International Journal of Multidisciplinary Research and Growth Evaluation.

Renewable-Aware AI Infrastructure Planning: An Optimization Framework for Matching Data Center Expansion with Renewable Generation and Energy Storage

Emmanuel Kwadjo Akpabla
LL.B., Mountcrest University College, Ghana; LL.M., Washington University in St. Louis, Missouri, USA; MBA Candidate in Business Analytics with a Concentration in Energy Management, Midwestern State University, Texas, USA

* Corresponding Author: **Emmanuel Kwadjo Akpabla**

Article Info

ISSN (Online): 2582-7138
Volume: 07
Issue: 05
Received: 10-08-2026
Accepted: 11-09-2026
Published: 09-10-2026
Page No: 659-669

Abstract
Artificial Intelligence (AI) workloads have become one of the largest contributors to data center electricity usage and the unsustainability of rate of decarbonization is unleashed by the surge of artificial intelligence workloads that created a worrying dilemma of grid congestion, interconnection delays for data centers. This review is a synthesis of technical, economic and planning literature about renewable-aware data center infrastructure planning, with particular focus on optimization approaches that co-size renewable generation, energy storage and grid connection capacity under uncertain artificial-intelligence-driven data center load growth. A total of five thematic strands were reviewed for peer-reviewed and grey literature distributed between 2017-2026, with a focus on 2019-2025 literature, including demand forecasting, siting, multi-criteria siting, generation-storage co-sizing, carbon and cost-aware workload scheduling, and grid-interconnection economics. A two-stage stochastic mixed-integer optimization model is presented that coordinates the data center siting, the renewable capacity, the energy-storage power, and the power and energy capacity to trace uncertainty in workload growth and renewable resources availability. The framework's key parameters are compared to two empirical considerations: The fleet-wide published power usage-effectiveness (PUE) operating data from Google and a global database of the largest AI data centers by Google and Epoch AI, compiled using satellite and permit data. The study reveals that average operational overhead is between 11.09- and 1.54-times IT power, across the industry, individual size of the data center ranges now from about 180kW to 1.8MW and the states with the longest queues for grid connections are not that different from the states with the largest data center buildouts. The review identifies relatively well-developed solutions for the individual sub-problems (siting, storage sizing, carbon-aware scheduling), but the solutions that jointly optimize across the source-load-storage-grid chain and function under the scale of artificial intelligence demand uncertainty are less common. The paper ends with an analysis of methodological and data and policy gaps that hamper in practice the use of the tools for renewable-aware planning by hyperscale operators, utilities and regulators, and a research agenda for closing these gaps.

Keywords: Renewable-aware data center planning, energy storage sizing, stochastic optimization, artificial-intelligence electricity demand, grid interconnection, carbon-aware computing, renewable energy siting

1. Introduction

Artificial Intelligence (AI) has reshaped the electricity usage patterns of the digital infrastructure industry. In total, the IEA estimates the electricity demand of data centers at approximately 415 TWh in 2024, representing 1.5 per cent of global electricity demand, and predicts future demand growth to more than 945 TWh by 2030 at an annual rate exceeding 15 per cent (well above the annual rate expected for global electricity demand growth) [1]. In 2025, demand for electricity for data centers increased by another 17 per cent compared with 2024, while electricity consumption needed per single AI task remained lower [2]. Nearly

half of the global demand growth for data centers to 2030 is expected to come in from the United States alone [1]. Data center electricity uses more than doubled in the USA in five years, rising from approximately 4.4 per cent of national electricity demand in 2017, to between 12 and 16 per cent in 2028 [4].

