Control-Oriented Thermal Dynamic Modeling for Hybrid-Cooled Data Centers*

Zhenyu Pu, Yu Yang*, Dawei Qiu, Xiaohong Guan, Costas J. Spanos

Abstract

Thermal modeling is essential for reliable and energy-efficient operation of modern data centers. However, the increasing deployment of high-density computing hardware and hybrid air-liquid cooling systems has fundamentally changed the thermal characteristics of modern data centers, giving rise to spatially heterogeneous and operating-dependent thermal dynamics and couplings. Existing data-driven and physics-based thermal models often struggle to accurately capture such operating-dependent thermal dynamics and interactions. To address the challenge, this paper proposes a hybrid physics-data framework for control-oriented thermal dynamic modeling of hybrid air-and-liquid cooling data centers and makes the following main contributions. *First*, a physics-based multi-node thermal dynamic model is developed to characterize the thermal inertia and airflow-induced thermal interactions among server racks, rear-door heat exchangers, and CRAC units. *Second*, a data-driven method is embedded into the physical model to infer operating-dependent airflow couplings from operational data without requiring direct airflow measurements. *Third*, the proposed framework is validated using real-world operational data of a large-scale hybrid-cooled data center with 49 racks and 980 servers. The results demonstrate that the proposed model can effectively predict rack inlet and CRAC return-air temperatures under varying operating conditions, achieving RMSEs of approximately 0.479 °C and 0.315 °C, respectively. The proposed model consistently outperforms the existing data-driven and physics-based thermal models in prediction accuracy. Further experiments demonstrate the benefits of the proposed approach in improving prediction accuracy with limited training data and over extended prediction horizons.

Keywords: Data Center, Hybrid air-liquid cooling, Physics-based thermal dynamic modeling, Data-driven airflow coupling learning, Control-oriented.

1. Introduction

Data centers are critical infrastructure underpinning today's rapidly evolving digital society. The rapid growth in data processing, storage, and computational demands has driven a substantial expansion of global data center infrastructures [1]. In 2024, global data centers consumed approximately 415 TWh of electricity, accounting for about 1.5% of global electricity demand and experiencing an annual growth of 12% over the past five years [2]. Thermal management is critical for the reliable and energy-efficient operation of data centers [3]. Cooling infrastructures constitute approximately 30-40% of data centers' total energy consumption for maintaining the required operating temperature of IT equipment [4]. While overcooling leads to substantial unnecessary energy consumption, undercooling often results in localized hotspots, computational performance throttling or even hardware damage [5]. Effective thermal management depends on a control-oriented thermal dynamic model that can predict the temperature evolution of data centers under time-varying workloads and cooling operation. Such a model enables the operation of cooling infrastructures to respond to the heat load variations to maintain the required thermal conditions with minimal energy consumption. Beyond thermal management, such a model is also fundamental to explore the thermal flexibility of data centers to shift cooling demand to low-price periods, or achieving thermal-aware workload scheduling to further improve the energy efficiency [6].

The computing and cooling architectures of data centers have undergone substantial transformations in recent years [7]. Conventional data centers are predominantly CPU-centric and operate at relatively low rack power densities (i.e., 7-10 kW/rack). In contrast, modern data centers increasingly integrate heterogeneous, power-intensive computing hardware, such as GPUs and other accelerators, substantially driving up rack power densities. Meanwhile, computing equipment is being deployed at increasing densities to meet the growing computational demands. In large-scale data centers, server racks are often organized into closely spaced rows, with each rack housing multiple servers with heterogeneous computing hardware. These trends have led to substantially increased, concentrated and spatially heterogeneous heat generation of data centers [8, 9]. Conventional air-cooling systems have become increasingly inadequate to meet the cooling requirements due to their limited cooling capacity. Liquid cooling and hybrid air-liquid cooling infrastructures have been increasingly deployed in modern data centers [10]. Compared with air, coolants offer higher specific heat capacities and thermal conductivities, enabling more

*This work supported by the National Natural Science Foundation of China (62403373, 72595834, 62192752).

*Yu Yang is corresponding author.

Email addresses: zhenyupu@stu.xjtu.edu.cn (Zhenyu Pu), yangyu21@xjtu.edu.cn (Yu Yang), dawei.qiu@ntu.edu.sg (Dawei Qiu), xhguan@tsinghua.edu.cn (Xiaohong Guan), spanos@berkeley.edu (Costas J. Spanos)

effective heat removal. For a hybrid-cooled data center, air-cooling systems usually depend on computer room air conditioning units (CRACs) to circulate conditioned air in data center to remove the heat, while liquid-cooling systems, such as rear-door heat exchangers (RDHxs), generally provide local heat removal through circulating coolant near server racks. The coexistence of hybrid air-liquid cooling infrastructures enables more effective heat removal, but also introduce more complex heat transfer paths and thermal couplings in modern data centers. In particular, airflow transport within data centers is complicated by cold air bypass, hot air recirculation, and airflow mixing among racks and cooling equipment. These airflow transport leads to complex thermal interactions among racks, RDHxs, and CRACs. Moreover, these thermal interactions vary with IT workloads and cooling operation, posing significant challenges to capture the thermal dynamics and couplings [11].

Thermal modeling of data centers has attracted extensive research interest [3, 5]. Existing studies can be broadly categorized into physics-based, data-driven, and hybrid physics-data approaches. Physics-based models rely on physical conservation laws to characterize the heat transfer processes, providing physically interpretable representations of data center thermal behaviors. Typically, computational fluid dynamics (CFD) models numerically solve the governing fluid-flow and heat transfer equations to provide high-fidelity simulations of airflow distributions and temperature fields within data centers [11–13]. However, constructing such models requires detailed information on data center geometry, equipment layout, airflow resistance characteristics, and boundary conditions. Their calibration can also be challenging due to the limited availability of detailed thermal measurements in real-world data centers. Moreover, the high computational cost of CFD simulations limits their applicability to real-time thermal prediction and control [14]. To reduce computational complexity, zonal models have been developed by dividing data centers into a limited number of interconnected thermal zones [15–17]. Nevertheless, accurately characterizing the airflow exchanges and thermal interactions among different zones remains challenging, particularly under varying operating conditions.

To facilitate control-oriented applications, multi-node thermal network models provide compact representations of thermal interactions among server racks and cooling systems. Early studies represented server racks as thermal nodes and characterized hot-air recirculation among racks using cross-interference coefficient matrices [14, 18, 19]. Related studies further developed temperature-rise distribution models to characterize the effects of rack power distribution on rack inlet temperatures [20]. However, these models primarily characterize steady-state thermal behaviors using fixed thermal coupling coefficients, limiting their ability to represent changes in thermal interactions under varying operating conditions. To address this limitation, [21] introduced a time-varying cross-interference coefficient matrix to characterize thermal couplings under different operating conditions and employed a data-driven method to estimate the coupling coefficients. Nevertheless, this model does not account for thermal inertia and the resulting temperature dynamics. Moreover, it relies on rack airflow information that may be difficult to obtain in practical data centers, and its evaluation was limited to a relatively small air-cooled data center using CFD-generated data. Server-level state-space models have also been developed to characterize thermal dynamics under varying workloads and cooling inputs [22, 23]. However, these models primarily focus on thermal dynamics within individual racks, rather than thermal interactions among multiple racks and cooling units.

Data-driven modeling provides an alternative approach to infer the complex thermal interactions in data centers directly from operational data [10, 24, 25]. Various machine learning models have demonstrated their capability to capture both steady-state and transient thermal behaviors of data centers. However, these models typically learn the relationships between operating conditions and thermal states directly from historical data, without explicitly characterizing the underlying heat-transfer processes and thermal interactions [24].

To leverage the complementary strengths of physics-based and data-driven approaches, hybrid physics–data thermal modeling has attracted increasing research attention. For example, [26] incorporated energy balance equations, airflow transport relationships, and spatial configuration information into a physics-informed neural network for data center thermal modeling, while [27] employed a data-driven approach to identify difficult-to-obtain parameters of a lumped-parameter resistance–capacitance (RC) model for predicting data center return-air temperatures.

Despite these advances, existing thermal models still face challenges in jointly characterizing thermal dynamics and operating-dependent thermal interactions among server racks and cooling infrastructures, particularly for hybrid-cooled data centers, as summarized in Table 1. Moreover, the airflow couplings governing these thermal interactions are difficult to determine directly from operational measurements. To address these challenges, this paper develops a control-oriented thermal dynamic modeling framework with the main contributions.

- A physics-based multi-node thermal dynamic model is developed for hybrid-cooled data centers to explicitly characterize thermal inertia and airflow-induced thermal interactions among server racks, RDHxs, and CRAC units. The model provides a compact state-space representation suitable for control-oriented thermal prediction.
- A data-driven method is developed to infer operating-dependent airflow couplings from operational data without requiring direct airflow measurements. The inferred couplings are incorporated into the energy balance equations to construct the system matrices, enabling the proposed thermal dynamic model to capture variations in thermal interactions under different operating conditions.
- The proposed hybrid physics-data framework is validated using real-world operational data from a large-scale hybrid-cooled data center with 49 racks and 980 servers. The results demonstrate that the proposed thermal dynamic model can effectively capture the operating-dependent thermal dynamics under varying operating conditions, with substantially improved prediction accuracy

