

Review

Critical Review of Artificial Intelligence and Machine Learning Methods for Data-Center Load Forecasting

Shivanshu Shekhar ¹, Shivanshu Tripathi ², Gajendra Singh Chawda ^{3,*}, Shivam Chaturvedi ⁴, Wencong Su ⁵, Mengqi Wang ⁵, Mandoye Ndoye ³ and Ben Oni ³

¹ Department of Electrical Engineering, Colorado School of Mines, Golden, CO 80401, USA; shivanshu_shekhar@mines.edu

² Department of Electrical & Computer Engineering, University of California, Riverside, CA 92521, USA; shivanshu.tripathi@email.ucr.edu

³ Department of Electrical & Computer Engineering, Tuskegee University, Tuskegee, AL 36088, USA; mndoye@tuskegee.edu (M.N.); boni@tuskegee.edu (B.O.)

⁴ Department of Electrical & Computer Engineering, George Mason University, Fairfax, VA 22030, USA; schatur2@gmu.edu

⁵ Department of Electrical & Computer Engineering, University of Michigan, Dearborn, MI 48128, USA; wencong@umich.edu (W.S.); mengqi@umich.edu (M.W.)

* Correspondence: gchawda@tuskegee.edu

Abstract

Modern data centers are consuming more energy than ever before due to the rapid growth of cloud services, artificial intelligence (AI), and large-scale digital applications. As energy demand continues to rise, accurate load forecasting has become an important tool for improving energy management and operational planning. However, predicting data-center power demand remains challenging because computing workloads, cooling systems, and facility operations are closely connected and constantly changing. This review examines the use of artificial intelligence (AI) and machine learning (ML) techniques for data-center load forecasting. The surveyed literature covers traditional statistical methods, supervised learning algorithms, deep learning models, probabilistic forecasting techniques, and physics-informed hybrid approaches. Important topics such as forecasting horizons, feature selection, performance evaluation, and practical deployment challenges are also discussed. The reviewed studies indicate that advanced approaches, including long short-term memory (LSTM), gated recurrent unit (GRU), transformer-based models, and digital twin (DT) frameworks, can improve forecasting performance and support more energy-efficient operation. These methods can also assist carbon-aware and grid-interactive data-center management. Several challenges remain. These include limited data availability, poor generalization, interpretability issues, and real-time implementation constraints. In this paper, the reviewed studies are classified and analyzed to provide a clear reference for researchers working in AI-based data-center energy forecasting.



Academic Editor: Valentina Colla

Received: 20 June 2026

Revised: 29 July 2026

Accepted: 31 July 2026

Published: 2 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: artificial intelligence (AI); machine learning (ML); data centers (DC); load forecasting; data-center load characteristics

1. Introduction

The growth of cloud computing, artificial intelligence (AI), and high-performance computing (HPC) has increased the size and energy use of modern data centers [1]. AI-driven computing is now a major contributor to global electricity demand and carbon emissions [2,3]. The use of large language models, generative AI, and accelerator-based

systems has further increased rack power density and cooling requirements [1]. Because of this, energy efficiency, thermal stability, and carbon reduction have become key concerns in addition to performance and reliability [2]. In this context, load forecasting has become very important for data-center operation. It is used for workload scheduling, power control, thermal management, and capacity planning. Accurate forecasting helps operators make better decisions. It improves efficiency and reduces energy cost and thermal stress. It is essential to note that load forecasting is also important for carbon-aware computing and grid-interactive operation [4]. Industry studies in [5] also highlight its role in reducing emissions in cloud systems. Unlike traditional power systems or buildings, data-center forecasting is more complex. Workloads change quickly due to software behavior, virtualization, and resource sharing. The IT load is also strongly linked with cooling and power systems. These interactions create nonlinear and time-varying behavior at multiple time scales. These range from seconds to weeks. Therefore, forecasting models must capture temporal patterns and system coupling together. To ensure transparency and reproducibility, the literature was identified and selected following the PRISMA guidelines. Each paper is assigned to exactly one category within each panel, so the counts in both panels sum to 132. The search strategy, eligibility criteria, and selection outcome are described in Figure 1.

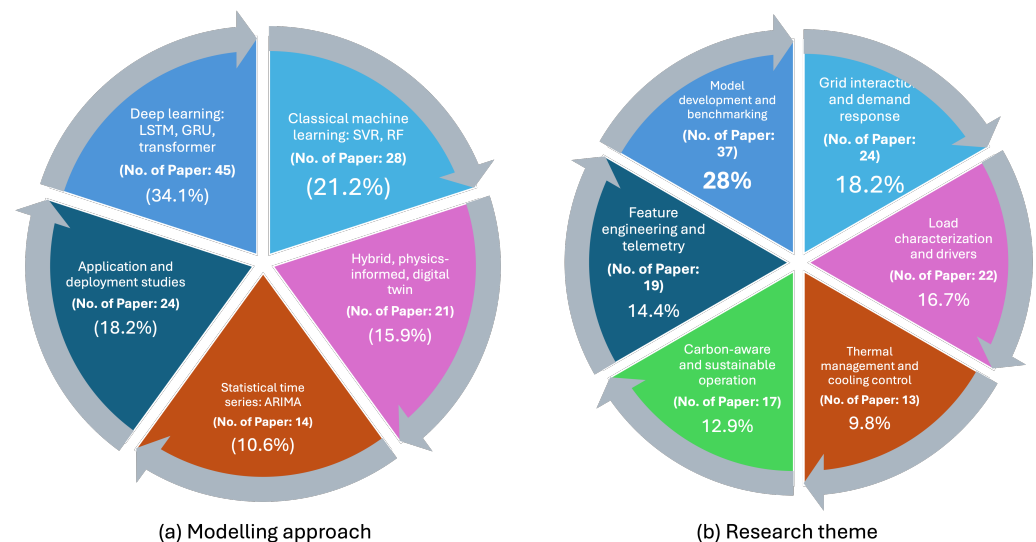


Figure 1. Figure shows PRISMA flow diagram with the composition of the 132 included papers. (a) shows distribution of the included papers by modeling approach. (b) shows distribution by research theme.

To address challenges discussed above, many studies have used AI and ML methods. Early work focused on statistical and regression models. Recent research has focused on deep learning methods such as LSTM, GRU, and transformers [6]. Some studies have also explored reinforcement learning for adaptive decision-making. Physics-informed models are used to include system constraints in learning. Digital twin frameworks are used to model real-time data-center behavior and support prediction. Several studies have focused on autoscaling and elasticity in cloud systems [7]. Other works studied workload prediction and resource estimation in virtualized environments [8]. Energy-aware scheduling and server management have also been widely studied [9]. Additional studies focus on virtual machine consolidation and energy reduction techniques [10]. Further research has explored server-level optimization for energy efficiency [11]. Workload consolidation strategies have also been proposed to reduce power consumption [12]. Thermal management has also received attention due to its impact on cooling cost and system reliability [13]. Recent studies also investigate advanced thermal modeling and optimization methods [14].

Although the studies discussed above provide useful insights, existing reviews are still fragmented. Many focus only on scheduling, auto scaling, or energy optimization. Load forecasting is often treated as a secondary topic. In many works, only CPU utilization or request-level metrics are predicted. Power demand, thermal behavior, and facility-level constraints are less studied. The connection between forecasting accuracy, prediction horizon, and real deployment is also not well explained. Another limitation is the lack of focus on modern AI-based data centers. GPUs, TPUs, and other accelerators create higher and variable power loads. These systems increase coupling between workload, cooling, and power delivery. This makes forecasting more difficult. At the same time, new research areas are emerging. These include explainable AI, physics-informed learning, digital twins, and carbon-aware forecasting [15]. However, these topics are often studied separately. There is limited effort to integrate them into a unified view.

Motivated by gaps discussed above, this paper presents a critical review of AI and ML methods for data-center load forecasting. It connects forecasting models with operational goals such as energy efficiency, thermal management, reliability, and sustainability. The review covers statistical methods, deep learning models, and physics-aware hybrid approaches. It also discusses practical challenges such as data scarcity, nonstationarity, generalization issues, interpretability, and real-time deployment. The reviewed studies presented in this paper are classified and compared based on forecasting targets, methods, and deployment characteristics. The goal is to provide a clear reference for researchers and engineers working on intelligent and sustainable data-center energy systems.

To strengthen the critical-review contribution, the surveyed forecasting methods are evaluated using common criteria, including accuracy, generalization, data requirements, computational complexity, interpretability, robustness, physical consistency, and operational relevance. Reported error metrics are interpreted cautiously because datasets, forecasting horizons, sampling intervals, and validation procedures differ across studies. Statistical models are generally interpretable and computationally efficient but may struggle with nonlinear workloads, whereas deep-learning models provide stronger nonlinear and temporal modeling capability at the cost of greater data and computational requirements. Physics-informed and digital twin approaches improve physical consistency and operational insight but require additional modeling, calibration, and synchronization. Therefore, this review emphasizes application-dependent trade-offs rather than ranking methods solely based on forecasting accuracy.

2. Characteristics of Data Center Loads

Data center loads have different characteristics unlike traditional commercial or industrial loads. These loads are the functions linked with computing activity, cooling systems, and power delivery infrastructure. In traditional buildings, consumption of electricity mainly depends on occupancy and weather conditions, while in data centers, the main driver is digital workload and server utilization presented in [16]. Modern data centers work as cyber-physical systems. Research in [17] discussed power consumption affected directly in proportion to computation activity. This also changes heat generation and cooling demand. Because of this coupling, load behavior becomes highly dynamic and non-stationary. This makes modeling and forecasting more difficult.

Figure 2 shows the structure of a modern data center and its main subsystems. In this figure, the IT layer represents the computing workload. The cooling system removes heat from servers. The power system distributes electricity to all components. In advanced data centers, renewable energy sources and grid signals also interact with system operation. Therefore, total power consumption is not controlled by a single factor. It depends on computing demand, thermal limits, power losses, and energy supply conditions. Because

of this reason, researchers recommended that load forecasting must consider IT workload, cooling behavior, auxiliary systems, and energy supply together.

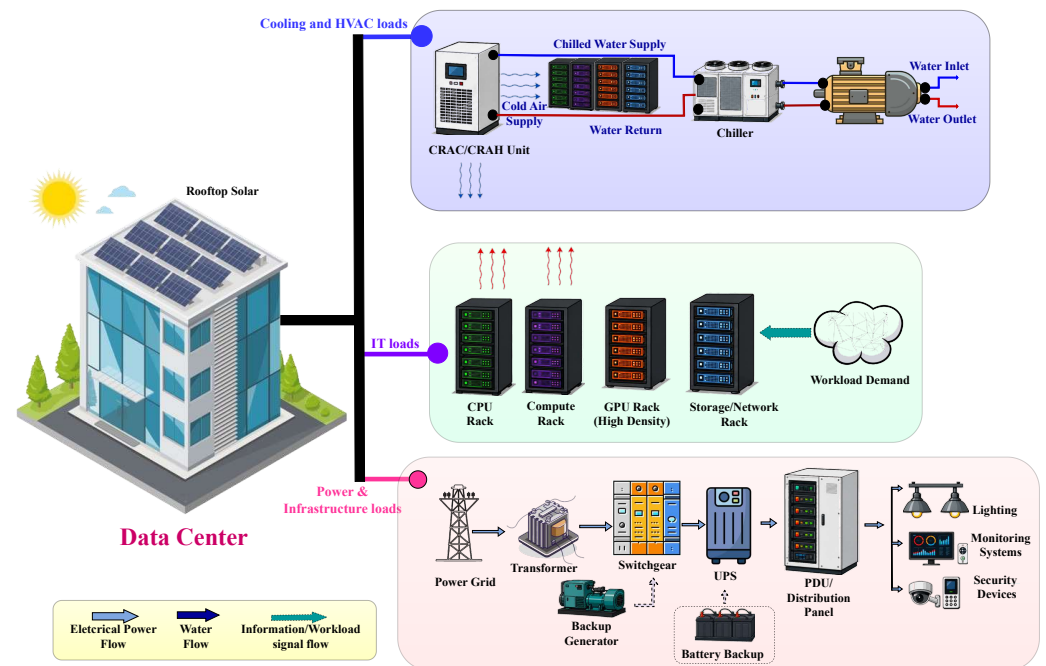


Figure 2. Physical layout of a modern data center showing interactions among IT load, cooling, power systems, renewables, and grid signals.

2.1. Electrical Load Components

The data center electricity use is usually divided into three main parts such as (1) IT equipment load, (2) cooling and HVAC load, and (3) auxiliary infrastructure load. This classification helps explain the electricity demand changes over time and its impact on energy efficiency, as discussed in the following subsections. Figure 3 illustrates the coupling between workload, IT power, thermal response, and scheduling decisions. As workload increases, heat generation increases as consumption of IT power increases, hence, the demand for more cooling, which, in turn, feeds back into operational control and scheduling policies.

(a) IT loads

The majority of energy consumption within modern data centers is attributed to IT equipment (e.g., CPUs, GPUs, accelerators, memory, storage, and networking). The amount of energy used by the IT layer is largely based on the intensity of the workload and the nature of the applications used to perform those workloads. Workloads that are CPU-based often create more consistent and stable power profiles. Conversely, workloads that are processed by GPUs and accelerators (especially for AI training and HPC) can experience a high degree of variability in their power profiles as they process data at multiple instances due to their parallel architecture and aggressive dynamic voltage and frequency scaling (DVFS) mechanisms, which is more common in data centers [18]. Recent AI workloads reported in [19] further amplify this effect due to large-scale model training and high computational density requirements. These workloads often generate sudden power variations and high peak-to-average ratios. As a result, IT load modeling becomes highly nonlinear and challenging, particularly for short-term forecasting.

(b) Cooling and HVAC loads

The cooling and HVAC systems represent the second largest energy consumer in data centers, reported in [20]. Their primary role is to remove heat generated by IT equipment and maintain safe operating temperatures. According to [21], the cooling demand increases with IT load, but it is also strongly influenced by environmental conditions such as temperature and humidity. In addition, system design, airflow management, and control strategies significantly affect cooling efficiency. Prior studies show that cooling energy does not scale linearly with IT load. In many cases, inefficient airflow and suboptimal thermal design can cause faster than linear growth in cooling energy consumption [22].

