applied system innovation

MDPI

Article

Proactive Cooling Control Algorithm for Data Centers Based on LSTM-Driven Predictive Thermal Analysis

Jieying Liu [1], Rui Fan [2], Zonglin Li [3], Napat Harnpornchai [4,*] and Jianlei Qian [1,*]

[1] Cybersecurity and Informatization Office, Shandong University of Finance and Economics, Jinan 250014, China; liujy@sdufe.edu.cn
[2] International Office, Shandong University of Finance and Economics, Jinan 250014, China; fanrui@sdufe.edu.cn
[3] Academic Affairs Office, Shandong University of Finance and Economics, Jinan 250014, China; lizonglin@sdufe.edu.cn
[4] Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand
* Correspondence: napateconcmu@gmail.com (N.H.); qianjl@sdufe.edu.cn (J.Q.)

Abstract

The conventional reactive cooling strategy, which relies on static thresholds, has become inadequate for managing dynamically changing heat loads, often resulting in energy inefficiency and increased risk of local hot spots. In this study, we develop a data center cooling optimization system that integrates distributed sensor arrays for predictive analysis. By deploying high-density temperature and humidity sensors both inside and outside server racks, a real-time, high-fidelity three-dimensional digital twin of the data center's thermal environment is constructed. Time-series analysis combined with Long Short-Term Memory algorithms is employed to forecast temperature and humidity based on the extensive environmental data collected, achieving high predictive accuracy with a root mean square error of 0.25 and an R^2 value of 0.985. Building on these predictions, a proactive cooling control strategy is formulated to dynamically adjust fan speeds and the opening degree of chilled-water valves in computer room air conditioning units, changing the cooling approach from passive to preemptive prevention of overheating. Compared with conventional proportional–integral–differential control, the developed system significantly reduces overall energy consumption and maintains all equipment within safe operating temperatures. Specifically, the framework has reduced the energy consumption of the cooling system by 37.5%, lowered the overall power usage effectiveness of the data center by 12% (1.48 to 1.30), and suppressed the cumulative hotspot duration (temperature 27 °C) by nearly 96% (from 48 to 2 h).

Keywords: data center cooling; predictive analysis; sensor array; algorithm; energy efficiency

Check for updates

Academic Editor: Emmanuel Karapidakis

Received: 6 December 2025
Revised: 31 December 2025
Accepted: 5 January 2026
Published: 12 January 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The ongoing digital transformation has driven unprecedented growth in the scale and computing capability of data centers, which serve as the core of global IT infrastructure [1]. Data centers consume 1–2% of global electricity, with cooling systems having the largest share, accounting for approximately 30–40% of the data center's total energy use [2]. The huge scale of high-power chips and high-density racks has considerably increased server heat generation, imposing stringent demands on cooling system performance [3]. Therefore, effective thermal management is essential for reducing operational costs and supporting

carbon neutrality [4], and for ensuring the stable operation of IT equipment and preventing performance degradation or hardware damage caused by overheating [5].

Thermal management of data centers has been transformed from using computational fluid dynamics (CFD) models to using data-driven paradigms. Although CFD offers high spatial resolution, its high computational cost is required for dynamic environmental parameter control [6]. Recently, sensors have been adopted to feed data to machine learning models to predict thermal behavior in milliseconds. Recent facilities adopt dense sensor arrays to capture the non-linear relationship between server workload, cooling airflow, and ambient humidity, which are critical for preventing local hot spots in high-density racks [7]. Advances have been further enabled by using Deep Reinforcement Learning (DRL) and Edge-Fog computing architectures to optimize cooling power usage effectiveness (PUE). By processing sensor data closer to the source (at the rack level), these systems adjust parameters dynamically without the latency that is caused by centralized cloud processing. Therefore, predictive humidity analysis is used to prevent equipment corrosion and electrostatic discharge and extend the operating life of equipment while reducing the energy consumption of humidification systems [8].

Conventional cooling control strategies, such as proportional–integral–differential (PID) control [9], rely on responsive mechanisms based on static temperature thresholds [10], meaning that cooling systems of data centers are activated only when sensor readings exceed preset limits [11]. The cooling system with this strategy suffers from response lag and cannot respond to peak heat loads, increasing the risk of instantaneous local hotspots that threaten equipment safety [12]. To avoid worst-case scenarios, operators adopt conservative thresholds, leading to excessive cooling and substantial energy waste [13]. Sensor measurement deviations or drift further compromise the efficient control, resulting in diminished system performance [14].

Recent advances in the IoT, big data, and AI have created new opportunities for intelligent thermal management in data centers [15]. Large-scale sensor networks enable real-time monitoring of thermal conditions, forming the basis for constructing digital twin models that dynamically replicate the operational state of physical data centers, providing a platform for complex analysis and simulation [16]. Leveraging machine learning, particularly DR time-series prediction models, enables the extraction of hidden patterns from numerous historical sensor data for accurate forecasting of future temperature trends [17]. Long short-term memory (LSTM) networks, a specialized type of recurrent neural network, are particularly well-suited for this task [18].

Beyond traditional data center architectures, intelligent thermal systems are increasingly implemented based on advanced predictive modeling and proactive control logic to manage complex energy dynamics. Data-driven proactive strategies significantly mitigate operational risks by anticipating thermal fluctuations before they manifest. Recent research highlights the efficacy of multi-scale predictive models in optimizing resource allocation and enhancing system resilience against volatile load profiles [19]. By situating data center cooling with intelligent energy management, the integration of high-density sensor data with robust forecasting algorithms offers an incremental improvement and industrial-scale thermal stability.

Their unique gating mechanisms enable effective learning and retention of long-term dependencies in time-series data. Given the strong temporal correlations and periodic characteristics of temperature and humidity in data centers, LSTMs provide robust predictive capabilities of temperature and humidity. By forecasting rack-level temperatures, cooling strategies can transition from reactive post-event remediation ones to proactive pre-event planning ones. Such cooling strategies also enable the preemptive allocation of cooling resources, dynamic adjustment of fan speeds, and precise control of chilled-water valves,

thereby achieving on-demand cooling that maximizes energy efficiency while ensuring operational safety.

Despite significant advancements in data center structures, the inherent latency between heat generation by high-performance computing loads and the subsequent cooling response from automated systems must be addressed. Traditional reactive cooling models often fail to reflect the non-linear relationship between temperature and humidity changes, leading to inefficient energy consumption and the formation of localized hot spots. To address such challenges, a predictive cooling system is developed in this study based on the principles of model predictive control (MPC) and real-time dynamic optimization (RTDO), based on a data-driven digital twin architecture powered by high-density sensor arrays. This system is designed to predict environmental fluctuations over a short-term horizon, allowing for proactive cooling adjustments. By forecasting thermal behavior using neural networks, the system enables proactive hardware adjustments. A structured sensor deployment strategy is adopted for precise data acquisition by implementing a neural network-based forecasting model for real-time environmental analysis and validating the energy efficiency and thermal stability of the proposed system against conventional threshold-based baselines [20].

2. Methodology

A multi-stage systematic workflow was employed for real-time data acquisition and distributed sensor arrays. The raw data underwent localized preprocessing at the Fog layer to ensure high fidelity through denoising and spatial–temporal fusion. The processed data was used to train and update a neural network-based Digital Twin. This model generates the prediction of temperature and humidity. The prediction was fed for control optimization to adjust the cooling intensity of the computer room air handler (CRAH) before reaching thermal thresholds. The structured method was adopted to ensure scalable and responsive control (Figure 1) [19].

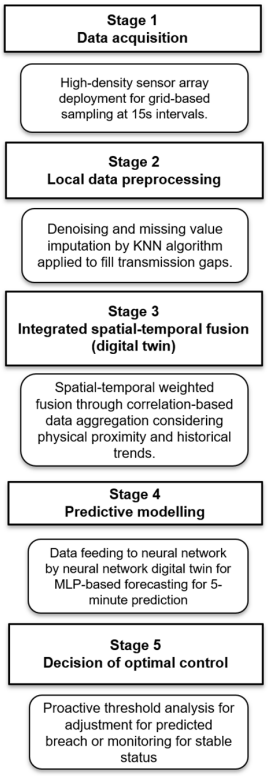


Figure 1. Flowchart of the developed predictive cooling system, integrating the spatial–temporal fusion logic and the end-to-end data processing.

2.1. System Architecture

We developed a cooling optimization system for data centers that integrates perception, analysis, decision-making, and execution. Its main objective is to enable refined and proactive management of cooling resources through precise environmental sensing and accurate modeling of future states. The architecture of the system is shown in Figure 2.

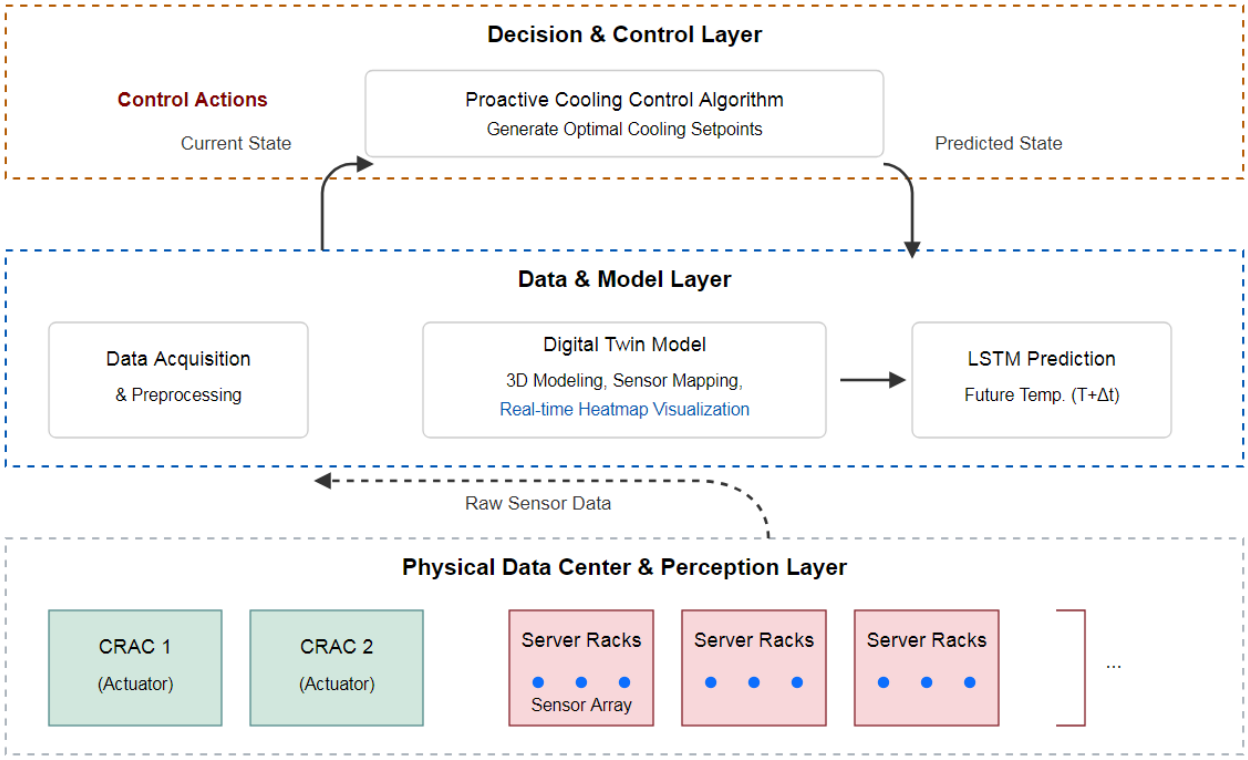


Figure 2. Architecture of the cooling optimization system developed in this study.