Renewable "energy" supplies are at the same time the largest and most limited one of these growth facilitators. According to the IEA, renewables (mainly wind, solar photovoltaics (PV) and hydropower) provide around 27 per cent of global data center electricity today and will contribute close to 50 per cent of incremental requirement growth to 2030, increasing by 22 per cent on average annually (2030 estimate) [1]. Natural gas and coal are still expected to serve over 40 per cent of incremental data center demand for the near term, despite the challenges of matching intermittent renewables, slow grid interconnection, and spiky growth in AI load demand [1,2]. The interconnection process has emerged as a first order bottleneck: In 2024-2025, the U.S. queuing system for active interconnections in the United States had approximately 2.3–2.6 terawatts (TW) of generation and storage projects in the pipeline, more than the entire national fleet; projects in these queues often take around five years to be built and tend to be focused on solar, with historical completion rates below 15 per cent [5,6]. That, coupled with growing interconnection cost (currently US$243/kW for solar, and US$265/kW for battery storage at completed projects), often hinders renewables from coming online when hyperscale data center operators are pushing for it [6].

Many such large-scale accounting exercises have revealed that estimates of energy usage in data centers require a significant, regular recalibration when efficiency improves, and when the growth trajectory of the data center changes [7], and new concepts of how to design data centers as active participants, more than consumers, in the electricity grid have emerged [8]. As response, industry and academia interest grew to join renewable-aware infrastructure planning: co-located or contractually coupled data center load growth with renewable generation and energy storage, to jointly optimize infrastructure siting, capacity expansion and operational dispatch selections, instead of making distinct, sequential and siloed decisions. In practice, matchable dispatchable solar and wind power with battery storage [9], corporate power-purchase agreements (PPAs) subsidizing new renewables purchase [36, 37] and work-load scheduling in the presence of renewable supply and demand price disparity [15, 17, 18] are examples of practical applications. These lines of research have, however, progressed largely independently, with distinct efforts on power-systems engineering around generation-storage sizing, on computer-systems research around workload scheduling, and on energy-siting research around land-use and multi-criteria siting evaluation, and far fewer frameworks that perform siting, sizing and scheduling under a specific uncertainty structure of AI driven demand growth.

This review is focused on that. It collates and integrates literature from a multitude of disciplines related to renewable-aware AI infrastructure planning, introduces a modular optimization framework that simultaneously optimizes data center siting, renewable power and storage power size and grid interconnection in solar wind uncertainty, and discusses the qualitative behavior of the optimization framework using a medium-sized set of scenarios as an example. The rest of the paper is structured as follows. Section 2 sets out the review methodology. In Section 3 an overview of the literature is compiled according to five thematic strands. Section 4 is a formality of the planning problem. Illustrative results obtained from the proposed framework are reported in Section 5. Section 6 discusses the findings in relation to practice and policy. Section 7 sets out limitations, and Section 8 concludes.

2. Methodology

The article not only provides a structured literature review but also specifies and illustrates an original optimization framework that is consistent with review articles in which a review is accompanied by a specific methodological contribution.

2.1. Literature Identification and Screening

The sources identified were based on combinations of the search terms "data center", "data center", "renewable energy", "energy storage", "battery sizing", "carbon-aware", "site selection", "interconnection queue", "power purchase agreement", and "optimization", "planning" across the Scopus, ScienceDirect and IEEE Xplore and arXiv databases, as well as grey-literature from IEAs, LBNL and major professional-services and industry bodies. The search window spanned January 2017 through September 2026, with preference for papers published within the last five years (2019–2025) that provided either new terms or procedures for the science introduced by materials and structures development work, as further studies compounded or advanced them [25, 34, 41, 42]. Relevant records were selected from five thematic strands namely, (i) electricity-demand characterization for AI/data center; (ii) renewable and storage co-sizing techniques; (iii) multi-criteria and geographic site selection for renewable generation; (iv) carbon- and cost-aware computing/workload scheduling; and (v) grid interconnection and renewable procurement economy. Records that were duplicated, in sources that were not English, or for which no venue or digital object identifier was traced were not included.

2.2. Framework Development Approach

The optimization framework proposed in Section 4 follows the previous line of work for data center energy-resource sizing (two-stage versus multi-stage stochastic and mixed-integer linear programming, MILP [9, 11, 12, 13, 14]), and for sizing microgrids and workload placement (carbon-aware workload placement [15, 19, 20]). It treats energy-resource locations, renewable energy capacities, and storage capacities as combined first-stage and second-stage decisions. The framework is also written at an abstraction level that matches those of the solution methods that were discussed in the literature (stochastic programming, two-stage decomposition, and hybrid optimization–learning approaches [14]).