(c) Auxiliary and infrastructure losses

Auxiliary and infrastructure loads include uninterruptible power supply (UPS) systems, power distribution units (PDUs), transformers, lighting, monitoring equipment, and security systems. These loads are generally more stable than IT and cooling loads. Early studies of warehouse-scale computing facilities highlighted power conversion and distribution equipment, which can contribute a substantial portion of non-IT energy consumption [23]. Reliability-oriented analyses have further shown, in [24], that the UPS systems and related electrical infrastructure introduce additional energy losses that must be considered during facility operation. Auxiliary loads are often viewed simply as “background” consumption; however, they represent a continual base/fixed level of energy demand. This baseline directly influences key efficiency metrics, particularly power usage effectiveness (PUE). For example, [25] examined the influence of different types of infrastructure and cooling configurations on PUE performance and the contribution of support systems to the sustainability of the data center, discussed in [26]. A recent report by [27] also found that non-IT related infrastructure still plays an integral role in overall facility energy consumption at large-scale data centers. Therefore, accurate characterization of auxiliary loads is very important to assess energy efficiency, plan capacity, and long-term operational optimization.

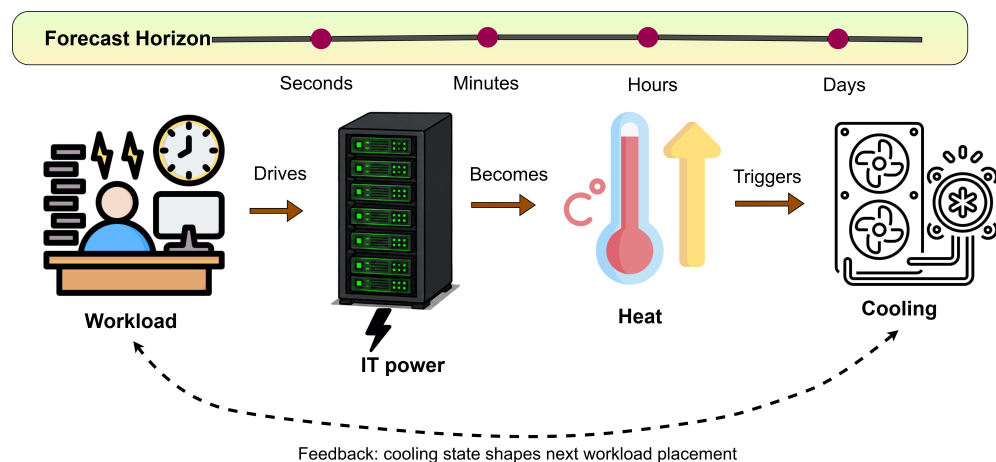


Figure 3. Data-center load dynamics showing workload demand drives IT power, heat generation, cooling response, and feedback to workload scheduling.

2.2. Temporal and Workload-Driven Patterns

Data center workloads show clear temporal patterns. These patterns depend on user demand, application behavior, scheduling policies, and environmental factors [27]. Unlike commercial buildings, which are occupied only on weekends, data centers remain operational 365 days per year, 24 h a day. As a result, energy demand is characterized by workload variation, scheduling decisions, and system usage that fluctuate with these three variables [28,29]. Changes in energy level occur in several ways over different time frames,

ranging from immediate to weeks, and can span years to decades depending on business need. Energy management, renewable energy collection/dispatch, and grid utilization impact (interactions) among today's data centers have a large effect due to the differences in time cycles.

(a) Diurnal, seasonal, and bursty behavior

Data centers typically experience diurnal and seasonal variations in temperature. Hyper-scale data centers often exhibit characteristics that deviate from expected patterns due to their geographic distribution or the presence of multiple workload sites, detailed in [30]. In addition, today's workloads regularly experience bursty behavior. Batch processing, analytics job completion times, and real-time service requests can all create "bursts" in demand that are very short in duration (ranging from seconds to several minutes) and can pose a significant challenge for accurately forecasting future demand, discussed in [28,31].

(b) AI training vs. inference workloads

The addition of AI workloads brings further complexity to temporal modeling. Workloads required for training AI typically have long durations and high computational demands and can therefore produce sustained high power consumption and thermal stress, reported in [32]. In contrast, the workloads required for AI inference are non-continuous and sensitive to latency, leading to highly variable, bursty demand profiles. The coexistence of training and inference workloads within the same infrastructure significantly increases forecasting complexity and resource planning uncertainty, discussed in [33].

2.3. Thermal–Electrical Coupling

The important feature of data center operation is the strong link between electrical power and thermal behavior. Most of the electrical energy used by IT equipment is converted into heat. This heat must be removed to keep systems safe and reliable [34]. Because of this, energy use and thermal management are closely connected. They cannot be studied separately when analyzing data center performance. Recent studies in [20] show that high-density computing and AI workloads have increased this coupling. They also increase cooling demand in modern data centers.

(a) Cooling–power interdependence

The cooling systems are closely linked with IT workloads. They operate in a continuous feedback loop; when computing demand increases, servers use more power. This produces more heat. As a result, cooling demand also increases. However, cooling energy is not only based on heat generation. It also depends on airflow, supply-air temperature, equipment layout, and control strategies, reported in [35,36]. Because of these factors, electrical and thermal systems are strongly connected. It is difficult to model or predict them separately.

(b) Spatial and rack-level effects

Power and temperature are not evenly distributed in a data center. Workload placement and server usage create differences across racks and zones, discussed in [37]. These differences can lead to local hotspots. Extra cooling is needed to keep safe operating conditions. Airflow recirculation and bypass airflow can also increase temperature imbalance in the system. Because of this, average facility measurements do not always show local thermal behavior. This limitation has led to multi-scale, rack-level, and zone-level modeling methods for better forecasting and control, reported in [27].

2.4. System-Level Load Components and Drivers

Data center load is not only determined by IT equipment and cooling systems. Total facility demand also includes overhead from UPS units, PDUs, transformers, and power distribution losses. It also includes auxiliary loads such as lighting, monitoring, security, and control systems. External factors also affect overall operation. These include electricity price, grid carbon intensity, renewable availability, and demand–response signals, widely reported in [38–41]. Because of these factors, system-level indicators are important. These include PUE, overhead ratio, and cooling efficiency [42]. They help in understanding total facility behavior and improving forecasting models. Figure 4 shows forecasting horizons and their operational roles. Short-term forecasts (seconds) support real-time control such as DVFS and power capping. Medium-term forecasts (minutes) support workload placement and thermal pre-cooling. Hour-level forecasts are used for demand response and energy optimization. Long-term forecasts (days to weeks) support capacity planning and infrastructure expansion.

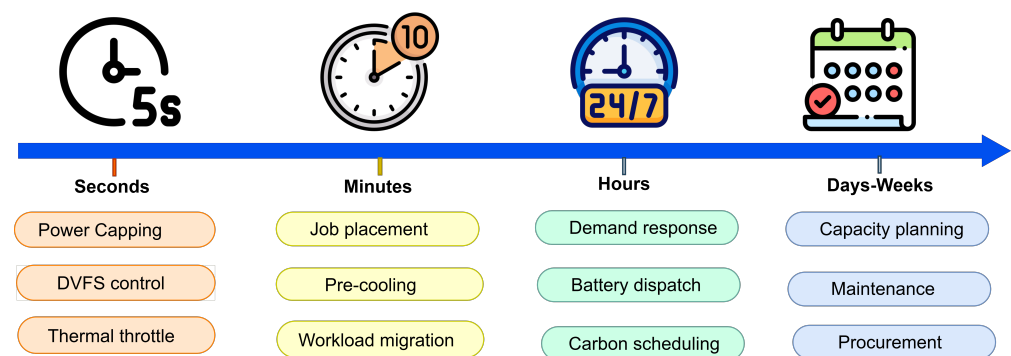


Figure 4. Forecasting horizons ranging from seconds to weeks and their related operational actions such as power control, workload placement, demand response, and capacity planning.

Table 1 summarizes key load components and input features used in AI-based forecasting models. IT features include CPU and GPU utilization, virtual machine (VM) density, and workload scheduling data. Thermal features capture cooling behavior and delayed heat effects. Grid and renewable features support carbon-aware and price-aware forecasting. Each feature group plays a different role in prediction. IT features mainly affect short-term behavior. Thermal features reflect delayed system response. Grid-related features support long-term optimization and sustainability.

Table 1. Summary of data-center load components and key input features for AI-based forecasting.

Load Component	Representative Features	Interpretability	Complexity	Forecasting Relevance	References
IT-related loads	CPU/GPU utilization, memory and I/O activity, VM/container count	High	Low to medium	Dominant driver of short-term and peak demand; critical for workload-aware forecasting.	[17,18,33]
	Job scheduling metadata, task type, batch size	Medium	Medium	Captures abrupt load changes and workload-induced variability.	[29,31]
Cooling and thermal loads	CRAC/CRAH supply and return temperatures, chilled-water temperature	High	Low to medium	Strongly correlated with cooling power and HVAC control regimes.	[20,25,35,36]
	Rack inlet temperature, thermal hotspot indicators	Medium	Medium to high	Improves forecasting under spatial thermal heterogeneity.	[34,35,37]

Table 1. Cont.

Load Component	Representative Features	Interpretability	Complexity	Forecasting Relevance	References
Electrical overheads	UPS input/output power, transformer loading, PDU losses	High	Low	Captures nonlinear electrical losses; essential for facility-level and PUE-aware forecasting.	[23,24]
Auxiliary facility loads	Lighting schedules, pumps, humidification, occupancy indicators	High	Low	Improves robustness and baseline stability during operational regime changes.	[25–27]
Grid and external signals	Electricity price, demand response events, carbon intensity	High	Low to medium	Enables grid-interactive, price-aware, and carbon-aware forecasting.	[38–40]
On-site energy resources	Battery SOC/SOH, PV output, generator dispatch	Medium	Medium	Required for net-load forecasting and closed-loop energy management.	[39,41]
System-level efficiency metrics	PUE, IT power fraction, cooling ratio, climate indicators	High	Low	Supports decomposition-based modeling and cross-site generalization.	[25,26,42]

3. Feature Engineering for Load Forecasting in Data Centers

Feature engineering plays a central role in data-center load forecasting. Unlike traditional systems, data-center electricity demand is not generated by a single physical process. It emerges from the interaction of multiple subsystems, including IT workloads, cooling infrastructure, power delivery systems, and increasingly, grid and carbon-intensity signals. Prior studies show that data-center load behavior is strongly influenced by computational activity and thermal coupling between IT and cooling systems. In addition, modern workloads introduce fast temporal variations driven by scheduling decisions and control actions. These characteristics make data center load fundamentally different from conventional building demand profiles.

Because of this complexity, models based only on historical power consumption or time-of-day patterns often fail under changing operating conditions. This limitation becomes more pronounced across different climates and workload regimes [43]. Classical forecasting frameworks also struggle to generalize under strong non-stationarity in system behavior [44]. Empirical results consistently show that physically informed and operationally meaningful feature sets improve both accuracy and robustness. This improvement is observed across multiple forecasting horizons, from minutes to days. In practice, total data-center power consumption is composed of multiple interacting components. These include IT load, cooling demand, electrical infrastructure losses, and auxiliary consumption [16,17]. This relationship can be expressed as

$$P_{\text{facility}}(t) = P_{\text{it}}(t) + P_{\text{cooling}}(t) + P_{\text{elect}}(t) + P_{\text{auxl}}(t), \quad (1)$$

where each term represents a physically distinct energy component. $P_{\text{facility}}(t)$ is the total data center facility power at time t . The term $P_{\text{it}}(t)$ represents the IT power consumed by servers, CPUs, GPUs, memory, storage, and networking devices. The cooling power $P_{\text{cooling}}(t)$ includes computer room air conditioner (CRAC)/computer room air handler (CRAH) units, chillers, pumps, and fans. The electrical infrastructure power $P_{\text{elect}}(t)$ represents UPS, PDUs, transformers, and distribution losses, while $P_{\text{auxl}}(t)$ includes auxiliary loads such as lighting, monitoring systems, security, and other support equipment. This

decomposition also leads to widely used efficiency metrics. One of the most important is power usage effectiveness (PUE), defined as

$$\text{PUE}(t) = \frac{P_{\text{facility}}(t)}{P_{\text{it}}(t)}. \quad (2)$$

$\text{PUE}(t)$ is not constant. It varies with ambient temperature, cooling design, and operational strategies.

Here, $\text{PUE}(t)$ denotes an instantaneous operational indicator used for short-term forecasting and control, whereas conventionally reported PUE is calculated from accumulated facility and IT energy over a longer reporting period, typically one year [45,46].

3.1. IT Utilization Metrics

IT usage metrics are the key features for determining the load on a data center. They are a reflection of workload intensity and will directly impact both the amount of power consumed and heat produced by a data center. CPU usage is often used as a first-order proxy for IT usage because of its strong correlation with dynamic power-related behaviors. The most frequently used baseline-plus-activity model is given by

$$P_{\text{server}}(t) = P_{\text{idle}} + (P_{\text{peak}} - P_{\text{idle}})u_{\text{CPU}}(t), \quad (3)$$

where P_{idle} represents the idle power, P_{peak} is peak server power, and $u_{\text{CPU}}(t)$ is the normalized CPU utilization, which varies from 0 and 1. It helps in determining the incremental power [47].

Modern data centers are increasingly dependent on a diverse range of accelerators. AI training and inference workloads that run on GPUs have different power profiles than those that run solely on CPUs. The peak power of these GPU workloads typically exceed much higher than their CPU counterpart, and the frequency at which they operate at peak power also varies significantly faster. Hence, these workloads often exhibit higher peak power and faster variations, discussed in [48]. For this reason, CPU-only modeling is no longer sufficient. GPU utilization, memory bandwidth, and accelerator-level metrics are now essential for accurate forecasting.

