Eq.(4-3)). This constraint arises from the fact that the rates at which the system leaves the current state must balance out with the rates at which it enters other states, maintaining the overall transition rate conservation in the system.

Here, this study employs CTIMC to capture the time-inhomogeneity of occupancy by allowing the transition rate matrix to vary in time (Eq.(4-6)).

It can be noticed that the number of parameters in Eq.(4-6) still grows quadratically with the increase in the number of states. In the following subsection, we explain how to construct a band structure based on the physical process of changes in occupant number in a room, thereby reducing the number of model parameters to O(s).

$$\boldsymbol{Q_t} = \begin{bmatrix} q_{t,11} & q_{t,12} & \cdots & q_{t,1s} \\ q_{t,21} & q_{t,22} & & \vdots \\ \vdots & & \ddots & \\ q_{t,s1} & \cdots & & q_{t,ss} \end{bmatrix} \tag{4–6}$$

4.1.1 Band structure

In the transition rate matrix, a band structure is established to represent the continuous variation of occupant numbers within a room, reflecting the physical characteristics of occupancy movement. It is assumed that individuals entering or exiting the room one-by-one in continuous time, resulting in a continuous change in occupant number. In DTMC models, such assumption does not hold. Between two discrete time steps, theoretically, there is no restriction on changes in the number of occupants. This is the reason why it is necessary to model almost each element within the state transition probability matrix individually in DTMC model.

$\boldsymbol{Q_t}$ in Eq.(4-6) can be written as Eq.(4-7), which shows the band structure. The elements in Eq.(4-7) are also illustrated in Figure 4-2. In this structure, the state represents directly the number of people, and it can only transition from *i* to its neighboring states *i+1* or *i-1* in continuous time. The last state corresponds to the room being fully occupied.

$$\boldsymbol{Q_{t,1}} = \begin{bmatrix} -q_{t,1} & q_{t,1} & 0 & \cdots & 0 & 0 \\ w_{t,2}q_{t,2} & -q_{t,2} & (1-w_{t,2})q_{t,2} & & 0 & 0 \\ \vdots & & \ddots & & & \\ 0 & \cdots & w_{t,i}q_{t,i} & -q_{t,i} & (1-w_{t,i})q_{t,i} & 0 \\ \vdots & & & & \ddots & \\ 0 & 0 & \cdots & 0 & q_{t,s} & -q_{t,s} \end{bmatrix} \tag{4–7}$$

where $q_{t,i}$ on the diagonal is the transition rate from *i* to other states at time *t*, $0 \leq q_{t,i} \leq 1$. The negative sign on the diagonal elements indicates that the direction of the rate transition is outward. Taking the first row as an example, $q_{t,1}$ represents the rate of flow from state 1 to state 2 at time *t*. $w_{t,i}$ is the ratio of total outward flow from *i* to the lower state (here to *i-1*) at time *t*, $1 < \mathrm{i} < \mathrm{s}$, $0 \leq w_{t,i} \leq 1$. Conversely, $1 - w_{t,i}$ is the ratio from *i* to the higher state (*i+1*) at time *t*. The fact that the sum of the row equals 0 represents the net zero rate of transition among all states, which can be interpreted as the law of conservation of transition.

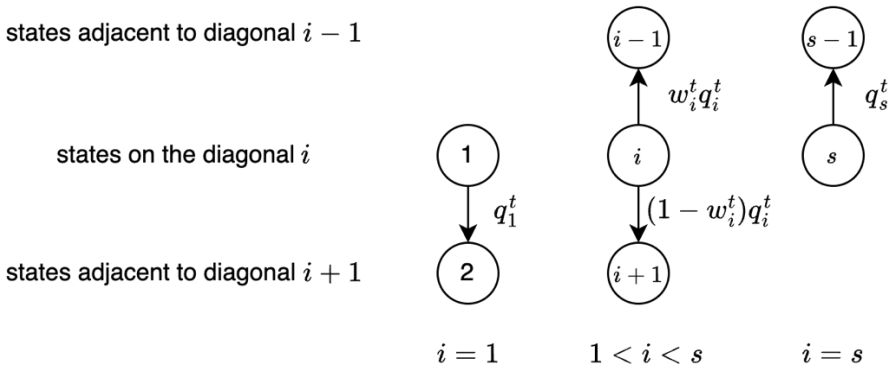


Figure 4-2. Illustration of band structure of Eq. (4-7)

Compared with the DTMC model, the number of parameter of CTMC model can be considerably reduced. The model complexity of the CTIMC with band structure model (Eq.(4-7})) is O(2s-2), i.e. 2s-2 number of parameters including $\mathrm{q_i}, \mathrm{i} \in 1, .., \mathrm{s}$ and $\mathrm{w_i}, \mathrm{i} \in 2, .., \mathrm{s} - 1$.

In band structure $\boldsymbol{Q_{t,1}}$, each specific state corresponds to an occupant number. However, in large meeting rooms or classrooms, the number of occupants can range from 0 to tens or even hundreds. A significant number of people may enter and exit in a short period, and it is also possible for more than one person to enter or leave the large room with larger door simultaneously. The building manager may not focus on the exact number of occupants, but rather on a general range. Therefore, a certain occupant counts

is clustered into an "occupancy level." In this scenario, an extension is made to the structure of $\boldsymbol{Q_{t,1}}$ in Eq.(4-6), allowing the occupancy levels to transition between multiple adjacent states: from *i* to either *i+1*, *i+2* or *i-1*, *i-2*. The extended band structure ($\boldsymbol{Q_{t,2}}$) is thus written as Eq.(4-8). Figure 4-3 illustrates the structure ($\boldsymbol{Q_{t,2}}$).

The meanings of $q_{t,i}$ and $w_{t,i}$ in Eq.(4-8) are similar to those of Eq.(4-7). Specifically, $q_{t,i}$ is the transition rate from *i* to other states at time *t*, and $w_{t,i}$ is the frequency of decreasing from *i* to other states (here to *i-1* and *i-2*) at time *t*; $1 - w_{t,i}$ is the frequency of increasing from *i* to other states (here to *i+1* and *i+2*) at time *t*. $e_{t,i}$ represents the frequency of one adjacent states transition from *i* to either *i-1* or *i+1* at time *t*, $0 \leq e_{t,i} \leq 1.1 - e_{t,i}$is the frequency of two adjacent states transition from *i* to either *i-2* or *i+2* at time *t*. Here, the model complexity of the CTIMC with band structure 2 is O(3s-2).

4.1.2 B-splines for modelling time-inhomogeneity

The time-inhomogeneity, i.e. the diurnal variation of the parameters of CTIMC ($q_{t,i}$, $w_{t,i}$, and $e_{t,i}$) is characterized using B-Spline-based function approximation method. It should be noted that in this chapter, the uniform B-spline method is used as the approximation function as detailed in Section 3.2.2.1 , rather than the adaptive B-spline method.

A parameter transformation between a constrained and unconstrained domain is necessary to ensure $q_{t,i}, w_{t,i}, e_{t,i} \in [0,1]$. The inverse logit transformation is also applied to these parameters. $q_{t,i}, w_{t,i}, e_{t,i}$ are then be written as $q_{t,i} = \text{expit}\left(\sum_{n=1}^{N} \phi_{q,i,n} B_{n,D}(t)\right), i \in \{1,2,\dots,s\}$, $w_i^t = \text{expit}\left(\sum_{n=1}^{N} \phi_{w,i,n} B_{n,D}(t)\right), i \in \{2,\dots,s-1\}$, $e_i^t = \text{expit}\left(\sum_{n=1}^{N} \phi_{e,i,n} B_{n,D}(t)\right), i \in \{1,\dots,s\}$.

$$\boldsymbol{Q_{t,2}} = \begin{bmatrix} -q_{t,1} & e_{t,1}q_{t,1} & (1-e_{t,1})q_{t,1} & \cdots & 0 & 0 \\ w_{t,2}q_{t,2} & -q_{t,2} & e_{t,2}(1-w_{t,2})q_{t,2} & (1-e_{t,2})(1-w_{t,2})q_{t,2} & \cdots & 0 \\ (1-e_{t,3})w_{t,3}q_{t,3} & e_{t,3}w_{t,3}q_{t,3} & -q_{t,3} & e_{t,3}(1-w_{t,3})q_{t,3} & (1-e_{t,3})(1-w_{t,3})q_{t,3} & 0 \\ \vdots & & & \ddots & & \\ 0 & \cdots & (1-e_{t,s-1})w_{t,s-1}q_{t,s-1} & e_{t,s-1}w_{t,s-1}q_{t,s-1} & -q_{t,s-1} & e_{t,s-1}(1-w_{t,s-1})q_{t,s-1} \\ 0 & 0 & \cdots & (1-e_{t,s})q_{t,s} & e_{t,s}q_{t,s} & -q_{t,s} \end{bmatrix} \quad (4\text{–}8)$$

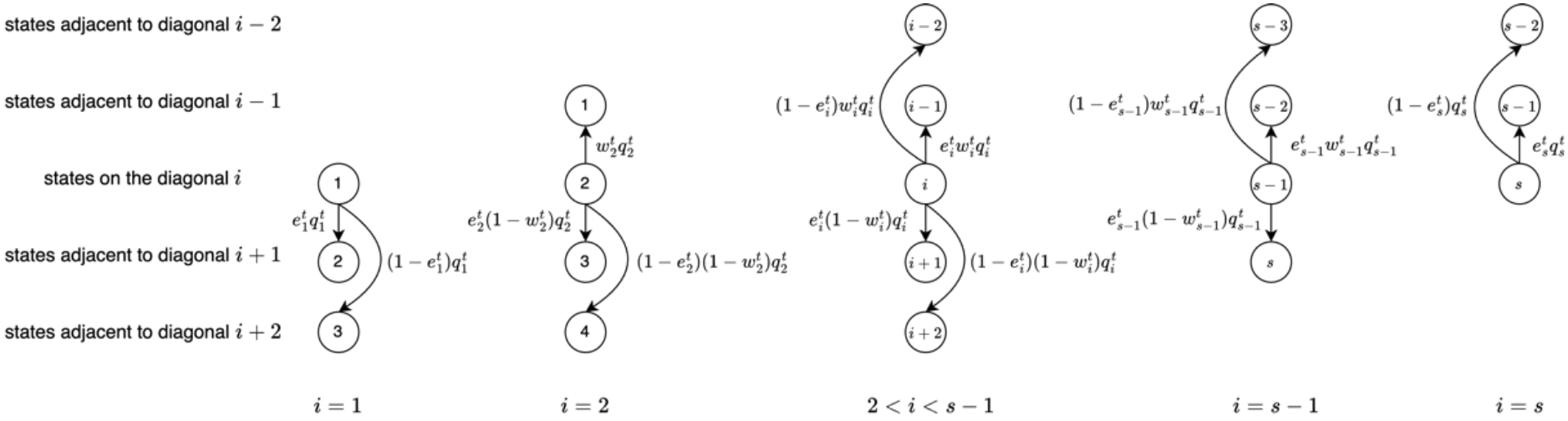


Figure 4-3. Illustration of band structure of Eq.(4-8)

4.1.3 Matrix exponential

The matrix exponential serves as a linkage between CTIMC and DTIMC. By applying matrix exponential operation, CTIMC can be transformed to DTIMC, and the parameters of CTIMC can thus be estimated using discrete measurement data. Below is a brief introduction of the matrix exponential. For more detailed elaboration, please refer to [44]. The solution to continuous-time forward Kolmogorov equations with t_0 be the initial time and at time $\mathrm{t_1 = t_0 + \Delta t}$ is then the discretised transition probability matrix $\mathrm{P(t_1) = \{p_{i,j}\}}$ with initial condition $\mathrm{P(t_0) = I}$. A simple solution is

$$\boldsymbol{P_t} = \exp(\boldsymbol{Q_t}\Delta \mathrm{t}) \tag{4–9}$$

assuming $\boldsymbol{Q_t}$ is constant for $\mathrm{t} \in [t_0, t_1[$. $P(t_1) = \{p_{ij}\}_{ij}$ are thus the transition probabilities between states *i* and *j* in the time interval $\Delta\, t$.

The maximum loglikelihood method as detailed in Section 3.2.4 is then adopted to identify the parameters of the continuous-time Markov chains model $\boldsymbol{Q_t}$.

4.2 Case study

This section provides a brief description of the dataset used for model development and validation tests. The occupant number data originates from the same office meeting room detailed in Section 3.3. The analysis is conducted on a nine-month occupancy dataset from 16th March to 31st December 2023. Note that during the nine months, the room has been consistently used as a meeting room, with no significant changes in room function. The analysis specifically focuses on workdays since the meeting room receives fewer visitors on weekends.

The nine-month data is divided into training (16th March-18th September, 133 days) and testing datasets (19th September-31st December, 74 days). Figure 4-4 illustrates the distribution of the different states of the training and testing dataset. The training set is used to identify model parameters of the proposed CTIMC with two band structures $\boldsymbol{Q_{t,1}}, \boldsymbol{Q_{t,2}}$ and the DTIMC identified using conventional method. The

accuracy of the models is then evaluated on the testing set to demonstrate their generalization capability.

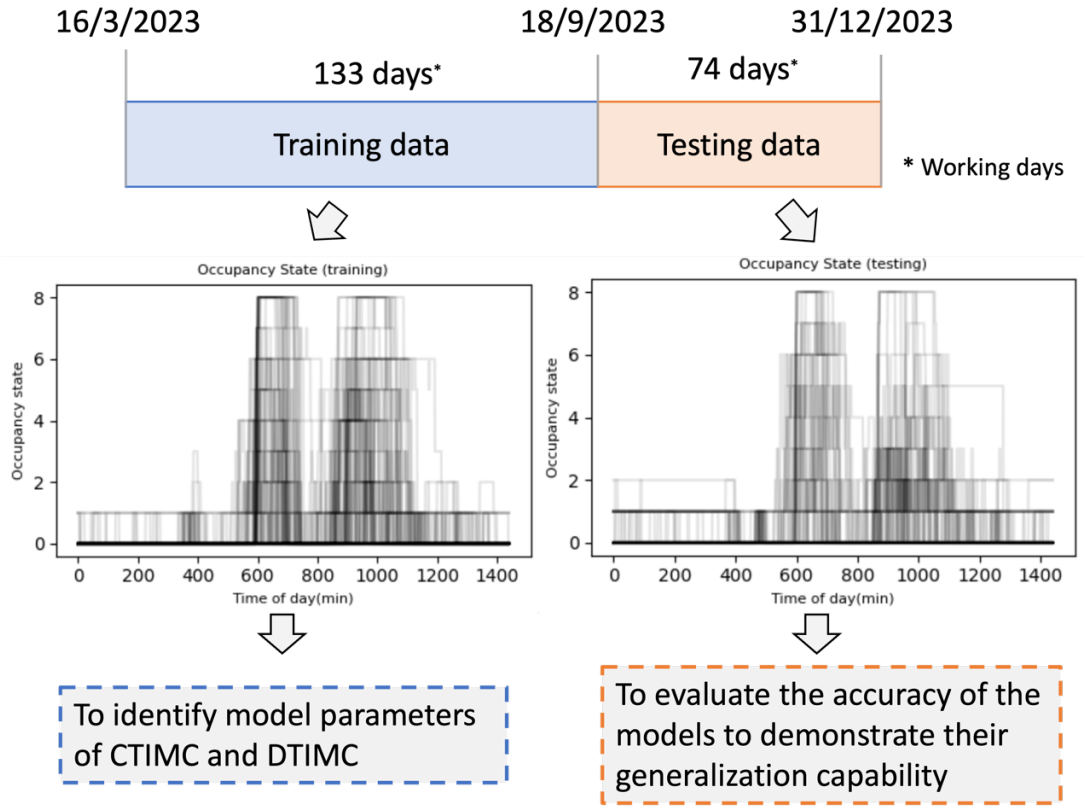


Figure 4-4. Training and testing data split

4.3 Results

This chapter compares the performances of the proposed CTIMC with band structure and DTIMC modelling methods. The comparison is conducted from three aspects. The first aspect involves examining the time variation in state transition probabilities obtained from the CTIMC and DTIMC models. The second aspect quantitatively analyzes the number of model parameters for each model to illustrate their model complexity. The third aspect evaluates the accuracy of the models on both training and testing datasets, highlighting their generalization capabilities. B-splines of different methods with a range of scale coefficients (N from 4 to 12) are identified for comparison.

4.3.1 Variation in transition matrix

In this section, the time-varying transition rates under the CTIMC model with band structures $\boldsymbol{Q_{t,1}}, \boldsymbol{Q_{t,2}}$ are first presented. These transition rates are converted into transition probabilities using the matrix exponential operation, which are compared with the transition probabilities obtained from the conventional DTIMC method.

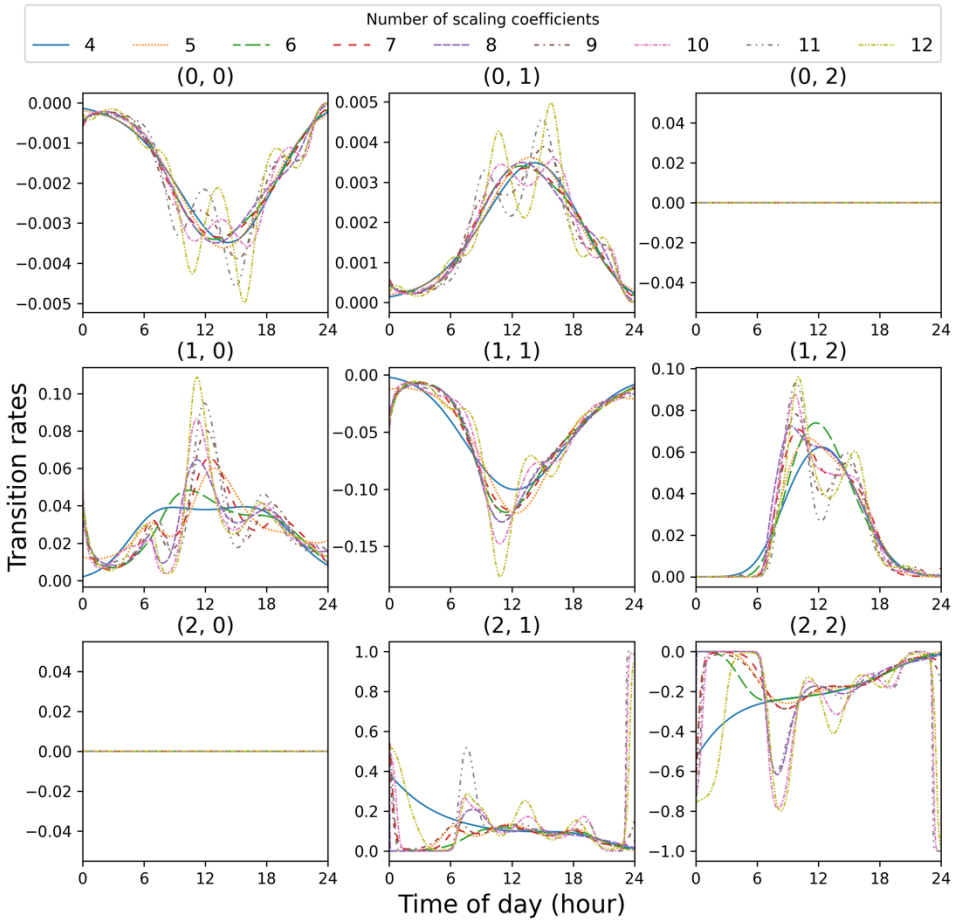


Figure 4-5. Inhomogeneous transition rates (y axis) of CTIMC with band structure $\boldsymbol{Q_{t,1}}$ *one state on the main diagonal can only jump to its one adjacent neighbouring state) from states 0,1,2 to states 0,1,2 as a function of time of day (x axis) identified using the uniform B-Spline method (the plot at top-left titled (0,0) represents the transition rates from state 0 to state 0). Each curve is a B-spline with a specific number of scaling coefficients (N)*

Figure 4-5 illustrates the variation of the transition rate matrix over time for CTIMC with band structure 1. In this structure, a state on the main diagonal can only transition to one of its adjacent neighboring states in continuous time. Note that the transition rates at positions (0,2) and (2,0) remain at 0. The figure compares the transition rates identified using a varying number of scale coefficients (N) ranging from 4 to 12.

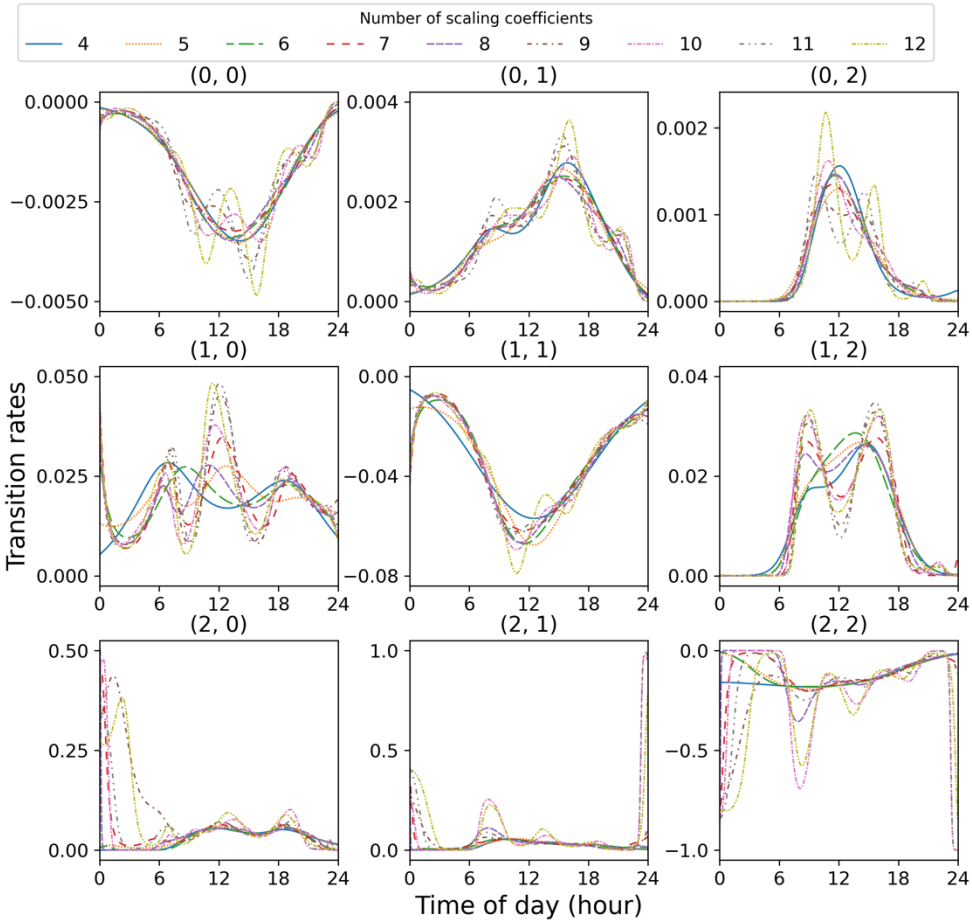


Figure 4-6. Inhomogeneous transition rates (y axis) of CTIMC with band structure 2 $\boldsymbol{Q_{t,2}}$ *one state on the main diagonal can jump to its two neighbouring states) from states 0,1,2 to states 0,1,2 as a function of time of day (x axis) identified using the uniform B-Spline method (the plot at top-left titled (0,0) represents the transition rates from state 0 to state 0). Each curve is a B-spline with a specific number of scaling coefficients (N)*

For N=4, the general trend of the transition rates characterized by a single peak, indicating an increase in occupancy during the day and a decrease in the evening, can

be captured. However, it fails to capture finer structures, such as the variation during the lunch break at noon. As the number of scale coefficients (N) increases, the identified transition rates can gradually and more accurately reflect the bimodal pattern.

Figure 4-6 illustrates the variation of the transition rate matrix over time for CTIMC with band structure 2. In this structure, a state on the main diagonal can transition to its two neighboring states in continuous time. The transition rates at positions (0,2) and (2,0) also vary with time.

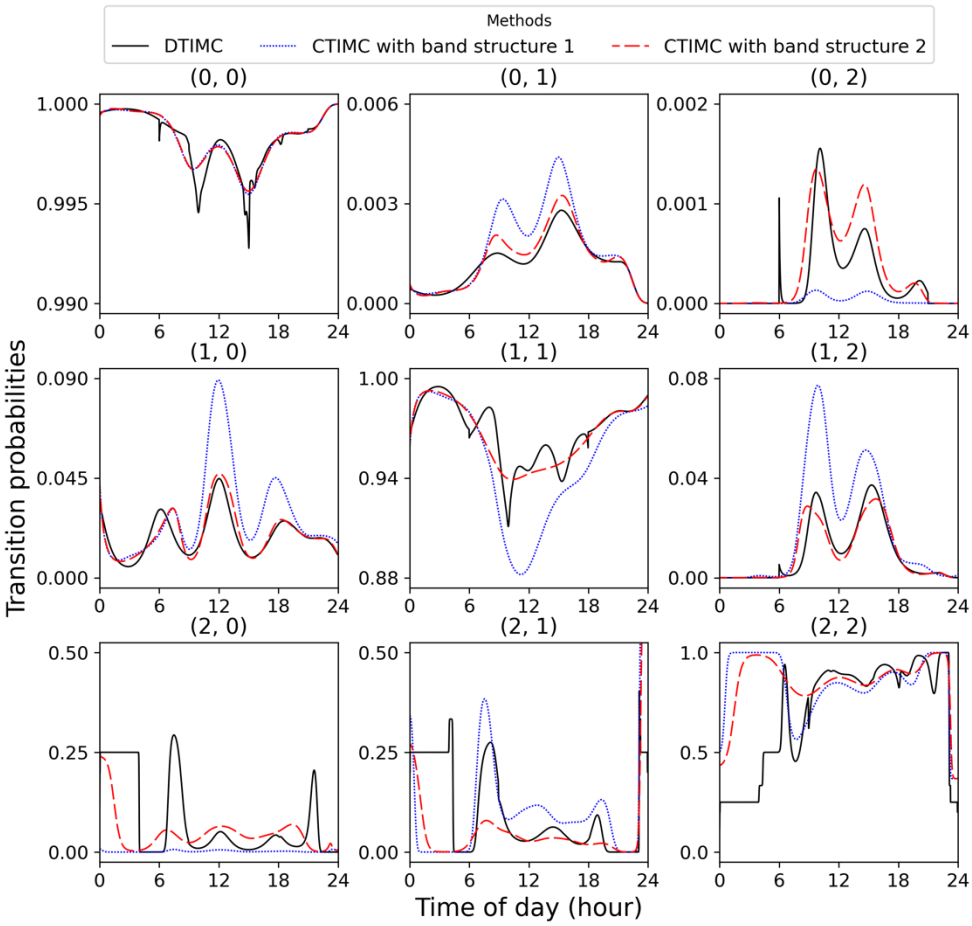


Figure 4-7. Comparison of transition probability matrix obtained using the proposed CTIMC with band structure $\boldsymbol{Q_{t,1}}$ *(blue dotted line), structure* $\boldsymbol{Q_{t,2}}$ *(red dashed line) after the matrix exponential operation and the conventional DTIMC* $\boldsymbol{P_t}$ *(black line) from states 0,1,2 to states 0,1,2 as a function of time of day (x axis). The comparison is based on the same number of scaling coefficients (N=9) in the identified B-splines*

Figure 4-7 contrasts the transition probabilities derived from the CTIMC with band structure model and those obtained directly from the DTIMC model. It is important to note that the comparison maintains the same number of scaling coefficients (N=9) and same training data for the B-spline estimation in both the DTIMC and CTIMC models to ensure consistency. Also note that the transition probabilities of DTIMC are identical with those in Figure 3-16 identified using uniform B-spline method.

It is evident that the transition probabilities obtained from the DTIMC model exhibit numerous spikes and are particularly less smooth, especially in time periods with scarce data points. It can be observed that the figure at diagonal titled (0,0), (1,1), and (2,2), exhibits multiple spikes. Note the spike around 6 o'clock in the (0,2) plot. These spikes indicate that DTIMC fit the training data too closely and are prone to capturing noise and random fluctuations. Additionally, during the early morning and evening periods (0-6 a.m.), when the occupant number remains close to zero, with few data transitions from zero to other numbers, the graph appears highly irregular and non-smooth. It is clear that the transition probability matrix derived from the CTIMC model with band structure is smoother compared to the DTIMC model. Specifically, band structure 2 better approximates the general shape of the DTIMC than structure 1 without any spikes seen in the DTIMC model, indicating that structure 2 may be better at capturing the patterns in the number of occupants.

4.3.2 Number of parameters

Figure 4-8 provides a visual comparison of the number of model parameters required by different models. It is evident that the CTIMC with band structure modeling approach requires significantly fewer parameters compared to the DTIMC model. Assuming both modeling methods employ B-spline for estimating model parameters for an IMC with s states, with N scaling coefficients in the B-spline, the model complexity for DTIMC is O(Ns(s-1)), while for CTIMC with band structure 1, it is O(N(2s-2)), and for CTIMC with band structure 2, it is O(N(3s-2)). In this study, with 9 states, the number of parameters required for CTIMC with band structure 1 and

structure 2 is about 1/4 and 1/3, respectively, compared to DTIMC.

4.3.3 Model accuracy on the training and testing dataset

Figure 4-9 and Figure 4-10 compare the log-likelihood of different models on the training and testing sets, reflecting the models' accuracy. A log-likelihood closer to 0 indicates higher model accuracy. It can be observed that overall, the DTIMC model performs better on the training set compared to the CTIMC model, and as the number of scaling coefficients increases, i.e., the number of nodes increases, the accuracy of the DTIMC model improves.

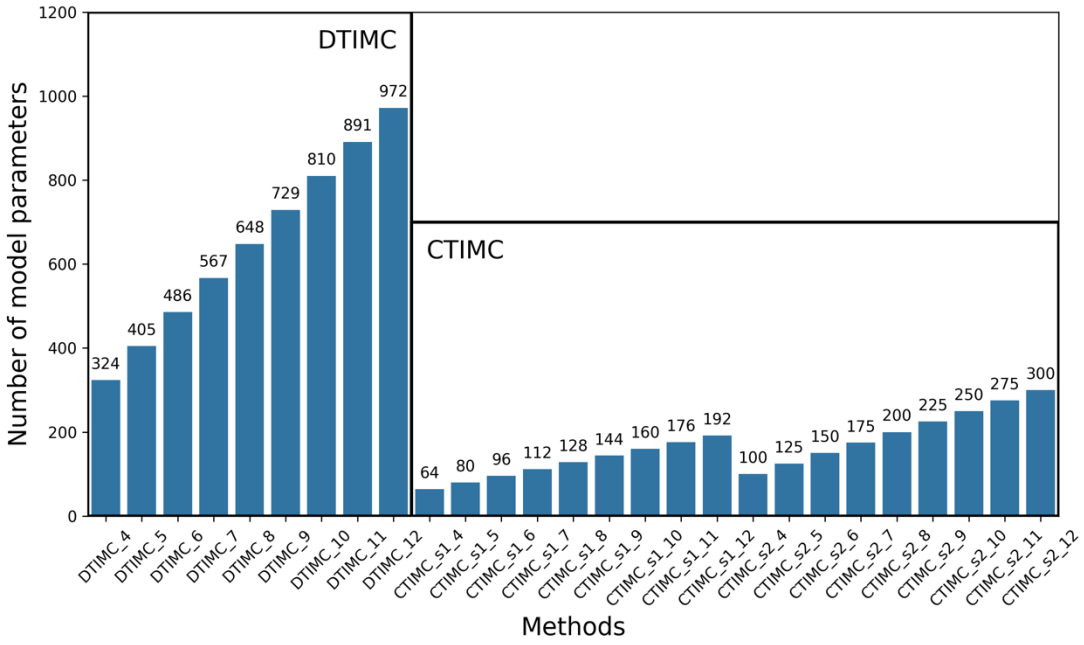


Figure 4-8. Comparison of number of parameters

However, when comparing the performance on the test data set, the log-likelihood of the CTIMC model significantly surpasses that of the DTIMC model. Furthermore, as the number of scaling coefficients increases, the log-likelihood of the DTIMC model on the test data set decreases, suggesting a decrease in accuracy. This indicates that the DTIMC model overfits the training set, and a higher number of scaling coefficients leads to increased level of overfitting. In contrast, for the CTIMC model, there is no

significant variation in log-likelihood as the number of scaling coefficients increases, indicating that this modeling approach effectively suppresses overfitting issues, greatly enhancing the model generalization ability.

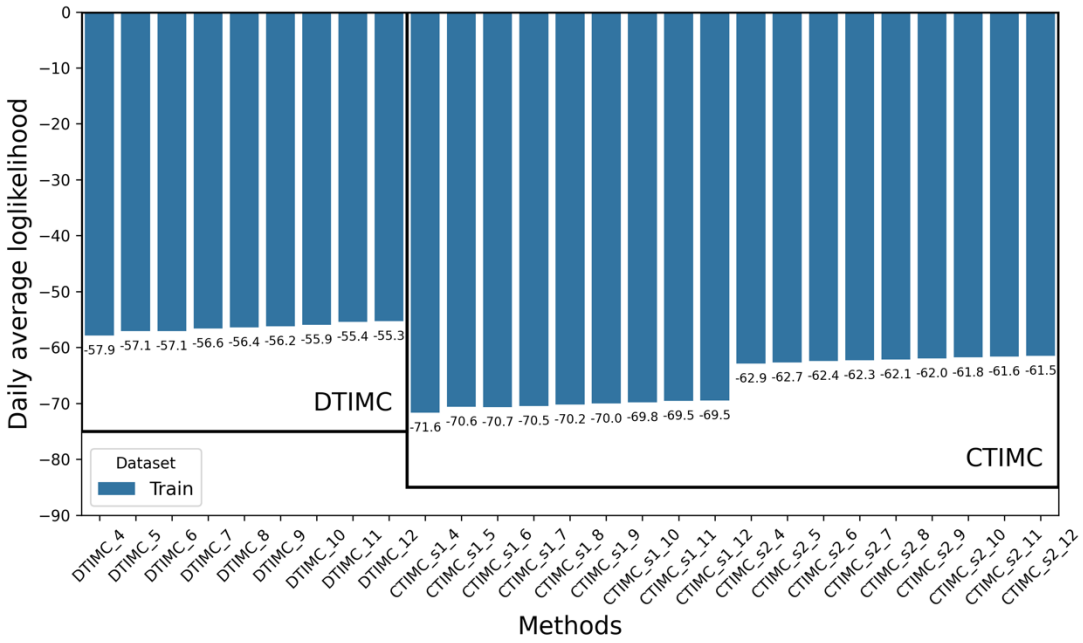


Figure 4-9. Comparison of loglikelihood on training dataset

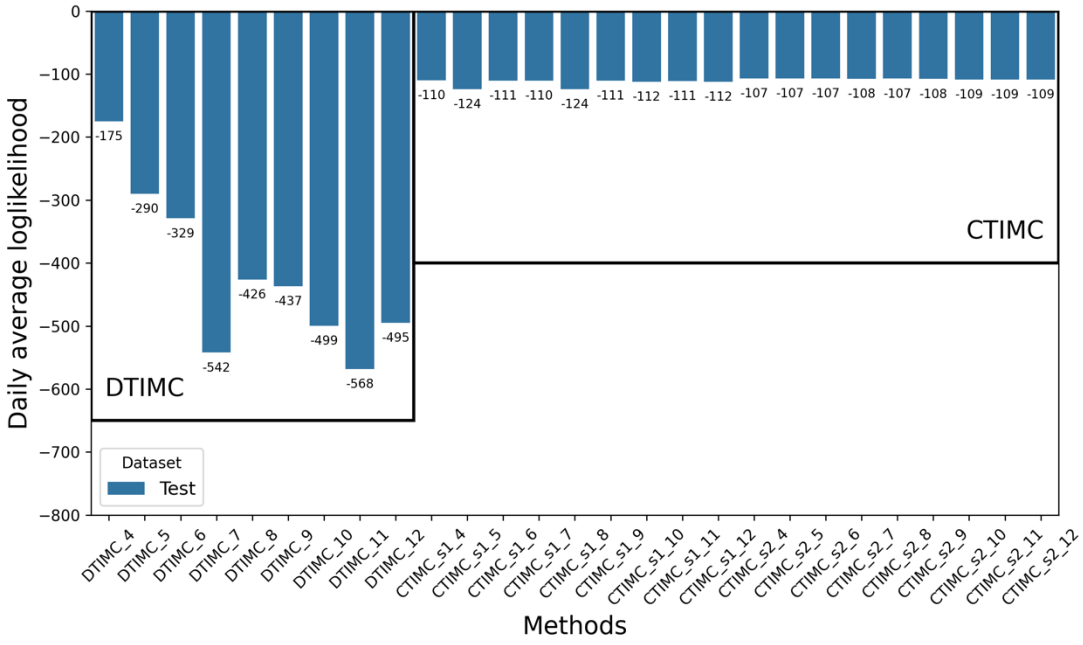


Figure 4-10. Comparison of loglikelihood on testing dataset

4.4 Summary

This chapter presented Continuous time Inhomogeneous Markov chain models with band structure to model the occupant number in a meeting room of an office building in Hong Kong. The proposed Continuous-Time Markov Chain with band structure can reduce the original quadratic complexity of the Discrete-Time Markov chain model to linear complexity. The models were tested and compared against each other. Results suggest that the proposed Continuous-Time Markov Chain models with band structure learns the intrinsic occupancy patterns in the data during training, rather than specific noises of the training data, demonstrating superior generalization capability.

5 Development and validation of stochastic occupancy-integrated MPC

In Chapters 3 and 4, we develops IMC-based stochastic occupancy modeling methods. In this section, we integrate the IMC-based stochastic occupancy predictions into the MPC strategy to achieve multi-objective optimal control of built environments. The overarching goal is to align building energy system operations with highly stochastic and intermittent occupancy patterns, thereby improving thermal comfort and IAQ while minimizing energy consumption within operational constraints. By addressing the challenges posed by dynamic and uncertain occupancy behaviors, the MPC strategy developed in this section seeks to enhance both energy efficiency and occupant satisfaction in built environment management.

The overview of the proposed stochastic occupancy-integrated MPC strategy is depicted in Figure 5-1. Three prediction models are embedded within the MPC for online predictions, as highlighted with red borders. These models include the stochastic occupancy prediction model, the stochastic differential equations (SDE)-based building thermal dynamics model, and the CO_2 dynamic model. Note that the IMC model developed in Section 3 is utilized in this section for occupancy prediction. Additionally, a white-box building energy system model, highlighted with a blue border, is constructed in TRNSYS as a simulation environment to represent a more realistic building energy system and to test and validate the effectiveness of the MPC strategy.

The methodology comprises the following three main parts:

(1) Utilization of the developed IMC model for stochastic occupancy prediction: The process begins with real-time occupancy detection. The detected occupant number is taken as the initial value of the IMC-based occupancy model (Section 3.2). The forward Kolmogorov equations are employed to predict the future occupant number trajectories within a prediction horizon using the IMC model (Section 5.1). The anticipated future trajectory of occupant number and presence is obtained and

integrated into the MPC.

(2) Formulation of the stochastic occupancy-integrated MPC: The stochastic occupancy prediction is incorporated into the MPC scheme in two key ways. Firstly, the expected future occupant number is utilized to predict occupancy-related disturbances including occupancy heat gains and CO_2 generation, which are then embedded in the SDE-based grey-box building dynamic model to enhance the prediction accuracy of dynamic building environment (Section 5.2). Secondly, the occupancy presence is integrated into the cost function of MPC as a weighting factor for thermal comfort and IAQ objectives (Section 5.3). The goal is to optimally allocate sufficient energy for pre-conditioning the built environment when occupancy is anticipated. The multi-objective rolling optimization problem is formulated and solved using the Gurobi optimizer, with the first control inputs applied to the TRNSYS white-box building energy systems model.

(3) Development of a white-box building energy system model in TRNSYS for simulation tests: This study uses different models in the stochastic occupancy-integrated MPC scheme and in simulation for performance evaluation to demonstrate the effectiveness of the stochastic occupancy-integrated MPC control strategy. A white-box building energy model is developed within TRNSYS, encompassing building thermal dynamics and typical building HVAC systems (Section 5.4.1-5.4.3). In this study, the white-box model of HVAC systems used for the case study refers to the air conditioning and ventilation systems model. Specifically, the fan coil unit model is developed for indoor air temperature control, and the primary air unit with variable air volume model is developed for fresh air flow rate control. The stochastic occupancy-integrated MPC provides supervisory control setpoints to the HVAC systems. Real occupancy data from the meeting room detailed in Section 3.3 is used for simulation tests in TRNSYS.