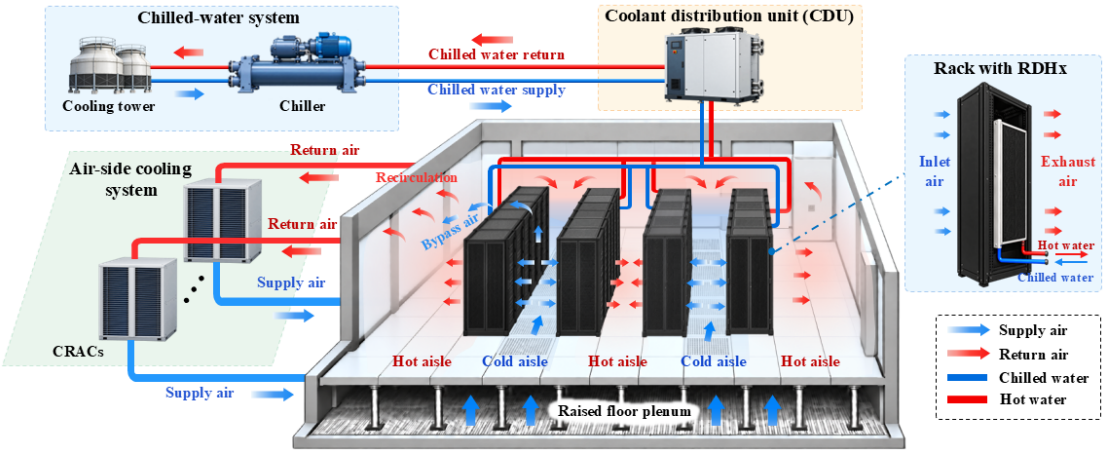


Figure 1: Illustration of a hybrid air-liquid cooling data center

Table 1: Existing studies of thermal modeling of data centers

Modeling Approaches	Thermal Models	Related Works	Inter-Rack Thermal Couplings	Dynamic Modeling	Operating-dependent Couplings	Control Oriented	Cooling Architecture	Evaluation Data
Physics-based	CFD models	[11–13]	Explicit	✗	✓	✗	Air	Simulated
	Zonal models	[15–17]	Inter-zone	✓	✓	✓	Air	Simulated
	Multi-node thermal network models	[14, 18, 20] [21]	Explicit	✗	✗ ✓	✗ ✓	Air	Simulated
	Server-level state-space models	[22, 23]	✗	✓	✓	✓	Air	Simulated
Data-driven	SVR, GPR, XGBoost, Neural Networks	[24, 25] [10]	Implicit	Implicit	Implicit	✓	Air Air Hybrid	Simulated Real Real
Hybrid physics-data	Physics-informed neural network	[26]	Implicit	Implicit	Implicit	✓	Air	Real
	Lumped-parameter RC model	[27]	✗	✓	✗	✓	Air	Real
	Hybrid physics-data multi-node thermal network model	**This paper**	**Explicit**	✓	✓	✓	**Hybrid**	**Real**

compared with existing models. Further experiments demonstrate improved prediction performance under limited training data and extended prediction horizons.

The remainder of this paper is organized as follows. Section 2 presents the physics-based multi-node thermal dynamic modeling, Section 3 introduces the data-driven thermal coupling learning, Section 4 presents the model identification and validation, and Section 5 concludes the paper.

2. Physics-Based Multi-Node Thermal Dynamic Model

2.1. Hybrid Air-liquid Cooling Data Center

As illustrated in Figure 1, a data center equipped with hybrid air-liquid cooling system is considered. The server racks are arranged in rows with alternating cold and hot aisles. The cooling system integrates room-level air cooling provided by computer room air conditioning (CRAC) units and rack-level liquid cooling provided by rear-door heat exchangers (RDHxs).

For room-level air cooling, multiple CRAC units deliver conditioned air from their supply-air outlets into a shared raised-floor plenum, from which the air is distributed to the cold aisles through perforated floor tiles. The distributed air enters the racks and absorbs the heat generated by the IT equipment. Due to the shared air distribution and return-air paths, rack inlet air is a mixture of the supply air from multiple CRAC units and the recirculated air from server racks, resulting in complex thermal interactions.

For rack-level liquid cooling, RDHxs installed at the rear of individual racks remove heat from the rack exhaust air before it enters the hot aisles. Specifically, the exhaust air passes through the corresponding RDHx, where part of its heat is transferred to the circulating coolant (i.e., cold water). The air leaving the RDHxs enters the shared return-air space, where it mixes with air from other racks and is subsequently drawn into the CRAC return-air inlets. The coolant supplied to the RDHxs is circulated through a coolant distribution unit (CDU) connected to the chilled-water system.

The thermal dynamics of the hybrid-cooled data center are governed by IT heat generation process, cooling system operation, and the airflow-induced heat transport, giving rise to complex spatial thermal couplings and temporal dynamics. In particular, variations in IT workloads and cooling operations can alter the airflow distributions and thermal interactions. For thermal management and control, the supply-air temperatures and fan speeds of CRAC units, together with the supply-water temperatures and water flow rates of RDHx units, often serve as the major cooling control inputs, while the IT heat generation

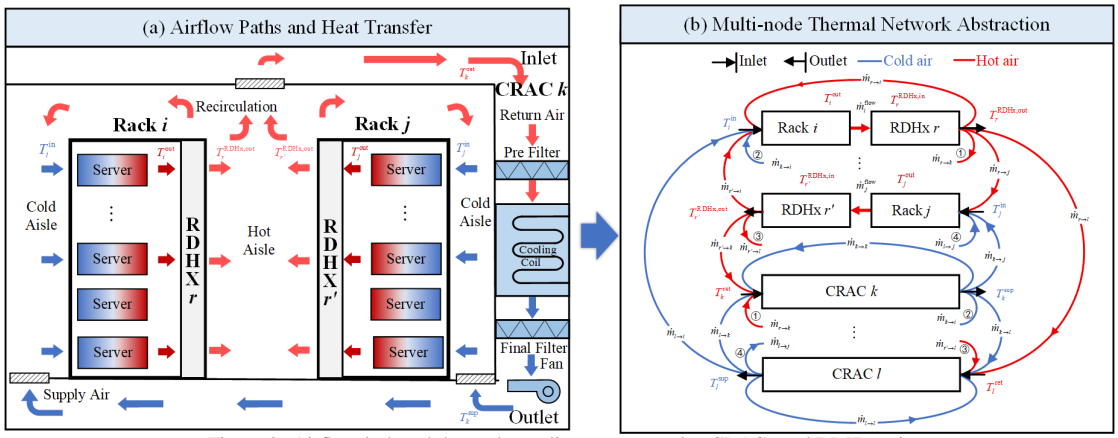


Figure 2: Airflow-induced thermal couplings across racks, CRACs and RDHx units

acts as major thermal disturbances.

For control-oriented thermal modeling, reconstructing the complete temperature field of data center is generally unnecessary and computationally expensive. Instead, characterizing the thermal interactions between server racks and cooling systems to capture how IT workloads and cooling operations affect rack inlet temperature dynamics is often more relevant for thermal management and control. In particular, the evolution of rack inlet air temperatures are of particular importance as they serve as the thermal safety indicators of IT equipment and are recommended to be maintained within 18–27°C by both GB 50174-2017 [28] and ASHRAE TC 9.9 [29].

Motivated by the thermal characteristics of data centers and control-oriented applications, this paper represents the hybrid-cooled data center as a multi-node thermal network consisting of rack, RDHx, and CRAC nodes. The thermal dynamics and interactions among these nodes are modeled based on energy balance equations, with operating-dependent airflow couplings introduced to characterize the variations in thermal interactions under different operating conditions. The resulting thermal network model and airflow coupling relationships are introduced in the following sections.

2.2. Multi-Node Thermal Network Representation

2.2.1. Thermal Network Abstraction

Figure 2(a) describes the airflow paths and associated heat transfer process within the hybrid-cooled data center, which are further represented by a directed multi-node thermal network, as shown in Figure 2(b). The network consists of N racks, N associated RDHx units, and M CRAC units, mapped from the hybrid-cooled data center. Each node represents the thermal behavior of a physical component, while the directed edges characterize airflow-induced heat transport among different components. Specifically, rack nodes describe the inlet and outlet air temperature dynamics governed by IT heat generation and airflow-induced heat transport; RDHx nodes characterize the outlet air temperature dynamics resulting from air-to-liquid heat exchange; and CRAC nodes characterize the return-air temperature dynamics governed by airflow mixing and heat transport, with supply-air temperature as a control input. This abstraction captures the dominant thermal dynamics and interactions among server racks and hybrid cooling components without reconstructing the complete temperature field.

2.2.2. Thermal State Definition

For the thermal network, let $i, j \in \{1, \ldots, N\}$, $r, r' \in \{1, \ldots, N\}$ and $k, l \in \{1, \ldots, M\}$ denote the indices of rack, RDHx, CRAC nodes, respectively. For rack node i, the inlet and outlet air temperatures are denoted by T_i^{in} and T_i^{out} (before entering the RDHx unit), respectively. For the associated RDHx node r, the inlet and outlet air temperatures are denoted by $T_r^{\text{RDHx,in}}$ and $T_r^{\text{RDHx,out}}$, respectively. For CRAC node k, the supply-air and return-air temperatures are denoted by T_k^{sup} and T_k^{ret}, respectively. These temperatures constitute the key thermal state variables of the data center. Among them, the CRAC supply-air temperatures are control variables, while the others evolve dynamically with system operation. Their temporal evolution is governed by IT heat generation, cooling operations and the airflow-induced heat transfer, which can be captured by energy balance equations.

2.3. Multi-Node Thermal Dynamic Model

2.3.1. Rack Inlet and Outlet Air Temperature

As illustrated in Figure 2, the inlet air temperature of rack i is influenced by the supply-air of CRAC units and the recirculated exhaust air from the RDHx units. To account for the thermal inertia of rack inlet, an effective air volume with finite thermal capacitance is introduced. Based on the energy balance equation, the rack inlet temperature dynamics can be formulated as

$$c_p m_i^{\text{in}} \frac{dT_i^{\text{in}}}{dt} = \sum_{k=1}^{M} c_p \dot{m}_{k\to i}(T_k^{\text{sup}} - T_i^{\text{in}}) + \sum_{r=1}^{N} c_p \dot{m}_{r\to i}(T_r^{\text{RDHx,out}} - T_i^{\text{in}}). \quad (1)$$

where t denotes time, m_i^{in} denotes the air mass of the effective air volume, and c_p is the specific heat capacity of air. $\dot{m}_{k\to i}$ and $\dot{m}_{r\to i}$ denote the airflow rates from CRAC unit k and RDHx unit r to rack i. These two terms characterize the effects of supply-air from the CRAC units and the recirculated exhaust air from the RDHx units, respectively.

Accordingly, the rack outlet air temperature is determined by the airflow through the rack and the heat generated by the IT equipment. Similarly, an effective air volume with finite thermal capacitance is introduced to represent the thermal inertia of rack outlet. Based on the energy balance, the rack outlet tem-

perature dynamics can be formulated as

$$c_p m_i^{\text{out}} \frac{dT_i^{\text{out}}}{dt} = c_p \dot{m}_i^{\text{flow}} (T_i^{\text{in}} - T_i^{\text{out}}) + Q_i^{\text{IT}} \tag{2a}$$

$$\dot{m}_i^{\text{flow}} = \sum\nolimits_{k=1}^{M} \dot{m}_{k\to i} + \sum\nolimits_{r=1}^{N} \dot{m}_{r\to i}. \tag{2b}$$

where m_i^{out} denotes the air mass of the effective air volume. Q_i^{IT} represents the heat generated by IT equipment of rack i. $\dot{m}_i^{\text{flow}}$ denotes the airflow through rack i, mixed from the supply-air of CRAC units and recirculated exhaust air from the RDHx units.