The system operates in a closed-loop architecture comprising perception, data and modeling, decision-making and control, and physical execution layers.

- Perception layer: Consisting of distributed temperature and humidity sensor arrays deployed throughout the data center, the perception layer promptly captures environmental status information.
- Data and model layer: Responsible for data collection, transmission, storage, and processing, this layer contributes to the construction of a digital twin of the data center and the implementation of the LSTM prediction model to support decision-making.
- Decision-making and control layer: A predictive cooling control algorithm runs in this layer to generate optimal instructions for cooling equipment based on both current conditions and future forecasts.
- Physical execution layer: Comprising cooling infrastructure such as CRAC units, this layer executes control instructions and regulates the physical environment of the data center.

To build a high-fidelity digital twin of the thermal environment, the deployment location and density of sensors are critical. Sensors in the developed cooling system are deployed following the comprehensive coverage and key monitoring principle to build a three-dimensional network, as shown in Figure 3. Figure 3 presents a symbolic description of the facility's thermal status. By mapping discrete sensor readings onto a 3D coordinate system, the Digital Twin provides an intuitive interface for monitoring hotspots. The specific color scale is calibrated to the ASHRAE-recommended thermal envelope, and the

interpolation parameters are optimized to reflect real-world airflow patterns within the containment zones.

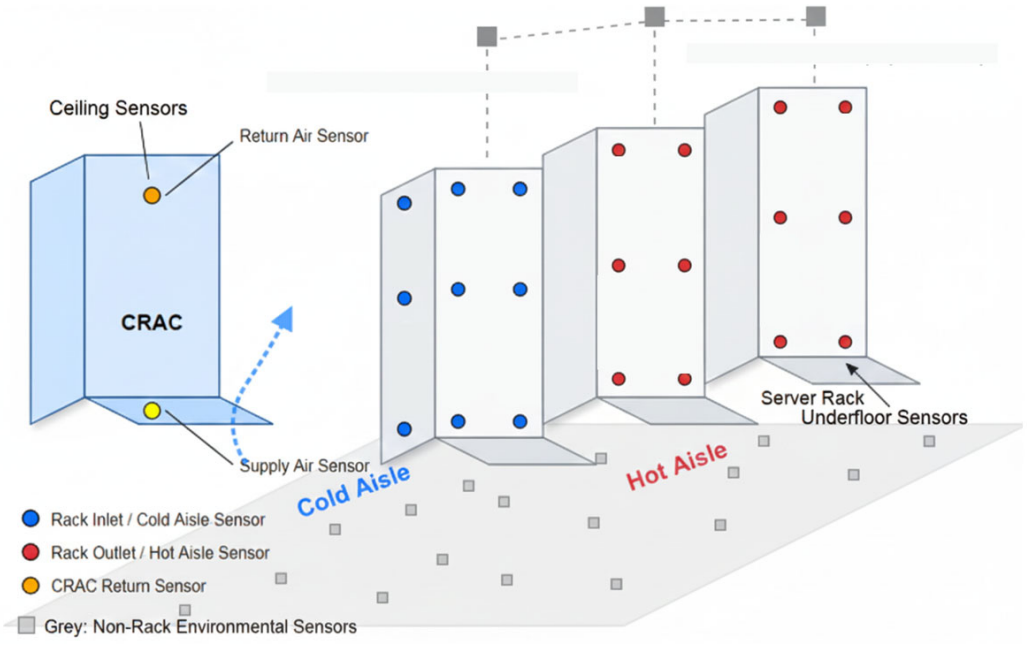


Figure 3. Representative 3D thermal mapping snapshot generated by the Digital Twin. This visualization illustrates the interpolated temperature field across the server racks. The color gradient represents the temperature range from 18 °C (deep blue/cool) to 30 °C (bright red/hot). The spatial distribution is generated using Kriging interpolation based on 120 distributed sensor nodes, providing a continuous thermal state estimation for proactive cooling adjustments.

Rack-level sensor deployment is the most critical in the cooling system. In the developed system, sensors are installed vertically according to the height of each server rack, positioned at the front (air inlet) and rear (air outlet) of the racks. Sensing points are placed at the top, middle, and bottom of each rack, resulting in a minimum of six temperature and humidity sensor arrays (3×2). For high-density racks, additional sensing points in the middle are added to capture vertical temperature gradients caused by uneven server loads, which are essential for identifying potential hotspots.

Sensors are deployed along the cold and hot aisles. Sensors in cold aisles monitor whether chilled air supplied by computer room air conditioning (CRAC) units effectively reaches server inlets, while those in hot aisles measure the temperature of exhaust air for evaluating cooling performance and calculating IT heat load. Temperature and humidity sensors are installed at the return air inlet and outlet of each CRAC unit. Return air temperature serves as the reference for formulating control strategies, whereas outlet air temperature is used to monitor CRAC cooling capacity. The difference between these two values is a critical indicator of CRAC efficiency estimation. In addition, additional sensors are deployed in non-rack areas of the data center, such as beneath the ceiling, under the floor, and at other key positions, to monitor overall environmental conditions and prevent issues such as airflow short circuits or stagnation.

2.2. Data Acquisition and Preprocessing

Efficient and stable data acquisition and communication is essential to connect the physical and digital domains. In this study, we employed industrial-grade digital temperature and humidity sensors, with their parameters presented in Table 1. High precision and low drift in sensing data are essential requirements for ensuring data quality. The system adopts a hierarchical architecture in which sensors are connected to the data acquisition

gateway via the Recommended Standard (RS) 485 using the Modbus remote terminal unit (RTU). The gateway performs protocol conversion and uploads data to the central server through Ethernet using either Modbus transmission control protocol (TCP) or Message Queuing Telemetry Transport (MQTT). Such a design offers scalability, simplified wiring, and strong anti-interference capability. A time-series database is constructed in the central server for efficient storage and querying of large volumes of sensor data. Leveraging this database, a data processing and visualization module performs model training and system monitoring.

Table 1. Technical parameters of temperature and humidity sensors.

Parameter	Value	Remark
Temperature measurement range	−20 to +80 °C	Covers all possible operating temperatures in data centers
Temperature measurement accuracy	±0.2 °C	Meets the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) TC 9.9 requirements for A1–A4 environmental classes
Humidity measurement range	0 to 100%	Measures relative humidity (RH)
Humidity measurement accuracy	±2% RH	—
Sampling frequency	1/60 Hz (once per minute)	Balances data timeliness and processing load
Communication interface	Modbus RTU/TCP	Industry standard protocols, easy to integrate
Long-term stability	<0.03 °C/year	Ensures reliability of long-term monitoring data

The raw data collected from sensors often contain noise, missing values, or outliers, which significantly degrade performance. Therefore, rigorous data preprocessing is indispensable for using the data. For short-term missing data points, linear interpolation or averaging of adjacent time points is applied, while for prolonged continuous data loss, which might indicate sensor malfunction, the affected period is flagged and excluded from the training set. Outliers are identified using statistical methods when values exceed the normal fluctuation range. The outliers are corrected using the Z-score method as follows.

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

Here, x represents the observed value, μ is the mean of the dataset, and σ is the standard deviation. Values exceeding a predefined threshold ($Z > 3$) are regarded as outliers.

Neural networks, such as long short-term memory (LSTM) models, are highly sensitive to the scale of input data. To accelerate model convergence and enhance prediction accuracy, all input variables—including temperature, humidity, and IT load—are normalized to the interval [0, 1] using the min–max normalization method.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}$$

Here, X_{max} and X_{min} represent the maximum and minimum values of the training dataset. After prediction, reverse normalization is applied to restore the outputs to their original scale.

To enrich the model's input, periodic features are extracted from timestamps: hour of the day (0–23), day of the week (0–6), and working-day indicators. They are separately encoded and incorporated as additional variables. The time-series data are then transformed into a supervised learning format. Specifically, the past L in time steps is used to predict

the future L out time steps. The original time series T is thus represented as sample pairs (X,Y), defined as follows.

$$X_i = \{t_i, t_{i+1}, \ldots, t_{i+L_{in}-1}\} \tag{3}$$

$$Y_i = \{t_{i+L_{in}}, \ldots, t_{i+L_{in}+L_{out}-1}\} \tag{4}$$

On this basis, the model learns the mapping relationship from X to Y for accurate prediction of future thermal status.

In this study, the input sequence L_{in} was set to 60 min, and the prediction horizon L_{out} to 30 min. These values were determined through a sensitivity analysis to balance the predictive accuracy (root mean square error (RMSE) and the coefficient of determination (R^2) against the mechanical response time of the cooling infrastructure.

2.3. Predictive Modeling and Digital Twin Construction

We adopt the LSTM network as the core prediction model. LSTM regulates the flow of information through its input, forget, and output gates to effectively address the gradient vanishing and explosion problems, which are commonly encountered by traditional recurrent neural networks when processing long sequences.

