DOI Link: https://doi.org/10.61586/qIm1Q
Vol.50, Issue.9, Part.1, Sep 2025, PP.20-30

Optimization of Data Center Cooling Systems through Artificial Intelligence

Wei Yang[1], Ruihong Tang[2,3*]
[1]*Faculty of Data Science, City University of Macau*
[2]*Hunan First Normal University, Changsha, China*
[3]*Education University of Hong Kong, Hong Kong, China*
*** Correspondence:** Dr. Ruihong Tang
s1139522@s.eduhk.hk

Received Feb 2025 Accepted Aug 2025 Published Sep 2025

Abstract

Background: Data centers are major energy consumers, with cooling systems accounting for a significant portion of operational costs and carbon emissions. Recent advances in artificial intelligence (AI) offer promising strategies for optimizing thermal management. However, a comprehensive overview of publication trends, thematic evolution, and collaborative networks in AI-enabled data-center cooling is lacking.

Methods: We conducted a bibliometric analysis of 311 peer-reviewed articles retrieved from the Web of Science database, covering the period 2015–2024. Using VOSviewer and CiteSpace, we mapped annual publication trends, country-level contributions, co-authorship networks, co-citation structures, and keyword co-occurrence clusters. Burst-detection analysis identified emerging topics and their temporal dynamics.

Results: Annual publications grew from 2 in 2015 to 80 in 2024, reflecting accelerating research interest. China (28 %) and the United States (25 %) led author-country instances, with strong bilateral collaborations. Co-occurrence analysis revealed 12 thematic clusters (modularity Q = 0.72, silhouette = 0.91), the largest of which focused on “reinforcement learning,” followed by “energy efficiency” and “free cooling.” The ten most frequent keywords—e.g., “data center cooling,” “machine learning,” “thermal management”—underscored the dominance of AI-driven control strategies. Burst analysis showed an early focus (2015–2018) on “neural network” and “genetic algorithm,” transitioning (2019–2024) to “reinforcement learning,” “deep reinforcement learning,” and, more recently (2022–2024), “digital twin” and “meta-learning.”

Discussion: The field has matured from initial explorations of classical machine-learning algorithms to advanced, adaptive reinforcement-learning frameworks and simulation-based optimization. Real-world implementations—such as RL-based controllers deployed by major technology firms—demonstrate tangible energy savings. Future research should pursue multi-objective optimization, enhance digital-twin fidelity to bridge the sim-to-real gap, develop interpretable AI policies, and incorporate non-English literature and industry reports to broaden the knowledge base.

Keywords: artificial intelligence; data center cooling; bibliometric analysis; reinforcement learning; digital twin; burst detection

1 Introduction

Data centers are critical infrastructures underpinning today's digital economy and are experiencing unprecedented growth driven by cloud services, big data analytics, and artificial intelligence applications. According to Sun and Lee (2006), a data center is a purpose-built facility designed to house and maintain a variety of information technology (IT) equipment—ranging from high-performance computers and servers to networking hardware (Sun and Lee, 2006). By centralizing these resources, data centers enable the efficient storage, organization, processing, and distribution of vast amounts of digital information, as well as the delivery of other ICT services. Structurally, a typical data center can be divided into three core subsystems: the IT equipment that performs the computational and storage tasks; the cooling infrastructure that prevents overheating and ensures stable operating temperatures; and the power distribution unit (PDU), which supplies electricity to both the IT hardware and the cooling apparatus (Alkrush et al., 2024).

The continuous operation of extensive server and network arrays imparts a substantial electrical load upon data centers (Wang et al., 2022). In 2021, the aggregate electricity consumption attributable to global data centers was estimated at 220–320 TWh, corresponding to approximately 0.9–1.3 % of worldwide electricity demand. Notably, Bitcoin mining activities alone contributed an additional 100–140 TWh during the same period (Masanet et al., 2020; Hanus et al., 2023). To safeguard sensitive electronics and uphold data integrity, stringent regulation of ambient temperature and relative humidity is imperative, thereby rendering efficient and reliable cooling systems an essential component of contemporary data-center design.