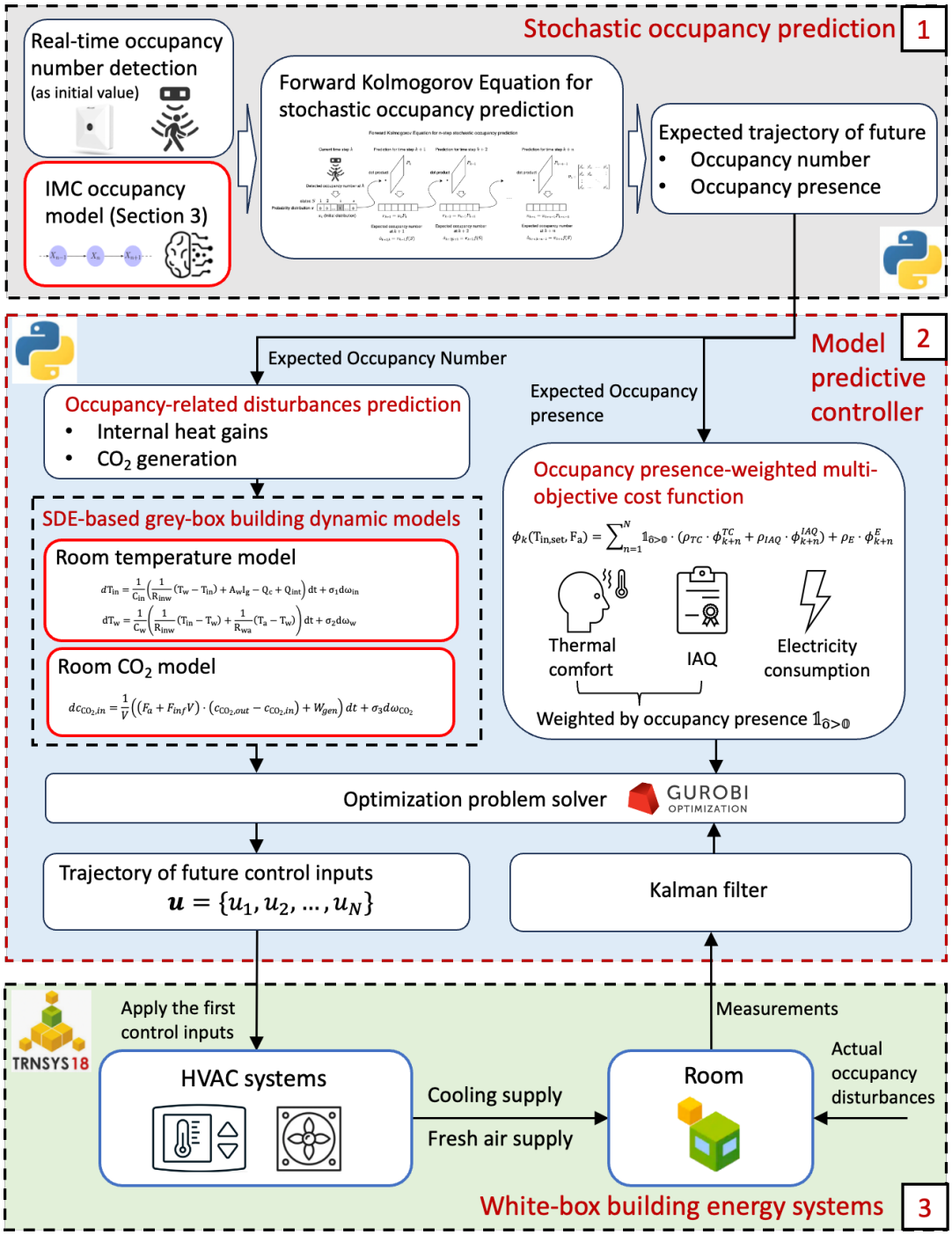


Figure 5-1. The overview of the proposed stochastic occupancy-integrated MPC strategy

In this study, the effectiveness of the method is validated through TRNSYS simulations by comparing different control strategies under the same outdoor weather conditions and occupancy scenarios . Additionally, a parameter sensitivity analysis was performed

by altering the optimization weight values in the MPC, and Pareto charts are generated to evaluate the control performance under different weight combinations. The proposed method was validated by comparing its performance with baseline control strategies.

5.1 Stochastic occupancy prediction

This part presents the offline stochastic occupancy modelling method and online stochastic occupancy prediction method, which are illustrated in Figure 5-2.

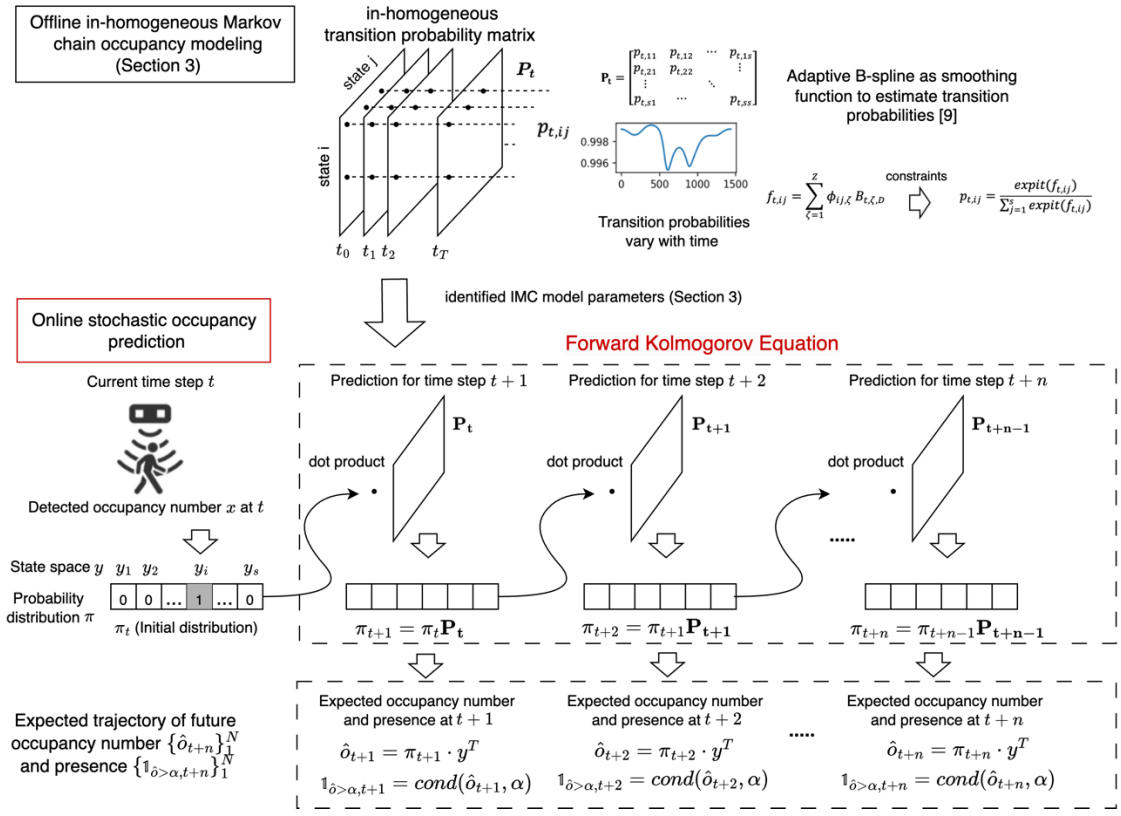


Figure 5-2. Illustration of forward Kolmogorov equation for n-step stochastic occupancy prediction

Our study adopts the IMC with adaptive B-spline method for stochastic occupancy modelling method, which is elaborated in Section 3. In online occupancy prediction, the real-time occupant number is detected at time step *t* using occupancy sensor, which corresponds to a specific state of occupancy level y_i. There are *s* number of occupancy

state, and the occupancy state vector is $\boldsymbol{y} = [y_1, y_2, \dots, y_s]$. The probability distribution vector of states at time step t ($\pi_t = [p_{t,1}, p_{t,2}, \dots, p_{t,s}]$) is initialized as that the probability for state y_i is 1($p_{t,i} = 1$)), while the probabilities for other states are 0 ($p_{t,j} = 0, j \neq i$). The forward Kolmogorov equations are used for determining the evolution of probability distribution of states for next n steps. The probability distribution of states at the time step $t+1$ (π_{t+1}) is calculated using the dot product between π_t and the transition probability matrix at time step t ($\mathbf{P_t}$), i.e. 1-step forward Kolmogorov equation (Eq.(5-1)). The predictions for next n step are given by the n-step forward Kolmogorov equation [44] (Eq.(5-2)), and the probability distribution vector of states at $t+n$ π_{t+n} can be obtained.

$$\pi_{t+1} = \pi_t \cdot \mathbf{P_t} = [p_{t,1}, p_{t,2}, \dots, p_{t,s}] \cdot \begin{bmatrix} p_{t,11} & p_{t,12} & \cdots & p_{t,1s} \\ p_{t,21} & p_{t,22} & & \vdots \\ \vdots & & \ddots & \\ p_{t,s1} & \cdots & & p_{t,ss} \end{bmatrix} \tag{5–1}$$

$$\pi_{t+n} = \pi_t \cdot \prod_{k=t}^{t+n-1} \mathbf{P_k} \tag{5–2}$$

The expected occupant number at time $t+n$ can is determined using Eq.(5-3) [44].

$$\hat{o}_{t+n} = \pi_{t+n} \cdot \boldsymbol{y^T} \tag{5–3}$$

where $\hat{o}_{t+n}$ is the expected occupant number at time $t+n$. The estimated trajectory of occupant number over the prediction horizon N is denoted as $\{\hat{o}_{t+n}\}_1^N$. $\hat{o}_{t+n}$ is used to estimate the occupancy heat gains ($\hat{Q}_{int,t+n}$) and CO_2 generation ($\hat{W}_{gen,t+n}$) using Eq. (5-4) and Eq. (5-5), respectively.

$$\hat{Q}_{int,t+n} = r_{Q_{int}} \cdot \hat{o}_{t+n} \tag{5–4}$$

$$\hat{W}_{gen,t+n} = r_{CO_2} \cdot \hat{o}_{t+n} \tag{5–5}$$

where $r_{Q_{int}}$ is the heat dissipate rate per occupant, r_{CO_2} is the CO_2 generation rate per occupant. In practical scenarios, these two parameters undergo complex variations due to various external factors. However, for the sake of simplification in this study, these two parameters are set as constant values, which are selected as 0.07 kW/person [97], and 0.018 m^3/h/person [98], respectively.

$\hat{o}_{t+n}$ is also utilized for predicting the occupancy presence $\mathbb{1}_{\hat{o}>\alpha,t+n}$, which is equal to 1 when condition $\hat{o} > \alpha$ is true, and equal to 0 otherwise (Eq.(5-6)). This is equivalent to a simple filter to $\hat{o}$.

$$\mathbb{1}_{\hat{o}>\alpha,t+n} = cond(\hat{o}_{t+n}, \alpha) \tag{5–6}$$

Considering the precision of numerical computation, this study does not directly set 0 as the threshold. After tuning, it was found that 0.5 is an appropriate value, which is chosen as the threshold in this study. It means that when the expected number of people ($\hat{o}_{k+n}$) is greater than 0.5 (α), the room is considered as occupied; when less than or equal to 0.5, considered as unoccupied.

5.2 Building dynamic modelling based on stochastic differential equations and stochastic occupancy prediction

This study introduces SDE-based building dynamic models [79] to forecast the evolution of indoor temperature and CO_2 concentration. In this study, the generalized form of SDE models in continuous-time that embed the stochastic occupancy disturbance models is formulated as:

$$\mathrm{d}\boldsymbol{x_t} = f_s(\boldsymbol{x_t}, \boldsymbol{u_t}, \boldsymbol{d_t})\mathrm{d}t + g_s(\boldsymbol{x_t}, \boldsymbol{u_t}, \boldsymbol{d_t})\mathrm{d}\boldsymbol{\omega_{s,t}} \tag{5–7}$$

$$\mathrm{d}\boldsymbol{d_t} = f_d(\boldsymbol{o_t Q_t})\mathrm{d}t \tag{5–8}$$

$$\boldsymbol{y_{s,t}} = h_s(\boldsymbol{x_t}) + \boldsymbol{v}_{s,k} \tag{5–9}$$

$$\boldsymbol{y_{d,t}} = h_d(\boldsymbol{d_t}) + \boldsymbol{v}_{d,k} \tag{5–10}$$

where $\boldsymbol{x}$, $\boldsymbol{u}$, $\boldsymbol{d}$ are the building system states, the input and the disturbances respectively. f_s and f_d are the drift functions for system states and disturbances, and g_s are the diffusion functions for the system states. Eqs.(5-9, 5-10) are the observation equations for the observable states and disturbances. The occupancy-related disturbances are driven by an in-homogeneous Markov chain with forward Kolmogorov equation in continuous-time. $\boldsymbol{Q_t}$ represents the continuous-time in-homogeneous transition rate matrix. It can be discretized and transformed to transition probability matrix $\mathbf{P_t}$ using the matrix exponential operation.

$\boldsymbol{\omega_{s,t}}$ in Eq.(5-7) is standard Brownian motion to model process uncertainties. $\boldsymbol{v}_{s,k}\sim N(\mathbf{0},\ R_s)$ and $\boldsymbol{v}_{d,k}\ \sim N(\mathbf{0},\ R_d)$ in Eqs.(5-9, 5-10) are the observation noises. SDEs driven by Brownian motion denoted ω_t (in continuous time) can inherently represent unmeasured process uncertainties and explicitly describe measurement and model errors [99], which has independent increments with Gaussian distribution $\omega_t - \omega_s \sim N(0, t-s)$ [100].

Here, we further assume a linear-form equations of f_s and g_s as follows.

$$f_s(\boldsymbol{x_t}, \boldsymbol{u_t}, \boldsymbol{d_t})\mathrm{d}t = A\boldsymbol{x_t} + B\boldsymbol{u_t} + E\boldsymbol{d_t} \quad (5\text{–}11)$$

$$g_s(\boldsymbol{x_t}, \boldsymbol{u_t}, \boldsymbol{d_t}) = G \quad (5\text{–}12)$$

where A, B, E, and G are the state matrix, input matrix, disturbance matrix and diffusion matrix, respectively. These equations shall also be discretized to formulate MPC that are fast to solve. The model discretization is elaborated on Section 5.2.2.

SDEs provide a sound physical interpretation of parameters, utilizing knowledge of physical constants and balance relations. SDEs have been proven to be strong for modelling both building thermal dynamics using RC (resistance and capacity) thermal networks [99,101,102] and CO_2 concentration using mass balance equation [38,64], which we shall carry out in this chapter as well.

5.2.1 SDE models of room temperature and CO_2 concentration

This chapter considers heat dynamics of a room as an RC-diagram[99,101,102]. Figure 5-3 shows the thermal dynamics structure for the room. The equivalent linear SDE model has the form shown in Eqs.(5-13,5-14), which describes the physical dynamics and interactions governing the indoor air temperature.

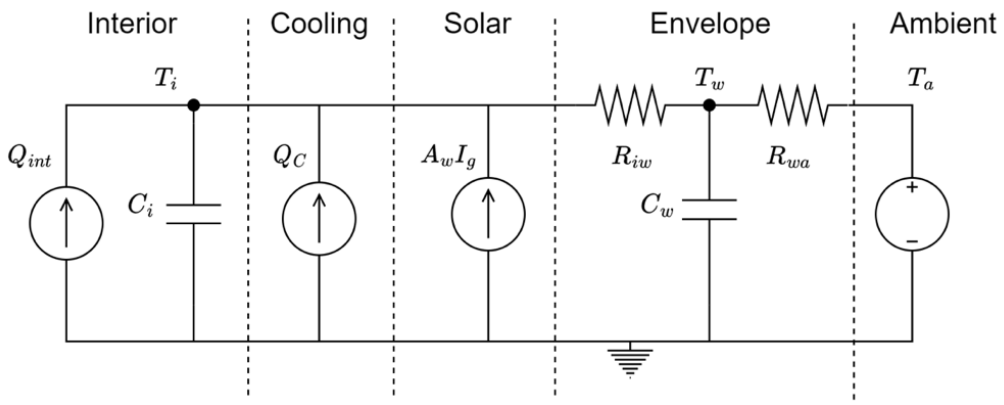


Figure 5-3. RC-diagram of thermal dynamics structure for the building

$$dT_{in,t} = \frac{1}{C_{in}}\left(\frac{1}{R_{inw}}\left(T_{w,t} - T_{in,t}\right) + A_w I_{g,t} - Q_{c,t} + r_{Q_{int}} \cdot \hat{o}_t\right) dt + \sigma_1 d\omega_{in,t} \quad (5\text{–}13)$$

$$dT_{w,t} = \frac{1}{C_w}\left(\frac{1}{R_{inw}}\left(T_{in,t} - T_{w,t}\right) + \frac{1}{R_{wa}}\left(T_{a,t} - T_{w,t}\right)\right) dt + \sigma_2 d\omega_{w,t} \quad (5\text{–}14)$$

Table 5-1 lists the interpretation of variables and parameters in Eqs. 5-13 and 14. The states vector of the thermal model $\boldsymbol{x_{T,t}} = [T_{in,t}, T_{w,t}]^T$ are given by the temperature $T_{in,t}$ of the room air with the capacity C_{in}, and by the temperature $T_{w,t}$ of the external wall with the heat capacity C_w.

Normally, we could measure only the indoor air temperature $T_{in,t}$, while the external wall temperatures are usually not measured, which means that they are hidden states. Therefore, the observation equation in Eq.(5-9) is $h_s(\boldsymbol{x_{T,t}}) = \mathrm{C}\boldsymbol{x_{T,t}}$ with $\mathrm{C} = [1,0]$. R_{inw} is the resistance against heat transfer between the room air and the external wall, while R_{wa} is the resistance against heat transfer from the external wall to the ambient air with the temperature $T_{a,t}$.

The control input vector of thermal model $\boldsymbol{u_{T,t}} = [-Q_{c,t}]$ is the cooling energy supply. The disturbance vector of the thermal model $\boldsymbol{d_{T,t}} = [\hat{Q}_{int,t}, I_{g,t}, T_{a,t}]$ acting on a building is occupancy heat gains and weather. The occupancy heat gains associate closely with occupant number. The solar radiation, $I_{g,t}$ penetrates through the windows, where A_w is the effective window area. Note that the weather disturbances are assumed

to be perfectly predictable and not modelled in this study, as the primary focus is on forecasting disturbances associated with occupants. $C_{in}, C_w, R_{inw}, R_{wa}, \sigma_1, \sigma_2, A_w$ are parameters to be estimated. For an introduction to how to develop a suitable SDE model and estimate model parameters for the thermal dynamics of a building/room, see [101]. The SDE model of room CO_2 concentration is constructed using mass balance equation, which has the following form [38,64],

$$dc_{\mathrm{CO}_2,in,t} = \frac{1}{V}\Big(\big(F_{a,t} + F_{inf}V\big)\cdot\big(c_{\mathrm{CO}_2,out,t} - c_{\mathrm{CO}_2,in,t}\big) + r_{CO_2}\cdot\hat{o}_t\Big)\,dt + \sigma_3 d\omega_{\mathrm{CO}_2,t} \tag{5–15}$$

The interpretation of variables and parameters in Eq. (5-15) is listed in Table 5-1. The states of the CO_2 model $\boldsymbol{x_{c,t}} = [c_{\mathrm{CO}_2,in,t}]$ are given by the indoor CO_2 concentration. The control input of the CO_2 model $\boldsymbol{u_{c,t}} = [F_{a,t}]$ is the fresh air flow rate. The disturbances of the CO_2 model $\boldsymbol{d_{c,t}} = [\widehat{W}_{gen,t}]$ acting on a building is CO_2 generation, which also associate closely with occupant number.

Table 5-1 A list of the parameters in the SDE in Eqs. 17-19 and their interpretations

Variable	Description	Unit
$T_{in,t}$	The indoor air temperature	°C
$T_{w,t}$	The external wall temperature	°C
C_{in}	The heat capacity of the indoor air and part of inner wall and floor slab	kWh/ °C
C_w	The heat capacity of the external walls	kWh/ °C
R_{inw}	Thermal resistance between the indoor air and external walls	°C/kW
R_{wa}	Thermal resistance between the external walls and outdoor air	°C/kW
$T_{a,t}$	Outdoor air temperature	°C
$\omega_{in,t}$	Brownian motion driving the indoor air temperature	$\sqrt{\mathrm{t}}$
$\omega_{w,t}$	Brownian motion driving the external wall temperature	$\sqrt{t}$
σ_1	Scaling parameter to scale the Brownian motion of the indoor air temperature	°C /$\sqrt{t}$

σ_2	Scaling parameter to scale the Brownian motion of the wall temperature	°C /$\sqrt{t}$
A_w	Effective area of the solar radiation in the room	m^2
$I_{g,t}$	Solar radiation on a horizontal surface	kW/ m^2
$Q_{c,t}$	Cooling supply to the room air	kW
$\hat{o}_t$	Occupant number	person
$r_{Q_{int}}$	Heat dissipate rate per occupant, which is selected as 0.07 kW/person [97]	kW/person
$c_{\mathrm{CO}_2,in,t}$	The indoor CO_2 concentration	ppm
$c_{\mathrm{CO}_2,out,t}$	The outdoor CO_2 concentration, assumed to be 450ppm in this study	ppm
V	The room volume	m^3
$F_{a,t}$	The fresh air flow rate	m^3/h
F_{inf}	The infiltration rate, which is assume to be 0.2 in this study [98]	$1/h$
r_{CO_2}	The CO_2 generation rate per occupant, which is selected as 0.018 m^3/h/person [98]	0.018 m^3/h/person

It is important to note that in hot summer, when supplying fresh air to the room, the temperature of fresh air (T_{fresh}) is typically pre-conditioned to a certain level in the PAU, which is usually lower than the indoor temperature setpoint of the FCU. Therefore, the supply of fresh air also introduces a certain amount of cooling to the room. This cooling effect can be calculated using $Q_{c,v,t} = c_a\, \rho_a\, F_{a,t} \cdot (T_{fresh} - T_{in,t})$. This factor would introduce a bilinear non-linear term ($F_{a,t} \cdot T_{in,t}$) to the MPC optimization, requiring more time for optimization. Considering that the cooling of the fresh air is significantly less than that of the FCU, and to maintain linearity and enhance computational efficiency in solving the optimization problem, this non-linear factor is neglected in the room RC model and MPC optimization. However, in the performance evaluation for control strategies in the white-box model, the cooling effect of fresh air supply on indoor temperature is considered.

5.2.2 Model discretization

Due to the linearity of the temperature- and CO_2 -models in Eqs. 5-13-15, the

continuous-time models are discretized to formulate optimal control problems that are faster to solve. The stochastic discrete-time models for room temperature and CO_2 have the forms:

$$\mathbf{T}_{t+1} = A_{T,d}\mathbf{T}_t + B_{T,d}u_{T,t} + E_{T,d}d_{T,t} + w_{T,t} \quad (5\text{–}16)$$

$$c_{CO_2,in,t+1} = A_{C,d}c_{CO_2,in,t} + B_{C,d}u_{C,t} + E_{C,d}d_{C,t} + w_{C,t} \quad (5\text{–}17)$$

where, $\boldsymbol{T_t}$ is the vector of system states of room temperature $[T_{in,t}, T_{w,t}]$ at discrete time step t, $u_{T,t}$ is the control variable, which is the cooling supply to the room $-Q_{c,t}$. This control variable is constrained by the maximum cooling supply capacity of the FCU.

$$Q_{c,t} \leq Q_{c,max} \quad (5\text{–}18)$$

In a corresponding implementation of this algorithm, the cooling supply is not implemented since this is not directly controllable, instead the air temperature (T_{in}) could be directly adjusted to the set-points ($T_{in,set,t}$) in thermostat of FCU and linked to the cooling supply through process controller [78,79]. The relationship can be written like

$$Q_{c,t} \propto \dot{m}(T_{in,set,t} - T_{in,t}) \quad (5\text{–}19)$$

where $\dot{m}$ is the mass flow rate of the air in the FCU. In such a model, it is also necessary to know the dynamic behaviour of the mass flow rate (e.g. if it depends on set point or is controllable itself).

Here, $d_{T,t}$ is the disturbance vector, $[\widehat{Q}_{int,t}, I_{G,t}, T_{a,t}]$. $c_{CO_2,in,t}$ is the system states of indoor CO_2 concentration at time step t, $u_{C,t}$ is the supplied fresh air flow rate $F_{a,t}$, which is also constrained by the maximum and minimum fresh air flow rates,

$$F_{a,min} \leq F_{a,t} \leq F_{a,max} \quad (5\text{–}20)$$

$d_{C,t}$ is CO_2 generation $\widehat{W}_{gen,t}$. $A_{T,d}$, $B_{T,d}$, $E_{T,d}$, $A_{C,d}$, $B_{C,d}$, and $E_{C,d}$ are the corresponding matrices of the discrete-time model, and $w_{T,t}$ and $w_{C,t}$ represent process noises of indoor air temperature and CO_2 concentration.

5.2.3 Kalman Filter

In a system that is encumbered with noise or where not all of states are directly observed, e.g. the temperature of external wall $T_{w,t}$, it is often necessary to apply a *filter*. This has the purpose of estimating the underlying system including hidden (non-observed) states. It is important for the performance of the optimal controller to have knowledge about all the system states; e.g., walls made of concrete stores significant amounts of heat and thus 1°C differences in the estimated and real temperature of the wall might skew the estimated amount of energy in the system critically.

For linear systems with state-independent diffusion function, the *Kalman Filter* provides optimal state estimations by combining measurements and one-step predictions using the dynamic model. Technically speaking, the Kalman filter is adopted to estimate the expectation of the underlying Gaussian system in an optimal way, that is the conditional mean and variance, $\mathbf{E}[x_t|\mathcal{Y}_t]$and $\mathbf{E}[P_t|\mathcal{Y}_t]$, respectively. Here, the estimates are conditioned on all historical observations, $\mathcal{Y}_t = \{y_1, y_2, \dots, y_t\}$ up till time *t*. The Gaussianity of the system is a key assumption, and the estimator is only approximated if this assumption does not hold. However, it is often a good approximation and is widely applied in the form of the extended Kalman filter due to its simplicity and fast numerical properties while usually giving good results.

The Kalman filter method is widely used for state estimation and parameter identification. For an introduction to Kalman filter methods, refer to [103]. In previous research, the Kalman filter has been extensively applied to online state estimation of building SDE models [65,70].

In the online prediction using the SDE model, not all state variables are directly measurable in real-world applications. For the closed-loop online control case study presented here, only the indoor air temperature is measured. The Kalman filter offers a computationally efficient solution for estimating process states while minimizing the prediction mean squared error. This method operates recursively, with each updated state estimate calculated based on the previous estimate and the newly measured data. Each cycle of the Kalman filter consists of two main steps. The first step is time update

(prior estimate) which predicts the current state based on the previous state and its uncertainty. The second step is measurement update (posterior estimate), which adjusts the predicted state using the new measurement data to refine the estimate. This recursive cycle of prediction and correction ensures that the Kalman filter provides an optimal estimate of the system state, even in the presence of noise and uncertainty.

Offline parameter identification involves estimating the parameters of a system model based on historical data, ensuring the model accurately represents the underlying system dynamics, in which the continuous-discrete extended Kalman filter (CDEKF) is widely used. The literature [104] and [79] introduced the CDEKF for offline parameter identification of SDE models. The CDEKF includes two steps, a prediction step and an update step. In the first step, the state evolution is predicted using the continuous-time SDE model. In the second step, the parameters are updated whenever new discrete measurements are available, using an CDEKF framework to linearize the nonlinear system dynamics around the current state estimates.

5.3 Stochastic occupancy-integrated model predictive control scheme

5.3.1 MPC structure

The optimal controller formulated in this study aims to optimizing the trajectory of indoor air temperature set-point ($\{\mathrm{T_{in,set,t+n}}\}_1^N$) and the fresh air flow rate ($\{\mathrm{F_{a,t+n}}\}_1^N$) over the prediction horizon *N* at time step *t*. This is achieved by solving a rolling optimization problem at each time step *t* to minimize thermal comfort penalty (ϕ^{TC}), IAQ penalty (ϕ^{IAQ}) and energy consumption penalty (ϕ^{E}). The overall MPC structure can be written as Eqs.(5-21,22). The subscript *t*+*n* means the values at *t*+*n* time step. The weighted sum approach is adopted to convert the multi-objective optimization problem into a single objective one. ρ_{TC}, ρ_{IAQ} and ρ_E are factors to weight the importance of the terms relative to each other. How to choose the weights is not trivial and depends on the specific preferences of the stakeholders of the meeting room. The

effects of the combination of the weights are explored in Section 5.5.3.

$$\min_{\{\mathrm{T_{in,set,t+n}}\}_1^N,\{\mathrm{F_{a,t+n}}\}_1^N} \sum_{n=1}^{N} \mathbb{1}_{\hat{o}>\alpha,t+n} \cdot (\rho_{TC} \cdot \phi_{t+n}^{TC} + \rho_{IAQ} \cdot \phi_{t+n}^{IAQ}) + \rho_E \cdot \phi_{t+n}^{E} \tag{5–21}$$

subject to

$$\begin{cases} \text{occupancy prediction} & \pi_{t+n} = \pi_t \cdot \prod_{k=t}^{t+n-1} \mathbf{P_k} \\ & \hat{o}_{t+n} = \pi_{t+n} \cdot \boldsymbol{y^T} \\ & \mathbb{1}_{\hat{o}>\alpha,t+n} = cond(\hat{o}_{t+n}, \alpha) \\ & \hat{Q}_{int,t+n} = r_{Q_{int}} \cdot \hat{o}_{t+n} \\ & \hat{W}_{gen,t+n} = r_{CO_2} \cdot \hat{o}_{t+n} \\ \text{building dynamics} & \mathbf{T}_{\mathrm{t+n}} = \mathrm{A_{T,d}}\mathbf{T}_{t+n-1} + \mathrm{B_{T,d}}\mathbf{u_{T,t+n-1}} + \mathrm{E_{T,d}}\mathbf{d_{T,t+n-1}} \\ & \mathrm{c_{CO_2,in,t+n}} = \mathrm{A_{C,d}c_{CO_2,in,t+n-1}} + \mathrm{B_{C,d}u_{C,t+n-1}} + \mathrm{E_{C,d}d_{C,t+n-1}d_{C,t+n-1}} \\ \text{operational constraints} & \mathrm{Q_{c,t+n}} \le \mathrm{Q_{c,max}} \\ & \mathrm{F_{a,min}} \le \mathrm{F_{a,t+n}} \le \mathrm{F_{a,max}} \\ \text{initial condition} & \pi_t = [p_1^t, p_2^t, \dots, p_s^t] \\ & \mathbf{T}_t = \widehat{\mathbf{T}}_t \\ & \mathrm{c_{CO_2,in,t}} = \hat{c}_{\mathrm{CO_2,in,t}} \end{cases} \tag{5–22}$$

The occupancy presence ($\mathbb{1}_{\hat{o}>\alpha,t+n}$) is embedded in the cost function of MPC serving as an additional weighting factor for the thermal comfort and IAQ objectives, which hold meaningful when the room is predicted to be occupied, as this indicates the actual need for the built environment to be controlled. With the occupancy presence as weighting factor, violation of the desired room temperature is only considered when occupants are predicted to present in the room, enabling flexible energy supply, such as pre-cooling, during times with no occupants. The occupancy presence is also weighted for IAQ, violation of the desired IAQ is only considered when occupants are predicted to present in the room.

The constraints (Eq. (5-22)) of the optimal control problem include the occupancy

prediction constraints, building dynamic constraints, operational constraints, and initial condition. The occupancy presence ($\mathbb{1}_{\hat{o}>\alpha,t+n}$) is calculated based on the predicted results of occupant number and a threshold α. If the predicted occupant number is higher than the threshold, the room is predicted to be occupied. This constraint can be considered as a switch: thermal comfort and IAQ objectives are only considered when the room is predicted to be occupied. For example, if the room is predicted to be unoccupied for the next half hour but occupied afterward, the thermal comfort and IAQ objectives will only be included in the objective function starting from the time the room is predicted to be occupied. The initial conditions for optimization are obtained through real-time measurements from sensors and the Kalman filter. Specifically, the initial values of observable states, including indoor temperature and CO_2 concentration are obtained through real-time detection, while the initial value of the hidden state, i.e. wall temperature is obtained through a Kalman filter.

5.3.2 Cost function

(1) Thermal comfort penalty (ϕ_t^{TC})

Thermal comfort is represented by a quadratic penalty function as shown in Eq.(5-23).

$$\phi_t^{TC} = \left(\mathrm{T_{in,t}} - \mathrm{T_{comf}}\right)^2 \tag{5–23}$$

The penalty function calculates the deviation between the indoor air temperature and desired reference temperature, T_{comf}. The latter is chosen as 24°C in this study. The quadratic thermal comfort penalty function is commonly used to formulate the cost function of MPC in literature[39,105].

(2) IAQ penalty (ϕ_t^{IAQ})

IAQ is represented by the concentration of CO_2 as the representative of indoor pollutant [106,107]. Due to metabolic processes of human, indoor CO_2 concentration level varies as the number of occupants changes. The cost function is formulated as shown in Eq.(5-24) [106].

$$\phi_t^{IAQ} = \max\left(0, \mathrm{c_{CO_2,in,t}} - \mathrm{c_{TH,CO_2}}\right) \tag{5–24}$$

$c_{\mathrm{TH,CO_2}}$ is the control threshold of CO_2 in the optimization, which adopts the value of its prescribed level (800ppm) in the IAQ control standard [108]. The max operation can be linearized by replacing $\max\left(0, \mathrm{c_{CO_2,in,t}} - \mathrm{c_{TH,CO_2}}\right)$ with variables v, δ and adding the following constraints.

$$\mathrm{v}_t = \mathrm{c_{CO_2,in,t}} - \mathrm{c_{TH,CO_2}} + \delta_t \tag{5–25}$$

$$\mathrm{v}_t \geq 0 \tag{5–26}$$

$$\delta_t \geq 0 \tag{5–27}$$

(3) Energy consumption (ϕ_t^E)

The energy consumption refers to the electricity used to supply cooling to the room and to process the fresh air, as shown in Eq.(5-28).

$$\phi_t^E = \frac{\mathrm{Q_{c,t}} + \mathrm{Q_{f,t}}}{\mathrm{COP}} = \frac{\mathrm{Q_{c,t}}}{\mathrm{COP}} + \left(\mathrm{H_{out,t}} - \mathrm{H_{sup}}\right)\frac{\mathrm{F_{a,t}}}{\mathrm{COP}} \tag{5–28}$$

The cooling of fresh air $\mathrm{Q_{f,t+n}}$, which is the product of the specific enthalpy difference between supplied fresh air and outdoor air and the fresh air flow rate $\mathrm{F_{a,t+n}}$. $\mathrm{H_{sup}}$ and $\mathrm{H_{out,t+n}}$ is the specific enthalpy of supply fresh air and outdoor air, respectively. This study assumes that the fresh air is conditioned to 18°C and 60% RH (relative humidity), and thus $\mathrm{H_{sup}}$ is a constant value, which is 37.61 kJ/kg. $\mathrm{H_{out}}$is calculated using the predicted outdoor temperature and RH. COP is the coefficient of performance of the chiller. A typical constant value of COP, i.e. 3, is chosen for case study in this chapter. Nevertheless, if a higher COP were used in the MPC, say 7, it would reduce the energy consumption objective (ϕ_t^E), which corresponds to lower the weight of energy consumption objective (ρ_E). Therefore, it can potentially improve building thermal comfort and air quality, assuming ρ_{TC} and ρ_{IAQ} remain constant.

The MPC structure shown in Eqs.(5-21, 22) can be converted to a quadratic programming problem, and is solved using Gurobi software. The optimization result at each time step is a trajectory of control signals, with only the first signal implemented.

5.4 Simulation case study

5.4.1 Description of the target space and its air-conditioning system

The target space is the meeting room described in Section 3.3 as the prototype in the simulation study. The layout is presented in Figure 5-4 for convenience.

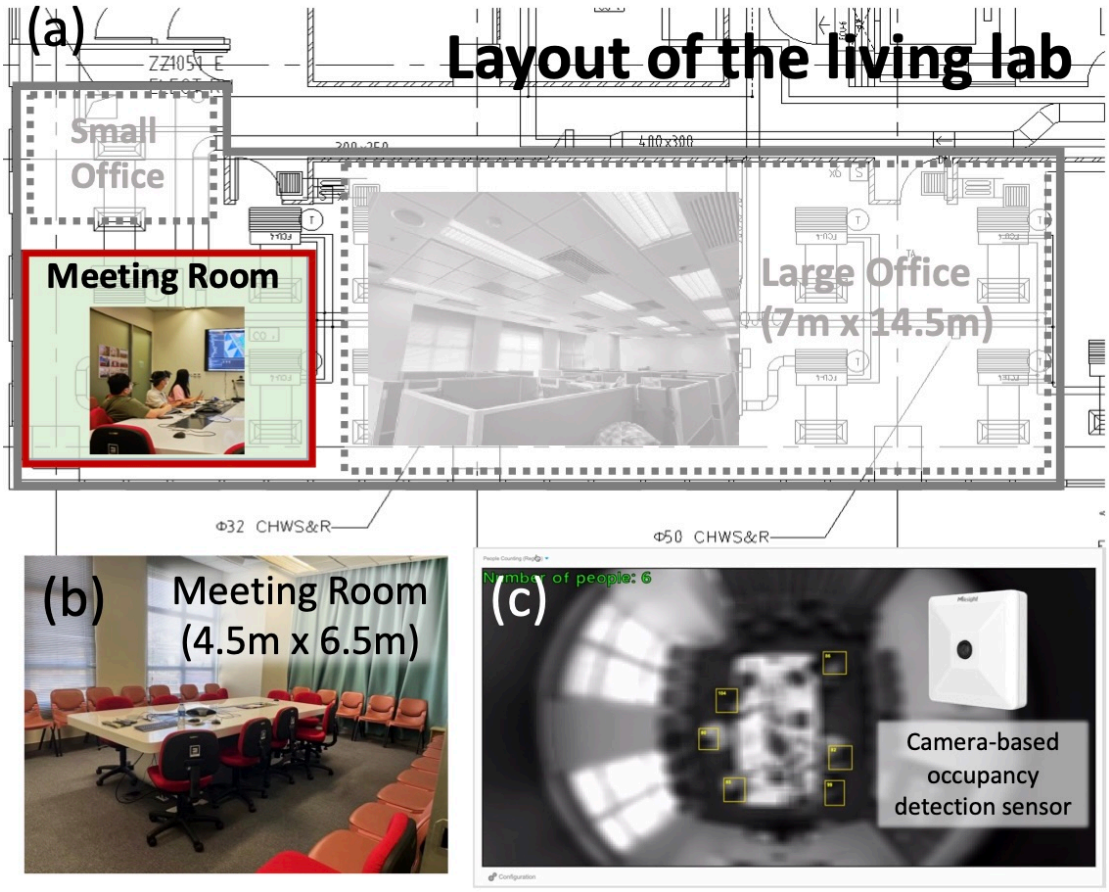


Figure 5-4. The target meeting room. (a) Layout of the office suite, which includes a large open-plan office, a small private office, and a meeting room, which is the target space adopted in this study. (b) Photo of the meeting room. (c) Camera-based occupancy detection sensor installed in the meeting room detecting the occupant number during a meeting.

The energy system of the meeting room is shown in Figure 5-5. The indoor air temperature is controlled by a FCU which is connected with a thermostat to adjust the temperature set-point. In FCU, there is a heat exchanger for air and chilled water. The

opening degree of chilled water valve in FCU is controlled through FCU process controller to maintain the return air temperature, i.e. the room air temperature at setpoint. Chilled water is supplied from the central chiller plant through chilled water supply pipelines. The fresh air is conditioned and supplied by a PAU/VAV system. The VAV box is connected to the BAS, which can send the fresh air flow rate setpoint to the VAV box. The VAV process controller compares the measured fresh air flow rate with the setpoint to control the opening of the fresh air valve.

Real occupant number data was collected in the meeting room. The details of occupancy sensor and data collection process can be found in Section 3.3. In this study, the proposed MPC strategy is deployed in a TRNSYS-Python co-simulation platform, which is elaborated in the next section. And the performance of MPC is compared with baseline strategies, including fixed-schedule strategy and OBC.

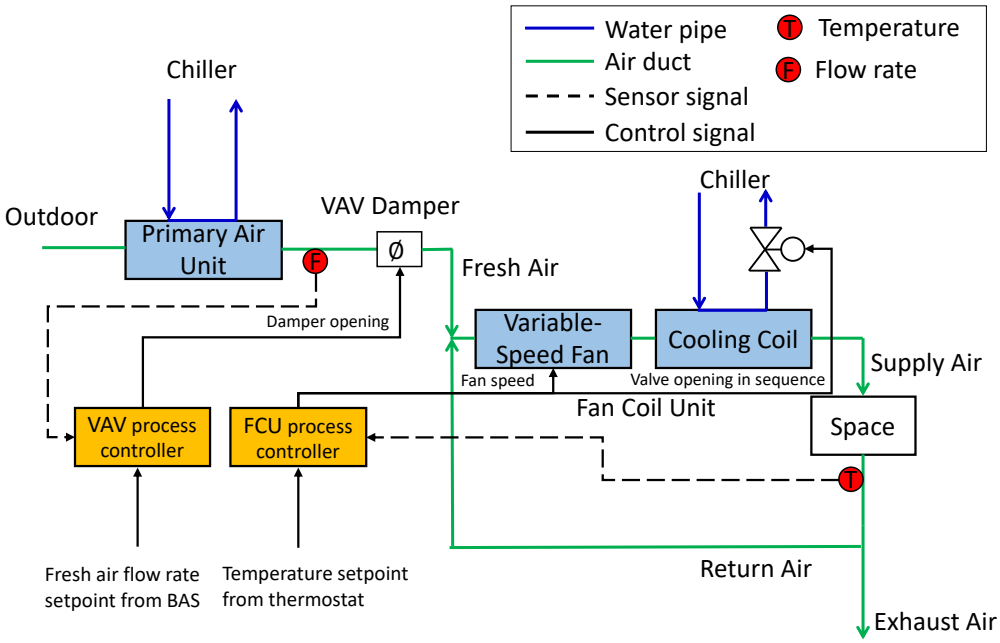


Figure 5-5. HVAC systems of the target meeting room. Indoor air temperature is controlled by a fan coil unit (FCU). The fresh air is conditioned and supplied by a primary air unit (PAU) and the fresh air flow rate is controlled by a VAV box

5.4.2 TRNSYS-Python co-simulation platform

Building energy simulation program such as TRNSYS [109–112] is commonly used in simulation of building thermal performance and energy performance of HVAC system and control strategy. As illustrated in Figure 5-6, we combine TRNSYS with Python to embed optimal supervisory controller, providing optimal setpoints to the process controllers. In this study, the Type 163 is used to build the communication between the models in TRNSYS 18 (64-bit) and Python (64-bit).

By coupling TRNSYS with Python via co-simulation, we could combine the strengths of both tools. TRNSYS handles the accurate representation of building physics and HVAC dynamics, while Python is used to implement advanced MPC algorithms in a flexible and efficient manner. This co-simulation approach enables real-time communication between the two platforms, allowing the MPC to interact dynamically with the building model during simulations.