While gated recurrent units (GRU), temporal convolutional networks, and Transformer-based models were considered, LSTM was selected for this study. Compared with GRUs, LSTMs contain a dedicated memory cell, which provides granular control over long-term information retention, which is important, given the slow-moving thermal inertia and long-cycle periodic patterns observed in data centers [21]. While Transformer models capture global dependencies through attention mechanisms, they require larger datasets to outperform recurrent models in short-to-medium term forecasting (30–60 min) and are computationally intensive for real-time edge-layer deployment [22]. Therefore, we selected an LSTM encoder–decoder architecture to balance between predictive stability and the millisecond-level inference latency required for proactive cooling adjustments.

The computation of an LSTM unit at a time step t proceeds as follows.

At the forget gate, information to be discarded from the cell state is determined.

$$f_t = \sigma\Big(W_f \cdot [h_{t-1}, x_t] + b_f\Big) \tag{5}$$

At the input gate, new information to store in the cell state is identified.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{7}$$

Cell status is updated as follows.

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t \tag{8}$$

At the output gate, values to output are determined.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{9}$$

$$h_t = o_t \odot tanh(C_t) \tag{10}$$

Here, x_t is the input at time t, h_{t-1} is the hidden state at the previous time step, and C_{t-1} is the previous cell state, W and b denote the weight matrices and bias terms of each gate, σ is the sigmoid activation function, and $\odot$ represents element-wise multiplication.

The prediction model employs an encoder–decoder architecture with stacked LSTM layers to enhance its ability to capture complex temporal patterns (Figure 4). The encoder, composed of two stacked LSTM layers, encodes the input time series into a fixed-length context vector that summarizes the information contained in the sequence. The decoder, consisting of two stacked LSTM layers, receives this context vector as its initial state and progressively generates the prediction sequence. Finally, a fully connected layer maps the decoder's output dimension to the target prediction dimension, producing the final predicted values.

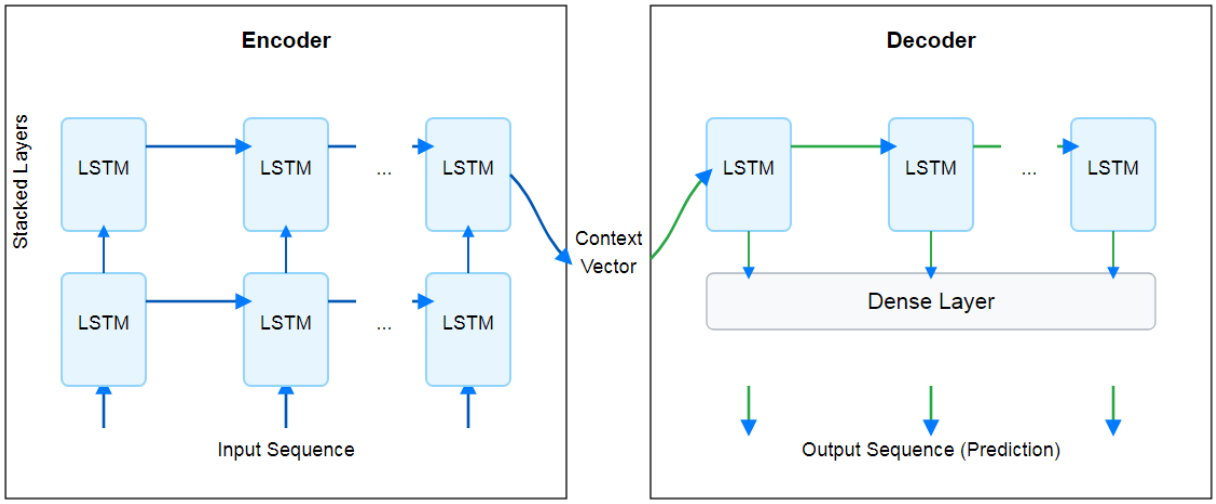


Figure 4. LSTM model architecture comprising an encoder and decoder.

The prediction model employs the mean square error (MSE) as the loss function to quantify the difference between predicted and true values.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_{true}^{(i)} - Y_{pred}^{(i)})^2 \quad (11)$$

Here, N denotes the number of samples, $Y_{true}^{(i)}$ represents the ground truth, and $Y_{pred}^{(i)}$ is the corresponding predicted value.

For optimization, the Adaptive Moment Estimation (Adam) optimizer is adopted, as it integrates the advantages of the Adaptive Gradient Algorithm and root mean square propagation (RMSProp). The Adam optimizer adaptively adjusts the learning rate and provides stable and efficient performance during training. The dataset is split into three subsets: 70% for training, 15% for validation, and 15% for testing, ensuring effective model learning and robust performance evaluation, and maintaining the temporal integrity of the environmental time series to prevent look-ahead bias. In accordance with the time-series validation principles [23], Min-Max scaling and outlier detection parameters were derived from the training partition and applied to the validation and test sets. This ensures that the model's performance reflects true out-of-sample generalization rather than temporal leakage.

This algorithm transforms future temperature forecasts generated by the LSTM model into specific, executable control actions for the CRAC unit for accurate prediction and control. The algorithm calculates the minimum cooling output required to address anticipated heat loads while ensuring compliance with safety constraints. The algorithm workflow is illustrated in Figure 5.

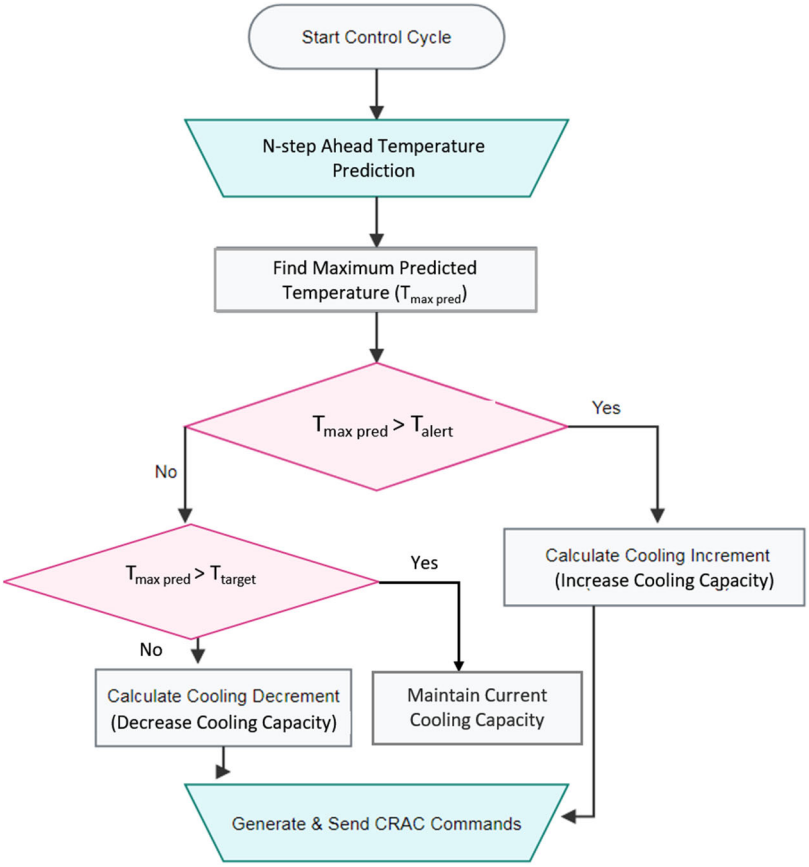


Figure 5. Workflow of the cooling control algorithm.

In each control cycle, the algorithm acquires the predicted temperature sequence pred for all key measurement points over the future L_{out} time steps from the LSTM model. It then identifies the maximum predicted value, T_{max_pred}, and determines the target cooling capacity. If T_{max_pred} exceeds the warning threshold T_{alert}, the system anticipates a risk of overheating and increases cooling output. Conversely, if T_{max_pred} is significantly lower than the target operating temperature T_{target}, the system identifies overcooling and reduces cooling capacity. When T_{max_pred} falls within the optimal range [T_{target}, T_{alert}], the current cooling level is maintained. The algorithm then calculates new setpoints for the CRAC unit, focusing on two primary parameters: fan speed and chilled-water valve opening. Fan speed affects the volume of cold air delivered, with an approximate cubic relationship between fan speed, airflow, and energy consumption, making fine-tuning essential for energy efficiency. The chilled-water valve opening regulates the flow of chilled water through the CRAC coil, thereby determining outlet air temperature. Control instructions are generated using either a simplified thermodynamic model or a pre-trained reinforcement learning policy network, which captures the effects of different control combinations on the thermal field.

A simplified proportional–integral–derivative (PID) adjustment logic is expressed as follows.

$$Error = T_{max_pred} - T_{target} \tag{12}$$

Here, T_{max_pred} is the maximum predicted temperature within the forecast horizon, obtained from the LSTM model, T_{target} is the desired operating temperature setpoint for the data center, and *Error* is the deviation between the predicted maximum temperature and the target temperature. A positive error indicates overheating risk, while a negative error suggests overcooling.

$$\Delta_{Control} = K_p \cdot Error + K_i \int Error \cdot dt \tag{13}$$

Here, $\Delta_{Control}$ is the adjustment amount applied to cooling parameters (e.g., fan speed or chilled-water valve opening), K_p is the proportional gain coefficient that determines the immediate response strength to the error, K_i is the integral gain coefficient that accounts for accumulated error over time to eliminate steady-state deviations, $\int Error \cdot dt$ is the integral of the error over time, representing the cumulative deviation from the target temperature.

$$New_Setpoint = Current_Setpoint + \Delta_{Control} \tag{14}$$

Here, *New_Setpoint* is the updated control value for the CRAC unit (e.g., adjusted fan speed or valve opening), and *Current_Setpoint* is the existing control value before adjustment.

The digital twin model of the data center is designed to provide a real-time, visual reproduction of its thermal status. The construction process of the model involves four steps.