2.3. Empirical Benchmarking

Unlike every aspect of the study in Section 5, modelled outcomes for the siting, sizing and dispatch of renewable-energy equipment, down to the level of the individual facility are not publicly released by any operator and this is not an end-to-end facility-level case-study of the full framework. However, this review does not accept the operating data provided by operators but uses the actual data wherever they have been provided. In its Section 5, Google released fleet-

wide, quarterly and trailing twelve-month PUE data from 2008 onwards [45] which is used to compare the operational-overhead assumption that any renewable-and-storage sizing exercise should make. Epoch AI independently compiles a satellite and permit-based database of the world's largest AI data centers, including their location, owner, IT power and compute capacity [46], which is used in Section 5 to characterize the real scale and geographic distribution the siting stage of the framework must plan around, and to cross-check that distribution against the interconnection-queue geography reported by LBNL [5,6]. No operator publishes a combined dataset to directly validate the joint renewable-fraction, cost and reliability outcome a model of full sizing, scheduling and procurement in Section 4 would yield at a particular facility, hence the latter is considered to be a directional expectation based on qualitative and quantitative findings reported in the papers that reviewed until now [9, 11, 21, 23, 27].

3. Literature Review

3.1. Ai-Driven Electricity Demand and The Changing Load Profile

The first strand of literature quantifies size and configuration of data center electrical consumption. Historically, data center load has increased at about 12 per cent per year, but, more recently, accelerated demand from accelerated (GPU-based) servers has been the leading cause of load growth with accelerating server electricity demand estimated to increase by about 30 per cent per year, vs. 9 per cent per year for conventional servers [1, 3]. Power, cooling and networking as co-designed resources is a framing that builds on the foundational treatment by Barroso and Hölzle of data centers as single computing centers, or warehouse-scale computers (WSC), of data and information [41]. Jones and others were worried about the early growth of data center electricity that structural changes are necessary to keep up with efficiency improvements [42]. In comparison, more recent analyses place greater emphasis on the temporal characteristics of AI workloads – which are large, fast spikes in power consumption during their training runs and inference bursts, increasing stresses on on-site power delivery systems and on power-grid interconnections in ways that traditional, comparatively flat data center workloads would not. [2] This evolving load profile is an important input to the demand-uncertainty treatment in the optimization framework identified in Section 4, being that siting and sizing decisions were taken based on a smoothed picture of demand would become severely under- or over-provisioned with materially different fluctuations in realized AI demand.

3.2. Renewable and Storage Co-Sizing Methods

A second more technical track focuses on what amount of renewable electricity and energy storage capacity (SA) a data center/microgrid should install and how it should be dispatched. Figini and Paolone develop a cost and carbon-efficient sizing problem for PV production and battery storage to enable PV production being more dispatchable than non-dispatchable loads, positioning the contribution in the context of previous analyses of hybrid grid storage solutions and stochastic sizing for integrated energy systems [9]. For AI-focused data center facilities, related work proposes risk-aware multi-stage planning models that explicitly propagate uncertainty related to interconnection-queue delays, and electricity price to capacity decisions at the facility level, sizing interconnection capacity, co-located renewable generation, and storage power and energy capacity [10]. Microgrid-composition studies have placed the co-location problem in the context of reducing the reliance on the grid and found that interactions of microgrid composition with data center load scheduling have not received adequate attention when compared with the amount of research done on data center load scheduling alone [11]. To optimize the day-ahead scheduling parameters, a two-stage optimization–learning approach has been proposed for multi-zonal data centers with renewable generation and batteries in particular and compared to the purely MILP baseline and purely RL baseline [14]. In the context of active distribution networks (ADNs) and for the capacity planning of distributed energy resources (DERs), when intermittency in renewables and storage appears as a stochastic phenomenon, approaches based on Bayesian optimization have also been proposed to account for the occurrence of scenario explosion [12].