In addition to simple utilization, performance counters can give more profound insight into the characteristics of the workloads. For example, the instructions per cycle (IPC), cache miss rate, memory bandwidth, and I/O activity can quantify system-level stress [49]. When used together with the signals for memory and I/O, these measurements also enhance short-term forecasting capability, especially with mixed workloads [50]. At a higher level, workload descriptors such as job arrival rate, queue length, and number of active virtual machines provide demand-side information [40]. Temporal feature engineering is also critical. Lag features, rolling statistics, and ramp rates help capture inertia and bursty behavior in workload evolution. A generalized IT feature vector can be represented as

$$\mathbf{x}_{\text{IT}}(t) = \left[u_{\text{CPU}}, u_{\text{GPU}}, b_{\text{mem}}, q_{\text{jobs}}, n_{\text{VM}}, \Delta u_{\text{CPU}}, \bar{u}_{\text{CPU}}^{(w)} \right]_t, \quad (4)$$

where $\mathbf{x}_{\text{IT}}(t)$ is the IT-related feature vector at time t , $u_{\text{CPU}}(t)$ is CPU utilization, $u_{\text{GPU}}(t)$ is GPU utilization, $b_{\text{mem}}(t)$ is memory bandwidth, $q_{\text{jobs}}(t)$ is job-queue length, $n_{\text{VM}}(t)$ is number of active virtual machines, $\Delta u_{\text{CPU}}(t)$ refers to change in CPU utilization, and $u_{\text{CPU}}^{(w)}(t)$ refers to window-based CPU utilization. Here, $\Delta u_{\text{CPU}}(t) = u_{\text{CPU}}(t) - u_{\text{CPU}}(t - \Delta t)$ represents the short-term variation in CPU utilization, while $u_{\text{CPU}}^{(w)}(t) = \frac{1}{w} \sum_{k=0}^{w-1} u_{\text{CPU}}(t - k)$ represents the average CPU utilization over a window of w previous time steps, where the

discussion on temporal and statistical operators captures both short-term fluctuations and longer-term workload trends.

To represent the nonlinear and coupled power behavior of heterogeneous servers, the CPU-based linear formulation is shown, as discussed in [51]:

$$\begin{aligned}
 P_{\text{server}}^{\text{NL}}(t) = & P_{\text{idle}} + \alpha_1 u_{\text{CPU}}(t) + \alpha_2 u_{\text{CPU}}^2(t) + \alpha_3 u_{\text{CPU}}(t) f_{\text{CPU}}^3(t) \\
 & + \beta_1 u_{\text{GPU}}(t) + \beta_2 u_{\text{GPU}}^2(t) + \beta_3 u_{\text{GPU}}(t) f_{\text{GPU}}^3(t) \\
 & + \gamma_1 b_{\text{mem}}(t) + \gamma_2 b_{\text{mem}}^2(t) + \delta_1 u_{\text{uncore}}(t) \\
 & + \eta_1 u_{\text{CPU}}(t) b_{\text{mem}}(t) + \eta_2 u_{\text{GPU}}(t) b_{\text{mem}}(t) + \kappa s_{\text{fan}}^3(t) + \varepsilon(t),
 \end{aligned} \quad (5)$$

where u_{CPU} and u_{GPU} denote CPU and GPU utilization; f_{CPU} and f_{GPU} are their normalized operating frequencies; b_{mem} is normalized memory-bandwidth utilization; u_{uncore} represents cache, memory-controller, and interconnect activity; and s_{fan} is normalized fan speed. The coefficients α_i , β_i , γ_i , δ_i , η_i , and κ quantify the CPU, GPU, memory, uncore, interaction, and fan-power contributions, respectively, while $\varepsilon(t)$ captures measurement noise and unmodeled effects. All normalized utilization, frequency, bandwidth, and fan-speed variables are bounded between zero and one.

The linear CPU-utilization model may lose accuracy under heterogeneous HPC and parallel AI-training workloads because the same CPU-utilization level can occur under different operating frequencies, memory-bandwidth demands, GPU loads, uncore activity, and cooling requirements. In contrast, the quadratic, frequency-dependent, and interaction terms in Equation (5) account for nonlinear resource behavior, DVFS effects, and CPU–GPU–memory coupling. This nonlinear formulation is therefore more suitable for bursty and accelerator-intensive workloads, although it requires additional measurements and parameter calibration [51]. The server-level power modeling framework can be integrated into a hierarchical forecasting architecture for facility-level load prediction by grouping servers according to hardware configuration, workload characteristics, or operational conditions. The power estimates from individual servers can then be aggregated to generate cluster-level and data-center-scale power profiles. To address real-time telemetry and computational limitations, representative server telemetry can be utilized instead of collecting fine-grained parameters from every node, thereby reducing data transfer requirements while maintaining effective macro-scale forecasting performance.

3.2. Environmental and Thermal Features

The environmental and thermal features strongly influence cooling-related energy consumption. Cooling systems account for a significant fraction of total facility power, often between 30% and 50% [52]. Their behavior is highly dependent on external conditions. Important environmental variables include outdoor dry-bulb temperature, wet-bulb temperature, humidity, and airflow conditions. These indicators directly influence economizer efficiency and cooling performance [53]. Recent studies also show that spatial temperature variation plays a major role in thermal optimization [54]. Thermal telemetry provides insight into system behavior. Server inlet and outlet temperatures, along with hot-aisle and cold-aisle measurements, capture heat distribution patterns. Thermal data are important for short-term forecasting because localized hotspots can increase fan speed or compressor operation even when the IT load remains constant [35]. Cooling power can be approximated using thermodynamic efficiency relationships [20,21]:

$$P_{\text{cool}}(t) = \frac{Q_{\text{IT}}(t)}{\text{COP}(t)}, \quad (6)$$

where $P_{\text{cool}}(t)$ is the cooling-system power at time t , the heat load generated by the IT equipment is $Q_{\text{IT}}(t)$, and the coefficient of performance of the cooling system is $\text{COP}(t)$; IT heat generation is approximately proportional to IT power consumption.

3.3. Operational and Control Parameters

Operational control parameters define process of cooling, and electrical systems are managed in real time. These include temperature setpoints, airflow control, valve positions, fan speeds, and compressor stages [55]. Such control variables directly determine system operating points. Therefore, they provide information that cannot be captured using only IT or environmental data. The electrical infrastructure controls also influence power consumption. Examples include UPS operating modes, transformer loading, and battery charge/discharge cycles [56]. These control decisions introduce additional variability into facility-level power measurements. Including them as features helps distinguish workload-driven variation from infrastructure-induced variation, detailed in [42,57].

A commonly used derived feature is the overhead ratio:

$$r_{\text{oh}}(t) = \frac{P_{\text{facility}}(t) - P_{\text{it}}(t)}{P_{\text{it}}(t)} = \text{PUE}(t) - 1. \quad (7)$$

where $r_{\text{OH}}(t)$ is the normalized facility overhead at time t . This equation captures the amount of cooling, electrical infrastructure, and auxiliary system overhead that is greater than the amount of IT load. Higher values indicate greater percentage of non-IT overheads as compared to computing power requirement, discussed in [16].

3.4. Grid and Carbon-Intensity Signals

The modern data center is becoming more grid-aware and low-carbon operation. Therefore, external signals such as electricity costs and carbon intensity are now significant predictors. Time-unique pricing mechanisms, such as time-of-use tariffs and locational marginal prices, affect the schedule of workloads [50]. These signals directly influence the observed demand behaviors. Another key driver is the carbon intensity (CI). Switching of workloads at times of low-carbon periods is now widely studied as a decarbonization strategy [58]. Recent literature shows that marginal emission factors provide more accurate operational insights than average values [59].

The emissions involved in operations of data center can be written as

$$E_{\text{CO}_2}(t) = P_{\text{facility}}(t) \Delta t \cdot CI(t), \quad (8)$$

where $E_{\text{CO}_2}(t)$ is defined as the carbon emission involved in data-center electricity consumption during particular time interval Δt . $P_{\text{facility}}(t)$ is total facility power, and $CI(t)$ is grid carbon intensity at that particular time t , expressed in kgCO_2/kWh or gCO_2/kWh . Power is converted into energy by factor of time interval Δt ; hence, this equation identifies the relationship between the electricity consumption and the grid carbon intensity at that time [4,5].

3.5. Practical Feature Engineering Workflow

Data-centered prediction uses a step-by-step approach with the data engineering pipeline. The first step is to align different sources of data, including IT telemetry, cooling signal, and grid data, into a common temporal resolution [60]. The second step is to add lag variables, rolling statistics, and ramp-rate variables to capture the time-dependent relations between workloads through inertia and burst behavior [61]. The third step is to create meaningful interaction features among the different physical characteristics of the data center, such as a thermal-aware metric that incorporates IT utilization and ambient

temperature [62]. The fourth step is to apply quality-processing techniques, such as outlier detection and missing-value processing, using physically valid operating ranges as defined by the thermal guidelines [22]. The effective feature engineering requires both data-driven learning and physical system understanding. This combined approach is increasingly emphasized in modern data-center forecasting literature [47].

4. Forecasting Methods

4.1. Traditional Statistical Models

Traditional statistical approaches continue to have a major impact on load forecasting in the data center, even though their impact has changed over time; for example, some of the most frequently used methods (AR, ARMA, ARIMA, SARIMA, and Holt–Winters smoothing) presented in [63] are still being used primarily as a basis for forecasting prediction but also serve as a basis for quantifying their predictive accuracy. All of those methods provide a transparent basis, helping to separate predictable structure from complex operational variability. Figure 5 shows load forecasting models from current statistical approaches to machine learning, deep learning, and physics-based models to hybrid models.

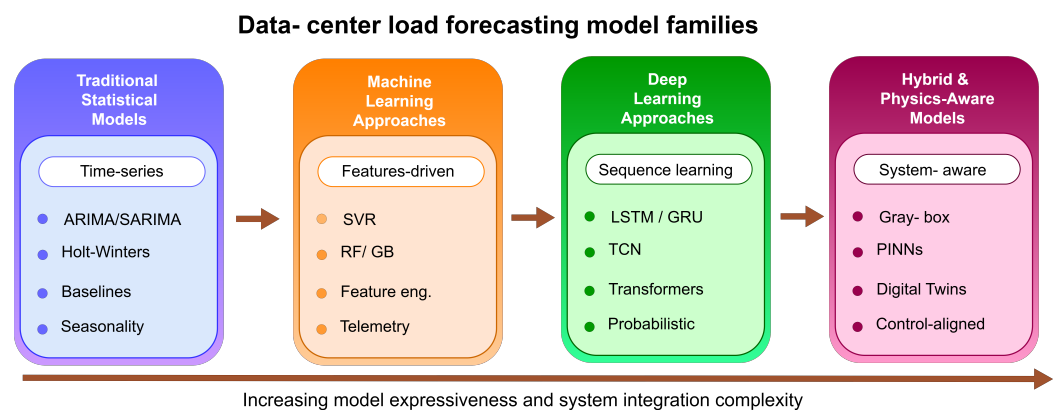


Figure 5. Data-center load forecasting model families from statistical to hybrid methods.

Power needs in data centers usually have regular predictable patterns, but these become difficult to model using fixed-parameter statistical models as workload characteristics change over time and data centers begin incorporating more complex workloads, such as artificial intelligence training and inference, autoscaling, and migration behavior, seen in more traditional deployments discussed in [64]. Consequently, demand for power in today's data centers experiences far more abrupt and unpredictable changes than previously seen and is thus not suited to traditional fixed-parameter models. In addition to these workload shifts, there are also non-linear dependencies between the IT load on the data center and the power consumed by the data center's cooling and infrastructure equipment through the chiller system, economizer system, and the control systems, which allow cooling and infrastructure systems to adjust based upon changes in IT load, illustrated in [65,66]. Therefore, traditional fixed-parameter time-series models cannot fully capture the interrelationships between the IT load on the data center and the overall power consumed by the data center.

Despite these limitations, statistical models remain important in practice [67]. They are commonly used as benchmark models when evaluating more complex approaches [68]. They can also serve as stability references in ensemble forecasting systems. In some cases, they help reduce extreme prediction errors, even when their overall accuracy is lower than learning-based models [69].

4.2. Machine Learning Approaches

Machine learning techniques offer greater flexibility by modeling nonlinear relationships between telemetry data and future load [70]. Unlike statistical models, they can accommodate many different types of input data sources. For example, IT utilization metrics, cooling system data, and environmental conditions can all be used as input [71]. Input features typically include, but are not limited to, CPU and GPU utilization, memory, storage, network activity, queue length, and workload description. Some features included are of cooling-related variables such as temperature, airflow, fan speed, and valve position [72]. Datasets with similar feature types have been used in prior research on data center workloads and thermal modeling [73].

Among machine learning techniques, regularized regression, support vector regression, random forests, and gradient boosting methods are widely adopted. Tree-based ensembles are among the most commonly used machine learning methods for capturing complex nonlinear feature interactions while being easy to train and deploy [74,75]. They rely heavily upon feature engineering to provide domain knowledge to the model, beyond simple pre-processing functions. For example, at a given level of CPU utilization, the actual power consumed by the system depends on both the cooling condition and temperature of the environment, and as such, the overall characteristics of the relationship are learned by machine learning models that receive well-engineered features [16,17,76]. In order to capture the temporal behavior associated with a data element or series of events, lagged feature values, rolling statistics, and workload descriptions are often added as feature values. These additional variables can assist with representing the inherent inertia and burstiness of demand patterns associated with the data to be modeled. A typical IT feature representation can be written as Equation (4). Researchers [17,18,70,71] have used this form of workload and server power modeling methods extensively, and the terms used are explained at the end of Equation (4). The biggest benefit of using machine learning models is their efficiency: they typically require fewer data points than deep learning models and can be retrained quickly. Interpretability methods, such as feature importance and SHAP analysis, increase the transparency of how the models perform in an operational setting [77]. However, generalization remains a challenge as models created in one facility do not transfer well to those at a different facility due to variations in cooling architecture, hardware configuration, and workloads between data centers. Additionally, the same model may perform differently as a result of changes to the system after upgrading or modifying the control policy, discussed in [78].

