

Grid-Interactive AI Data Centres for Renewable Smart Grids: A Systematic Review

Larbi Kwadwo*

Department of Renewable Engineering, University of Mines and Technology (UMaT), Tarkwa, Ghana

*Corresponding Author: Larbi Kwadwo, larbikwadwo123@gmail.com

ORCID: 0000-0002-1859-3145

Article History

Received: 06 August 2026

Revised: 01 September 2026

Accepted: 07 September 2026

Available Online: 28 October 2026

Abstract. With the increasing number of artificial intelligence workloads being run in hyperscale and cloud data centres, there is a growing demand for electricity in power systems with high shares of variable renewable energy. This systematic review analyzes how grid-interactive AI data centres can help renewable smart grids with flexible computing and energy-management strategies. There has been a search of studies published between the years 2022 and August 2026 on Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Google Scholar. A search was done with various terms to get better relevancy. Out of the total number of studies, 43 studies were finally included. There are four evidentiary themes relating to architectures (enabling), coordination (renewable-energy), performance (multi-dimensional) and barriers (implementation). Insights reveal that the shifting workload, relocation, battery-backed UPS, thermal control enables an increase in renewable-energy use, peak demand decrease, operating costs decline, and supports grid services. Nevertheless, the outcomes are heavily dependent upon the forecast accuracy, workload flexibility, battery degradation, converter inertia, service level requirements and electricity market conditions. There are times when maximising economic savings does not lead to maximising carbon reductions and therefore we need to value costs, carbon intensity, grid reliability and computing joint policy evaluation. Simulation-based studies and limited field verification are the dominant evidence at present. It is therefore essential to develop greater standardisation, lifecycle carbon assessment, uncertainty-aware control and operational demonstrations for reliable deployment.

Key words. Artificial intelligence, Data centres, Demand response, Renewable energy, Smart grids, Workload flexibility

1. Introduction

The rapid growth of artificial intelligence (AI), cloud computing and high-performance computing is transforming data centres from conventional digital infrastructure into participants in electricity systems (Figure 1). AI-oriented facilities contain dense clusters of graphics processing units, high-capacity cooling systems and power-conditioning equipment that create large, variable electrical loads. Rather than treating these facilities solely as inflexible consumers, the grid-interactive data-centre concept positions them as controllable energy resources capable of coordinating computing demand with renewable generation and grid conditions. Paananen [1] explains that grid-interactive uninterruptible power supplies can provide rapid frequency services, while Zhou et al. [2] demonstrate that GPU-intensive AI facilities may offer substantial power flexibility at lower opportunity cost than conventional high-performance computing centres. This establishes a research frontier connecting smart grids, renewable energy, sustainable computing and AI infrastructure.

This transition is especially relevant to renewable-dominant grids, where solar and wind generation vary across time and location. Data centres possess flexibility that can help accommodate this variability. Deferrable training and batch-processing jobs can be shifted to periods of abundant renewable electricity, geographically distributed workloads can be migrated toward cleaner regions, and noncritical computing demand can be curtailed during grid emergencies. Radovanović et al. [3] showed that carbon-aware capacity controls can delay flexible workloads toward periods with lower marginal carbon intensity. Zhao and Zhou [4] similarly developed energy- and carbon-aware placement methods that increase renewable-energy utilisation across distributed

cloud facilities, while Piontek et al. [5] demonstrated carbon-responsive orchestration for containerised workloads. These approaches suggest that software scheduling can provide virtual energy storage by shifting electricity consumption without relying entirely on physical batteries.

Physical infrastructure provides complementary flexibility. Data centres commonly operate battery-backed uninterruptible power supplies, cooling systems and backup equipment that can be coordinated with on-site renewables and grid electricity. Acun et al. [6] argue that renewable procurement, battery deployment and workload scheduling should be evaluated jointly because operational carbon reductions can introduce embodied-carbon and investment trade-offs. Zhang et al. [7] show that battery and cold-storage capacity can support flexibility services while reducing operating costs, whereas Nkwawir et al. [8] extend this perspective through multi-energy integration linking computing, electrical and thermal resources. The grid-interactive model is therefore a cyberphysical architecture integrating workload control, storage, cooling, power electronics, forecasting and electricity-market participation.

Despite rapid progress, the literature remains fragmented across computing, power systems, thermal engineering and sustainability. Existing reviews generally examine individual components rather than the full AI-grid interaction. Verma et al. [9] survey energy-efficient workflow scheduling in cloud environments, while Yuan et al. [10] and Hao et al. [11] focus on data-centre waste-heat recovery and thermal utilisation. Crozier and Liska [12] discuss the grid-flexibility potential of data-centre demand, and Takci et al. [13] review data centres as flexibility resources for power systems. However, they do not collectively integrate AI-specific workload characteristics, computational scheduling, electrical storage, converter behaviour, thermal flexibility, renewable coordination, carbon accounting and grid-service participation within one systematic framework. They also leave limited distinction between direct evidence from AI- or GPU-intensive facilities and findings transferred from conventional cloud, HPC or general data-centre studies.

This review addresses that gap through cross-domain synthesis and by distinguishing AI-specific evidence from

transferable evidence derived from broader data-centre contexts. It integrates computational, electrical, thermal and spatial flexibility within one taxonomy; compares methods used to coordinate these resources; and evaluates outcomes across grid reliability, renewable-energy utilisation, operating cost, carbon emissions and computing quality of service. It also considers the strength and validation context of the evidence, recognising that simulation, optimisation, experimental and field studies provide different levels of operational support. This distinction is important because flexibility is not automatically equivalent to decarbonisation. Senga et al. [14] show that flexible data centres can reduce power-system costs while increasing emissions under some generation mixes, while Chen and Zheng [15] demonstrate that temporal deferral and geographical migration produce location-dependent grid-planning benefits. Williams et al. [16] therefore emphasise architectures combining grid signals, power telemetry and performance-aware orchestration.

Accordingly, this paper makes four contributions:

1. Develops an integrated taxonomy of computational, electrical, thermal and spatial flexibility in AI data centres.
2. Compares forecasting, optimisation, machine-learning and workload-orchestration techniques used for renewable-aware operation.
3. Synthesises evidence on demand response, ancillary services, storage utilisation, renewable integration, emissions, operating cost and computing performance.
4. Identifies technical, market, regulatory, cybersecurity, measurement and validation gaps and proposes priorities for grid-interactive deployment.

The review aims to characterise enabling architectures, evaluate computational, storage and thermal coordination strategies, compare multidimensional grid and computing outcomes, and identify unresolved methodological and implementation challenges. The research also clarifies when, in this respect, the flexibility of an AI data centre can help the renewable smart grid and where more operational validation is needed.

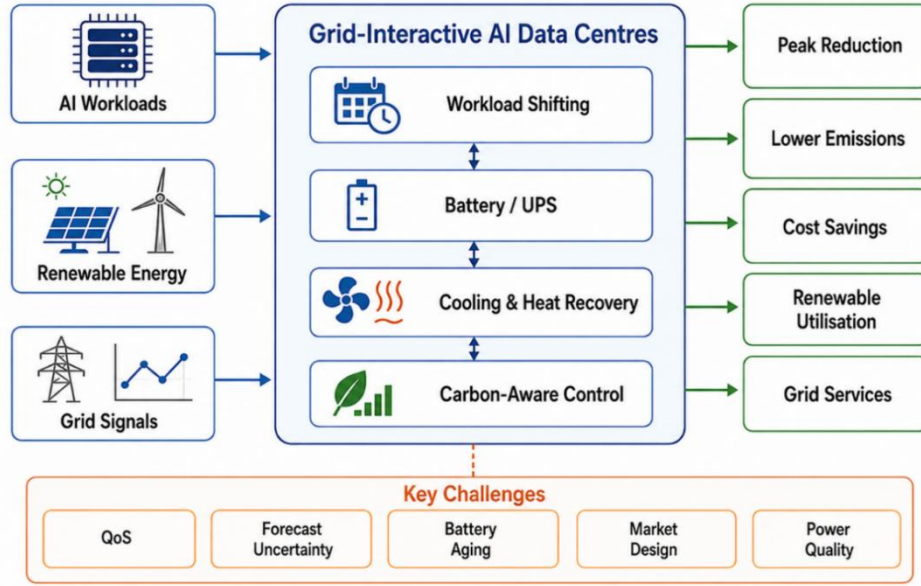


Figure 1. Conceptual overview of grid-interactive AI data centres in renewable smart grids.

2. Related Studies

Studies regarding grid-interactive “artificial intelligence” data centres have started to treat their computing infrastructures as a cyberphysical energy resource. This means these infrastructures are not passive electrical loads drawing power from the grid as most data centres, but an active component of the energy resource system. Covering the spectrum of demand response and ancillary services, electrical and thermal flexibility, intelligent control, power-quality management and more. The evidence, however, is heterogeneous in relation to AI. While some studies explore explicitly AI or GPU intensive facilities, others look into high-performance computing (HPC), geographically distributed cloud systems, edge-cloud settings or traditional data centres. This distinction is important because findings from general cloud or HPC environments may inform AI data-centre operation but cannot automatically be interpreted as direct evidence for contemporary GPU-dense AI facilities. The present review therefore distinguishes direct AI/GPU-specific evidence from transferable evidence generated in broader computing environments.

2.1 Comparison with Existing Reviews

Previous reviews have addressed important components of sustainable data-centre operation but generally from narrower disciplinary perspectives. Verma et al. [9] reviewed energy-efficient workflow-scheduling algorithms in cloud computing, concentrating on computational resource management and optimisation rather than grid-service participation or the electrical characteristics of AI facilities. Yuan et al. [10] reviewed waste-heat recovery pathways, including district heating, cooling and other forms of thermal reuse, while Hao et al. [11] compared contemporary waste-heat recovery technologies and their economic and thermodynamic suitability. These studies provide detailed thermal evidence but do not integrate workload flexibility,

electricity-market participation, battery coordination and AI-specific computing constraints within the same analytical framework. Power-system-oriented reviews provide a complementary perspective. Crozier and Liska [12] examined the potential of data-centre electricity demand to provide grid flexibility through mechanisms such as workload management, processor control and cooling flexibility. Takci et al. [13] similarly considered data centres as flexibility resources capable of supporting peak reduction, congestion management and renewable-energy integration. Although these reviews establish the relevance of data centres to future power systems, they provide less emphasis on the distinctive characteristics of AI workloads, GPU-dense infrastructure, workload migration, carbon-aware orchestration and the interaction between computational service requirements and electrical flexibility. The present review extends this literature by integrating computational, electrical, thermal and spatial flexibility within one taxonomy and by evaluating grid reliability, renewable utilisation, operating cost, carbon performance and computing quality of service together. It also distinguishes evidence obtained directly from AI / GPU environments from findings transferred from HPC, cloud and conventional data-centre studies.