2.3.2. RDHx Inlet and Outlet Air Temperature

Each rack i is associated with a RDHx unit r to directly remove heat from its exhaust air. As the RDHx is directly installed near rack outlet, its inlet air temperature is assumed to be equal to the rack outlet air temperature, i.e.,

$$T_r^{\text{RDHx,in}} = T_i^{\text{out}}, i = r \tag{3}$$

Similar to the rack outlets, an effective air volume is introduced to account for the thermal inertia associated with air-to-liquid heat exchange. Based on the energy balance, the RDHx outlet air temperature can be captured by

$$c_p m_r^{\text{RDHx}} \frac{dT_r^{\text{RDHx,out}}}{dt} = c_p \dot{m}_r^{\text{flow}} (T_i^{\text{out}} - T_r^{\text{RDHx,out}}) - Q_r^{\text{RDHx}} \tag{4}$$

where m_r^{RDHx} denotes the air mass of the effective air volume. Q_r^{RDHx} denotes the heat removal rate of the RDHx, and $\dot{m}_r^{\text{flow}}$ is the airflow rate through RDHx r, which equals the airflow rate through rack i as in (2b).

The heat removal rate of RDHx r depends on the inlet air temperature and airflow rate, as well as the coolant supply temperature $T_r^{\text{sup,wat}}$ and volumetric flow rate $V_r^{\text{sup,wat}}$. It can be calculated using the effectiveness–number of transfer units (ε–NTU) method [30, 31] as

$$Q_r^{\text{RDHx}} = \varepsilon_r C_r \left(T_i^{\text{out}} - T_r^{\text{sup,wat}}\right) \tag{5a}$$

$$C_r = \min\left(c_p \dot{m}_r^{\text{flow}}, c_p^w \rho_w V_r^{\text{sup,wat}}\right) \tag{5b}$$

where ε_r is the heat exchanger effectiveness. c_p^w is the specific heat of coolant, ρ_w denotes the coolant density, and C_r denotes the minimum heat capacity rate of the air and coolant streams.

2.3.3. CRAC Supply and Return Air Temperature

The supply-air temperatures of CRAC units are controllable variables, while their return-air temperatures are influenced by the mixing of bypassed supply air from CRAC units and recirculated exhaust air from the RDHx units. Similarly, an effective air volume is introduced to account for the thermal inertia of CRAC inlets. Based on energy balance equation, the return-air temperature dynamics of CRAC k are given by

$$\begin{aligned} c_p m_k^{\text{CRAC}} \frac{dT_k^{\text{ret}}}{dt} = &\sum\nolimits_{l=1}^{M} c_p \dot{m}_{l\to k} \left(T_l^{\text{sup}} - T_k^{\text{ret}}\right) \\ &+ \sum\nolimits_{r=1}^{N} c_p \dot{m}_{r\to k} \left(T_r^{\text{RDHx,out}} - T_k^{\text{ret}}\right) \end{aligned} \tag{6}$$

where m_k^{CRAC} denotes the air mass of the effective air volume. $\dot{m}_{l\to k}$ and $\dot{m}_{r\to k}$ denote the airflow rate from CRAC unit l and RDHx unit r to CRAC k. These two terms characterize the effects of CRAC supply-air bypass and RDHx exhaust-air recirculation on the return-air temperature dynamics, respectively.

2.4. Discrete-time State-Space Thermal Model

For control-oriented applications, a discrete-time thermal dynamic model is often required. To this end, this section derives a discrete-time state-space thermal model from (1)–(6) using a sampling interval of Δt.

2.4.1. Thermal States, Cooling Variables, and Disturbances

Based on the multi-node thermal dynamic model (1)–(6), the thermal states are defined as rack inlet and outlet air temperatures, RDHx outlet air temperatures, and CRAC return-air temperatures:

$$\begin{aligned} \mathbf{T}_t = [\quad & \underbrace{T_{1,t}^{\text{in}}, \cdots, T_{N,t}^{\text{in}}}_{\text{Rack inlet-air temperatures}}, \quad \underbrace{T_{1,t}^{\text{out}}, \cdots, T_{N,t}^{\text{out}}}_{\text{Rack outlet-air temperatures}}, \\ & \underbrace{T_{1,t}^{\text{RDHx,out}}, \cdots, T_{N,t}^{\text{RDHx,out}}}_{\text{RDHx outlet-air temperatures}}, \quad \underbrace{T_{1,t}^{\text{ret}}, \cdots, T_{M,t}^{\text{ret}}}_{\text{CRAC return-air temperatures}} \quad]^\top \end{aligned} \tag{7}$$

where the subscript t denotes discrete time steps. $\mathbf{T}_t \in \mathbb{R}^{3N+M}$ represents the thermal states of the data center at time step t.

The evolution of these thermal states is influenced by the cooling operating variables of the hybrid-cooled system, including the CRAC supply-air temperatures and the heat-removal rates of RDHx units, which are collected as

$$\mathbf{U}_t = [\quad \underbrace{T_{1,t}^{\text{sup}}, \cdots, T_{M,t}^{\text{sup}}}_{\text{CRAC supply air temperatures}}, \quad \underbrace{Q_{1,t}^{\text{RDHx}}, \cdots, Q_{N,t}^{\text{RDHx}}}_{\text{RDHx heat removal rates}} \quad]^\top \tag{8}$$

where $\mathbf{U}_t \in \mathbb{R}^{M+N}$. The CRAC supply-air temperatures are directly controllable, while the RDHx heat removal rates are regulated by the control inputs of coolant supply temperatures and volumetric flow rates, as discussed in (5a)-(5b).

The IT heat generation represents the major exogenous thermal disturbances of the data center, which are collected as

$$\mathbf{D}_t = \left[Q_{1,t}^{\text{IT}}, \cdots, Q_{N,t}^{\text{IT}}\right]^\top \tag{9}$$

where $\mathbf{D}_t \in \mathbb{R}^N$ and $Q_{i,t}^{\text{IT}}$ denotes the heat generation rate of IT equipment in rack i at time step t.

2.4.2. Airflow-induced Thermal Couplings

According to (1)–(6), the thermal dynamics of thermal nodes are coupled through the airflow transports among the racks, CRAC units, and RDHx units. However, the airflow rates are not independent control variables. They depend on the physical layout of racks and cooling equipment, and vary with cooling operations and IT workload distribution. Specifically, CRAC fan speeds affect the supply airflow rates and air distribution across different racks. Meanwhile, variations in IT workloads alter the built-in server fan operations and the airflow through individual racks. These variations further influence the bypass and recirculation airflow among racks, RDHxs, and CRAC units. To represent these airflow-induced thermal couplings in a compact form, the airflow rates involved in (1)–(6), are collected into an airflow coupling matrix

$$\mathbf{\Phi}_t = \begin{bmatrix} \dot{\mathbf{M}}_{\text{CRAC}\to\text{Rack,t}} & \dot{\mathbf{M}}_{\text{CRAC}\to\text{CRAC,t}} \\ \dot{\mathbf{M}}_{\text{RDHx}\to\text{Rack,t}} & \dot{\mathbf{M}}_{\text{RDHx}\to\text{CRAC,t}} \end{bmatrix} \tag{10}$$

where $\mathbf{\Phi}_t \in \mathbb{R}^{(M+N)\times(M+N)}$ and the four blocks characterize the airflow rates along the CRAC-to-Rack, RDHx-to-Rack, CRAC-to-CRAC, and RDHx-to-CRAC paths, respectively. Specifically, $\dot{\mathbf{M}}_{\text{CRAC}\to\text{Rack,t}} = [\dot{m}_{k\to i}]_{k,i} \in \mathbb{R}^{M\times N}$ represents the airflow rates from the CRAC outlets to rack inlets, characterizing conditioned air supply; $\dot{\mathbf{M}}_{\text{RDHx}\to\text{Rack,t}} = [\dot{m}_{r\to i}]_{r,i} \in \mathbb{R}^{N\times N}$ represents the airflow rates from the RDHx outlets to rack inlets, characterizing hot-air recirculation; $\dot{\mathbf{M}}_{\text{CRAC}\to\text{CRAC,t}} =$

$[\dot{m}_{l\to k}]_{l,k} \in \mathbb{R}^{M\times M}$ represents the airflow rates from the CRAC outlets to their inlets, characterizing cold air bypass; $\dot{\mathbf{M}}_{\text{RDHx}\to\text{CRAC,t}} = [\dot{m}_{r\to k}]_{r,k} \in \mathbb{R}^{N\times M}$ represents the airflow rates from RDHx outlets to the CRAC inlets, characterizing the exhaust hot air recirculation. The airflow coupling matrix is operating-dependent, which will be discussed in Section 3.

2.4.3. Discrete-time Parameter-Varying State-Space Model

The effective air masses associated with the rack inlet, rack outlet, RDHx outlet and CRAC inlets for capturing the thermal inertia can be collected as

$$\begin{aligned}\mathbf{m} =&[m_1^{\text{in}}, \cdots, m_N^{\text{in}}, m_1^{\text{out}}, \cdots, m_N^{\text{out}},\\ &m_1^{\text{RDHx}}, \cdots, m_N^{\text{RDHx}}, m_1^{\text{CRAC}}, \cdots, m_M^{\text{CRAC}}]^\top \in \mathbb{R}^{3N+M}\end{aligned} \tag{11}$$