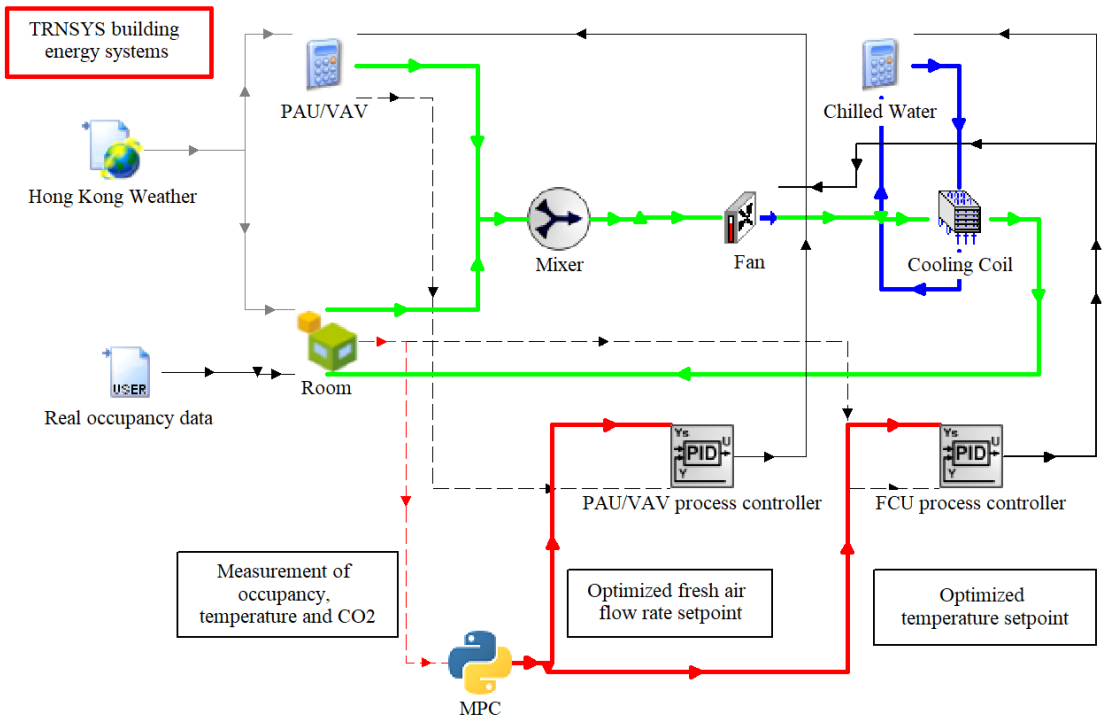


Figure 5-6. TRNSYS-Python co-simulation platform

Other building energy simulation tools (e.g., EnergyPlus, DeST) also support detailed building modeling but lack the same modular co-simulation flexibility with Python for

advanced controls. Tools like MATLAB/Simulink could also be used for control implementation, but they may not provide the same ease of integration with building simulation platforms or the scalability and versatility of Python for advanced algorithm development. TRNSYS-Python co-simulation offers a unique combination of detailed building modeling and advanced control implementation, which is essential for the objectives of this study.

In the simulation, the control horizon of the MPC is set to 5 minutes, and the prediction horizon is set to 2 hours with a 5-minute time interval. The horizon for occupancy prediction is the same as the MPC prediction horizon.

On average, the occupancy prediction for each run takes around one second, while solving the optimization problem in Eqs. (5-21, 5-22) requires approximately 1 minute per run.

In the integrated building energy system, the building thermal dynamic model (Type 56) in TRNSYS is used to characterize the building thermal performance under the influences of weather in Hong Kong and real occupancy in the meeting room. The information about the building envelope is listed in Table 5-2. The air-conditioned room is integrated with FCU and PAU/VAV systems. The specification of the HVAC systems is listed in Table 5-3.

The outputs of the MPC controller are the temperature set-point for the thermostat of FCU and fresh air flow rate set point for the VAV box. The PID parameters of FCU process controller and VAV process controller have been manually tuned to make sure that the control variables can well track the setpoints. The issue of PID parameter tuning is beyond the scope of this study; therefore, this study does not investigate the impact of PID parameters on control performances.

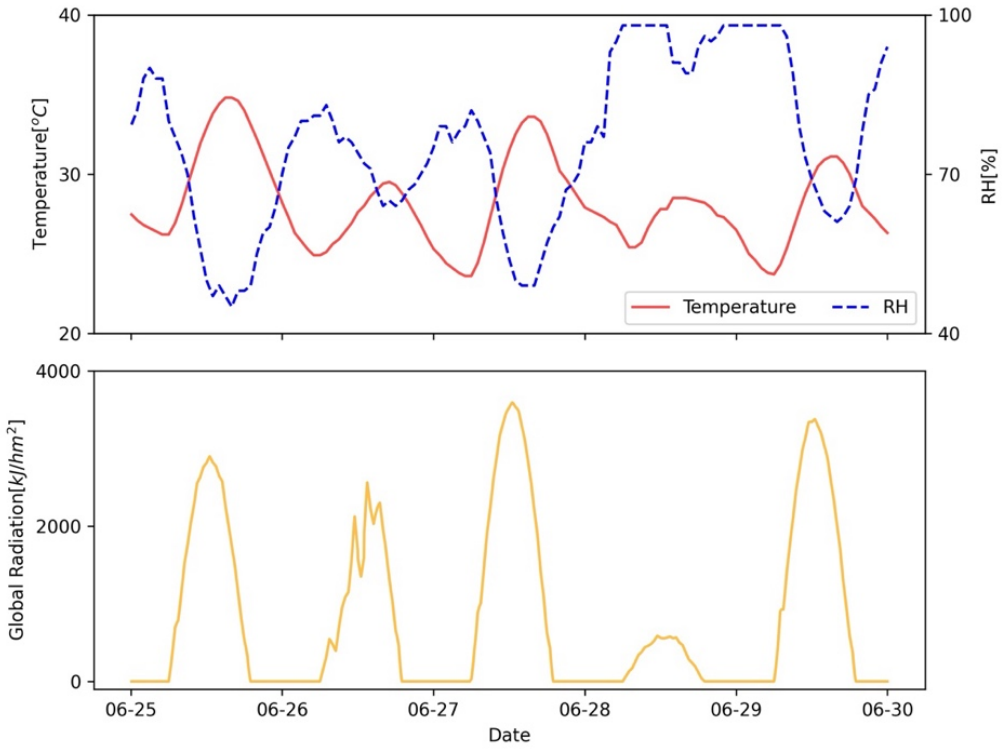


Figure 5-7. Weather condition on the 5 testing days

Table 5-2 Information about the building envelope

Construction type	Layers	Thickness (mm)	U-value ($W/m^2 \cdot K^{-1}$)
External walls	Brick+ insulation + plaster	355	0.339
Internal walls	Gypsum + insulation + gypsum	74	0.652
Floor/ceiling	Concrete+ insulation	400	0.233
External windows	Double glazing	12	1.10

The test case is conducted on 5 consecutive typical hot and humid summer days in Hong Kong (from June 25 to June 29). The data of typical meteorological year (TMY) of Hong Kong is adopted. Figure 5-7 shows the outdoor air temperature, RH and global solar radiation on the test days. The real measurement of occupant number of the same days is adopted for the test, which is shown in Figure 5-12. The initial indoor

temperature is set as 25°C in Type 56 building thermal dynamic model.

Table 5-3 Specification of the equipment of HVAC systems

Description	Unit	Value
COP of chiller	-	3
Maximum fresh air flow rate of VAV	kg/h	500
Minimum fresh air flow rate of VAV	kg/h	10
Rated flow rate of VSF	kg/h	2000
Rated power of VSF	kW	0.75
Motor efficiency of VSF	%	90
Maximum chilled water flow rate of coil	kg/h	200
Chilled water supply temperature of coil	°C	7
Specific Heat of chilled water	kJ/(kg °C)	4.18

5.4.3 System model identification

The parameters of SDE-based grey-box room temperature model integrated in MPC for predicting building responses under disturbances are identified using the extracted data from the Type 56 building model in TRNSYS. The extracted data includes indoor air temperature, cooling supply, internal heat gains, outdoor temperature, and solar radiation. The maximum likelihood-based R-package *ctsmr* [113] is applied for the SDE parameter estimation, which has been validated in numerous real projects [100,114,115]. The identified parameters of SDE-based room temperature model are shown in Table 5-4.

Table 5-4 Identified and estimated parameters in the SDE model in Eqs.17-18

Variable	C_{in}	C_w	R_{inw}	R_{wa}	σ_1	σ_2	A_w
Unit	kWh/ °C	kWh/ °C	°C/kW	°C/kW	°C /$\sqrt{t}$	°C /$\sqrt{t}$	m^2
Identified value	0.123	8.602	2.568	13.716	0.102	0.254	1.15

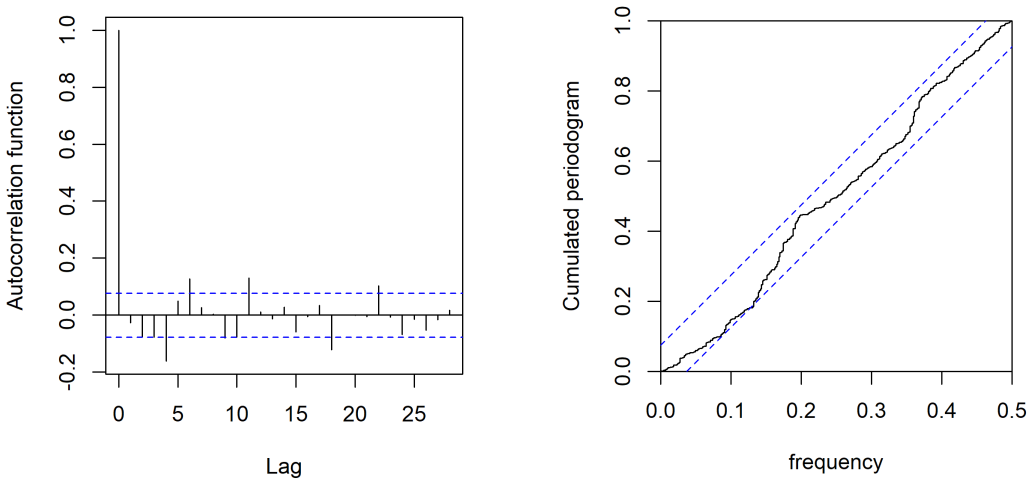


Figure 5-8. The ACF and the CPG of the prediction residuals of the SDE-based room temperature model.

Figure 5-8 shows the autocorrelation function (ACF) and the cumulated periodogram (CPG) of the prediction residuals of the SDE-based room temperature model. The ACF and the CPG of the SDE model indicate that the residuals can be viewed as white noise. The TRNSYS data and RC model data of indoor temperature are compared as shown in Figure 5-9. The prediction errors between are evaluated using root mean square error (RMSE). The RMSE of indoor air temperature between TRNSYS data and SDE model data are 0.26 °C, indicating the SDE model is able to accurately predict the room thermal dynamics.

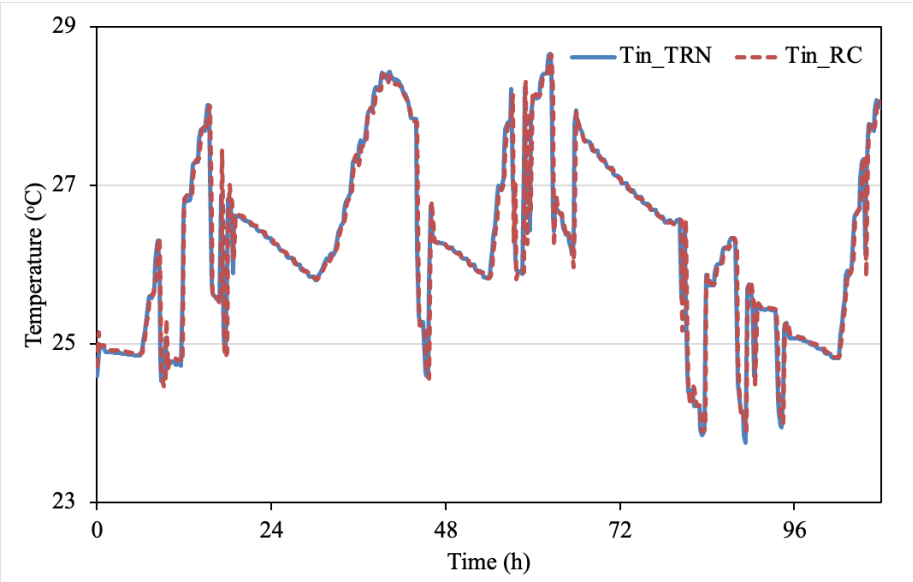


Figure 5-9. SDE room temperature model performance from TRNSYS and RC model.

5.4.4 Control strategies used in comparison test

This section describes two baseline strategies and the proposed strategy for built environment control. Strategy 1 is fixed-schedule strategy, which is commonly adopted in current commercial BAS, but often lacks responsiveness to realistic and dynamic occupancy. Strategy 2 is the OBC. Based on a set of rules, it gives setpoint responding to real-time occupancy detection. It is a straightforward way of considering real occupancy in built environment control in literature [29,59,60], but may cause suboptimal thermal comfort and IAQ, particularly in spaces with highly stochastic occupancy patterns. The baseline strategies are compared with the proposed MPC strategy. The three control strategies are summarized in the Table 5-5.

Table 5-5 Comparison of baseline and proposed strategies for built environment control

	Temperature control setpoints	Fresh air flow rate control setpoints
Strategy 1_FIX (Fixed-schedule control)	24°C at working hours*; 30°C at non-working hours	Eq.(33) using designed occupant number
Strategy 2_OBC (Occupancy-based control)	24°C if the room is occupied; 30°C if the room is empty.	Eq.(33) using real occupant number
Strategy 3_MPC (Occupancy-integrated MPC)	Optimal setpoints	

* Working hours is from 8 am to 12am and from 2pm to 6pm; 8 hours in total

Strategy 1 *(Fixed-schedule control, FIX)*: A simple control strategy commonly adopted in commercial BAS. Both the temperature set-point and fresh air flow rate is controlled

based on pre-set schedules. In this study, the indoor air temperature set-point is set as 24°C at working hours from 8am to 12am and from 2pm to 6pm, and 30°C at non-working hours. The fresh air flow rate is determined using the Eq. (5-29)

$$V_{bz,t} = R_p \times P_{z,t} + R_a \times A_z \quad (5\text{–}29)$$

where A_z is room floor area; P_z is designed number of people in the room, which is chosen as 15 in working hours and 0 in non-working hours based on the recommended occupancy density; R_a is the fresh air flow rate per floor area, 0.3 L/s/m^2 and R_p fresh air flow rate per people, 2.5 L/s/person, both of which is determined based on the ASHERAE recommendation for meeting room. It results 47L/s at working hours and 9L/s at non-working hours [97]. This strategy applies no optimization, and no occupancy related data.

Strategy 2 *(Occupancy-based control, OBC)*: Strategy 2 is a rule-based control strategy reactive to real-time occupancy. Both temperature set-point and fresh air flow rate set-point are determined in real-time based on the real-time occupant number. When the room is detected as occupied, the temperature set-point of rooms is set to 24°C. When the room is detected as unoccupied, the temperature set-point of rooms is set to 30°C. The fresh air flow rate set-point is determined based on real-time occupant number detection using Eq.(33), where $P_{z,t}$ is real occupant number. This strategy applies no optimization but requires real-time occupant number data. The schematic diagram of the strategy 2 is shown in Figure 5-10.

Strategy 3 *(Proposed occupancy-integrated MPC)*: The MPC can be viewed as a high-level controller sending the optimized indoor air temperature set-point and fresh air flow rate set-point to the process controllers. The prediction horizon and prediction interval are chosen as 2 hours and 5 minutes, respectively. The MPC is run in closed loop such that the first value of the sequence of optimized control values are implemented in the TRNSYS model every 5 minutes in simulation time. Figure 5-11 shows the proposed MPC strategy for the office energy systems. The weighting factor of ρ_{TC} , ρ_{IAQ} and ρ_E are chosen as 0.1, 0.1 and 0.1 respectively. Note that the choice of weights for objectives is not trivial and depends on the specific preferences of the

stakeholders of the room. The effects of the combination of the weights are further explored in Section 5.5.3.

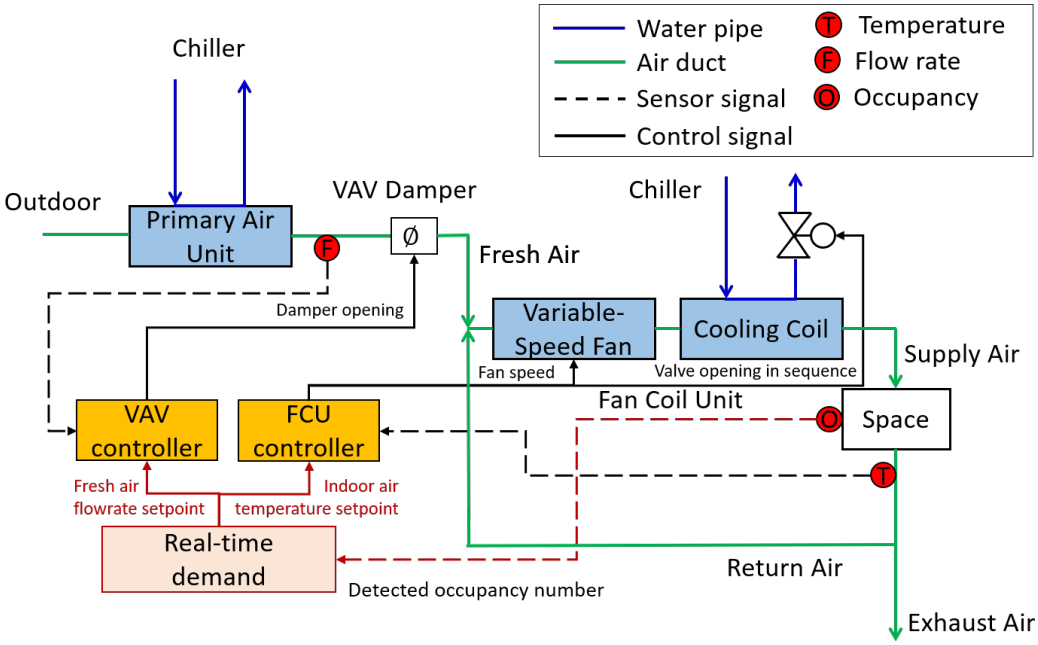


Figure 5-10. Schematic of OBC strategy (Strategy 2)

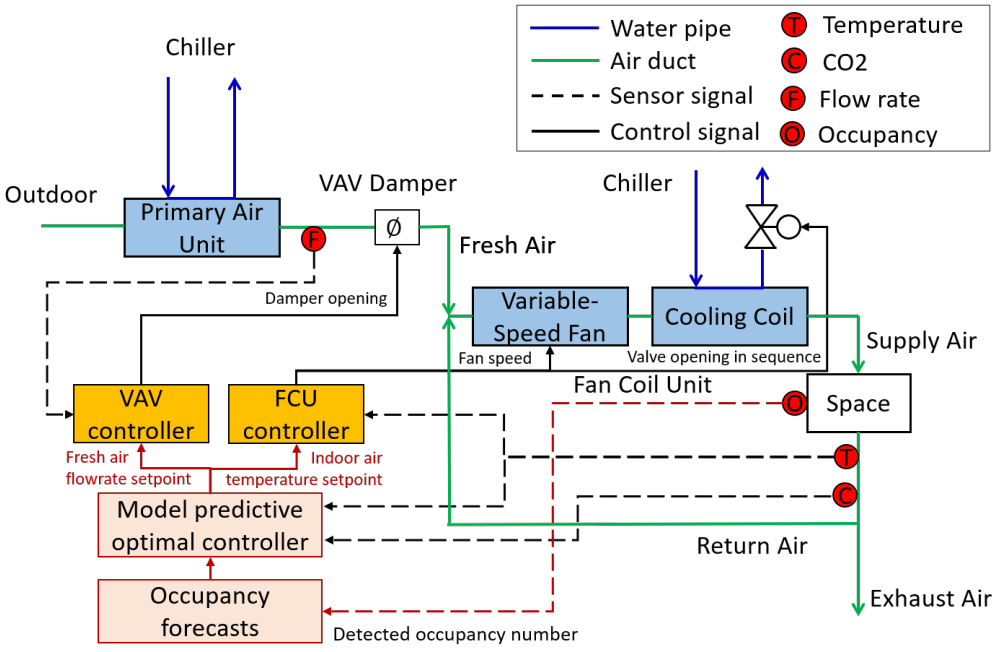


Figure 5-11. Schematic of the proposed occupancy-integrated MPC (Strategy 3)

5.4.5 Performance evaluation indicators

In this study, to quantitively compare indoor comfort levels under different strategies,

temperature deviations during occupied periods from the most comfortable temperature, i.e. 24°C are evaluated. It is important to note that the comparison does not utilize the thermal comfort penalty with a quadratic term in the cost function (Eq.(5-23)). Instead, it involves taking the square root of the deviation, creating a linearized deviation calculation similar to RMSE (Root Mean Square Error), referred to as $RMSE_T$ in this study (Eq.(5-30)). The interpretation of $RMSE_T$ is the square root deviation between the indoor temperature and the most comfortable temperature during occupied periods. By using $RMSE_T$, both temperatures higher than 24 °C and lower than 24 °C are considered as deviations.

$$RMSE_T = \sqrt{\frac{\sum_{k=1}^{K} \mathbb{1}_{\hat{o}>\alpha,k} \cdot (\mathrm{T_{in,k}} - \mathrm{T}_{comf})^2}{\sum_{k=1}^{K} \mathbb{1}_{\hat{o}>\alpha,k}}} \tag{5–30}$$

The performance index for indoor IAQ is referred to as AVG_{CO_2} in this study (Eq.(5-31)), which is the average value of indoor CO_2 during occupied periods.

$$AVG_{CO_2} = \frac{\sum_{k=1}^{K} \mathbb{1}_{\hat{o}>\alpha,k} \cdot \mathrm{c_{CO_2,in,k}}}{\sum_{k=1}^{K} \mathbb{1}_{\hat{o}>\alpha,k}} \tag{5–31}$$

The energy consumption is used to demonstrate the energy efficiency of different strategies. It encompasses the power needed for cooling supply to the FCU, PAU, and the power consumption of the fan.

$$E_{total} = \frac{Q_c + Q_f}{COP} + E_{fan} \tag{5–32}$$

5.5 Simulation results and Pareto analysis

5.5.1 Results of online stochastic occupancy prediction

Figure 5-12 shows the results of occupancy prediction of the 5 testing days and comparison with the ground true values (red line) and the pre-set schedule (blue dashed line). The real occupant number in the meeting room exhibits highly stochastic occupancy pattern, i.e. the room is unoccupied for most of the time with occasional periods of high occupancy when a large number of people enter and exit within a short duration.

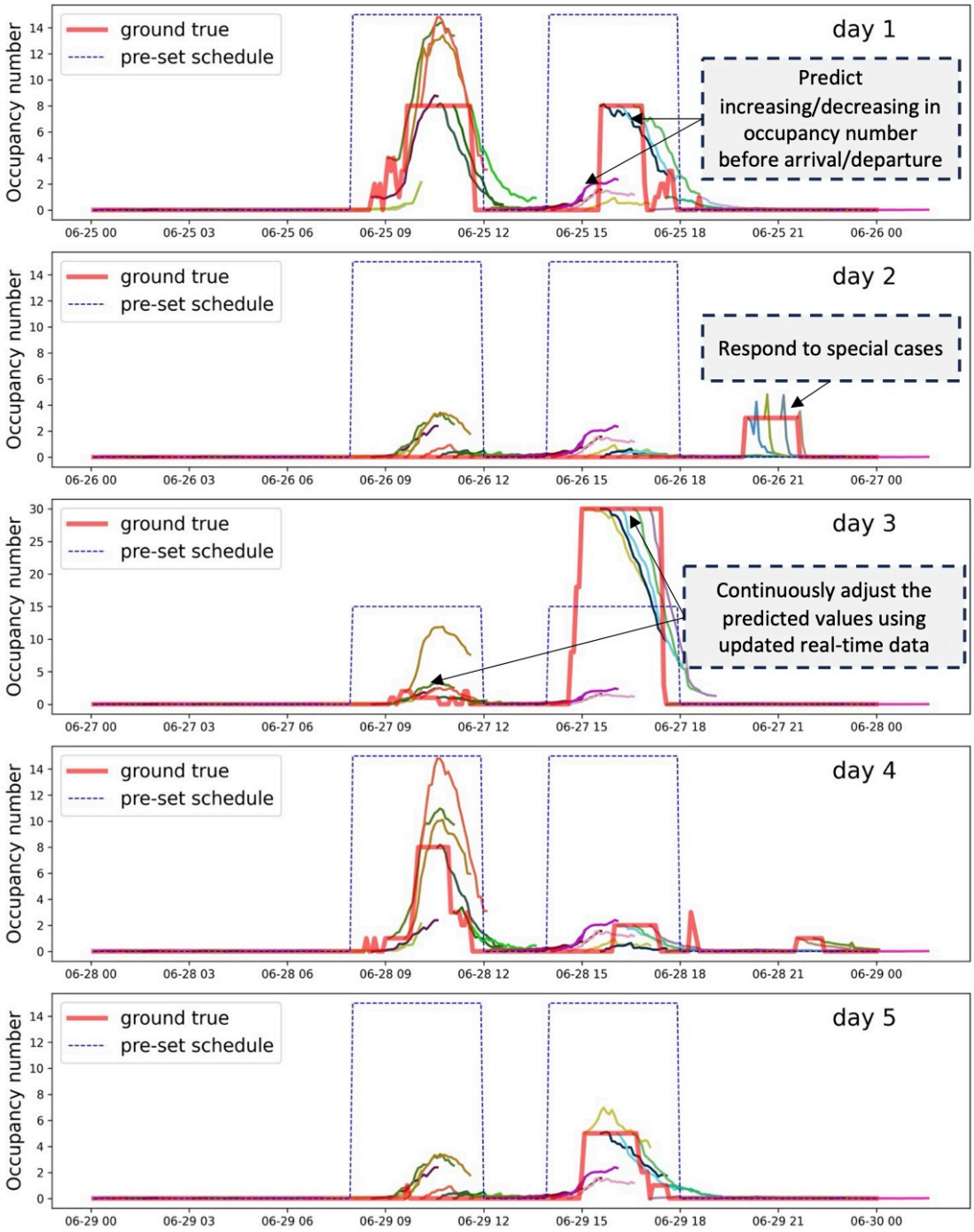


Figure 5-12.IMC-based occupant number prediction results. The red line represents the real measurement; the blue dashed line is the pre-set fixed schedule, which is adopted in the FIX strategy; the curves with various colors stand for the expected 2-hour trajectories of occupant number at different time steps using the real-time measurement as the initial values.

The FIX strategy lacks responsiveness to realistic and dynamic occupancy. Specifically, for most of the time, the FIX strategy estimates a higher number of occupants than the actual values, potentially leading to oversupply of fresh air and energy wastage during unoccupied time. In addition, the FIX strategy is unable to handle the unscheduled situation.

The curves with various colors stand for the expected 2-hour trajectories of occupant number using IMC model at different time steps. The MPC strategy utilizes the stochastic occupancy prediction that combines the stochastic occupancy model and real-time measurement. It can predict increasing/decreasing trend of occupant number before occupants' arrival/departure based on historical patterns, enabling proactive control to handle occupancy disturbances in advance and reduce the risk of discomfort when occupants are expected to arrive in future. It can also address the control issue when meetings are held outside of normal hours (see day 2 and day 4). Though deviations between prediction and real values cannot be avoided, it can continuously adjust the prediction based on real-time updated occupant number data.

5.5.2 Performance comparison under different control strategies

5.5.2.1 Comparison of thermal comfort

Figure 5-13 and Figure 5-14 illustrate the indoor air temperature set-points and the room temperature simulation results. The FIX strategy could maintain the indoor temperature at a comfortable level during working hours. However, continuously maintaining temperature at comfortable temperature during unoccupied hours results in energy wastage. In addition, it fails to condition the room when some urgent meetings happen during non-working hours. For instance, during the meetings on the evenings of day 2 and day 4, the temperature of room reaches 26°C, significantly deviating from the comfortable temperature.

The OBC strategy could response to real-time occupancy, leading to significant energy savings when the room is unoccupied. However, it overlooks the building's thermal inertia and weather condition, potentially resulting in thermal discomfort when

occupants come in the room. As observed in Figure 5-14, on day 1 afternoon, day 3 morning and afternoon, and day 5 afternoon, the indoor temperature reaches nearly 30°C before meetings. Despite immediate adjustments of the indoor temperature setpoint to 24°C upon occupants entering the room, the room maintains an uncomfortably hot environment for more than half an hour.

The proposed MPC strategy achieves an optimal trade-off between energy efficiency and thermal comfort. The MPC strategy implements pre-cooling before occupants' arrival, effectively removing accumulated heat and minimizing thermal discomfort, as shown in Figure 5-14 on day 1 afternoon, day 2 morning and afternoon, mornings and afternoons on day 3 and 5. During pre-cooling, MPC optimizes the indoor temperature setpoint in anticipation of future occupancy, reducing the risk of discomfort. The optimized setpoints during pre-cooling dynamically varies with exogeneous and endogenous factors such as current indoor temperature, outdoor temperature, and solar radiation. For instance, the pre-cooling on day 3 morning is implemented earlier than day 2 morning due to a higher outdoor temperature and solar radiation on day 3 morning. The pre-cooling is not implemented on day 4 afternoon, but the MPC can still swiftly maintain the indoor temperature at a comfortable level after occupants' arrival, since the low outdoor temperatures and solar radiation on day 4 afternoon.

5.5.2.2 Comparison of IAQ

Figure 5-15 and Figure 5-16 illustrate the fresh air flow rate set-points and the room CO_2 simulation results. The FIX strategy, which gives a constant fresh air flow rate during scheduled periods, results in oversupply during unoccupied or low-occupancy periods. It also fails to ensure a good IAQ during non-working hours. As shown in Figure 5-16, on day 2 evening, CO_2 concentration of the FIX strategy is significantly higher than the threshold. The OBC strategy adjust fresh air flow rate responsively to real-time occupant number and per-person flow rate (R_p).

Figure 5-16 shows the simulation results of CO_2 with R_p set at 2.5 L/s/person. The CO_2 concentration of the OBC strategy is also higher than the threshold. The MPC strategy

on the other hand can ensure the CO_2 concentration within 800ppm for the most of time. Even in the most unfavorable condition during group meeting period (on day 3 afternoon) when the room accommodate for more than 30 people, the CO_2 concentration of MPC strategy is the lowest among the three strategies.

Figure 5-17 and Figure 5-18 show the fresh air flow rate and CO_2 concentration of OBC strategy with varying R_p from 2.5 to 12.5 L/s/person and the comparion with MPC strategy. It can be observed that as R_p increases, the CO_2 concentration gradually decreases. However, using a fixed R_p can lead to over- or under-supply of fresh air. For instance, as shown in Figure 5-18, when R_p adopts 10 L/s/person, the CO_2 concentration exceeds 800 ppm on day 1 morning, while below 800 ppm on day 2 evening, suggesting under- and over-supply respectively. In contrast, the MPC strategy could dynamically adjust fresh air flow rate achieving optimal trade-off between CO_2 concentration and energy consumption ensuring CO_2 concentration below the threshold of 800 ppm. Notice that over estimation of occupancy in MPC can sometimes lead to excessive supply of fresh air, as seen during the morning meetings on day 1 and day 3, when the fresh air flow rates are maintained at the maximum for a short period. However, the predictions are continuously refined, allowing for prompt refinement of fresh air flow rate.

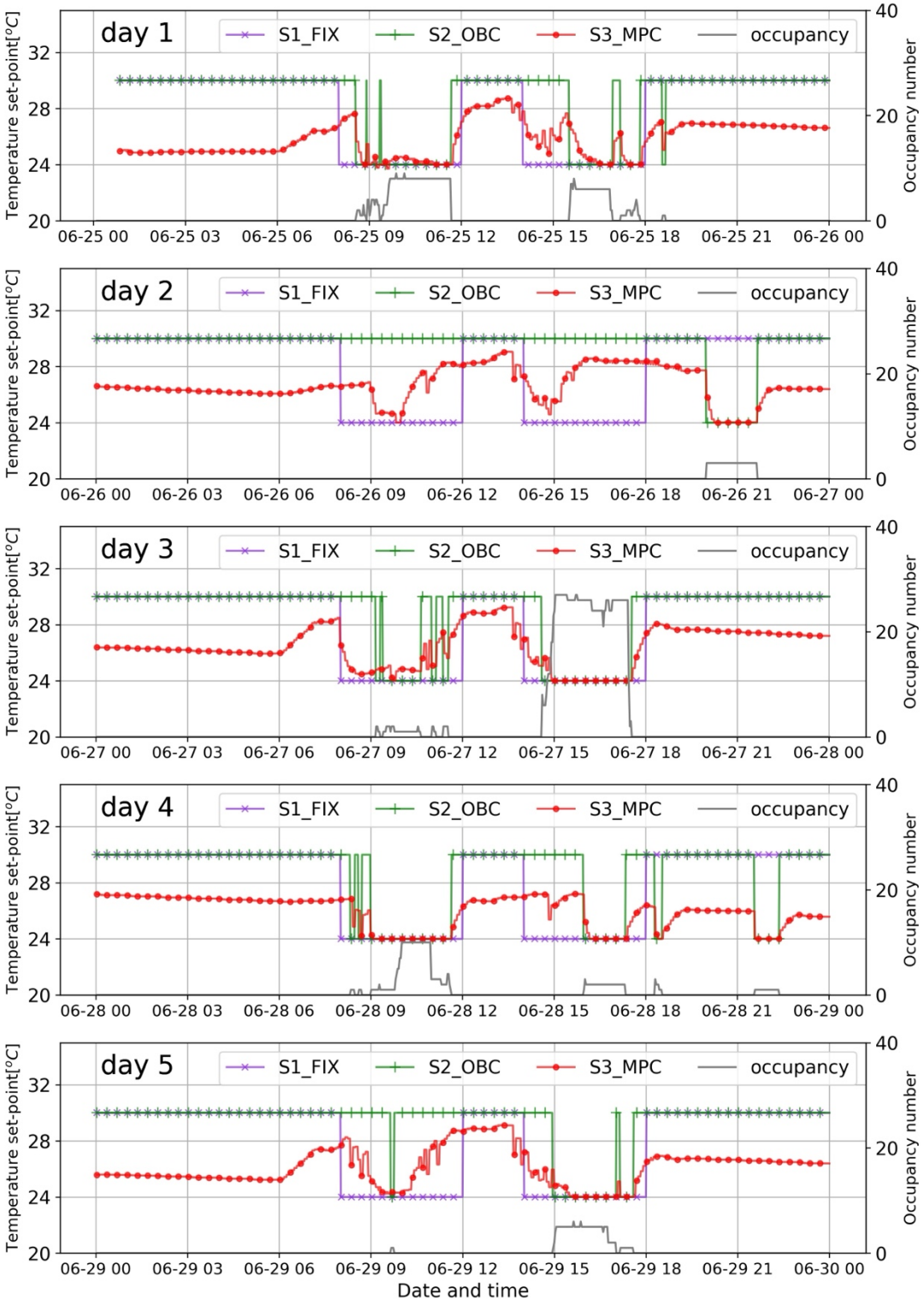


Figure 5-13.Comparison of indoor air temperature set-point of the three control strategies

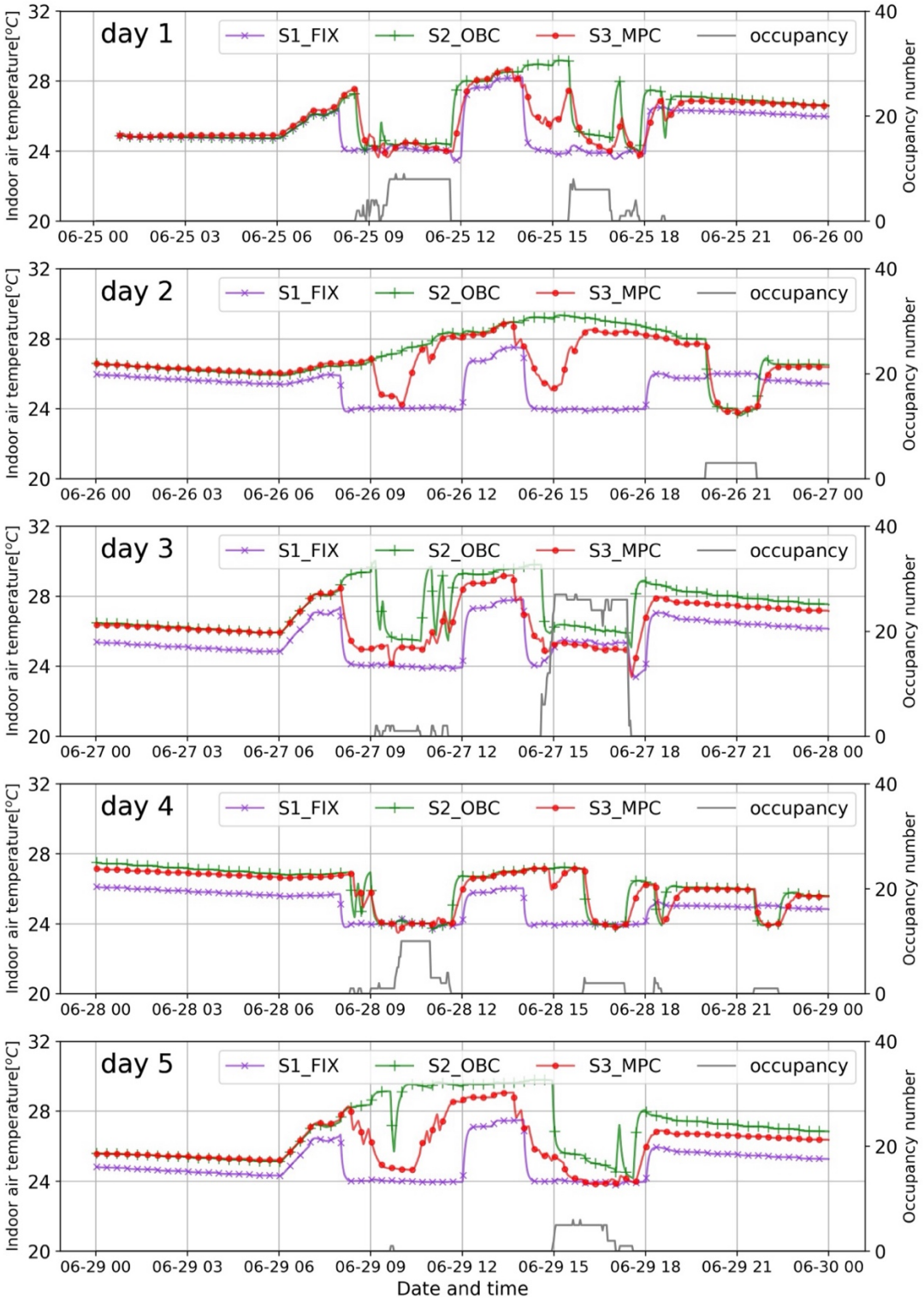


Figure 5-14. Comparison of indoor air temperature simulated TRNSYS

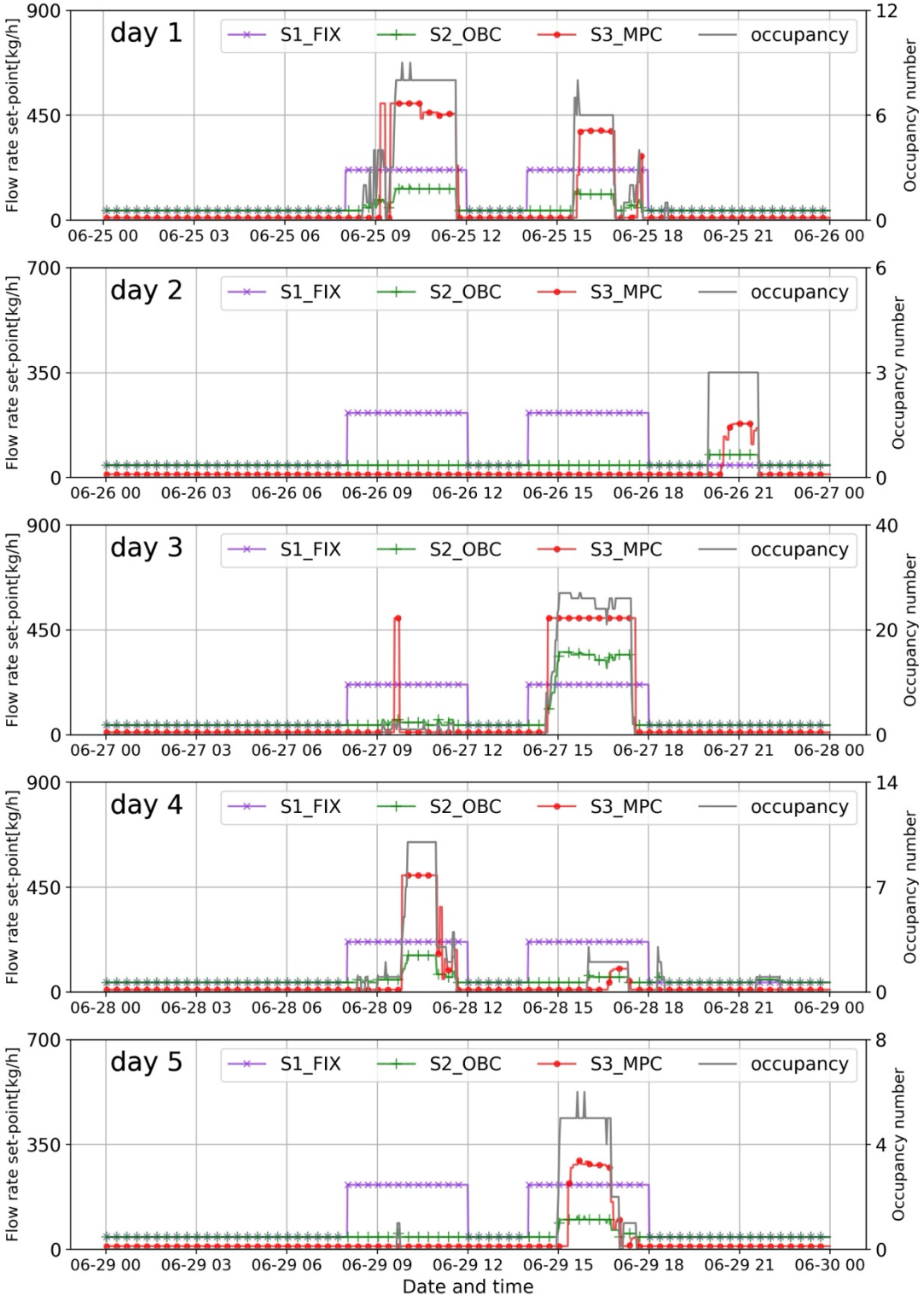


Figure 5-15. Comparison of fresh air flow rate set-point of the three control strategies