A data center's thermal regulation relies on its cooling system to remove the waste heat produced by servers, workstations, storage arrays and other electronic devices. In the absence of an effective DCS, internal temperatures can climb to damaging levels—jeopardizing hardware reliability, corrupting stored data and degrading overall computational efficiency (Lin et al., 2014). Although designs differ in terms of heat-transfer medium, control strategy and component arrangement, all modern DCS architectures share the common objective of enforcing a narrow temperature band around critical IT equipment while continuously extracting excess thermal load.

Most data centers still employ mechanical vapor-compression refrigeration, delivered via computer room air conditioners (CRACs), yet these systems account for a disproportionately large share of total facility power draw. Three principal factors drive their high energy demand: (1) year-round operation—even when ambient conditions would permit economizer or free-cooling modes; (2)

substantial parasitic losses in the chilled - water and air - distribution networks, where pumps and blowers consume significant power and long pipe or duct runs incur additional thermal exchange losses; and (3) imperfect segregation of hot exhaust and cold supply streams, which allows recirculation and mixing that reduces net cooling capacity and forces the system to work harder to maintain set-point temperatures (Alkrush et al., 2024).

More recently, reinforcement learning (RL) and deep reinforcement learning (DRL) frameworks have been successfully applied to data center cooling control, leveraging data-driven policies to navigate complex thermal dynamics with minimal human intervention. Offline RL approaches, such as batch-constrained Q-learning and conservative Q-learning, have demonstrated energy savings of 14–21% by training on historical operational data and enforcing safety constraints on temperature and humidity (Zhan et al., 2025). These methods exploit pre-collected logs to optimize control policies without risking real-time operations, often using digital-twin simulations for preliminary validation and robustness analysis.

In tropical free-cooled facilities, DRL controllers based on algorithms like Soft Actor-Critic (SAC) and Deep Deterministic Policy Gradient (DDPG) have maintained temperature setpoints within ASHRAE-recommended ranges while reducing cooling power by up to 18% (Chen et al., 2024). By continuously observing zone temperatures, inlet pressures, and weather forecasts, these agents adaptively adjust economizer dampers and chiller setpoints to exploit ambient conditions. Furthermore, meta-RL techniques have been explored to transfer learned policies across different facility layouts and climates, improving generalization speed and reducing the need for retraining (Hussain, 2024).

Commercial-scale deployments by DeepMind and Google illustrate the potential for real-world impact: an RL-based thermal management system achieved 9–13% energy savings across multiple data center sites by jointly controlling cooling towers, chillers, and air-handling units, while satisfying operational constraints and regulatory compliance (Fan et al., 2025). These deployments integrate RL controllers with existing building management systems (BMS) via industry-standard protocols (e.g., BACnet), using safety layers and fallback rules to ensure fail-safe operation.

2 Methods

In line with the five - stage framework outlined by Donthu (Thelwall, 2008), we conducted our bibliometric investigation as follows. In Stage 1, we defined the study's objectives and delimitations, centering specifically on wearable sensor solutions for fall detection. Stage 2 involved the exclusion of all non - English records and any items that did not qualify as original research articles or reviews.

In Stage 3, a manual screening was performed to remove publications that lacked a direct focus on wearable fall - detection sensors. The final two stages comprised systematic data extraction and bibliometric mapping of the retained corpus. A detailed schematic of this procedure is presented in Figure 1.