Using the forward Euler method, we approximate $dT/dt \approx (T_{t+1} - T_t)/\Delta t$. Accordingly, the thermal dynamics in (1)–(9) can be expressed in a compact discrete-time state-space form

$$\mathbf{T}_{t+1} = \mathbf{A}(\mathbf{\Phi}_t, \mathbf{m})\mathbf{T}_t + \mathbf{B}(\mathbf{\Phi}_t, \mathbf{m})\mathbf{U}_t + \mathbf{E}(\mathbf{m})\mathbf{D}_t \tag{12}$$

where $\mathbf{A}(\mathbf{\Phi}_t, \mathbf{m}) \in \mathbb{R}^{(3N+M)\times(3N+M)}$, $\mathbf{B}(\mathbf{\Phi}_t, \mathbf{m}) \in \mathbb{R}^{(3N+M)\times(N+M)}$, and $\mathbf{E}(\mathbf{m}) \in \mathbb{R}^{(3N+M)\times N}$ are the state, input, and disturbance matrices, respectively. The matrices are determined by the airflow coupling matrix $\mathbf{\Phi}_t$ and the effective air masses $\mathbf{m}$. They can be readily derived from (1)–(6), and have the following block structures:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} & \mathbf{A}_{13} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{43} & \mathbf{A}_{44} \end{bmatrix},\ \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{32} \\ \mathbf{B}_{41} & \mathbf{0} \end{bmatrix},\ \mathbf{E} = \begin{bmatrix} \mathbf{0} \\ \mathbf{E}_2 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{13}$$

where the parameter dependences are omitted for brevity. The non-zero submatrices $\mathbf{A}_{11}, \mathbf{A}_{13}, \mathbf{A}_{21}, \mathbf{A}_{22}, \mathbf{A}_{32}, \mathbf{A}_{33}, \mathbf{B}_{32}, \mathbf{E}_2$ have dimensions $N\times N$; $\mathbf{A}_{44}, \mathbf{B}_{41}$ are $M\times M$, $\mathbf{A}_{43}$ are $M\times N$; and $\mathbf{B}_{11}$ is $N\times M$. Among them, $\mathbf{A}_{11}, \mathbf{A}_{21}, \mathbf{A}_{22}, \mathbf{A}_{32}, \mathbf{A}_{33}, \mathbf{A}_{44}, \mathbf{B}_{32}, \mathbf{E}_2$ are diagonal matrices, with their diagonal elements given by

$$\begin{aligned}
&[\mathbf{A}_{11}]_{ii} = 1 - \Delta t[\sum\nolimits_{k=1}^{M} \frac{\dot{m}_{k\to i}}{m_i^{\text{in}}} + \sum\nolimits_{r=1}^{N} \frac{\dot{m}_{r\to i}}{m_i^{\text{in}}}]\\
&[\mathbf{A}_{21}]_{ii} = \frac{\dot{m}_i^{\text{flow}}}{m_i^{\text{out}}}\Delta t,\ [\mathbf{A}_{22}]_{ii} = 1 - \frac{\dot{m}_i^{\text{flow}}}{m_i^{\text{out}}}\Delta t,\\
&[\mathbf{A}_{32}]_{rr} = \frac{\dot{m}_r^{\text{flow}}}{m_r^{\text{RDHx}}}\Delta t,\ [\mathbf{A}_{33}]_{rr} = 1 - \frac{\dot{m}_r^{\text{flow}}}{m_r^{\text{RDHx}}}\Delta t,\\
&[\mathbf{A}_{44}]_{kk} = 1 - \Delta t[\sum\nolimits_{l=1}^{M} \frac{\dot{m}_{l\to k}}{m_k^{\text{CRAC}}} + \sum\nolimits_{r=1}^{N} \frac{\dot{m}_{r\to k}}{m_k^{\text{CRAC}}}]\\
&[\mathbf{B}_{32}]_{rr} = -\frac{\Delta t}{c_p m_r^{\text{RDHx}}},\ [\mathbf{E}_2]_{ii} = \frac{\Delta t}{c_p m_i^{\text{out}}}
\end{aligned} \tag{14}$$

The submatrices $\mathbf{A}_{13}, \mathbf{A}_{43}, \mathbf{B}_{11}, \mathbf{B}_{41}$ are given by

$$\begin{aligned}
[\mathbf{A}_{13}]_{ir} &= \frac{\dot{m}_{r\to i}}{m_i^{\text{in}}}\Delta t,\ [\mathbf{A}_{43}]_{kr} = \frac{\dot{m}_{r\to k}}{m_k^{\text{CRAC}}}\Delta t\\
[\mathbf{B}_{11}]_{ik} &= \frac{\dot{m}_{k\to i}}{m_i^{\text{in}}}\Delta t,\ [\mathbf{B}_{41}]_{kl} = \frac{\dot{m}_{l\to k}}{m_k^{\text{CRAC}}}\Delta t,
\end{aligned} \tag{15}$$

Here $[\mathbf{X}]_{ab}$ denotes the (a, b)-th entry of $\mathbf{X}$, with $i, r \in \{1, 2, \cdots, N\}$ and $l, k \in \{1, 2, \cdots, M\}$.

The model (12) establishes a discrete-time linear parameter-varying state-space representation for compactly characterizing the thermal dynamics of the data center under cooling operations, IT heat generation, and airflow-induced thermal couplings. The system matrices $\mathbf{A}(\mathbf{\Phi}_t, \mathbf{m})$, $\mathbf{B}(\mathbf{\Phi}_t, \mathbf{m})$ and $\mathbf{E}(\mathbf{m})$ are determined by the operating-dependent airflow coupling matrix $\mathbf{\Phi}_t$, and time-invariant effective air masses $\mathbf{m}$. However, the airflow couplings are generally unavailable from direct measurements and vary with cooling operations and IT workloads. Therefore, identifying $\mathbf{\Phi}_t$ and $\mathbf{m}$ is essential for constructing the state-space model and predicting thermal responses under different operating conditions.

3. Airflow-Induced Thermal Couplings Learning

Explicitly modeling the airflow coupling matrix $\mathbf{\Phi}_t$ is generally impractical due to the complex airflow dynamics within a data center. A promising solution is to use data-driven approach to implicitly learn the airflow coupling matrix $\mathbf{\Phi}_t$ and the time-invariant effective air masses $\mathbf{m}$ to construct the system matrices $\mathbf{A}(\mathbf{\Phi}_t, \mathbf{m})$, $\mathbf{B}(\mathbf{\Phi}_t, \mathbf{m})$ and $\mathbf{E}(\mathbf{m})$. An alternative way is to directly learn the system matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{E}$ from the operational data. In principle, both approaches can capture the airflow-induced thermal couplings from operational data. However, directly learning the system matrices by treating their entries as independent model parameters does not necessarily preserve the physical airflow pathways established in the multi-node thermal network. In contrast, learning $\mathbf{\Phi}_t$ and $\mathbf{m}$ allows these pathways to be explicitly represented when constructing the system matrices for the energy balance equations (12).

Therefore, this section focuses on learning the airflow coupling matrix $\mathbf{\Phi}_t$ and the effective air masses $\mathbf{m}$ to construct the system matrices $\mathbf{A}(\mathbf{\Phi}_t, \mathbf{m})$, $\mathbf{B}(\mathbf{\Phi}_t, \mathbf{m})$ and $\mathbf{E}(\mathbf{m})$. Directly learning the system matrices will be discussed in our experiments.

3.1. Airflow-Induced Thermal Couplings Learning

Airflow transport in data centers is jointly shaped by the physical layout of the data center, the operation of cooling system and IT workloads [25, 32]. For a given data center, the rack layout, aisle configuration, perforated-tile arrangement, and plenum configuration determine possible airflow pathways and their associated flow resistances [32]. The operation of CRAC units, particularly their fan speeds, affects the supplied airflow rates and pressure distribution within the cold and hot aisles of the data center, thereby influencing airflow transport [32]. Meanwhile, the liquid-cooling system, through RDHx operation, modifies rack exhaust-air temperatures, which indirectly influence airflow distribution through temperature-dependent buoyancy effects [33, 34]. Moreover, IT workloads affect server fan speeds through variations in heat generation, thereby influencing airflow rates within the data center [35].

The above analysis suggests that the airflow coupling matrix $\mathbf{\Phi}_t$ depends on cooling operations and IT workloads, given the physical layout of a data center. To capture such operating-dependent airflow couplings, a neural network is developed to infer $\mathbf{\Phi}_t$ from the observable operating variables, with the effects of the physical layout implicitly embedded in the learned mapping.

Specifically, the hybrid cooling operation is characterized by the physical control inputs of the hybrid cooling system, including the supply-air temperatures and fan speeds of CRAC units,

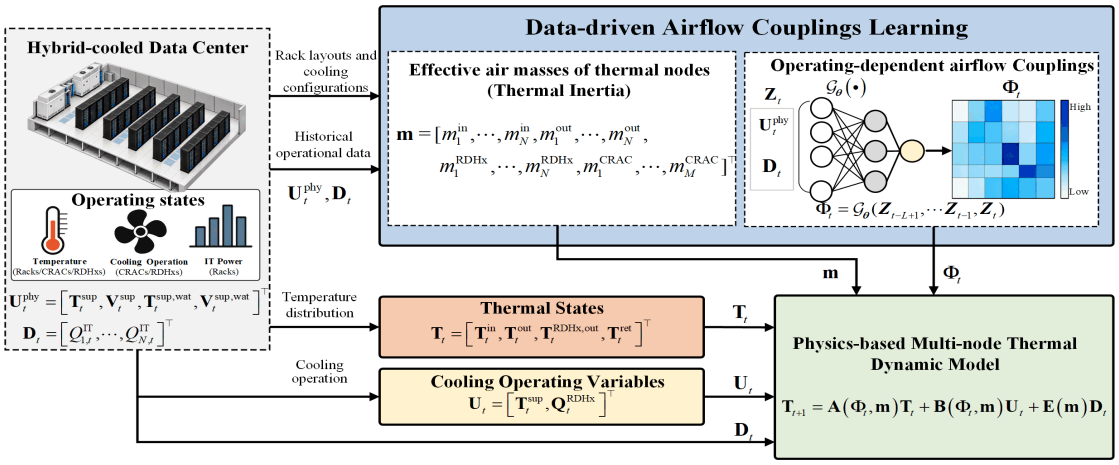


Figure 3: Hybrid physics-data thermal modeling framework for hybrid-cooled data center

as well as the supply-water temperatures and volumetric flow rates of RDHx units, which are collected as

$$\mathbf{U}_t^{\text{phy}} = [\underbrace{\mathbf{T}_t^{\text{sup}}, \mathbf{V}_t^{\text{sup}}}_{\text{Air cooling system}}, \underbrace{\mathbf{T}_t^{\text{sup,wat}}, \mathbf{V}_t^{\text{sup,wat}}}_{\text{Liquid cooling system}}]^\top \tag{16}$$

where $\mathbf{T}_t^{\text{sup}} = [T_{1,t}^{\text{sup}}, \cdots, T_{M,t}^{\text{sup}}]$ and $\mathbf{V}_t^{\text{sup}} = [V_{1,t}^{\text{sup}}, \cdots, V_{M,t}^{\text{sup}}]$ denote the supply-air temperatures and fan speeds of CRACs units. $\mathbf{T}_t^{\text{sup,wat}} = [T_{1,t}^{\text{sup,wat}}, \cdots, T_{N,t}^{\text{sup,wat}}]$ and $\mathbf{V}_t^{\text{sup,wat}} = [V_{1,t}^{\text{sup,wat}}, \cdots, V_{N,t}^{\text{sup,wat}}]$ denote the supply-water temperatures and volumetric flow rates of the RDHx units. It should be emphasized that the cooling operating variables $\mathbf{U}_t$ are governed by the physical control inputs $\mathbf{U}_t^{\text{phy}}$ of the hybrid cooling system, where RDHx heat-removal rates are determined through (5).