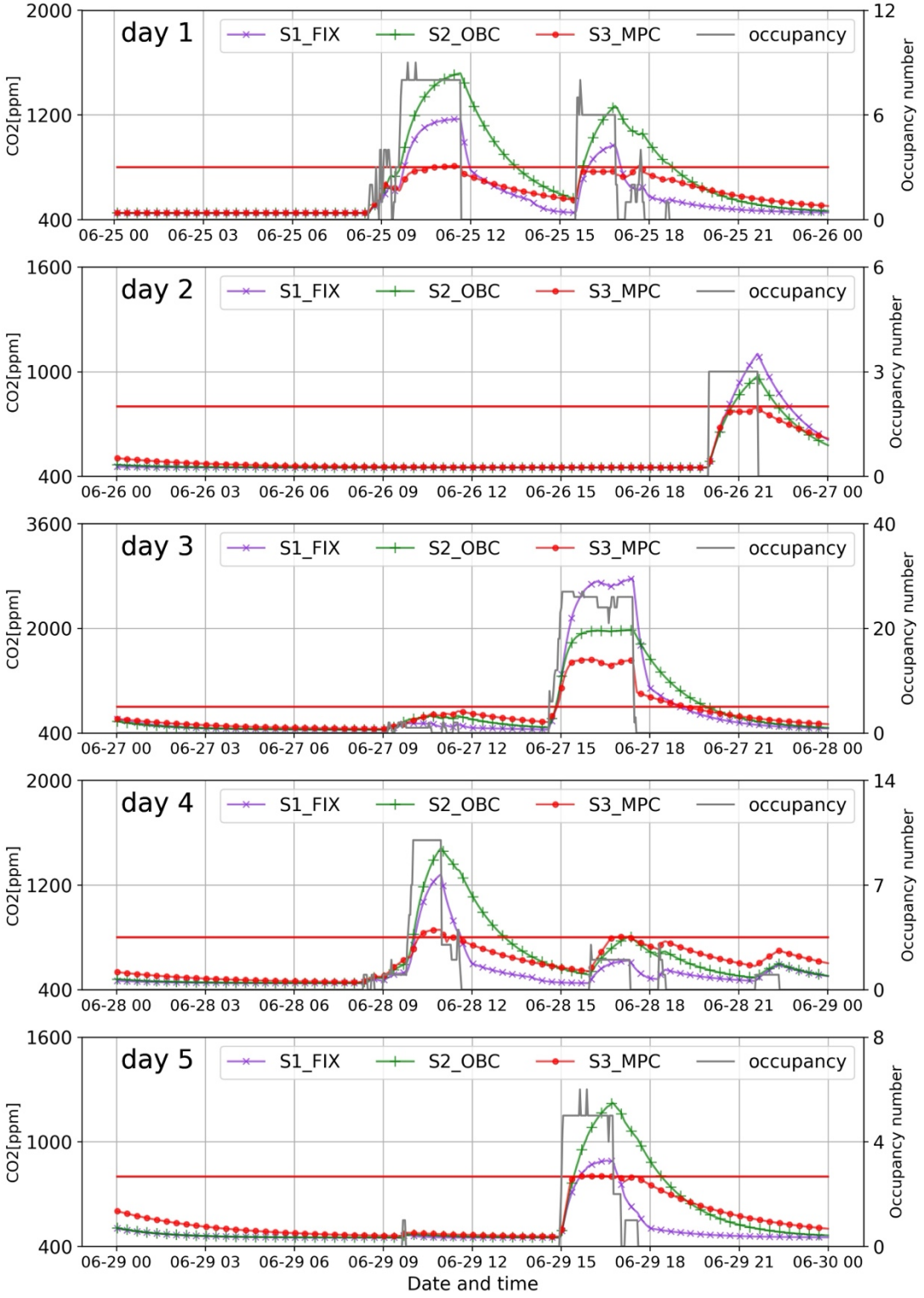


Figure 5-16.Comparison of CO_2 concentration simulated in TRNSYS (the red horizontal line represents the limit of 800 ppm)

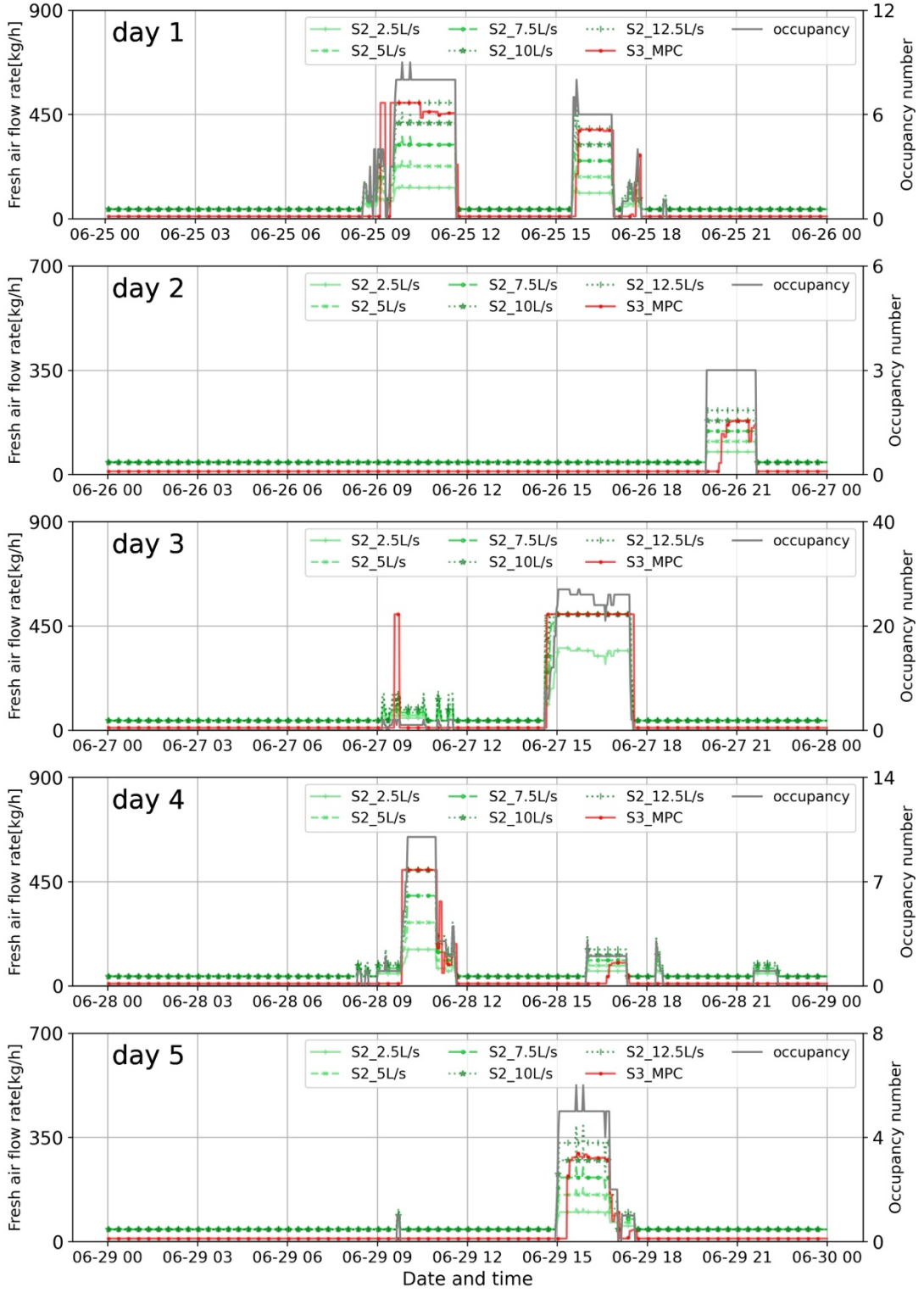


Figure 5-17. Comparison of fresh air flow rate of the OBC strategy with varying fresh air flow rate per person (R_p) from 2.5L/s/person to 12.5L/s/person

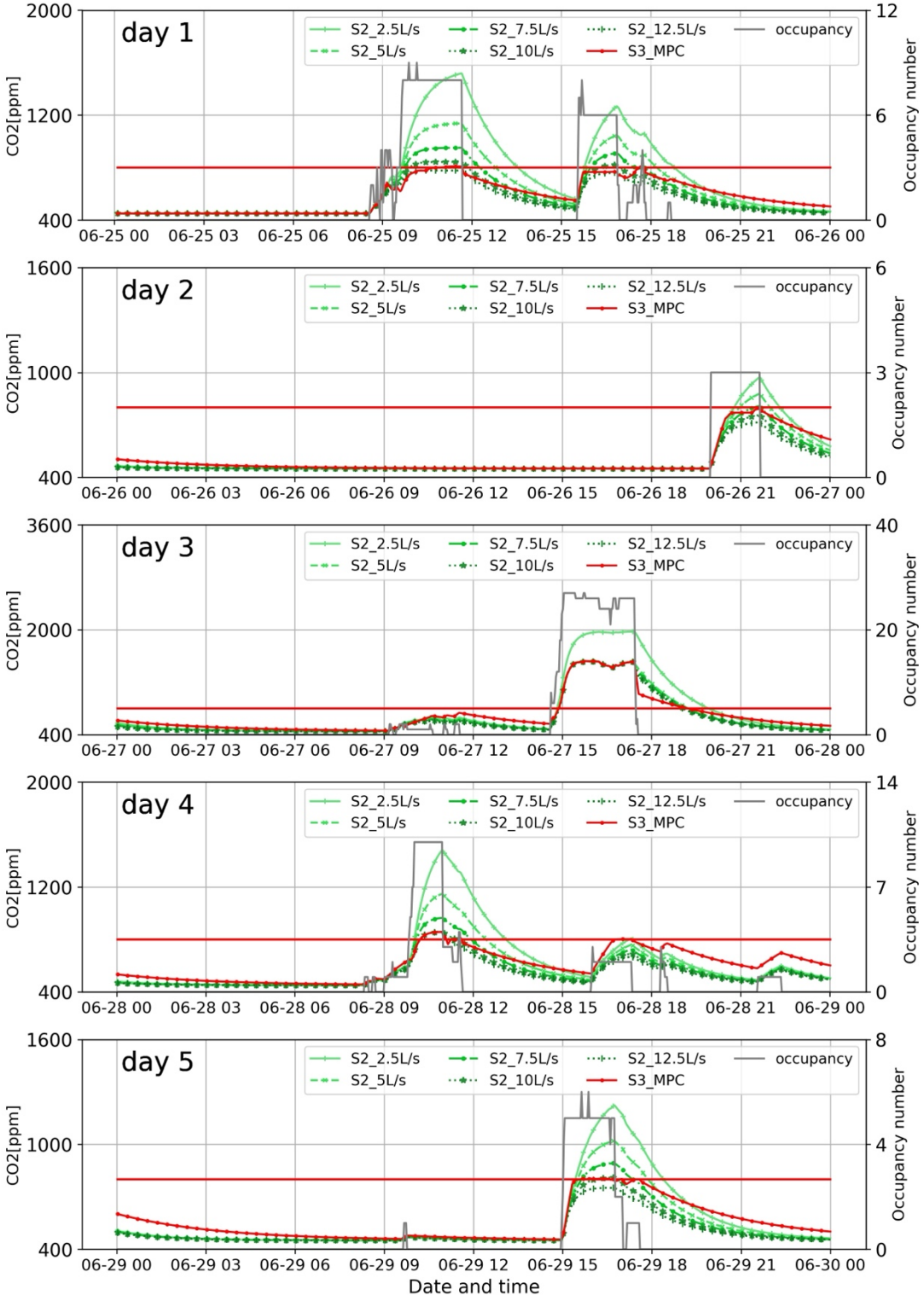


Figure 5-18. Comparison of CO_2 concentration of the OBC strategy with varying fresh air flow rate per person (R_p) from 2.5L/s/person to 12.5L/s/person

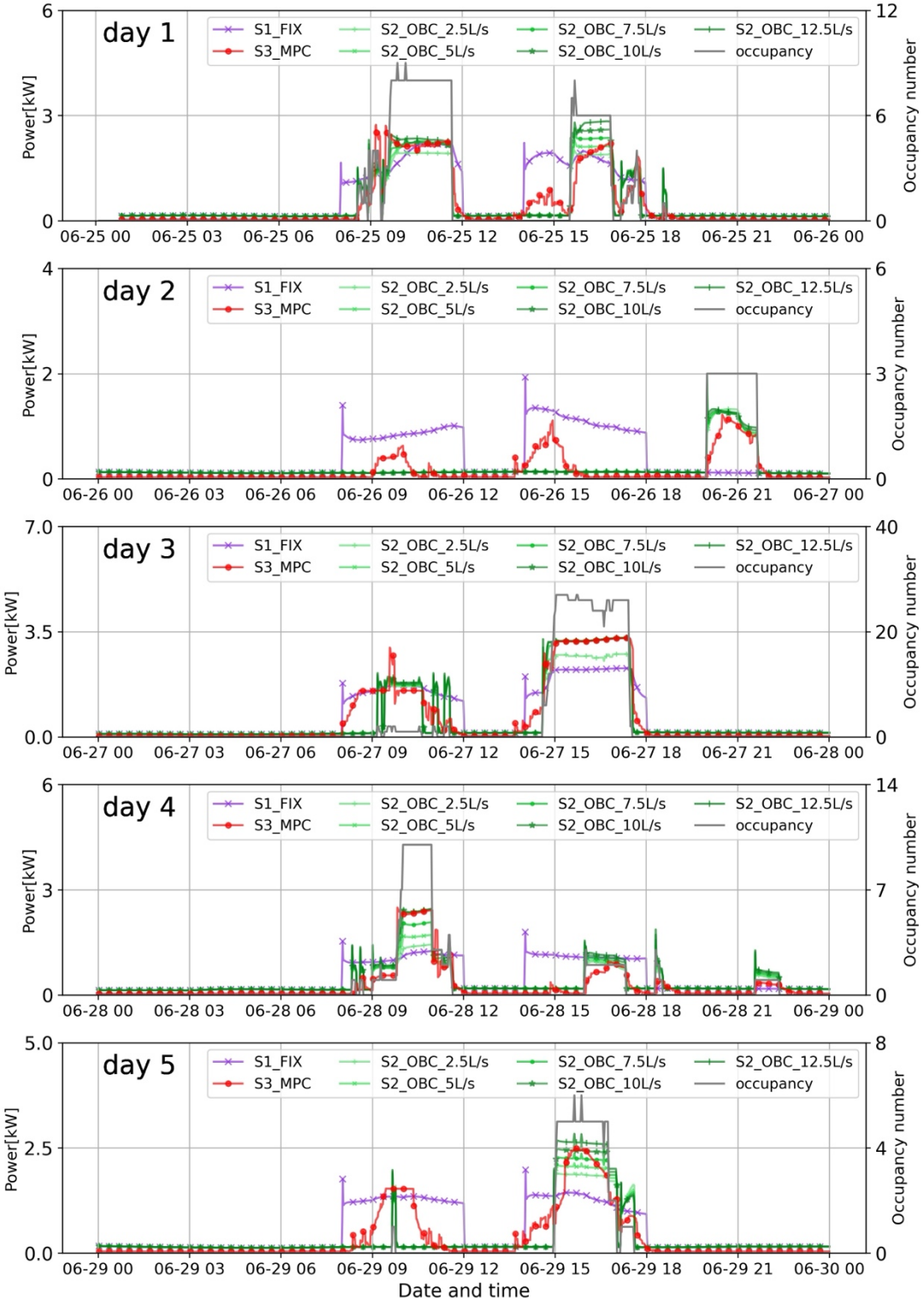


Figure 5-19. Comparison of energy consumption of the three control strategies

5.5.2.3 Comparison of energy consumption

Figure 5-19 compares the total energy consumption for the three strategies over five

testing days. It can be observed that FIX strategy consumes a significant amount of energy when there are no occupants presented in the room, while OBC strategy provides cooling only when there are occupants in the room. The proposed MPC strategy offers optimized control, achieving an optimal trade-off between energy consumption and comfort as well as IAQ. It provides the pre-cooling strategies and optimally adjusts the fresh air flow rate to enhance both thermal comfort and IAQ while ensure energy efficiency.

5.5.2.4 Statistical analysis of the results

Table 5-6 presents the statistical results of the three control strategies. The FIX strategy achieves the lowest $RMSE_T$ among all strategies. However, this comes at the cost of significant energy consumption, total 64.6 kWh, the highest among all strategies, with a lot of energy wasted during unoccupied periods. The R_p for the FIX strategy is set as 2.5L/s, yielding an AVG_{CO_2} of 979.8 ppm. It can be anticipated that increasing the R_p for FIX strategy would reduce the AVG_{CO_2} be reduced, but energy consumption would increase even more.

Table 5-6 Statistical results of the three control strategies

Control strategy	R_p	Thermal comfort			IAQ			Energy		
		$\boldsymbol{RMSE_T}$	IMP*	RIMP*	$\boldsymbol{AVG_{CO_2}}$	IMP*	RIMP*	$\boldsymbol{E_{total}}$	IMP*	RIMP*
		°C	°C	%	ppm	ppm	%	kWh	kWh	%
S1_FIX	-	0.82	-0.18	-22%	980	**168**	**17%**	64.6	**20.7**	**32%**
S2_OBC	2.5L/s	1.43	**0.43**	**30%**	1025	**213**	**21%**	44.2	0.2	0.5%
	5L/s	1.26	0.26	21%	882	70	8%	47.4	3.5	7%
	7.5L/s	1.2	0.2	17%	827	15	2%	48.9	5	10%
	10L/s	1.16	0.16	14%	790	-21	-3%	50.4	6.5	13%
	12.5L/s	1.13	**0.13**	**12%**	768	-43	-6%	51.8	**7.9**	**15%**
S3_MPC	-	**1**	-	-	**812**	-	-	**43.9**	-	-

* IMP and RIMP stand for improvement and relative improvement of MPC compared with two baseline strategies, respectively

In contrast, the OBC strategy demonstrates relatively low energy consumption.

However, it has the highest $RMSE_T$ and AVG_{CO_2} among all strategies, indicating the poorest thermal comfort and IAQ. Table 6 also presents the performance of the Strategy 2 with varying R_p. As R_p increases, both thermal comfort and IAQ improve, although energy consumption rises simultaneously. When R_p reaches 7.5 L/s, the AVG_{CO_2} of OBC strategy (827ppm) becomes comparable to that of MPC strategy (812ppm). Nevertheless, the $RMSE_T$ remains higher than that of MPC strategy, indicating worse thermal comfort, alongside an additional 10.2% in energy consumption compared to MPC strategy. Further increases in R_p result in the AVG_{CO_2} dropping below 800 ppm, but energy consumption continues to rise, indicating an overall oversupply of fresh air. The MPC strategy can achieve an optimal trade-off among the conflicting objectives of energy consumption, indoor thermal comfort and IAQ. It offers a 32% energy saving and a 17.2% improvement in IAQ compared to the FIX strategy. MPC strategy uses nearly the same amount of electricity as OBC strategy at 2.5L/s of R_p, while achieving improvements of 30.1% in thermal comfort and 20.8% in IAQ. As R_p increases, the $RMSE_T$ and AVG_{CO_2} of OBC strategy approach those of MPC strategy, but the energy performance of OBC strategy deteriorates.

5.5.3 Pareto analysis on the effects of weights combination

The cost function in Eq. (5-21) is a linear combination of multiple objectives with different weights assigned to them. The selection of these weights is non-trivial and has a significant impact on the optimal control performances. In this section, we explore the effects of weight combinations on the three control performances (i.e. $RMSE_T$, AVG_{CO_2}, E_{total}) using Pareto analysis by modifying two weighting factors of thermal comfort and IAQ (ρ_{TC}, ρ_{IAQ}), respectively. Table 5-7 shows the 11 cases with different weighting factors. Note that the weighting factor of energy consumption (ρ_E) remains at 0.1. In Cases 1-6, ρ_{TC} varies from 0.1 to 4, while keeping ρ_{IAQ} constant at 0.1. In Cases 7-10, 4, 11, ρ_{IAQ} varies from 1e-4 to 1, while ρ_{TC} remains at 1. The statistical results of two baseline strategies, i.e. FIX and OBC are also included in the Pareto analysis for a holistic comparison.

Table 5-7 Cases of different combination of weights in the cost function

Case	1	2	3	4	5	6	7	8	9	10	11
ρ_{TC}	0.1	0.3	0.5	1	2	4	1	1	1	1	1
ρ_{IAQ}	0.1	0.1	0.1	0.1	0.1	0.1	1e-4	3e-4	1e-3	1e-2	1

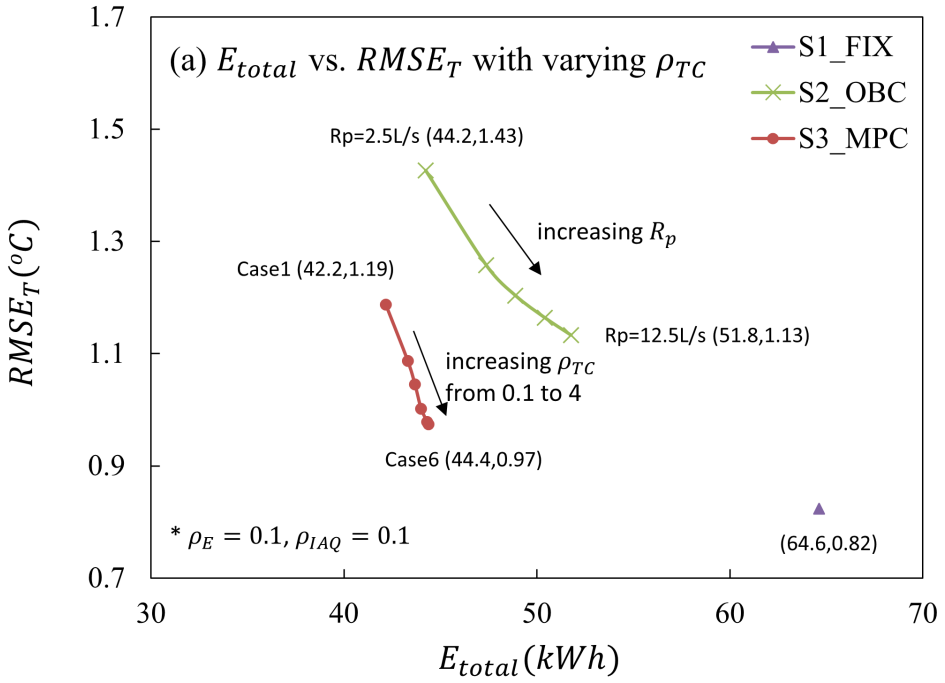
5.5.3.1 Impacts of $\boldsymbol{\rho_{TC}}$ on the control performances

Figure 5-20(a) illustrates the trade-offs between thermal comfort and energy consumption with varying ρ_{TC}. As ρ_{TC} increases, which reflects increasing importance of thermal comfort, the $RMSE_T$ of MPC decreases while the E_{total} increases. The results clearly demonstrate the conflict between maximizing thermal comfort and minimizing energy consumption. The proposed MPC strategy achieves superior performance in balancing the two objectives compared to the OBC strategy, which maintains a same level of thermal comfort using significantly higher amount of energy. The FIX strategy, though ensures good thermal performance, exhibits even higher energy usage.

Figure 5-20 (b) illustrates the trade-offs between IAQ and energy consumption with varying ρ_{TC}. As ρ_{TC} increases, the AVG_{CO_2} of MPC maintains a consistent value of approximately 812 ppm, which is close to the threshold 800 ppm. This indicates that the MPC strategy can dynamically adjust the fresh air flow rates avoiding both over- and under-supply of fresh air. In contrast, the OBC strategy with fixed setup of R_p values results in under-supply when R_p.is low, and over-supply when R_p.is high. FIX strategy is the worst among the three strategies with the highest E_{total} (64.6 kWh) and rather high AVG_{CO_2} (980ppm), due to the oversupply during unoccupied periods and undersupply when the occupants show up during unscheduled periods.

5.5.3.2 Impacts of $\boldsymbol{\rho_{IAQ}}$ on the control performances

Figure 5-21(a) illustrates the trade-offs between thermal comfort and energy consumption with varying ρ_{IAQ}. As ρ_{IAQ} increases, the $RMSE_T$ of MPC gradually reduces. This improvement in thermal comfort is attributed to the cooling effect provided by the fresh air, which is conditioned to 18°C in PAU.



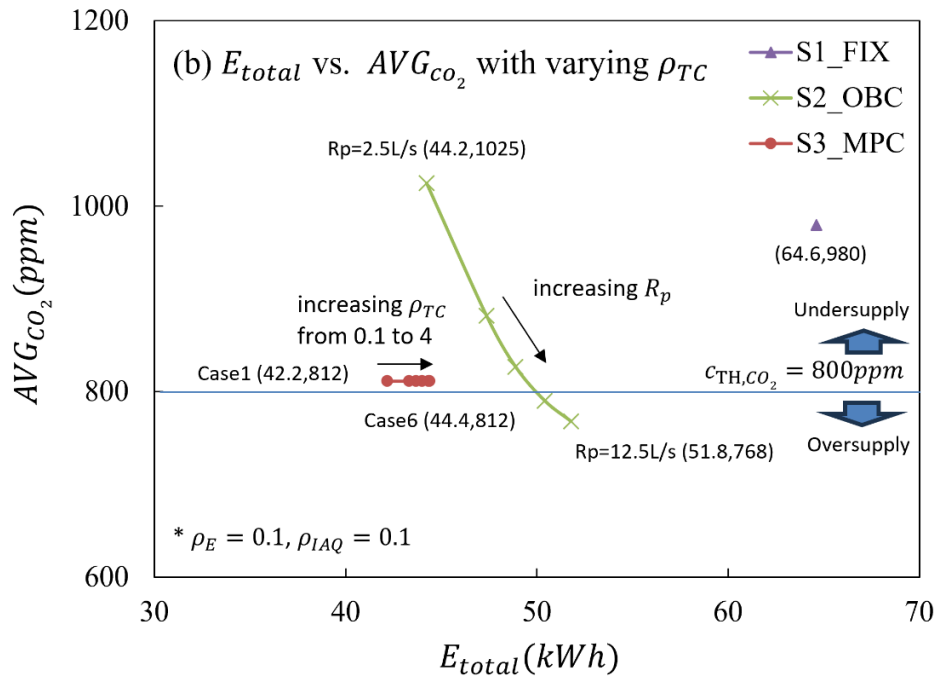


Figure 5-20. The impact of $\boldsymbol{\rho_{TC}}$ *values on the control performances. (a) shows the energy performance* $\boldsymbol{E_{total}}$ *against thermal comfort performance* $\boldsymbol{RMSE_T}$ *under FIX strategy, OBC strategy with varying* R_p *and MPC strategy with varying* $\boldsymbol{\rho_{TC}}$*; (b) shows the energy performance* $\boldsymbol{E_{total}}$ *against IAQ performance* $\boldsymbol{AVG_{CO_2}}$ *under the three strategies. Numbers enclosed in parentheses indicate the coordinates of points.*

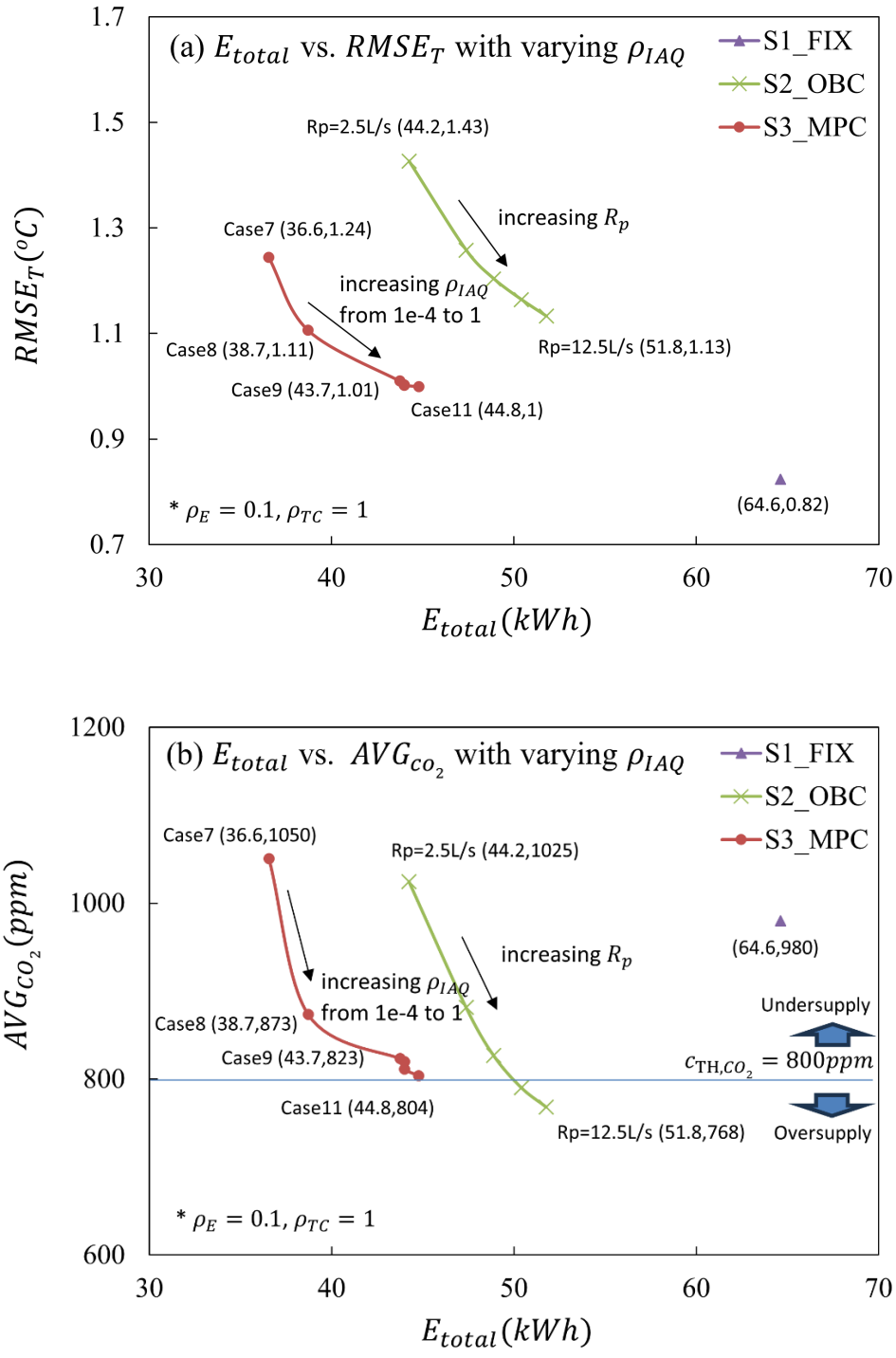


Figure 5-21. The impact of $\boldsymbol{\rho_{IAQ}}$ *values on the control performances. (a) shows the energy performance* $\boldsymbol{E_{total}}$ *against thermal comfort performance* $\boldsymbol{RMSE_T}$ *under FIX strategy, OBC strategy with varying* R_p *and MPC strategy with varying* $\boldsymbol{\rho_{IAQ}}$*; (b) shows the energy performance* $\boldsymbol{E_{total}}$ *against IAQ performance* $\boldsymbol{AVG_{CO_2}}$ *under the three strategies.*

This effect can be also observed in OBC with varying R_p. The MPC strategy could achieve better energy performance with same level of thermal comfort compared with OBC strategy. Figure 5-21(b) illustrates the trade-offs between IAQ and energy consumption with varying ρ_{IAQ}. The AVG_{CO_2} of MPC rapidly decreases as ρ_{IAQ} increases from 1e-4 to 1e-3 and gradually approaches to the threshold as ρ_{IAQ} increases from 1e-3 to 1. The MPC could achieve best energy performance with the same level of IAQ compared with OBC and FIX strategies.

5.6 Discussion and limitations

5.6.1 Applicability and scalability in multiple zones

The proposed stochastic occupancy-integrated MPC framework demonstrates promising results for single-zone systems, but its extension to large building complexes with multiple zones and centralized HVAC systems introduces additional challenges and opportunities. In this subsection we discuss key considerations and potential strategies for scaling up the framework.

In large building complexes, each zone (e.g., offices, conference rooms, lobbies) has unique occupancy patterns, thermal characteristics, and ventilation requirements. Extending the MPC requires accounting for the interactions between these zones. Zones often interact thermally (e.g., heat transfer through walls) and through the shared airflow. Some studies have explored modelling the multi-zone building dynamics. However, it is acknowledged that accurately identifying the large amount of parameters of the large thermal and airflow network is still a critical challenge. Complexity of solving the optimization problem of MPC with such large networks also grows dramatically with the increase of amount of zones.

Hierarchical MPC structure could be a potential solution. Independent controllers for each zone (lower-level MPCs) can be established to ensure computational scalability. Each controller can observe its local information and exchange information with adjacent zones to build localized building dynamic models. The lower-level MPCs

manage individual zone-level objectives like thermal comfort and IAQ, based on individual occupancy predictions. The upper-level MPC can be established in central chiller plant to optimize global objectives such as total energy consumption, total peak load and overall HVAC efficiency. The hierarchical MPC structure can thus dynamically allocate energy resources (e.g., chilled water, airflow) to zones based on their predicted occupancy and environmental conditions under physical and operational constraints and also coordinate zone-level demands to avoid simultaneous peaks and improve overall system efficiency.

5.6.2 Integration of booking system

It is important to note when the booking information is available, the OBC performance may have some improvement. However, using the OBC method still requires determining a pre-cooling time in advance. Predefining a fixed pre-cooling time may still result in insufficient or excessive cooling. A longer pre-cooling time result in excessive cooling, while a shorter pre-cooling time may still result in insufficient pre-cooling and deteriorate thermal comfort. Therefore, the fundamental challenge of conflict goals between thermal comfort and energy consumption is still not addressed.

In contrast, MPC can dynamically optimize the pre-cooling time based on indoor and outdoor information after receiving the booking details. Additionally, it should be noted that even if booking data is available, there may still be discrepancies between the booking information and the actual number of attendees or meeting times. The likelihood of energy waste or discomfort when using the OBC method cannot be avoided. However, booking data can further refine the occupancy prediction model by embedding the booking information as an external parameter, thereby reducing prediction errors. Therefore, compared to OBC, the stochastic occupancy-integrated MPC with booking system still holds significant advantages.

5.6.3 Limitations

5.6.3.1 Extension in diverse condition

The current research is limited to meeting rooms in office buildings, equipped only with cooling and ventilation systems, and is conducted in the tropical climate of Hong Kong.

This study will further comprehensively evaluate the applicability of the stochastic occupancy-integrated MPC across different building types, building systems, and climate conditions. In the future, the focus will expand to explore different types of buildings, including residential buildings, commercial buildings, and school buildings. We will collect more occupancy data from diverse space types such as classrooms, lecture halls, and lobbies and study the entirely different occupant characteristics and building dynamic properties. Additionally, the integration of stochastic occupancy prediction in the control of various building systems, such as lighting and shading systems, is also worth exploring, since they are also closely related to occupancy patterns. The study will also investigate other climate types, such as cold regions and areas with hot summers and cold winters.

5.6.3.2 Evaluation using comprehensive indoor environment quality factors

Currently, indoor environmental quality is evaluated using temperature as the indicator of thermal comfort and CO_2 concentration as the indicator of air quality, which has certain limitations. In the future, the incorporation of additional environmental parameters will aim to achieve a more comprehensive and complete evaluation of indoor environments, as well as to develop more robust indoor environment prediction models. For example, thermal comfort evaluation could integrate indoor temperature and relative humidity, while IAQ evaluation could consider a combination of parameters such as CO_2, TVOCs, and PM2.5.

5.6.3.3 Handling variable occupant activities and metabolic rates

In this study, we focus on office building where activity levels are typically within the light to moderate range, as most occupants are engaged in activities such as working at desks or attending meetings. As such, the variability in activity levels is not as drastic in these settings. For this reason, we adopt the CO_2 and heat gain rates per person presented in standard and commonly used in prior research, which represent the majority activity levels of occupants in the typical office environments.

However, during certain special occasions, the indoor space may not be used in its standard manner or for its intended function, which could result in different occupant activity levels and different metabolic rates. Addressing such scenarios would further enhance the robustness and applicability of our approach. In future work, we plan to explore extensions to handle variable occupant activities and metabolic rates to better account for these exceptional cases and improve the generalizability of our findings.

5.7 Summary

This chapter presents a novel stochastic occupancy-integrated MPC scheme for built environment control of spaces characterized by highly stochastic and intermittent occupancy patterns. The proposed MPC strategy effectively balances energy efficiency, thermal comfort, and IAQ through dynamic optimization based on predicted occupant number and presence. It demonstrated superior performance compared to traditional fixed-schedule and occupancy-based control strategies under complex and uncertain occupancy conditions, achieving significant reductions in energy consumption and improvements in IAQ while enhancing thermal comfort. Simulation results demonstrate that the proposed MPC achieves 32% energy savings and 17% IAQ improvement compared to the FIX strategy, and 30% thermal comfort improvement and 20% IAQ improvement with the same energy consumption compared to OBC. The robustness of the MPC approach against stochastic occupancy pattern and its ability to adapt to real-world challenges highlight its potential for advancing occupant-centric and energy-efficient building control systems.

The proposed strategies have undergone testing on a simulation platform. As the next step, we further validate the effectiveness of this strategy in a real system. Through the testing phase in the actual system, we seek to evaluate the practical performance and effectiveness of the MPC control strategy. The data and insights gathered from this experiment contribute to refining and improving the control strategy, ultimately leading to more efficient and intelligent temperature control in building environments.

6 Field implementation of IoT-enabled Occupant-in-the-loop MPC in a living lab

This chapter presents the implementation of the IoT-enabled Occupant-in-the-loop (OITL) MPC strategy based on the control scheme developed in Chapter 5 in a real-world office living lab with active occupancy participation. Real-time occupancy data from camera-based IoT sensors and personalized feedback from QR code-based voting systems are integrated in the OITL MPC to dynamically adjust control setpoints, which are applied to the HVAC systems through the IoT infrastructure, enabling a closed-loop, adaptive, and occupant-centric control system.

6.1 IoT-enabled architecture for built environment management

The proposed IoT-enabled implementation architecture is illustrated in Figure 6-1. This architecture enables high-level supervisory control of optimized setpoints through an IoT framework, eliminating the need for modifications to existing low-level controllers and legacy BAS during IoT retrofits. The IoT-enabled implementation architecture follows a classic three-layer structure, consisting of the data acquisition and field control layer, data transmission layer, and edge/cloud-based application layer, with an additional physical world layer. A key focus of the proposed architecture is the collection and modeling of occupancy data, as well as the implementation of occupancy-centric control strategies to enhance both energy efficiency and occupant comfort.

(1) Physical World Layer

This layer represents the various physical entities and dynamic processes within the system. These include the target rooms where environmental conditions are regulated, the HVAC systems responsible for temperature and ventilation control, the outdoor environment that introduces external thermal influences, and the occupants whose presence and activity significantly impact indoor conditions. The

interactions among these entities involve processes such as cooling and fresh air supply from HVAC systems, passive thermal storage within the room's thermal mass, and disturbances from weather conditions and occupant heat gains. These elements are fundamental to the data-driven modeling and model predictive optimization, allowing the system to dynamically adjust to real-world conditions.

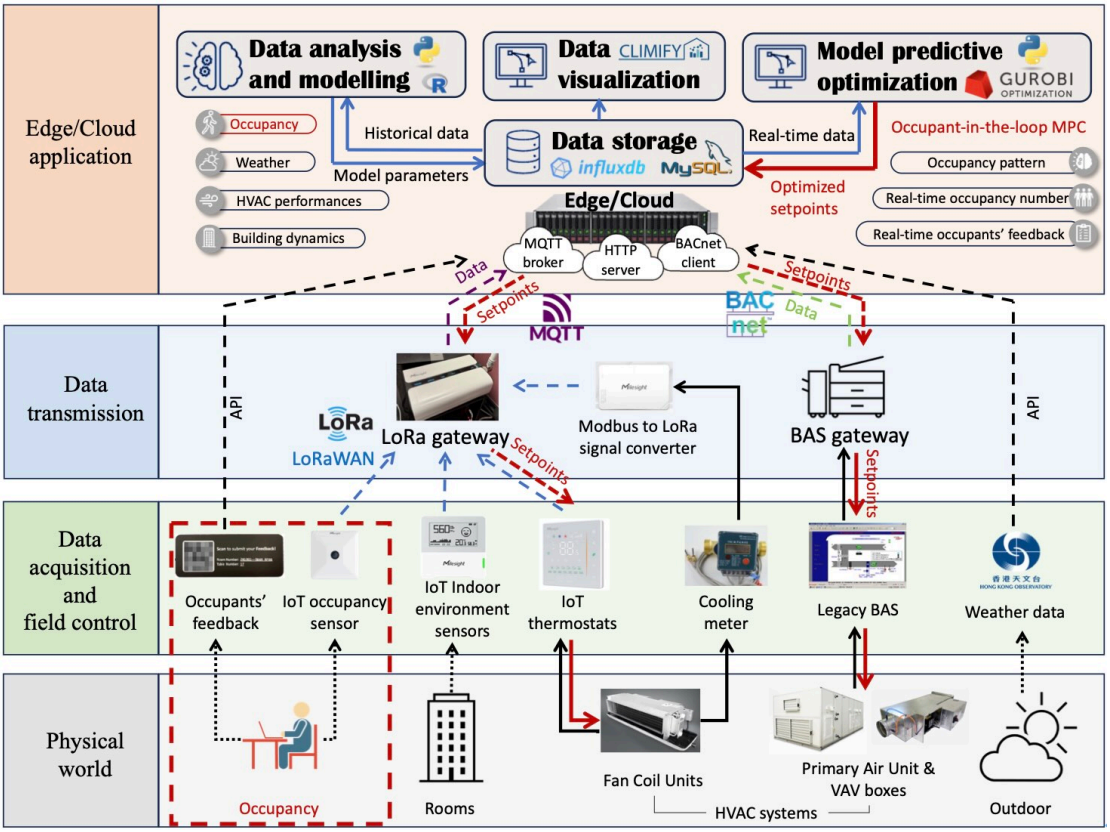


Figure 6-1. IoT-enabled architecture of the OILT MPC.