4.3. Deep Learning Models

Due to their ability to model multi-dimensional data over longer time periods, deep learning models have become instrumental in identifying temporal relationships in multi-dimensional data over time. LSTM and GRU model [79] architectures are commonly utilized to represent delayed correlation between changes in workload and responses to power. Another alternative for learning local temporal patterns in time series is through temporal convolutional networks (TCN), discussed in [80]. Recently, there has been a focus on transformer-based and attention-based models as they can learn temporal correlations across many telemetry data streams over long time horizons [81]. In addition to improved accuracy through deep learning models, deep learning enables structured predictive forecasting. Unlike traditional predictive modeling methods, deep learning models can not only produce a single prediction for a single time period but also produce a series of multi-step trajectory predictions. This capability is of particular importance with respect to tasks performed in operational environments (i.e., demand response, peak-load management, optimizing energy) where an understanding of historical trends is critical [82,83].

Probabilistic forecasting is another key capability. Quantile regression, Bayesian approximation, and other techniques allow models to express uncertainty in their predictions rather than simply producing point estimates. This is critical for data centers because an underestimation of the overall load will create operational risk [84]. Multi-task learning is also enabled through deep learning, as multiple variables can be predicted together using one model (e.g., IT power, cooling load, thermal state, and efficiency). This works very well because the physical systems in data centers are all interrelated. However, using deep models requires large datasets and careful tuning of hyperparameters. In addition, deep models are harder to interpret than other prediction methods, which hinders their use in environments where safety is critical. To be useful, deep models must be accurate; however, they must also be reliable, interpretable, and stable with respect to changing conditions, discussed in [85].

4.4. Hybrid and Physics-Aware Models

Hybrid and physics-aware models address a key limitation of purely data-driven methods: poor performance outside observed operating conditions [86,87]. Data center behavior follows physical laws. These include energy conservation, thermal dynamics, airflow, and equipment efficiency. Ignoring these relationships can lead to unreliable predictions during rare events such as heat waves or workload spikes. One common approach is gray-box modeling. In this approach, simple physical equations are combined with data-driven parameter learning. Another approach is residual learning. The first prediction is made by using a physics-based model. The second prediction is made using a learning model that corrects the first prediction [88]. Learning models may also include physics-based constraints to produce physically valid outputs. For example, constraints include non-negative power, bounded temperature, and monotonic relationships. Digital twin systems take the concept further. They simulate both IT and facility operations in a virtual environment and enable analysis of workload scheduling, cooling optimization, and carbon-aware operations [89,90].

Following the physics-informed learning framework used for data center airflow and thermal modeling, a representative composite loss function can be formulated as discussed in [91]:

$$\begin{aligned} \mathcal{L}_{\text{PI}} = & \lambda_d \frac{1}{N} \sum_{i=1}^N \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 \\ & + \lambda_E \frac{1}{N} \sum_{i=1}^N \left[\hat{Q}_{\text{cool},i} - c_p \hat{m}_i (\hat{T}_{\text{ret},i} - \hat{T}_{\text{sup},i}) \right]^2 \\ & + \lambda_m \frac{1}{N} \sum_{i=1}^N (\hat{m}_{\text{sup},i} - \hat{m}_{\text{ret},i})^2 \\ & + \lambda_T \frac{1}{N} \sum_{i=1}^N \left[\left[T_{\min} - \hat{T}_{\text{in},i} \right]_+^2 + \left[\hat{T}_{\text{in},i} - T_{\max} \right]_+^2 \right]. \end{aligned} \quad (9)$$

Here, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ denote the measured and predicted cooling-system variables, respectively; c_p is the specific heat capacity of the cooling fluid; $\dot{m}$ is the mass-flow rate; and $[z]_+ = \max(0, z)$. The weighting coefficients λ_d , λ_E , λ_m , and λ_T balance prediction accuracy and physical consistency. These physical residuals regularize model training by penalizing violations of energy conservation, mass-flow continuity, and permissible server-inlet temperature limits.

The main hybrid approaches to forecasting, as well as the probabilistic approach and the digital twin approach, have been summarized in Table 2. This table illustrates the way in which these approaches combine aspects of physical knowledge, uncertainty

modeling, and system level constraints to enhance both the performance of the forecast and its operational relevance. All three approaches have distinct advantages, including reduced effort required to calibrate the models, increased accuracy achieved through sensor coverage, and continuous maintenance associated with the forecasts. On the other hand, developing and operating a digital twin can be quite expensive. Nevertheless, these approaches offer a promising way forward in developing forecasting systems that are reliable, interpretable, and operationally safe. Hybrid and physics-informed forecasting approaches integrate physical constraints with machine learning and deep learning methodologies to improve the physical consistency and reliability of predictions. However, incorporating such constraints typically increases model development complexity, including challenges associated with model formulation, training, and deployment. The benefits associated with developing hybrid systems will include improved ability to support decision-making for data centers in the real world, reported in [90,92]. In addition to Table 2, Tables 3 and 4 summarize the quantitative evidence in the literature at different levels. Table 3 shows a comparison on a single AI/HPC trace, where accuracy, model size, and inference latency are measured under the same target and horizon and are therefore comparable. It shows that attention-based modeling reduces MAE from 5% for recurrent baselines to 3.14% and that a knowledge-distilled MLP retains this accuracy while using fewer parameters. Table 4 shows the reported results across studies in AI and HPC facilities. The contrast between the two tables is informative as the field has converged on a common set of error metrics but not on a common benchmark, which is why this review emphasizes trade-offs rather than a ranking by reported accuracy.

Table 2. Hybrid, probabilistic, and digital-twin-based forecasting approaches for data centers.

Approach Type	Embedded Physical Knowledge	Forecasted Variables	Data Requirements	Advantages	Application Context	Limitations	References
Probabilistic forecasting models	Statistical uncertainty distributions, temporal correlations, and prediction intervals	IT load, total facility power, cooling demand, and net load	Historical load, weather, cooling-demand, and workload data with sufficient samples for uncertainty calibration	Provide prediction intervals, quantify uncertainty, and support risk-aware operational decisions	Capacity planning, reserve allocation, demand response, and uncertainty-aware scheduling	Sensitive to distribution assumptions, tail-risk estimation, and computational cost	[68,84]
Hybrid physics–ML models	Energy-balance equations, thermal dynamics, PUE decomposition, and data-driven residual learning	Cooling power, HVAC states, facility power, temperature, and PUE	Historical power and temperature data, equipment parameters, cooling-system measurements, airflow information, and operating constraints	Improve physical consistency and generalization while reducing dependence on large training datasets	Cooling forecasting, HVAC optimization, thermal management, and control-oriented operation	Model mismatch, uncertain physical parameters, parameter identification, and loss-weight selection	[86,92]
Physics-aware neural networks	Soft or hard physical constraints embedded in network architectures or training loss functions	Power consumption, airflow, thermal states, cooling demand, and efficiency metrics	Power, temperature, airflow, and mass-flow measurements together with governing equations and permissible operating limits	Penalize physically infeasible predictions and improve robustness, interpretability, and extrapolation under sparse data	Thermal-state forecasting, airflow modeling, cooling-load prediction, and safety-aware operation	Higher training complexity and possible conflict between data-fitting and physics-based loss terms	[91,92]

Table 2. Cont.

Approach Type	Embedded Physical Knowledge	Forecasted Variables	Data Requirements	Advantages	Application Context	Limitations	References
Digital-twin-based forecasting	High-fidelity representations of IT, cooling, and electrical systems synchronized with real-time operational data	Power, temperature, PUE, cooling states, and equipment conditions	Real-time sensor streams, system topology, equipment specifications, control signals, and model-calibration data	Enable continuous monitoring, virtual experimentation, fault diagnosis, and evaluation of alternative operating scenarios	Predictive maintenance, anomaly detection, capacity planning, real-time monitoring, and what-if analysis	High modeling, integration, calibration, maintenance, and computational costs	[89,90]
Hybrid digital twin and probabilistic models	Physics-based digital twins combined with uncertainty quantification and probabilistic inference	Probabilistic power, cooling, temperature, and equipment-state trajectories	Digital twin states, real-time measurements, historical disturbances, uncertainty distributions, and scenario data	Provide physically grounded uncertainty estimates and support robust, risk-aware operational decisions	Robust optimization, resilience analysis, contingency assessment, and risk-aware energy management	Complex uncertainty propagation, continuous calibration requirements, and high computational burden	[84,86,89,92]

Table 3. Forecasting accuracy and computational cost on the MIT SuperCloud AI/HPC benchmark, with normalized-load errors compiled and cross-validated from Refs. [61,70,81]; the final column identifies representative studies applying each model.

Model	MAE [%]	RMSE [%]	Parameters	FP32 Memory [KB]	Model Size [KB]	Train/Val/Test Split [%]	Related Studies
GRU (baseline)	5.35	14.09	64.0K	250.00	430.00	70/15/15	[29,79,86]
LSTM (baseline)	5.12	13.64	81.9K	320.00	520.00	70/15/15	[6,48,82]
Transformer	3.14	9.96	69K	270.00	410.00	70/15/15	[71,81]
Distilled MLP	3.25	10.18	6.1K	23.82	27.74	70/15/15	[61,70]

Notes: We observe that the LSTM is best in memory size. The knowledge-distilled MLP matches the transformer teacher's accuracy while cutting the parameter count by approximately 13 times.

Table 4. The table shows the inventory of the predicted performance across several data centers and cloud workload.

Ref.	Method	Dataset	Target	Horizon	Metric	Reported Result
(A) Facility with AI and HPC data centers						
[61]	Transformer → MLP	MIT SuperCloud (public)	GPU power	1 min	MAE, RMSE	MAE 3.25%, RMSE 10.18%; transformer 3.14% / 9.96%; GRU (5.35% / 14.09%)
[82]	LSTM / 1D-CNN	GPU load	Node power	Short-term	MAE, RMSE	LSTM performs the best
(B) Cloud workload and resource utilization						
[29]	CNN-GRU + attention layers	Google (public)	multivariate workload	Multi-step	MAE, RMSE, MAPE	Outperforms CNN, LSTM baseline
[79]	GRU + multi-head attention	Google/Alibaba (public)	Cloud workload	Multi-step	MAE, RMSE, MAPE	Outperforms baselines
[6]	VSBG	Large-scale	Cloud Workload	Multi-step	RMSE, MAPE	Outperforms ARIMA
(C) Facility total						
[48]	Multi-feature LSTM	DC demand	Peak demand	Short-term	RMSE, MAPE	depends on the architecture
[86]	DLinear + GRU	DC load	cooling load	Short-term	RMSE, MAPE	depends on the architecture
[35]	GNN	DC load	Temporary outlet	Short-term	RMSE, MAE	depends on the architecture
[36]	Data-driven models	Building-district system	Cooling and heating load	Short-term	RMSE, MAPE, R ²	different domain

5. Applications of Load Forecasting in DC

Forecasting load has transitioned from only being a support tool for capacity planning to being part of modern DC operations, as discussed in [93]. In former days, load forecasting was used principally as a tool for infrastructure planning and for estimating capacity. Today, however, forecasting is a component of many data center operational workflows, including workload scheduling, cooling control, grid interaction, and carbon-aware optimization [94]. This transition represents an overall shift in the DC functioning. DCs are no longer just passive consumers of electricity but are becoming more and more integrated into the larger energy system and into electricity markets.

As highlighted in [4], modern facilities must also respond dynamically to external signals such as grid demand, pricing, and sustainability constraints. Because of this shift, forecasting is no longer limited to a single time scale. It must support both real-time operational control and day-ahead or long-term planning decisions [95].

5.1. Demand Response and Grid Interaction

A key application of load forecasting is participation in demand response programs and grid-interactive operation. Prior work has shown that data centers can actively support grid stability by adjusting computational loads under operational constraints [96]. This is possible because a portion of the workload can be shifted or delayed without violating service requirements. Short-term forecasting plays a critical role in this process. These projections enable proactive adjustments to system settings under uncertainty about future demand. A common approach to the management of systems through forecasting includes implementing workload redistribution on an hourly basis and cooling control strategies. Furthermore, if energy resources available at the site could be managed to coincide with these forecasts, they could provide a benefit by improving demand response program participation while maintaining service level commitments [97].

From an economic standpoint, forecasting is essential for predicting future behavior. For example, the use of dynamic pricing models (e.g., time-of-use tariffs and locational marginal pricing) have shown that the benefits of predictive scheduling exist. Specifically, through the use of forecasting, the ability of operators to shift their flexible workloads away from high-priced time frames results in lower operating costs and reduced peak demand charges [98]. Additionally, the use of forecasting will improve the operation of local energy storage systems. By using forecasts to match battery charge and discharge decisions with anticipated load peaks, electricity price patterns can help improve operating efficiency as well as enhance system reliability, as demonstrated in prior work on coordinated energy management [99].

5.2. Energy-Aware Workload Scheduling

The energy-aware scheduling method relies on load forecasting to determine when and where to run computation tasks. Instead of waiting to observe actual conditions in real time and respond to them, energy-aware scheduling uses pre-calculated (forecasted) load/energy information to optimize the schedule in advance, presented in [100]. Forecasts are used to determine whether to run workloads immediately, later, move them to a different site, or throttle them. This is very different from traditional performance-oriented schedulers that only consider system utilization at the time a task is executed and do not account for future utilization [101,102]. By taking advantage of temporal variations in both energy pricing and environmental conditions, predictive scheduling enables workloads to be executed at times that reduce cooling demand in the DC or when electricity is less expensive. Previous researchers consistently show that predictive scheduling is superior to reactive scheduling when the workload is delay-tolerant, such as in batch analytics and

large-scale data processing [103]. A fundamental requirement of these system types is reliable, accurate short- to mid-term predictions of both workload changes over time and the infrastructure's impact on those changes, from the perspective of cooling and electricity related limitations.