The IT workloads across the racks are characterized by IT heat generation rates in racks have been collected as $\mathbf{D}_t$.

The operating variables that influence the airflow couplings at time t are collected as

$$\mathbf{Z}_t = [(\mathbf{U}_t^{\text{phy}})^\top, \mathbf{D}_t^\top]^\top \tag{17}$$

While $\mathbf{Z}_t$ describes the current operating conditions, the airflow couplings may also depend on operating history. In particular, the server fan speeds are often regulated based on equipment thermal conditions, which are influenced by the history of IT workloads and cooling operations [36]. Therefore, both the current and historical operating variables are incorporated into the data-driven model to account for such dependencies, and the airflow coupling matrix is inferred as

$$\mathbf{\Phi}_t = \mathcal{G}_{\boldsymbol{\theta}}\,(\mathbf{Z}_{t-L+1}, \cdots \mathbf{Z}_{t-1}, \mathbf{Z}_t) \tag{18}$$

where $\mathcal{G}_{\boldsymbol{\theta}}(\cdot)$ denotes a learnable nonlinear mapping (e.g.,neural network) parameterized by $\boldsymbol{\theta}$, and L denotes the historical window size. The inferred airflow coupling matrix $\mathbf{\Phi}_t$ can be subsequently used to construct the system matrices for discrete-time linear parameter-varying state-space representation (12).

3.2. Multi-step Thermal Dynamic Model Parameter Learning

In addition to the operating-dependent airflow coupling matrix $\mathbf{\Phi}_t$, the effective air masses $\mathbf{m}$ is also needed to be identified for capturing the thermal inertia in the multi-node thermal dynamic model. Unlike $\mathbf{\Phi}_t$ that varies with the operating conditions, $\mathbf{m}$ can be treated as time-invariant model parameters. Other physical parameters, including the specific heat capacities of air c_p and water c_p^w, are known physical constants and are readily available.

Since the airflow couplings are generally unavailable from direct measurements, the airflow coupling model (18) cannot be trained through direct supervision. To address this issue, the airflow coupling model, together with the effective air masses $\mathbf{m}$, is learned in an end-to-end manner. Specifically, the inferred airflow couplings $\mathbf{\Phi}t$ and effective air masses $\mathbf{m}$ are incorporated into the multi-node thermal dynamic model to recursively predict the thermal states $\{\hat{\mathbf{T}}_{t+1}, \ldots, \hat{\mathbf{T}}_{t+H}\}$, where t denotes the current time step, H denotes the length of prediction horizon, and $\hat{\mathbf{T}}_{t+k}$ denotes the predicted thermal state of the data center at time $t+k$. The model parameters are then optimized by minimizing the multi-step thermal prediction loss, defined as

$$\begin{aligned}\mathcal{L}_{\text{loss}} = \frac{1}{H}\sum_{\tau=1}^{H}\Big(&\lambda_{\text{in}}\left\|\hat{\mathbf{T}}_{t+\tau}^{\text{in}} - \mathbf{T}_{t+\tau}^{\text{in}}\right\|^2 + \lambda_{\text{out}}\left\|\hat{\mathbf{T}}_{t+\tau}^{\text{out}} - \mathbf{T}_{t+\tau}^{\text{out}}\right\|^2 \\ &+ \lambda_{\text{ret}}\left\|\hat{\mathbf{T}}_{t+\tau}^{\text{ret}} - \mathbf{T}_{t+\tau}^{\text{ret}}\right\|^2 + \lambda_{\text{RDHx}}\left\|\hat{\mathbf{T}}_{t+\tau}^{\text{RDHx,out}} - \mathbf{T}_{t+\tau}^{\text{RDHx,out}}\right\|^2 \\ &+ \lambda_{\text{Q}}\left\|\hat{Q}_{t+\tau-1}^{\text{RDHx}} - Q_{t+\tau-1}^{\text{RDHx}}\right\|^2\Big)\end{aligned} \tag{19}$$

where λ_{in}, λ_{out}, λ_{ret}, λ_{RDHx}, and λ_{Q} are non-negative weighting coefficients that balance the contributions of different terms. $Q_{t+\tau-1}^{\text{RDHx}}$ denotes the total heat removal rates of all RDHx units at time step $t+\tau-1$, calculated by summing the heat removal rates of individual RDHx units given in (5). The RDHx heat removal terms are included as an auxiliary supervision terms to further constrain the modeled air–liquid heat transfer process. For the H-step thermal prediction, we employ the established discrete-time state-space model (12) to generate the predictions in a recursive manner. Specifically, the predicted thermal state together with the corresponding cooling operating variables and IT heat generation at each time step will be used to generate predictions for the next time step.

In principle, the model parameters can be learned by minimizing the prediction loss over a given horizon H. However, a long recursive prediction horizon often leads to large accumulated prediction errors, especially in the early stages of training. Moreover, training directly with a long prediction horizon often results in slow convergence of the training loss [37]. To address the issue, a curriculum learning strategy proposed in [38] is adopted. Specifically, for a H-step prediction task, the prediction horizon is first initialized to one step and pro-

gressively increased to H during training process. This allows the model to first focus on short-term prediction accuracy and gradually adapt to longer prediction horizons, facilitating training stability and convergence. Overall, the proposed hybrid physics-data thermal modeling framework that integrates physics-based multi-node thermal dynamic model with data-driven airflow couplings learning is illustrated in Figure 3.

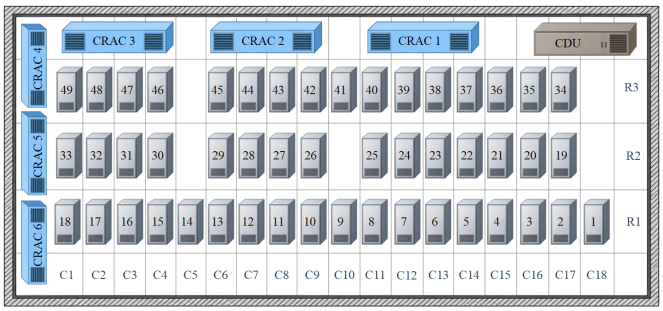


Figure 4: The rack and cooling system layout of M100 data center.

4. Experimental Results

This section first investigates the operational and thermal characteristics of modern hybrid-cooled data center. Afterwards, the proposed thermal modeling framework is evaluated.

4.1. Experiment Setup

Our experiments are conducted based on the operational data of Marconi 100 (M100), a high-performance hybrid-cooled data center located in Bologna, Italy [39]. M100 consists of 49 racks with 980 servers. Each server integrates two 16-core IBM CPUs and four NVIDIA V100 GPUs. M100 adopts a hybrid air-liquid cooling system that consists of 6 CRAC units and rack-level RDHx units. The rack and cooling system layout is shown in Figure 4, where R1-R3 and C1-C18 are used to indicate rack rows and columns, respectively.

Historical operational data of the data center from March 9, 2020 to Sept. 28, 2022 (i.e., two and a half year) is publicly available [40]. The datasets contain inlet air temperatures of racks, supply-air temperatures, return-air temperatures and fan speeds of CRAC units, supply-water temperatures and volumetric flow rates of RDHx units, as well as the power consumption of IT and cooling equipment, recorded at either 10s or 20s intervals. However, the rack and RDHx outlet air temperatures are not available in this dataset. To address this limitation, the quasi-steady equations are used to infer these temperatures as

$$\begin{aligned} T_i^{\text{out}} &\approx T_i^{\text{in}} + Q_i^{\text{IT}}/(c_p \dot{m}_i^{\text{flow}}), \\ T_r^{\text{RDHx,out}} &\approx T_i^{\text{out}} - Q_r^{\text{RDHx}}/(c_p \dot{m}_r^{\text{flow}}), r = i. \end{aligned} \tag{20}$$

where $i, r \in \{1, 2, \cdots, N\}$.

For control-oriented thermal modeling, the raw data is cleaned, temporally aligned and resampled to a uniform 15-min interval. Given the large raw dataset size, 7-month operational datasets spanning April to June 2021, February to April 2022, and September 2022, are adopted for the subsequent data analysis, model development and evaluation. Among them, 5-month datasets including April to May 2021, February to March 2022, and September 2022 are used for training, while June 2021 and April 2022 are reserved for testing. These datasets span seasons and years, making them suitable for evaluating the proposed modeling framework under different operating conditions. The following thermal modeling approaches are compared:

- **NN**: a pure neural network model for thermal states prediction proposed in [41], where current and historical thermal states, cooling control inputs and IT heat generation are used as inputs to predict next-step thermal states.
- **SS-TNM**: a steady-state thermal network model proposed in [21]. A time-varying cross-interference coefficient matrix is learned using a similar data-driven approach as in this study to characterize thermal couplings among nodes.
- **DML-TDM**: the proposed thermal dynamic model (12), where the system matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{E}$ are directly learned using the data-driven approach.
- **ACL-TDM**: the proposed thermal dynamic model (12), where the airflow coupling matrix $\mathbf{\Phi}_t$ and effective air masses $\mathbf{m}$ are learned using the data-driven approach to construct system matrices $\mathbf{A}(\mathbf{\Phi}_t, \mathbf{m})$, $\mathbf{B}(\mathbf{\Phi}_t, \mathbf{m})$ and $\mathbf{E}(\mathbf{m})$.

For all the thermal models, the input window sizes of the data-driven modules are set as $L = 4$ steps (1h). The prediction horizons are set as $H = 24$ steps (6h). All models are trained and evaluated under the same training and testing datasets. For **NN**, **DML-TDM** and **ACL-TDM**, the multi-step predictions are achieved in a recursive manner, while **SS-TNM** is a steady-state model without explicitly considering temporal dynamics. The implementation details of the thermal models are summarized in Appendix A.

4.2. Operational and Thermal Characteristics of Data Center

This section investigates the operational and thermal characteristics of modern data center based on M100 operational data. In particular, the spatial distributions of rack power and rack inlet temperatures, operating-dependent thermal correlations, and thermal responses to workload variations are investigated.

Figure 5(a) shows the distributions of rack power and rack inlet air temperatures with the 49 racks over the experimental dataset. The results indicate that the IT workloads are unevenly distributed across the racks. Meanwhile, the rack inlet temperatures exhibit substantial spatial heterogeneity. Figure 5(b) further presents two snapshots of rack inlet temperature distributions, which imply noticeable spatial heterogeneity and temporal variations. These observations demonstrate that the thermal behaviors of data center show spatially heterogeneous and temporally dynamic characteristics.