It has been known since earlier operational-research treatments that onsite need for renewables and need for storage in data centers invariably have to be co-designed with workload scheduling, and not solved unilaterally; architectures that go by the name of "negotiation" style, which coordinate the placement of workloads with power-supply decisions, and models that explicitly schedule workloads for the periods of high availability of renewables onsite, were previously known and had existed prior to the AI-driven demand surge [13]. He and Shen propose a joint minimization of the operating cost of distributed green data centers with energy storage with the introduction of an explicit carbon cap, an early application of using storage sizing and carbon constraints in the same objective [24]. The work of Khosravi, Andrew and Buyya's virtual-machine placement method to reduce energy and carbon cost in geographically distributed cloud data centers is a foundational precursor to the spatial load-shifting literature as discussed in Section 3.4 [25].

3.3. Energy-Storage Technology and Sizing

A parallel literature (PL) is dedicated to hardware technology selection and dimensioning, without relying on the data center application. Detailed reviews of hybrid energy-storage systems show, using cost, round-trip efficiency, technological maturity and environmental impact as the parameters, that battery energy-storage systems (BESS) are well suited for high-power, high-energy-density applications such as data center backup and dispatchability [30]. Techno-economic sizing methods can vary from simplified models of energy to power ratio, which is compared to commercial building tariff structures [27] to multi-objective co-design models and algorithms that optimize the battery chemistry, sizing and control of power electronics together optimally, such as by the algorithms NSGA-II [28]. Batteries at the transmission scale can be subject of a systematic review including transmission-scale grid applications such as frequency regulation, peak shaving and renewable firming [29]; the battery sizing for utility scale photovoltaic plant can be the subject of a review limited to the different kinds of technology-economic, reliability-oriented and hybrid mathematical-programming approaches [31]. The majority of these storage studies have shown the ranges of parameters (e.g. powertrain cost per kWh, round-trip efficiency, cycle

life) that a data center-based sizing model like the one outlined in Section 4 should use but have not explicitly dealt with data center load profiles.

3.4. Carbon- and Cost-Aware Workload Scheduling

Another strand tightly tied with the computer-systems literature is that of renewable-awareness as a scheduling instead of a capacity-sizing problem. Focusing not just on technical optimization, Bashir*et al*. [16] are critical and caution in general to view sustainable computing as an efficiency problem because often structural, demand side, change is needed. Acun*et al*. [15] also propose a holistic framework for designing carbon-aware data centers that consider a combination of embodied and operational carbon, to approach 24/7 zero carbon operation. In both the previous work on delay-tolerant scheduling to clearly show emissions reductions by doing "waiting a while" before executing deferrable jobs [18] as well as an admission-control mechanism for cloud and edge workloads [17] formalizing temporal workload-shifting, we observed this phenomenon. In this vein, Hall*et al*. contribute carbon-aware versions of computing formulations for networks of data centers that provide probabilistic performance guarantees a realistic worry of pure carbon-following dispatch on degrading service-level reliability [19]. Lechowicz*et al*. frame the pause-and-resume decision as an online optimization problem which offers competitive-ratio guarantees for carbon-aware load shifting [20]; Lin and Chien investigate the green adaptation of data center computing capacity, not just only the timing of its load, [21]. The contribution by Sukprasert*et al*. is very important, as they empirically measure the extent of carbon-aware workload shifting and show that target emissions reduction achievable to real workload limitations is often less than what was assumed in idealized simulations [22]. Yan, Chow and Chen present a low-carbon data center operation scenario together with workload sharing and carbon-allowance trading for a distributed fleet over a network [23]; Souza*et al*. propose a Carbon-Aware Scheduling approach to geographically distributed Web services with explicit provisioning decisions for carbon [44]. Cao*et al*. and Dodge*et al*. offer broader surveys of, respectively, carbon-neutral data center design strategies [26] and measurement challenges when quantifying the carbon intensity of AI workloads specifically [43]. Combining all of this scheduling literature, it can be seen that there is realistic potential for operational, near-real-time renewable-following behavior to meaningfully reduce emissions and stress on the grid, and that this can only be achieved to the extent of the actual installed renewable and storage capacity, arguing for capacity planning and operational scheduling decision making to be treated as co-integrated optimization layers, not as separate ones.