First, computer-aided design or building information modeling (BIM, FreeCAD 1.0) software is used to establish a geometric model of the data center interior to accurately represent physical dimensions, including room dimensions, rack layout, CRAC units, and ventilation openings. Second, each sensor is assigned a unique identifier and bound to its precise three-dimensional coordinates. Third, the data acquisition module is installed to continuously stream sensor data, including timestamps and sensor identifiers, to the digital twin platform. Fourth, spatial interpolation algorithms, such as inverse distance weighting (IDW) or Kriging, are applied to estimate temperatures at unsensed positions based on known readings, since a continuous thermal field cannot be constructed with discrete sensor data. The IDW algorithm is expressed as the following equation.

$$T(x_0) = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} T(x_i)}{\sum_{i=1}^{n} \frac{1}{d_i^p}} \tag{15}$$

Here, $T(x)$ is the estimated temperature at location x, T_i is the temperature measured at sensor i, $d(x,x_i)$ is the distance between x and sensor i, and p is the weighting power parameter.

The resulting continuous temperature field is rendered onto the 3D model using color mapping technology to generate an intuitive dynamic heat map. For example, in Figure 6, blue indicates low-temperature regions, while red highlights high-temperature areas. This visualization enables the observation of the real-time thermal distribution from any angle and rapidly identifies hot or cold spots.

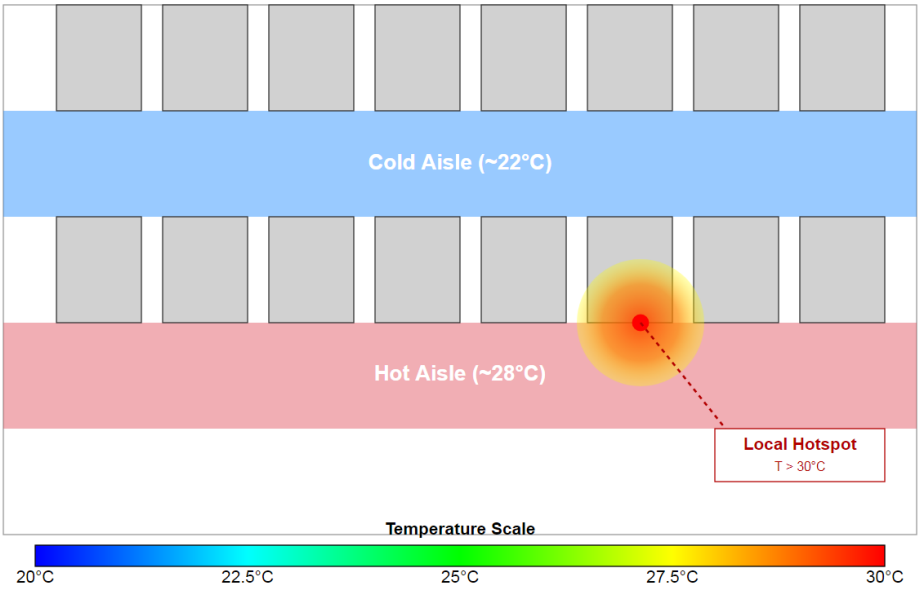


Figure 6. Example of a heatmap drawn in the Digital Twin model.

2.4. Control Logic and Security Constraints

To ensure the stability and security of the system, constraints are embedded in the control logic. First, no control decision is permitted to cause the immediate temperature reading of any sensor to exceed the upper limit of the equipment's safe operating range as defined by ASHRAE. In addition, restrictions are imposed on the adjustment rates of fan speed and valve opening to prevent mechanical shock to the cooling equipment and to avoid drifts in the system. The control logic guarantees a minimum level of cooling, ensuring adequate ventilation even under low-load conditions and providing resilience against sudden load fluctuations. Finally, in the event of significant prediction errors or communication failures, the system automatically and seamlessly reverts to the existing PID control mode based on the monitored air temperature to ensure operational safety.

The performance evaluation in this study was conducted based on data-driven simulation. High-fidelity historical sensor data (IT load and thermal patterns) were used for a thermodynamic model of the data center's cooling infrastructure and a direct comparison between the baseline PID control and the proactive strategy in identical environments. While the cooling response is simulated, the underlying dynamics were investigated by using real operational datasets.

2.5. Comparison with MPC and RTDO

The methodology employed in this study aligns with the principles of MPC and RTDO. The developed system aligns with MPC, which uses sensor data to project future states rather than reacting to current violations. The system continuously recalculates parameters in real-time, mirroring the RTDO's operation of maintaining optimal setpoints under changing constraints [24]. While traditional MPC often relies on explicit physical or state-space CFD-based models, the developed system utilizes a neural network Digital Twin, which is mainly data-driven. This allows the system to learn complex, non-linear thermal interactions in the data center that are difficult to characterize with physics-based equations. While RTDO solves complex differential equations, neural network-based prediction offers near-instantaneous inference, making it more appropriate for high-frequency edge deployments where resource-constrained nodes must make rapid decisions. While traditional physics-based models are rigid to specific layouts, the data-driven system maintains robustness by continuously recalculating parameters in real-time, mirroring RTDO operations to maintain optimal setpoints under changing IT loads and constraints.

The proposed proactive strategy was benchmarked against traditional PID control, as it has been widely deployed in data centers. Advanced methods such as MPC and RL require high-fidelity physical models or extensive training periods that are difficult to implement in live environments. The strategy in this study was developed to provide the look-ahead capability of MPC using a data-driven LSTM, which is easier to deploy than physical models, and offer the high-quality state predictions that could eventually serve as the input for a more complex RL model.

3. Experiment and Implementation

To verify the effectiveness of the cooling system developed in this study, we carried out a four-month comparative test (from January to April 2025).

3.1. Operating Conditions

We selected different operating conditions of the developed cooling system to evaluate the performance of the cooling system. The parameters of the cooling environment are presented in Table 2. 6 sensors per rack were implemented to ensure high-fidelity

monitoring, exceeding the identified minimum threshold of 2 sensors required for stable predictive modeling.

Table 2. Parameters of data center environment.

Category	Parameter	Value
Physical space	Room size	15 × 10 × 3.5 m
	Number of racks	20 (2 rows, 10 racks per row)
	Rack size	2U standard rack (2 × 4.83 m)
IT equipment	Total IT load	100 kW (average), peak 150 kW
	Server type	Deployment of 1U (44.45 mm)/2U servers (88.9 mm in height)
	Load pattern	Simulated real-world daily and weekly cyclic variations
Cooling system	Cooling type	Chilled-water CRAC
	Number of CRAC units	2 units, with N + 1 redundancy
	Layout	Hot/Cold aisle containment
Sensor network	Total number of sensors	150 sensors (6 per rack, 4 for CRAC units, 26 other locations)
	Dataset time span	1 January 2025–30 April 2025 (hourly collecting data)

3.1.1. Baseline Load Operating Condition

In this condition, the stable operation of a data center during nighttime or non-peak business hours is performed. The IT load fluctuates within a narrow range of ±10% of the average value of 100 kW, providing a baseline for evaluating cooling system performance under low power consumption while avoiding excessive cooling capacity. During the night period (00:00–07:00), temperature variations are generally mild. The interval from 6 January to 8 January 2025, was selected in this condition, during which the ambient temperature in the computer room remained stable without abrupt changes (Table 3).

Table 3. Environmental data in baseline load operating condition.

Time	Temperature (°C)	Relative Humidity (%)	Remarks
7 January 2025 08:00:18	20.8	18.0	Weekday morning; the load begins to stabilize.
7 January 2025 12:00:48	20.6	17.3	Stable operation during the day.
7 January 2025 18:00:26	20.8	18.0	End of workday; load slightly decreases.
7 January 2025 23:00:21	20.9	18.1	Night-time low; ambient temperature is stable.

3.1.2. Humidity Stability and Control

Thermal management requires much energy consumption. Maintaining stable relative RH is essential for the reliability of the management system. The system monitors humidity sensors across the rack arrays to preemptively adjust the computer room air handler humidification cycles. The results of Table 4 show that the system effectively reduced humidity fluctuations. While the related strategy occasionally presented RH above 65% (increasing the risk of hygroscopic dust failures), the neural network-driven model maintained RH from 45 to 55% with a standard deviation of 1.8%. These results show that the prediction results of the Digital Twin are effectively used for moisture and temperature control. By

predicting humidity trends based on the intake air temperature and server fan speeds, the system prevents the sawtooth humidity patterns needed for threshold-based on/off controls, reducing the mechanical wear on humidification components [25].

Table 4. Humidity control performance.

Metric	Reactive Baseline	Developed System
Mean RH (%)	52.4	49.8
Maximum RH (%)	68.2	54.1
Minimum RH (%)	36.5	44.7
Standard deviation (%)	7.4	1.8
Stability in ASHRAE range	82% of time	99.4% of time

3.1.3. Peak Load Condition

At peak business periods, such as the initiation of large-scale computing tasks or promotional events on e-commerce platforms, the IT load rapidly increased to 150 kW, which served as a test of the cooling system's responsiveness and ability to identify hot spots. Data from 18:00 to 21:00 on 25 January 2025 were chosen for this condition, during which the temperature exhibited a sharp and rapid rise, reflecting the sudden escalation in heat load (Table 5).

Table 5. Environmental data in peak load operating conditions.

Time	Temperature (°C)	Relative Humidity (%)	Remarks
25 January 2025 18:00:13	22.6	21.5	Load begins to climb rapidly.
25 January 2025 19:00:29	23.2	21.2	Load approaches peak, temperature rises sharply.
25 January 2025 20:00:46	23.2	21.3	Load maintained at peak, testing the cooling system.
25 January 2025 21:00:58	21.5	22.9	After the peak, the load begins to fall back.

3.1.4. Dynamic Load Fluctuation Condition

We captured frequent and irregular IT load variations, including dynamic migration of virtual machines in cloud computing environments. Data from 27–28 January 2025 were selected, during which both temperature and humidity demonstrated significant and irregular fluctuations (Table 6).

Table 6. Environmental data in a dynamic load fluctuation condition.

Time	Temperature (°C)	RH (%)	Remarks
27 January 2025 17:00:41	19.9	13.3	Both temperature and humidity are at a low point.
27 January 2025 20:00:18	20.5	14.6	Temp and humidity rebound slightly in a short period.
28 January 2025 12:00:47	20.4	16.0	Temp and humidity rise again, with frequent fluctuations.
28 January 2025 15:00:14	20.1	14.5	Temp and humidity fall back again, showing an oscillating trend.