Beyond that, complex thermal correlations among the racks are observed as shown in Figure 5(c), where the correlations of rack inlet temperatures under low, medium, and high levels of IT workloads are analyzed. The results show that noticeable thermal correlations exist among the rack inlet temperatures. Moreover, the correlations are operating-dependent and vary with the workloads. Under a low-workload operating condition, strong positive correlations are observed across the rack inlet air temperatures. This can be attributed to the dominant effects of shared air cooling operation under low workload conditions, where the inlet air temperatures show a synchronized variation pattern. However, as the workload increases, negative correlations are observed among certain racks. This may

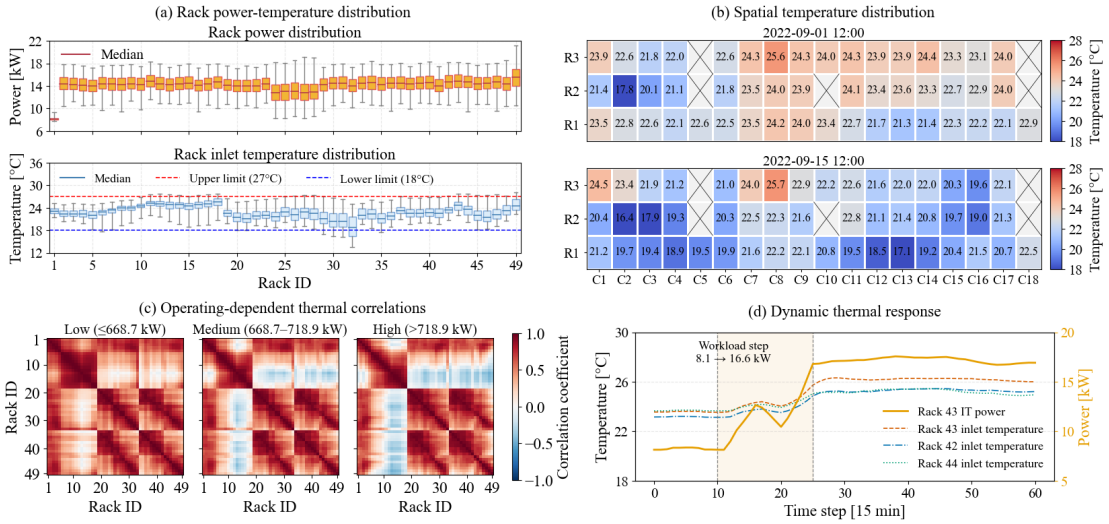


Figure 5: Operational and thermal characteristics of M100 data center: (a) Rack power and rack inlet temperature distributions; (b) Snapshots of rack inlet air temperature distributions; (c) operating-dependent rack inlet air temperature correlations; and (d) Rack inlet air temperature responses to workload variations.

be attributed to increased server fan speeds under higher workloads, which alter the airflow distribution and hot air recirculation patterns within the data center, leading to opposite rack inlet air temperature variations. In addition, the thermal correlations exhibit block-like patterns among Racks 1–18, 19–33 and 34–49. These actually highlight the stronger thermal correlations among racks of the same rows caused by their relatively similar airflow paths.

The complex thermal interactions can be further inferred from the thermal response of rack inlet temperature to IT workload changes. As shown in Figure 5(d), the workload variations in Rack 43 lead to the rack inlet temperature rises of Rack 43 and its neighboring racks, i.e., Rack 42 and Rack 44. This may be attributed to the hot air recirculation from rack outlets to inlets. Specifically, the workload increase of Rack 43 leads to a rise of its outlet temperature, which further affects the inlet temperature of itself and its neighbors through hot air recirculation.

These observations overall highlight that modern hybrid-cooled data centers exhibit operating-dependent thermal interactions and couplings.

4.3. Overall Performance of Thermal Dynamic Modeling

This section evaluates the performance of the thermal models. Specifically, the multi-step predictions of rack inlet temperatures and CRAC return-air temperatures obtained from **NN**, **SS-TNM**, **DML-TDM** and **ACL-TDM** are compared with the measurements. The measurements of rack outlet and RDHx outlet air temperatures are not available and therefore not evaluated. The overall prediction accuracies among the 49 racks and 6 CRAC units are assessed using MAPE, RMSE, MAE and R^2 on both the training and testing datasets, with the results reported in Table 2. In particular, the results of the proposed **DML-TDM** and **ACL-TDM** are highlighted with colored shading, while the best metrics on either training or testing datasets are indicated in bold.

As shown in Table 2, the proposed **ACL-TDM** achieves the best performance across almost all metrics on both the training and testing datasets. Notably, the pure data-driven

Table 2: Rack inlet and CRAC return-air temperature prediction performance by the different thermal models

Thermal State	Method	Dataset	MAPE [%]	RMSE [°C]	MAE [°C]	R^2
Rack inlet temperature	NN	Train	1.461	0.435	0.325	0.921
		Test	2.384	0.805	0.524	0.721
	SS-TNM	Train	2.042	0.695	0.454	0.771
		Test	2.182	0.714	0.484	0.761
	DML-TDM (Proposed)	Train	1.446	0.454	0.320	0.911
		Test	1.655	0.534	0.365	0.869
	ACL-TDM (Proposed)	Train	**1.252**	**0.419**	**0.278**	**0.928**
		Test	**1.476**	**0.479**	**0.326**	**0.901**
CRAC return-air temperature	NN	Train	0.691	0.241	0.178	0.957
		Test	1.523	0.689	0.383	0.563
	SS-TNM	Train	0.956	0.364	0.249	0.885
		Test	1.264	0.437	0.319	0.811
	DML-TDM (Proposed)	Train	0.698	**0.232**	0.180	0.959
		Test	1.084	0.357	0.272	0.870
	ACL-TDM (Proposed)	Train	**0.591**	0.233	**0.154**	**0.967**
		Test	**0.935**	**0.315**	**0.234**	**0.908**

NN model demonstrates comparable performance with the proposed **DML-TDM** and **ACL-TDM** on the training datasets. Nevertheless, its performance degrades substantially on the testing datasets, while the proposed models maintain relatively consistent prediction accuracy.

The advantages of **ACL-TDM** are further demonstrated through comparisons with the **SS-TNM**. Specifically, while **SS-TNM** exhibits relatively stable performance across the training and testing datasets, it shows limited prediction accuracies as indicated by all the performance metrics compared with the proposed **ACL-TDM**. In particular, the test RMSE, MAE and R^2 for rack inlet air temperatures are 0.714 °C, 0.484 °C and 0.761, compared with 0.479 °C, 0.326 °C and 0.901 for **ACL-TDM**. Similar performance gaps of the two models can be observed from the CRAC return-air temperatures. These results highlight the benefits of explicitly modeling multi-node thermal dynamics and operating-dependent thermal couplings.

Furthermore, it is observed that **ACL-TDM** achieves higher prediction accuracy than **DML-TDM** across almost all evaluation metrics for both the rack inlet air and CRAC return-air temperature predictions. Since both models share the same thermal dynamic structure, this improvement highlights the benefits of learning airflow couplings to construct system matrices, rather than directly learning these matrices.

Collectively, the results demonstrate that the proposed **ACL-TDM**, by integrating physics-based thermal dynamic modeling with data-driven airflow couplings learning, can substantially improve the multi-step prediction accuracies under the varying operating conditions compared with existing data-driven and steady-state thermal models.

To further highlight the performance of the proposed **ACL-TDM**, we investigate the rack-level and CRAC-level prediction accuracies on the testing datasets. Figure 6 presents the distributions of MAE and RMSE for rack inlet air temperature predictions, mapped on the physical rack layouts. It is observed that both MAE and RMSE for rack-level predictions show consistent spatial patterns, with the relatively higher prediction errors concentrated in Columns C6–C10 and C14–C17. Moreover, neighboring racks tend to exhibit similar error levels, indicating spatial correlations. However, the MAE and RMSE of rack inlet air temperatures approximately remain within 0.4 °C and 0.6 °C across all the racks.

Accordingly, Figure 7 presents the prediction errors of individual CRAC return-air temperatures. Similar to rack inlet temperature predictions, the prediction errors vary across the CRAC units. Specifically, CRAC 2 and 4 exhibit relatively higher MAE and RMSE, whereas CRAC 5 and 6 show the lowest errors. Nevertheless, both the MAE and RMSE approximately remain within 0.3 °C and 0.4 °C for all CRAC units.

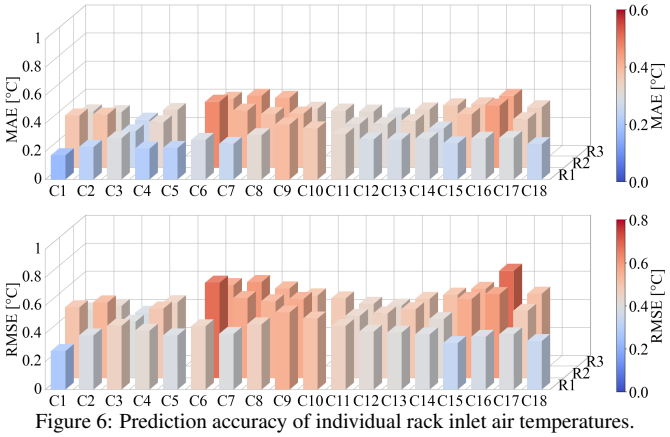


Figure 6: Prediction accuracy of individual rack inlet air temperatures.

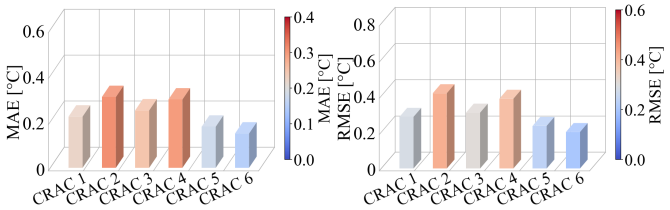


Figure 7: Prediction accuracy of individual CRAC return-air temperatures.

4.4. Prediction Performance Across Temperature Variability

This section further evaluates the performance of the thermal models for capturing the operating-dependent thermal dynamics and couplings across the different temperature variability. Specifically, the operating periods of testing datasets are categorized into the low, medium and high temperature variability levels. We further investigate the predictive performance of the thermal models on the rack inlet air temperatures and CRAC return-air temperatures across the different variability levels.

Figure 8 compares the measured and predicted rack inlet temperature trajectories under the temperature variability levels. Under the low variability, all thermal models generally can follow the measured temperatures. However, as the temperature variability increases, **NN** and **SS-TNM** exhibit increasingly pronounced prediction deviations, particularly under high variability time periods. In contrast, the proposed **DML-TDM** and **ACL-TDM** maintain closer agreement with the measured temperature trajectories across the different variability levels.