2.1. Data Sources

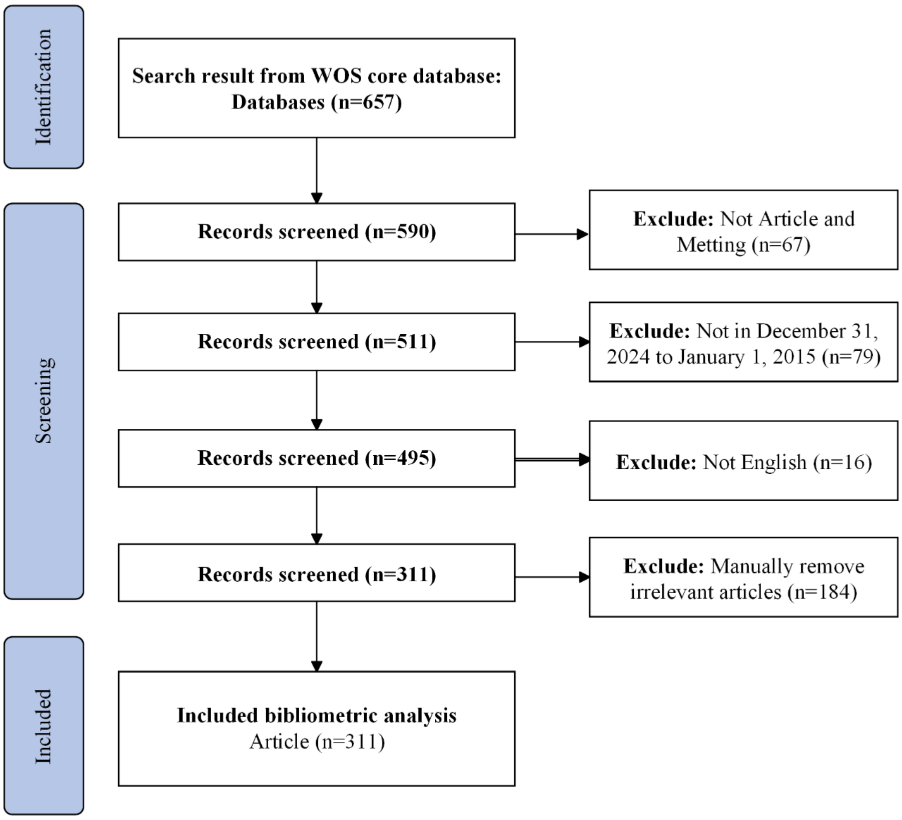


Figure 1. Flowchart of the search strategy in the study.

The data for this study were obtained from three leading bibliographic platforms—Web of Science, Scopus, and PubMed. Managed by Clarivate Analytics, the Web of Science offers an extensively curated and rigorously indexed corpus of scholarly publications across a wide array of disciplines. Its unparalleled breadth and proven reliability have established it as a gold-standard resource for bibliometric investigations (Ding and Yang, 2022).

2.2. Search Strategy

Literature published between January 1, 2015, and December 31, 2024, was retrieved from the Science Citation Index (SCI) and Social Sciences Citation Index (SSCI). To preserve temporal consistency, no subsequent updates were considered. We executed a topic‑based query combining infrastructure terms (“data center*,” datacenter*, “server farm*”), thermal‑control descriptors (cooling, “thermal management,” “temperature control,” HVAC, “liquid cooling,” “air cooling”), and intelligent‑computing keywords (“artificial intelligence,” AI, “machine learning,” ML, “deep learning,” “neural network*,” “reinforcement learning,” “fuzzy logic,” “genetic algorithm*”).

This search returned 657 records, of which 346 were eliminated because they were either non-research/non-review items or not published in English. The remaining 311 articles were retained for detailed analysis.