5.3. Carbon-Aware Load Shifting

Carbon-aware computing has become a major driver for modern load forecasting research due to increasing pressure to reduce emissions from large-scale computing infrastructures, detailed in [104]. In carbon-aware load shifting, forecasts of both electricity demand and grid carbon intensity are combined to determine optimal scheduling times for flexible workloads. This allows computation to be shifted toward periods of lower carbon intensity or higher renewable energy availability [105]. In addition to that, accurate load forecasting is essential in this context because emissions depend not only on grid carbon intensity but also on the actual power consumption profile of the facility. Even small forecasting errors can significantly affect emission estimates when evaluating alternative scheduling strategies.

Recent studies in [106] emphasize the importance of marginal carbon intensity over average carbon intensity because marginal signals better capture the system-level impact of incremental demand changes. This makes marginal-based forecasting more suitable for operational decision-making in data centers.

5.4. Integration with Real-Time Control Systems

Load forecasting is becoming an important emerging use case for real-time control systems. In real time, modern data centers use forecasts as input for closed-loop control systems that dynamically manage power, cooling, and workload. Forecast information allows operators to proactively adjust control variables (e.g., airflow rates, power caps, and energy storage dispatch) rather than react after a deviation occurs, discussed in [107].

The application of MPC is an excellent illustration of this type of integrated system. With MPC, a system uses its load and environmental forecast to determine the best control action for its future usage over a set period (the predictive horizon). This control action is then continually updated as new information is received (e.g., updated forecasts) [108]. Compared to traditional control methods (such as reactive), forecast-assisted methods, discussed in [109], show superior performance (in reducing energy consumption, maintaining temperature safety, and improving overall system stability) when there are frequent and rapid fluctuations in workload and limited thermal tolerance for heat output. Due to the strong interrelationship between forecasting and control, the research community is now taking a more integrated approach to both, treating them as a single unified system rather than two separate entities. This is particularly important for large-scale, grid-interactive data centers that are subject to volatile energy and workload conditions, highlighted in [106].

The framework in Figure 6 summarizes the comprehensive methodology for forecasting data center loads and the practical use of estimated load values generated from those forecasts. The components utilized to develop the forecasted load values are largely historical data collected from the sites containing all available inputs: electrical characteristics such as IT loads, cooling/HVAC loads, auxiliary load, facility loads (net, averages, min/max), workload utilization like CPU/GPUs, memory usage, and environmental and thermal features such as inlet/rack temperature, humidity, ambient temperature, as well as signals sent from the grid, utility, and carbon markets like electricity price, demand response and carbon intensity. These are then processed through a series of feature-engineering steps for each input to create a dataset of processed time-series data. These steps involved are

(1) data cleaning; (2) handling missing values; (3) normalizing datasets; (4) generation of lagged features; (5) temporal encoding; (6) sliding-window formation (for example, moving averages of input characteristics); and (7) feature extraction. These processed time-series data are then used as input for the deep learning forecasting model (e.g., LSTM/GRU) to learn temporal dependencies in data center load patterns, generating multi-horizon forecasts for IT load, cooling load, total facility load, and PUE and predicting ahead of time, as shown in the figure. $t + 10$ min, $t + 3$ h, and $t + 48$ h are multi-horizon predictions in minutes, hours, and days. The forecasted load values produced can, subsequently, be utilized for a number of operational applications such as demand response and grid interaction, energy-aware workload Scheduling, carbon-aware load shifting, integration with real-time control systems, as well as contribute to cost optimization, renewable energy, and thermal management within DCs (long-term capacity planning) in today's AI-enabled data centers.

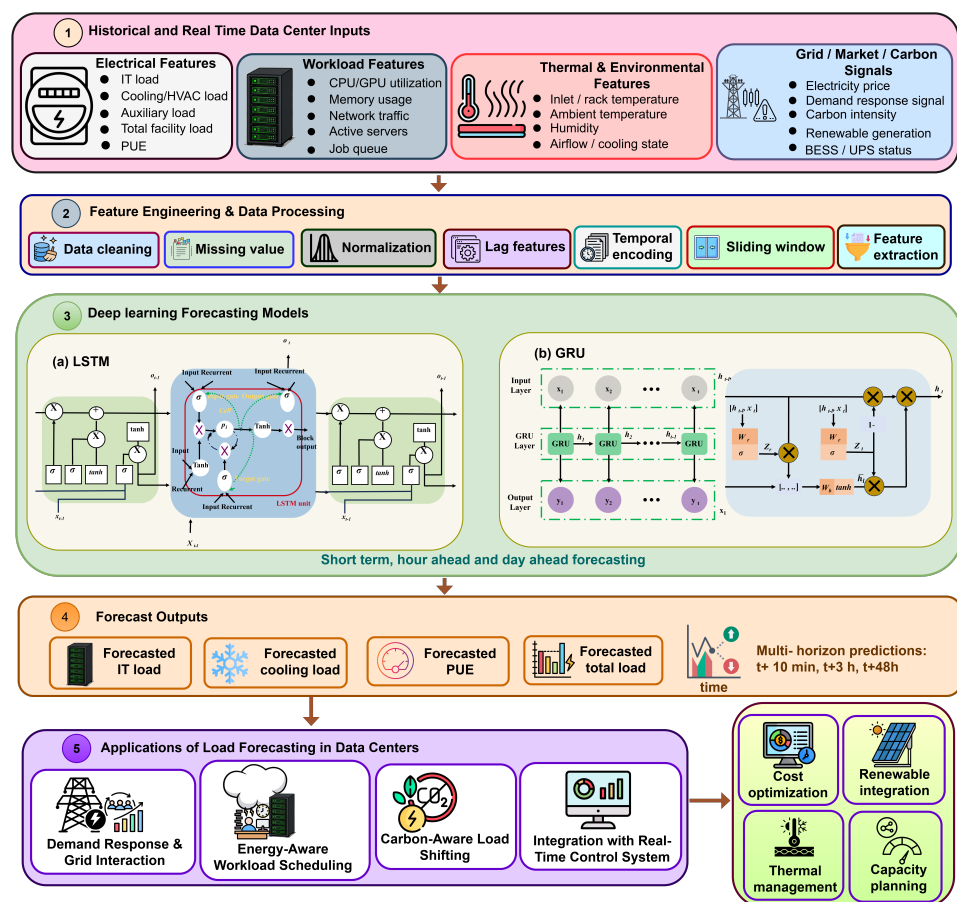


Figure 6. Multi-source data-driven load forecasting framework for data center control and scheduling.

6. Challenges and Research Gaps

Despite significant progress in data center load forecasting, several important challenges remain. Existing studies have improved forecasting accuracy using advanced machine learning, deep learning, and physics-informed approaches. Challenges associated with real-time data access and availability, model generalization, and interpretability of model outputs are currently limiting the practical implementation of machine learning. In addition, the need for improved forecasting models to support energy flexibility, enable interaction with the electric grid, and enable low-carbon operations for contemporary DCs has been highlighted by recent studies. Resolving these hurdles will increase energy efficiency, enabling intelligent DC operation.

6.1. Limited Availability of Public Datasets

The lack of available high-quality public datasets continues to hinder the forecasting of DC loads. Due to concerns about security, privacy, and commercial use, many data center operators limit access to operational data, discussed in [110]. Thus, many studies have utilized proprietary datasets, synthetic workloads, or experimental workloads from other sources. Some publicly available datasets contain only aggregated energy consumption measurements but generally lack important information on workload characteristics, cooling system performance, environmental conditions, control settings, and grid carbon intensity. As a result of this limited amount of publicly available data, many models may not be able to improve accuracy when using spatial, operational, and physics awareness, as features used in the model will capture site characteristics, airflow, and thermal constraints that result from the operation of a DC [111]. Lack of comprehensive datasets has limited the use of feature engineering and prevented researchers from comparing forecasting methods. As such, developing standard benchmark datasets is an important research focus.

6.2. Generalization Across Data Centers

Forecasting models are typically designed and validated using information from only one facility. Although these models may achieve high accuracy for the target site, their performance often degrades when applied to other DCs. This occurs as a result of different DCs having varying infrastructures, cooling technologies, operational policies, and geographical environments, making forecasting performance at a data center difficult, and deploying it across several data centers is a major challenge. In order to resolve this concern, researchers have begun utilizing transfer learning and domain adaptation methods to ascertain the influence of spatial configuration change on model accuracy [112]. There is still no systematically developed framework for cross-project prediction; therefore, future investigations should concentrate on establishing a model that requires little or no retraining and/or calibration when transitioning from one site to another.

6.3. Accuracy vs. Interpretability Trade-Offs

Forecasting research faces the challenge of mediating prediction accuracy vs. model transparency. Deep learning models have shown good performance in forecasting because they can represent nonlinear relationships in the data that are hard to understand intuitively [113]. Ref. [114], demonstrated very good predictive performance using advanced data-driven techniques, but these models are often treated as black-box systems and therefore lack transparency in their predictions. On the other hand, physics-informed and hybrid models allow for the use of known relationships between IT load, cooling systems, and environmental conditions in the modeling process, making them generally more interpretable and explainable than pure physics-based approaches. However, they generally do not have the level of accuracy associated with deep learning as they are not able to capture rapid changes in operational conditions. Ref. [115] emphasized the need for transparency in the models used to make operational decisions and provided examples of data-driven processes to enable intelligent management of DCs. The current research area of bridging the gap between interpretability and accuracy remains an ongoing challenge.

6.4. Real-Time Deployment Constraints

Most of the models used in forecasting are tested in an offline situation, where the amount of computation and length of time consumed are not significant considerations. However, this added complications when forecasting is carried out in a real-time manner. Ref. [116] states that operational forecasting system requirements must comply with requirements on latency, reliability, and computing efficiency. Real-world data streams can

experience errors because of sensor malfunctions, communication delays, missing values, and noise within the measurements made by sensors; therefore, these factors cause a large decrease in the accuracy of the forecasts. In addition, periodic retraining may not always be possible in a production environment due to scale. Because of the variability of workloads and operating conditions [117], forecasting must be both adaptive and robust to enable real-time control. Therefore, developing self-learning and adaptive forecasting models is an important future research direction.

An important but often overlooked issue is the computational and environmental burden associated with deploying advanced forecasting models in data centers. Large transformer-based models and deep recurrent networks may require considerable GPU or TPU resources for initial training, periodic retraining, and continuous inference. The electricity consumed by these processes, together with the additional cooling demand, hardware utilization, and related carbon emissions, may partially reduce the energy savings obtained through improved forecasting and control. Therefore, future studies should evaluate not only forecasting accuracy but also training and inference energy consumption, GPU-hours, computational latency, retraining frequency, and associated carbon emissions. In addition, lightweight model architectures, model compression, quantization, knowledge distillation, and event-triggered retraining should be further investigated to improve the overall energy efficiency and environmental sustainability of AI-based data center management [118,119].

6.5. Integration with Control and Optimization Systems

In addition, one of the largest gaps in current research is the lack of integration between forecasting models and optimization/control frameworks. Most of the available studies only measure the performance of forecasts using statistical metrics, e.g., MAE/RMSE/MAPE without considering the impact of forecast errors on operational decisions. In practice, forecasts are an input to demand response programs and workload scheduling algorithms, as well as MPC. The accuracy of the forecasted value allows the MPC/robust controller to work within the thermal limits of a system, manage risk associated with uncertainty, and operate according to the other physical/engineering limits of a system [120,121].

Although there has been an increase in interest regarding integrated forecasting and control frameworks, methodologies capable of jointly optimizing forecasting accuracy and control goals are still limited. Future research into this area should view forecasting uncertainty, energy efficiency, operational reliability, and control performance as interconnected goals. It is also clear that just improving forecasting accuracy will not be enough for the next generation of intelligent DCs. Ref. [122] focused on the need for data-aware and sustainability-motivated computing strategies and discussed the benefits of integrating forecasting with energy management systems. Further progress in this area depends upon improved datasets, greater generalization of models, enhanced interpretability, reliable real-time deployment, and more tightly connected forecasting, control, and optimization systems.

7. Future Directions

Based on the reviewed literature, the identified research directions can be organized into near-term, medium-term, and long-term priorities for future data-center load forecasting studies. Near-term directions address limitations that can be investigated using currently available datasets and models. Medium-term directions require closer integration between forecasting and operational systems, whereas long-term directions involve autonomous and fully coordinated data center energy management frameworks.

7.1. Near-Term Research Directions

1. Future forecasting systems should quantify prediction uncertainty to support risk-aware operational decisions under uncertain workloads and grid conditions. This discussion is also highlighted in [123].
2. Explainable forecasting models are needed to improve operator trust and support decision-making in critical data center operations. This discussion is also highlighted in [124].
3. Future models should simultaneously predict multiple variables such as IT load, cooling demand, PUE, and carbon emissions. Such approaches can improve prediction accuracy and reduce model complexity. This discussion is also highlighted in [125].

7.2. Medium-Term Research Directions

1. Combining AI models with physical and thermodynamic constraints can improve forecasting robustness, reliability, and generalization under varying operating conditions. This discussion is also highlighted in [126].
2. Transfer learning can enable model deployment across different data centers with limited training data. Domain adaptation methods can further address variations caused by climate conditions, workloads, and cooling technologies. This discussion is also highlighted in [127,128].
3. Future forecasting models should be integrated with energy management systems, demand response programs, and smart-grid operations instead of functioning as standalone prediction tools. This discussion is also highlighted in [129,130].
4. Future forecasting frameworks should incorporate carbon intensity, renewable energy availability, and sustainability metrics to support low-carbon data center operation. This discussion is also highlighted in [131].
5. Forecasting models should consider electricity prices, demand response signals, renewable variability, and grid constraints to support flexible and grid-friendly data center operation. This discussion is also highlighted in [123].

7.3. Long-Term Research Directions

1. Future data centers are expected to use forecasting outputs for real-time control, optimization, workload scheduling, and cooling management. Digital twins, MPC, and reinforcement learning are expected to play significant roles. This discussion is also highlighted in [130,132].
2. A major research challenge is the development of unified frameworks that combine forecasting, control, optimization, and decision-making while ensuring stability, safety, reliability, and interpretability. This discussion is also highlighted in [4].
3. Future studies should validate forecasting models using data from multiple data centers and different operating conditions to improve model generalization and reliability.