3.1.5. Equipment Maintenance Condition (N + 1 Redundancy Test)

To predict the temporary shutdown of a single CRAC unit, we tested the collaborative capacity of the remaining cooling units and the robustness of the system under an $N + 1$ redundancy configuration. Anomalous data from 10 January 2025 were selected to represent this condition. During this period, relative humidity exhibited an instantaneous peak of 34.20% (Table 7). Such anomalies were caused by equipment start–stop events or sensor disturbances and were used to test the filtering and error-correction capabilities of the control logic.

Table 7. Environmental data in equipment maintenance condition.

Time	Temperature (°C)	Relative Humidity (%)	Remarks
10 January 2025 12:00:12	20.6	18.1	The system is operating stably.
10 January 2025 13:00:35	20.4	23.5	Humidity begins to rise abnormally.
10 January 2025 14:00:58	20.5	34.2	Humidity reaches an abnormal peak, testing system stability.
10 January 2025 15:00:14	20.2	24.8	After the abnormal state, the humidity drops back quickly.

3.2. Control Logic and Execution

The implemented controller operates as a supervisory predictive control. The execution logic is summarized in the pseudocode as follows (Algorithm 1) [26].

Algorithm 1: Proactive Cooling Control Logic

1. Input: Current states S_t, Forecast horizon L_{out} (30 min) from LSTM;
2. State aggregation: Mean rack inlet temperature $T_{in,avg}$ and local hotspot T_{max};
3. Setpoint generation: If $T_{pred} > T_{target}$, calculate proactive adjustment as $\Delta\text{Setpoint} = \text{Kp}(T_{pred} - T_{target})$.
4. Actuator mapping: Fan speed (f): $f_{t+1} = f_t + \Delta f$, where $\Delta f \propto \Delta\text{Setpoint}$.
5. Valve opening (v): $v_{t+1} = v_t + \Delta v$, subject to ramping constraint $|\Delta v| < 5\%/\text{min}$
6. Fallback: If $\left|T_{pred} - T_{Target}\right| < 1.5\ ^\circ\text{C}$, switch to the baseline PID.

This mapping ensures that the high-latency mechanical response of the CRAC unit is initiated 30 min before the predicted thermal peak.

To maintain physical transparency, the effective heat transfer and power relations are governed by parameters shown in Table 8. These values were estimated using steady-state operational data and validated against the facility's historical cooling logs.

Table 8. Thermodynamic and system parameters.

Parameter	Description	Assumption
Rth	Effective thermal resistance	0.05–0.08 K/W
H	Convective heat transfer coefficient	20–50 W/(m^2·K)
η_{fan}	Fan power relation	Cubic law ($P \propto f^3$)
$\in$	Radiative emissivity (fins)	0.85 (constant)

3.3. Implementation and Reproducibility Details

To ensure the reproducibility of this study, technical parameters for data preprocessing, model architecture, and training are presented in Table 9. The dataset was split into training (70%), validation (15%), and testing (15%) sets using a strict time-ordered sequence to prevent data leakage and preserve the temporal dependencies inherent in thermal dynamics. Feature sets include raw sensor values and time-of-day/day-of-week indicators, which were encoded using sine–cosine cyclic transformation to capture periodic load patterns. The LSTM model adopted a direct mapping strategy for multi-step forecasting to avoid the error propagation typically associated with recursive outcomes.

Table 9. Summary of implementation and reproducibility details.

Category	Parameter	Specification
Dataset	Total duration/samples	122 days/175,680 samples
	Sampling interval	1 min (1/60 Hz)
	Data split (70/15/15)	Strict time-ordered (chronological)
Features	Input/target channels	120 rack sensors + 4 ambient sensors
	Temporal encoding	Sine/cosine encoding for hour and day
Model	Architecture	Seq2Seq LSTM (encoder–decoder)
	L_{in} (input length)	60 steps (60 min)
	L_{out} (horizon)	30 steps (30 min)
	Hidden layers/sizes	2 layers/128 units per layer
Training	Optimizer/learning rate	Adam/0.001 (with Decay)
	Batch size/epochs	64/100
	Stopping criterion	Early stopping (patience = 10)
	Random seed	42

4. Results and Discussion

From the four-month operation of the cooling system, we assessed temperature prediction accuracy, hot spot suppression, energy consumption performance, system dynamic response, and robustness as follows.

4.1. Temperature Prediction

The predictive performance of the LSTM model in the baseline load operating condition was assessed for rack outlet temperature, which is the foundation of proactive control. An outlet sensor located at the top of a high-density rack was selected for analysis. Figure 6 illustrates the comparison between the model's predicted outlet temperatures for the subsequent 30 min and the actual measurements during the final week of the test set. The predicted curve aligned with the observed values (Figure 7). The model captured both the day–night periodic fluctuations in temperature and the random disturbances induced by variations in IT load. To further demonstrate the model's predictive capability, three distinct time intervals, stable, rising, and fluctuating periods, were randomly selected from the dataset, and the predicted values were compared with the corresponding ground truth.

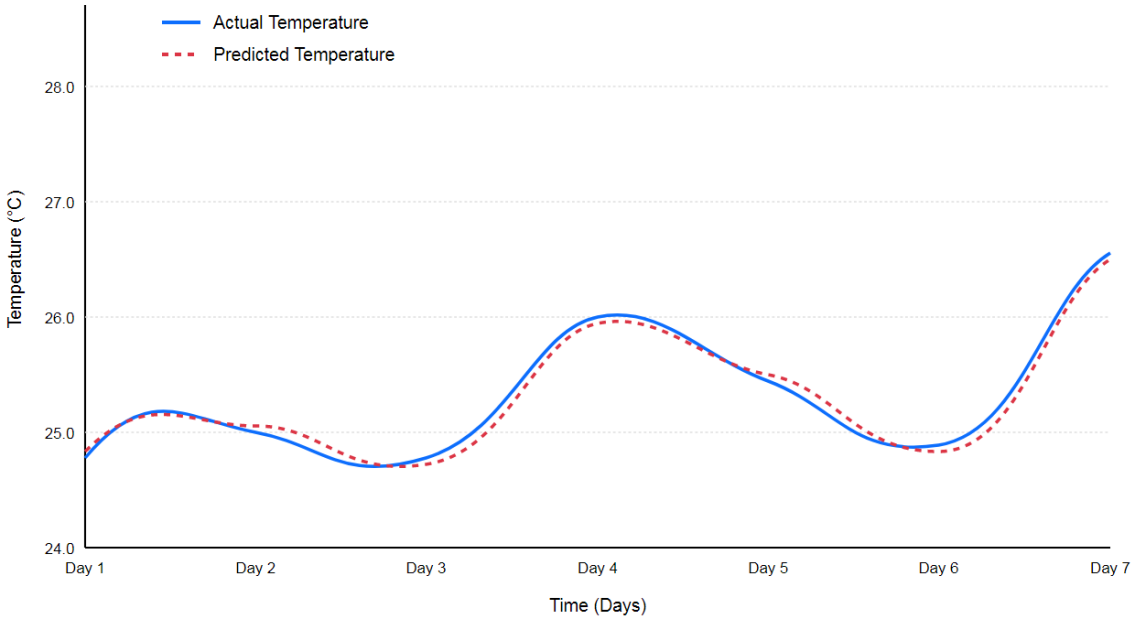


Figure 7. The temperature prediction and the actual result of the LSTM model at the outlet sensor located at the top of a rack.

The sensitivity analysis was conducted during the model development phase to evaluate the impact of different L_{in} and L_{out} on RMSE and R^2. As shown in Table 10, various configurations were tested to find the optimal balance between accuracy and lead time.

Table 10. Sensitivity analysis of L_{in} and L_{out} on model performance.

L_{in} (Minutes)	L_{out} (Minutes)	RMSE (°C)	R^2
30	30	0.38	0.962
60	15	0.18	0.991
60 (selected)	30 (selected)	0.25	0.985
60	60	0.44	0.948
120	30	0.26	0.983

Table 11 shows the model's performance during a stable operating period, when both IT load and ambient temperature remained relatively constant. Under these conditions, the prediction error was negligible, confirming the model's high accuracy in stable environments.

Table 11. Prediction error of outlet temperatures under the baseline load operating condition.

Time	Actual Temperature (°C)	Predicted Temperature (Next 30 min) (°C)	Absolute Error (°C)
04 January 2025 03:00:50	20.8	20.8	0.0
04 January 2025 11:00:37	20.9	20.9	0.0
04 January 2025 16:00:06	21.0	21.0	0.0
04 January 2025 22:00:00	20.9	20.9	0.0

Table 12 presents the model's predictive ability in the peak load condition. Even when the temperature rose rapidly within a short time frame, the model closely tracked the changes in the actual values, confirming its effectiveness in dynamically varying situations.

Table 12. Prediction error of outlet temperatures under the peak load condition.

Time	Actual Temperature (°C)	Predicted Temperature (Next 30 min) (°C)	Absolute Error (°C)
21 January 2025 16:00:01	22.1	21.9	0.2
21 January 2025 17:00:15	21.9	21.9	0.0
21 January 2025 18:00:28	21.9	21.7	0.2
21 January 2025 19:00:53	21.9	21.7	0.2

Table 13 illustrates the model's performance in the dynamic load fluctuation condition, particularly at lower temperatures. In this condition, the model's capacity to capture nonlinear and irregular changes was tested. The results showed that the model maintained high accuracy even under fluctuating operating conditions.

Table 13. Prediction error of outlet temperatures under the dynamic load fluctuation condition.