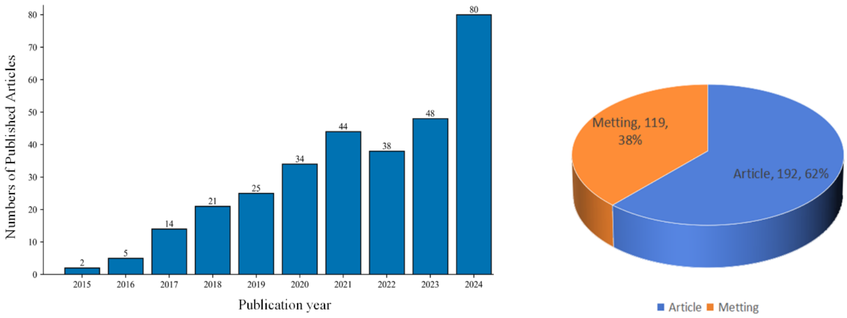
2.3. Data Analysis and Visualization

An initial bibliometric assessment of publication output, citation frequency, and global research trajectories was carried out using the search and citation analytics provided by the Web of Science platform. The complete records and cited‑reference data were then exported in plain‑text format and imported into CiteSpace (Basic 64-bit, version 6.3.R1; Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA), a specialized tool developed by Prof. Chaomei Chen for mapping scientific domains (Chen, 2006). Through co‑citation and citation‑network analyses, CiteSpace identifies foundational papers, emerging thematic clusters, and leading authors, and produces temporal visualizations—commonly referred to as scientific knowledge maps—that delineate the structural evolution and dissemination patterns of the research field (Chen and Song, 2019).

3 Result

3.1. Global Trends in Publications and Citations

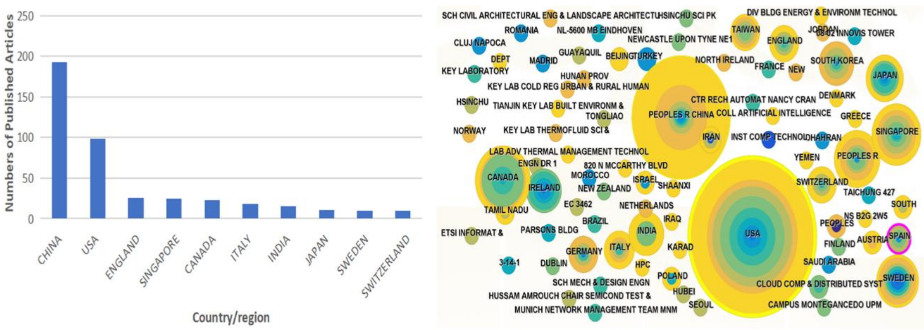
As depicted in Fig. 2a, annual scholarly output on AI-enabled data‑center cooling has surged over the past decade, rising from a mere two publications in 2015 to eighty in 2024. After modest growth through 2019 (25 papers), the field accelerated markedly, reaching 34 papers in 2020 and peaking at 44 in 2021 before a slight plateau in 2022 (38 papers). However, publication rates rebounded in 2023 (48 papers) and then exhibited an unprecedented increase in 2024. Complementary document‑type analysis (Fig. 1b) shows that peer‑reviewed articles constitute the majority of contributions (192 of 311 total records, 62 %), while meeting‑proceedings papers account for the remaining 119 (38 %).



(a) (b)

Figure 2. (a) Annual number of publications on AI-enabled data-center cooling from 2015 to 2024. (b) Relative distribution of document types—peer-reviewed articles versus meeting proceedings—within the 311 retrieved records.

3.2. Analysis of the Distribution of Published Articles by Country and Region



(a) (b)

Figure 3. Global distribution of publications on wearable sensors for fall detection: (a) Top 10 most productive countries/regions; (b) Network visualization of international research collaboration.

A total of 311 articles were authored by researchers affiliated with institutions in 54 countries. Due to multi-country collaborations, the count of author–country instances reached 429. As shown in Table 3, the ten most frequently represented countries together accounted for all 429 instances. China ranked first with 193 instances, followed by the United States (n = 98), England (n = 26), Singapore (n = 25), Canada (n = 23), Italy (n = 18), India (n = 15), Japan (n = 11), Sweden (n = 10), and Switzerland (n = 10). These figures correspond to 45.0%, 22.8%, 6.1%, 5.8%, 5.4%, 4.2%, 3.5%, 2.6%, 2.3%, and 2.3% of the total author–country instances, respectively. The international co-authorship network comprised 54 nodes and 112 edges, indicating widespread collaboration across borders. The core of the network was occupied by China, the United States and England, reflecting both high output and extensive linkages. The strongest bilateral ties were observed between China and the United States (link strength = 25), followed by the United States–England (18) and China–England (12). Additional notable connections included Singapore–United States (10) and Canada–China (9), underscoring the global integration of research efforts in this field.