2.2 Renewable-Aware and Carbon-Aware Workload Orchestration

Renewable-aware workload orchestration seeks to align computing activity with the temporal and geographical availability of low-carbon electricity. Evidence in this area is dominated by general cloud and edge-cloud environments rather than dedicated AI facilities. Miao et al. [17], for example, scheduled distributed machine-learning tasks across a multi-renewable edge-cloud continuum, providing relatively direct evidence for AI-related workloads while retaining an edge-cloud system context. Khodayarsesht et al. [18] examined energy- and carbon-aware virtual-machine placement across

geographically distributed cloud data centres, whereas Bahreini et al. [19] developed a carbon-aware cloud workload dispatcher. These studies are therefore best interpreted as transferable cloud evidence rather than direct validation in GPU-intensive AI facilities. Broader carbon-aware research strengthens this foundation. Asadov et al. [20] combined temporal deferral, geographical migration, embodied-carbon considerations and workload forecasting, while Hewage et al. [21] showed that spatio-temporal shifting must account for deadlines, resource availability and data-transfer emissions. Hussain et al. [22] incorporated electricity-price variation and workflow dependencies, demonstrating that low-cost and low-carbon schedules may favour different locations or execution periods. Collectively, this literature establishes workload flexibility as a major grid-interaction mechanism but also reveals limitations in renewable forecasting, migration overhead, common benchmarks and AI-specific operational validation. Evidence from cloud and edge-cloud studies is therefore informative for AI data centres, but its transferability depends on whether GPU utilisation, training-job duration, communication intensity and service-level constraints are adequately represented.

2.3 Demand Response and Grid-Service Coordination

A second research stream models data centres as controllable demand-response resources. Much of this evidence originates from geographically distributed data-centre models rather than AI-specific facilities. Zhao et al. [23] coordinated workloads, local energy resources and demand-response incentives across distributed data centres, showing potential reductions in operating cost and peak demand. Zhang et al. [24] linked data centres with load aggregators, while Zhang et al. [25] integrated traffic, data and power networks to reduce grid stress associated with proactive workload migration. S. Zhou et al. [26] similarly balanced electricity expenditure and distributed workload allocation. These studies demonstrate the technical potential of flexible computing demand, although most rely on modelling assumptions rather than operational AI data-centre deployment. Storage and fast-response studies extend flexibility toward shorter grid-service timescales. Ye et al. [27] showed that uninterruptible-power-supply losses materially affect the value of joint workload and storage scheduling, while Y. Zhou et al. [28] modelled optimal energy-consumption characteristics for improving both energy efficiency and demand-response quality. Tao and Gadh [29] further examined fast frequency response through coordinated workload modulation and UPS operation. More direct AI-specific evidence is provided by Zhou et al. [2], who focused on AI-oriented HPC data centres and reported that their workload characteristics could provide considerable grid flexibility at comparatively low opportunity cost. This evidence is particularly important because it directly connects AI-intensive computing with grid services rather than inferring performance from conventional data-centre models. Nevertheless, uncertainty remains regarding baseline measurement, rebound demand, battery degradation, market remuneration and the ability to

sustain grid-service commitments under real AI workloads.

2.4 Energy Storage, Cooling and Waste-Heat Integration

Physical flexibility arises from batteries, cooling systems, thermal storage and recoverable heat. Most evidence in this area comes from general data-centre infrastructure and is therefore transferable rather than AI-specific. Xu et al. [30] optimised a multi-energy portfolio for zero-carbon data-centre buildings under proactive demand-response programmes, demonstrating the value of jointly sizing and dispatching electrical and thermal resources. Yuan et al. [10] found that waste-heat utilisation depends strongly on outlet temperature, heat-pump configuration, transport distance and nearby heat demand. Lin et al. [31] proposed an integrated carbon-dioxide heat pump, mechanical-subcooling and absorption-refrigeration system to increase the usefulness of low-grade server heat. Experimental and dynamic-system studies provide stronger physical validation. Chen et al. [32] experimentally demonstrated energy-saving potential from integrated cooling and waste-heat recovery, while Cui et al. [33] investigated adsorption cooling driven by data-centre waste heat. F. Zhou et al. [34] analysed the dynamic behaviour of a heat-pump system integrating free cooling and heat recovery, and Hao et al. [11] compared the suitability of different recovery technologies. Although these studies are not specifically centred on AI workloads, their infrastructure-level findings are relevant because GPU-dense AI facilities produce substantial cooling demand and thermal output. Their direct transfer to AI environments nevertheless requires consideration of higher rack densities, changing workload profiles and tighter thermal-management requirements.

2.5 Intelligent Control, Measurement and Power Quality

The final research stream concerns intelligent control, measurement, uncertainty and electrical compatibility. Dinesh Reddy et al. [35] applied imitation-based optimisation to data-centre resource management while considering migration overhead and service-level violations. This represents general data-centre evidence. Shim [36], by contrast, used measured HPC traces and a multilayer perceptron to estimate job-level emissions, providing evidence from an HPC environment that is more closely related to AI-scale computing but still distinct from direct GPU-intensive AI data-centre validation. Ruparel [37] combined carbon-intensity forecasting, battery scheduling and distributionally robust optimisation, highlighting the need for control strategies that remain effective when workload, renewable output and carbon intensity deviate from forecasts. At the electrical interface, Sun, Xu, Cespedes, and Kauffman [38] developed impedance models for data-centre power supplies, while Sun, Cespedes, Wong, and Kauffman [39] extended this work to system-level stability. These studies show that aggregated converter-controlled loads cannot always be represented as conventional static demand. Sutaria and Rönnberg [40] further identified

supraharmonic propagation under different UPS operating modes, demonstrating that grid-service participation can create power-quality considerations within converter-dominated facilities. Chen et al. [41] addressed online carbon-aware backup and job scheduling in geographically distributed analytics under long-term emission constraints.

3. Methodology

This systematic review applied a structured approach to identify, evaluate and synthesise evidence on grid-interactive artificial intelligence (AI) data centres within renewable smart-grid environments. The review covered computational flexibility, renewable-aware workload scheduling, demand response, battery-supported uninterruptible power supplies (UPSs), thermal management, waste-heat recovery, ancillary services, carbon-aware operation and power-quality considerations. Reporting was organised according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) framework to improve transparency across literature identification, screening, eligibility assessment and study inclusion. The final evidence base consisted of 43 studies published between 2022 and 2026. Of these, 4 studies were published in 2022, 11 in 2023, 9 in 2024, 14 in 2025 and 5 in 2026. Because the literature spans computer science, power systems, energy engineering, thermal engineering and sustainable computing, the review accommodated different study designs while distinguishing methodological quality from the strength of operational evidence.

3.1 Review Design

A systematic literature review design was adopted because research on grid-interactive data centres is distributed across disciplines that use different workloads, datasets, electricity-system models, experimental environments and outcome measures. The review addressed four objectives: identifying architectures and enabling technologies for grid-interactive AI data centres; examining computational, electrical-storage and thermal strategies for coordinating demand with renewable generation; comparing effects on grid reliability, carbon emissions, operating cost, renewable-energy utilisation and computing quality of service; and identifying methodological, technical, regulatory and implementation barriers.

A review protocol was developed around the research scope, information sources, search concepts, eligibility criteria, screening process, extraction variables, methodological appraisal and synthesis procedure. The protocol was not registered in an external systematic-review registry, and this is acknowledged as a methodological limitation rather than implying retrospective registration. Quantitative meta-analysis was not undertaken because the included studies differed substantially in system boundaries, optimisation objectives, datasets, electricity markets, temporal resolution, baseline assumptions and outcome units. A structured narrative and comparative synthesis was

therefore considered more appropriate. The review also recognised that evidence from different research designs should not be interpreted equivalently. We assessed studies on simulation and optimization for modeling rigor and realistic constraints, whereas in addition to hardware implementation, measurement procedures and operational validation, also assessed experimental and field studies. When assessing quantitative outcomes, review studies were largely used to provide context to wider research developments rather than as stand-alone primary observations.

3.2 Information Sources and Search Strategy

The identification of literature was done using seven sources of information. These were Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink and Google Scholar. The chosen resources encompass a wide range of subjects including computing, artificial intelligence, cloud systems, electrical engineering, renewable energy, smart grids, thermal systems, sustainability and interdisciplinary research. A publication period ranging from January 2022 – August 2026 was specified. The 2022 start date was chosen because the review focuses on the recent convergence of AI-intensive computing, GPU-dense infrastructure, carbon-aware workload management and grid-responsive data-centre operation. Although earlier publications were used for contextual background where applicable, studies published before 2022 were not selectively incorporated into the systematic evidence set. The set review period remained consistent with this. The search was structured around three concept groups: computing infrastructure, AI and computational workload, and grid / energy flexibility. Variations of search terms used were “data centre”, “data center”, “artificial intelligence”, “machine learning”, “GPU”, “high-performance computing”, “cloud”, “edge-cloud”, “grid interactive”, “smart grid”, “renewable energy”, “carbon aware”, “workload shifting” etc. “demand response”, “ancillary services”, “frequency response”, “energy storage”, “battery”, “UPS”, “cooling”, “thermal storage”, “waste heat”, “heat recovery”, “power quality” and “converter”.

A common Boolean structure was formulated as:

("data centre*" OR "data center*" OR datacenter*) AND ("artificial intelligence" OR "machine learning" OR GPU* OR HPC OR "high-performance computing" OR cloud OR "edge-cloud") AND ("grid-interactive" OR "smart grid*" OR "renewable energy" OR "carbon-aware" OR "workload shift*" OR "demand response" OR "ancillary service*" OR "frequency response" OR "energy storage" OR batter* OR UPS OR cooling OR "thermal storage" OR "waste heat" OR "heat recovery" OR "power quality")

Database syntax and field restrictions were adapted to the capabilities of each information source. The reconstructed database-specific strategy used for reproducibility is presented in Supplementary Table S1.

Supplementary Table S1. Database-specific search strategy.