Similar trends are observed with CRAC return-air temperatures in Figure 9. Specifically, all thermal models achieve relatively accurate predictions under low temperature variability, whereas the prediction deviations of **NN** and **SS-TNM** become increasingly pronounced as the temperature variability increases. In contrast, both the proposed **DML-TDM** and **ACL-TDM** show relatively stable tracking performance across the different temperature variability levels.

Furthermore, **ACL-TDM** shows slightly closer agreement with the measured rack inlet air and CRAC return-air temperatures than **DML-TDM** across the different variability levels. Since both models share the same thermal dynamic model, this suggests that learning the operating-dependent airflow coupling matrix and effective air masses to construct the system matrices can better capture the operating-dependent thermal dynamics and couplings.

Overall, these results demonstrate the effectiveness of the proposed model in capturing operating-dependent thermal dynamics and couplings under the varying operating conditions.

4.5. Performance Under Different Training Data Size

This section further investigates the effects of training data size on the prediction performance of different thermal models. Specifically, the models are trained using different amounts of historical operational data, ranging from one to five months, and evaluated on the same testing datasets.

Figure 10 presents the monthly distributions of rack inlet air temperature and CRAC return-air temperature of both training and testing dataset. To investigate the effects of training data size, five training datasets containing 1–5 months of data are constructed by progressively adding data from May 2021 to Sept. 2022, while the testing datasets are fixed to June 2021 and April 2022.

Figure 11 presents the MAE and RMSE of rack inlet temperature predictions on the testing datasets for models trained with different amount of data. Overall, the prediction errors of all thermal models decline as the training data size increases. However, the pure data-driven **NN** exhibits relatively high prediction errors, particularly when limited training data are available. In contrast, **SS-TNM**, **DML-TDM**, and **ACL-TDM** achieve relatively low prediction errors even with limited training data.

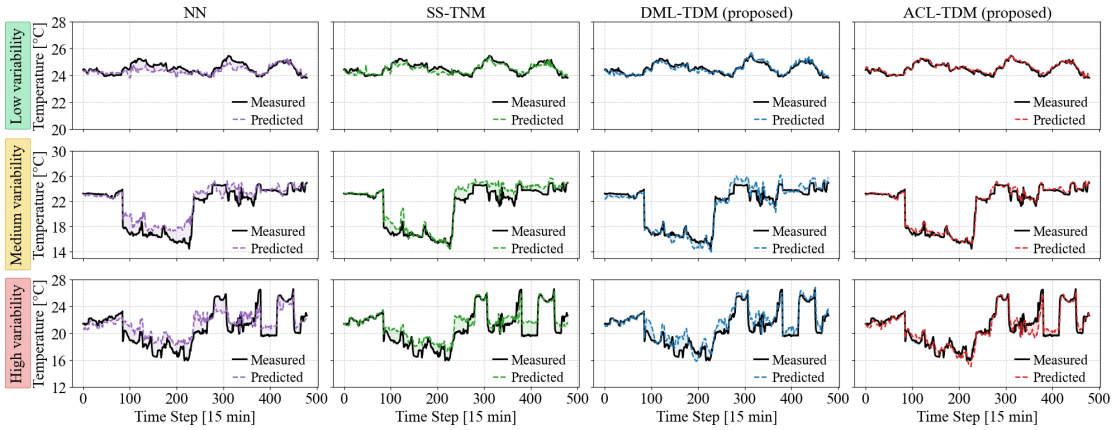


Figure 8: Measured and predicted rack inlet temperature trajectories across varying variability.

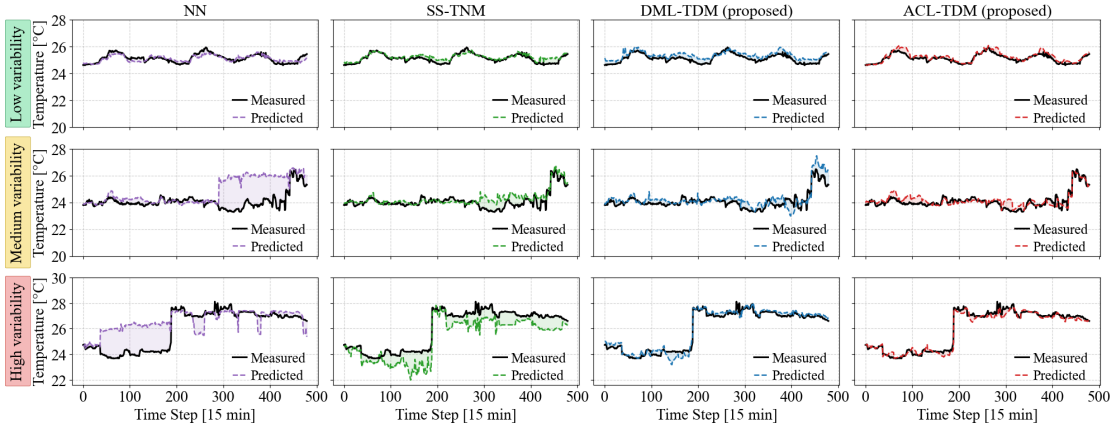


Figure 9: Measured and predicted CRAC return-air temperature trajectories across varying variability.

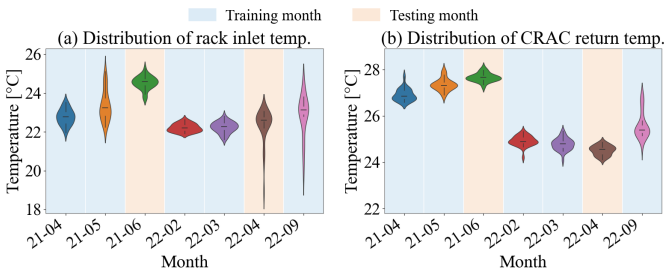


Figure 10: Monthly distributions of rack inlet temperature and CRAC return-air temperature.

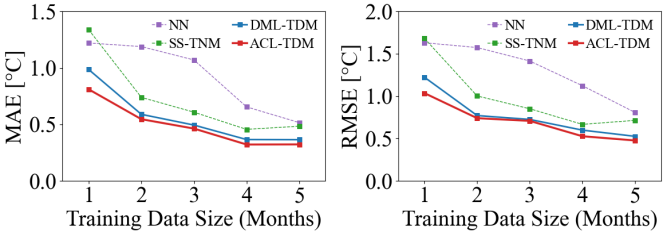


Figure 11: Prediction errors in MAE and RMSE for rack inlet temperature under different training data sizes.

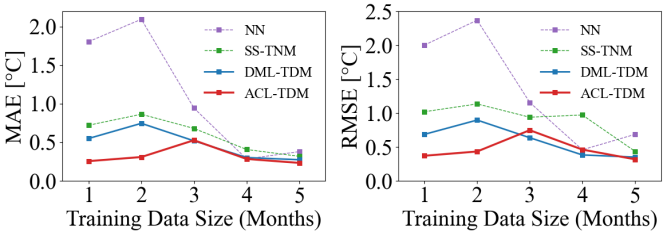


Figure 12: Prediction errors in MAE and RMSE for CRAC return-air temperature under different training data sizes.

Among them, **DML-TDM** and **ACL-TDM** generally outperform **SS-TNM** across different training data sizes, with **ACL-TDM** achieving slightly better prediction than **DML-TDM**.

Similar trends are observed with CRAC return-air temperature predictions, as shown in Figure 12. Specifically, **DML-TDM** and **ACL-TDM** exhibit lower prediction errors than **NN** and **SS-TNM** across most training data sizes. Moreover, **ACL-TDM** generally achieves lower prediction errors compared with **DML-TDM**, especially under the small training data sizes.

These results highlight the benefits of explicitly modeling thermal dynamics for accurate temperature prediction across different training data sizes. Furthermore, learning airflow couplings and constructing the system matrices based on the underlying energy balance equations can further improve prediction accuracy compared with directly learning the system matrices, particularly with limited training data.

4.6. Prediction Performance Under Different Horizons

To investigate the effect of prediction horizons on model performance, the thermal models trained for a prediction horizon of $H = 24$ steps (6 h) are evaluated over different horizons without retraining. Specifically, we test the trained models under varying prediction horizons of $H = 1, 8, 24, 48, 96, 192$ steps (i.e., 15 min, 2h, 6h, 12h, 1 day and 2 days).

Figure 13 presents the RMSE of rack inlet air tempera-

tures and CRAC return-air temperatures across different prediction horizons. Overall, the proposed **DML-TDM** and **ACL-TDM** consistently achieve much lower prediction errors than **NN** and **SS-TNM**, especially for the rack inlet air temperatures. However, the four models show different trends as the prediction horizon increases. Specifically, **SS-TNM** exhibits nearly unchanged prediction errors across the prediction horizons, whereas the prediction errors of **NN**, **DML-TDM** and **ACL-TDM** slightly increase with the prediction horizons. This is because SS-TNM predicts temperatures directly from the operating conditions without recursive state updates, whereas the other three models perform recursive multi-step predictions, allowing prediction errors to accumulate over successive steps. Despite the accumulated prediction errors, the proposed **DML-TDM** and **ACL-TDM** maintain relatively low prediction errors over extended horizons, demonstrating their effectiveness for long-horizon thermal prediction. Furthermore, **ACL-TDM** achieves slightly lower prediction errors than **DML-TDM**. This indicates that learning airflow couplings to construct the system matrices can further improve prediction accuracy compared with directly learning the system matrices.

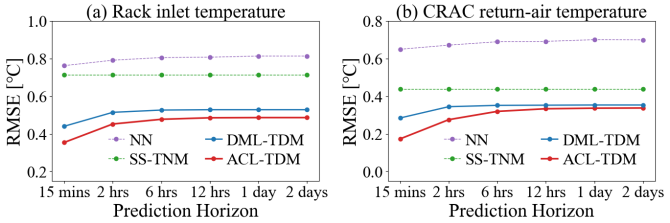


Figure 13: Prediction error in RMSE for rack inlet and CRAC return-air temperature under different prediction horizons.

5. Conclusion

This paper developed a hybrid physics–data framework for control-oriented thermal dynamic modeling of hybrid-cooled data centers. A physics-based multi-node model was established to characterize thermal dynamics and interactions among racks, RDHxs, and CRACs, and further a data-driven method was embedded to learn the airflow couplings without direct measurements. By integrating data-driven airflow coupling learning into the physics-based thermal dynamic model, the proposed framework captures operating-dependent thermal dynamics and interactions while preserving the underlying physical relationships, enabling accurate multi-step temperature prediction. Validation using real-world operational data from a large-scale hybrid-cooled data center with 49 racks and 980 servers demonstrated that the proposed ACL-TDM achieves RMSEs of 0.479 °C and 0.315 °C, respectively. Further analyses demonstrated its advantages under limited training data and extended prediction horizons. The developed model provides a foundation for future thermal management and thermal-aware workload scheduling of hybrid-cooled data centers.