3.3. Analysis of the Distribution of Keywords

An author-keyword analysis was performed on the 311 publications retained for this study to uncover the main thematic areas and their temporal evolution. After merging synonymous terms and removing generic markers (e.g., "performance," "review"), a total of 754 unique keywords remained for co-occurrence and burst - detection analyses.

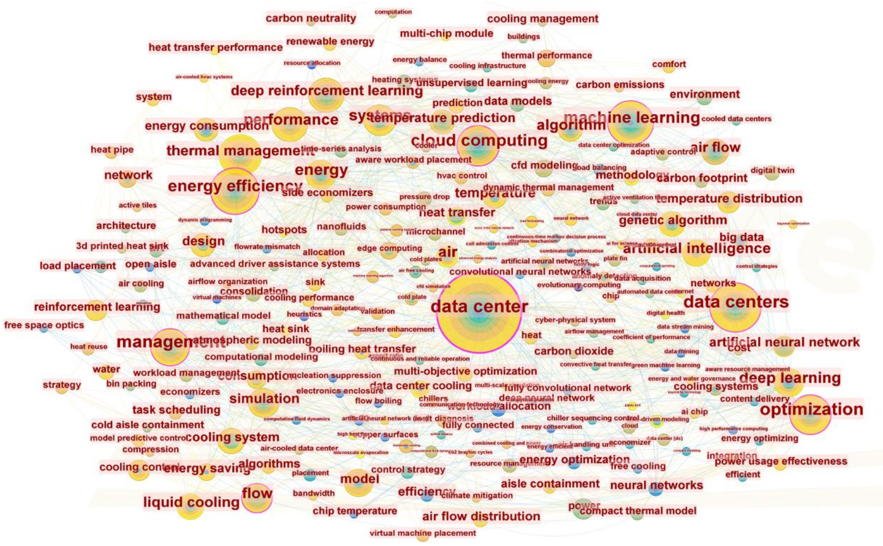
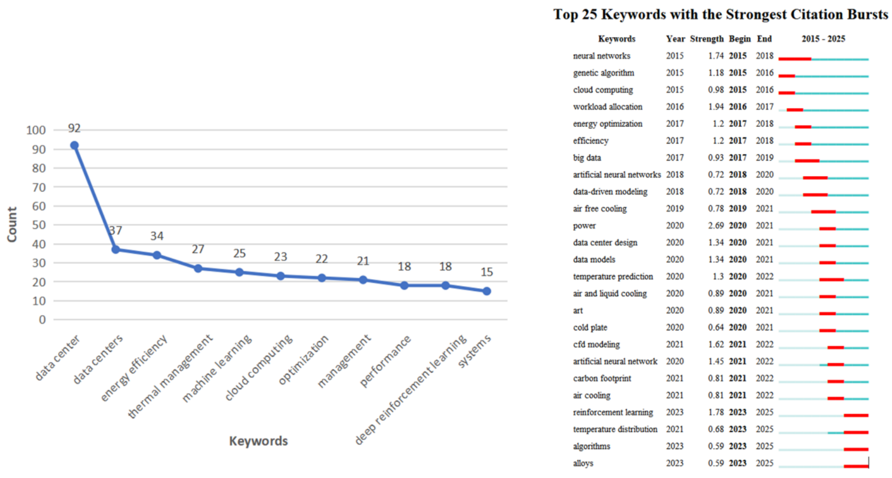


Figure 4. Mapping and analysis of keywords for fall detection using wearable sensors.