(2) Data acquisition and field control layer

This layer serves two primary functions. First, it collects real-time data from physical entities through IoT sensors, IoT thermostats, meters, legacy BAS, and QR codes. Second, it implements high-level supervisory control setpoints by transmitting optimized parameters to low-level actuators in the HVAC system.

The study employs LoRa-based IoT environmental sensors, which are characterized by low power consumption and flexible deployment, allowing them to be repositioned as needed to ensure accurate monitoring of target room conditions.

Additionally, a LoRa-based IoT thermostat is used to enable wireless reception of optimal temperature setpoints, ensuring seamless HVAC control without requiring modifications to the existing BAS infrastructure. Meters play a crucial role in measuring energy consumption, providing essential data for system modelling and optimization. The legacy BAS serves as an interface for integrating existing control systems, allowing the IoT-enabled architecture to operate without disrupting established building automation processes. Meanwhile, QR codes facilitate occupant feedback collection, enabling users to report their thermal comfort preferences in real time. This feedback mechanism enhances the adaptability of the system, allowing the OILT MPC to dynamically refine control strategies based on real-world user inputs.

(3) Data transmission layer

This layer is responsible for managing real-time data exchange and protocol conversion. It facilitates the transmission of real-time data and control signals through gateways and APIs using various communication protocols, including MQTT, HTTP, and BACnet. The layer also serves as a bridge for protocol conversion, ensuring seamless interoperability between different subsystems. MQTT plays a crucial role in enabling efficient, lightweight communication between the server and the LoRa gateway in IoT environments. Meanwhile, HTTP APIs are used to receive occupancy feedback and weather forecasts, while BACnet enables bidirectional communication between the server and the BAS gateway, allowing for the monitoring and optimal control of HVAC systems integrated with legacy BAS.

(4) Edge/Cloud-based application layer

This layer integrates a variety of software services, including real-time data processing, data storage, data visualization, advanced data analytics, and model predictive optimization. An MQTT broker, BACnet client, and HTTP server function as real-time data drivers, facilitating bidirectional communication across different protocols. Data storage is managed through InfluxDB, which stores time-series data, and MySQL, which retains structured information for further analysis.

Advanced data analytical modeling leverages historical data to develop data-driven models that capture stochastic occupancy patterns (Chapters 3 and 4), building dynamics (Chapter 5), HVAC performance, and weather conditions. The model predictive optimization component integrates real-time measurements with these data-driven models (Chapter 5), allowing the system to optimize temperature and ventilation setpoints dynamically. This approach ensures an optimal balance between thermal comfort, IAQ, and energy efficiency by making proactive adjustments based on predictive insights.

6.2 Occupant-in-the-loop MPC integrating real-time occupants' feedback

The proposed OILT MPC scheme, which integrates real-time occupant numbers, real-time occupant feedback, and a stochastic occupancy model, is illustrated in Figure 6-2. Below is a concise overview of the OILT MPC.

6.2.1 Occupant-in-the-loop MPC scheme

To enable online occupancy prediction, the scheme employs an inhomogeneous Markov chain (IMC)-based stochastic occupancy model that we developed in Chapter 3. This stochastic occupancy model forecasts the expected future trajectory of occupant numbers and presence, using the real-time occupant number as the initial value. A more detailed explanation of the online stochastic occupancy prediction process is provided in Section 5.1.

The occupant number is used to estimate occupancy-related disturbances, including internal heat gains and CO_2 generation. These predicted disturbances, together with weather forecasts for outdoor temperature, and solar radiation, are integrated into SDE-based grey-box building dynamic models. The SDE models for the real implementation are similar with the model developed in Section 5.2. Therefore, they are not repeated in this section. While, the identified parameters are different, since this section uses real measured data for parameter identification. The identified results and model validation

are presented in Appendix B.

The continuous-time SDE models are converted into discrete-time state-space models, which are embedded in the MPC for temperature and CO_2 prediction. Real-time measurements of indoor temperature and CO_2 concentration are used as initial values for these predictions. Additionally, hidden states in the state-space equations, such as wall temperature, are estimated using a Kalman filter, ensuring more accurate predictions. Details on the Kalman filter for hidden state estimation can be found in [65,79].

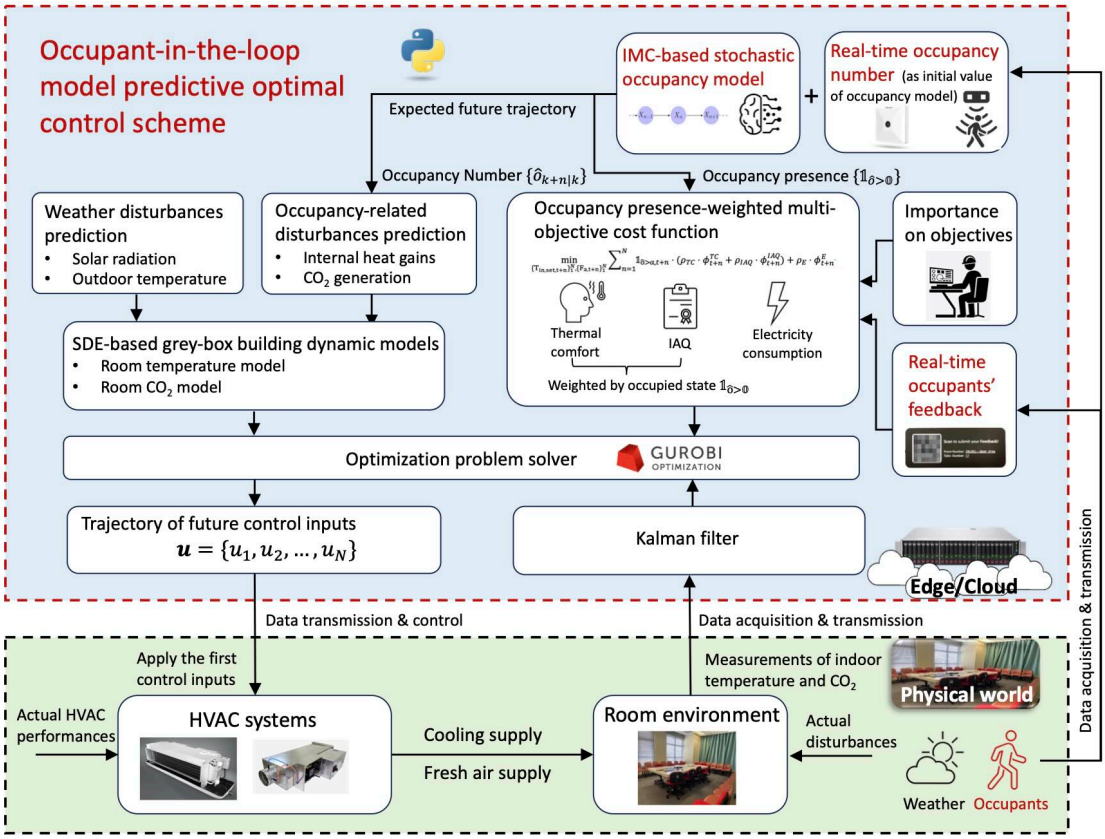


Figure 6-2 The OILT MPC scheme integrating stochastic occupancy model, real-time occupant number and real-time occupants' feedback

The primary objectives of the MPC are to enhance thermal comfort and IAQ while minimizing electricity consumption, all within the constraints of real-world operation. To achieve this, a weighted sum approach is employed to transform the multi-objective

optimization problem into a single-objective function. The parameters ρ_{TC}, ρ_{IAQ}, and ρ_E are used to assign weights to thermal comfort, IAQ, and energy efficiency, respectively. These weightings are configured by the operator to reflect specific operational priorities. For example, if an operator prioritizes thermal comfort and IAQ over energy savings, they can increase the values of ρ_{TC} and ρ_{IAQ}. The detail model structure and solution methods in the real implementation is similar in Section 5.3. The difference part is that the real-time comfort feedback from users is embedded in the thermal comfort objective function, thereby affecting the optimization of temperature setpoints. This part is detailed in Section 6.2.2.

The multi-objective rolling optimization problem is formulated and solved using the Gurobi optimizer [116], generating optimized trajectories for indoor temperature setpoints and fresh air flow rate setpoints. The first set of values from these optimized trajectories serves as the supervisory optimal control setpoints, which are then applied to the real HVAC control systems in the physical environment via the IoT infrastructure. The HVAC system subsequently supplies cooling and fresh air to the target room, dynamically responding to real-time weather conditions and occupancy disturbances.

In this study, the prediction horizon is set as 2 hours, and prediction interval is 5 minutes. The MPC runs every 3 minutes. At the next time step, the optimization is recomputed using updated real-time data. This receding horizon strategy ensures that the control decisions continuously adapt to real-time occupancy, environmental conditions, and occupant feedback, optimizing thermal comfort, IAQ, and energy efficiency dynamically.

6.2.2 Integration of real-time occupants' feedback

In conventional MPC systems [27,65,77,78], the optimization process typically assumes a fixed most comfortable indoor temperature or a predefined comfortable temperature range. These systems continuously send optimized setpoints to the thermostat, ensuring tight control over indoor conditions. However, this approach presents a significant limitation: if occupants manually adjust the temperature setpoint,

their changes are overridden in the next optimization cycle. As a result, occupants lose control over their environment, and the system fails to accommodate personalized comfort preferences.

To address this issue, this study proposes an OILT MPC, designed to integrate real-time occupant feedback into the control process. The implementation of this system is illustrated in Figure 6-3. Occupants can provide personalized comfort feedback by scanning a QR code and completing a brief survey. This feedback is then transmitted to a cloud server via an API, where it is immediately processed. The collected voting data is stored in a database, and the desired reference temperature (T_{comf}) is dynamically updated based on user input. In this study, the specific relationship between occupant voting and reference temperature adjustments is detailed in Table 6-1.

The optimization process incorporates the updated T_{comf} to refine HVAC control setpoints, ensuring that the system continuously adapts to occupant preferences. This approach fosters a responsive and occupant-centric environment within the building. To prevent long-term drift in temperature preferences, the reference temperature is reset to 24°C at 00:00 each day.

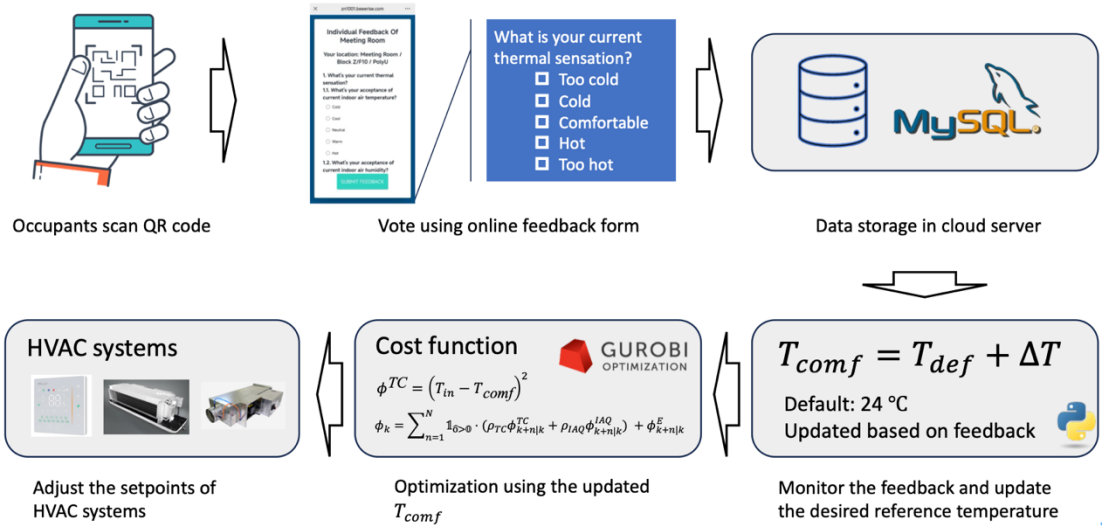


Figure 6-3 QR code-based real-time occupants' feedback system

Table 6-1 Relationship between voting and the adjustment of the reference temperature

Vote	Cold	Cool	Neutral	Warm	Hot
ΔT	+1 ºC	+0.5 ºC	0	-0.5 ºC	-1 ºC

6.3 Experiment study in the living lab

6.3.1 Description of the target office living lab

The proposed IoT-enabled implementation architecture has been successfully deployed in the office living lab as described in Section 3.3 and Section 5.4.1. The layout of the living lab is also presented in Figure 6-4 for convenience. The HVAC systems in the living lab consist of Fan Coil Units (FCUs) combined with Primary Air Units (PAUs) and Variable Air Volume (VAV) systems, as illustrated in Figure 6-4 (2).

The indoor air temperature in these three rooms is regulated by a total of 13 FCUs: one dedicated to the small office, two serving the meeting room, and the remaining ten allocated to the large office. Each FCU is equipped with a thermostat. The coil water valve operates based on an on/off control mechanism, while the fan offers three speed levels. The on/off control logic functions as follows: when the measured temperature exceeds the set temperature by 1°C, the water valve opens, initiating cooling. Conversely, when the measured temperature returns to the setpoint, the valve closes. However, due to the thermal inertia of the cooling coil heat exchanger, residual cooling may continue to be supplied even after valve closure, resulting in a temporary drop in room temperature beyond the setpoint.

Fresh air is initially conditioned in the central PAU room on each floor before being distributed to individual offices. The fresh air flow rate in the living lab is controlled by two VAV boxes—one serving the large office and the other dedicated to the meeting room. Although the small office is physically connected to the fresh air pipeline, it does not receive fresh air from the central PAU/VAV system. The PAU/VAV system is fully integrated with the legacy BAS, allowing fresh air flow rate setpoints to be transmitted

to the VAV boxes via BACnet. Inside each VAV box, a PID process controller continuously compares the setpoint with the measured flow rate, adjusting the damper opening accordingly.

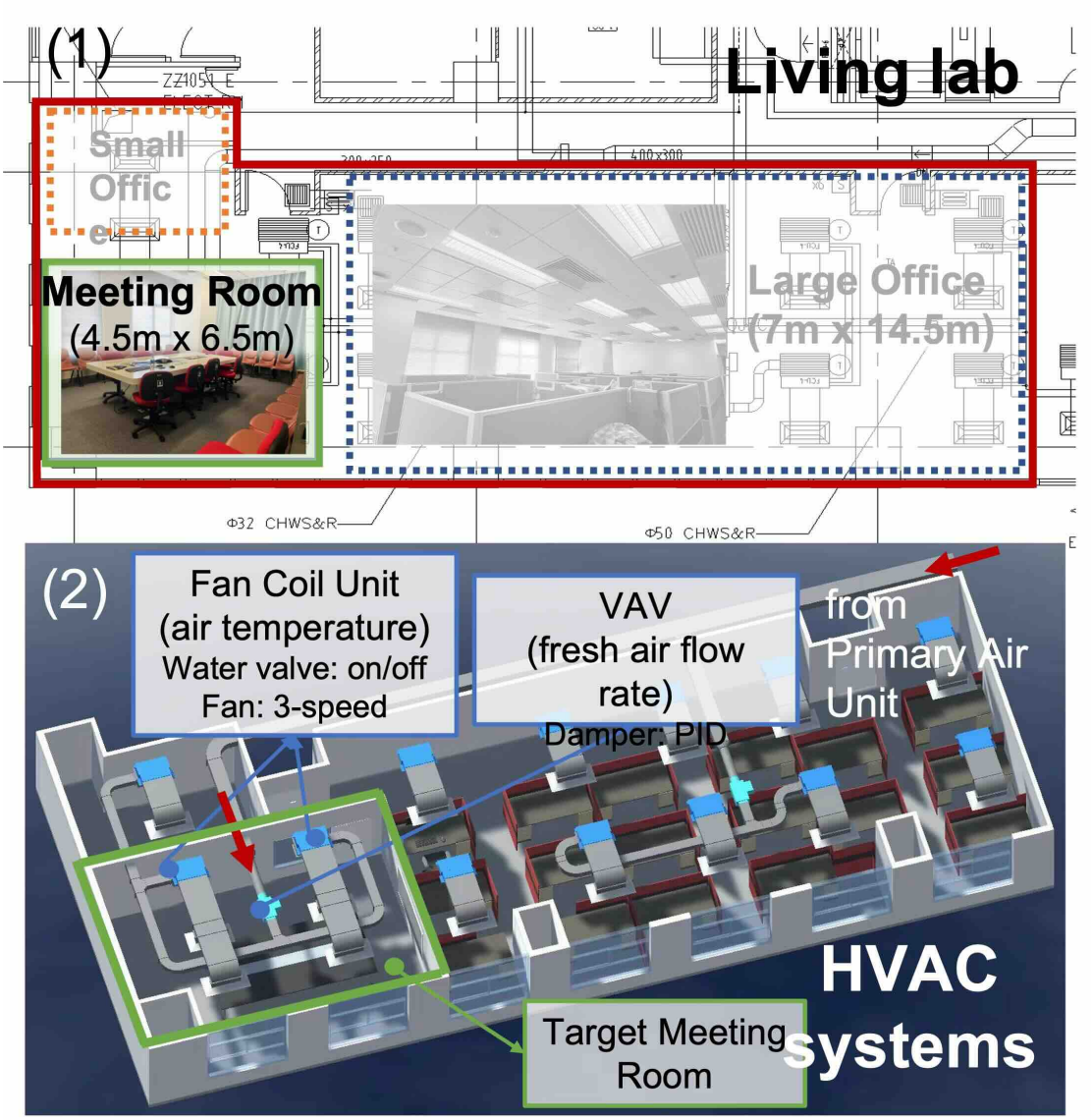


Figure 6-4 Layout and HVAC systems of the office living lab.

6.3.2 IoT retrofit

The IoT retrofit for the living lab was carried out from 2023 to mid-2024. A schematic of the retrofit, along with the installed IoT devices, is presented in Figure 6-5 and listed in Appendix A.

Prior to the retrofit, the measured return air temperature from the ceiling-mounted FCU was used in the local thermostat to regulate the FCU's operation. However, due to vertical temperature stratification, the ceiling temperature often deviated significantly from the actual temperature at the working plane, where occupants are present. To address this issue, the IoT thermostat introduced in the retrofit allows specification of a temperature, termed the "server temperature", for FCU control. In this study, an IoT environmental sensor placed on a desk continuously measures the working plane temperature, which is then transmitted to the IoT thermostat. This refinement ensures that the FCU maintains the working plane temperature at the designated setpoint.

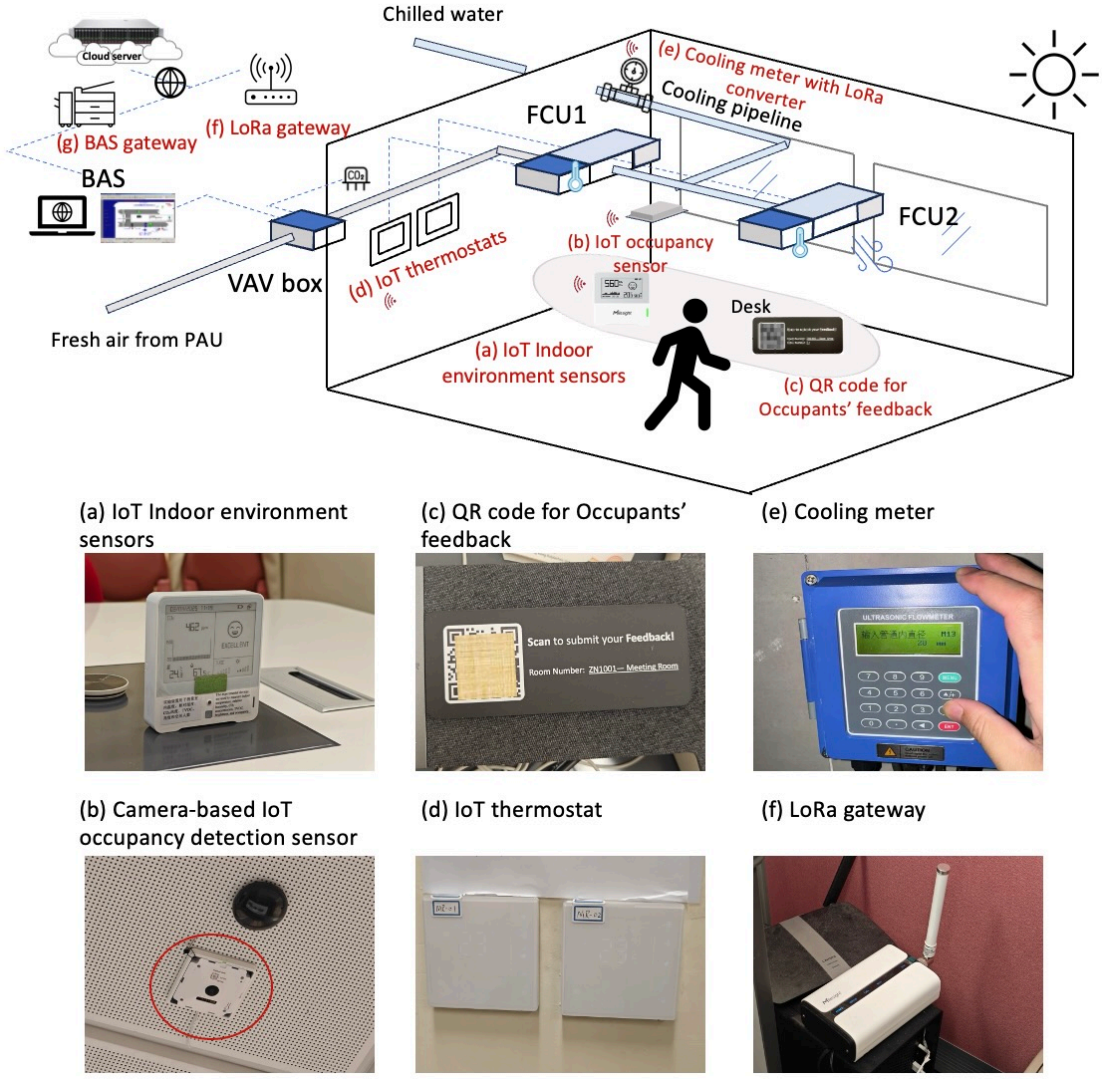


Figure 6-5 Schematic of IoT retrofit and installed devices

The IoT retrofit has been implemented across all three rooms in the living lab. However, the meeting room was selected as the target space for testing the OILT MPC due to its highly irregular occupancy pattern. The meeting room often remains unoccupied for extended periods, punctuated by sudden bursts of high occupancy, where multiple individuals enter and exit within a short timeframe. This random and intermittent usage pattern presents significant challenges for indoor environmental control. The OILT MPC was tested under these demanding conditions to rigorously evaluate its performance, demonstrating its ability to dynamically adapt and optimize control strategies in response to real-time occupancy fluctuations.

6.3.3 Experiment setup

The IoT retrofit was completed in early June, 2024 enabling data collection for system testing. The experiment evaluated the proposed OILT MPC against two baseline control strategies: (1) FIX strategy, a conventional control strategy maintaining a fixed setpoint; and (2) OBC strategy, a rule-based control strategy adjusting setpoints based on occupancy detection. Details of these baseline strategies are outlined in Table 6-2, and the testing periods are illustrated in Figure 6-6.

Before July 1, 2024, the FIX strategy was in continuous operation. From July 9 to July 21, data was collected for system identification of the SDE-based building dynamic models. From August 5 to August 9, the OBC strategy was tested. From September 25 to November 1, 2024, the proposed MPC system was tested on weekdays, covering the late summer and autumn seasons in Hong Kong. During this period, a total of 28 days of MPC operation data was collected, encompassing a range of small, medium, and large meetings.

To illustrate the control behaviors of the three strategies, nine typical cases were selected for analysis. These cases are summarized in Table 6-3, providing insight into how each control strategy performed under different occupancy conditions.

Table 6-2 Comparison of baseline and proposed strategies for built environment control

Control strategy	Indoor temperature setpoints	Fresh air flow rate setpoints
Baseline 1: FIX	The temperature setpoint is adjust by the occupants	Determined by designed occupant number*
Baseline 2: OBC	24°C when the room is occupied; 30°C when the room is empty.	Determined by real-time occupant number*
Proposed: OILT MPC	Optimized setpoint	Optimized setpoint

* The fresh air flow rate is determined using the equation $V_{bz,t} = R_p \times P_{z,t} + R_a \times A_z$, where A_z is room floor area; R_a is the fresh air flow rate per floor area, 0.3 L/s/m^2; for the FIX strategy, P_z is the designed number of people in the room, which is chosen as 15, for the OBC strategy, P_z is real-time measured number of people in the room; R_p fresh air flow rate per people, 5 L/s/person. Note that the R_p is chosen with higher margin than the ASHERAE recommendation for meeting room (minimum 2.5 L/s/person).

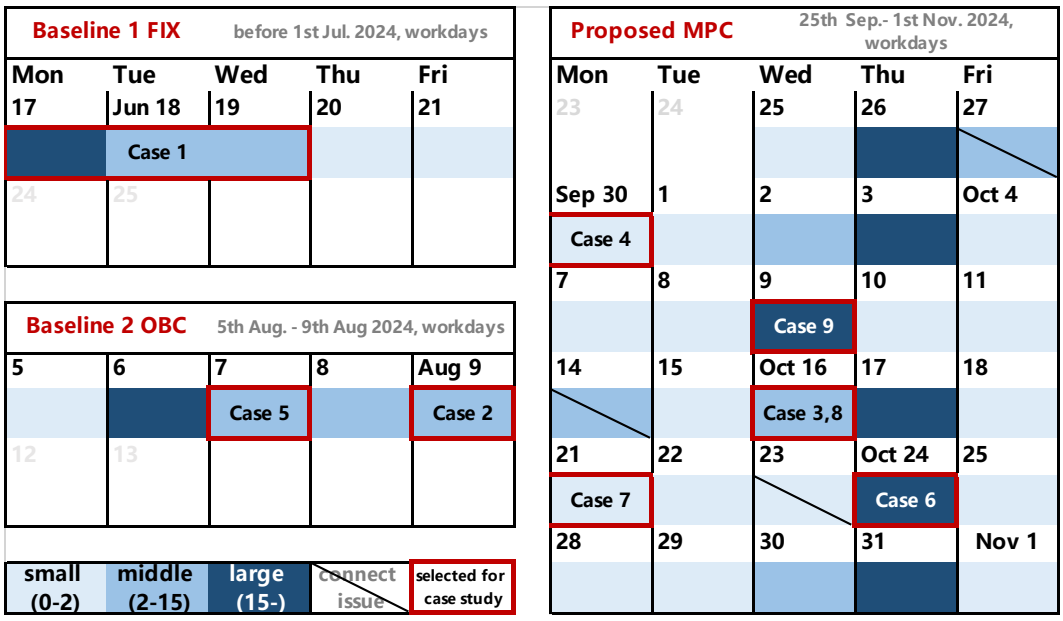


Figure 6-6 Experiment periods of different control strategies

Table 6-3 Case explanation

Case	Testing date	Explanation	Control strategy	Control behavior

1	Jun 18	The temperature setpoint is adjust by the occupants, and the BAS do not interfere the setpoint, and the HVAC system operates 24/7 to meet the room setpoint.	FIX	Indoor temperature
2	Aug 9	Adjusts indoor temperature setpoints based on real-time indoor occupancy presence.	OBC	
3	Oct 16	Optimal precooling is implemented based on the stochastic occupancy prediction.	MPC	
4	Sep 30	The precooling dynamically varies with internal and external factors, including current indoor air temperature, outdoor temperature and solar radiation.	MPC	
5	Aug 7	Adjusts fresh air flow rate setpoints based on real-time indoor occupant number	OBC	CO_2 concentration
6	Oct 24	The MPC optimizes fresh air flow rate avoiding over- and under-ventilation	MPC	
7	Oct 21	The MPC responds to “cold” feedback in small meeting.	MPC	Real-time feedback
8	Oct 16	The MPC responds to “hot” feedback in middle meeting.	MPC	
9	Oct 9	The MPC responds to “cold” feedback in large meeting.	MPC	

6.4 Results

6.4.1 Indoor air temperature

6.4.1.1 FIX control strategy (Case 1)

Figure 6-7 illustrates the control results of indoor air temperature under the conventional control strategy, where occupants manually adjust the temperature setpoint.

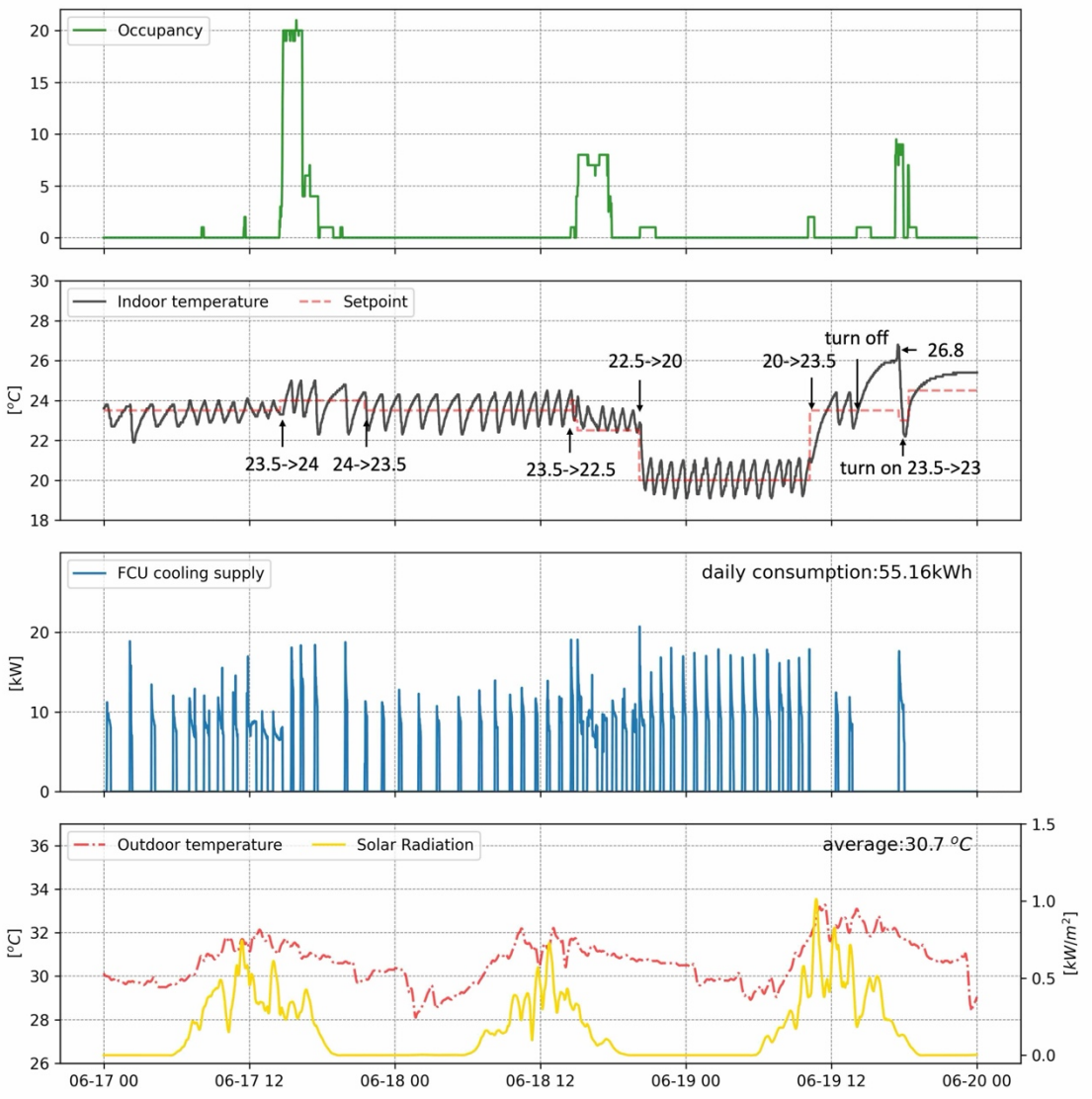


Figure 6-7 Operation data of the target meeting room under conventional control strategy in Case 1

On June 17, at the beginning of a group meeting, an occupant slightly increased the setpoint from 23.5°C to 24°C. The following day, June 18, at the start of a medium-sized meeting, another occupant lowered the setpoint from 23.5°C to 22.5°C. However, under this control strategy, users frequently forget to turn off the thermostat when leaving the meeting room, causing the FCU to continuously maintain the setpoint.

A notable energy waste scenario occurred on the night of June 18, when an occupant entered the meeting room and set the thermostat to 20°C. After staying for approximately one hour, the person left without turning off the thermostat, resulting in

the FCU running overnight, consuming excessive energy. A similar situation was observed on June 17, further highlighting the inefficiency of this manual control strategy. Over these three days, the average outdoor temperature was 30.7°C, and the daily average cooling consumption reached 55.16 kWh.

On the morning of June 19, another occupant—possibly feeling too cold—adjusted the setpoint from 20°C back to 23.5°C. Later that afternoon, an energy-conscious occupant turned off the thermostat upon leaving. However, this caused the indoor temperature to quickly rise to 26.8°C, making it uncomfortably warm for attendees of a night meeting. In response, they turned the thermostat back on and immediately lowered the setpoint, but still had to endure the heat for a period before the room cooled down.

This series of events demonstrates the inconsistencies and inefficiencies of conventional manual control, where frequent adjustments, forgetfulness, and lack of coordination lead to energy waste and thermal discomfort.

6.4.1.2 OBC control strategy (Case 2)

Figure 6-8 presents the control results of the OBC strategy, where the temperature setpoint is dynamically adjusted based on real-time indoor occupancy detection. A key advantage of this approach is that the FCU operates only when the room is occupied, leading to significant energy savings, particularly during unoccupied periods.

On August 9, the average outdoor temperature was 32.2°C, which was 1.5°C higher than during the FIX control strategy tested in Case 1 periods, and solar radiation levels were also elevated. Despite these more demanding environmental conditions, the daily cooling consumption was only 22.7 kWh, representing a 58.7% reduction compared to the FIX control strategy.

However, while OBC strategy effectively reduces unnecessary cooling, it can sometimes lead to thermal discomfort, particularly when occupants first enter the room. The right-hand side of Figure 6-8 highlights a medium-sized meeting held between 13:30 and 15:00. The first attendee arrived at approximately 14:02, triggering the OBC strategy to activate after a short delay of about 3 minutes, at which point the setpoint

was adjusted to 24°C.

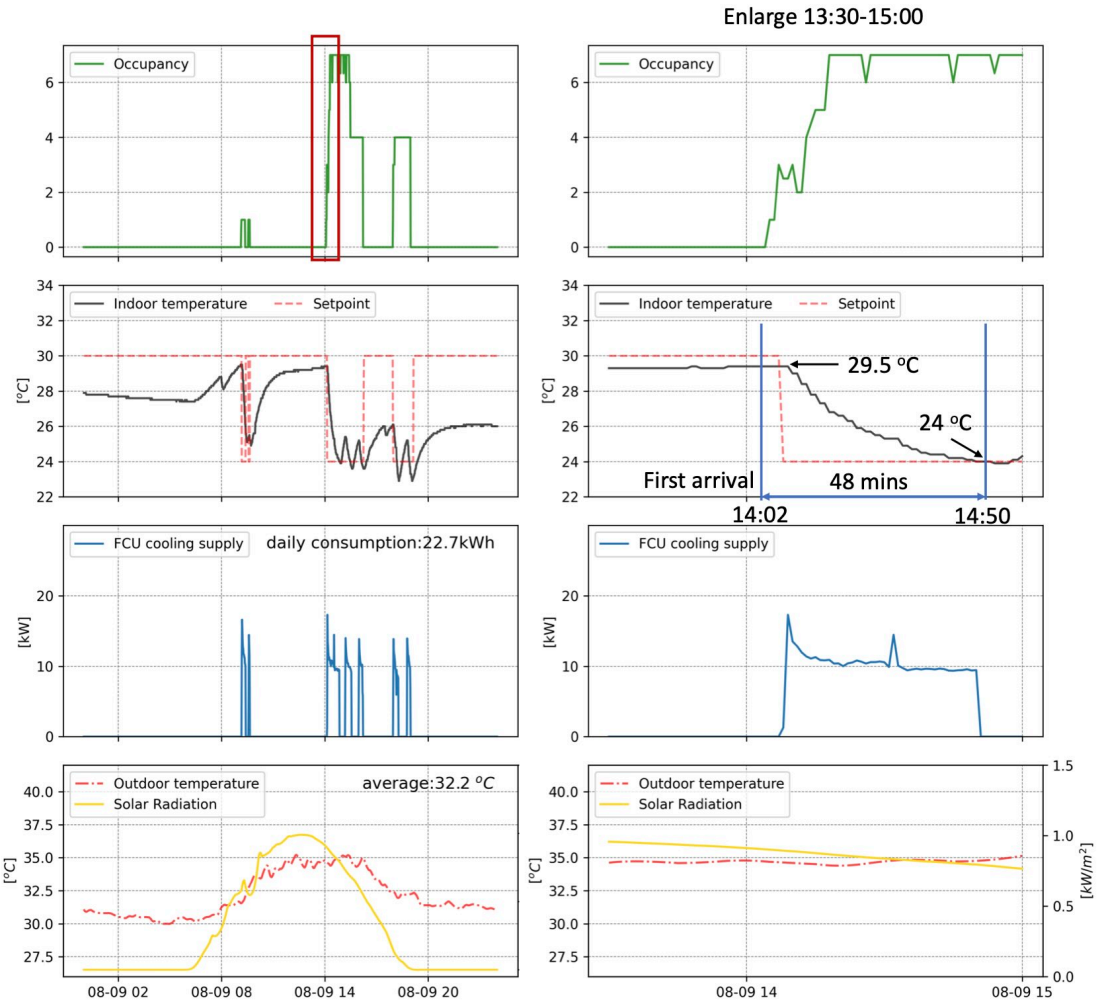


Figure 6-8 Operation data of the target meeting room under the OBC control strategy

At that moment, however, the indoor temperature had already risen to approximately 29.5°C, while the outdoor temperature was extremely high, reaching around 35°C. As a result, the FCU required a significant amount of time to cool the room down to 24°C, taking approximately 48 minutes. This prolonged cooling period left occupants enduring uncomfortable heat for an extended time, highlighting a key limitation of the OBC strategy.

6.4.1.3 Proposed OILT MPC control strategy (Case 3-4)

Figure 6-9 illustrates the control results of the proposed MPC strategy in Case 3. In the first plot, the red line represents the actual measured occupant number, while the colored lines indicate the expected occupant number trajectory predicted at different time steps. For a clearer visualization of the occupancy prediction results during the morning optimal precooling period, Figure 6-10 provides a mesh plot.

The right side of Figure 6-9 presents an enlarged view of the period from 8:30 to 10:30, when a medium-sized meeting was held. At 9:06, the occupancy prediction model (represented by the red dot in Figure 6-10) anticipated that occupants would arrive within 15 minutes. Based on the building's thermal dynamic model, the estimated time required to lower the indoor temperature from its current level to the optimized setpoint (25.5°C in this case) was also approximately 15 minutes. Consequently, precooling was initiated at that moment.

It is important to note that this optimized setpoint (25.5°C) is not the same as the reference temperature (24°C) in the MPC system. Instead, the optimized setpoint is determined by balancing the trade-off between thermal comfort and energy savings, while also accounting for the uncertainty of future occupancy. The final optimized setpoint depends on the weight assigned to thermal comfort by the operator: a higher weight on thermal comfort (ρ_{TC}) results in a lower optimized setpoint, prioritizing occupant comfort; while a higher weight on energy savings (ρ_E) leads to a higher optimized setpoint, reducing energy consumption, but may compromise thermal comfort.

The precooling process began at 9:07 by setting the optimized setpoint. After approximately 13 minutes, the room's temperature dropped from 27.5°C to 25°C, consuming a total of 2.1 kWh of cooling energy. Although no occupants had arrived yet, the setpoint remained at the optimized value since the room was still expected to be occupied within the next 15 minutes (as shown by the red line in Figure 6-10).

By 9:42, the first occupant arrived, at which point the indoor temperature had slightly increased to 25.7°C. Without optimal precooling, the temperature could have been

significantly higher, especially given the rising outdoor temperature and increasing solar radiation. This demonstrates the effectiveness of the MPC strategy in reducing thermal discomfort upon occupant arrival.