Database	Search strategy
Scopus	TITLE-ABS-KEY(("data centre*" OR "data center*" OR datacenter*) AND ("artificial intelligence" OR "machine learning" OR GPU* OR HPC OR "high-performance computing" OR cloud OR "edge-cloud") AND ("grid-interactive" OR "smart grid*" OR "renewable energy" OR "carbon-aware" OR "workload shift*" OR "demand response" OR "ancillary service*" OR "frequency response" OR "energy storage" OR batter* OR UPS OR cooling OR "thermal storage" OR "waste heat" OR "heat recovery" OR "power quality"))
Web of Science	TS=(("data centre*" OR "data center*" OR datacenter*) AND ("artificial intelligence" OR "machine learning" OR GPU* OR HPC OR "high-performance computing" OR cloud OR "edge-cloud") AND ("grid-interactive" OR "smart grid*" OR "renewable energy" OR "carbon-aware" OR "workload shift*" OR "demand response" OR "ancillary service*" OR "frequency response" OR "energy storage" OR batter* OR UPS OR cooling OR "thermal storage" OR "waste heat" OR "heat recovery" OR "power quality"))
IEEE Xplore	("data center" OR "data centre" OR datacenter) AND ("artificial intelligence" OR "machine learning" OR GPU OR HPC OR cloud) AND ("smart grid" OR "renewable energy" OR "grid-interactive" OR "carbon-aware" OR "workload shifting" OR "demand response" OR "ancillary services" OR "frequency response" OR "energy storage" OR UPS OR cooling OR "waste heat" OR "power quality")
ACM Digital Library	("data center" OR "data centre" OR datacenter) AND ("artificial intelligence" OR "machine learning" OR GPU OR HPC OR cloud OR "edge-cloud") AND ("renewable energy" OR "carbon-aware" OR "grid-interactive" OR "smart grid" OR "workload shifting" OR "demand response" OR "energy storage" OR UPS OR cooling OR "waste heat")
ScienceDirect	("data center" OR "data centre" OR datacenter) AND ("artificial intelligence" OR "machine learning" OR GPU OR HPC OR cloud) AND ("renewable energy" OR "carbon-aware" OR "smart grid" OR "demand response" OR "workload shifting" OR "energy storage" OR UPS OR cooling OR "thermal storage" OR "waste heat" OR "power quality")
SpringerLink	("data center" OR "data centre" OR datacenter) AND ("artificial intelligence" OR "machine learning" OR GPU OR HPC OR cloud) AND ("renewable energy" OR "carbon-aware" OR "smart grid" OR "demand response" OR "workload shifting" OR "energy storage" OR UPS OR cooling OR "waste heat")
Google Scholar	Supplementary combinations included "AI data center" "grid flexibility", "AI data centre" "renewable energy", "data center" "carbon-aware" workload, "data center" "demand response", "data center" UPS "frequency response", "data center" "waste heat" and "data center" "power quality" UPS.

Backward reference searching and forward citation tracking were additionally used to identify eligible studies not captured by the principal database queries. Google Scholar was essentially a secondary discovery source due to its differing ranking and retrieval behaviour from conventional bibliographic databases.

3.3 Eligibility Criteria

Eligibility criteria were defined to maintain a direct relationship between included evidence and the objectives of the review. Studies were eligible when they investigated data-centre or computing flexibility in

relation to renewable-energy integration, smart-grid participation, demand response, grid services, storage, carbon-aware operation, thermal flexibility or electrical compatibility. Studies concerned exclusively with computing performance or general energy efficiency without a renewable, carbon, grid or flexibility dimension were excluded. Eligible evidence included empirical studies, experiments, field demonstrations, operational or workload-trace studies, simulation and optimisation studies and technically relevant reviews. Only English-language publications containing sufficient methodological information for meaningful interpretation were considered (Table 1).

Table 1. Eligibility criteria.

Inclusion criteria	Exclusion criteria
Published from 2022 to 2026	Published before 2022
Data-centre, AI / GPU, HPC, cloud or edge-cloud context	Unrelated residential, commercial or industrial energy systems
Renewable energy, carbon-aware computing, smart-grid participation, demand response, storage or grid-service relevance	Computing performance or generic energy efficiency without a relevant grid, carbon or renewable-energy dimension
Experimental, empirical, field, trace-based, modelling, optimisation or relevant review evidence	Editorials, commentaries, news articles and unsupported opinion pieces
Sufficient methodological and outcome information	Inaccessible or methodologically insufficient publications
English-language publication	Non-English publication
Relevant technical, environmental, economic, grid or computing outcomes	Duplicate reports presenting no additional evidence

AI specificity:

1. AI / GPU-specific: direct evaluation of AI training, inference, machine-learning workloads, GPU-dense infrastructure or AI-focused HPC.
2. HPC: high-performance-computing environments or workloads not explicitly identified as AI.
3. General cloud / data centre: conventional cloud, virtual-machine, container or general data-centre environments.
4. Edge / cloud: geographically distributed edge-cloud or related distributed-computing systems.

Only the first category was treated as direct AI-specific evidence. HPC, cloud and edge-cloud findings were considered transferable evidence and interpreted with appropriate caution. The restriction to English-language publications may have reduced coverage of studies from regions where relevant work is published in other languages. This possible language and geographic bias was considered when interpreting the evidence and is discussed explicitly in the limitations.

3.4 Study Selection and Screening

Records obtained from the selected databases and supplementary searches were consolidated into a common reference-management dataset. Duplicate publications were identified using combinations of title, author names, publication year, journal or conference information and digital object identifier. Titles and abstracts were then screened against the predefined eligibility criteria. Publications clearly unrelated to data-centre flexibility, renewable integration, carbon-aware operation, storage, thermal management or grid interaction were excluded at this stage. Potentially eligible publications underwent full-text assessment. Full-text decisions considered the computing context, energy or grid relationship, flexibility mechanism, study methodology and reported technical, environmental, economic or computing outcomes. Studies discussing data-centre energy efficiency without a clear relationship to renewable integration, grid services, demand response, storage or carbon-aware operation were excluded. Duplicate publications describing the same underlying investigation were retained only when the later publication supplied materially additional evidence. The screening and eligibility process resulted in 43 unique studies for the final synthesis. The retained review records did not permit reliable reconstruction of all intermediate numerical screening counts; therefore, unverified intermediate counts are not reported.

3.5 Data Extraction and Evidence Coding

A structured extraction framework was applied consistently to the 43 included studies. Information extracted from each study included author and publication year, study design, geographical or electricity-market context, data-centre type, AI-specificity category,

workload characteristics, renewable-energy source, grid-interaction mechanism, storage technology, cooling or heat-recovery configuration, optimisation or control approach, validation level, performance indicators, quantitative outcomes, major findings, assumptions and reported limitations. Computational flexibility was coded as temporal workload deferral, job scheduling, virtual-machine or container placement, machine-learning task allocation and processor-power modulation. Spatial flexibility covered geographical workload migration and multi-site allocation. Electrical flexibility included battery storage, UPS coordination, converter operation and demand-response interfaces. Thermal flexibility included cooling control, thermal storage, heat pumps and waste-heat recovery.

Outcome measures were grouped into renewable-energy utilisation, peak-demand reduction, operational carbon emissions, lifecycle or embodied carbon, electricity cost, ancillary-service capability, frequency response, power quality, storage performance, latency, service-level compliance and workload completion. Where a study reported several outcomes, each outcome was extracted separately to avoid representing the publication only through its most favourable result. A complete characteristics matrix for all 43 studies is reported in Supplementary Table S3 and includes study type, data-centre context, AI specificity, workload, geographical or grid setting, renewable-energy context, flexibility mechanism, methodology, validation level, performance indicators, principal results and limitations.

3.6 Study Appraisal and Operational Evidence Strength

Study appraisal considered study design, validation context, reported assumptions and limitations, and relevance to grid-interactive data-centre operation. A separate numerical methodological-quality score was not used in the final synthesis because reproducible study-level domain ratings were not available for the complete evidence set. Operational evidence strength was therefore represented using four validation levels: Level I, operational or field evidence; Level II, experimental, prototype, hardware-in-the-loop or realistic operational-trace evidence; Level III, simulation, optimisation or analytical modelling evidence; and Level IV, secondary or conceptual evidence. These validation levels indicate the degree of operational demonstration and do not imply methodological inferiority. Study-level design and validation classifications are reported in Supplementary Table S4.

3.7 Evidence Strength and Data Synthesis

Operational evidence strength was interpreted using the four validation levels described in Section 3.6:

Level I – Operational / field evidence: results obtained from deployed systems or operational data-centre environments.

Level II – Experimental or realistic trace-based evidence: laboratory experiments, prototypes, hardware-in-the-loop studies or analyses based on operational workload traces.

Level III – Simulation / optimisation evidence: numerical models, scheduling simulations, optimisation models or synthetic system evaluations.

Level IV – Secondary / conceptual evidence: review papers or primarily conceptual investigations.

This hierarchy did not imply that lower-level studies were methodologically inferior. Instead, it represented the degree to which findings had been demonstrated under operationally representative conditions. The evidence was synthesised under four themes: architectures and enabling technologies; renewable-energy and flexibility coordination; multidimensional performance outcomes; and methodological and implementation barriers. Findings within each theme were interpreted according to AI specificity, study design, reported limitations, validation level, number of supporting studies and consistency of results.

Evidence was considered stronger when several methodologically sound studies produced consistent findings and at least part of the evidence involved operational, experimental or realistic trace-based validation. Evidence was considered moderate when findings were consistent but predominantly derived from simulation or optimisation. Evidence was considered limited when supported by a small number of studies, inconsistent results, predominantly indirect cloud / HPC evidence or weak operational validation. Quantitative results were compared only when outcome definitions, units, baselines and system boundaries were sufficiently similar. Where appropriate, reported ranges were used for operating-cost reduction, carbon reduction, renewable-energy utilisation, peak-demand effects and grid-service performance. Results were not statistically pooled across incompatible metrics. Review papers were used to provide contextual synthesis but were not treated as additional independent primary observations when calculating the number of studies supporting a quantitative finding.

3.8 Reliability, Transparency and Ethical Considerations

Several measures were incorporated to strengthen transparency. Search concepts were linked directly to the review objectives, predefined eligibility criteria were applied during screening, duplicate records were checked using multiple bibliographic identifiers, and study characteristics were extracted using a common framework. In order to avoid conventional cloud or simulated systems' findings being presented as direct field evidence for AI-intensive facilities, AI specificity and strength of operational evidence were documented separately. The assessment also took possible bias sources into account. These seven sources have a wide multidisciplinary coverage which could not promise the identification of all relevant publications. Focusing exclusively on English may severely limit research on some countries and regions.

The concentration on academic literature can also obscure grey literature, industry reports and unpublished operational demonstrations. Studies demonstrating successful reductions in energy cost, emissions, or grid demand may be favoured by publication bias over null or negative findings. The fast advancement of technology adds another risk that evidence can become outdated as GPU architectures, cooling technologies and electricity-market agreements change. These concerns will be elaborated on in Section 6. This review was based on published scholarly material and did not involve human participants, identifiable personal information, clinical records or confidential organisation data. As a result, formal ethics approval and informed consent were not applicable. Integrity in research was maintained through appropriate attribution, transparent study-selection criteria, material treatment of indirect evidence and avoiding unsupported generalisation from conventional cloud or HPC systems to AI-specific data centres.