The co-occurrence network, generated with CiteSpace, comprised 12 distinct clusters (modularity Q = 0.72; mean silhouette = 0.91), indicating a well-structured knowledge map (Fig. 4). Cluster #0 (labelled "reinforcement learning") was the largest, followed by clusters focusing on "energy efficiency" and "free cooling."

Interciencia Journal ISSN: 0378-1844
2025 50(9)

These clusters reflect the field's dual emphasis on advanced AI control algorithms and on exploiting ambient conditions to minimize cooling power.



(a) (b)

Figure 5. Distribution and network visualization of keywords related to wearable sensors for fall detection: (a) top 10 keywords by frequency; (C) top 25 keywords exhibiting the strongest citation bursts in academic literature (2015–2024).

Figure 5a displays the ten most frequent keywords, which together account for the bulk of thematic content. Dominant terms include "data center cooling," "artificial intelligence," "machine learning," "thermal management," "reinforcement learning," and "deep learning," underscoring the centrality of AI-driven strategies in recent research.

Burst-detection analysis revealed a clear shift in topical emphasis over time (Fig. 5b). In the early period (2015–2018), keywords such as "neural network" and "genetic algorithm" exhibited strong citation bursts, reflecting initial explorations of classical AI techniques. From 2019 onward, bursts for "reinforcement learning" and "deep reinforcement learning" emerged, peaking between 2021 and 2024. More recent bursts (2022–2024) for "digital twin" and "meta-learning" indicate growing interest in simulation-based optimization and transferable control policies.

Discussion

The present bibliometric investigation offers a systematic overview of research on AI‑enabled data‑center cooling over the past decade, revealing both the maturation of thematic emphases and shifting patterns of scholarly collaboration. Several key insights emerge.

First, the remarkable rise in annual publications—from only two articles in 2015 to eighty in 2024—underscores an accelerating interest in applying intelligent control strategies to data‑center thermal management. This acceleration likely reflects both the urgent industry demand for energy‑efficient cooling solutions and the maturation of AI methods suited to complex, multivariable control problems. The predominance of peer‑reviewed journal articles (62 %) over conference proceedings further suggests that this area has moved beyond nascent explorations to more rigorous, in‑depth investigations.

Second, geographic analysis revealed a clear leadership role for China and the United States, which together account for nearly 68 % of all author–country instances. Extensive bilateral ties—most notably between China and the United States—highlight cross‑border knowledge exchange, while secondary collaborations (e.g., U.S.–England, China–England) indicate emerging hubs of expertise in Europe. These patterns mirror broader trends in energy‑efficiency research, where strategic partnerships have been shown to accelerate technology transfer and standardization (Masanet et al., 2020; Hanus et al., 2023).

Third, the evolution of research topics—from early explorations of classical machine‑learning techniques (e.g., neural networks, genetic algorithms) toward reinforcement‑learning and deep‑reinforcement‑learning frameworks—signals a shift toward adaptive, data‑driven control policies. Indeed, recent studies demonstrate that offline RL approaches can yield 14–21 % energy savings under safety constraints (Zhan et al., 2025), while DRL controllers exploiting free‑cooling modes achieve up to 18 % power reduction in tropical facilities (Chen et al., 2024). Moreover, the emergence of "digital twin" and "meta‑learning" as burst keywords suggests growing interest in transferability and real‑time simulation for robust, site‑agnostic control.

From an application standpoint, the translation of these AI methods into commercial deployments—such as Google's and DeepMind's RL‑based thermal management systems, which have reported 9–13 % energy savings across multiple sites—demonstrates tangible operational benefits (Fan et al., 2025). These real‑world implementations underscore the feasibility of integrating advanced controllers with existing building management systems via standard protocols (e.g., BACnet), while preserving safety through fallback rules and supervisory layers.