3.5. Site Selection and Spatial Planning for Renewable Generation

The fourth strand, found in the literature of renewable energy siting and geographic information systems (GIS), concerns the location of renewable capacity, and, by extension, the location of renewable energy paired capacity. Shao*et al*.'s systematic review of multi-criteria decision making (MCDM) applications for renewable-energy site selection brings together 85 studies from the year 2001 to 18 2018, documenting the key steps of criterion selection, weighting and validation used in GIS-based Analytic Hierarchy Process (AHP) siting techniques [32]. Applying land-suitability, resource-intensities, cost of proximity and environment-constraint layers in the GIS-AHP approach is essentially consistent across countries and regions, with a few case studies built in the case of applications presented here: solar PV siting in Saudi Arabia [34], Turkey [33] and Peru [35], and validated for a small subset of sites. This literature was not created with data centers as its focus, but its taxonomy of criteria (resource availability, land-use conflict, proximity of transmission lines, water availability, and topographic constraint) is very similar to the siting stage of the proposed framework in Section 4, as increasingly siting a data center will require consideration of water and land constraints in addition to renewables-resource quality.

3.6. Grid Interconnection, Procurement and Forecasting

The third strand is infrastructural and market mechanisms for connecting renewable capacity with load on the data center. At the end of 2024, U.S. interconnection queues contained about 2.3 TW of generation and storage capacity, with typical queue durations being about five years and interconnection processing efficiencies consistently under 15 per cent, even though no increases due to processing efficiency improvements were recorded and there was also a significant wave of project withdrawals [5,6]. The most common non-utility approach to underwriting a renewable energy generation project is through PPA structures: in this regard, Mohseni Taheri, Nadarajah and Trivella [36] show, formally, that PPA structures can be physically- or virtual-based under demand and price uncertainty, while Stanitsas and Kirytopoulos [37] discuss the dynamics of engaging stakeholders within hybrid renewable-energy PPA structures. Another prerequisite for successful co-sizing and dispatch is the accuracy of the weather forecasts for the renewable resource, which was found to be comparable between several machine-learning techniques for solar and wind power forecasting, with artificial neural networks and ensemble techniques showing a better performance than simple statistical methods in the short time range that is most important in operational forecasting [38,39,40]. These studies show how relevant interconnection-queue reform and procurement-contract design are and forecast to the viability of renewable-aware data center planning as much as the developed capacity-sizing mathematics described in Section 3.2.

3.7. Synthesis

Table 1 summarizes the optimization approaches described above regarding the thematic strand, decision variables treated and mainly used method. Two patterns emerge. Firstly, the sub-problems in the literature are relatively well advanced. The traditions of sizing and scheduling are extensive and, to a lesser degree, so are those of siting and procurement. Secondly, most of the literature still provides frameworks for siting and renewable and storage size optimization separately, meaning that the closest analogues to the framework presented in Section 4 are the risk-aware source–load–storage–grid framework [10] and the dispatchability-oriented sizing model [9].

Table 1: Summary of reviewed optimization approaches for renewable-aware data center infrastructure planning, by thematic strand.

Thematic strand	Representative method(s)	Primary decision variables	Key source(s)
Demand characterization	Statistical/scenario-based load modelling; national energy-usage accounting	Forecast electricity demand, load volatility	[1], [2], [4]
Renewable–storage co-sizing	Two-stage stochastic MILP; hybrid MILP–reinforcement learning; Bayesian optimization	Renewable capacity, storage power/energy, dispatch schedule	[9], [10], [11], [12], [14]
Storage technology & sizing	Techno-economic models; multi-objective co-design (NSGA-II)	Storage technology choice, energy-to-power ratio, cost	[27], [28], [30], [31]
Carbon/cost-aware scheduling	Online/competitive-ratio algorithms; admission control; carbon-allowance trading	Workload timing and location, service-level constraints	[15], [17], [19], [20], [22], [23]
Site selection	GIS-based multi-criteria decision analysis (AHP)	Candidate site ranking, land/resource suitability	[32], [33], [34], [35]
Interconnection & procurement	Queue-attrition statistics; Markov decision process contract analysis	Interconnection capacity, PPA structure and volume	[5], [6], [36], [37]