4. Results

The systematic review included 43 studies published between 2022 and 2026. The evidence was analysed according to publication period, study design, validation level, geographical context, AI specificity, workload type, flexibility mechanism and reported outcomes. This broader characterisation was undertaken before the thematic synthesis to distinguish evidence obtained directly from AI-intensive facilities from evidence transferred from HPC, cloud and conventional data-centre environments. The results were subsequently organised into four themes: architectures and enabling infrastructure, renewable-energy coordination, multidimensional performance outcomes, and implementation constraints.

4.1 Characteristics of Included Studies

Research activity increased markedly across the review period. Four studies (9.3%) were published in 2022, 11 (25.6%) in 2023, 9 (20.9%) in 2024, 14 (32.6%) in 2025 and 5 (11.6%) in 2026. The concentration of studies in 2024–2026 reflects the recent expansion of interest in carbon-aware computing, AI-related electricity demand and grid-responsive data-centre operation. The evidence remained dominated by modelling. Twenty-seven studies (62.8%) relied principally on simulation, optimisation or analytical modelling and were classified as Level III validation. Seven studies (16.3%) provided experimental, prototype, implementation or operational-trace evidence and were classified as Level II. Three studies (7.0%) provided operational or field evidence and were classified as Level I, while six studies (14.0%) were reviews, surveys or technical perspectives classified as Level IV. Consequently, approximately three-fifths of the evidence concerning grid-interactive data-centre operation remained model-based rather than operationally demonstrated.

The distinction between AI-specific and transferable evidence was equally important. Six studies (14.0%) directly examined AI, machine-learning or GPU-intensive computing; one study (2.3%) focused primarily on HPC;

10 studies (23.3%) examined cloud or edge-cloud systems; and 26 studies (60.5%) addressed general data-centre infrastructure or operation. Direct AI evidence was therefore a relatively small part of the overall corpus. Findings concerning batteries, UPS systems, cooling, waste-heat recovery and power quality were particularly dependent on conventional data-centre evidence and should be interpreted as technically transferable rather than as direct validation in GPU-dense AI facilities.

Geographic coverage was uneven. Explicit operational or modelling contexts included Phoenix, Arizona; Texas, the Mid-Atlantic and Western Interconnection in the United States; Dallas, Elizabeth City and Boston; southwest China; Beijing; German electricity-carbon conditions; and Google's geographically distributed production fleet. Alibaba production traces provided additional real-world workload evidence, although they were not tied to a single grid model. No included study provided a named operational demonstration in an African power system,

and evidence from other developing regions was also limited. The geographic distribution therefore favoured North American, European and Chinese electricity and computing environments.

Table 2 demonstrates that the evidence base contains a clear validation imbalance. Modelling provides broad coverage of scheduling, storage, markets and grid interactions, but direct operational evidence is concentrated in a small number of studies. Radovanović et al. [3] provide production-fleet evidence for temporal carbon-aware computing, while Colangelo et al. [42] and Williams et al. [16] provide direct evidence from GPU-intensive systems. Trace-based studies by Caprara et al. [43] and Zhou et al. [2] further strengthen the case for workload flexibility but do not reproduce all constraints associated with live electricity-market participation. Accordingly, evidence for the existence of computational flexibility is stronger than evidence for sustained, market-integrated grid-service delivery.

Table 2. Characteristics of the 43 studies included in the systematic review.

Study	Evidence context	Study design	Validation	Geographic / grid context	Principal mechanism / outcome	Main evidence limitation
Acun et al. [6]	General data centre	Framework and modelling	Level III	Multi-site / not location-specific	Joint hardware, renewables, workload and lifecycle carbon	Model-dependent assumptions
Asadov et al. [20]	Edge-cloud	Review and algorithmic evaluation	Level III	Geo-distributed	Spatio-temporal carbon-aware shifting	Limited operational validation
Bahreini et al. [19]	Cloud	Scheduling algorithm / simulation	Level III	Not location-specific	Carbon-aware workload dispatch	Simulation-based
Caprara et al. [42]	AI / ML and GPU-related	Operational-trace analysis and optimisation	Level II	Alibaba production traces	Workload deferral and demand response	Trace-to-power modelling required
Chen et al. [32]	General data centre	Experimental	Level II	Laboratory setting	Cooling and waste-heat recovery	Component / facility-scale evidence
Chen et al. [41]	Geo-distributed cloud / data centre	Online optimisation	Level III	Multi-region	Carbon-aware backup and scheduling	Model-based grid representation
Chen and Zheng [15]	AI data centre	Grid-planning optimisation	Level III	Multi-location grid scenarios	Temporal deferral and geographic shifting	Planning-model assumptions
Colangelo et al. [43]	AI / GPU-specific	Field demonstration	Level I	Phoenix, Arizona, USA	Software-controlled power flexibility	Single operational deployment
Crozier and Liska [12]	General data centre	Review / perspective	Level IV	International	Grid-flexibility mechanisms	Secondary evidence
Cui et al. [33]	General data centre	Laboratory experiment	Level II	Laboratory setting	Waste-heat-driven adsorption cooling	Component-level validation
Dinesh Reddy et al. [35]	General data centre	Optimisation / simulation	Level III	Not location-specific	Resource consolidation and migration	Simulation-based
Hao et al. [11]	General data centre	Review	Level IV	International	Waste-heat recovery technologies	Secondary evidence
Hewage et al. [21]	Edge / cloud	Optimisation / simulation	Level III	Geo-distributed	Spatio-temporal workload shifting	Migration assumptions
Hussain et al. [22]	Cloud	Workflow optimisation / simulation	Level III	Geo-distributed	Price- and energy-aware scheduling	Simulated workloads / grid signals
Khodayarsesh et al. [18]	Cloud	VM-placement optimisation	Level III	Geo-distributed	Carbon- and energy-aware placement	Limited field validation
Lin et al. [31]	General data centre	Thermodynamic modelling	Level III	Not location-specific	Heat pump and absorption refrigeration	Model-level validation
Miao et al. [17]	AI / GPU-specific DML-edge-cloud	Scheduling optimisation	Level III	Dallas, Elizabeth City and Boston	Multi-renewable DML scheduling	Modelled infrastructure
Nkwawir et al. [8]	General data centre	Multi-energy optimisation	Level III	Not location-specific	Carbon-aware multi-energy operation	Simulation-based

Paananen [1]	General data centre	Technical perspective	Level IV	International	UPS-based grid services	Limited primary experimental evidence
Piontek et al. [5]	Cloud / Kubernetes	Implemented experimental scheduling	Level II	German / European carbon-intensity scenarios	Temporal carbon-aware job shifting	Small-scale Kubernetes evaluation
Radovanović et al. [3]	Hyperscale data centre	Production-system evaluation	Level I	Google's multi-regional fleet	Carbon-intelligent temporal shifting	Proprietary operational environment
Ruparel [37]	Cloud	Robust optimisation	Level III	Not location-specific	Forecasting and battery scheduling	Simulation / model dependence
Senga et al. [14]	General data centre with AI-demand context	Capacity-expansion modelling	Level III	Texas, Mid-Atlantic and WECC, USA	Temporal flexibility, cost and emissions	Results sensitive to generation mix
Shim [36]	HPC	Empirical trace / data analysis	Level II	PM100 HPC dataset	Job-level carbon-efficiency estimation	HPC rather than AI-specific
Sun et al. [38]	General data centre	Electrical modelling	Level III	Not location-specific	Power-supply impedance	Model / component transferability
Sun et al. [39]	General data centre	System modelling	Level III	Not location-specific	Converter and network stability	Limited operational grid validation
Sutaria and Rönnberg [40]	General data centre	Measurement / experimental analysis	Level II	Data-centre installation	Supraharmonic propagation	Facility-specific electrical conditions
Takci et al. [13]	General data centre	Review / synthesis	Level IV	International	Power-system flexibility	Secondary evidence
Tao and Gadh [29]	General data centre	Dynamic grid simulation	Level III	Modified IEEE 39-bus system	Workload and UPS frequency response	Test-system rather than field validation
Verma et al. [9]	Cloud	Survey	Level IV	International	Energy-efficient workflow scheduling	Secondary evidence
Williams et al. [16]	AI / GPU-specific	Operational / field demonstration	Level I	Real-world GPU cluster	Grid-responsive compute orchestration	Limited number of deployments
Xu et al. [30]	General data centre	Multi-energy optimisation	Level III	Southwest China	Renewable portfolio and demand response	Regional model assumptions
Ye et al. [27]	General data centre	Optimisation / simulation	Level III	Geo-distributed	Workload and UPS coordination	Battery / UPS model dependence
Yuan et al. [10]	General data centre	Review	Level IV	International	Waste-heat recovery	Secondary evidence
Zhang et al. [7]	General data centre	Storage design / optimisation	Level III	Not location-specific	Energy-storage dispatch and grid services	Simulation-based
Zhang et al. [24]	General data centre	Optimisation and event-based simulation	Level III	Multi-data-centre / aggregator	Demand response and workload sharing	Simulated market environment
Zhang et al. [25]	General data centre	Integrated network optimisation	Level III	Power-traffic-data network model	Workload migration and grid-stress reduction	Model complexity and assumptions
Zhao and Zhou [4]	Cloud	Scheduling optimisation	Level III	Distributed cloud	Renewable-aware task placement	Simulation-based
Zhao et al. [23]	General / geo-distributed data centre	Optimisation	Level III	Geo-distributed	Workload, energy and demand response	Forecast and market assumptions
Zhou et al. [34]	General data centre	Validated dynamic modelling	Level III	Beijing, China	Heat pump, free cooling and heat recovery	Case-specific thermal model
Zhou et al. [26]	General data centre	Optimisation / simulation	Level III	Geo-distributed	Energy-aware coordinated operation	Simulation-based
Zhou et al. [2]	AI / GPU-specific HPC	Real workload-trace analysis	Level II	Fourteen AI / general HPC data centres	Grid flexibility and opportunity cost	Trace-based rather than live grid response
Zhou et al. [28]	General data centre	Load modelling / optimisation	Level III	Not location-specific	Energy efficiency and demand-response quality	Model-based validation

4.2 Grid-Interactive Architectures and Enabling Infrastructure

The architecture literature indicates that grid-interactive data centres depend on four interacting layers: computational orchestration, electrical conversion and storage, thermal management, and grid-facing

measurement and control. Software controls determine when and where flexible workloads execute, while batteries and UPS systems provide rapid electrical response. Cooling systems and recoverable heat create slower thermal flexibility, and grid-facing control links these resources to carbon signals, electricity prices, frequency measurements or demand-response commands (Table 3).

Table 3. Evidence on grid-interactive architectures and enabling infrastructure.