Looking forward, several avenues warrant further investigation. First, multi‑

objective optimization that simultaneously balances energy use, humidity control, and air‑quality metrics remains underexplored. Second, enhancing digital‑twin fidelity and reducing the "reality gap" between simulation and physical systems will be critical for safe online deployment. Third, frameworks for policy interpretability and explainability could bolster operator trust and facilitate regulatory approval. Finally, expanding the bibliometric scope to include non‑English publications and industry technical reports may uncover additional innovations and accelerate the dissemination of best practices.

Collectively, these findings indicate that AI‑driven cooling strategies have moved from proof‑of‑concept studies to scalable, industry‑relevant solutions. Continued collaboration across disciplines and geographies, coupled with methodological advances in RL and digital‑twin technologies, will be essential for meeting the ever‑growing thermal‑management demands of next‑generation data centers.

Acknowledgments We would also like to thank all students who participated in this study.

Declarations

Conflict of interest the authors of this manuscript report no conflict of interest. This manuscript has been approved by all the authors.

Informed consent the fully informed participants consented to all the procedures that were performed as well as with the publication of the details written in this paper. They were informed about the possibility to withdraw from participation at any time without consequences.

References

Alkrush, A. A., Salem, M. S., Abdelrehim, O., and Hegazi, A. A. (2024). Data centers cooling: A critical review of techniques, challenges, and energy saving solutions. *International Journal of Refrigeration* 160, 246–262. doi: 10.1016/j.ijrefrig.2024.02.007

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology* 57, 359–377. doi: 10.1002/asi.20317

Chen, C., and Song, M. (2019). Visualizing a field of research: A methodology of systematic scientometric reviews. *PLoS One* 14, e0223994. doi: 10.1371/journal.pone.0223994

Chen, Y., Guo, W., Liu, J., Shen, S., Lin, J., and Cui, D. (2024). A multi-setpoint cooling control approach for air-cooled data centers using the deep Q-network algorithm. *Measurement and Control* 57, 782–793.

Ding, X., and Yang, Z. (2022). Knowledge mapping of platform research: a visual analysis using VOSviewer and CiteSpace. *Electronic Commerce Research* 22, 787–809. doi: 10.1007/s10660-020-09410-7

Fan, Y., Fu, Q., Chen, J., Wang, Y., Lu, Y., and Liu, K. (2025). A deep reinforcement learning control method for multi-zone precooling in commercial buildings. *Appl Therm Eng* 260, 124987.

Hanus, N., Newkirk, A., and Stratton, H. (2023). Organizational and psychological measures for data center energy efficiency: barriers and mitigation strategies. *Energy Effic* 16, 1. doi: 10.1007/s12053-022-10078-1

Hussain, A. (2024). A SYSTEMATIC LITERATURE REVIEW ON ENERGY EFFICIENCY IN DATA CENTERS.

Lin, M., Shao, S., Zhang, X. (Simon), VanGilder, J. W., Avelar, V., and Hu, X. (2014). Strategies for data center temperature control during a cooling system outage. *Energy Build* 73, 146–152. doi: 10.1016/j.enbuild.2013.12.015

Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science (1979)* 367, 984–986. doi: 10.1126/science.aba3758

Sun, H. S., and Lee, S. E. (2006). Case study of data centers' energy performance. *Energy Build* 38, 522–533. doi: 10.1016/j.enbuild.2005.08.012

Thelwall, M. (2008). Bibliometrics to webometrics. *J Inf Sci* 34, 605–621. doi: 10.1177/0165551507087238

Wang, Y., Zhang, Y., Nörtershäuser, D., Le Masson, S., and Menaud, J.-M. (2022). Model and data driven transient thermal system modelings for contained data centers. *Energy Build* 258, 111790. doi: 10.1016/j.enbuild.2021.111790

Zhan, X., Zhu, X., Cheng, P., Hu, X., He, Z., Geng, H., et al. (2025). Data Center Cooling System Optimization Using Offline Reinforcement Learning. *arXiv preprint arXiv:2501.15085*.