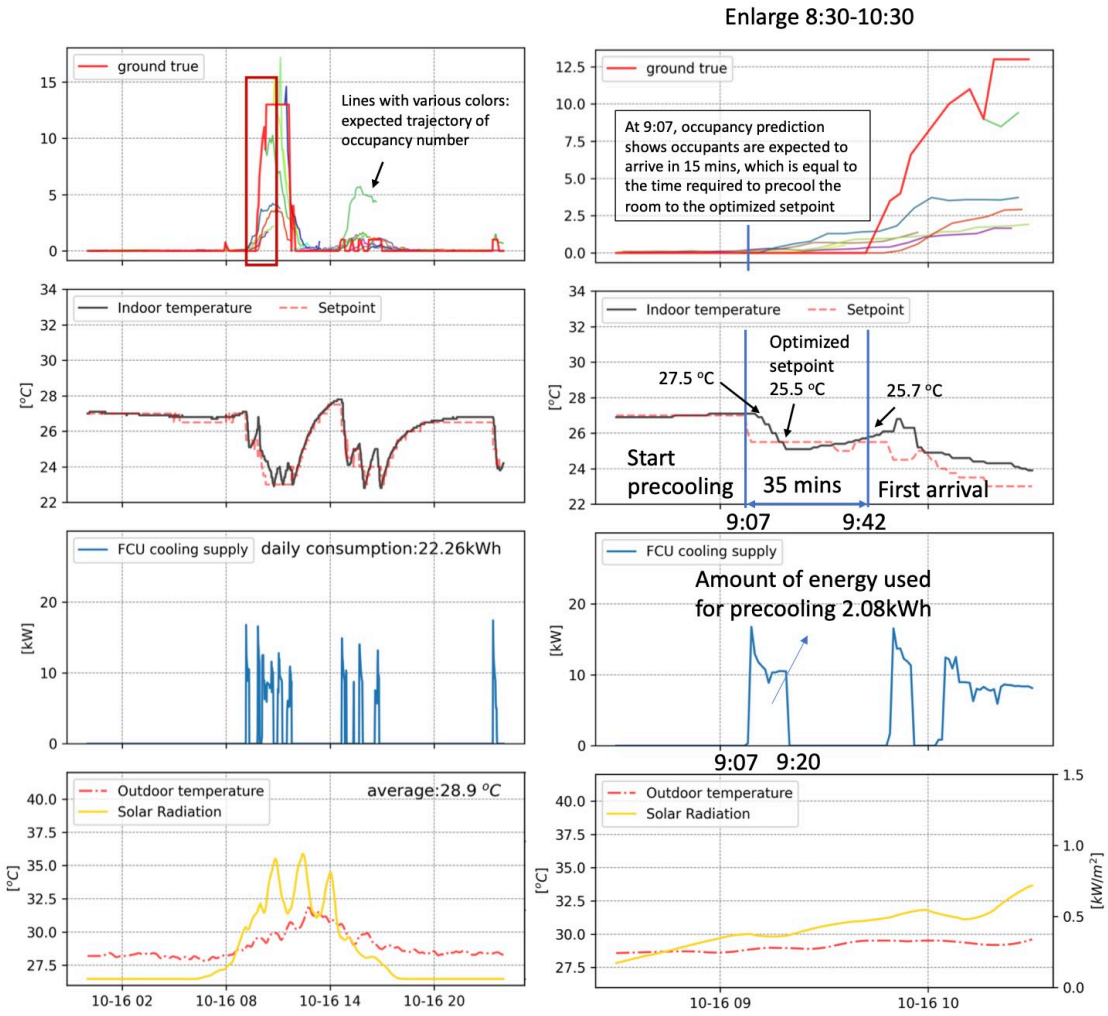


Figure 6-9 Operation data of the target meeting room under the OBC control strategy in Case 3

Figure 6-11 presents the control results of the MPC strategy in Case 4, while Figure 6-12 provides the occupancy prediction mesh plot for the morning optimal precooling period. Unlike Case 3, the meeting room remained largely unoccupied throughout the day. However, precooling was still implemented in the morning to mitigate potential thermal discomfort in case occupants arrived.

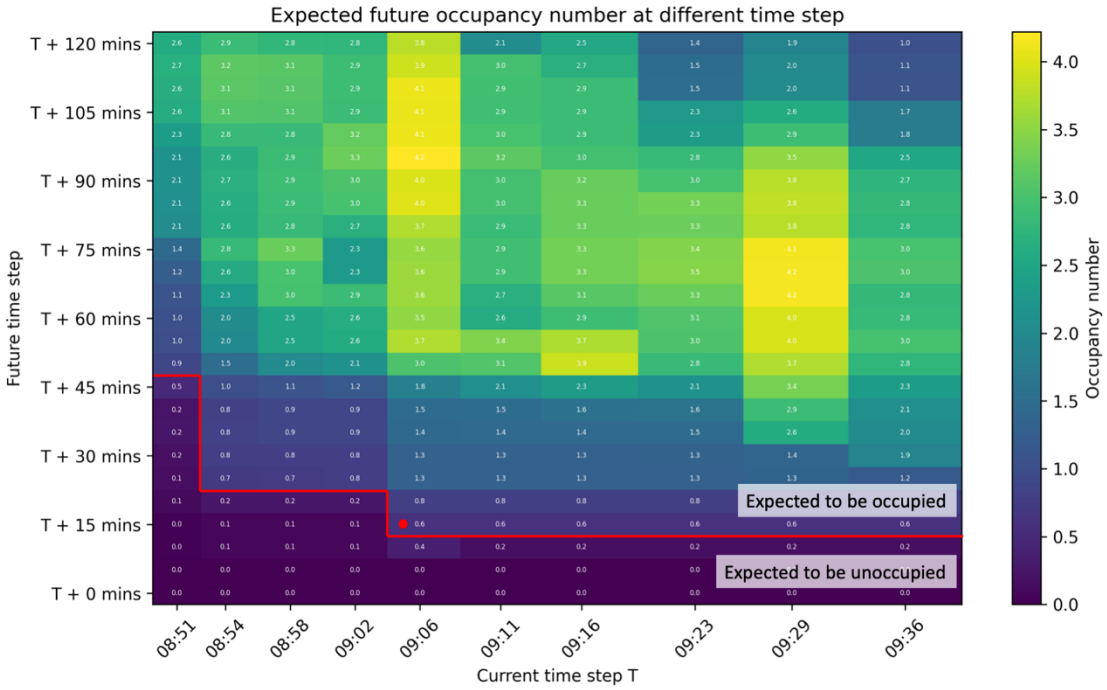


Figure 6-10 The expected occupant number and occupancy presence predicted at different time steps during morning optimal precooling period in Case 3 (16th Oct.)

The duration of the precooling period varied based on indoor air temperature, outdoor temperature, and solar radiation levels. At 8:57, the indoor air temperature was 29.6°C, which was 2.1°C higher than in Case 3, indicating greater heat accumulation overnight and in the early morning. Additionally, the weather in Case 4 was hotter, with an average outdoor temperature of 32.6°C—3.7°C higher than in Case 3. According to the occupancy prediction model (as indicated by the red dot in Figure 6-12), occupants were expected to arrive in 25 minutes. This prediction closely aligned with the time required for precooling to bring the indoor temperature down from 29.6°C to the optimized setpoint, considering the higher thermal load and outdoor heat disturbance.

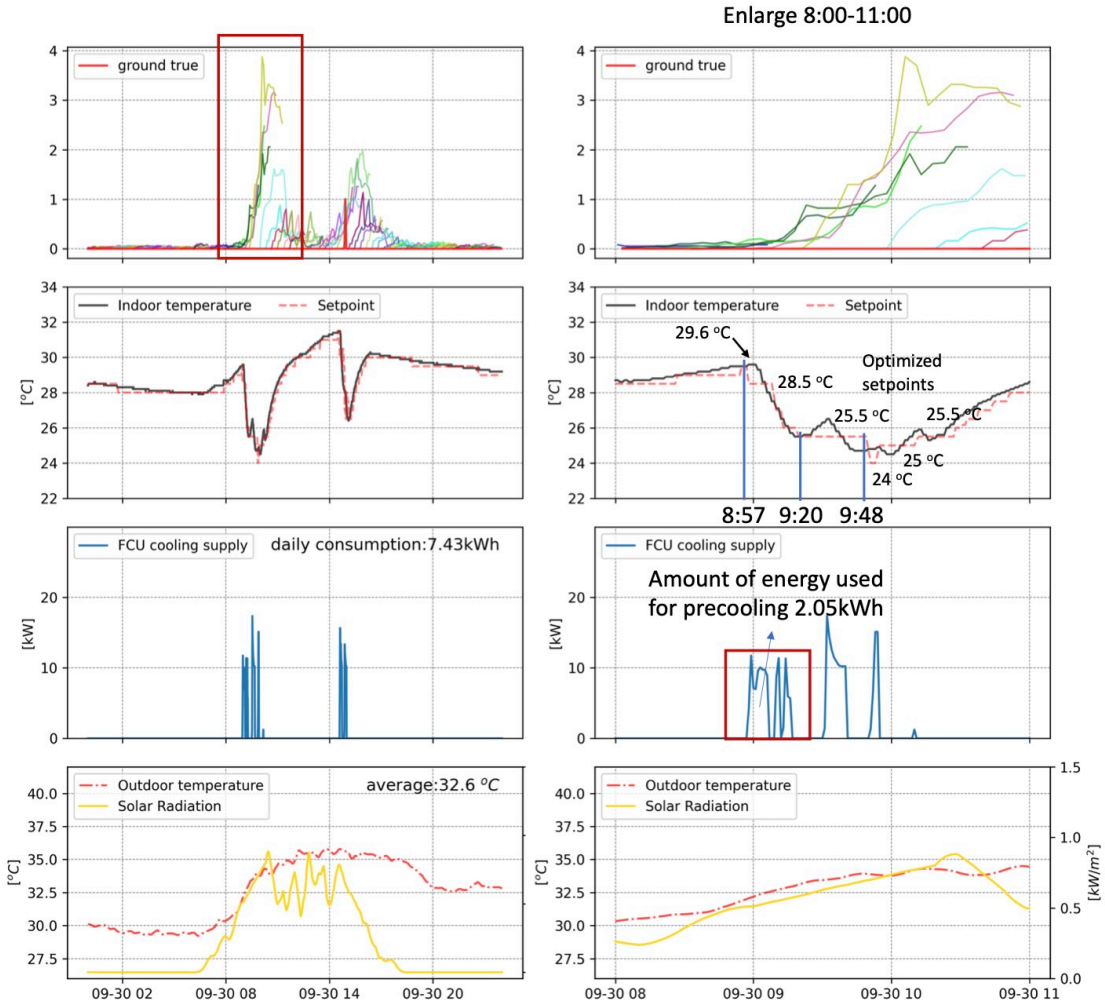


Figure 6-11 Operation data of the target meeting room under the OBC control strategy in Case 4

Consequently, the MPC initiated precooling earlier—10 minutes sooner than in Case 3. After approximately 23 minutes, the indoor temperature was successfully lowered to 25.5°C. This optimized setpoint was maintained until 9:48, when the room was expected to be occupied within the next 10 minutes. At this point, the MPC further lowered the setpoint to 24°C for about 5 minutes.

However, since no occupants arrived, the MPC gradually increased the setpoint, eventually leading the FCU to stop providing cooling, thereby minimizing unnecessary energy consumption.

These results highlight the adaptive intelligence of the proposed MPC strategy, which

not only optimizes thermal comfort but also dynamically adjusts cooling operation based on real-time occupancy predictions and environmental conditions, ensuring both comfort and energy efficiency.

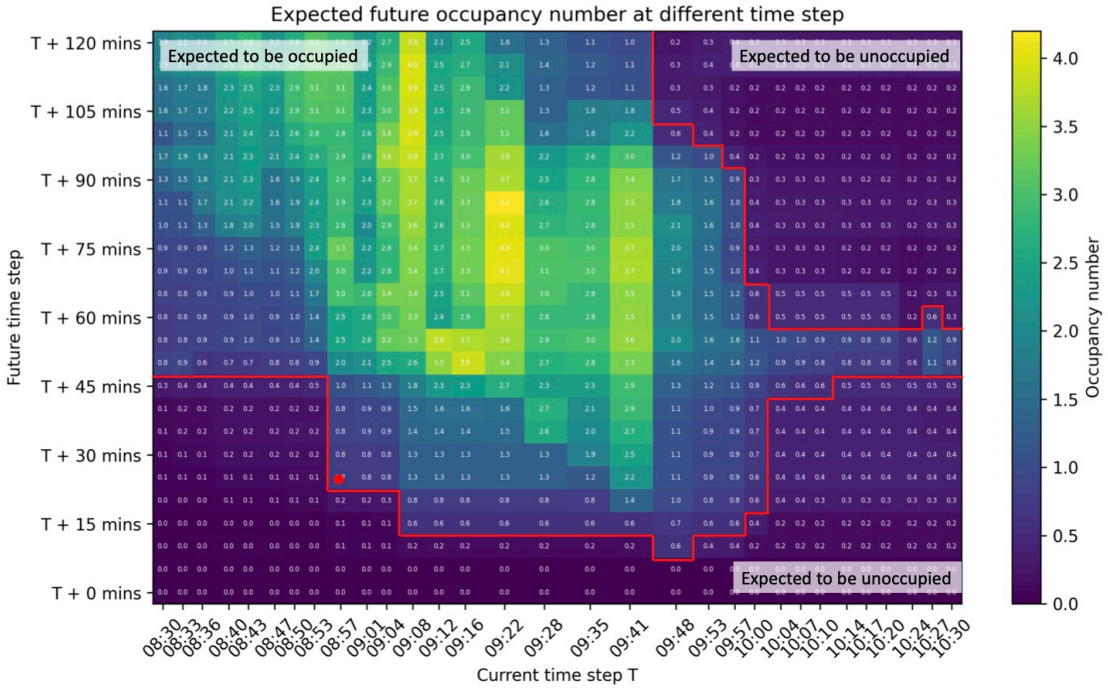


Figure 6-12 The expected occupant number and occupancy presence predicted at different time steps during morning optimal precooling period in Case 4 (30th Sep.)

6.4.2 CO_2 concentration

6.4.2.1 OBC control strategy (Case 5)

Figure 6-13 presents the fresh air flow rate control results under the OBC (Occupancy-Based Control) strategy. In this approach, the fresh air flow rate setpoint is adjusted based on real-time indoor occupancy detection. When the room is unoccupied, the PAU/VAV system supplies a minimum amount of fresh air.

For occupied periods, the fresh air flow rate per person (R_p) was set at 5 L/s per person, which is higher than the ASHRAE-recommended minimum of 2.5 L/s per person for meeting rooms. The right side of Figure 6-13 provides an enlarged view of the period from 14:00 to 16:00, during which a medium-sized meeting took place. Throughout the entire meeting, the average CO_2 concentration remained around 750 ppm, which is below the 800 ppm threshold, indicating that the system was over-supplying fresh air.

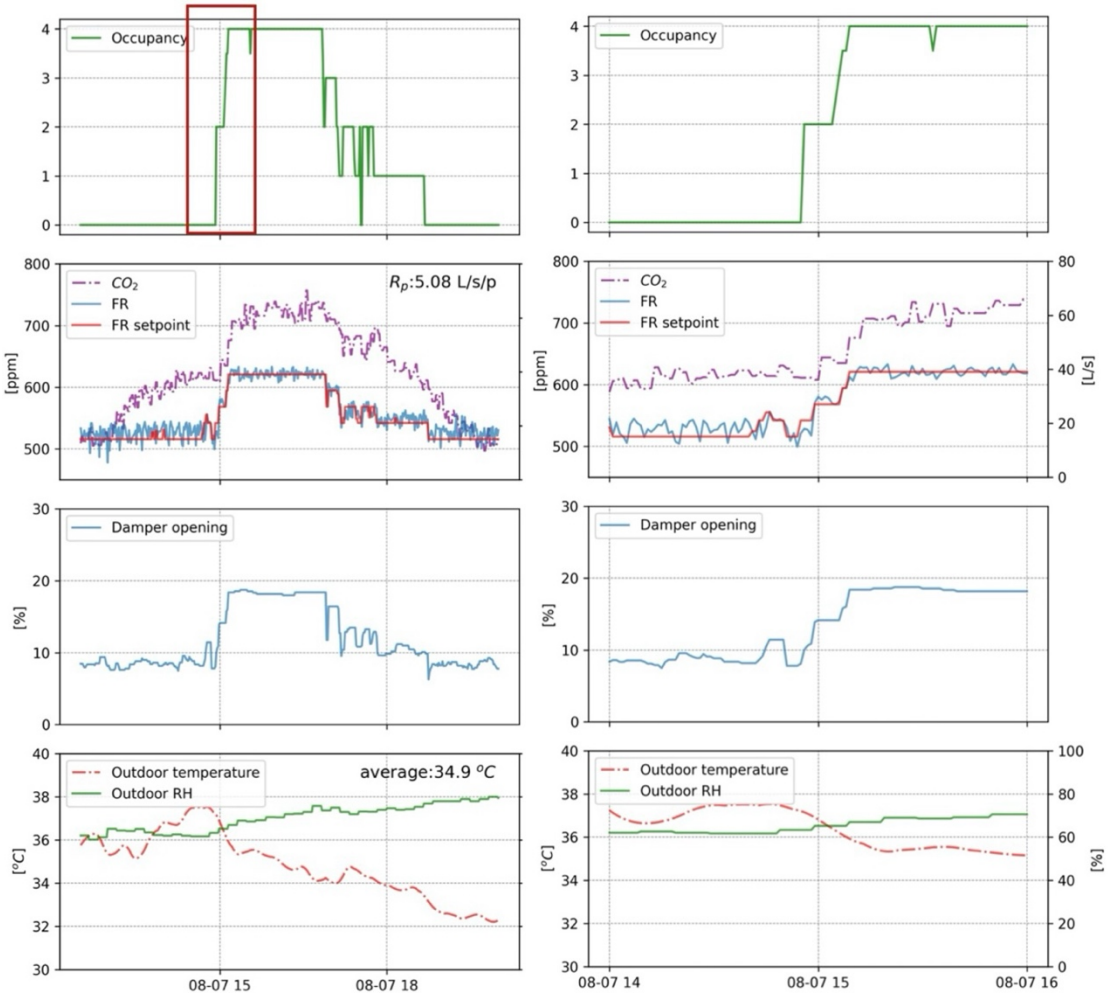


Figure 6-13 Operation data of the target meeting room under the OBC control strategy

6.4.2.2 Proposed OILT MPC control strategy (Case 6)

Figure 6-14 illustrates the fresh air flow rate control results under the proposed MPC strategy. Unlike the OBC strategy, which relies on a fixed (R_p) value, the MPC dynamically adjusts the fresh air flow rate based on real-time CO_2 measurements, occupancy data, and occupancy predictions. Instead of maintaining a constant fresh air supply, the MPC sets a CO_2 threshold and modulates airflow adaptively to maintain IAQ while optimizing energy efficiency.

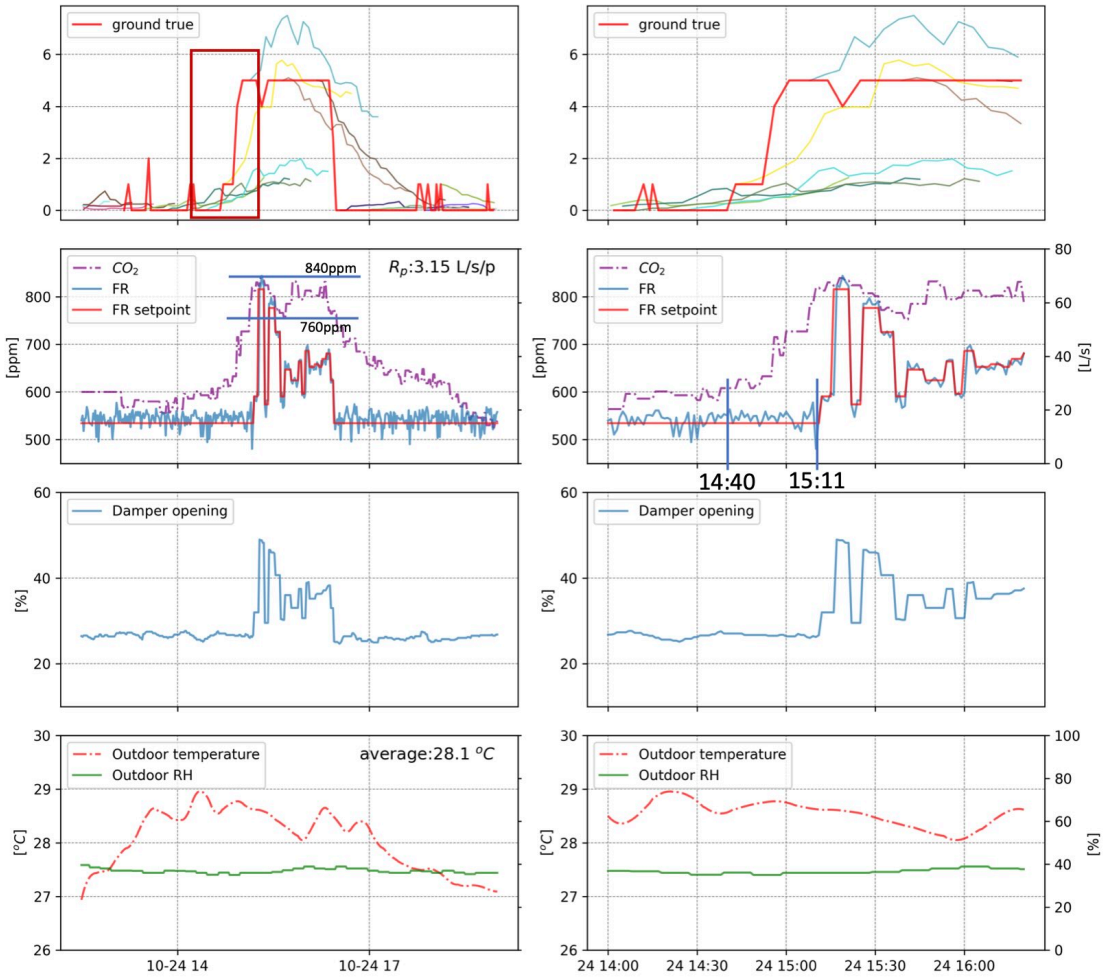


Figure 6-14 Operation data of the target meeting room under the MPC control strategy

During the medium-sized meeting, the MPC effectively maintained CO_2 levels around the 800 ppm threshold (±40 ppm), considering the CO_2 sensor's accuracy of ±55 ppm. This adaptive approach reduced the average R_p value to 3.15 L/s per person, which is 38% lower than the OBC strategy's fixed setting. However, if the ASHRAE-recommended minimum of 2.5 L/s per person had been applied, it could have resulted in under-ventilation, highlighting the advantage of the MPC's dynamic control strategy. The right side of Figure 6-14 provides an enlarged view of the 14:00 to 16:00 period, while Figure 6-15 presents a mesh plot of the occupancy prediction results during the medium-sized meeting in Case 6. Between 14:40 and 15:11, occupancy gradually increased to five people. Notably, from 14:46 to 15:08, the occupancy prediction model

overestimated the future occupant number by 30 minutes. However, the fresh air flow rate setpoint did not increase immediately; instead, it remained at the minimum level until CO_2 concentrations approached the threshold. Unlike temperature control, where the relatively large thermal mass requires precooling considerations, the MPC does not need to pre-supply fresh air, as the room's air volume is relatively small compared to the large fresh air flow capacity of the VAV system.

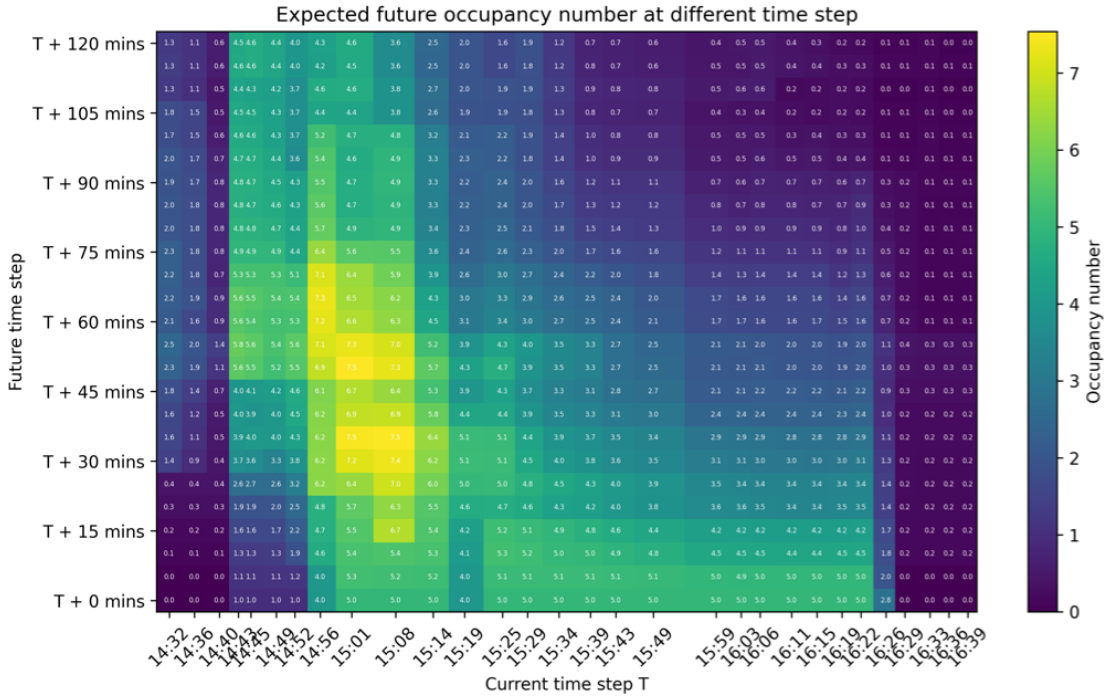


Figure 6-15 The expected occupant number predicted at different time steps in Case 6 (24th Oct.)

6.4.3 Response to real-time occupant's feedback

During the MPC strategy testing period, a total of six thermal comfort votes were recorded, as summarized in Table 6-4. The votes included three instances of "cold," one "cool," one "neutral," and one "hot." This study selected three representative cases for detailed analysis, each corresponding to a different meeting size, as illustrated in Figure 6-16.

Table 6-4 Summary of occupant's feedback

time	vote	Reference temperature	size	Case
2024/10/9 10:25	cold	24 -> 25	large	9
2024/10/16 10:03	hot	24 -> 23	middle	8
2024/10/21 10:18	cold	24 -> 25	small	7
2024/10/24 09:42	cool	24 -> 24.5	large	-
2024/10/25 22:36	cold	24 -> 25	small	-
2024/10/30 14:44	neutral	-	middle	-

In Case 7, the vote was cast when only one person was present in the room. That morning, the MPC did not initiate precooling, as both indoor and outdoor temperatures were relatively low. As a result, the temperature setpoint before the occupant's arrival was maintained at approximately 25.5°C. At 10:13, when the occupant entered the room, the setpoint was immediately lowered to 24°C. However, despite the indoor temperature remaining at 25.5°C, the occupant felt cold due to the cooling supply, leading to a "cold" vote. In response, the MPC reference temperature was adjusted from the default 24°C to 25°C, and the setpoint was modified accordingly. The FCU was then turned off, preventing further cooling.

In Case 8, the vote occurred during a middle-sized meeting. The precooling period for this day had been previously analyzed in Case 3. Thanks to optimal precooling, the indoor temperature had already been lowered by 2.5°C before occupants arrived (see Figure 6-9). As more people entered, a significant amount of heat was introduced into the room, causing the temperature to rise. However, with the cooling system actively operating, the room temperature quickly dropped back to 25°C. During this time, one participant felt warm upon entering and submitted a "hot" vote.

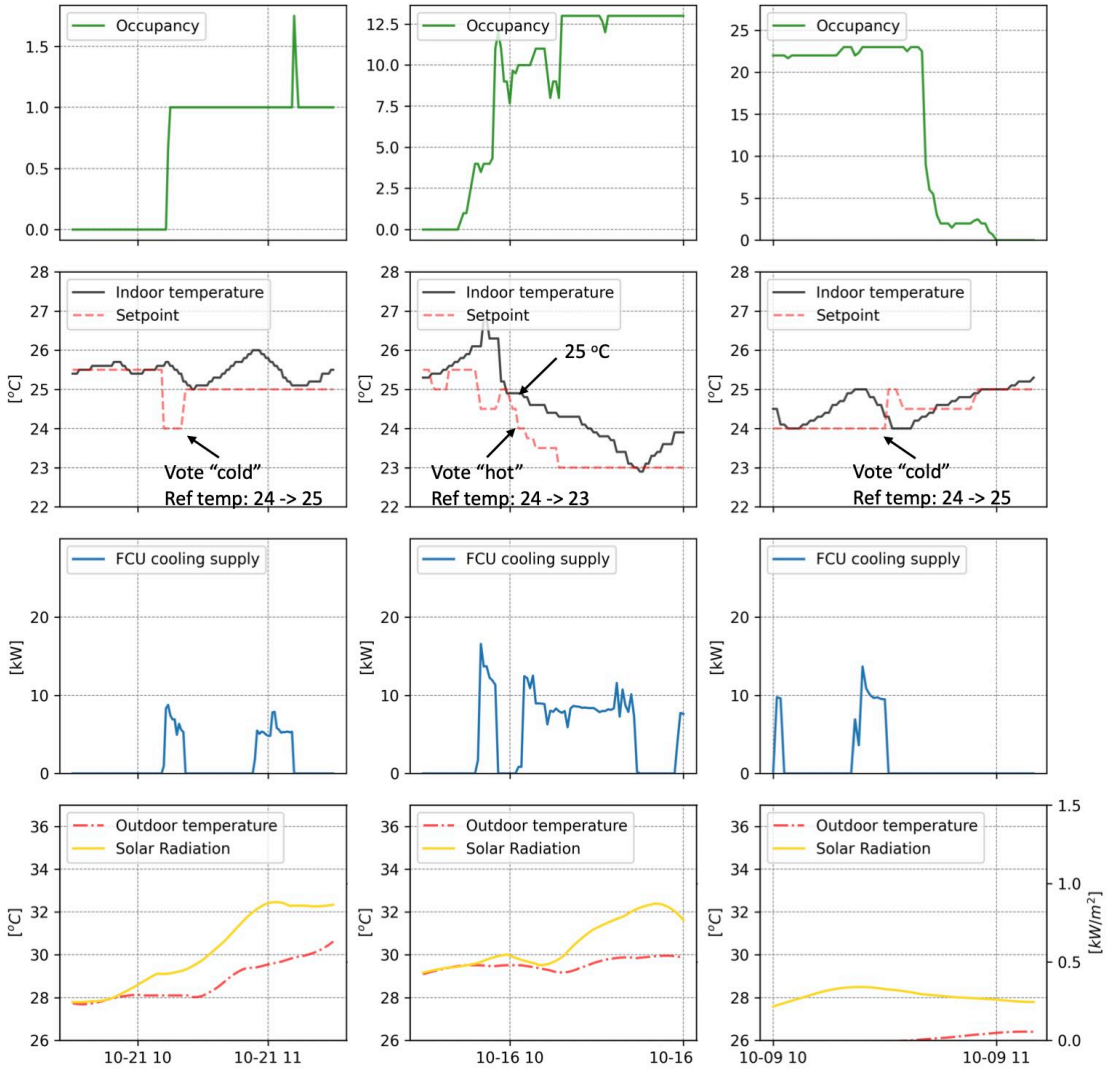


Figure 6-16 Control results of MPC responding to real-time occupants' feedback (Case 7,8,9)

In response, the MPC reference temperature was lowered from the default 24°C to 23°C, and the setpoint was gradually reduced to match the new reference. The FCU continued providing cooling until the indoor temperature reached the new setpoint, ensuring improved thermal comfort.

In Case 9, the vote occurred during a large-sized meeting. Throughout the meeting, the MPC maintained the temperature setpoint at 24°C. However, just as the meeting was concluding, an occupant submitted a "cold" vote. In response, the MPC reference temperature was raised from 24°C to 25°C, and the FCU was immediately turned off.

These cases highlight the adaptive nature of the proposed MPC strategy, which dynamically adjusts temperature setpoints based on real-time occupant feedback, ensuring personalized thermal comfort.

6.4.4 Energy performance

One of the key challenges in verifying the energy-saving potential of the proposed strategy against baseline strategies in real-world implementation tests is the difficulty of finding comparable weather and occupancy conditions when testing different control strategies. Given the highly stochastic nature of both the timing and number of participants in the target meeting room, achieving identical occupancy conditions across tests is particularly challenging. Additionally, weather conditions varied significantly throughout the testing periods, further complicating direct comparisons.

To address this issue, this study employs two key factors—the daily average outdoor temperature and the daily average occupant number—to quantify the testing conditions for each control strategy. These factors enable the identification of similar scenarios for a more equitable energy performance comparison. Figure 6-17 visualizes the daily testing conditions for all test dates across the three control strategys: FIX (5 days), OBC (5 days), and MPC (28 days). From this dataset, two groups of daily conditions were identified as the most comparable across the three strategies: One group represents large meetings; the other represents small to medium-sized meetings. These two groups are summarized in Table 6-5.

In the first group, the average outdoor temperature for the FIX strategy is 30.57 °C, with an average occupant number of 1.71, both of which are the lowest among the three strategies. However, the FCU cooling consumption is the highest, reaching 58.03 kWh. In contrast, the average outdoor temperature for the OBC strategy is 32.74 °C, with an average occupant number of 2.23, both of which are the highest among the three strategies. Notably, the FCU cooling consumption for OBC is the lowest at 18.81 kWh. The average outdoor temperature for the MPC strategy is 31.32 °C, with an average occupant number of 2.2, both lower than those of OBC, while the MPC consumed more

energy than OBC. This increased energy usage is due to the implementation of precooling to enhance comfort, as analyzed in Cases 2, 3, and 4. Nevertheless, the energy consumption of MPC is significantly lower than that of the FIX strategy, achieving a 54.9% energy savings compared to FIX strategy.

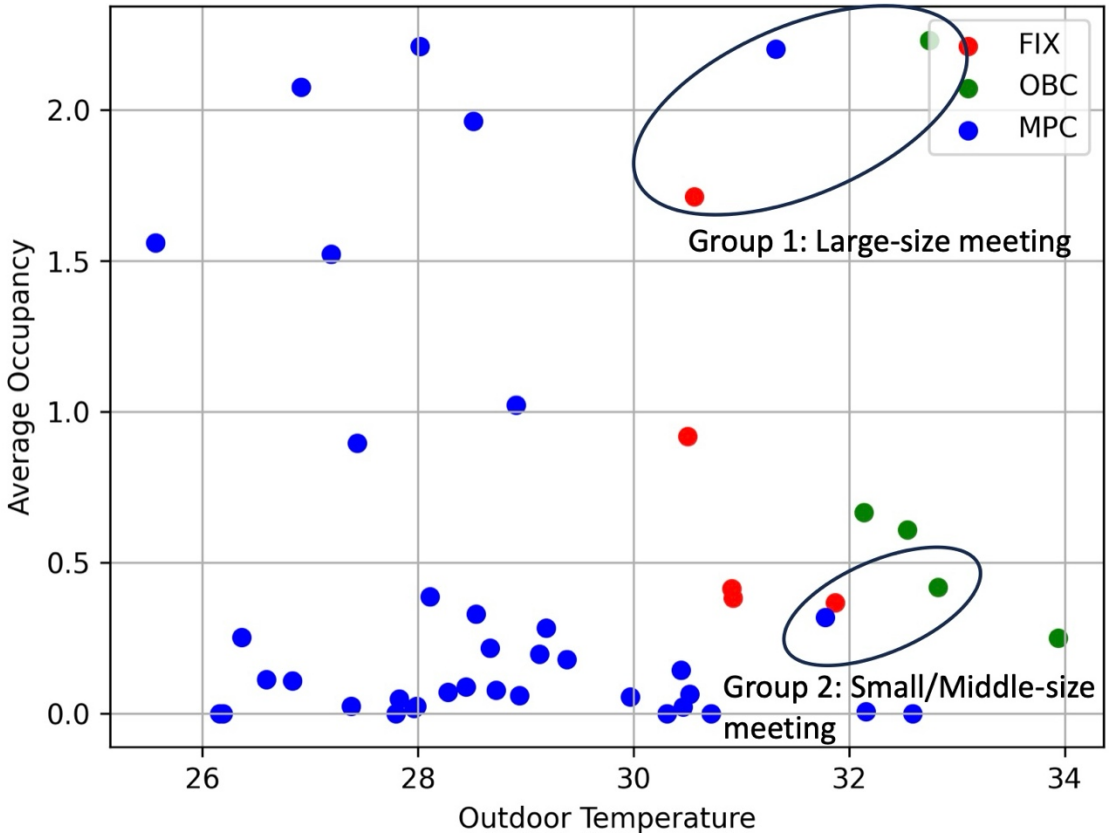


Figure 6-17 Daily testing condition of all the testing dates under the three control strategy

Table 6-5 Energy performance of the proposed MPC

Group	Date	Control strategy	Average outdoor temperature	Average occupant number	FCU cooling consumption
			°C		kWh
1: Large meeting	2024/6/17	FIX	30.57	1.71	58.03
	2024/8/6	OBC	32.74	2.23	18.81
	2024/9/26	MPC	31.32	2.2	**26.17**
2: Small/Middle	2024/6/21	FIX	31.87	0.37	51.56
	2024/8/7	OBC	32.82	0.42	14.93

meeting	2024/9/27	MPC	31.77	0.32	**14.98***

* This value is equal to the measured value 12.93 kWh and the estimated precooling energy consumption 2.05 kWh

In the second group, the average outdoor temperature for the FIX strategy is 31.87 °C, with an average occupant number of 0.37. The FCU cooling consumption remains high at 51.56 kWh, which is 11.1% lower than in the first group. The average outdoor temperature for the OBC strategy is 32.82 °C, with an average occupant number of 0.42, leading to an FCU cooling consumption of 18.81 kWh, which is 20.62% lower than in the first group. During the MPC test on September 27, the weather and occupancy conditions were suitable for comparison within this second group. However, a communication error on that date resulted in no cooling being implemented before 9 AM. This issue is analyzed in Section 6.5.2.2. To estimate the cooling energy that should have been consumed during the missed precooling period, we referenced data from Case 4, which had similar weather conditions. In Case 4, the precooling energy consumption was 2.05 kWh. Adding this to the recorded energy usage, the total estimated FCU cooling consumption for MPC in Group 2 was 14.98 kWh. The energy consumption of MPC in Group 2 is also significantly lower than that of the FIX strategy, achieving a 70.9% energy savings compared to FIX strategy.

6.4.5 Analysis of occupancy modelling accuracy on the control performance

The accuracy of occupancy forecasting, plays a critical role in the performance of the proposed MPC strategy. This section we provide an analysis of how overestimation and underestimation of occupancy predictions impact the performance of the proposed method, and how these effects are mitigated.

(1) Impact of occupancy prediction accuracy on thermal comfort and energy efficiency

When the occupancy is overestimated, the MPC strategy may implement pre-cooling earlier or maintain a lower indoor temperature than necessary. This could lead to unnecessary cooling, which reduces energy efficiency without significantly improving thermal comfort. For example, in Figure 6-11, the MPC begins pre-cooling earlier due

to higher outdoor temperatures and solar radiation, but overestimations of occupancy result in longer cooling periods than required. Overestimations increase cooling energy consumption as the system unnecessarily conditions the space for more occupants than present. This is evident during certain periods when MPC maintains lower temperature setpoints than required due to over-predicted occupancy.

Underestimating occupancy can delay or reduce the extent of pre-cooling, leading to thermal discomfort for occupants. However, our results (e.g., Figure 6-9) show that the MPC strategy can still quickly adapt to real-time feedback and rapidly adjust setpoints to restore thermal comfort after occupants arrive. Underestimation reduces pre-cooling delivery, which may save energy in the short term but risks energy-intensive recovery efforts to quickly reach comfortable conditions after occupants arrive.

(2) Impact of occupancy prediction accuracy on IAQ

Overestimating occupancy leads to excessive fresh air flow rates, resulting in over-ventilation. This ensures that CO_2 levels remain well below the threshold (800 ppm), but at the cost of increased ventilation energy consumption. For example, in Figure 6-14, the fresh air flow rate is maintained at high level for short periods due to overestimated occupancy. Over-ventilation increases fan energy consumption and cooling energy requirements, especially when outdoor air needs to be conditioned.

Underestimating occupancy reduces fresh air flow rates, risking higher CO_2 concentrations. However, the MPC strategy continuously refines its predictions and adjusts ventilation rates dynamically. This ensures that CO_2 concentrations remain below the 800 ppm threshold, even when occupancy is initially underestimated. Underestimation may initially save ventilation energy but can lead to increased consumption if corrective actions (e.g., rapid ventilation increases) are required to meet IAQ thresholds.

(3) Experimental evidence and mitigation strategies

The MPC strategy continuously refines its occupancy predictions based on real-time feedback. This allows the control system to adjust setpoints and fresh air flow rates promptly, mitigating the effects of both overestimation and underestimation. The MPC

strategy balances thermal comfort, IAQ, and energy consumption through its multi-objective cost function. This ensures that even with prediction errors, the system prioritizes maintaining IAQ thresholds and minimizing thermal discomfort over energy savings.

6.5 Discussion and limitations

6.5.1 Cost analysis of the IoT retrofit and payback period

Table 6-6 presents the estimated costs for the IoT retrofit in the meeting room, covering labor costs, equipment costs, and cloud service costs. The total cost of the IoT retrofit amounts to approximately 25,000 HKD (~3,205 USD).

Based on the energy-saving potential analysis in Section 6.4.4, the MPC strategy is projected to achieve energy savings of around 60%. Given that the daily cooling consumption under conventional control strategy is approximately 55 kWh, this suggests a potential reduction of 33 kWh per day. Assuming an overall COP of 3 for the cooling system, this translates to roughly 11 kWh of electricity savings per day. With the electricity price for non-residential users in Hong Kong at approximately 1.5 HKD/kWh, the daily cost savings amount to 16.5 HKD. Considering that Hong Kong experiences an extended cooling season of about 10 months (~300 days per year), the annual cost savings total 4,950 HKD. This results in an estimated capital payback period of approximately 5.1 years.

It is important to note that the labor cost per room may decrease if the retrofit and MPC system are implemented across multiple rooms, thereby shortening the payback period.

Table 6-6 Estimated costs of IoT retrofit in the meeting room

Category	Item	Total costs of IoT retrofit in the meeting room (HKD)
Labor	Estimated infrastructure installation and commissioning	10000
Equipment	IoT indoor environment sensor	1400
	IoT occupancy sensor	1680
	IoT thermostat	500

	Cooling meter	10100
Cloud service	Cloud server (per year)	1500
Total		25180

The cooling meter, while providing precise measurements of the cooling supply, constitutes the largest portion of the retrofit costs, both in terms of equipment and labor expenses. However, the on/off state data of the cooling system can serve as a substitute for precise cooling meter data [117]. The on/off state can be estimated using the IoT thermostat by comparing the temperature setpoint with the measured temperature. If the cooling meter is omitted—resulting in an estimated cost reduction of 15,100 HKD—the total retrofit cost can be significantly lowered, leading to a shorter payback period of approximately 2 years.

6.5.2 Limitations

6.5.2.1 Information delay issues

The current environmental sensor is configured to upload data every 2 minutes; however, an analysis of the timestamps retrieved from the database indicates that the actual upload intervals range from 2 to 4 minutes. This discrepancy may be attributed to fluctuations in the wireless communication network signal.

Similarly, while the MPC system is designed to run every 3 minutes, logged data reveals that the actual execution intervals vary between 3 to 8 minutes. This inconsistency is also reflected in the horizontal coordinate intervals of Figure 6-10 and Figure 6-12. Notably, around 9:30 AM, the MPC execution intervals appear to increase, possibly due to a higher likelihood of occupant arrivals, which complicates the optimization problem and extends the solution time. While this remains speculative, further investigation is required.