Appendix A. Implementation Details

The common hyperparameters and model-specific neural network architectures are summarized in Table A.1 and Table A.2. For the recursive multi-step models (**NN**, **DML-TDM**, and **ACL-TDM**), a curriculum learning strategy is adopted to improve training stability and convergence. Specifically, the prediction horizon is initially set to one step for the first 150 epochs and then progressively increased to 24 steps over the following 200 epochs.

Table A.1: Common Training Hyperparameter Settings

Param.	**Value**	**Param.**	**Value**
Sampling interval	15 min	Historical length	$L = 4$ (1 h)
Prediction horizon	$H = 24$ (6 h)	Batch size	64
Optimizer	Adam	Learning rate	1×10^{-3}
Max. epochs	400		

Table A.2: Model-Specific Neural Network Architectures

Model	**Neural Network**	**Learned Component**	**Multi-step**
NN	TCN-64 + BiGRU-64 + Attention-64	Thermal states	Recursive
SS-TNM	3-layer MLP-64	Cross-interference coefficient matrix	Steady-state
DML-TDM	2-layer MLP + GRU-64	State-space system matrices $\mathbf{A}, \mathbf{B}, \mathbf{E}$	Recursive
ACL-TDM	2-layer MLP + GRU-64	Airflow coupling matrix	Recursive

References

[1] International Energy Agency, Energy and ai, Tech. rep., International Energy Agency, Paris, https://iea.blob.core.windows.net/assets/de9dea13-b07d-42c5-a398-d1b3ae17d866/EnergyandAI.pdf (2025).

[2] International Energy Agency, Electricity 2024: Analysis and forecast to 2026, Tech. rep., IEA, Paris, available: https://www.iea.org/reports/electricity-2024, Licence: CC BY 4.0 (2024).

[3] A. H. Khalaj, S. K. Halgamuge, A review on efficient thermal management of air-and liquid-cooled data centers: From chip to the cooling system, Applied energy 205 (2017) 1165–1188.

[4] A. Devasia, Data center energy consumption statistics & data (2026), The Network Installers, accessed 2026-01-22; includes global and regional data center energy consumption statistics. Available at: https://thenetworkinstallers.com/blog/data-center-energy-consumption-statistics/ (2026).

[5] J. Lin, W. Lin, H. Huang, W. Lin, K. Li, Thermal modeling and thermal-aware energy saving methods for cloud data centers: A review, IEEE Transactions on Sustainable Computing 9 (3) (2023) 571–590.

[6] J. Ran, Q. Zhang, Y. Zhu, J. Zhai, J. Li, Z. Guo, T. Wang, Co-optimization of thermal-aware workload scheduling with deep reinforcement learning-based cooling control in data centers, Energy (2026) 139965.

[7] S. Mishra, The evolution of data centers in the age of ai, International Journal of Scientific Research in Computer Science, Engineering and Information Technology 10 (5) (2024) 363–368.

[8] Y. Wang, Q. Guo, M. Chen, Providing load flexibility by reshaping power profiles of large language model workloads, Advances in Applied Energy (2025) 100232.

[9] X. Chen, X. Wang, A. Colacelli, M. Lee, L. Xie, Electricity demand and grid impacts of ai data centers: Challenges and prospects, arXiv preprint arXiv:2509.07218 (2025).

[10] W. Lin, W. Lin, J. Lin, L. He, J. Wang, H. Jiang, A multi-agent reinforcement learning-based method for cooling control optimization in hybrid cooling data centers, Journal of Building Engineering 108 (2025) 112811.

[11] N. Wang, Y. Guo, C. Huang, B. Tian, S. Shao, Multi-scale collaborative modeling and deep learning-based thermal prediction for air-cooled data centers: An innovative insight for thermal management, Applied Energy 377 (2025) 124568.

[12] Y. T. Lee, C. Y. Wen, Y. C. Shih, Z. Li, A. S. Yang, Numerical and experimental investigations on thermal management for data center with cold aisle containment configuration, Applied Energy 307 (2022) 118213. doi:10.1016/j.apenergy.2021.118213.

[13] W. Liu, S. Lian, X. Fang, Z. Shang, H. Wu, H. Zhu, S. Hu, An open-source and experimentally guided cfd strategy for predicting air distribution in data centers with air-cooling, Building and Environment 242 (2023) 110542.

[14] Q. Tang, T. Mukherjee, S. K. Gupta, P. Cayton, Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters, in: 2006 Fourth international conference on intelligent sensing and information processing, IEEE, 2006, pp. 203–208.

[15] Z. Song, B. T. Murray, B. Sammakia, A dynamic compact thermal model for data center analysis and control using the zonal method and artificial neural networks, Applied thermal engineering 62 (1) (2014) 48–57.

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

[17] S. Asgari, H. Moazamigoodarzi, P. J. Tsai, S. Pal, R. Zheng, G. Badawy, I. K. Puri, Hybrid surrogate model for online temperature and pressure predictions in data centers, Future Generation Computer Systems 114 (2021) 531–547.

[18] Q. Tang, S. K. S. Gupta, G. Varsamopoulos, Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach, IEEE Transactions on Parallel and distributed systems 19 (11) (2008) 1458–1472.

[19] L. Parolini, B. Sinopoli, B. H. Krogh, Reducing data center energy consumption via coordinated cooling and load management, in: Proceedings of the 2008 conference on Power aware computing and systems, HotPower, Vol. 8, 2008, pp. 14–14.

[20] Z. He, Z. He, X. Zhang, Z. Li, Study of hot air recirculation and thermal management in data centers by using temperature rise distribution, in: Building simulation, Vol. 9, Springer, 2016, pp. 541–550.

[21] Q. Fang, J. Zhou, S. Wang, Y. Wang, Control-oriented modeling and optimization for the temperature and airflow management in an air-cooled data-center, Neural Computing and Applications 34 (7) (2022) 5225–5240.

[22] J. Wang, X. Tong, C. Yue, W. Liu, Q. Zhang, L. Zeng, G. Huang, Real-time temperature distribution reconstruction via linear parameter-varying state-space model and kalman filter in rack-based cooling data centers, Building and Environment 242 (2023) 110601.

[23] X. Tong, J. Wang, W. Liu, H.-A. Samah, Q. Zhang, L. Zhang, A time-varying state-space model for real-time temperature predictions in rack-based cooling data centers, Applied Thermal Engineering 230 (2023) 120737.

[24] J. Athavale, M. Yoda, Y. Joshi, Comparison of data driven modeling approaches for temperature prediction in data centers, International Journal of Heat and Mass Transfer 135 (2019) 1039–1052.

[25] J. Lin, W. Lin, W. Lin, J. Wang, H. Jiang, Thermal prediction for air-cooled data center using data driven-based model, Applied Thermal Engineering 217 (2022) 119207.

[26] L. Kachun, C. Song, H. Chao, L. Chen, A. Guo, High temporal-resolution hvac control in gpu-centric data centers via reinforcement learning and a knowledge-embedded physics-informed neural network, Journal of Building Engineering (2026) 115499.

[27] D. Chen, C.-K. Chui, P. S. Lee, Adaptive physically consistent neural networks for data center thermal dynamics modeling, Applied Energy 377 (2025) 124637.

[28] GB 50174-2017: Code for Design of Data Centers (2017).

[29] ASHRAE Technical Committee 9.9, Thermal Guidelines for Data Processing Environments, 5th Edition, ASHRAE, Peachtree Corners, GA, USA, 2021.

[30] A. P. Wemhoff, A. Ortega, An exergy-based analysis of the effects of rear door heat exchange systems on data center energy efficiency, in: Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), IEEE, 2014, pp. 1129–1136.

[31] M. Del Valle, C. Caceres, A. Ortega, Transient modeling and validation of chilled water based cross flow heat exchangers for local on-demand cooling in data centers, in: 2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), IEEE, 2016, pp. 727–736.

[32] H. Lu, Z. Zhang, L. Yang, A review on airflow distribution and management in data center, Energy and Buildings 179 (2018) 264–277.

[33] J. Cho, Numerical analysis of rack-based data center cooling with rear door heat exchanger (rdhx): Interrelationship between thermal performance and energy efficiency, Case Studies in Thermal Engineering 63 (2024) 105247. doi:10.1016/j.csite.2024.105247.

[34] E. Wibron, A.-L. Ljung, T. S. Lundström, Computational fluid dynamics modeling and validating experiments of airflow in a data center, Energies 11 (3) (2018) 644. doi:10.3390/en11030644.

[35] K. Cao, Z. Li, H. Luo, Y. Jiang, H. Liu, L. Xu, P. Gao, H. Liu, Comprehensive review and future prospects of multi-level fan control strategies in data centers for joint optimization of thermal management systems, Journal of Building Engineering 94 (2024) 110021. doi:10.1016/j.jobe.2024.110021.

[36] B. Durand-Estebe, C. Le Bot, J. N. Mancos, E. Arquis, Data center optimization using pid regulation in cfd simulations, Energy and Buildings 66 (2013) 154–164.

[37] D. Kochkov, J. Yuval, I. Langmore, P. Norgaard, J. Smith, G. Mooers, M. Klöwer, J. Lottes, S. Rasp, P. Düben, S. Hatfield, P. Battaglia, A. Sanchez-Gonzalez, M. Willson, M. P. Brenner, S. Hoyer, Neural general circulation models for weather and climate, Nature 632 (8027) (2024) 1060–1066. doi:10.1038/s41586-024-07744-y.

[38] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 41–48. doi:10.1145/1553374.1553380.

[39] A. Borghesi, C. Di Santi, M. Molan, M. S. Ardebili, A. Mauri, M. Guarrasi, D. Galetti, M. Cestari, F. Barchi, L. Benini, et al., M100 exadata: a data collection campaign on the cineca's marconi100 tier-0 supercomputer, Scientific Data 10 (1) (2023) 288.

[40] ECS Lab, M100 ExaData, https://gitlab.com/ecs-lab/exadata, accessed 22 October 2026 (2023).

[41] J. Lin, W. Lin, W. Lin, T. Liu, J. Wang, H. Jiang, Multi-objective cooling control optimization for air-liquid cooled data centers using tcn-bigru-attention-based thermal prediction models, in: Building simulation, Vol. 17, Springer, 2024, pp. 2145–2161.