Time	Actual Temperature (°C)	Predicted Temperature (Next 30 min) (°C)	Absolute Error (°C)
28 January 2025 13:00:57	20.1	20.1	0.0
28 January 2025 14:00:05	20.1	20.1	0.0
28 January 2025 15:00:14	20.1	19.9	0.2
28 January 2025 16:00:26	19.9	20.2	0.3

For model selection, the LSTM model was compared with standard GRU and 1D-Transformer-based models. The LSTM model demonstrated a higher stability in prediction even during peak load fluctuations due to its complex gating structure, which effectively captures the non-linear relationship between server workload and ambient humidity. To assess prediction accuracy from a broader perspective, three evaluation metrics were calculated: mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2). As shown in Table 10, the RMSE was only 0.25 °C, and the R^2 value reached 0.985. These results confirmed that the LSTM model delivered accurate predictions, sufficient to support subsequent proactive cooling control decisions. To assess the model's generalizability across the data center, a spatial performance analysis was conducted using all 120 sensors (six per rack across 20 racks). The results demonstrate strong consistency, with a median R^2 of 0.982 and a 95th percentile RMSE of 0.32 °C, indicating that even sensors located in high-turbulence zones maintain acceptable accuracy. To evaluate robustness against temporal fluctuations in IT workloads, the analysis was extended across the full 18-day test period. A weekly breakdown of metrics (Table 14)

confirms stable performance despite diurnal load variations, with RMSE deviating by less than 0.05 °C between peak and off-peak weeks. This comprehensive evaluation ensures that predictive capability is not localized but effectively represents the facility's overall thermal dynamics.

Table 14. Temporal stability of model performance across the full test period under varying IT load patterns.

Time	Mean RMSE (°C)	Median R^2	95th Percentile RMSE
Week 1	0.24	0.985	0.31
Week 2	0.26	0.981	0.33
Overall	0.25	0.982	0.32
Week 1	0.24	0.985	0.31

The Digital Twin's fidelity was also examined over four months by analyzing error distributions across all 120 sensors. Median R^2 values consistently exceeded 0.97, demonstrating reliable generalization regardless of rack position or sensor height. Sensitivity analysis of sensor density revealed that the current high-density configuration is critical for capturing vertical temperature gradients. When the dataset was reduced to simulate lower sensor density, performance remained stable down to two sensors per rack (inlet and outlet). Below this threshold, RMSE increased by approximately 22%, establishing the one-inlet-one-outlet configuration as the minimum deployment necessary for reliable proactive control.

To evaluate the spatial generalization of the framework, a leave-one-group-out cross-validation was conducted (Table 15). Data from Racks 15–20 were excluded from the training phase and used exclusively for testing. The model achieved a median RMSE of 0.28 °C on these unseen physical locations, which was within 10% of the performance on the training racks. This indicates that the LSTM layers captured underlying thermal dynamics rather than memorizing site-specific sensor IDs. Furthermore, temporal generalization was verified through a walk-forward validation across the four-month dataset. As shown in the results in Table 15, the performance metrics remained stable as the model transitioned from the data from January to April, despite shifting IT load patterns and ambient temperature variations. This suggests the model is resilient to temporal drift and seasonal fluctuations.

Table 15. Performance comparison against baseline forecasting models.

Model	MAE (°C)	RMSE (°C)	MASE
Persistence	0.58	0.72	1
Seasonal Naïve	0.49	0.65	0.88
Proposed LSTM	0.21	0.25	0.42

To contextualize the LSTM's forecasting performance, two baselines were implemented: a Persistence model (predicting for $t + 30$) and a Seasonal Naïve model (using the previous day's value at the same time). We adopted the mean absolute scaled error (MASE) [27], which provides a scale-independent measure of forecast accuracy. The LSTM developed achieved a MASE of 0.42, significantly outperforming both the persistence (MASE = 1.0) and the Seasonal Naïve (MASE = 0.88). Furthermore, the error distribution across all 120 sensors showed that even in the 95th percentile (worst-case tails), the model maintains an RMSE below 0.35 °C, providing the reliable accuracy required for proactive hotspot mitigation.

4.2. Hotspot Suppression Effect

Hotspots are a critical concern in data center operations. To evaluate the suppression capabilities of the cooling system, we conducted a statistical analysis of the duration during which rack outlet temperatures exceeded the ASHRAE-recommended upper limit of 27 °C across the entire experimental period.

The predictive system maintained a lower mean relative humidity of 49.8%, compared with 52.4% under the reactive baseline of the PID control strategy, showing that it kept conditions closer to the target. The maximum relative humidity with the developed system was 54.1%, which is substantially lower than the 68.2% under the baseline, indicating that the new approach prevented excessive humidity peaks. At the same time, the minimum relative humidity was higher with the predictive system (44.7%) than with the baseline (36.5%), suggesting that it avoided the PID control strategy. The variability in humidity was reduced, while the standard deviation dropped from 7.4% in the baseline to just 1.8% in the predictive system, reflecting enhanced stable control. The system kept humidity within the ASHRAE-recommended range for 99.4% of the time, compared with only 82% of the time with the baseline, demonstrating a significant improvement in compliance with industry standards.

Together, these results show that the predictive system not only reduces extremes in humidity but also ensures greater stability and comfort by keeping conditions within recommended limits almost all the time. The cumulative duration of hotspot events under the conventional PID and the developed system's strategy was estimated through simulation. Under the PID control, response delays prevented timely cooling during sudden load increases, resulting in approximately 48 h of hotspot events over the four-month test period. In contrast, the control strategy of this study, which anticipates increasing temperature trends and adjusts cooling capacity in advance, nearly eliminated such events. The cumulative hotspot duration was reduced to 2 h, most of which were short enough that the temperature marginally exceeded the threshold. These results highlight the advantage of shifting from passive response to active prevention in safeguarding IT equipment.

To further characterize hotspot behavior, we analyzed the total duration and the severity of over-temperature events. Temperatures exceeding 27 °C were classified into the Grade I category (slight over-temperature, 27.1–28.0 °C), the Grade II category (moderate over-temperature, 28.1–29.0 °C), and the Grade III category (severe over-temperature, >29.0 °C). The detailed analysis results of hotspot events under the conventional PID and the developed system's strategy were summarized in Table 16.

The control strategy developed in this study reduced the total duration of hotspot events by nearly 96% and significantly decreased both their frequency and individual duration. Under the PID control, moderate (Grade II) and severe (Grade III) over-temperature events occurred, with maximum temperatures reaching 29.5 °C, thereby posing a direct threat to equipment safety. In contrast, the developed control strategy eliminated all over-temperature events above 28 °C and maintained the maximum temperature within the safety margin at 27.8 °C. According to the risk score index, such a proactive strategy reduced the potential risk of thermal load damage to equipment by 98.7%.

To manage the inherent uncertainty in deep learning predictions, the control algorithm incorporates a dual-layer safety mechanism. First, a residual monitoring system continuously compares the LSTM predictions with real-time sensor data. If the prediction error exceeds a safety threshold, the proactive control is suspended, and the system reverts to conservative PID control to ensure equipment safety. Second, actuator ramping constraints are applied to the cooling fans and chilled-water valves. By limiting the maximum rate of change per control cycle, the system filters out potential prediction noise and prevents me-

chanical oscillations. These safety margins ensure that the cooling infrastructure operates reliably even during transient periods of high model uncertainty.

Table 16. Comparison of two sensor control strategies.

Performance Metric	Parameter	PID Strategy	Developed Strategy in This Study
Frequency of occurrence	Total hotspot events (count)	12	2
	Average duration per Event (hours)	4.0	1.0
Duration (hours)	Total	48.0	2.0
	Grade I (27.1–28.0 °C)	25.5	2.0
	Grade II (28.1–29.0 °C)	16.3	0
	Grade III (>29.0 °C)	6.2	0
Severity	Highest recorded Temperature (°C)	29.5	27.8
	Average overtemperature Margin (°C)	1.3	0.4
Risk score ($\Sigma(T_{\text{actual}} - T_{\text{threshold}})$)	Degree-hours (°C·h)	62.4	0.8

4.3. Energy Consumption

Energy consumption is an important indicator for evaluating the performance of a data center's cooling system. In this study, we used the total energy consumption of the cooling system and the overall power usage effectiveness (PUE) of the data center as metrics to assess the energy consumption level. A PUE value closer to 1 indicates higher energy efficiency.

Figure 8a presents the monthly energy consumption of the cooling system under the conventional PID and the developed system's strategy. Across all months, the developed system's control strategy consistently consumed less energy than the PID control. This improvement owes to its ability to avoid the excessive cooling commonly observed in the PID control, instead achieving cooling on demand. Figure 8b shows the average PUE values over four months. The average PUE under the PID control was 1.48, whereas the forward-looking control strategy reduced this value to 1.30.

The energy-saving rate was calculated as follows.

$$Energy\ Saving\ Rate = \frac{(1.48 - 1.30)}{1.48 - 1} \times 100\% \approx 37.5\%\ (cooling\ units) \tag{16}$$

$$Overall\ PUE\ Reduction = \frac{1.48 - 1.30}{1.48} \times 100\% \approx 12.16\% \tag{17}$$

These results demonstrate that the proposed framework achieved more than 37% energy savings at the cooling system level and reduced the overall PUE of the data center by approximately 12%, confirming a highly significant energy-saving effect. The 37.5% reduction in cooling energy consumption was estimated from January to April. During the period, diurnal load variations and early seasonal transitions were observed. Energy savings are attributed to the increased efficiency of the proactive control logic due to reduced over-cooling and fan oscillations, rather than external climatic factors. Year-long assessments need to be quantified to elucidate the impact of extreme summer ambient temperatures on the system's total annual efficiency.