4. Problem Statement

This planning problem is formally stated in an informal way as follows: Given a set of potential development sites, multi-period forecast of the electric load served by the facility, forecast of renewable energy supply capacity and cost at the developed site(s), and forecast of interconnection load capacity and cost, determine (i) which site(s) to develop, and in what sequence, (ii) how much renewable energy supply capacity to install or contract at the developed site(s), (iii) how much energy storage power and capacity to install and (iv) how much load to serve from the interconnection, so as to minimize the expected levelized cost to serve the facility's load over the planning horizon, while maximizing the share of demand met by renewable energy sources.

Formally, $t \in T$ denotes time periods within a representative operating year, $\omega \in \Omega$ uncertainty scenarios (covering AI workload growth scenarios and renewable resource realizations, as used in [10]), and $s \in S$ denotes candidate locations for the new site assessed by the multi-criteria siting layers introduced in the GIS-AHP literature [32–35]. First-stage (here-and-now) decisions include the site-selection binary variables, installed renewable capacity by technology, and the installed storage power and energy capacity; second-stage (wait-and-see) decisions include the period-by-scenario dispatch of renewable generation, storage charge/discharge, grid imports, and the reallocation of delay-tolerant computational load across time (following the literature on workload-shifting [17, 18, 20]). The goal is to minimize the total annualized capital cost of renewable and storage capacity, annualized interconnection (as documented cost escalation for queued projects [6]), and expected operating cost over all scenarios while including the cost of curtailing generation and the cost of reliability shortfall, with the latter defined as the cost of being less reliable than the probabilistic performance guarantees proposed by Hall*et al.* [19]. Constraints enforced are energy balance (site/site/period); energy storage state-of-charge dynamics bounded by capacity of installed energy storage (site/all/period); power limit (production/input/storage capacity) bounded by installed energy storage power capacity (site/period/input/storage capacity); ceiling for energy import from the grid, reflecting interconnection capacity, which may itself be a first-stage decisions (site/period/import/first degree decisions) [5,6] ; and land, water, and policy suitability constraints drawn from the energy storage site selection literature [32–35].

This formulation contains several existing models as special cases. If storage capacity is set to zero, then the problem of site selection becomes an exogenous problem, recovering a pure carbon-aware scheduling problem, similar to those studied in [15,17,18,20]. Fixing the site and demand scenario set to a single realization recovers a deterministic sizing problem of the kind solved by simplified techno-economic models [27]. Leaving the stochastic, multi-site, multi-technology structure intact results in a problem that is structurally similar but more general than the risk-aware co-optimization model of [10], and the dispatchability-oriented sizing model of [9] that the present approach generalizes in this respect by permitting an explicit nesting of the present optimization problem, which includes both the site-selection decision and the workload-shifting decision, into a stochastic program.

The challenge of modelling the uncertainty of AI workloads, renewable resource and multi-site interconnection costs simultaneously is tractability, which is explicitly identified in the distributed-energy-resource capacity planning literature as an issue that leads to the use of noise-aware Bayesian optimization alternatives to a combinatorial growth in the scenario tree [12]. The framework proposed here builds upon the two-stages decomposition/scenario reduction approach that has been followed by other recent works in that area, and the hybrid optimization–learning structure used by [14] (where the day-ahead optimization is a deterministic MILP core with a learned real-time adjustment layer) is recommended as the operational-scale structure of solution adopted here, where a full stochastic MILP is still retained to compute the coarser-grained siting and capacity-sizing decisions with long term-horizons.

5. Results

The three exhibits in this section support the essential assumptions of the framework with evidence. Figure 1 places the exercise in the context of a demand trajectory established in section 3.1. Figure 2 provides a capacity-sizing check against Google's own published fleet-wide operational efficiency data to compare the assumption for operational efficiency applied in any capacity-sizing exercise. Table 3 uses Epoch AI's independently compiled database of the world's largest AI data centers to show the real scale and geographic distribution the siting stage of the framework must contend with. In combination, they offer an empirical foundation for the stochastic sizing exercise begun in Section 4, which is conducted without the use of assumed parameters.