Study	Evidence context	Validation	Architectural focus	Principal finding
Paananen [1]	General DC	Level IV	Grid-interactive UPS	UPS batteries and bidirectional interfaces can support rapid grid services
Sun et al. [38]	General DC	Level III	Power-supply impedance	Converter behaviour requires explicit representation rather than static-load assumptions
Sun et al. [39]	General DC	Level III	Electrical stability	Aggregated electronic loads and network conditions influence system stability
Sutaria and Rönnberg [40]	General DC	Level II	Power quality	UPS operating modes alter supraharmmonic behaviour
Radovanović et al. [3]	Hyperscale DC	Level I	Carbon-intelligent control	Production control can shape temporal computing capacity using carbon forecasts
Acun et al. [6]	General DC	Level III	Lifecycle architecture	Workloads, batteries, renewables and embodied carbon require joint evaluation
Khodayarsesht et al. [18]	Cloud	Level III	Distributed VM control	Carbon and energy signals can guide geographic resource allocation
Miao et al. [17]	AI / DML	Level III	Renewable edge-cloud	Distributed ML scheduling links computation to multiple renewable sources
Xu et al. [30]	General DC	Level III	Multi-energy architecture	Generation, conversion, storage and computing can be represented as a coupled energy hub
Yuan et al. [10]	General DC	Level IV	Thermal infrastructure	Heat recovery extends data-centre integration to neighbouring thermal demand
Colangelo et al. [42]	AI / GPU	Level I	Grid-responsive software control	Direct GPU-cluster power modulation can preserve priority-service requirements
Williams et al. [16]	AI / GPU	Level I	Integrated power-flexible architecture	Grid telemetry and workload orchestration enable multiple forms of compute flexibility

The strongest architectural evidence concerns software-controlled computing flexibility. Radovanović et al. [3] demonstrated production-scale temporal control across Google's fleet, while the Phoenix field demonstration [42] used 256 GPUs and achieved a 25% reduction in cluster power for three hours without violating the specified AI quality-of-service requirements. Williams et al. [16] subsequently demonstrated rapid and sustained power control on a 130-kW GPU cluster. These operational results provide stronger support for AI-specific computational flexibility than the predominantly simulation-based evidence available for integrated battery, market and thermal coordination. Electrical evidence is narrower but important. Power-supply impedance, system-level stability and supraharmmonic measurements [38-40] show that converter-dominated facilities cannot always be represented as conventional static loads. Consequently, the feasibility of workload flexibility does not by itself establish the safe delivery of grid services. Electrical dynamics, UPS operating states, battery reserve requirements and facility protection remain part of the grid-interaction problem.

4.3 Renewable-Energy Coordination Strategies

The second theme examined how computing, storage and thermal resources were coordinated with renewable-energy availability. Temporal workload deferral was the most frequently represented mechanism, followed by geographical migration, virtual-machine or container

placement, battery / UPS scheduling, multi-energy optimisation and thermal recovery. The evidence indicates that no individual mechanism addresses all timescales of flexibility (Table 4).

The evidence for computational coordination ranges from operational implementation to optimisation. Piontek et al. [5] implemented carbon-aware scheduling in Kubernetes and reported carbon reductions of 0.37-2.41% across four workload scenarios, with an average reduction of approximately 1.3% while maintaining service-level requirements. Miao et al. [17] reported renewable-energy utilisation of 90.8% in their distributed machine-learning scheduling framework, together with an average response time of 12.6 ms and a task-failure rate of 1.25%. These results indicate that renewable-aware scheduling can improve environmental outcomes, but the magnitude depends strongly on workload flexibility, carbon-signal variability and system assumptions. Storage and thermal mechanisms provided complementary flexibility. Xu et al. [30] reported that the optimised multi-energy portfolio could reduce system portfolio cost by up to 17.8% in the southwest China case study. At the component level, Cui et al. [33] experimentally achieved a coefficient of performance of 0.37 and a specific cooling power of 102.35 W/kg using a waste-heat-driven adsorption chiller. These findings support the technical viability of multi-energy and thermal pathways, although neither constitutes direct evidence of closed-loop thermal coordination in an operational AI data centre.

Table 4. Comparative evidence on renewable-energy coordination strategies.

Study	AI specificity	Validation	Flexibility mechanism	Evidence synthesis
Radovanović et al. [3]	Transferable hyperscale DC	Level I	Temporal workload deferral	Production workloads were shifted away from carbon-intensive periods while preserving daily computing capacity
Acun et al. [6]	General DC	Level III	Joint hardware-software planning	Joint optimisation reveals lifecycle trade-offs hidden by single-resource approaches
Zhao and Zhou [4]	Cloud	Level III	Renewable-aware placement	Geographic task placement increases alignment with renewable availability
Khodayarseresht et al. [18]	Cloud	Level III	VM placement	Energy / carbon signals guide placement subject to resource and QoS constraints
Miao et al. [17]	Direct AI / DML	Level III	Multi-renewable DML scheduling	Spatio-temporal coordination improved renewable utilisation while retaining service constraints
Hewage et al. [21]	Edge / cloud	Level III	Spatio-temporal shifting	Combining spatial and temporal flexibility provides more options than either mechanism alone
Ye et al. [27]	General DC	Level III	UPS and workload coordination	Explicit UPS losses materially alter optimal schedules
Xu et al. [30]	General DC	Level III	Multi-energy portfolio	Hydro, solar, wind, conversion and storage were coordinated with demand response
Piontek et al. [5]	Cloud	Level II	Carbon-aware Kubernetes scheduling	Practical container scheduling shifted flexible jobs toward lower-carbon periods
Lin et al. [31]	General DC	Level III	Heat recovery and cooling	Server heat can be upgraded for useful thermal applications
Chen et al. [32]	General DC	Level II	Cooling and heat recovery	Experimental evidence supports integrated cooling / heat-recovery operation
Cui et al. [33]	General DC	Level II	Waste-heat-driven cooling	Low-grade server heat can drive additional cooling through adsorption technology

4.4 Multidimensional Performance Outcomes

Performance outcomes were evaluated across economic, environmental, grid and computing dimensions. Considerable heterogeneity in system boundaries prevented statistical pooling. Cost reduction could refer to a data-centre electricity bill, aggregator social cost, generation-system cost or long-term investment cost, while carbon outcomes were variously expressed using operational emissions, grid carbon intensity or lifecycle considerations. Quantitative results were therefore retained at study level rather than combined into a single effect estimate.

Several conclusions emerge from Table 5. First, the strongest direct evidence concerns short-duration power flexibility. Colangelo et al. [42] demonstrated a 25% reduction in GPU-cluster demand for three hours, while Caprara et al. [43] found that more than 20% of estimated GPU-side power in the Alibaba workload traces was associated with tasks suitable for deferral and achieved up to 22% load shedding during flexibility windows. Zhou et al. [2] found that AI-focused HPC facilities could provide flexibility at approximately 50% lower opportunity cost than general-purpose HPC facilities. Together, these studies provide a stronger basis for AI-specific workload flexibility than earlier general cloud simulations.

Second, economic gains were more consistent than carbon gains. Zhang et al. [24] reported an average 5% reduction

in data-centre energy cost and social-cost reductions above 30% in most simulated aggregator scenarios. Senga et al. [14] similarly found that temporal flexibility lowered total system costs across Texas, the Mid-Atlantic and WECC, with maximum reductions of approximately 5%, 4% and 2%, respectively. However, the corresponding carbon effect depended on regional generation composition. In the high-renewable Texas case, emissions declined by as much as 40%, whereas greater flexibility could increase emissions in regions where shifting favoured carbon-intensive baseload generation.

Third, longer flexibility windows did not automatically produce proportionally greater benefits. Chen and Zheng [15] reported location-dependent reductions of approximately 3–21% in grid investment and operating costs and found diminishing returns from increasing deferral duration. This reinforces the conclusion that flexibility value depends on where a facility is connected, when demand can be shifted, what generation is marginal during the receiving period and how much workload is actually deferrable. No pooled percentage reduction was calculated across these studies because the reported quantities were not methodologically interchangeable. A 5% reduction in a data-centre electricity bill, a 5% reduction in total power-system cost and a 17.8% reduction in multi-energy portfolio cost represent different denominators and system boundaries. Preserving these distinctions avoids implying a common effect size where none exists.

Table 5. Quantitative and qualitative evidence on multidimensional performance.

Study	Evidence context / validation	Performance dimension	Reported outcome	Interpretation
Piontek et al. [5]	Cloud, Level II	Carbon / QoS	0.37-2.41% CO ₂ reduction; ~1.3% average	Implemented temporal shifting produced modest but consistent carbon savings
Zhang et al. [24]	General DC, Level III	Cost / demand response	~5% average DC energy-cost reduction; >30% social-cost reduction in most cases; 6% greater power reduction	Aggregator coordination improves joint grid / DC value
Xu et al. [30]	General DC, Level III	Multi-energy cost	Up to 17.8% portfolio-cost reduction	Benefits depend on regional renewable and infrastructure assumptions
Miao et al. [17]	AI / DML, Level III	Renewable utilisation / QoS	Renewable utilisation 90.8%; response time 12.6 ms; failure rate 1.25%	Directly AI-related but simulation-based
Zhou et al. [2]	AI-HPC, Level II	Flexibility cost	Approximately 50% lower flexibility cost than general-purpose HPC	Real traces indicate distinctive flexibility value of GPU-heavy AI computing
Colangelo et al. [42]	AI / GPU, Level I	Power reduction / QoS	25% cluster-power reduction for 3 h on 256 GPUs	Strong direct operational evidence
Caprara et al. [43]	AI / ML traces, Level II	Demand flexibility	>20% of estimated GPU-side power associated with deferrable tasks; up to 22% load shedding	Strong trace-based evidence for short-duration flexibility
Chen and Zheng [15]	AI DC, Level III	Grid planning / cost	Approximately 3–21% reduction in grid investment and operating costs depending on location and flexibility	Strong location dependence
Senga et al. [14]	General DC / AI-demand context, Level III	System cost / emissions	Cost reductions up to 4% Mid-Atlantic, 5% Texas and 2% WECC; emissions from substantial reduction to increases under some mixes	Demonstrates cost-carbon divergence
Ye et al. [27]	General DC, Level III	UPS efficiency	Approximately 5% UPS losses materially affected schedules	Storage losses reduce idealised flexibility
Cui et al. [33]	General DC, Level II	Thermal performance	COP 0.37; 102.35 W/kg specific cooling power	Experimental component-level thermal evidence
Tao and Gadh [29]	General DC, Level III	Frequency response	Coordinated workload / UPS control improved frequency nadir and recovery in the IEEE 39-bus test system	Promising grid-service evidence remains simulation-based

4.5 Implementation Constraints and Evidence Gaps

Implementation barriers were distributed across computational, electrical, environmental, market and thermal domains. The most recurrent issues were inconsistent evaluation methods, forecast uncertainty, simplified physical models, battery and UPS constraints, workload-service requirements, electricity-market assumptions, power-quality effects and limited operational validation (Table 6).