As a result, the so-called "real-time" data utilized by the MPC during operation cannot be considered truly real-time, as there is an inherent delay in data collection. Additionally, the variability in MPC execution intervals means that during longer

intervals, new data may not be incorporated in a timely manner, potentially impacting setpoint adjustments and reducing overall control effectiveness. In future research, we will examine the impact of information delays on control performance while continuing to optimize the settings of the LoRa wireless communication network and improving the computational efficiency of the MPC solution process.

6.5.2.2 Communication and program failures

Operational data during failure periods is presented in Figure 6-18. Three main issues were identified:

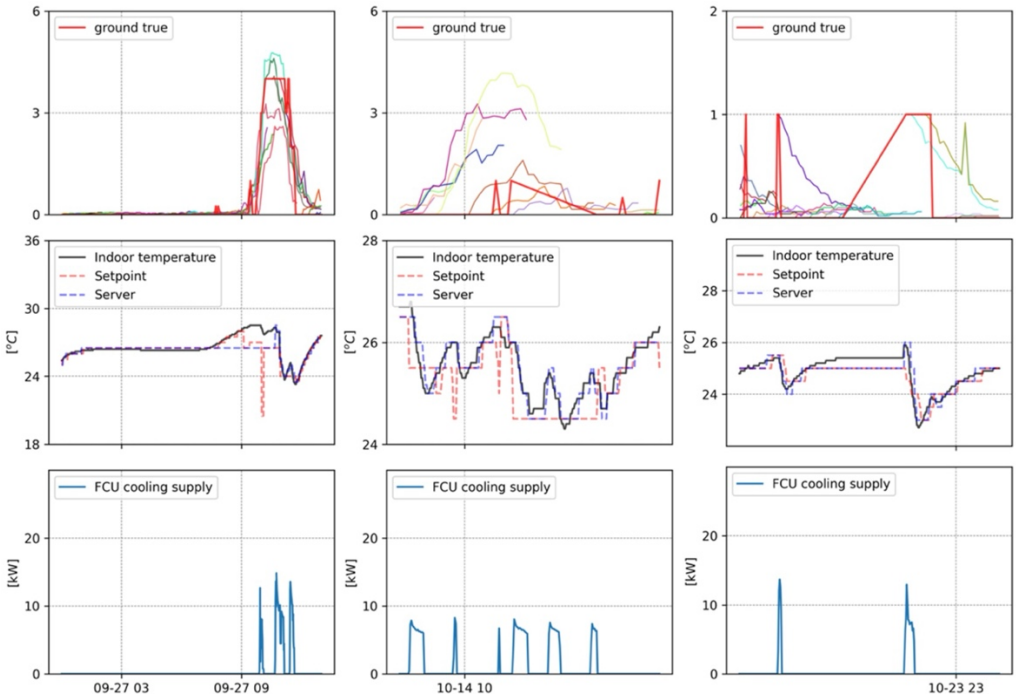


Figure 6-18 Illustration of communication failure during the test of MPC strategy

1. Server temperature sending program failure (September 27th, early morning)

   The IoT thermostat relies on a specified temperature (referred to as "server temperature") to control the FCU. The working plane temperature, measured by the IoT environmental sensor, is sent to the IoT thermostat via a server-based program. However, this program failed during early morning, causing the server temperature

received by the thermostat to remain constant at 26°C, which did not reflect the actual indoor air temperature. As a result, the FCU could not be activated, even when the setpoint was lower than the actual indoor temperature, leading to a lack of cooling.

2. Occupancy data communication failure (October 14th, around 10:30 AM)

   The power plug supplying the occupancy sensor became unstable, disrupting the occupancy signal. At that moment, one person was present in the room, and the FCU was already running. The FCU remained on until the plug stabilized and the occupancy data was restored.

3. LoRa gateway breakdown (October 23th, afternoon)

   A breakdown of the LoRa gateway resulted in the loss of all LoRa-related data, including: environmental sensor data, thermostat data, occupancy sensor data, and cooling data.

These failures highlight the need for further enhancements to the proposed IoT-enabled implementation architecture to improve system robustness and reduce the likelihood of communication and program failures in the future.

6.5.2.3 Occupancy privacy concerns

The current system utilizes a camera-based occupancy sensor, which provides accurate real-time measurements of the number of occupants in the room. This data is crucial for modeling occupancy patterns and providing real-time occupant number data. However, this capability raises privacy concerns among users. While the smart sensor employs image-blurring algorithms to prevent facial recognition, privacy apprehensions persist. To address these concerns, future implementations will explore alternative occupancy measurement methods, such as estimating occupancy based on CO_2 levels.

6.5.2.4 Real-time occupant feedback limitations

Under the current system strategy, during a large meeting, if one participant submits a

comfort vote, it could influence the overall temperature settings, potentially leading to discomfort for other occupants. Although this issue has not yet arisen during testing, it is important to develop a refined control strategy that can balance individual comfort preferences while ensuring overall thermal comfort for all occupants. Future work will focus on formulating an adaptive strategy for modifying the reference temperature when conflicting comfort preferences are detected.

6.5.2.5 Scalability and applicability of the strategy in large-scale deployment

In Section 6.5.1, we analyze the costs of deploying the strategy in the target room. If the same IoT retrofit scheme is directly applied to different rooms in other buildings at scale, the expensive cooling meter may lower the applicability of the method. The cooling data is used to construct the SDE-based building thermal dynamic model, as described in Section 5.2.1 and Appendix B. However, if low-cost on/off state data is used as a substitute for constructing the model, substantial cost savings can be achieved, significantly enhancing the scalability and applicability of the method for large-scale deployment. However, one critical issue remains unresolved: to what extent the use of on/off state data as a substitute affects the accuracy of the SDE-based building thermal dynamic model. This thesis investigates this issue in Chapter 7, focusing on the scalability and applicability of the strategy for large-scale deployment by examining the trade-off between costs and model accuracy.

6.6 Summary

This section proposes an IoT-enabled implementation architecture for intelligent built environment management and implements the proposed architecture in a real office living lab. The architecture integrates advanced data-driven modeling methods to analyze stochastic occupancy patterns and building dynamics, alongside model predictive optimization. Optimized control setpoints are applied to the HVAC systems through the IoT infrastructure in the living lab.

The proposed architecture emphasizes the integration of real-time occupancy information and the uncertainties related to occupancy behaviors. Camera-based occupancy sensing technology was employed to gather occupant number data, while a QR code-based voting system was established to obtain real-time occupants' feedback. These real-time data, combined with stochastic occupancy model trained using historical data, are integrated into an OILT MPC control strategy for the optimal control of indoor air temperature and fresh air flow rate setpoints. The MPC dynamically adjusts temperature setpoints and fresh air flow rate setpoints, balancing conflicting objectives of thermal comfort, air quality and energy savings while considering varying weather condition and future occupancy uncertainties.

Evaluation over a one-month experimental test showed that the proposed MPC achieved more than 54.9% energy savings compared to the conventional control strategy. The MPC lowered indoor temperatures from 27.5 °C to 25 °C prior to occupant arrival through optimal precooling, effectively enhancing thermal comfort. It effectively maintained CO_2 levels at 800 ppm ± 40 ppm during occupied periods, preventing both over- and under-ventilation. The MPC could also promptly respond to the personalized comfort feedback through the real-time voting system.

This chapter contributes to the field by developing an MPC control strategy that effectively integrate occupancy information and the uncertainties of occupancy behaviors, as well as demonstrating in field implementation. The insights learned from the field implementation pave ways to future improvements in communication robustness, privacy-preserving, and cost-effectiveness of the IoT-enabled intelligent built environment management systems.

7 Development of IoT retrofit schemes for the scalable and applicable strategy deployment concerning trade-off between costs and model accuracy

7.1 Trade-off between costs and model accuracy

Prediction models form the fundamental basis for deploying MPC strategies. The IoT-enabled implementation architecture developed in Chapter 6 incorporates a comprehensive data collection scheme, utilizing various sensors and external meteorological services to construct multiple prediction models. However, achieving higher model accuracy inevitably requires more fine-grained data measurements, which often result in increased equipment and labor costs.

As discussed in Section 6.5.1, building thermal dynamics modeling relies on cooling measurement. While the cooling meters provide fine-grained measurements of cooling energy (Q_c), they account for the majority of retrofit costs, both in terms of equipment and labor. Installing cooling meters in every room significantly increases sensor costs and extends the payback period for MPC deployment, ultimately limiting the scalability of the proposed MPC system.

Alternatively, the on/off state ($\mathbb{1}_{Q_c}$) of the fan coil unit's water valve could be estimated using data from IoT thermostats, serving as a potential substitute for cooling meter data [117]. Leveraging on/off state data to develop SDE models can substantially reduce sensor costs, though it may have some negative impact on model accuracy.

Therefore, there is a trade-off between costs and model accuracy. Investigating these trade-offs is fundamental for the large-scale, cost-effective deployment of the MPC systems. This section proposes a the SDE-based building thermal dynamic modeling method using low-cost cooling substitute data. Instead of using expensive cooling meters, this study proposes a modeling method that utilizes relatively inexpensive on/off state data from the IoT thermostats to construct the SDE-based building thermal dynamic models. This research systematically compares both the modeling accuracy

and associated costs of three approaches: not using cooling data, using precise cooling meter data, and using on/off state substitute cooling data. The first approach, where no cooling data is used, serves as a baseline control group, which all other variables that could affect model accuracy are kept consistent with the second and third approaches to ensure a fair comparison. Furthermore, the marginal improvement in model accuracy and the marginal cost incurred by incorporating each type of data against the baseline are analyzed.

It should be noted that the measurement of cooling energy by cooling meters cannot be 100% accurate and is inevitably subject to errors and uncertainties. In this study, however, it is still assumed that the cooling energy measured by the cooling meters is accurate, as the meters have been calibrated and verified according to the specifications. The intrinsic measurement errors are not the main focus of this research.

7.2 SDE-based building thermal dynamic modelling using substitute data

The cooling meter data is used to construct the SDE-based building thermal dynamic model adopted in Chapter 6. The thermal dynamics of a room is described using an RC diagram with two time constant describing the dynamics of indoor air temperature and wall temperature. Three different approaches for handling cooling supply are considered in the modeling of the heat dynamics. Figure 7-1 shows the thermal dynamics structure for the room using different approaches. The equivalent SDE models of the indoor temperature ($T_{in,t}$) for the three approaches has the forms shown in Eq.(7-1), Eq.(7-2), Eq.(7-3) respectively. The equivalent SDE models of the wall temperature ($T_{w,t}$) for the three cases share the same form shown in Eq. (7-4).

For the SDE model 0, the cooling load term is not considered. For the SDE model 1, which is adopted in Chapter 6, the module $-Q_{c,t}$ is added in the Eq. (7-2), where $Q_{c,t}$ is obtained from actual measurements from the cooling meter. It contains finer information about the variation of the cooling supply, but the cost of obtaining the data

is high.

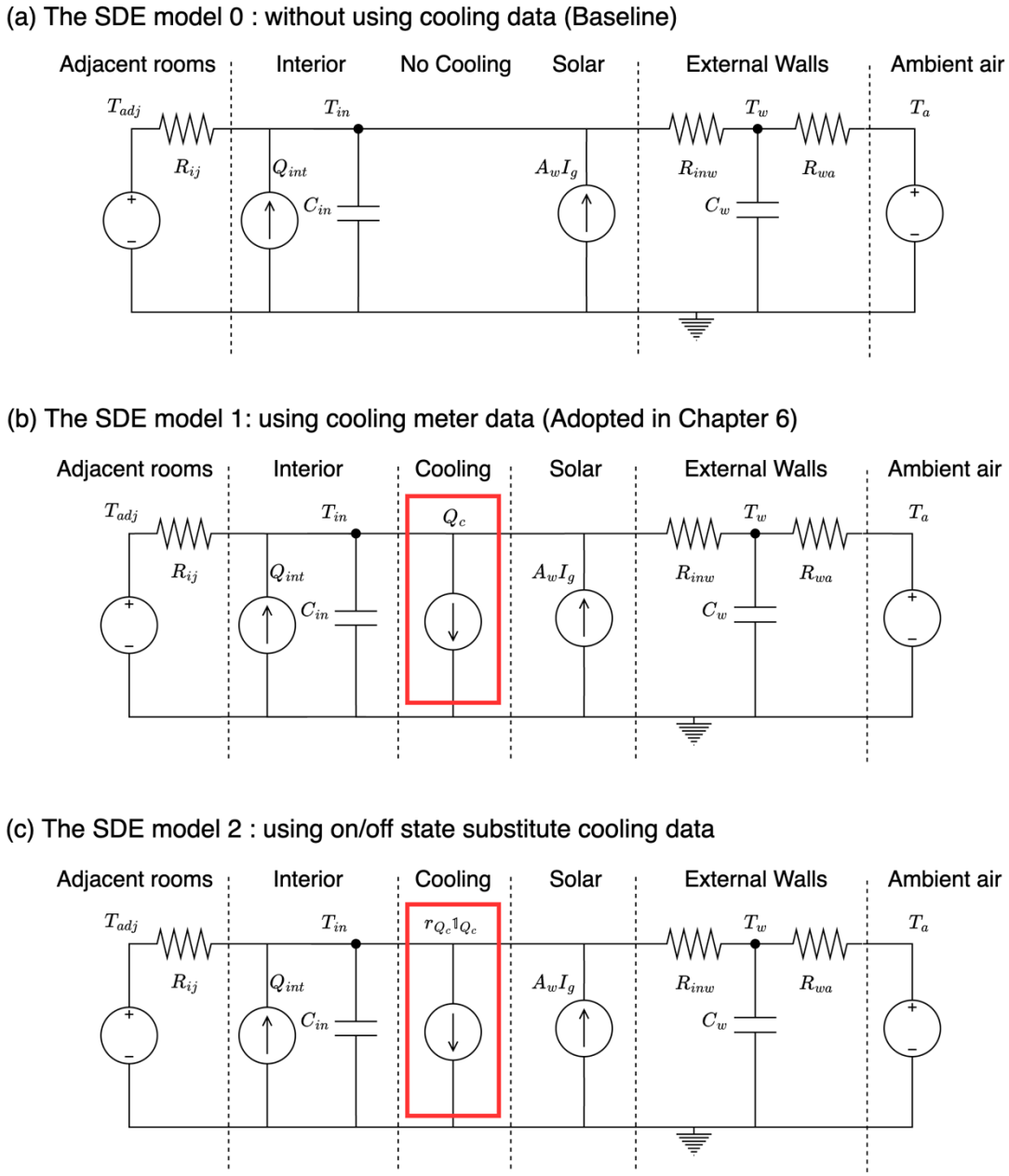


Figure 7-1 Illustration of SDE-based building thermal dynamic models using different cooling data (a) no cooling data, (b) cooling meter data, and (c) on/off state substitute cooling data

For the SDE model 2, the module $-r_{Q_c} \cdot \mathbb{1}_{Q_c,t}$ is added to the Eq. (7-3), where $\mathbb{1}_{Q_c,t}$ is the on/off state of the cooling valve, which is Boolean value, and can be obtained

through the IoT thermostat. The timing of the cooling supply can be determined, but the detailed variations in the cooling supply cannot be known. r_{Q_c} is a fixed parameter to be identified. This parameter ensures that the unit of cooling supply remains in kW. The interpretation of parameters in Eqs. (7-1)-(7-4) is listed in Table 7-1. The R-language *ctsmr* package [113] is utilized for SDE-based building dynamic modeling and system identification.

$$dT_{in,t} = \frac{1}{C_{in}}\left(\frac{1}{R_{inw}}\left(T_{w,t} - T_{in,t}\right) + \frac{1}{R_{ij}}\left(T_{adj} - T_{in,t}\right) + A_w I_{g,t} + r_{Q_{int}} \cdot o_t\right)dt + \sigma_1 d\omega_{in,t} \tag{7–1}$$

$$dT_{in,t} = \frac{1}{C_{in}}\left(\frac{1}{R_{inw}}\left(T_{w,t} - T_{in,t}\right) + \frac{1}{R_{ij}}\left(T_{adj} - T_{in,t}\right) + A_w I_{g,t} - Q_{c,t} + r_{Q_{int}} \cdot o_t\right)dt + \sigma_1 d\omega_{in,t} \tag{7–2}$$

$$dT_{in,t} = \frac{1}{C_{in}}\left(\frac{1}{R_{inw}}\left(T_{w,t} - T_{in,t}\right) + \frac{1}{R_{ij}}\left(T_{adj} - T_{in,t}\right) + A_w I_{g,t} - r_{Q_c} \cdot \mathbb{1}_{Q_c,t} + r_{Q_{int}} \cdot o_t\right)dt + \sigma_1 d\omega_{in,t} \tag{7–3}$$

$$dT_{w,t} = \frac{1}{C_w}\left(\frac{1}{R_{inw}}\left(T_{in,t} - T_{w,t}\right) + \frac{1}{R_{wa}}\left(T_{a,t} - T_{w,t}\right)\right)dt + \sigma_2 d\omega_{w,t} \tag{7–4}$$

Table 7-1 A list of the parameters in the SDEs in Eqs. (7-1)-(7-4) and their interpretations

Variable	Description	Unit	Type
$T_{in,t}$	The indoor air temperature	°C	state
$T_{w,t}$	The external wall temperature	°C	state
T_{adj}	The average temperature of adjacent rooms	°C	parameter

C_{in}	The heat capacity of the indoor air and part of inner wall and floor slab	kWh/ ºC	parameter
C_w	The heat capacity of the external walls	kWh/ ºC	parameter
R_{inw}	Thermal resistance between the indoor air and external walls	ºC/kW	parameter
R_{wa}	Thermal resistance between the external walls and outdoor air	ºC/kW	parameter
R_{ij}	Thermal resistance between the adjacent room and indoor air	ºC/kW	parameter
$T_{a,t}$	Outdoor air temperature	ºC	input
$\omega_{in,t}$	Brownian motion driving the indoor air temperature	-	-
$\omega_{w,t}$	Brownian motion driving the external wall temperature	-	-
σ_1	Scaling parameter to scale the Brownian motion of the indoor air temperature	ºC	parameter
σ_2	Scaling parameter to scale the Brownian motion of the wall temperature	ºC	parameter
A_w	Effective area of the solar radiation in the room	m^2	parameter
$I_{g,t}$	Solar radiation on a horizontal surface	kW/ m^2	input
$Q_{c,t}$	Cooling supply to the room air	kW	input
o_t	Occupant number	person	input
$r_{Q_{int}}$	Heat dissipate rate per occupant	kW/person	parameter
r_{Q_c}	The average value of cooling supply when cooling is active	kW	parameter
$\mathbb{1}_{Q_c,t}$	The on/off state of the cooling valve	-	input

7.3 The experimental data

This section describes the experimental data used to identify the parameters of the SDE models. The experimental setup is the same as described in Section 6.3. Figure 7-2 shows the experimental data in the period July 9 through July 21. The upper graph displays the cooling load ($Q_{c,t}$) and the on/off state ($\mathbb{1}_{Q_c,t}$) of the FCU in the room. The cooling supply varies from day to day. It seems to be characterized by a large peak whenever the cooling turns on, before return to a lower and steady level.

The second graph shows the indoor and outdoor air temperature. The indoor

temperature fluctuates between 24°C and 27°C. When cooling is supplied, the indoor temperature quickly drops to the setpoint of 24°C, and after the cooling stops, the temperature gradually rises. It can be observed that on July 14th, when the FCU was not turned on, the indoor temperature rose to nearly 30°C. On July 11th, due to a malfunction, cooling was continuously supplied, causing the indoor temperature to drop to nearly 18°C. During this period, the outdoor temperature fluctuated between 28°C and 34°C with significant variations.

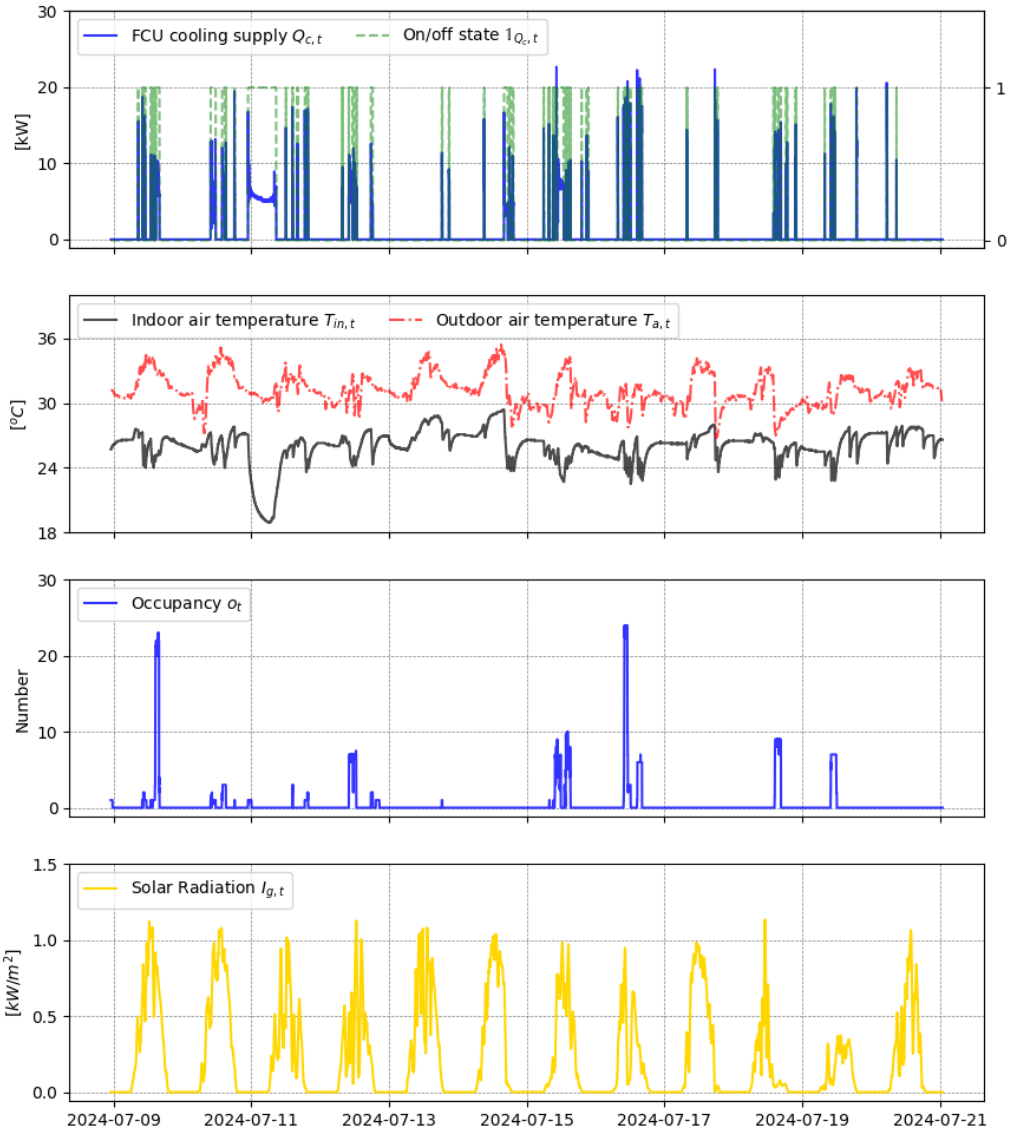


Figure 7-2 The data from the experiment performed in the period July 9 through July 21

The third graph shows the occupant number. As mentioned earlier, the number of people in this meeting room changes very randomly. Most of the time, the room is unoccupied, while during a few periods, a large number of people enter or leave the room.

The fourth graph shows the solar radiation. The fourth graph shows the solar radiation. The daily variation in solar radiation differs significantly. For example, on sunny days, the fluctuations are relatively small, while on days with cloudy weather or short periods of rain, such as July 11th and July 12th, the fluctuations in solar radiation are larger. On July 19th, which was overcast all day, the solar radiation remained relatively low.

7.4 Results

This section details the results of parameter estimates, residual analysis, and prediction accuracy of the models.

7.4.1 Model parameters

Table 7-2 shows the results of parameter estimation of the model in Section 7.2.

Table 7-2 Identified parameters of the three SDE-based building dynamic models

Parameter	Unit	Estimation		
		Model 0	**Model 1**	**Model 2**
T_{adj}	°C	25.4	25.8	25.4
C_{in}	kWh/ °C	0.4	1.78	1.06
C_w	kWh/ °C	1.56	0.50	0.36
R_{inw}	°C/kW	1.93	0.56	0.93
R_{wa}	°C/kW	9.85	8.77	6.11
R_{ij}	°C/kW	1.01	1.14	1.21
A_w	m^2	2.08	2.245	1.88
$r_{Q_{int}}$	kW/person	0.0012	0.167	0.09
r_{Q_c}	kW	-	-	5.32

From the identified results, it can be seen that the C_{in} and R_{inw} identified by Model 0 are highly skewed. This is due to the lack of cooling data as a constraint, making accurate identification impossible. The parameters of Model 2 are relatively closer to

those of Model 1, although C_{in} and R_{inw} still exhibit deviations.

7.4.2 Residuals analysis

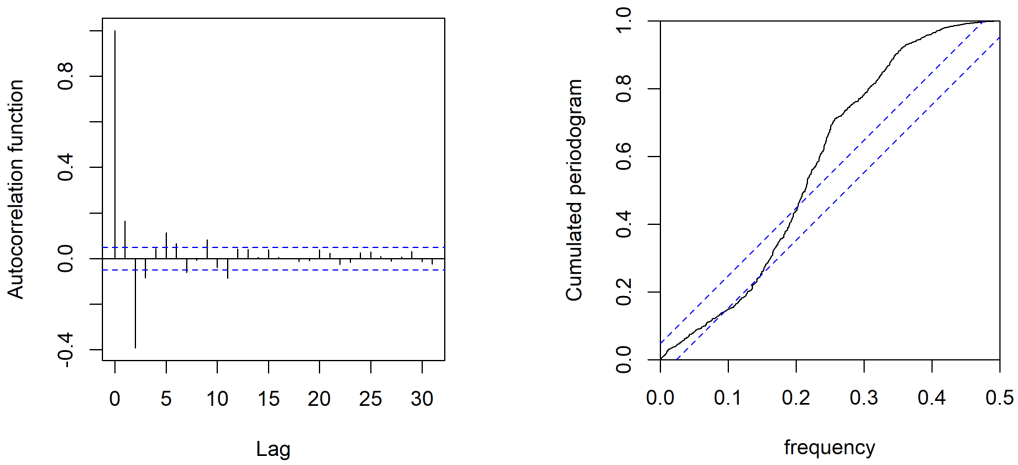
Figure 7-3 illustrates the autocorrelation function (ACF, left panel) and the cumulative periodogram (CPG, right panel) plots of 1-step prediction residuals for three SDE models. The theoretical reference line with 95% confidence intervals (blue dashed) is included for comparison in both panels.

The ACF measures how strongly a time series is correlated with itself at different lags. If the autocorrelations remain close to 0 across all lags, this suggests that the residuals may resemble white noise, indicating the developed SDE model effectively captures the temporal patterns in the data, leaving no detectable structured errors. Conversely, significant autocorrelations at certain lags, i.e. exceeding the confidence interval, indicate the presence of trends or seasonality, suggesting that the residuals deviate from white noise, potentially pointing to unaccounted temporal dependencies or an oversimplification of the underlying dynamics in the SDE model.

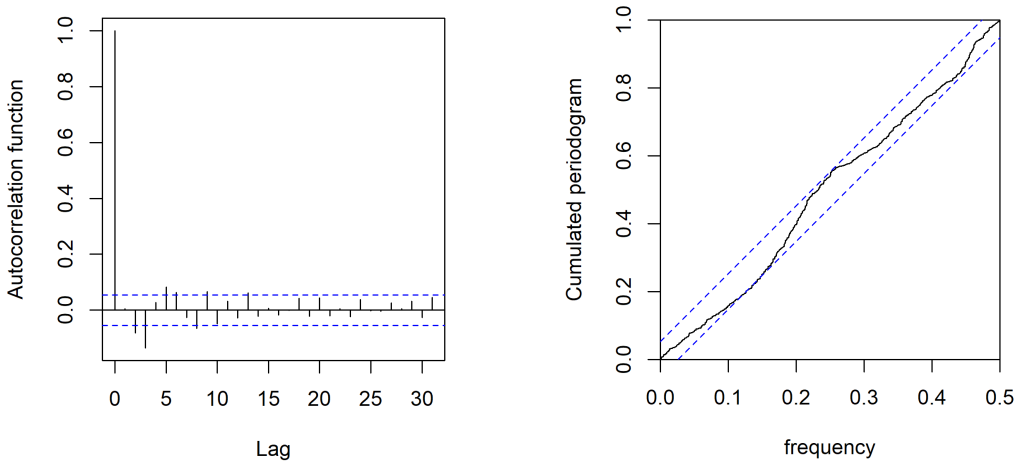
The CPG is another method used to test whether a time series can be considered white noise. Unlike the ACF, which operates in the time domain, the CPG analyzes the frequency domain, focusing on the distribution of spectral power across frequencies. The CPG is particularly effective at identifying periodicity or seasonal components, which appear as non-uniform spectral power, and can detect frequency-domain features that may not be easily visible in the time domain. If the CPG lies close to the theoretical reference line (or within its confidence intervals), the time series can be considered white noise. However, deviations from the reference line at specific frequencies indicate periodicity, seasonal components, or structured patterns.

While the ACF and CPG evaluate white noise through different approaches, they are complementary tools. By leveraging both methods, this study provides a more robust and comprehensive analysis of the residuals.

(a) *Model 0*



(b) *Model 1*



(c) *Model 2*

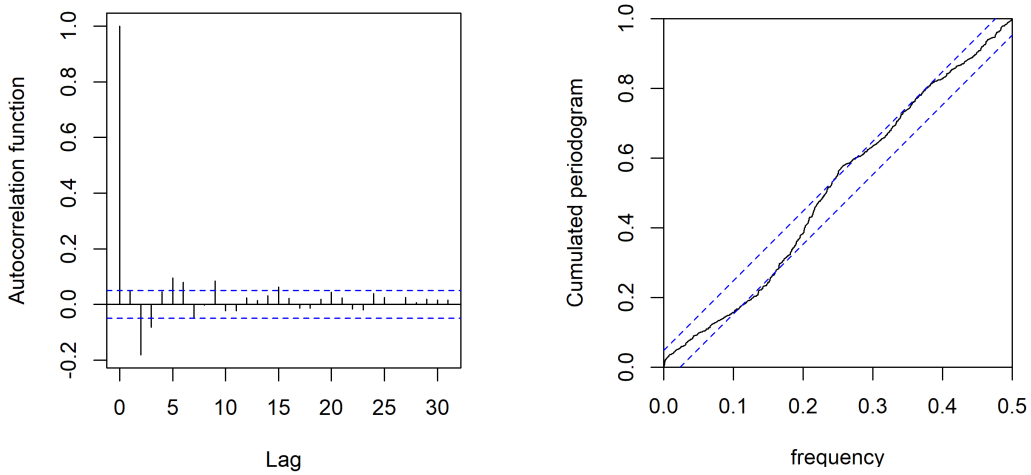


Figure 7-3 The ACF and the CPG of the prediction residuals of the three SDE models

In contrast, Model 1 demonstrates very small autocorrelations across all lags, with only minor exceedances of the confidence intervals at the third and fifth lags. This suggests that Model 1, which incorporates data reflecting almost complete cooling supply dynamics, produces residuals that can be considered white noise. Furthermore, its CPG closely follows the theoretical line across all frequencies, with only minor deviations in the very low-frequency range (0.1–0.2). This strong alignment with the uniform distribution confirms that the residuals are close to white noise. Therefore, Model 1 effectively captures the underlying dynamics.

For Model 2, the autocorrelation at the second lag slightly exceeds the confidence level, with a value around -0.2. While this is higher than that of Model 1, it remains much smaller than that of Model 0. The CPG for Model 2 shows an upward trend in the low-frequency range, with values slightly exceeding the confidence level around frequencies 0.15 and 0.25, indicating mild residual periodicity at these frequencies. This suggests that while Model 2, which uses on/off state cooling substitute data, lacks some cooling-related dynamic characteristics, the residuals can still be approximately considered as white noise. Consequently, Model 2 provides a reasonably good representation of the underlying dynamics.

7.4.3 Analysis on prediction accuracy and marginal costs

Figure 7-4 illustrates the RMSE results for indoor temperature predictions at various time horizons (N-minutes ahead) for three building thermal dynamic models. These include short-term predictions, such as 10 or 20 minutes ahead, as well as long-term predictions, such as 5 hours ahead. As the prediction horizon increases, the RMSE of all three models rises. However, the slope of the RMSE growth decreases over time, eventually stabilizing and flattening out.

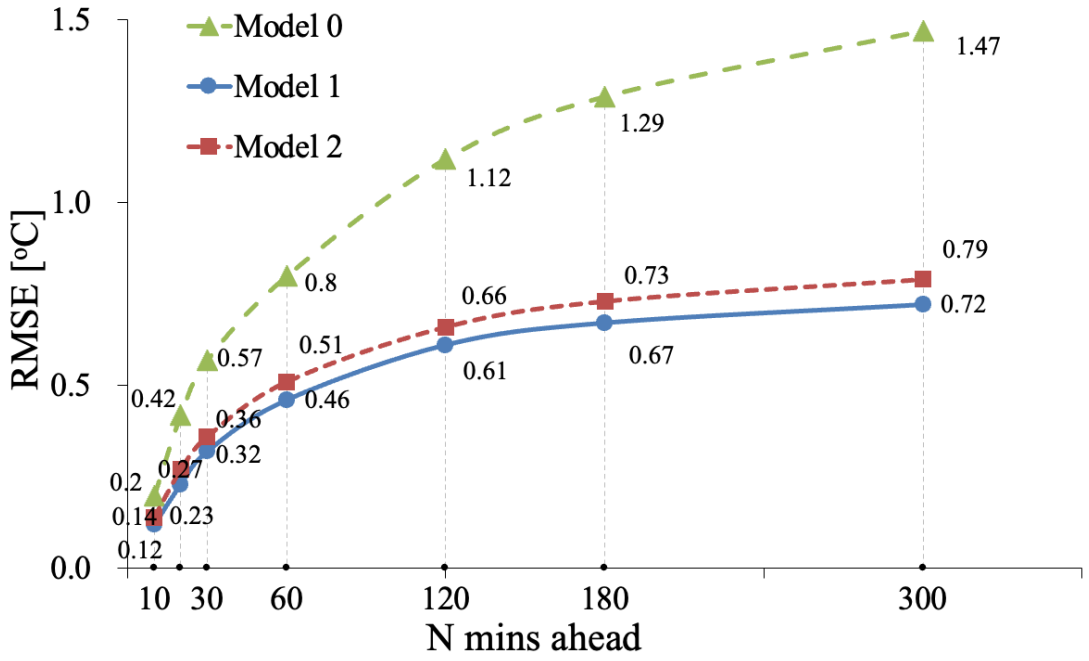


Figure 7-4 RMSE of N-minutes ahead prediction of the indoor air temperature of the three SDE models

The RMSE of Model 2 is slightly higher than that of Model 1 across all prediction horizons, indicating that Model 2 is marginally less accurate. However, Model 2 only relies on the more affordable on/off state cooling substitute data. Compared to Model 0, the baseline model that does not account for cooling data, the marginal cost of Model 2 is around 1500 HKD (IoT thermostat 500 HKD, installation around 1000 HKD) as shown in Table 6.6., while it achieves significant improvements in accuracy.

Model 1 consistently achieves the lowest RMSE, delivering the most accurate predictions, it does so by utilizing expensive yet precise cooling meter data. Compared to Model 0, the marginal cost of Model 1 is around 15100 HKD (Cooling meter 10100 HKD, installation around 5000 HKD), more than 10 times of the marginal cost of Model 2. However, the marginal improvement of Model 1 over Model 2 is not substantial.

Take the 60-minute-ahead prediction as an example. Model 0 has an RMSE of 0.8°C. By utilizing the cooling substitute data, Model 2 reduces the RMSE to 0.51°C, achieving a 36.25% improvement in accuracy. Further incorporating precise cooling meter data, Model 1 lowers the RMSE to 0.46°C, representing a 42.5% accuracy

improvement over Model 0. However, this is only 6.25% better than Model 2, a marginal improvement given the much higher cost.

7.5 Summary

This study proposes an SDE-based building thermal dynamic modeling method using low-cost cooling substitute data, which facilitates the cost-effective deployment of MPC systems. The research delves into the trade-off between costs and model accuracy for SDE prediction models used in MPC, by constructing and analyzing three SDE models—Model 0 (no cooling data), Model 1 (high-precision cooling meter data), and Model 2 (low-cost on/off valve state substitute data).

The results demonstrate that the SDE-based indoor thermal dynamic model using low-cost cooling substitute data significantly improves prediction accuracy compared to models that do not incorporate cooling data. However, further utilizing high-precision cooling meter data yields only marginal improvements in prediction accuracy over the model that uses low-cost substitute data, despite the substantially higher costs. The key findings are summarized as follows:

(1) Residual analysis using the ACF and the CPG indicates that the residuals of Model 2 can be approximately considered as white noise, suggesting that it reasonably captures the underlying dynamics.

(2) Model 2 reduces the RMSE to 0.51°C, achieving a 36.25% improvement in accuracy over the baseline model. Incorporating precise cooling meter data, Model 1 further lowers the RMSE to 0.46°C, representing only a 6.25% further improvement over Model 2. This marginal accuracy gain comes at a significantly higher cost, underscoring the cost-effectiveness of Model 2.

8 Conclusions and future research

8.1 Conclusions

This thesis develops and validates advanced stochastic modeling and predictive control strategies for intelligent built environment management, addressing the challenges posed by irregular and uncertain occupancy patterns. The Adaptive B-Spline-based IMC method introduced in this research offers a novel approach to occupancy modeling by dynamically adjusting knot density to capture occupancy variations in different types of spaces. This method achieves significantly improved prediction accuracy with fewer model parameters compared to conventional uniform B-Spline and counting methods. By automatically increasing knot density in areas of high occupancy variation and reducing it in areas of low variation, the Adaptive B-Spline method provides a tailored solution for diverse occupancy scenarios.

The CTIMC model with band structure introduces a physically interpretable and computationally efficient framework for continuous-time occupancy forecasting. The band structure ensures that transitions between occupancy states are consistent with physical constraints, while the linear scaling of model complexity with the number of states reduces computational costs. The CTIMC model outperforms traditional discrete-time Markov Chain models in both accuracy and generalization, as demonstrated by its ability to capture intrinsic occupancy patterns rather than overfitting to noise in training data.

The integration of these models into stochastic occupancy-integrated MPC frameworks demonstrates their practical utility in optimizing building energy management. By leveraging accurate occupancy forecasts, the MPC effectively balances thermal comfort, IAQ, and energy efficiency. The incorporation of IoT-enabled occupant feedback systems further enhances personalization and adaptability, allowing real-time adjustments based on occupant preferences. Field testing in a living lab environment confirms the effectiveness of these methods, achieving substantial energy savings, improved thermal comfort, and better IAQ compared to baseline control strategies.

Overall, this thesis provides a comprehensive solution to the challenges of managing dynamic indoor environments, combining advanced modeling, predictive optimization, and real-time interaction to achieve sustainable and occupant-centric building operations.

8.2 Summary of new contributions

This thesis makes several significant contributions to the field of intelligent occupant-centric built environment management as follows.

First, this research introduces the Adaptive B-Spline-based IMC method, which dynamically adjusts knot density to effectively capture diverse occupancy patterns across different spaces. This method achieves higher prediction accuracy with fewer model parameters, providing a scalable and efficient solution for occupancy forecasting in scenarios with varying degrees of occupancy dynamics.

Second, this research proposes the CTIMC model with band structure for the first time, offering a physically interpretable and computationally efficient framework for continuous-time occupancy forecasting. By incorporating physical constraints into the transition rate matrix and reducing model complexity from quadratic to linear scaling, the CTIMC model enhances generalization capabilities and makes continuous-time modeling feasible for large state spaces.

Third, this research proposes the stochastic occupancy-integrated MPC strategy, leveraging the IMC-based occupancy predictions to precondition indoor environments, and balancing multiple conflicting control objectives of thermal comfort and IAQ, and energy consumption.

Fourth, the research proposes an IoT-enabled architecture incorporating advanced data-driven modeling and predictive optimization for intelligent built environment management, and realizes the real-world implementation of the IoT-enabled OITL MPC strategy with active occupancy participation.

Finally, the research proposes a data-driven modeling method using low-cost data,

balancing costs and model accuracy, which enhances the scalability and applicability for strategy deployment in diverse building spaces.

8.3 Summary of limitations and recommendations for future research

8.3.1 Limitations

The current research is limited to office meeting rooms in tropical climates (e.g., Hong Kong) with cooling and ventilation systems. Its applicability to diverse building types (e.g., residential, commercial, schools), systems (e.g., lighting, shading), and climate conditions (e.g., cold regions, mixed climates) remains unexplored.

The TRNSYS simulation study assumes uniform occupant activities and metabolic rates typical of office settings, without accounting for variability during non-standard uses or exceptional scenarios. Future work could address these limitations by expanding to diverse conditions, incorporating comprehensive environmental factors, and accommodating variable occupant activities to enhance robustness and generalizability.

In the real-world tests, the communication and program failures, such as server temperature sending issues, occupancy sensor disruptions, and LoRa gateway breakdowns, highlight the need for improved system robustness. Besides, the use of camera-based occupancy sensors raises privacy concerns despite employing image-blurring algorithms, necessitating exploration of alternative methods like CO_2-based occupancy estimation.

8.3.2 Recommendations for future research

(1) Stochastic occupancy modelling incorporating external inputs

The proposed in-homogeneous model uses only time as external input. Even though it supplies significant improvements over the time-homogeneous model, expanding the model to depend also on external inputs/signals could be a natural next step for improving it. The meeting room in the present study is coupled with a booking system

where employees can book the room for a future meeting. Using this could be a significant indicator of when the meeting room is occupied, and at the least where occupation begins. The booking system is however only able to tell if a meeting is scheduled or not (and not how many people are going to attend). The weather is another candidate for being an explanatory variable, however, it may not have a significant effect since meetings inside an office building may not be too affected by the weather.