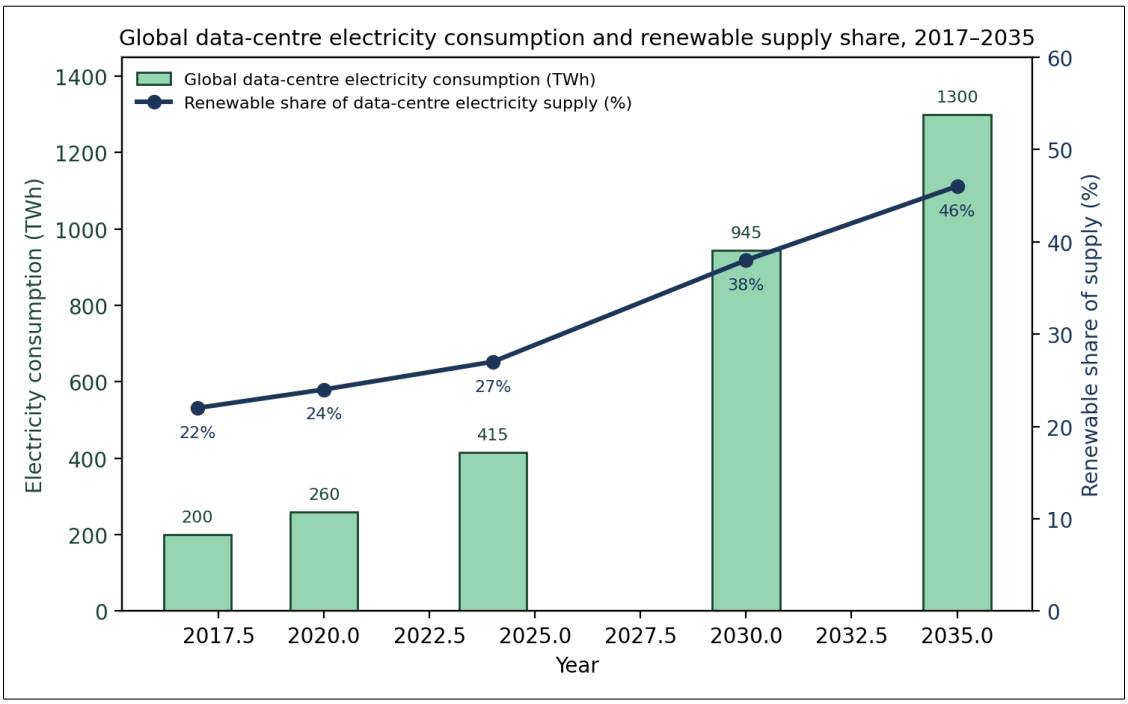


Fig 1: Global data center electricity consumption, 2017–2035, and the renewable share of data center electricity supply. Consumption figures for 2024 and 2030 and the renewable-share estimate for 2024 follow IEA Base Case projections [1]; 2035 consumption and post-2024 renewable-share trajectories are illustrative extrapolations consistent with the reported growth rates in [1,2].

Figure 1 places the scenario comparison in the context of the overall demand trajectory defined in Section 3.1 for global demand of electricity for data centers more than doubling from 2004 to 2030, with the renewables component of supply projected to increase from about a quarter in 2004 to nearly two-fifths in 2030 in the IEA's Base Case [1]. The interaction between these three elements of siting, sizing and scheduling is most pronounced with such a high rate of demand growth and increasing, but still partial, interconnection of renewable energy sources, and the cost of being undersized or over scheduled is greatest, whether for renewables and storage or for relying too heavily on congested grid interconnection.