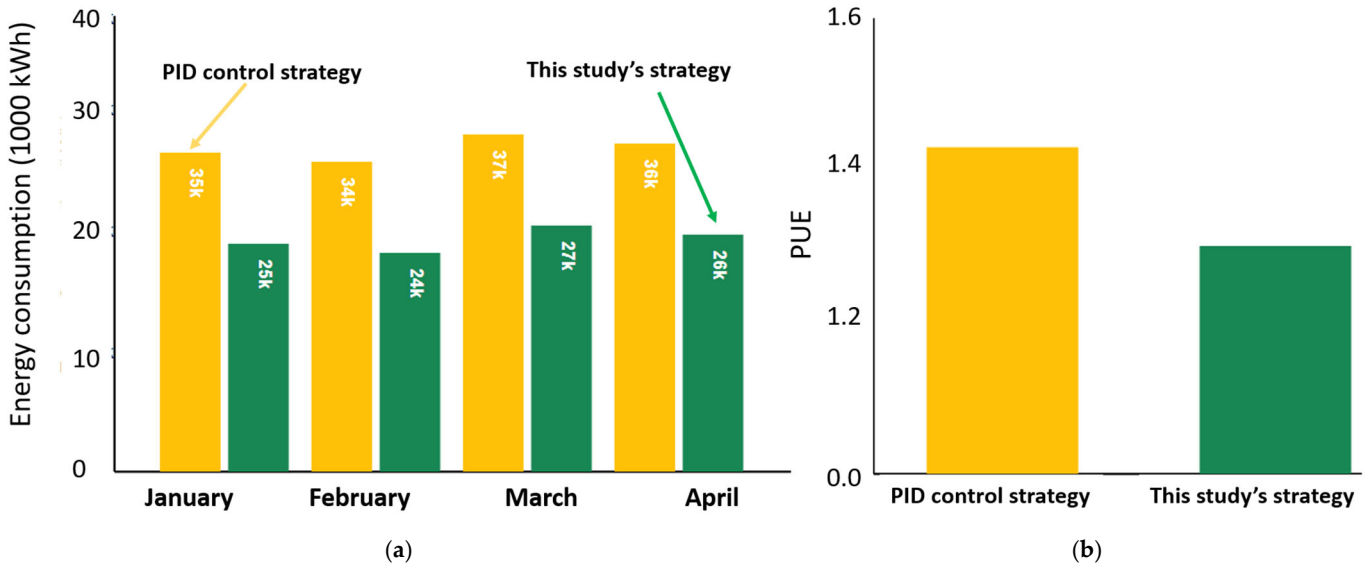


Figure 8. (**a**) Energy consumption of the cooling system each month; (**b**) Comparison of the average PUE values over four months (The yellow and green bars present the energy consumption under the PID control strategy and under the developed strategy in this study, respectively).

4.4. Dynamic Response of CRAC Operation

To understand the mechanisms underlying energy conservation, we analyzed the dynamic changes in fan speed and chilled-water valve opening of the CRAC unit under both strategies over a 24-h cycle.

As shown in Figure 9, under the PID control, the CRAC fan speed (orange dotted line) fluctuated minimally and remained at a high level (approximately 80%) for most of the time to ensure that the return air temperature did not exceed the set point. In contrast, the developed control strategy exhibited a strong correlation between fan speed (blue solid line) and IT load (gray area). During low-load nighttime periods, fan speed was reduced to less than 40%. Before daytime peak loads, the system developed proactively and smoothly increased fan speed to accommodate the anticipated heat, thereby achieving more efficient operation.

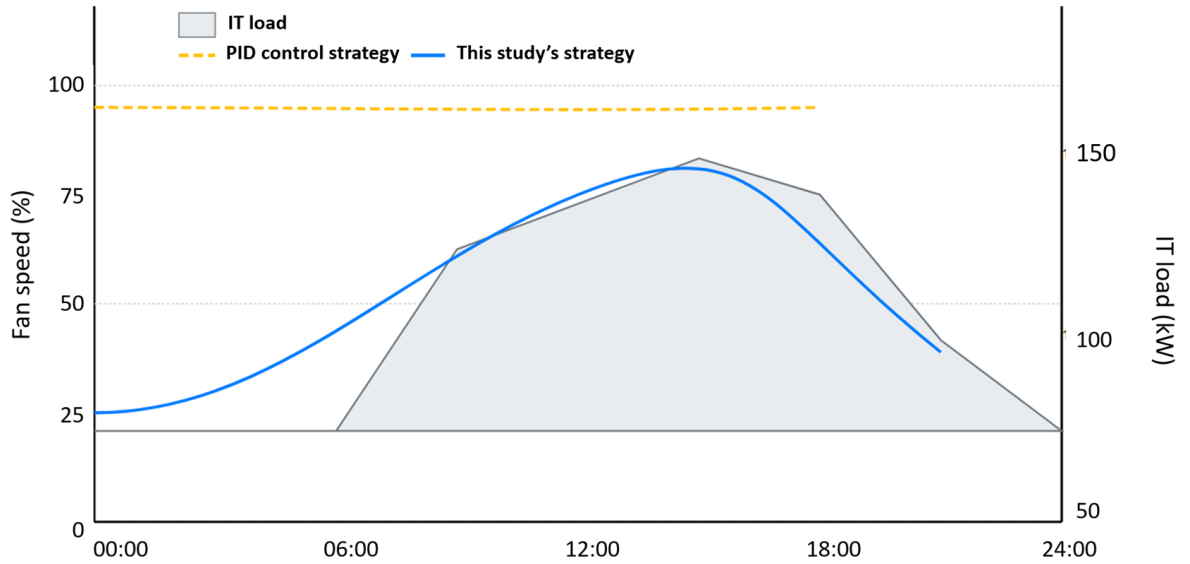


Figure 9. Dynamic response comparison of IT load and the average rotational speed of CRAC fans under two strategies within 24 h.

4.5. System Robustness

System robustness was evaluated by simulating the failure of a single CRAC unit. In the third month of experiments, one CRAC unit was set offline for 24 h. Under this single-point failure condition, the PID control system experienced severe overheating, with maximum temperatures reaching 29.5 °C. By contrast, the developed control system leveraged its global regulation capability to intelligently allocate remaining cooling resources, maintaining the maximum temperature at 27.8 °C (Table 17). The recovery time to system stability was shorter under the developed control system, which provides superior stability and reliability under abnormal operating conditions.

Table 17. System performance of CRAC units in a single-point failure condition.

Performance Metric	PID Strategy	Developed Strategy in This Study
Maximum hotspot temperature during failure (°C)	29.5 °C	27.8 °C
Duration of hotspot during failure (hours)	6.2 h	0.5 h
Average load rate of remaining CRAC units (%)	95%	88%
System stabilization time after recovery (minutes)	45 min	20 min

A sensor sparsity analysis was performed to determine the minimum deployment density required for high-fidelity Digital Twin operation. The developed system reduced the input feature set from 120 to 30 sensors. R^2 remained above 0.95 until the density dropped below 40 sensors (2 sensors per rack). Below the threshold, the model's ability to reconstruct the 3D thermal field degraded significantly (RMSE increased by more than 30%), identifying the two-sensor-per-rack configuration as the threshold for maintaining control benefits.

4.6. Ablation and Sensitivity Analysis

To verify the robustness of the proposed system, a multi-factor sensitivity analysis was performed. The impact of its design on prediction accuracy (RMSE) is summarized in Table 18.

Table 18. Sensitivity analysis and ablation study of key design parameters.

Parameter	Variation	RMSE (°C)	R^2
Z-score threshold	2.5/3.0 (base)/3.5	0.26/0.25/0.25	0.985
Mapping method	IDW/Kriging (base)	0.27/0.25	0.981
Horizon (L_{out})	30 m (base)/60 m	0.25/0.31	0.985/0.942
Model capacity	Small/medium (base)/large	0.34/0.25/0.24	0.965/0.985/0.986

Varying the Z-score threshold from 2.5 to 3.5 resulted in a negligible RMSE fluctuation (<0.02 °C), indicating that the model is not overly sensitive to the specific parameters of initial data cleaning. A comparison between IDW and Kriging interpolation revealed that Kriging achieved a 4.2% lower mean reconstruction error due to its ability to account for spatial correlation. However, downstream LSTM accuracy remained consistent, suggesting the Digital Twin pipeline is robust to the choice of interpolation method. Extending the forecast horizon (L_{out}) from 30 to 60 min increased RMSE by 24%. However, the accuracy remained within the acceptable safety threshold for cooling adjustments. Testing results of smaller (1 layer, 64 units) and larger (3 layers, 256 units) architectures showed that the selected medium configuration (2 layers, 128 units) captured the necessary non-linear dynamics without the overfitting risks or high computational costs of deeper models.

4.7. Economic and Practical Considerations

The developed system involves three costs: (1) hardware costs for industrial-grade sensor arrays and cabling, (2) data infrastructure costs for acquisition gateways and time-series databases, and (3) computational resources for LSTM training and real-time inference. While high-density deployment increases initial CAPEX, the resulting 12% improvement in PUE offers substantial savings. In a medium-scale data center (e.g., 100 kW load), the reduction in cooling energy costs enables a projected return on investment (ROI) within 12 to 36 months. As sensor technology costs continue to decline and cloud-based computational resources become accessible, the predictive system becomes more economically viable for larger, multi-megawatt facilities due to economies of scale.

To confirm the validity of the improvement claims, we conducted a statistical significance analysis on the daily results collected over the 18-day test set. A paired-sample t-test was employed to compare the daily PUE and energy consumption of the proactive strategy against the PID baseline. The analysis yielded a p-value of less than 0.01 for both metrics, indicating that the observed 37.5% energy saving was statistically significant. Cohen's d was 1.85, representing a large effect size. This confirmed the benefits of the proactive strategy and its robustness, which proved to outperform traditional reactive methods with various IT loads.

4.8. Robustness and Sensitivity Check

Following the efficiency analysis principles of Sadri et al. [28] and Babaelahi et al. [29], we performed a sensitivity check on the reported 37.5% energy savings. By varying the effective heat-transfer coefficients (h) and thermal resistances (Rth) by $\pm$20%, the energy saving margin fluctuated between 34.2 and 39.1%. This confirms that even with realistic variability in thermal pathways and convective/radiative contributions, the proactive strategy remains significantly more efficient than the baseline PID.

5. Conclusions

We developed and implemented a data center cooling system based on distributed sensor arrays and LSTM predictive analysis. Through a four-month operation and comparative experiment, the following results were obtained. To address the diversity of data center environments, transfer learning and the Digital Twin model were integrated to adapt to different facilities with minimal data requirements. DRL was also adopted for autonomous optimization for long-term changes such as seasonal climatic variations and mechanical degradation. These features, combined with the system's inherent scalability, demonstrate that the developed system methodology has large application potential for thermal management across various industrial scales.

The system employs a high-precision LSTM model for temperature prediction, achieving an RMSE of only 0.25 °C and an R^2 of 0.985. This predictive capability enables the system to anticipate future heat load variations and adjust cooling strategies proactively, thereby overcoming the response lag inherent in traditional reactive control. Compared with conventional PID control, the proposed framework reduces cooling system energy consumption by approximately 37.5% and lowers the overall PUE of the data center from 1.48 to 1.30, corresponding to a reduction of more than 12%. Hotspot occurrence is effectively suppressed, with cumulative hotspot duration (temperature > 27 °C) reduced from 48 to 2 h. Furthermore, all over-temperature events above 28 °C were eliminated, significantly enhancing IT equipment safety.