Forecast uncertainty already limits the practical performance of many proposed control strategies. Several studies assume that renewable output, electricity prices, carbon intensity, and workload demand are known in

advance or can be predicted with high accuracy. In practice, errors in these forecasts can change when workloads are scheduled and may reduce the expected economic or environmental benefits. Robust and distributionally aware optimisation offers one way of handling this uncertainty, although evidence from operational data centres remains limited. Battery condition and UPS efficiency create additional constraints. UPS systems are designed primarily to maintain computing continuity, so using them for grid services reduces the capacity available for backup and may increase battery cycling. Ye et al. [27] reported that UPS losses of about 5% were sufficient to influence optimal workload and storage scheduling. Models that ignore these losses or battery ageing may therefore overestimate the flexibility that can be sustained in practice.

Table 6. Consolidated implementation constraints and strength of supporting evidence.

Study	Constraint	Validation	Implication
Acun et al. [6]	Incomplete lifecycle boundaries	Level III	Operational carbon savings may be offset by embodied impacts of servers, batteries and renewable infrastructure
Asadov et al. [20]	Benchmark inconsistency	Level III	Carbon signals, workloads and shifting windows limit cross-study comparability
Verma et al. [9]	Heterogeneous metrics	Level IV	Workflow algorithms cannot be ranked reliably across incompatible benchmarks
Khodayarsesht et al. [18]	Migration / QoS trade-off	Level III	Carbon-efficient placement may increase latency and communication overhead
Zhao et al. [23]	Forecast dependence	Level III	Benefits depend on workload, price and renewable-generation forecasts
Zhang et al. [24]	Market coordination	Level III	Baselines, incentives and aggregator objectives influence realised benefits
Ye et al. [27]	UPS losses and degradation	Level III	Ideal storage assumptions overestimate available flexibility
Sun et al. [38]	Converter-model accuracy	Level III	Static-load approximations can misrepresent dynamic behaviour
Sun et al. [39]	Electrical stability	Level III	Network impedance, imbalance and converter interaction affect safe operation
Sutaria and Rönnberg [40]	Supraharmonic propagation	Level II	Changes in UPS mode can create power-quality consequences
Yuan et al. [10]	Heat-demand matching	Level IV	Waste heat has limited value without suitable nearby thermal demand
Hao et al. [11]	Thermal technology selection	Level IV	Feasibility varies with temperature, exergy, climate and economics
Colangelo et al. [42]	Scale / generalisation	Level I	Field evidence is strong but remains limited to a small number of deployments
Senga et al. [14]	Regional carbon response	Level III	Cost-optimal flexibility can increase rather than reduce emissions
Caprara et al. [43]	Deferrable workload identification	Level II	Grid flexibility depends on the proportion and duration of latency-tolerant computing

Rebound demand also requires attention. Workloads postponed during a grid event still have to be completed later. If too many jobs resume at the same time, the resulting increase in demand can create a new peak and reduce part of the benefit achieved during the original response period. Geographical migration can additionally increase data transfer, communication delay and network energy consumption. Direct AI evidence indicates that flexible operation can preserve quality of service under controlled conditions, but the extent of permissible deferral differs between training, batch processing, latency-critical inference and other computing tasks. Flexibility therefore cannot be represented as a uniform percentage of total facility load. Market participation remains less mature than technical load control. Data centres require measurable baselines, dispatch rules, remuneration and penalties that reflect both grid requirements and computing-service obligations. Aggregator-based models show potential economic value, but most evidence remains simulated. Fast frequency response is also technically promising, although dynamic modelling using the IEEE 39-bus test system cannot substitute for verification of converter stability, battery response, protection coordination and communication latency in an operational facility.

Power quality and converter stability constitute a separate implementation requirement. Evidence from impedance modelling and supraharmonic measurement shows that rapid changes in converter-controlled demand and UPS operating mode can interact with facility and upstream network characteristics. These effects are particularly relevant for GPU-dense AI data centres because power density and converter penetration are increasing. However, direct electrical measurements from large operational AI facilities remain scarce. Finally, thermal integration is technically feasible but geographically constrained. Waste-heat utilisation depends on heat quality, temperature upgrading, distance to consumers, local climate and seasonal heat demand. Experimental adsorption cooling and validated heat-pump modelling demonstrate that useful recovery is possible, but there is little evidence of simultaneous coordination among AI workloads, cooling, battery condition, renewable forecasts and external grid-service commitments. Taken together, the evidence strength is uneven across the four themes. Computational workload flexibility has the strongest support, combining production, field, trace-based and modelling evidence. Renewable-aware scheduling has moderate evidence, with consistent benefits but continued reliance on simulation. Thermal

flexibility has moderate component-level evidence but limited AI-specific integration evidence. Ancillary-service and market participation remain promising but less mature, as much of the evidence relies on power-system models rather than sustained commercial operation. The most significant research gap is therefore not whether data centres possess flexibility, but how much flexibility can be delivered repeatedly under realistic AI workloads, electrical-security constraints, battery ageing, forecast error, carbon-aware objectives and market obligations.

5. Discussion

This review shows that the value of grid-interactive AI data centres depends less on the availability of any single flexibility mechanism than on how computational, electrical, thermal and grid-facing resources are coordinated. The evidence is strongest for workload-based flexibility, where production, trace-based and modelling studies consistently show that some computing demand can be shifted across time or location. Evidence for batteries, cooling systems, waste-heat recovery and ancillary services is also promising, but much of it is derived from conventional data centres, cloud systems or simulation environments rather than operational AI facilities. The principal implication is therefore one of conditional transferability: established data-centre flexibility mechanisms are relevant to AI infrastructure, but their performance cannot be assumed to scale directly to GPU-dense facilities with different power densities, workload durations and service requirements. The evidence also demonstrates that flexibility is multidimensional. A strategy that improves electricity cost may not reduce carbon emissions, and a strategy that supports the grid may create computing, battery or electrical-security penalties. This makes single-objective evaluation inadequate for grid-interactive AI data centres. The following discussion therefore focuses on the implications, contradictions and confidence of the evidence rather than repeating the mechanisms reported in Section 4.

5.1 Grid-Interactive Architectures and Evidence Transferability

The architectural evidence supports a cyberphysical interpretation of the data centre in which workload schedulers, storage systems, cooling equipment, power converters and grid signals operate as interacting control layers. The practical significance of this architecture is that flexibility cannot be evaluated solely from the amount of computing demand that appears technically shiftable. A workload scheduler may identify deferrable jobs, but actual grid-service capability remains constrained by converter behaviour, UPS state, cooling requirements, communication delays and service-level obligations. The power-system studies by Sun et al. [38,39] and the measurements of Sutaria and Rönnberg [40] are therefore important because they show that controllable digital demand must still satisfy electrical stability and power-quality requirements. The evidence also requires careful distinction between direct AI findings and transferred knowledge. AI / GPU-oriented studies provide the most

relevant information on computational flexibility because they represent the power density and workload characteristics increasingly associated with AI infrastructure. In contrast, much of the knowledge concerning UPS systems, heat recovery, converter stability and building-level energy management originates from conventional data centres. These findings remain useful because the underlying electrical and thermal components are shared, but their magnitude and operating constraints may differ in AI facilities. The review therefore supports transfer of mechanisms more strongly than transfer of numerical performance estimates.

This distinction affects confidence in the overall evidence. Workload flexibility can be regarded as comparatively well supported because consistent findings appear across production, trace-based and optimisation studies. By contrast, an integrated architecture in which AI workloads, batteries, cooling systems, power converters and external grid markets are coordinated simultaneously has limited operational validation. The major engineering challenge is consequently no longer demonstrating that individual flexibility mechanisms exist, but establishing that they can operate together reliably under realistic computing and electricity-system conditions.

5.2 Renewable-Energy, Storage and Thermal Coordination

The evidence indicates that computational, electrical and thermal flexibility operate over different timescales and should be treated as complementary rather than interchangeable resources. Temporal workload shifting is most appropriate for delay-tolerant training, batch processing and other tasks whose completion time can be adjusted. Geographical migration can exploit regional differences in renewable availability and carbon intensity, but it introduces communication, latency and data-transfer constraints. Batteries and UPS systems provide faster electrical response, whereas cooling systems and thermal storage provide slower flexibility through thermal inertia. The practical value of coordination therefore lies in matching each resource to a timescale and service for which it is technically suitable. The storage literature also demonstrates why idealised flexibility estimates require caution. Ye et al. [27] showed that UPS conversion losses of approximately 5% were sufficient to alter optimal workload and storage schedules. This illustrates a broader issue: storage available on paper is not equivalent to storage available for grid participation. UPS batteries must retain sufficient capacity for their primary backup function, and repeated grid-service use can change their economic value through additional cycling and ageing. Although battery degradation is frequently recognised as a constraint, it is not quantified consistently across the reviewed studies. Evidence on sustained grid participation over battery lifetime is therefore weaker than evidence based on short optimisation horizons.

Thermal flexibility has a similarly conditional character. Xu et al. [30], Chen et al. [32], Yuan et al. [10], Cui et al. [33] and Hao et al. [11] show that cooling optimisation and heat recovery can reduce rejected thermal energy or

provide useful heating and cooling services. However, heat quantity alone does not determine practical value. Temperature quality, heat-pump requirements, distance to external consumers, climatic conditions and temporal matching between available waste heat and thermal demand determine whether recovery is technically and economically useful. Thermal integration is therefore highly location dependent, and current evidence is stronger at component and system-design level than for simultaneous thermal, computational and grid-service coordination in operational AI facilities.

5.3 Cost, Carbon and Location-Dependent Trade-offs

One of the most important findings is that economic and environmental objectives are not automatically aligned. Zhang et al. [24] reported approximately 5% reductions in data-centre energy cost under aggregator coordination, demonstrating that flexibility can generate measurable economic benefits. However, studies such as Senga et al. [14] show that lower system cost can coexist with higher emissions under some electricity-generation mixes. A controller responding only to price may therefore move workload toward inexpensive electricity that is not necessarily low carbon. This explains why cost reduction should not be interpreted as evidence of decarbonisation.

Carbon accounting also requires a distinction between average and marginal carbon intensity. Average carbon intensity describes the emissions associated with the overall electricity mix during a period, whereas marginal carbon intensity reflects the emissions consequences of the generator responding to an incremental change in demand. For grid-interactive scheduling, the marginal signal is often more directly related to the emissions consequence of shifting additional computing load. A period with a relatively clean average generation mix can still have a carbon-intensive marginal generator, while a relatively carbon-intensive system may have periods when additional renewable generation would otherwise be curtailed. Carbon-aware scheduling based on different signals can therefore produce different decisions even when applied to the same workload.