Load forecasting is evolving from a standalone prediction task into a key intelligence layer for autonomous, sustainable, and grid-interactive data center energy management systems.

8. Conclusions

This paper presented a critical review of AI and ML methods for load forecasting in modern data centers. It covered forecasting models, feature engineering, applications, deployment issues, and recent research forecasting trends. The review shows that accurate load forecasting is very important. It helps improve energy efficiency, thermal management, system reliability, and carbon-aware operation in modern data centers. Recent research

in deep learning, physics-informed models, and digital twins have improved forecasting performance. These methods help in better coordination of workloads, cooling systems, and power systems. However, some challenges still remain. These include limited good-quality data, poor generalization across sites, low model interpretability, real-time deployment issues, and weak integration with control systems. As data centers become more AI-driven, load forecasting will play a key role in future energy management. Future work should focus on explainable models, physics-based learning, adaptive forecasting, multi-objective optimization, and better integration with control and carbon-aware strategies. Therefore, it can be said that AI-based forecasting will become an important part of smart and energy-efficient data centers.

Author Contributions: All authors contributed equally to this work. Conceptualization, investigation, methodology, writing—original draft preparation, reviewing, and editing were collectively performed by S.S., S.T., G.S.C., S.C., W.S., M.W., M.N., and B.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT-4o and Grammarly to assist with grammar and spelling corrections, as well as to enhance sentence clarity and overall readability. No content was generated by AI tools beyond language refinement.

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

Abbreviations

AI	Artificial Intelligence
AR	Auto Regressive Model
ARIMA	Auto Regressive Integrated Moving Average model
ARMA	Auto Regressive Moving Average model
CI	Carbon Intensity
CPU	Central Processing Unit
CRAC	Computer Room Air Conditioner
CRAH	Computer Room Air Handler
DC	Data Center
DT	Digital Twin
DVFS	Dynamic Voltage and Frequency Scaling
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HPC	High Performance Computing
HVAC	Heating, Ventilation, and Air Conditioning
IPC	Instructions Per Cycle
IT	Information Technology
LMP	Locational Marginal Pricing
LSTM	Long Short-Term Memory
ML	Machine Learning
MPC	Model Predictive Control
PDU	Power Distribution Unit
SARIMA	Seasonal Auto Regressive Integrated Moving Average model
SHAP	Shapley Additive Explanations
SOC	State of Charge
SOH	State of Health
TCN	Temporal Convolutional Network
TPU	Tensor Processing Unit

UPS Uninterruptible Power Supply
VM Virtual Machine

References

- De Roucy-Rochegonde, L.; Buffard, A. AI, Data Centers and Energy Demand: Reassessing and Exploring the Trends. Institut Français des Relations Internationales. 2025. Available online: <https://www.ifri.org/en/papers/ai-data-centers-and-energy-demand-reassessing-and-exploring-trends-0> (26 February 2026)).
- Raimi, D.; Joiner, E.; Hubbell, B.; Lohawala, N.; Robertson, M. Global Energy Outlook 2026: How the World Lost the Goal of 1.5 °C. Technical Report, Resources for the Future. 2026. Available online: <https://www.rff.org/publications/reports/global-energy-outlook-2026/> (23 April 2026).
- Liu, J.; He, X.; Li, K.; Xue, W. A Review of State-of-the-Art AI and Data-Driven Techniques for Load Forecasting. *Energies* **2025**, *18*, 4408. <https://doi.org/10.3390/en18164408>.
- Radovanović, A.; Koningstein, R.; Schneider, I.; Chen, B.; Duarte, A.; Roy, B.; Xiao, D.; Haridasan, M.; Hung, P.; Care, N.; et al. Carbon-Aware Computing for Datacenters. *IEEE Trans. Power Syst.* **2023**, *38*, 1270–1280. <https://doi.org/10.1109/TPWRS.2022.3173250>.
- Buchanan, W.; Foxon, J.; Cooke, D.; Iyer, S.; Graham, E.; DeRusha, B.; Binder, C.; Chiu, K.; Corso, L.; Richardson, H.; et al. Carbon-Aware Computing: Measuring and Reducing the Carbon Intensity Associated with Software in Execution. Technical Report, Microsoft, UBS, WattTime, and Green Software Foundation. 2023. Available online: <https://news.microsoft.com/de-ch/2023/01/10/carbon-aware-computing-whitepaper/> (accessed on 12 May 2026).
- Yuan, H.; Bi, J.; Li, S.; Zhang, J.; Zhou, M. An Improved LSTM-Based Prediction Approach for Resources and Workload in Large-Scale Data Centers. *IEEE Internet Things J.* **2024**, *11*, 22816–22829. <https://doi.org/10.1109/JIOT.2024.3383512>.
- Lorido-Botran, T.; Miguel-Alonso, J.; Lozano, J.A. A review of auto-scaling techniques for elastic applications in cloud environments. *J. Grid Comput.* **2014**, *12*, 559–592. <https://doi.org/10.1007/s10723-014-9314-7>.
- Masdari, M.; Khoshnevis, A. A survey and classification of the workload forecasting methods in cloud computing. *Clust. Comput.* **2020**, *23*, 2399–2424. <https://doi.org/10.1007/s10586-019-03010-3>.
- Beloglazov, A.; Abawajy, J.; Buyya, R. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Gener. Comput. Syst.* **2012**, *28*, 755–768. <https://doi.org/10.1016/j.future.2011.04.017>.
- Hamdi, N.; Chainbi, W. A survey on energy aware VM consolidation strategies. *Sustain. Comput. Inform. Syst.* **2019**, *23*, 80–87. <https://doi.org/10.1016/j.suscom.2019.06.003>.
- Chaurasia, N.; Kumar, M.; Chaudhry, R.; Verma, O.P. Comprehensive survey on energy-aware server consolidation techniques in cloud computing. *J. Supercomput.* **2021**, *77*, 11682–11737. <https://doi.org/10.1007/s11227-021-03760-1>.
- Helali, L.; Omri, M.N. A Survey of Data Center Consolidation in Cloud Computing Systems. *Comput. Sci. Rev.* **2021**, *39*, 100366. <https://doi.org/10.1016/j.cosrev.2021.100366>.
- Zhao, R.; Du, Y.; Yang, X.; Zhou, Z.; Wang, W.; Yang, X. A Critical Review on the Thermal Management of Data Center for Local Hotspot Elimination. *Energy Build.* **2023**, *297*, 113486. <https://doi.org/10.1016/j.enbuild.2023.113486>.
- Liu, H.; Aljbri, A.; Song, J.; Jiang, J.; Hua, C. Research Advances on AI-Powered Thermal Management for Data Centers. *Tsinghua Sci. Technol.* **2022**, *27*, 303–314. <https://doi.org/10.26599/TST.2021.9010019>.
- Athavale, J.; Bash, C.; Brewer, W.; Maiterth, M.; Milojicic, D.; Petty, H.; Sarkar, S. Digital Twins for Data Centers. *Computer* **2024**, *57*, 151–158. <https://doi.org/10.1109/MC.2024.3436945>.
- Dayarathna, M.; Wen, Y.; Fan, R. Data Center Energy Consumption Modeling: A Survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 732–794. <https://doi.org/10.1109/COMST.2015.2481183>.
- Jin, C.; Bai, X.; Yang, C.; Mao, W.; Xu, X. A Review of Power Consumption Models of Servers in Data Centers. *Appl. Energy* **2020**, *265*, 114806. <https://doi.org/10.1016/j.apenergy.2020.114806>.
- Mei, X.; Wang, Q.; Chu, X. A Survey and Measurement Study of GPU DVFS on Energy Conservation. *Digit. Commun. Netw.* **2017**, *3*, 89–100. <https://doi.org/10.1016/j.dcan.2016.10.001>.
- Sheng, Y.; Zhang, C.; Zhu, Z.; Xu, H.; Wen, J.; Wang, R.; Yang, J.; Wang, Q.; Bu, S. Power for AI Data Centers: Energy Demand, Grid Impacts, Challenges and Perspectives. *Energies* **2026**, *19*, 722. <https://doi.org/10.3390/en19030722>.
- Al Kez, D.; Foley, A.M.; Hasan Wong, F.W.B.M.; Dolfi, A.; Srinivasan, G. AI-Driven Cooling Technologies for High-Performance Data Centres: State-of-the-Art Review and Future Directions. *Sustain. Energy Technol. Assess.* **2025**, *82*, 104511. <https://doi.org/10.1016/j.seta.2025.104511>.
- Zhao, J.; Chen, Z.; Li, H.; Liu, D. A Model Predictive Control for a Multi-Chiller System in Data Center Considering Whole System Energy Conservation. *Energy Build.* **2024**, *324*, 114919. <https://doi.org/10.1016/j.enbuild.2024.114919>.
- ASHRAE Technical Committee 9.9. *Thermal Guidelines for Data Processing Environments*, 5th ed.; ASHRAE: Peachtree Corners, GA, USA, 2021.

23. Fan, X.; Weber, W.D.; Barroso, L.A. Power Provisioning for a Warehouse-Sized Computer. In Proceedings of the 34th Annual International Symposium on Computer Architecture, San Diego, CA, USA, 9–13 June 2007; pp. 13–23. <https://doi.org/10.1145/1273440.1250665>.
24. Ahmed, K.M.U.; Alvarez, M.; Bollen, M.H. Reliability Analysis of Internal Power Supply Architecture of Data Centers in Terms of Power Losses. *Electr. Power Syst. Res.* **2021**, *193*, 107025. <https://doi.org/10.1016/j.epsr.2021.107025>.
25. Kim, J.; Kim, H.G.; Shin, D.U. Data Center Energy Evaluation Tool Development and Analysis of Power Usage Effectiveness with Different Economizer Types in Various Climate Zones. *Buildings* **2024**, *14*, 299. <https://doi.org/10.3390/buildings14010299>.
26. Safari, A.; Sorouri, H.; Rahimi, A.; Oshnoei, A. A Systematic Review of Energy Efficiency Metrics for Optimizing Cloud Data Center Operations and Management. *Electronics* **2025**, *14*, 2214. <https://doi.org/10.3390/electronics14112214>.
27. Shehabi, A.; Smith, S.J.; Hubbard, A.; Newkirk, A.; Lei, N.; Siddik, M.A.; Holecsek, B.; Koomey, J.G.; Masanet, E.R.; Sartor, D.A. *2024 United States Data Center Energy Usage Report*; Technical Report; Lawrence Berkeley National Laboratory: Berkeley, CA, USA, 2024. <https://doi.org/10.71468/P1WC7Q>.
28. Caprara, A.; Yu, Y.; Teng, F.; Junyent-Ferré, A.; Bullich-Massagué, E.; Aragüés-Peñalba, M. Data center workload flexibility for power system demand response: Evidence from Alibaba traces. *Int. J. Electr. Power Energy Syst.* **2026**, *178*, 111940. <https://doi.org/10.1016/j.ijepes.2026.111940>.
29. Dogani, J.; Khunjush, F.; Mahmoudi, M.R.; Seydali, M. Multivariate workload and resource prediction in cloud computing using CNN and GRU by attention mechanism. *J. Supercomput.* **2023**, *79*, 3437–3470. <https://doi.org/10.1007/s11227-022-04782-z>.
30. Parsa, Z.; Oleksandr, H.; Jacek, R. Data Centers as a Driving Force for the Renewable Energy Sector. *Energies* **2025**, *19*, 236. <https://doi.org/10.3390/en19010236>.
31. Duc, T.L.; Nguyen, C.; Östberg, P.O. Workload prediction for proactive resource allocation in large-scale cloud-edge applications. *Electronics* **2025**, *14*, 3333. <https://doi.org/10.3390/electronics14163333>.
32. Wang, Y.; Guo, Q.; Chen, M. Providing Load Flexibility by Reshaping Power Profiles of Large Language Model Workloads. *Adv. Appl. Energy* **2025**, *19*, 100232. <https://doi.org/10.1016/j.adapen.2025.100232>.
33. Liao, X.; Li, Y.; Zhang, S.; Wei, X.; Hu, J. Predicting GPU Training Energy Consumption in Data Centers Using Task Metadata via Symbolic Regression. *Energies* **2026**, *19*, 448. <https://doi.org/10.3390/en19020448>.
34. Tang, Q.; Mukherjee, T.; Gupta, S.K.S.; Cayton, P. Sensor-Based Fast Thermal Evaluation Model for Energy Efficient High-Performance Datacenters. In Proceedings of the 2006 Fourth International Conference on Intelligent Sensing and Information Processing, Bangalore, India, 15 October–18 December 2006; pp. 203–208. <https://doi.org/10.1109/ICISIP.2006.4286097>.
35. Sha, Q.; Yang, J.; Shao, R.; Wang, Y. Prediction of Air-Conditioning Outlet Temperature in Data Centers Based on Graph Neural Networks. *Energies* **2025**, *18*, 1803. <https://doi.org/10.3390/en18071803>.
36. Yu, H.; Zhong, F.; Du, Y. Short-Term Cooling and Heating Loads Forecasting of Building District Energy System Based on Data-Driven Models. *Energy Build.* **2023**, *298*, 113513. <https://doi.org/10.1016/j.enbuild.2023.113513>.
37. Li, X.; Li, M.; Zhang, Y.; Han, Z.; Wang, S. Rack-Level Cooling Technologies for Data Centers—A Comprehensive Review. *J. Build. Eng.* **2024**, *90*, 109535. <https://doi.org/10.1016/j.jobte.2024.109535>.
38. Wang, J.; Deng, H.; Liu, Y.; Guo, Z.; Wang, Y. Coordinated Optimal Scheduling of Integrated Energy System for Data Center Based on Computing Load Shifting. *Energy* **2023**, *267*, 126585. <https://doi.org/10.1016/j.energy.2022.126585>.
39. Liu, L.; Shen, X.; Chen, Z.; Sun, Q. Optimal Energy Management of Data Center Micro-Grid Considering Computing Workloads Shift. *IEEE Access* **2024**, *12*, 102061–102075. <https://doi.org/10.1109/ACCESS.2024.3432120>.
40. Wan, T.; Tao, Y.; Qiu, J.; Lai, S. Internet data centers participating in electricity network transition considering carbon-oriented demand response. *Appl. Energy* **2023**, *329*, 120305. <https://doi.org/10.1016/j.apenergy.2022.120305>.
41. Wang, K.; Ye, L.; Yang, S.; Deng, Z.; Song, J.; Li, Z.; Zhao, Y. A Hierarchical Dispatch Strategy of Hybrid Energy Storage System in Internet Data Center with Model Predictive Control. *Appl. Energy* **2023**, *331*, 120414. <https://doi.org/10.1016/j.apenergy.2022.120414>.
42. Flores-Martin, D.; Lemus-Prieto, F.; Rico-Gallego, J.A. PULSE: A Modular Framework for Predictive Energy Efficiency in Heterogeneous Data Centers. *SoftwareX* **2025**, *31*, 102313. <https://doi.org/10.1016/j.softx.2025.102313>.
43. Zhang, Y.; Liu, J. Prediction of Overall Energy Consumption of Data Centers in Different Locations. *Sensors* **2022**, *22*, 3704. <https://doi.org/10.3390/s22103704>.
44. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; OTexts: Melbourne, Australia, 2018. Available online: <https://otexts.com/fpp2/> (accessed on 12 May 2026).
45. Abdilla, A.; Borg, S.P.; Licari, J. Relating measured PUE to the cooling strategy and operating conditions through a review of a number of Maltese data centres. *Energy Rep.* **2025**, *13*, 2612–2623. <https://doi.org/10.1016/j.egyr.2025.02.013>.
46. Zhou, F.; Shen, C.; Ma, G.; Yan, X. Power Usage Effectiveness Analysis of a Liquid-Pump-Driven Hybrid Cooling System for Data Centers in Subclimate Zones. *Sustain. Energy Technol. Assess.* **2022**, *52*, 102277. <https://doi.org/10.1016/j.seta.2022.102277>.
47. Ismail, L.; Materwala, H. Computing Server Power Modeling in a Data Center: Survey, Taxonomy and Performance Evaluation. *Acm Comput. Surv.* **2020**, *53*, 1–34. <https://doi.org/10.1145/3390605>.