The system demonstrates precise dynamic regulation by closely matching the cooling supply with immediate IT load changes. It also exhibits stability and adaptability under abnormal operating conditions, even in equipment failure. In the simulated failure of a

single CRAC unit, the maximum hotspot temperature decreased from 29.5 °C (PID control) to 27.8 °C, while system stabilization time was shortened from 45 to 20 min.

In this study, we integrated transfer learning into the Digital Twin model to enable the system to adapt to different data centers with minimal retraining. DRL was introduced to enable the controller to autonomously discover efficient energy-saving strategies that account for the aging of mechanical cooling components and varying seasonal ambient conditions [22]. Beyond data center cooling, the predictive system has potential for alternative applications such as precision agriculture to manage greenhouse micro-climates or for medical storage facilities where strict temperature and humidity control is mandatory. Additionally, its ability to process enormous sensor data in real-time can be used for structural health monitoring in smart cities, where early detection of environmental stressors is critical [30]. However, it is necessary to integrate computational fluid dynamics simulations into digital twin models for more detailed airflow analysis, as well as the application of algorithms to autonomously learn optimal cooling control strategies. In addition, the concepts and methodology of the developed system demonstrate scalability and hold broad application potential for thermal management in other large-scale industrial facilities. While the developed system demonstrates significant efficiency gains, the results are obtained for the specific layout and high-density sensor configuration. Further validation with different airflow containment strategies (e.g., cold-aisle vs. hot-aisle containment) is necessary to generalize the prediction capability of the system. Deployment in new environments is also needed by using transfer learning and sparsity thresholds.

Author Contributions: Conceptualization, J.L. and R.F.; methodology, J.L.; software, J.L.; validation, J.L., R.F. and Z.L.; formal analysis, R.F.; investigation, J.L. and Z.L.; resources, N.H.; data curation, Z.L.; writing—original draft preparation, J.L.; writing—review and editing, J.L. and J.Q.; visualization, R.F.; supervision, N.H. and J.Q.; project administration, N.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets generated and/or analyzed during the current study are not publicly available due to institutional restrictions, but are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CFD	Computational Fluid Dynamics
CRAH	Computer Room Air Handler
DRL	Deep Reinforcement Learning
IIoT	Industrial Internet of Things
IoT	Internet of Things
LAN	Local Area Network
MAN	Metropolitan Area Network
MPC	Model Predictive Control
PUE	Power Usage Effectiveness
RH	Relative Humidity
RTDO	Real-Time Dynamic Optimization
TL	Transfer Learning
WAN	Wide Area Network
WSN	Wireless Sensor Network

References

1. Luo, H.; Li, Y.; Wang, X. A Brief Analysis of Temperature and Temperature Difference in Data Center Cooling Systems. *Refrig. Air Cond.* **2021**, *21*, 83–88. Available online: https://www.xueshuwenhai.com/zlykt (accessed on 23 December 2025).
2. Wang, J.; Huang, Z.; Liu, Z.; Yue, C.; Wang, P.; Yoon, S. In-situ sensor correction method for data center cooling systems using Bayesian Inference coupling with autoencoder. *Sustain. Cities Soc.* **2022**, *76*, 103514. [CrossRef]
3. Zhou, C.; Hu, Y.; Liu, R.; Liu, Y.; Wang, M.; Luo, H.; Tian, Z. Energy Performance Study of a Data Center Combined Cooling System Integrated with Heat Storage and Waste Heat Recovery System. *Building* **2025**, *15*, 326. [CrossRef]
4. Tian, Y.; Wang, J.; Qi, Z.; Yue, C.; Wang, P.; Yoon, S. Calibration method for sensor drifting bias in data center cooling system using Bayesian Inference coupling with Autoencoder. *J. Build. Eng.* **2023**, *67*, 105961. [CrossRef]
5. Wang, B.; Guo, Y.; Shao, S.; Tu, Q.; Li, Q. A Review of Energy Efficiency Evaluation Indicators Related to Data Center Cooling Systems. *Acta Refrigerol. Sin.* **2023**, *44*, 18–27. [CrossRef]
6. Nadjahi, C.; Louahlia, H.; Lemasson, S. A review of thermal management and innovative cooling strategies for data center. *Sustain. Comput. Inform. Syst.* **2018**, *19*, 14–28. [CrossRef]
7. Athavale, J.; Yoda, M.; Joshi, Y. Comparison of data driven modeling approaches for temperature prediction in data centers. *Int. J. Heat Mass Transf.* **2019**, *135*, 1039–1052. [CrossRef]
8. Li, Y.; Wen, Y.; Tao, D.; Guan, K. Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning. *IEEE Trans. Cybern.* **2020**, *50*, 2002–2013. [CrossRef]
9. Wang, J.; Tian, Y.; Qi, Z.; Zeng, L.; Wang, P.; Yoon, S. Sensor fault diagnosis and correction for data center cooling system using hybrid multi-label random Forest and Bayesian Inference. *Build. Environ.* **2024**, *249*, 111124. [CrossRef]
10. Zhang, Y.; Wei, Z.; Zhang, M. Free cooling technologies for data centers: Energy saving mechanism and applications. *Energy Procedia* **2017**, *143*, 410–415. [CrossRef]
11. Zhang, M.; An, Q.; Long, Z.; Pan, W.; Zhang, H.; Cheng, X. Optimization of Airflow Organization for a Small-scale Date Center Based on the Cold Aisle Closure. *Procedia Eng.* **2017**, *205*, 1893–1900. [CrossRef]
12. Du, S.; Zhang, Q.; Zou, S.; Meng, F.; Liu, L. Simulation analysis on energy consumption and economy of CPU cooling system based on loop heat pipe for data center. *Therm. Sci. Eng. Prog.* **2023**, *45*, 102115. [CrossRef]
13. Sulaiman, M.A.; Adham, A.M.; Hasan, H.F.; Benim, A.C.; Anjal, H.A. Performance analysis of novel dew point evaporative cooler with shell and tube design through different air-water flow configurations. *Energy* **2024**, *289*, 129922. [CrossRef]
14. Chang, Q.; Huang, Y.; Liu, K.; Xu, X.; Zhao, Y.; Pan, S. Optimization Control Strategies and Evaluation Metrics of Cooling Systems in Data Centers: A Review. *Sustainability* **2024**, *26*, 7222. [CrossRef]
15. Oh, H.; Jin, W.; Peng, P.; Winick, J.A.; Sickinger, D.; Sartor, D.; Zhang, Y.; Beckers, K.; Kitz, K.; Acero-Allard, D.; et al. Techno-economic performance of reservoir thermal energy storage for data center cooling System. *Appl. Energy* **2025**, *391*, 125858. [CrossRef]
16. Kula, A.; Dąbrowski, D.; Blachnik, M.; Sajkowski, M.; Smalcerz, A.; Kaminski, Z. Modelling the Temperature of a Data Centre Cooling System Using Machine Learning the Methods. *Energies* **2025**, *18*, 2581. [CrossRef]
17. Qu, S.; Wang, C.; Liu, Y.; Feng, Y.; Han, Y.; Xing, Z. System design, control strategy development and performance comparison analysis for a novel air-liquid hybrid data center cooling system based on cold plates. *Int. J. Refrig.* **2025**, *180*, 222–239. [CrossRef]
18. Huang, J.; Zou, S.; Li, Y.; Xiao, T.; Zhang, J. Performance and maintenance strategy of a data center cooling system under equipment Aging. *Appl. Therm. Eng.* **2025**, *281*, 128623. [CrossRef]
19. Babay, M.-A.; Adar, M.; Chebak, A.; Mabrouki, M. Forecasting green hydrogen production: An assessment of renewable energy systems using deep learning and statistical methods. *Fuel* **2025**, *381*, 133496. [CrossRef]
20. Lee, E.K.; Kulkarni, I.; Pompili, D.; Parashar, M. Proactive thermal management in green datacenters. *J. Supercomput.* **2010**, *60*, 165–195. [CrossRef]
21. Lin, J.; Lin, W.; Lin, W.; Wang, J.; Jiang, H. Thermal prediction for Air-cooled data center using data Driven-based model. *Appl. Therm. Eng.* **2022**, *217*, 119207. [CrossRef]
22. Kabir, M.M.; Begum, S.; Barua, S.; Ahmed, M.U. Taxonomy, challenges, and future directions for AI-driven industrial cooling systems. *Array* **2025**, *27*, 100448. [CrossRef]
23. Bergmeir, C.; Benites, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]
24. Sarkhosh, M.H.; Edrisnia, H.; Raveshi, M.R.; Sharbatdar, M. Prediction of time averaged wall shear stress distribution in coronary arteries' bifurcation varying in morphological features via deep learning. *Front. Physiol.* **2025**, *16*, 1518732. [CrossRef]
25. tempCube. Available online: https://tempcube.io/blogs/temperature-humidity/proactive-data-management-preventing-downtime-with-temperature-and-humidity-sensors (accessed on 25 December 2025).
26. Parolini, L.; Sinopoli, B.; Krogh, B.H.; Wang, Z. A Cyber–Physical Systems Approach to Data Center Modeling and Control for Energy Efficiency. *Proc. IEEE* **2012**, *100*, 254–268. [CrossRef]
27. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]

28. Sadri, S.; Raveshi, M.R.; Amiri, S. Efficiency analysis of straight fin with variable heat transfer coefficient and thermal conductivity. *J. Mech. Sci. Technol.* **2012**, *26*, 1283–1290. [CrossRef]
29. Babaelahi, M.; Raveshi, M.R. Analytical efficiency analysis of aerospace radiating fin. *Proc. IMechE Part C J. Mech. Eng. Sci.* **2014**, *228*, 3133–3140. [CrossRef]
30. Liu, W.; Song, J.; Ji, J.; Yan, R.; Yin, Z.; Li, Y.; Zhang, D. A machine learning based predictive control framework for energy efficient fog prevention and humidity management in indoor ice rinks. *Energy Build.* 2025, *in press*. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.