A second distinction concerns operational and embodied carbon. Operational carbon arises from electricity consumed during computing, cooling and associated facility operation. Embodied carbon is associated with manufacturing and deploying servers, batteries, power equipment and renewable-energy infrastructure. Acun et al. [6] demonstrates why this lifecycle boundary matters: adding batteries or replacing hardware may reduce operational emissions while introducing additional embodied impacts. Consequently, a strategy cannot be described as comprehensively low carbon merely because operational electricity emissions decline. Lifecycle consequences become particularly important where flexibility requires additional storage capacity or accelerated infrastructure replacement. Location further modifies these trade-offs. The environmental value of shifting depends on the generation mix, renewable curtailment, network conditions and carbon profile of both the sending and receiving location. Chen and Zheng [15]

show that the grid-planning value of temporal deferral and geographical migration varies by location and that longer flexibility windows do not necessarily produce proportionally larger benefits. Geographical migration can also introduce additional network energy consumption and latency. The review therefore does not support a universal numerical estimate for the cost or carbon benefit of grid-interactive operation. Outcome definitions and system boundaries are too heterogeneous for meaningful pooling. Quantitative findings are most defensible when retained within their original grid, workload and accounting context rather than combined into an artificial average effect.

5.4 Uncertainty, Rebound Effects and Operational Constraints

Forecast uncertainty is a central limitation because most flexibility decisions depend on information that is not perfectly known in advance. Renewable output, electricity prices, grid carbon intensity and workload arrivals all vary over time. Many optimisation studies use deterministic or highly accurate forecasts, creating conditions that may overstate achievable benefits. Ruparel [37] addresses this limitation more directly through distributionally robust optimisation, illustrating a transition from perfect-information scheduling toward control methods that remain feasible when forecasts deviate from expectations. The broader evidence nevertheless contains substantially more deterministic optimisation than operational testing under forecast error. Reported benefits should therefore be interpreted partly as performance under assumed information quality rather than guaranteed field outcomes.

Rebound demand creates a related operational problem. Curtailing or deferring computing does not necessarily eliminate the associated work; many jobs must be completed later. If a large volume of deferred demand resumes simultaneously after a grid event, a new peak may emerge and partly offset the original benefit. This problem is especially relevant where multiple facilities or aggregators respond to the same price or grid signal. The reviewed literature recognises workload deadlines and scheduling constraints, but explicit assessment of post-event rebound and synchronised recovery remains limited. Reliable participation therefore requires recovery scheduling to be treated as part of the flexibility event rather than as an independent post-event process. Market design further determines whether technically available flexibility becomes commercially usable. Zhao et al. [23] and Zhang et al. [24] consider demand-response participation and aggregator coordination, whereas Paananen [1] and Tao and Gadh [29] extend the discussion toward faster ancillary and frequency-related services. These market settings impose different requirements. Conventional demand response may tolerate slower scheduling decisions, while frequency services require rapid, measurable and dependable response. Both require credible baselines so that actual load reduction can be distinguished from normal workload variation. Remuneration must also compensate data-centre operators for lost scheduling freedom, battery use, migration costs or service risk. Current evidence

demonstrates plausible economic participation models but remains dominated by simulation rather than sustained commercial deployment.

Electrical constraints become increasingly important as response times become shorter. Sun et al. [38,39] show that converter impedance and network interactions influence data-centre stability, while Sutaria and Rönnberg [40] demonstrate that UPS operating modes affect supraharmonic propagation. These findings imply that rapid workload modulation or battery dispatch should not be qualified for ancillary services solely on the basis of available megawatts. Converter stability, harmonic behaviour, protection coordination and UPS operating state must also remain within acceptable limits. The literature presently connects these electrical issues to grid-interactive operation conceptually, but direct studies combining AI workload modulation with detailed converter and power-quality measurements remain scarce.

Cybersecurity and privacy constitute an additional evidence gap. Grid interaction requires exchange of telemetry relating to facility load, computing availability, storage state and external grid conditions. Such information may expose commercially sensitive workload patterns, while compromised control communication could create both computing and electrical consequences. However, cybersecurity and privacy were not systematically evaluated in most of the 43 included studies. The evidence base therefore does not support strong conclusions about the effectiveness of particular cybersecurity architectures. Rather, the limited coverage itself is important: secure communication, authentication, access control, privacy-preserving aggregation and resilience to manipulated grid or carbon signals require more direct evaluation before large-scale automated participation can be regarded as mature.

5.5 Practical Implications for Data-Centre Operators and Grid Planners

For data-centre operators, the evidence supports a staged approach to grid interaction. The first requirement is to classify workloads according to flexibility rather than treating total facility demand as dispatchable. Delay-tolerant training and batch workloads can provide temporal flexibility, whereas latency-critical inference and priority services require stricter protection. Operators also require sufficiently granular telemetry to observe workload demand, facility power, UPS state, cooling conditions and relevant grid signals. Flexibility limits should include battery reserve requirements, conversion losses and computing-service constraints so that grid participation does not compromise the facility's primary reliability function. Carbon-aware operation should use signals appropriate to the intended environmental objective. Electricity price, average carbon intensity, marginal carbon intensity and renewable availability are not interchangeable. Where carbon reduction is a contractual or policy objective, cost-only scheduling is insufficient. Operators considering geographical migration should additionally account for transfer overhead, latency and the carbon characteristics of the

receiving region. Similarly, battery participation should incorporate health and reserve limits rather than assuming the entire installed capacity is available for market services.

For grid planners and market operators, data centres should be evaluated as heterogeneous flexible resources rather than uniform controllable loads. Qualification procedures should distinguish short-duration workload modulation, longer workload deferral, battery discharge and thermal flexibility because each resource has different response, recovery and persistence characteristics. Demand-response programmes should include credible baseline methods, rebound management and clear remuneration for flexibility. Fast ancillary-service qualification should additionally consider converter stability and power-quality performance. Where waste heat is part of the proposed energy integration, planning should consider proximity to suitable heat consumers and local thermal demand rather than treating recovered heat as automatically useful.

The evidence therefore supports practical deployment, but initially within well-defined operational envelopes. Computational flexibility is sufficiently mature to justify controlled deployment and further field demonstration, whereas market-integrated battery, thermal and ancillary-service coordination requires stronger operational validation. For both operators and planners, the central principle is to value flexibility according to what can be delivered repeatedly while preserving computing quality of service, electrical security and environmental integrity, rather than according to the theoretical fraction of data-centre demand that appears technically adjustable.

6. Limitations and Future Work

This review has several limitations. Although seven major information sources were searched, coverage was not exhaustive; databases such as INSPEC, ProQuest, and Dimensions were not included and may contain additional relevant studies. Restricting the review to English-language publications may also have introduced language and geographical bias. In addition, the exclusion of grey literature, including industry reports, technical white papers, utility demonstrations, and unpublished operational studies, may have reduced representation of real-world deployment evidence. Publication bias is also possible because studies reporting successful reductions in cost, emissions, or peak demand may be more likely to appear in the scholarly literature than null or negative findings.

Generalisability is further constrained by the dominance of simulation and optimisation studies and the limited number of operational AI / GPU demonstrations. Geographic coverage was uneven, with most identifiable evidence concentrated in North America, Europe, and China, while Africa and other developing regions were underrepresented. Differences in workloads, carbon metrics, electricity systems, market rules, and performance indicators also prevented meaningful statistical pooling. Future research should prioritise

standardised datasets, common benchmarks, lifecycle carbon accounting, and uncertainty-aware optimisation. More hardware-in-the-loop and field demonstrations are needed to evaluate rebound demand, battery degradation, converter stability, cybersecurity, market participation, and thermal integration. Greater attention should also be given to underrepresented power systems and developing regions.

7. Conclusion

Grid-interactive AI data centres can support renewable smart grids, although the strength of evidence varies across flexibility mechanisms. Temporal workload shifting and geographical migration currently have the strongest support from production, field and trace-based studies. Battery-supported UPS systems, thermal management, waste-heat recovery and ancillary-service participation are also promising, but much of the available evidence remains simulation-based or derived from conventional data-centre environments. Economic, carbon, grid-reliability and computing-performance benefits do not necessarily improve simultaneously; outcomes depend on workload flexibility, carbon signals, regional generation mix, forecast accuracy, battery condition, converter stability and market design. Future research should prioritise standardised benchmarks, lifecycle carbon assessment, uncertainty-aware control and large-scale operational demonstrations across diverse power systems and regulatory contexts.

Conflict of Interest

The author declares that they have no competing interests.

Funding

This research received no specific funding.

Acknowledgements

The author has no additional acknowledgements to declare.

Author Contributions

The author conceived the review and contributed to the literature search, screening, data extraction, synthesis, interpretation, and manuscript preparation. The author reviewed the final manuscript and approved it for submission.

Supplementary Materials

Supplementary materials accompany the manuscript. Supplementary Table S1 contains the database search strategies; Table S2 presents eligibility criteria applied during full-text assessment; Table S3 provides characteristics of the 43 included studies; Table S4 reports study design and operational validation levels; and Supplementary File S5 contains the completed PRISMA 2020 checklist.

Data Availability Statement

The review is based on published and publicly accessible studies. Supporting materials, including the search strategies, eligibility criteria, study characteristics, study design and operational validation classifications, and PRISMA documentation, are supplied with the manuscript.

Ethics Statement

Ethical approval was not required because this study reviewed previously published literature and involved no human participants, personal data, clinical records, or animals.

Consent for Publication

Not applicable.

AI Statement

Generative AI tools were used during revision to assist with language editing and organization. The authors independently checked the scientific content, references, interpretations, and conclusions and remain fully responsible for the final manuscript.