48. Yang, Z.; Wu, C.; Liu, G.; Zhang, J. Data Center Peak Electrical Demand Forecasting: A Multi-Feature SARIMA-LSTM Model. In Proceedings of the Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2023 ACM International Symposium on Wearable Computing, Cancun, Mexico, 8–12 October 2023; pp. 700–705. <https://doi.org/10.1145/3594739.3612910>.
49. Ranganathan, P.; Leech, P.; Irwin, D.; Chase, J. Ensemble-level Power Management for Dense Blade Servers. In Proceedings of the 33rd International Symposium on Computer Architecture (ISCA'06), Boston, MA, 17–21 June 2006; pp. 66–77. <https://doi.org/10.1109/ISCA.2006.20>.
50. Islam, M.A.; Mahmud, H.; Ren, S.; Wang, X. A Carbon-Aware Incentive Mechanism for Greening Colocation Data Centers. *IEEE Trans. Cloud Comput.* **2020**, *8*, 4–16. <https://doi.org/10.1109/TCC.2017.2767043>.
51. Racedo, S.; Jaumard, B.; Glatard, T.; Delgado, O.; Masoudi, M. Advances in Power Consumption Model for Data Centers: Analytical Formulas vs. Machine Learning Models. *Future Gener. Comput. Syst.* **2026**, *176*, 108141. <https://doi.org/10.1016/j.future.2025.108141>.
52. Lei, N.; Masanet, E. Climate- and Technology-Specific PUE and WUE Estimations for U.S. Data Centers Using a Hybrid Statistical and Thermodynamics-Based Approach. *Resour. Conserv. Recycl.* **2022**, *182*, 106323. <https://doi.org/10.1016/j.resconrec.2022.106323>.
53. Jin, Y.; Bai, X.; Xu, X.; Mi, R.; Li, Z. Climate Zones for the Application of Water-Side Economizer in a Data Center Cooling System. *Appl. Therm. Eng.* **2024**, *250*, 123450. <https://doi.org/10.1016/j.applthermaleng.2024.123450>.
54. Wang, N.; Guo, Y.; Huang, C.; Tian, B.; Shao, S. Multi-Scale Collaborative Modeling and Deep Learning-Based Thermal Prediction for Air-Cooled Data Centers: An Innovative Insight for Thermal Management. *Appl. Energy* **2025**, *377*, 124568. <https://doi.org/10.1016/j.apenergy.2024.124568>.
55. Gheni, M.; Kerskes, H.; Stergiaropoulos, K. Operational analysis of the cooling system in a direct liquid-cooled data center: A measurement and simulation study on the impact of supply water temperature. *Appl. Energy* **2026**, *403*, 127061. <https://doi.org/10.1016/j.apenergy.2025.127061>.
56. Azizi, A.; Morovati, S.; Zamani, A.; Piruzza, J.; Guo, D.; Liu, Z.; Taimela, P. Strengthening Data Center Operations Using Grid-Forming Battery Energy Storage as a Line-Interactive Uninterruptible Power Supply. *Int. J. Electr. Power Energy Syst.* **2026**, *175*, 111638. <https://doi.org/10.1016/j.ijepes.2026.111638>.
57. Zhu, Y.; Zhang, Q.; Huang, G.; Wang, J.; Zou, S.; Ee, Y.J.; Sopian, K. Research on Collaborative Control Strategy of Cold Storage and IT Workload Migration in Data Center. *Energy* **2025**, *323*, 135802. <https://doi.org/10.1016/j.energy.2025.135802>.
58. Siler-Evans, K.; Azevedo, I.L.; Morgan, M.G. Marginal Emissions Factors for the U.S. Electricity System. *Environ. Sci. Technol.* **2012**, *46*, 4742–4748. <https://doi.org/10.1021/es300145v>.
59. Lindberg, J.; Lesieutre, B.C.; Roald, L.A. Using Geographic Load Shifting to Reduce Carbon Emissions. *Electr. Power Syst. Res.* **2022**, *212*, 108586. <https://doi.org/10.1016/j.epsr.2022.108586>.
60. Khan, T.; Tian, W.; Ilager, S.; Buyya, R. Workload Forecasting and Energy State Estimation in Cloud Data Centres: ML-Centric Approach. *Future Gener. Comput. Syst.* **2022**, *128*, 320–332. <https://doi.org/10.1016/j.future.2021.10.019>.
61. Wang, L.; Chen, J.; Sui, F.; Zhang, Y.; Shi, D. Deployment-Efficient Short-Term Load Forecasting in AI Data Centers via Sequence-to-Point Knowledge Distillation. *arXiv* **2026**, arXiv:2605.00681.
62. 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. <https://doi.org/10.1016/j.applthermaleng.2022.119207>.
63. Hong, J.H.; Lee, G.C. Monthly Load Forecasting in a Region Experiencing Demand Growth: A Case Study of Texas. *Energies* **2025**, *18*, 4135. <https://doi.org/10.3390/en18154135>.
64. Zhou, Y.; Wei, F.; Li, S.; Wang, Z.; Liu, J.; Yu, D. Data Center Load Modeling Through Optimal Energy Consumption Characteristics: A Path to Simultaneously Enhance Energy Efficiency and Demand Response Quality. *Appl. Energy* **2025**, *393*, 126095. <https://doi.org/10.1016/j.apenergy.2025.126095>.
65. Xie, K.; Giacomoni, A.; Deng, X.; Wu, Y. Bias Calibration and Error Propagation Adjustment for ML-Based Time Series Forecasting: A Systematic Study for PJM's Electricity Load Forecast Amid Virginia's Data Center Surge. *Energy* **2025**, *336*, 138411. <https://doi.org/10.1016/j.energy.2025.138411>.
66. Nagpal, H.; Soppelsa, A.; Madsen, H.; Fedrizzi, R. An Adaptive Multi-Seasonal ARIMA Approach for Domestic Hot Water Load Forecasting: A Pilot Study. *Energy Build.* **2025**, *347*, 116226. <https://doi.org/10.1016/j.enbuild.2025.116226>.
67. Karamolegkos, S.; Koulouriotis, D.E. Advancing Short-Term Load Forecasting with Decomposed Fourier ARIMA: A Case Study on the Greek Energy Market. *Energy* **2025**, *325*, 135854. <https://doi.org/10.1016/j.energy.2025.135854>.
68. Hong, T.; Pinson, P.; Fan, S.; Zareipour, H.; Troccoli, A.; Hyndman, R.J. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *Int. J. Forecast.* **2016**, *32*, 896–913. <https://doi.org/10.1016/j.ijforecast.2016.02.001>.
69. Misiurek, K.; Olkusi, T.; Zyśk, J. Review of Methods and Models for Forecasting Electricity Consumption. *Energies* **2025**, *18*, 4032. <https://doi.org/10.3390/en18154032>.

70. Yildirim, E.; Hussein, M.; Titov, M.; Kilic, O.O. Predicting Runtime and Resource Utilization of Jobs on Integrated Cloud and HPC Systems. *Future Gener. Comput. Syst.* **2026**, *176*, 108230. <https://doi.org/10.1016/j.future.2025.108230>.
71. Shao, S.; Ding, X.; Zhao, B.; Ye, P. Attention-Based Workload Prediction and Dynamic Resource Allocation for Heterogeneous Computing Environments. *Sci. Rep.* **2026**, *16*, 8571. <https://doi.org/10.1038/s41598-026-38622-4>.
72. Yuan, H.; Zhang, Z.; Yang, D.; Xue, T.; Wen, D.; Yao, G. An Intelligent Thermal Management Strategy for a Data Center Prototype Based on Digital Twin Technology. *Appl. Sci.* **2025**, *15*, 7675. <https://doi.org/10.3390/app15147675>.
73. Chen, X.; Tu, R.; Li, M.; Yang, X.; Jia, K. Hot Spot Temperature Prediction and Operating Parameter Estimation of Racks in Data Center Using Machine Learning Algorithms Based on Simulation Data. *Build. Simul.* **2023**, *16*, 2159–2176. <https://doi.org/10.1007/s12273-023-1022-4>.
74. Zhao, X.; Ma, Z.G.; Jørgensen, B.N. An End-to-End Data and Machine Learning Pipeline for Energy Forecasting: A Systematic Approach Integrating MLOps and Domain Expertise. *Information* **2025**, *16*, 805. <https://doi.org/10.3390/info16090805>.
75. Rabie, A.H.; Saleh, A.I.; Elkhaliq, S.H.A.; Takieldein, A.E. An Optimum Load Forecasting Strategy (OLFS) for Smart Grids Based on Artificial Intelligence. *Technologies* **2024**, *12*, 19. <https://doi.org/10.3390/technologies12020019>.
76. Wang, Z.; Xia, L.; Yuan, H.; Srinivasan, R.S.; Song, X. Principles, Research Status, and Prospects of Feature Engineering for Data-Driven Building Energy Prediction: A Comprehensive Review. *J. Build. Eng.* **2022**, *58*, 105028. <https://doi.org/10.1016/j.jobbe.2022.105028>.
77. Baur, L.; Ditschuneit, K.; Schambach, M.; Kaymakci, C.; Wollmann, T.; Sauer, A. Explainability and Interpretability in Electric Load Forecasting Using Machine Learning Techniques: A Review. *Energy AI* **2024**, *16*, 100358. <https://doi.org/10.1016/j.egyai.2024.100358>.
78. Arabzadeh, V.; Frank, R. A Four-Dimensional Analysis of Explainable AI in Energy Forecasting: A Domain-Specific Systematic Review. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 153. <https://doi.org/10.3390/make7040153>.
79. Zhao, F.; Lin, W.; Lin, S.; Zhong, H.; Li, K. TFEGRU: Time-Frequency Enhanced Gated Recurrent Unit With Attention for Cloud Workload Prediction. *IEEE Trans. Serv. Comput.* **2025**, *18*, 467–478. <https://doi.org/10.1109/TSC.2024.3517324>.
80. Shekhar, S.; Gauchia, L.; Amaris, H.; Pérez-Borondo, Á.; Hernández, C. Effect of Variable Temperatures on Machine Learning Battery SoH Estimation for Auxiliary Aircraft Batteries. *J. Power Sources* **2025**, *660*, 238451. <https://doi.org/10.1016/j.jpowsour.2025.238451>.
81. Zhao, P.; Hu, W.; Cao, D.; Dong, Z.; Zeng, Y.; Huang, Q.; Chen, Z. A Review on Deep Learning-Based Electrical Load Forecasting: From Perspectives of Learning Paradigms and Foundation Models. *J. Mod. Power Syst. Clean Energy* **2026**, *14*, 791–809. <https://doi.org/10.35833/MPCE.2025.000740>.
82. Mughees, M.; Li, Y.; Chen, Y.; Li, Y.R. Short-Term Load Forecasting for AI-Data Center. In Proceedings of the 2025 IEEE Power & Energy Society General Meeting, Austin, TX, USA, 27–31 July 2025; pp. 1–5. <https://doi.org/10.1109/PESGM52009.2025.11225317>.
83. Gassar, A.A.A. Short-Term Energy Forecasting to Improve the Estimation of Demand Response Baselines in Residential Neighborhoods: Deep Learning vs. Machine Learning. *Buildings* **2024**, *14*, 2242. <https://doi.org/10.3390/buildings14072242>.
84. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*; PMLR : New York, NY 19–24 June 2016; Volume 48, pp. 1050–1059. https://proceedings.mlr.press/v48/gal16.html?trk=public_post_comment-text.
85. Takci, M.T.; Qadrdan, M.; Summers, J.; Gustafsson, J. Data Centres as a Source of Flexibility for Power Systems. *Energy Rep.* **2025**, *13*, 3661–3671. <https://doi.org/10.1016/j.egyri.2025.03.020>.
86. Zhao, Z.; Liang, C.; Zhang, Y.; Zhu, L.; Wang, X.; Tang, J. Short-Term Data Center Load Forecasting Method Based on Data-Physics Hybrid Driven Framework. *Energy Rep.* **2026**, *15*, 108953. <https://doi.org/10.1016/j.egyri.2025.108953>.
87. Bahadori-Jahromi, A.; Room, S.; Paknahad, C.; Altekreeti, M.; Tariq, Z.; Tahayori, H. The Role of Artificial Intelligence and Machine Learning in Advancing Civil Engineering: A Comprehensive Review. *Appl. Sci.* **2025**, *15*, 10499. <https://doi.org/10.3390/app151910499>.
88. 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. <https://doi.org/10.1109/JPROC.2011.2161244>.
89. Cao, Z.; Wang, R.; Zhou, X.; Tan, R.; Wen, Y.; Yan, Y.; Wang, Z. Adaptive Capacity Provisioning for Carbon-Aware Data Centers: A Digital Twin-Based Approach. *IEEE Trans. Sustain. Comput.* **2025**, *10*, 730–743. <https://doi.org/10.1109/TSUSC.2025.3526192>.
90. Saad, A.; Faddel, S.; Mohammed, O. IoT-Based Digital Twin for Energy Cyber-Physical Systems: Design and Implementation. *Energies* **2020**, *13*, 4762. <https://doi.org/10.3390/en13184762>.
91. Hashemi, Z.; Gholampour, M.; Su, H.H.; Chiang, C.Y.; Chang, H.H.; Wang, C.C. Hybrid Physics-Informed Neural Networks for Airflow and Thermal Modeling in Data Center Environments. *Int. Commun. Heat Mass Transf.* **2025**, *168*, 109511. <https://doi.org/10.1016/j.icheatmasstransfer.2025.109511>.