(2) Investigating the impact of different occupancy prediction methods on control performance

In practice, the two methods developed in Chapter 3 and Chapter 4 are suitable for deployment under different scenarios. The adaptive B-spline method proposed in Chapter 3 is more appropriate for occupancy prediction in environments with high uncertainty and rapid changes, while the band structure-based method introduced in Chapter 4 is better suited for situations with anomalous occupancy measurements. The band structure method effectively enhances the generalization capability of the model predictions. In future work, we will further validate the robustness of the band structure-based MPC control strategy in the presence of sensor anomalies through both simulation and real-world tests.

(3) Long term testing of the MPC in diverse environments

The testing period in this study was limited to one month, which does not fully capture the seasonal variations in Hong Kong. To gain a comprehensive understanding of system performance, future research will conduct extended testing over an entire year to assess the MPC strategy's effectiveness across different climatic conditions. We will also expand testing beyond meeting rooms to examine different occupancy patterns in large open-plan offices, small private offices, classrooms, and lecture halls. By implementing the MPC strategy in diverse indoor environments, future studies will enable a more comprehensive comparison with baseline control strategies, yielding more accurate energy savings data across various space types.

(4) Improving IoT communication robustness

The current IoT communication infrastructure occasionally experiences failures or sensor malfunctions, leading to disruptions in data collection. Additionally, communication delays are sometimes observed, which can affect real-time decision-making and system performance. These challenges highlight the need for more robust and reliable communication solutions. In the future, the architecture of the LoRa communication infrastructure will be further optimized to address these issues. Enhancements may include implementing redundancy mechanisms to ensure data transmission continuity during failures, improving gateway placement to maximize coverage and signal strength, and adopting advanced protocols to minimize latency. Moreover, integrating error detection and correction algorithms will help mitigate the effects of data loss or corruption. By optimizing LoRa's communication infrastructure, the system's overall reliability, efficiency, and resilience can be significantly improved, ensuring uninterrupted IoT operations and better support for applications like indoor environment control and energy optimization.

(5) Extended IoT architecture incorporating grids, renewables and EVs

The IoT-enabled implementation architecture proposed in this study primarily targets indoor environment control, aiming to achieve energy savings while improving indoor comfort and air quality. In the future, this architecture can be further expanded to integrate grid signals, renewable energy, and electric vehicles (EVs). Grid signals, such as dynamic electricity pricing and dynamic carbon emission factors [118], can be incorporated to enable demand response based on grid signals, thereby reducing electricity costs and carbon emissions. Renewable energy sources (e.g., solar and wind) and batteries can also be integrated into the architecture. By optimizing battery charging/discharging and indoor HVAC systems, the utilization of renewable energy on the building side can be maximized. Additionally, the architecture can take into account people's mobility patterns to optimize the charging and discharging demands of EVs.

(6) Developing a multi-feedback response system for handling different votes with diverse preferences

While the current thesis focuses on a response system for single-feedback scenarios, it is important to acknowledge that individuals may have different thermal preferences even under the same indoor temperature. To better address this diversity, future research will explore the development of a multi-feedback response system. This could involve integrating various user feedback through methods such as calculating a weighted average of preferred temperatures, where weights may be assigned based on user-specific priorities, to determine a new comfort temperature that accommodates multiple preferences.

Appendices

A. Specification of IoT retrofit, devices and external services

Table A1 Specification of IoT retrofit, devices and external services

	Devices	Installation time	Brand	Type	Communication Protocol	Up/Dn cycle	Measurement	Unit
a	IoT indoor environment sensors	2023/3/16	Milesight	AM307	LoRaWAN	2 min	Temperature	°C
							Relative humidity	%
							Motion	
							Light	Lux
							TVOC	
							Barometric Pressure	hPa
							CO_2	ppm
b	IoT occupancy sensor	2023/3/16	Milesight	VS121	LoRaWAN	2 min	Regional people counting	
c	QR code for occupants' feedback	2024/1/18	CLIMIFY		HTTP API	Event	Thermal sensation	
d	IoT thermostat	2024/6/6	Milesight	WT301	LoRaWAN	2 min	Room temperature	°C
							Temperature setpoint	°C
							Fan speed	level
e	Cooling meter	2024/6/6	Midea	MUFR-2000A-G-01-SW-G	Modbus	2 min	Cooling energy	kWh
							Chilled water flow rate	kg/h
							Supply temperature	°C
							Return temperature	°C
	LoRa converter	2024/6/6	Milesight	UC100	Modbus to LoRaWAN	2 min	-	-

f	LoRaWAN gateway	2023/3/16	Milesight	UG65	LoRaWAN, MQTT		-	-
g	BAS gateway	2024/4/25	Johnson Controls		Bacnet		-	-
	Legacy BAS		Johnson Controls				PAU supply air temperature	°C
							PAU fan Frequency	hz
							PAU supply Air pressure	kPa
							VAV flow rate	kg/h
							VAV flow rate setpoint	kg/h
							VAV damper	%
							VAV CO_2	ppm
h	External services Real-time weather data		Hong Kong observatory		HTTP API	10mins	Outdoor air temperature	°C
							Outdoor relative humidity	%
							Global solar radiation	kW/m^2
							Direct solar radiation	kW/m^2
							Diffuse solar radiation	kW/m^2
							Barometric Pressure	hPa
							Wind speed	m/s
							Wind direction	
i	Weather prediction		OpenWeather	3-day hourly prediction	HTTP API	30mins	Outdoor air temperature	°C
							Outdoor relative humidity	%

	Barometric Pressure	hPa
	Wind speed	m/s
	Wind direction	
	UVI	mW/m^2
	Clouds	%
	Probability of precipitation	%

B. System identification of the SDE-based building dynamics models

Fig.B1 shows the data used for system identification of the SDE-based building dynamic models (Eqs. 5-7), which includes the state and input variables of indoor air temperature ($T_{in,t}$), outdoor air temperature ($T_{a,t}$), FCU cooling supply ($Q_{c,t}$), occupant number (o_t) and solar radiation ($I_{g,t}$) for thermal dynamic modelling; and indoor CO_2 concentration ($c_{CO_2,in,t}$) and fresh air flow rate ($F_{a,t}$) for CO_2 dynamic modelling. The *ctsmr* R-language package [113] is employed for SDE-based building dynamic modeling and system identification. Table B1 shows the identified parameters of the thermal dynamic model and CO_2 dynamic model.

Fig. B2 shows the autocorrelation function (ACF) and the cumulated periodogram (CPG) of the prediction residuals of (a) building thermal dynamic model and (b) CO_2 dynamic model. The autocorrelation function and the cumulated periodogram of the building thermal dynamic model indicate that the residuals of the indoor air temperature can be classified as white noise. However, the CO_2 residuals are governed by some minor autocorrelation in the first few lags, while this minor autocorrelation is not impact MPC performance significantly.

Fig. B3 shows the RMSE results of N-minutes ahead prediction of the building thermal dynamic model and CO_2 dynamic model, which includes short-term prediction, such as 10 or 20 minutes ahead, and long-term prediction, such as 5 hours ahead. With the increase of number of minutes ahead prediction, that RMSE also rises. However, the

slop of RMSE is decrease and gradually become stable and flat. The RMSE results show that both of the accuracy of temperature and CO_2 prediction is very high. The RMSE of prediction in short-term, which less than 1 hour head are less than 0.5 ºC and 53ppm respectively.

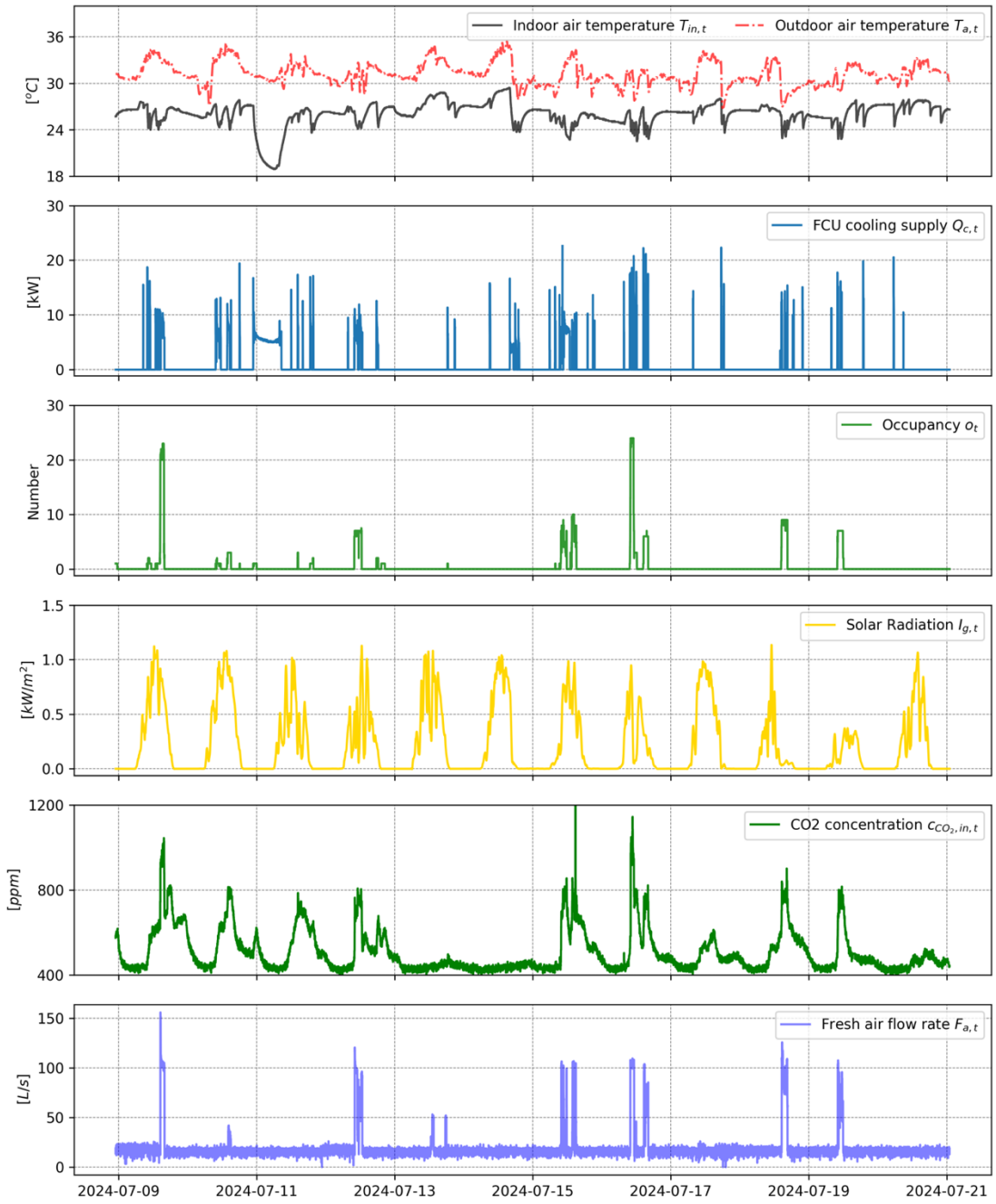
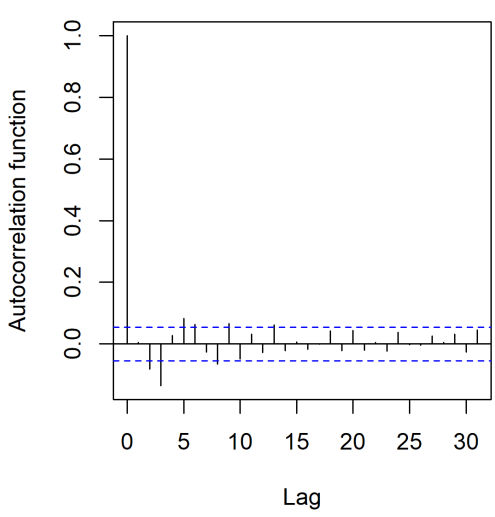


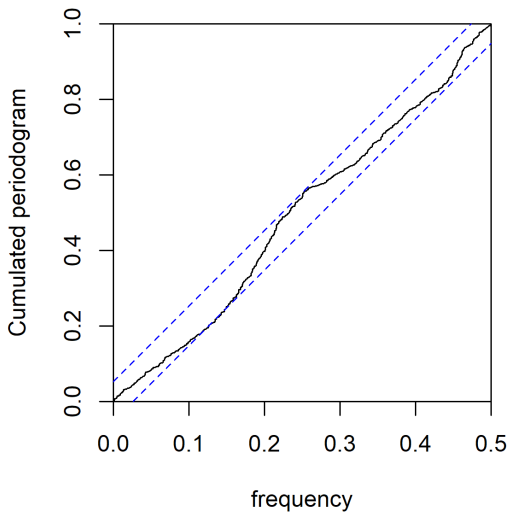
Fig. B1 Data used for system identification of the SDE-based building dynamic models (Eqs. 5-7)

Table B1 Identified parameters of SDE-based building dynamic models

Parameter	Unit	Estimate
T_{adj}	°C	25.8
C_{in}	kWh/ °C	1.78
C_w	kWh/ °C	0.50
R_{inw}	°C/kW	0.56
R_{wa}	°C/kW	8.77
R_{ij}	°C/kW	1.14
σ_1	°C	5.03E-10
σ_2	°C	1.68E-5
A_w	m^2	2.245
$r_{Q_{int}}$	kW/person	0.167
$c_{CO_2,out,t}$	ppm	457
V	m^3	170
F_{inf}	$1/h$	0.23
r_{CO_2}	m^3/h/person	0.0
σ_3	ppm	2.3E-6

(a) *Building thermal dynamic model*





(b) *Building CO2 dynamic model*

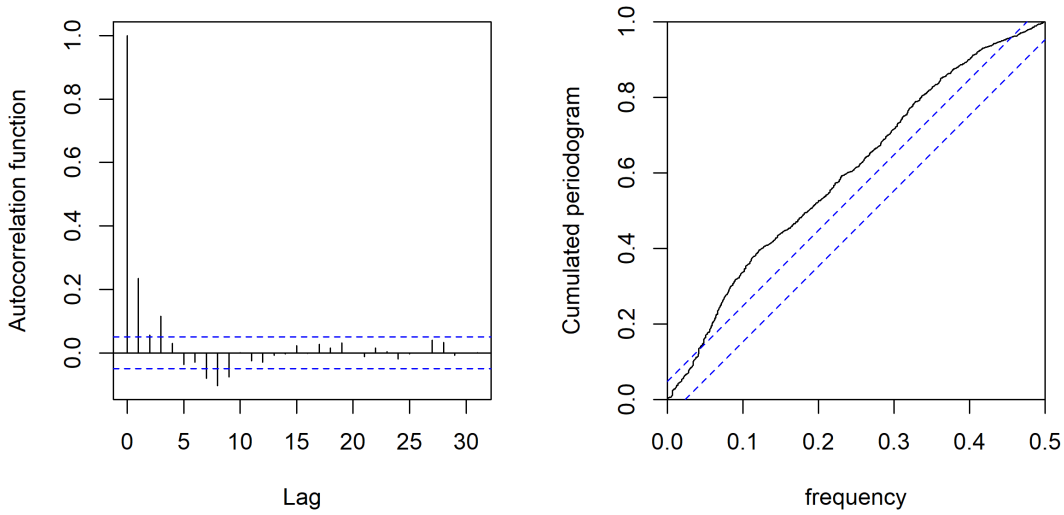


Fig. B2 The ACF and the CPG of the prediction residuals of (a) building thermal dynamic model and (b) CO2 dynamic model.

(a) Building thermal dynamic model *(b) Building CO2 dynamic model*

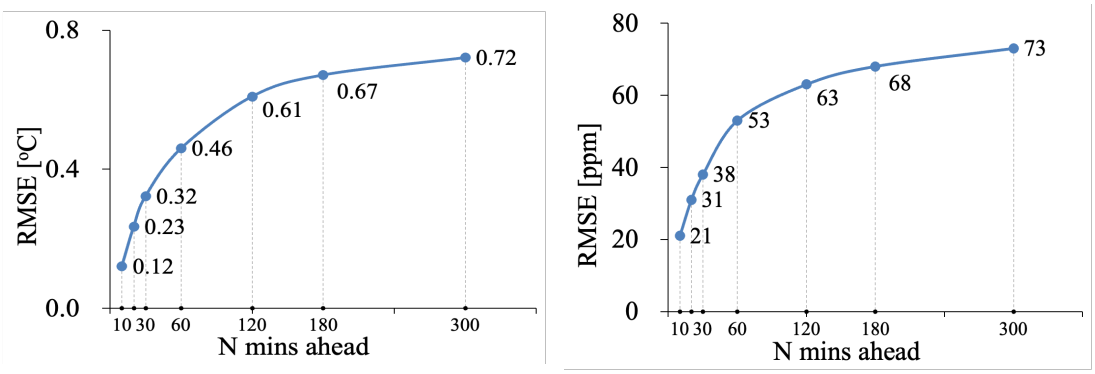


Fig. B3 RMSE of N-minutes ahead prediction of (a) building thermal dynamic model and (b) CO2 dynamic model.

References

[1] Huang X, Liu Y, Huang L, Onstein E, Merschbrock C. BIM and IoT data fusion: The data process model perspective. Automation in Construction 2023;149:104792. https://doi.org/10.1016/j.autcon.2023.104792.

[2] IoT Devices in Smart Commercial Buildings 2023 to 2028 n.d. https://memoori.com/portfolio/iot-devices-smart-commercial-buildings-2023/ (accessed May 9, 2025).

[3] Jia M, Komeily A, Wang Y, Srinivasan RS. Adopting Internet of Things for the development of smart buildings: A review of enabling technologies and applications. Automation in Construction 2019;101:111–26. https://doi.org/10.1016/j.autcon.2019.01.023.

[4] Moudgil V, Hewage K, Hussain SA, Sadiq R. Integration of IoT in building energy infrastructure: A critical review on challenges and solutions. Renewable and Sustainable Energy Reviews 2023;174:113121. https://doi.org/10.1016/j.rser.2022.113121.

[5] IEA. Buildings - Energy System. IEA 2023. https://www.iea.org/energy-system/buildings (accessed February 23, 2025).

[6] Electrical and Mechanical Services Department of Hong Kong. Hong Kong Energy End-use Data 2023. 2023.

[7] Pachouri V, Singh R, Gehlot A, Pandey S, Vaseem Akram S, Abbas M. Empowering sustainability in the built environment: A technological Lens on industry 4.0 Enablers. Technology in Society 2024;76:102427. https://doi.org/10.1016/j.techsoc.2023.102427.

[8] U.S. Energy Information Administration. INTERNATIONAL ENERGY OUTLOOK 2021. 2021.

[9] Yan D, O'Brien W, Hong T, Feng X, Burak Gunay H, Tahmasebi F, et al. Occupant behavior modeling for building performance simulation: Current state and future challenges. Energy and Buildings 2015;107:264–78. https://doi.org/10.1016/j.enbuild.2015.08.032.

[10] Meyers RJ, Williams ED, Matthews HS. Scoping the potential of monitoring and control technologies to reduce energy use in homes. Energy and Buildings 2010;42:563–9. https://doi.org/10.1016/j.enbuild.2009.10.026.

[11] Naylor S, Gillott M, Lau T. A review of occupant-centric building control strategies to reduce building energy use. Renewable and Sustainable Energy Reviews 2018;96:1–10. https://doi.org/10.1016/j.rser.2018.07.019.

[12] Esrafilian-Najafabadi M, Haghighat F. Occupancy-based HVAC control systems in buildings: A state-of-the-art review. Building and Environment 2021;197:107810. https://doi.org/10.1016/j.buildenv.2021.107810.

[13] J.R P. The Daunting Climate Change: Science, Impacts, Adaptation & Mitigation Strategies, Policy Responses. CRC Press; 2022.

[14] Bierwirth A, Thomas S. Energy sufficiency in buildings n.d.

[15] Fournier ED, Cudd R, Federico F, Pincetl S. On energy sufficiency and the need for new policies to combat growing inequities in the residential energy sector. Elem Sci Anth 2020;8:24. https://doi.org/10.1525/elementa.419.

[16] Darby S, Fawcett T. Energy sufficiency – an introduction: A concept paper for ECEEE 2018. https://doi.org/10.13140/RG.2.2.31198.08006.

[17] Peng RD, Bobb JF, Tebaldi C, McDaniel L, Bell ML, Dominici F. Toward a Quantitative Estimate of Future Heat Wave Mortality under Global Climate Change. Environ Health Perspect 2011;119:701–6. https://doi.org/10.1289/ehp.1002430.

[18] Hu S, Zhou X, Yan D, Guo F, Hong T, Jiang Y. A systematic review of building energy sufficiency towards energy and climate targets. Renewable and Sustainable Energy Reviews 2023;181:113316. https://doi.org/10.1016/j.rser.2023.113316.

[19] Chiou Y-S, Carley KM, Davidson CI, Johnson MP. A high spatial resolution residential energy model based on American Time Use Survey data and the bootstrap sampling method. Energy and Buildings 2011;43:3528–38. https://doi.org/10.1016/j.enbuild.2011.09.020.

[20] China National Bureau of Statistics. Communiqué on China's Third National

Time Use Survey (No. 2). 2024.

[21] Buttitta G, Turner W, Finn D. Clustering of Household Occupancy Profiles for Archetype Building Models. Energy Procedia 2017;111:161–70. https://doi.org/10.1016/j.egypro.2017.03.018.

[22] Aerts D, Minnen J, Glorieux I, Wouters I, Descamps F. Discrete occupancy profiles from time-use data for user behaviour modelling in homes, 2013, p. 2421–7.

[23] Shan H. Investigation and analysis of Chinese residential building occupancy with large-scale questionnaire surveys 2019.

[24] Wang Z. Individual difference in thermal comfort_ A literature review. Building and Environment 2018.

[25] Wiley J. Thermal comfort and gender: a literature review. Data Collection n.d.

[26] Brager GS, De Dear RJ. Thermal adaptation in the built environment: a literature review. Energy and Buildings 1998;27:83–96. https://doi.org/10.1016/S0378-7788(97)00053-4.

[27] Jiang Z, Deng Z, Wang X, Dong B. PANDEMIC: Occupancy driven predictive ventilation control to minimize energy consumption and infection risk. Applied Energy 2023;334:120676. https://doi.org/10.1016/j.apenergy.2023.120676.

[28] Li Z, Dong B. A new modeling approach for short-term prediction of occupancy in residential buildings. Building and Environment 2017;121:277–90. https://doi.org/10.1016/j.buildenv.2017.05.005.

[29] Jiang Z, O'Neill Z, Dong B. OCCUPIED: Long-term field experiment results from an occupant-centric control in an office building. Energy and Buildings 2023;297:113435. https://doi.org/10.1016/j.enbuild.2023.113435.

[30] Kanthila C, Boodi A, Beddiar K, Amirat Y, Benbouzid M. Markov Chain-based Algorithms for Building Occupancy Modeling: A Review. 2021 3rd International Conference on Smart Power & Internet Energy Systems (SPIES), 2021, p. 438–43. https://doi.org/10.1109/SPIES52282.2021.9633933.

[31] Liisberg J, Møller JK, Bloem H, Cipriano J, Mor G, Madsen H. Hidden Markov Models for indirect classification of occupant behaviour. Sustainable Cities

and Society 2016;27:83–98. https://doi.org/10.1016/j.scs.2016.07.001.

[32] Ding Y, Ma X, Wei S, Chen W. A prediction model coupling occupant lighting and shading behaviors in private offices. Energy and Buildings 2020;216:109939. https://doi.org/10.1016/j.enbuild.2020.109939.

[33] Zhou X, Liu T, Yan D, Shi X, Jin X. An action-based Markov chain modeling approach for predicting the window operating behavior in office spaces. Build Simul 2021;14:301–15. https://doi.org/10.1007/s12273-020-0647-9.

[34] Andersen PD, Iversen A, Madsen H, Rode C. Dynamic modeling of presence of occupants using inhomogeneous Markov chains. Energy and Buildings 2014;69:213–23. https://doi.org/10.1016/j.enbuild.2013.10.001.

[35] Page J, Robinson D, Morel N, Scartezzini J-L. A generalised stochastic model for the simulation of occupant presence. Energy and Buildings 2008;40:83–98. https://doi.org/10.1016/j.enbuild.2007.01.018.

[36] Sanchez J. Hidden Markov Models for Time Series - An Introduction Using *R*. J Stat Soft 2011;43. https://doi.org/10.18637/jss.v043.b04.

[37] Koralov L, Sinai YG. Theory of Probability and Random Processes. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. https://doi.org/10.1007/978-3-540-68829-7.

[38] Wolf S, Møller JK, Bitsch MA, Krogstie J, Madsen H. A Markov-Switching model for building occupant activity estimation. Energy and Buildings 2019;183:672–83. https://doi.org/10.1016/j.enbuild.2018.11.041.

[39] Dobbs JR, Hencey BM. Model predictive HVAC control with online occupancy model. Energy and Buildings 2014;82:675–84. https://doi.org/10.1016/j.enbuild.2014.07.051.

[40] Richardson I, Thomson M, Infield D. A high-resolution domestic building occupancy model for energy demand simulations. Energy and Buildings 2008;40:1560–6. https://doi.org/10.1016/j.enbuild.2008.02.006.

[41] Virote J, Neves-Silva R. Stochastic models for building energy prediction based on occupant behavior assessment. Energy and Buildings 2012;53:183–93.

https://doi.org/10.1016/j.enbuild.2012.06.001.

[42] Xie J, Pan Y, Jia W, Xu L, Huang Z. Energy-consumption simulation of a distributed air-conditioning system integrated with occupant behavior. Applied Energy 2019;256:113914. https://doi.org/10.1016/j.apenergy.2019.113914.

[43] Salimi S. Occupancy prediction model for open-plan offices using real-time location system and inhomogeneous Markov chain. Building and Environment 2019.

[44] Ching W-K, Huang X, Ng MK, Siu T-K. Markov Chains: Models, Algorithms and Applications. vol. 189. Boston, MA: Springer US; 2013. https://doi.org/10.1007/978-1-4614-6312-2.

[45] Li Z, Dong B. Short term predictions of occupancy in commercial buildings—Performance analysis for stochastic models and machine learning approaches. Energy and Buildings 2018;158:268–81. https://doi.org/10.1016/j.enbuild.2017.09.052.

[46] Volkov YuS. The General Problem of Polynomial Spline Interpolation. Proc Steklov Inst Math 2018;300:187–98. https://doi.org/10.1134/S0081543818020190.

[47] Li D, Cheng B, Xiang S. Direct cubic B-spline interpolation: A fuzzy interpolating method for weightless, robust and accurate DVC computation. Optics and Lasers in Engineering 2024;172:107886. https://doi.org/10.1016/j.optlaseng.2023.107886.

[48] Achite M, Tsangaratos P, Pellicone G, Mohammadi B, Caloiero T. Application of multiple spatial interpolation approaches to annual rainfall data in the Wadi Cheliff basin (north Algeria). Ain Shams Engineering Journal 2024;15:102578. https://doi.org/10.1016/j.asej.2023.102578.

[49] Rasmussen C, Frölke L, Bacher P, Madsen H, Rode C. Semi-parametric modelling of sun position dependent solar gain using B-splines in grey-box models. Solar Energy 2020;195:249–58. https://doi.org/10.1016/j.solener.2019.11.023.

[50] Thilker CA, Bacher P, Cali D, Madsen H. Identification of non-linear autoregressive models with exogenous inputs for room air temperature modelling. Energy and AI 2022;9:100165. https://doi.org/10.1016/j.egyai.2022.100165.

[51] Zhang X, Saelens D, Roels S. Estimating dynamic solar gains from on-site

measured data: An ARX modelling approach. Applied Energy 2022;321:119278. https://doi.org/10.1016/j.apenergy.2022.119278.

[52] Zhang H, Thilker CA, Madsen H, Li R, Xiao F, Ma T, et al. Stochastic occupancy modelling for spaces with irregular occupancy patterns using adaptive B-Spline-based inhomogeneous Markov Chains. Building and Environment 2024:111721. https://doi.org/10.1016/j.buildenv.2024.111721.

[53] Koeppl H, Setti G, Di Bernardo M, Densmore D, editors. Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology. New York, NY: Springer New York; 2011. https://doi.org/10.1007/978-1-4419-6766-4.

[54] Pautrel L, Moulherat S, Gimenez O, Etienne M. Analysing biodiversity observation data collected in continuous time: Should we use discrete- or continuous-time occupancy models? Methods Ecol Evol 2024;15:935–50. https://doi.org/10.1111/2041-210X.14314.

[55] Kellner KF, Parsons AW, Kays R, Millspaugh JJ, Rota CT. A Two-Species Occupancy Model with a Continuous-Time Detection Process Reveals Spatial and Temporal Interactions. JABES 2022;27:321–38. https://doi.org/10.1007/s13253-021-00482-y.

[56] Madsen H. Inhomogeneous Markov Chains model for the variation in cloud cover 1986.

[57] Kong M, Dong B, Zhang R, O'Neill Z. HVAC energy savings, thermal comfort and air quality for occupant-centric control through a side-by-side experimental study. Applied Energy 2022;306:117987. https://doi.org/10.1016/j.apenergy.2021.117987.

[58] Park JY, Ouf MM, Gunay B, Peng Y, O'Brien W, Kjærgaard MB, et al. A critical review of field implementations of occupant-centric building controls. Building and Environment 2019;165:106351. https://doi.org/10.1016/j.buildenv.2019.106351.

[59] Ye Y, Chen Y, Zhang J, Pang Z, O'Neill Z, Dong B, et al. Energy-saving

potential evaluation for primary schools with occupant-centric controls. Applied Energy 2021;293:116854. https://doi.org/10.1016/j.apenergy.2021.116854.

[60] Pang Z, O'Neill Z, Chen Y, Zhang J, Cheng H, Dong B. Adopting occupancy-based HVAC controls in commercial building energy codes: Analysis of cost-effectiveness and decarbonization potential. Applied Energy 2023;349:121594. https://doi.org/10.1016/j.apenergy.2023.121594.

[61] Emmerich SJ, Persily AK. State-of-the-art review of CO2 demand controlled ventilation technology and application. Gaithersburg, MD: National Institute of Standards and Technology; 2001. https://doi.org/10.6028/NIST.IR.6729.

[62] Rahman H, Han H. Real-time ventilation control based on a Bayesian estimation of occupancy. Build Simul 2021;14:1487–97. https://doi.org/10.1007/s12273-020-0746-7.

[63] Pistochini T, Ellis M, Meyers F, Frasier A, Cappa C, Bennett D. Method of test for CO2-based demand control ventilation systems: Benchmarking the state-of-the-art and the undervalued potential of proportional-integral control. Energy and Buildings 2023;301:113717. https://doi.org/10.1016/j.enbuild.2023.113717.

[64] Wolf S, Calì D, Krogstie J, Madsen H. Carbon dioxide-based occupancy estimation using stochastic differential equations. Applied Energy 2019;236:32–41. https://doi.org/10.1016/j.apenergy.2018.11.078.

[65] Hu M, Xiao F, Jørgensen JB, Li R. Price-responsive model predictive control of floor heating systems for demand response using building thermal mass. Applied Thermal Engineering 2019;153:316–29. https://doi.org/10.1016/j.applthermaleng.2019.02.107.

[66] Dong B, Lam KP. A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting. Build Simul 2014;7:89–106. https://doi.org/10.1007/s12273-013-0142-7.

[67] Goyal S, Ingley HA, Barooah P. Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance. Applied Energy

2013;106:209–21. https://doi.org/10.1016/j.apenergy.2013.01.039.

[68] Goyal S, Barooah P, Middelkoop T. Experimental study of occupancy-based control of HVAC zones. Applied Energy 2015;140:75–84. https://doi.org/10.1016/j.apenergy.2014.11.064.

[69] Killian M, Kozek M. Short-term occupancy prediction and occupancy based constraints for MPC of smart homes. IFAC-PapersOnLine 2019;52:377–82. https://doi.org/10.1016/j.ifacol.2019.08.239.

[70] Hu M, Xiao F, Jørgensen JB, Wang S. Frequency control of air conditioners in response to real-time dynamic electricity prices in smart grids. Applied Energy 2019;242:92–106. https://doi.org/10.1016/j.apenergy.2019.03.127.

[71] Oldewurtel F, Sturzenegger D, Morari M. Importance of occupancy information for building climate control. Applied Energy 2013;101:521–32. https://doi.org/10.1016/j.apenergy.2012.06.014.

[72] Bird M, Daveau C, O'Dwyer E, Acha S, Shah N. Real-world implementation and cost of a cloud-based MPC retrofit for HVAC control systems in commercial buildings. Energy and Buildings 2022;270:112269. https://doi.org/10.1016/j.enbuild.2022.112269.

[73] Shi J, Yu N, Yao W. Energy Efficient Building HVAC Control Algorithm with Real-time Occupancy Prediction. Energy Procedia 2017;111:267–76. https://doi.org/10.1016/j.egypro.2017.03.028.

[74] Stoffel P, Berktold M, Müller D. Real-life data-driven model predictive control for building energy systems comparing different machine learning models. Energy and Buildings 2024;305:113895. https://doi.org/10.1016/j.enbuild.2024.113895.

[75] Freund S, Schmitz G. Implementation of model predictive control in a large-sized, low-energy office building. Building and Environment 2021;197:107830. https://doi.org/10.1016/j.buildenv.2021.107830.

[76] De Coninck R, Helsen L. Practical implementation and evaluation of model predictive control for an office building in Brussels. Energy and Buildings

2016;111:290–8. https://doi.org/10.1016/j.enbuild.2015.11.014.

[77] Bünning F, Huber B, Heer P, Aboudonia A, Lygeros J. Experimental demonstration of data predictive control for energy optimization and thermal comfort in buildings. Energy and Buildings 2020;211:109792. https://doi.org/10.1016/j.enbuild.2020.109792.

[78] Ham SW, Paul L, Kim D, Pritoni M, Brown R, Feng J. Decarbonization of heat pump dual fuel systems using a practical model predictive control: Field demonstration in a small commercial building. Applied Energy 2024;361:122935. https://doi.org/10.1016/j.apenergy.2024.122935.

[79] Thilker CA, Madsen H, Jørgensen JB. Advanced forecasting and disturbance modelling for model predictive control of smart energy systems. Applied Energy 2021;292:116889. https://doi.org/10.1016/j.apenergy.2021.116889.

[80] Pan J, Jain R, Paul S, Vu T, Saifullah A, Sha M. An Internet of Things Framework for Smart Energy in Buildings: Designs, Prototype, and Experiments. IEEE Internet Things J 2015;2:527–37. https://doi.org/10.1109/JIOT.2015.2413397.

[81] Li X, Lu R, Liang X, Shen X, Chen J, Lin X. Smart community: an internet of things application. IEEE Commun Mag 2011;49:68–75. https://doi.org/10.1109/MCOM.2011.6069711.

[82] Kelly SDT, Suryadevara NK, Mukhopadhyay SC. Towards the Implementation of IoT for Environmental Condition Monitoring in Homes. IEEE Sensors J 2013;13:3846–53. https://doi.org/10.1109/JSEN.2013.2263379.

[83] Dragulinescu A-M, Dragulinescu A, Zamfirescu C, Halunga S, Suciu G. Smart Neighbourhood: LoRa-based environmental monitoring and emergency management collaborative IoT platform. 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), Lisbon, Portugal: IEEE; 2019, p. 1–6. https://doi.org/10.1109/WPMC48795.2019.9096192.

[84] Ragnoli M, Colaiuda D, Leoni A, Ferri G, Barile G, Rotilio M, et al. A LoRaWAN Multi-Technological Architecture for Construction Site Monitoring. Sensors 2022;22:8685. https://doi.org/10.3390/s22228685.

[85] Gallardo JL, Ahmed MA, Jara N. LoRa IoT-Based Architecture for Advanced Metering Infrastructure in Residential Smart Grid. IEEE Access 2021;9:124295–312. https://doi.org/10.1109/ACCESS.2021.3110873.

[86] Nur-A-Alam, Ahsan M, Based MdA, Haider J, Rodrigues EMG. Smart Monitoring and Controlling of Appliances Using LoRa Based IoT System. Designs 2021;5:17. https://doi.org/10.3390/designs5010017.

[87] Mataloto B, Ferreira JC, Cruz N. LoBEMS—IoT for Building and Energy Management Systems. Electronics 2019;8:763. https://doi.org/10.3390/electronics8070763.

[88] Moreno M, Úbeda B, Skarmeta A, Zamora M. How can We Tackle Energy Efficiency in IoT BasedSmart Buildings? Sensors 2014;14:9582–614. https://doi.org/10.3390/s140609582.

[89] Paul L, Pereira FDA, Ham SW, Pritoni M, Brown R, Feng JD. Open Building Operating System: An Open-Source Grid Responsive Control Platform for Buildings. 2023 ASHRAE Annual Conference, ASHRAE; 2024. https://doi.org/10.63044/s23pau99.

[90] Bashir MR, Gill AQ. Towards an IoT Big Data Analytics Framework: Smart Buildings Systems. 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Sydney, Australia: IEEE; 2016, p. 1325–32. https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0188.

[91] Kabbaj OA, Péan L-M, Masson J-B, Marhic B, Delahoche L. Occupancy states forecasting with a hidden Markov model for incomplete data, exploiting daily periodicity. Energy and Buildings 2023;287:112985. https://doi.org/10.1016/j.enbuild.2023.112985.

[92] Carl De Boor. A Practical Guide to Splines (Revised Edition). Springer London; 2001.

[93] Starkey D. Cox-deBoor Equations for B-Splines n.d.

[94] Cavanaugh JE, Neath AA. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. WIREs Computational Stats 2019;11:e1460. https://doi.org/10.1002/wics.1460.

[95] Claeskens G, Hjort NL. Model Selection and Model Averaging. Cambridge: Cambridge University Press; 2008. https://doi.org/10.1017/CBO9780511790485.

[96] Milesight. Milesight VS121 Occupancy Sensor 2023. https://www.milesight.cn/documents-download/.

[97] ANSI/ASHRAE Standard 62.1-2022 n.d.

[98] Xu X, Wang S, Sun Z, Xiao F. A model-based optimal ventilation control strategy of multi-zone VAV air-conditioning systems. Applied Thermal Engineering 2009;29:91–104. https://doi.org/10.1016/j.applthermaleng.2008.02.017.

[99] Henrik Madsen. Estimation of continuous-time models for the heat dynamics of a building. Building and Environment 1995.

[100] Thilker CA, Bacher P, Bergsteinsson HG, Junker RG, Cali D, Madsen H. Non-linear grey-box modelling for heat dynamics of buildings. Energy and Buildings 2021;252:111457. https://doi.org/10.1016/j.enbuild.2021.111457.

[101] Bacher P, Madsen H. Identifying suitable models for the heat dynamics of buildings. Energy and Buildings 2011;43:1511–22. https://doi.org/10.1016/j.enbuild.2011.02.005.

[102] Andersen KK, Madsen H, Hansen LH. Modelling the heat dynamics of a building using stochastic differential equations 2000.

[103] D Simon. Optimal state estimation: Kalman, H infinity, and nonlinear approaches. John Wiley & Sons; 2006.

[104] Kristensen NR, Madsen H, Jørgensen SB. Parameter estimation in stochastic grey-box models. Automatica 2004;40:225–37. https://doi.org/10.1016/j.automatica.2003.10.001.

[105] Drgoňa J, Arroyo J, Cupeiro Figueroa I, Blum D, Arendt K, Kim D, et al. All you need to know about model predictive control for buildings. Annual Reviews in Control 2020;50:190–232. https://doi.org/10.1016/j.arcontrol.2020.09.001.

[106] Li W. A real-time optimal control strategy for multi-zone VAV air-conditioning systems adopting a multi-agent based distributed optimization method. Applied Energy 2021:14.

[107] Li W, Li H, Wang S. An event-driven multi-agent based distributed optimal control strategy for HVAC systems in IoT-enabled smart buildings. Automation in Construction 2021;132:103919. https://doi.org/10.1016/j.autcon.2021.103919.

[108] The Government of the Hong Kong Special Administrative Region Indoor Air Quality Management Group. Guidance Notes for the Management of Indoor Air Quality in Offices and Public Places. 2019.

[109] Gök O, Çam NY, Alptekin E, Ezan MA, Erek A. Development of an integrated underfloor heating system model in TRNSYS and performance assessments. Journal of Energy Storage 2024;84:110993. https://doi.org/10.1016/j.est.2024.110993.

[110] Vera-García F, Rubio-Rubio JJ, López-Belchí A, Hontoria E. Modelling and real-data validation of a logistic centre using TRNSYS®: Influences of the envelope, infiltrations and stored goods. Energy and Buildings 2022;275:112474. https://doi.org/10.1016/j.enbuild.2022.112474.

[111] Wang M, Hu E, Chen L. Simulation of a radiation-enhanced thermal diode tank (RTDT) assisted refrigeration and air-conditioning (RAC) system using TRNSYS. Journal of Building Engineering 2024;82:108168. https://doi.org/10.1016/j.jobe.2023.108168.

[112] TRNSYS 2023. https://www.trnsys.com/ (accessed April 5, 2024).

[113] Juhl R, Møller JK, Madsen H. ctsmr - Continuous Time Stochastic Modeling in R 2016.

[114] Leprince J, Madsen H, Miller C, Real JP, Van Der Vlist R, Basu K, et al. Fifty shades of grey: Automated stochastic model identification of building heat dynamics. Energy and Buildings 2022;266:112095. https://doi.org/10.1016/j.enbuild.2022.112095.

[115] Stentoft PA, Munk-Nielsen T, Møller JK, Madsen H, Valverde-Pérez B,

Mikkelsen PS, et al. Prioritize effluent quality, operational costs or global warming? – Using predictive control of wastewater aeration for flexible management of objectives in WRRFs. Water Research 2021;196:116960. https://doi.org/10.1016/j.watres.2021.116960.

[116] https://www.gurobi.com/products/gurobi-optimizer/ n.d.

[117] Bünning F, Huber B, Schalbetter A, Aboudonia A, Hudoba de Badyn M, Heer P, et al. Physics-informed linear regression is competitive with two Machine Learning methods in residential building MPC. Applied Energy 2022;310:118491. https://doi.org/10.1016/j.apenergy.2021.118491.

[118] Zhang H, Xiao F, Zhang C, Li R. A multi-agent system based coordinated multi-objective optimal load scheduling strategy using marginal emission factors for building cluster demand response. Energy and Buildings 2023;281:112765. https://doi.org/10.1016/j.enbuild.2022.112765.