References

- [1] J. Paananen. Grid-interactive data centers enabling energy transition: Data center's hidden potential to provide essential grid services of a future power system. *IEEE Electrification Magazine*, 2023, 11(3), 26-34. DOI: 10.1109/MELE.2023.3291195
- [2] Y. Zhou, A. Paredes, C. Essayeh, T. Morstyn. AI-focused HPC data centers can provide more power grid flexibility and at lower cost. *ArXiv*, 2024. DOI: 10.48550/arXiv.2410.17435
- [3] A. Radovanović, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, et al. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems*, 2023, 38(2), 1270-1280. DOI: 10.1109/TPWRS.2022.3173250
- [4] D. Zhao, J. Zhou. An energy and carbon-aware algorithm for renewable energy usage maximisation in distributed cloud data centers. *Journal of Parallel and Distributed Computing*, 2022, 165, 156-166. DOI: 10.1016/j.jpdc.2022.04.001
- [5] T. Piontek, K. Haghshenas, M. Aiello. Carbon emission-aware job scheduling for Kubernetes deployments. *The Journal of Supercomputing*, 2024, 80, 549-569. DOI: 10.1007/s11227-023-05506-7
- [6] B. Acun, B. Lee, F. Kazhamiaka, K. Maeng, U. Gupta, M. Chakkaravarthy, et al. Carbon Explorer: A holistic framework for designing carbon-aware datacenters. *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2023, 2, 118-132. DOI: 10.1145/3575693.3575754
- [7] Y. Zhang, H. Tang, H. Li, S. Wang. Unlocking the flexibilities of data centers for smart grid services: Optimal dispatch and design of energy storage systems under progressive loading. *Energy*, 2025, 316, 134511. DOI: 10.1016/j.energy.2025.134511
- [8] B.W. Nkwawir, M.O. Kayalica, D. Guven, A.C. Duman, H.S. Erden. Carbon-aware workload management in data

- centers: A multi-energy integration approach. *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, 2025, 907-914. DOI: 10.1145/3679240.3735104
- [9] P. Verma, A.K. Maurya, R.S. Yadav. A survey on energy-efficient workflow scheduling algorithms in cloud computing. *Software: Practice and Experience*, 2024, 54(5), 637-682. DOI: 10.1002/spe.3292
 - [10] X. Yuan, Y. Liang, X. Hu, Y. Xu, Y. Chen, R. Kosonen. Waste heat recoveries in data centers: A review. *Renewable and Sustainable Energy Reviews*, 2023, 188, 113777. DOI: 10.1016/j.rser.2023.113777
 - [11] Y. Hao, H. Zhou, T. Tian, W. Zhang, X. Zhou, Q. Shen, et al. Data centers waste heat recovery technologies: Review and evaluation. *Applied Energy*, 2025, 384, 125489. DOI: 10.1016/j.apenergy.2025.125489
 - [12] C. Crozier, M. Liska. The potential of data center energy demand to provide grid flexibility. *Current Sustainable/Renewable Energy Reports*, 2025, 12. DOI: 10.1007/s40518-025-00258-9
 - [13] M.T. Takci, M. Qadrdan, J. Summers, J. Gustafsson. Data centres as a source of flexibility for power systems. *Energy Reports*, 2025, 13, 3661-3671. DOI: 10.1016/j.egyr.2025.03.020
 - [14] J.R.L. Senga, S.Wang, C.R. Knittel. Flexible data centers reduce power system costs but can increase emissions. *iScience*, 2026, 29(7), 116497. DOI: 10.1016/j.isci.2026.116497
 - [15] Y. Chen, X. Zheng. To defer or to shift? The role of AI data center flexibility on grid interconnection. *Proceedings of the 2026 ACM Sustainability Week*, 2026, 322-327. DOI: 10.1145/3765611.3815593
 - [16] C. Williams, P. Colangelo, A. Coskun, E. Levine, A. Neale, C. Roberts, et al. Power-flexible AI data centers: A new paradigm for grid-responsive compute. *ArXiv*, 2026. DOI: 10.48550/arXiv.2606.25098
 - [17] Z. Miao, L. Liu, H. Nan, W. Li, X. Pan, X. Yang, et al. Energy and carbon-aware distributed machine learning tasks scheduling scheme for the multi-renewable energy-based edge-cloud continuum. *Science and Technology for Energy Transition*, 2024, 79, 82. DOI: 10.2516/stet/2024076
 - [18] E. Khodayarsesht, A. Shameli-Sendi, Q. Fournier, M.R. Dagenais. Energy and carbon-aware initial VM placement in geographically distributed cloud data centers. *Sustainable Computing: Informatics and Systems*, 2023, 39, 100888. DOI: 10.1016/j.suscom.2023.100888
 - [19] T. Bahreini, A. Tantawi, A. Youssef. A carbon-aware workload dispatcher in cloud computing systems. In *Proceedings of the 2023 IEEE 16th International Conference on Cloud Computing*, 2023, 212-218. DOI: 10.1109/CLOUD60044.2023.00032
 - [20] N. Asadov, V.C. Coroamă, M. Franzil, S. Galantino, M. Finkbeiner. Carbon-aware spatio-temporal workload shifting in edge-cloud environments: A review and novel algorithm. *Sustainability*, 2025, 17(14), 6433. DOI: 10.3390/su17146433
 - [21] T.B. Hewage, S. Ilager, M.A. Rodriguez, R. Buyya. A Framework for Carbon-Aware Real-Time Workload Management in Clouds Using Renewables-Driven Cores. *IEEE Transactions on Computers*, 2025, 74(8), 2757-2771. DOI: 10.1109/TC.2025.3571495
 - [22] M. Hussain, L.F. Wei, A. Rehman, A. Hussain, M. Ali, M.H. Javed. An electricity price and energy-efficient workflow scheduling in geographically distributed cloud data centers. *Journal of King Saud University-Computer and Information Sciences*, 2024, 36(8), 102170. DOI: 10.1016/j.jksuci.2024.102170
 - [23] M.M. Zhao, X.Y. Wang, J.R. Mo. Workload and energy management of geo-distributed datacenters considering demand response programs. *Sustainable Energy Technologies and Assessments*, 2023, 55, 102851. DOI: 10.1016/j.seta.2022.102851
 - [24] Y.J. Zhang, A. Tsiligkaridis, I.C. Paschalidis, A.K. Coskun. Data center and load aggregator coordination towards electricity demand response. *Sustainable Computing: Informatics and Systems*, 2024, 42, 100957. DOI: 10.1016/j.suscom.2024.100957
 - [25] Y.S. Zhang, B.K. Zou, X. Jin, Y.F. Luo, M. Song, Y.J. Ye, et al. Mitigating power grid impact from proactive data center workload shifts: A coordinated scheduling strategy integrating synergistic traffic-data-power networks. *Applied Energy*, 2025, 377, 124697. DOI: 10.1016/j.apenergy.2024.124697
 - [26] S.B. Zhou, M. Zhou, Z.Y. Wu, Y.Y. Wang, G.Y. Li. Energy-aware coordinated operation strategy of geographically distributed data centers. *International Journal of Electrical Power and Energy Systems*, 2024, 159, 110032. DOI: 10.1016/j.ijepes.2024.110032
 - [27] G.S. Ye, F. Gao, J.Y. Fang, Q. Zhang. Joint workload scheduling in geo-distributed data centers considering UPS power losses. *IEEE Transactions on Industry Applications*, 2023, 59(1), 612-626. DOI: 10.1109/TIA.2022.3214186
 - [28] Y.C. Zhou, F.C. Wei, S.X. Li, Z.H. Wang, J.F. Liu, D.R. Yu. Data center load modeling through optimal energy consumption characteristics: A path to simultaneously enhance energy efficiency and demand response quality. *Applied Energy*, 2025, 393, 126095. DOI: 10.1016/j.apenergy.2025.126095
 - [29] X.J. Tao, R. Gadh. Fast frequency response potential of data centers through workload modulation and UPS coordination. *IEEE Access*, 2025, 13, 214511-214519. DOI: 10.1109/ACCESS.2025.3646120
 - [30] D. Xu, S.Z. Xiang, Z.Y. Bai, J. Wei, M.L. Gao. Optimal multi-energy portfolio towards zero carbon data center buildings in the presence of proactive demand response programs. *Applied Energy*, 2023, 350, 121806. DOI: 10.1016/j.apenergy.2023.121806
 - [31] X.X. Lin, L.J. Zuo, L.K. Yin, W. Su, S.D. Ou. An idea to efficiently recover the waste heat of data centers by constructing an integrated system with carbon dioxide heat pump, mechanical subcooling cycle and lithium bromide-water absorption refrigeration cycle. *Energy Conversion and Management*, 2022, 256, 115398. DOI: 10.1016/j.enconman.2022.115398
 - [32] X.X. Chen, X.Y. Wang, T. Ding, Z. Li. Experimental research and energy saving analysis of an integrated data center cooling and waste heat recovery system. *Applied Energy*, 2023, 352, 121875. DOI: 10.1016/j.apenergy.2023.121875
 - [33] Z.P. Cui, S. Du, T.H. Zhao, Z.H. Chen, R.Z. Wang. High-power-density adsorption chiller driven by data center waste heat using encapsulated composite as adsorbent. *Energy*, 2024, 311, 133391. DOI: 10.1016/j.energy.2024.133391
 - [34] F. Zhou, X.W. Tian, Y. Song, G.Y. Ma. Dynamic performance of an integrated heat pump system coupled free cooling and waste heat recovery in data centers. *Energy*, 2025, 323, 135838. DOI: 10.1016/j.energy.2025.135838
 - [35] V. Dinesh Reddy, G.S.V.R.K. Rao, M. Aiello. Energy efficient resource management in data centers using imitation-based optimization. *Energy Informatics*, 2024, 7, 106. DOI: 10.1186/s42162-024-00370-y
 - [36] H. Shim. Quantifying job-level carbon efficiency in HPC: An empirical study based on the PM100 dataset. *Energy Informatics*, 2025, 8, 136. DOI: 10.1186/s42162-025-00586-6

- [37] H. Ruparel. Carbon-aware scheduling and distributionally robust optimization for cloud systems: Forecasting, battery storage, and adaptive ambiguity refinement. *Journal of Cloud Computing*, 2026, 15, 84. DOI: 10.1186/s13677-026-00904-7
- [38] J. Sun, M.C. Xu, M. Cespedes, M. Kauffman. Data center power system stability—Part I: Power supply impedance modeling. *CSEE Journal of Power and Energy Systems*, 2022, 8(2), 403-419. DOI: 10.17775/CSEEJPES.2021.02010
- [39] J. Sun, M. Mihret, M. Cespedes, D. Wong, M. Kauffman. Data center power system stability—Part II: System modeling and analysis. *CSEE Journal of Power and Energy Systems*, 2022, 8(2), 420-438. DOI: 10.17775/CSEEJPES.2021.02020
- [40] J. Sutaria, S.K. Rönnberg. Propagation of supraharmonics in a data center with different operating modes of UPS. *IEEE Access*, 2023, 11, 36823-36833. DOI: 10.1109/ACCESS.2023.3266092
- [41] Y.T. Chen, L.L. Luo, D.K. Guo, Q. He. Carbon-aware energy cost optimization of data analytics across geo-distributed data centers. *Journal of Computer Science and Technology*, 2025, 40(3), 654-670. DOI: 10.1007/s11390-025-4636-4
- [42] P. Colangelo, A.K. Coskun, J. Megrue, C. Roberts, S. Sengupta, V. Sivaram, et al. Turning AI data centers into grid-interactive assets: Results from a field demonstration in Phoenix, Arizona. *ArXiv*, 2026, DOI: 10.48550/arXiv.2507.00909
- [43] A. Caprara, Y. Yu, F. Teng, A. Junyent-Ferré, E. Bullich-Massagué, M. Aragüés-Peñalba. Data center workload flexibility for power system demand response: Evidence from Alibaba traces. *International Journal of Electrical Power and Energy Systems*, 2026, 178, 111940. DOI: 10.1016/j.ijepes.2026.111940