92. Wang, R.; Zhang, Q.; Wen, Y.; Kennedy, S. Toward Multiphysics-Informed Machine Learning for Sustainable Data Center Operations: Intelligence Evolution with Deployable Solutions for Computing Infrastructure. *arXiv* **2025**, arXiv:2505.19414. <https://doi.org/10.48550/arXiv.2505.19414>.
93. Guo, C.; Luo, F.; Cai, Z.; Dong, Z.Y. Integrated energy systems of data centers and smart grids: State-of-the-art and future opportunities. *Appl. Energy* **2021**, *301*, 117474. <https://doi.org/10.1016/j.apenergy.2021.117474>.
94. Zhang, Y.; Tsiligkaridis, A.; Paschalidis, I.C.; Coskun, A.K. Data Center and Load Aggregator Coordination Towards Electricity Demand Response. *Sustain. Comput. Inform. Syst.* **2024**, *42*, 100957. <https://doi.org/10.1016/j.suscom.2024.100957>.
95. Zhang, H.; Xu, T.; Wu, H.; Liu, B.; Faqiry, M.N. Risk-Based Stochastic Day-Ahead Operation for Data Centre Virtual Power Plants. *IET Renew. Power Gener.* **2019**, *13*, 1665–1677. <https://doi.org/10.1049/iet-rpg.2018.5736>.
96. Pakbaznia, E.; Pedram, M. Minimizing Data Center Cooling and Server Power Costs. In Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design, San Francisco, CA, USA, 19 August 2009; pp. 145–150. <https://doi.org/10.1145/1594233.1594268>.
97. Mohsenian-Rad, A.H.; Wong, V.W.S.; Jatskevich, J.; Schober, R.; Leon-Garcia, A. Autonomous Demand-Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid. *IEEE Trans. Smart Grid* **2010**, *1*, 320–331. <https://doi.org/10.1109/TSG.2010.2089069>.
98. Sheikholeslami, S.M.; Rabiei, A.M.; Mohammad-Taheri, M.; Abouei, J. Cloud data center participation in smart demand response programs for energy cost minimisation. *IET Smart Grid* **2022**, *5*, 380–394. <https://doi.org/10.1049/stg2.12082>.
99. Zhang, W.; Zavala, V.M. Remunerating space–time, load-shifting flexibility from data centers in electricity markets. *Appl. Energy* **2022**, *326*, 119930. <https://doi.org/10.1016/j.apenergy.2022.119930>.
100. Liu, N.; Dong, Z.; Rojas-Cessa, R. Task Scheduling and Server Provisioning for Energy-Efficient Cloud-Computing Data Centers. In Proceedings of the 2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops, Philadelphia, PA, USA, 8–11 July 2013; pp. 226–231. <https://doi.org/10.1109/ICDCSW.2013.68>.
101. Liu, W.; Yan, Y.; Sun, Y.; Mao, H.; Cheng, M.; Wang, P.; Ding, Z. Online job scheduling scheme for low-carbon data center operation: An information and energy nexus perspective. *Appl. Energy* **2023**, *338*, 120918. <https://doi.org/10.1016/j.apenergy.2023.120918>.
102. Wang, F.; Lv, C.; Xu, J. Carbon awareness oriented data center location and configuration: An integrated optimization method. *Energy* **2023**, *278*, 127744. <https://doi.org/10.1016/j.energy.2023.127744>.
103. Deng, Q.; Meisner, D.; Ramos, L.; Wenisch, T.F.; Bianchini, R. MemScale: Active Low-Power Modes for Main Memory. In Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems, New York, NY, USA, 5 March 2011; pp. 225–238. <https://doi.org/10.1145/1950365.1950392>.
104. Lechowicz, A.; Christianson, N.; Zuo, J.; Bashir, N.; Hajiesmaili, M.; Wierman, A.; Shenoy, P. The Online Pause and Resume Problem: Optimal Algorithms and an Application to Carbon-Aware Load Shifting. *Proc. ACM Meas. Anal. Comput. Syst.* **2023**, *7*, 1–32. <https://doi.org/10.1145/3626776>.
105. Hawkes, A.D. Estimating Marginal CO₂ Emissions Rates for National Electricity Systems. *Energy Policy* **2010**, *38*, 5977–5987. <https://doi.org/10.1016/j.enpol.2010.05.053>.
106. Riepin, I.; Brown, T.; Zavala, V.M. Spatio-temporal Load Shifting for Truly Clean Computing. *Adv. Appl. Energy* **2025**, *17*, 100202. <https://doi.org/10.1016/j.adapen.2024.100202>.
107. Chen, D.; Chui, C.K.; Lee, P.S. Physics informed machine learning based predictive control for intelligent operation of edge datacenters. *Appl. Energy* **2026**, *402*, 126975. <https://doi.org/10.1016/j.apenergy.2025.126975>.
108. Rawlings, J.B.; Mayne, D.Q.; Diehl, M.M. *Model Predictive Control: Theory, Computation, and Design*, 2nd ed.; Nob Hill Publishing: Madison, WI, USA, 2017.
109. Wang, J.; Deng, W.; Yue, C.; Ding, W.; Zeng, L.; Su, W. MPC-Based Joint Optimization for Rack-Based Cooling Data Centers: Modeling, Performance Evaluation, and Time-Delay Characteristic Analysis. *Build. Environ.* **2025**, *286*, 113695. <https://doi.org/10.1016/j.buildenv.2025.113695>.
110. Crozier, C.; Liska, M. The potential of data center energy demand to provide grid flexibility. *Curr. Sustain. Energy Rep.* **2025**, *12*, 12. <https://doi.org/10.1007/s40518-025-00258-9>.
111. Artioli, M.; Borghesi, A.; Chinnici, M.; Ciampolini, A.; Colonna, M.; De Chiara, D.; Loreti, D. C6EnPLS: A High-Performance Computing Job Dataset for the Analysis of Linear Solvers' Power Consumption. *Future Internet* **2025**, *17*, 203. <https://doi.org/10.3390/fi17050203>.
112. Jiang, H.; Dong, Y.; Dong, Y.; Wang, J. Power Load Forecasting Based on Spatial–Temporal Fusion Graph Convolution Network. *Technol. Forecast. Soc. Change* **2024**, *204*, 123435. <https://doi.org/10.1016/j.techfore.2024.123435>.
113. Shamim, M.M.I.; Hamid, A.B.A.; Nyamasvisva, T.E.; Rafi, N.S.B. Advancement of Artificial Intelligence in Cost Estimation for Project Management Success: A Systematic Review of Machine Learning, Deep Learning, Regression, and Hybrid Models. *Modelling* **2025**, *6*, 35. <https://doi.org/10.3390/modelling6020035>.
114. Gou, T.; Xu, Y.; Sun, H. Carbon-aware day-ahead optimal dispatch for integrated power grid thermal systems with aggregated distributed resources. *Appl. Energy* **2025**, *389*, 125715. <https://doi.org/10.1016/j.apenergy.2025.125715>.

115. Chen, Z.; Xiao, F.; Guo, F.; Yan, J. Interpretable Machine Learning for Building Energy Management: A State-of-the-Art Review. *Adv. Appl. Energy* **2023**, *9*, 100123. <https://doi.org/10.1016/j.adapen.2023.100123>.
116. Ye, Z.; Gao, W.; Hu, Q.; Sun, P.; Wang, X.; Luo, Y.; Zhang, T.; Wen, Y. Deep Learning Workload Scheduling in GPU Datacenters: A Survey. *ACM Comput. Surv.* **2024**, *56*, 1–38. <https://doi.org/10.1145/3638757>.
117. Qin, D.; Wu, X.; Sun, D.; Liang, Z.; Zhang, N. Load Forecasting under Distribution Shift: An Online Quantile Ensembling Approach. *Appl. Energy* **2025**, *401*, 126812. <https://doi.org/10.1016/j.apenergy.2025.126812>.
118. Luccioni, A.S.; Jernite, Y.; Strubell, E. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 85–99. <https://doi.org/10.1145/3630106.3658542>.
119. Adamantiadou, D.S.; Tsironis, L. Leveraging Artificial Intelligence in Project Management: A Systematic Review of Applications, Challenges, and Future Directions. *Computers* **2025**, *14*, 66. <https://doi.org/10.3390/computers14020066>.
120. Zhao, J.; Xu, D.; Yuan, X.; Liu, D. Cooling Supply Allocation Model Predictive Control Strategy of Data Center Computer Room Air Conditioning Systems Considering Multi-Zone Thermal Coupling Effects. *Appl. Therm. Eng.* **2025**, *277*, 127103. <https://doi.org/10.1016/j.applthermaleng.2025.127103>.
121. Lv, X.; Basem, A.; Hasani, M.; Sun, P.; Zhang, J. The Potential of Combined Robust Model Predictive Control and Deep Learning in Enhancing Control Performance and Adaptability in Energy Systems. *Sci. Rep.* **2025**, *15*, 11187. <https://doi.org/10.1038/s41598-025-95636-0>.
122. Kumar, A.; Pindoriya, N.M. Toward Sustainable Data Center Operation: A Review on Existing Infrastructures, Integrated Smart Energy Management Frameworks, and Future Perspectives. *Renew. Sustain. Energy Rev.* **2026**, *230*, 116664. <https://doi.org/10.1016/j.rser.2025.116664>.
123. Knittel, C.R.; Senga, J.R.L.; Wang, S. *Flexible Data Centers and the Grid: Lower Costs, Higher Emissions?* Technical Report 34065; National Bureau of Economic Research: Cambridge, MA, USA, 2025. <https://doi.org/10.3386/w34065>.
124. Shadi, M.R.; Mirshekali, H.; Shaker, H.R. Explainable Artificial Intelligence for Energy Systems Maintenance: A Review on Concepts, Current Techniques, Challenges, and Prospects. *Renew. Sustain. Energy Rev.* **2025**, *218*, 115668. <https://doi.org/10.1016/j.rser.2025.115668>.
125. Li, K.; Qin, Z.; Mu, Y.; Wang, H.; Bie, Q.; Yin, X.; Yan, Y. Transfer Learning-Based Multi-Energy Load Forecasting Method for Integrated Energy System with Zero-Shot. *Appl. Energy* **2025**, *401*, 126769. <https://doi.org/10.1016/j.apenergy.2025.126769>.
126. Von Krannichfeldt, L.; Orehounig, K.; Fink, O. Combining Physics-Based and Data-Driven Modeling for Building Energy Systems. *Appl. Energy* **2025**, *391*, 125853. <https://doi.org/10.1016/j.apenergy.2025.125853>.
127. Liu, P.; Liang, Y.; Li, S.; Li, Y.; Sheng, W. Domain adaptation-enhanced transfer learning framework for cross-building cooling load forecasting: Case studies in metro stations. *Energy Build.* **2026**, 117059. <https://doi.org/10.1016/j.enbuild.2026.117059>.
128. Luo, C.; Xia, L.; Hong, S.H. A Method for Selecting the Appropriate Source Domain Buildings for Building Energy Prediction in Transfer Learning: Using the Euclidean Distance and Pearson Coefficient. *Energies* **2025**, *18*, 3706. <https://doi.org/10.3390/en18143706>.
129. Ginzburg-Ganz, E.; Lifshits, P.; Machlev, R.; Belikov, J.; Krieger, Z.; Levron, Y. Technical Challenges of AI Data Center Integration into Power Grids—A Survey. *Energies* **2026**, *19*, 137. <https://doi.org/10.3390/en19010137>.
130. Chawda, G.S.; Shekhar, S.; Hollweg, G.V.; Su, W.; Wang, M.; Ndoye, M.; Oni, B. Artificial Intelligence in Bidirectional Energy Management: A Critical Review of Control Strategies for V2G and G2V Systems. *IET Smart Grid* **2026**, *9*, e70093. <https://doi.org/10.1049/stg2.70093>.
131. Zou, J.; Liu, S.; Ouyang, L.; Ruan, J.; Tang, S. Carbon-aware demand response for residential smart buildings. *Electronics* **2024**, *13*, 4941. <https://doi.org/10.3390/electronics13244941>.
132. Chi, C.; Ji, K.; Song, P.; Marahatta, A.; Zhang, S.; Zhang, F.; Qiu, D.; Liu, Z. Cooperatively Improving Data Center Energy Efficiency Based on Multi-Agent Deep Reinforcement Learning. *Energies* **2021**, *14*, 2071. <https://doi.org/10.3390/en14082071>.

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.