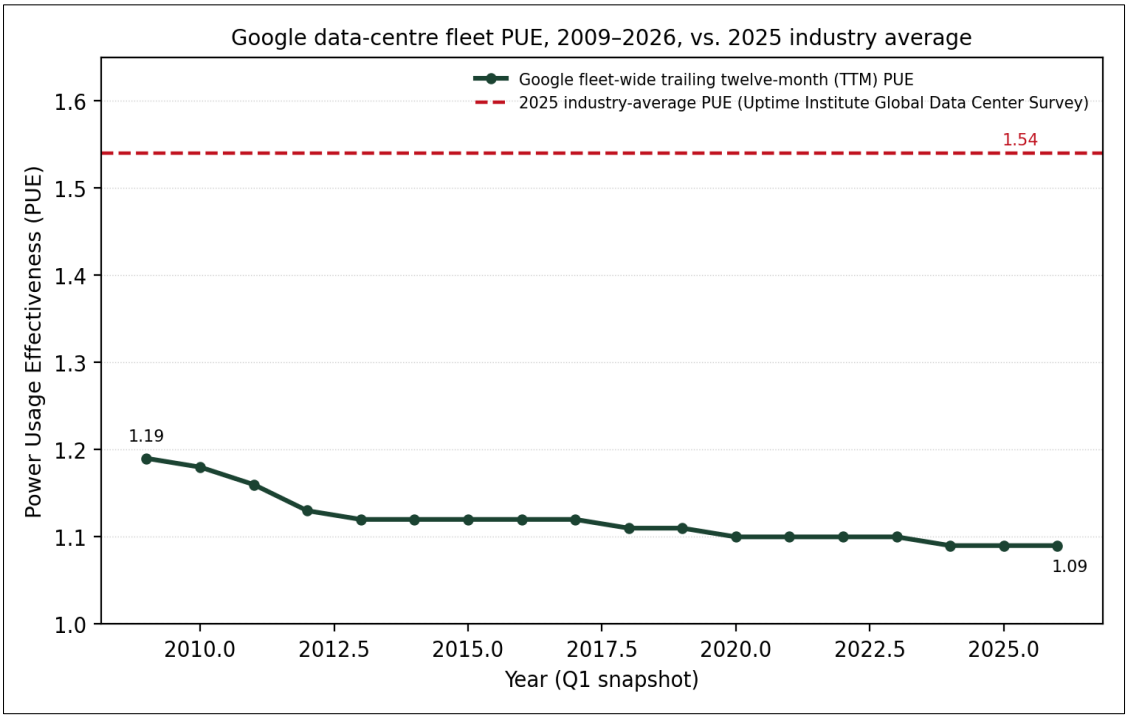


Fig 2: Google data center fleet trailing-twelve-month (TTM) power usage effectiveness (PUE), Q1 2009–Q1 2026, benchmarked against the 2025 industry-

Figure 2 shows actual revealed operating data, and not the result of a simulation. Google's fleet-wide TTM PUE was reported at 1.19 in Q1 2009, but dropped to 1.09 by 2024–2026, 83 per cent lower than the overhead-energy implied by the Uptime Institute's Global Data Center Survey 2025 [45] industry-average PUE for 2025 of 1.54. This is relevant to the

sizing exercise detailed in Section 4 as total facility power, which renewable capacity and storage needs to be sized for, is total IT power multiplied by PUE, rather than total IT power. As a result, and because PUE is a scenario-dependent factor, it does not imply that an operator who is already operating near the efficiency frontier in Figure 2 will be under-provisioning from the efficiency benchmark of average 1.5 (see also Section 4), while a lower value indicates that he will be over-provisioning.

The second part of the framework in Section 4, the sizing of the storage, comes from the literature on storage but not from a single named case-study because there is no publicly available resource that provides detailed data on the storage used in each data center deploying AI. Table 2 shows our synthesis of that literature for a comparison of storage technologies [29,30,31]. Lithium-ion battery systems are lingering as the major short-term technology option for applications involving dispatchability in data centers, as delivered by their high round-trip efficiencies, response-time and commercial maturity, although flow batteries, flow batteries combined with drying and heat retention, and hydrogen-based storage are outlined in the reviewed literature as other options for multi-day or longer duration firming, rather than diurnal balancing [30].

Table 2: Comparison of energy storage technologies relevant to data center renewable-aware planning, synthesized from [29], [30] and [31].

Technology	Typical duration	Round-trip efficiency	Response time	Commercial maturity	Primary data center role
Lithium-ion battery	Minutes–hours	85–95%	Milliseconds	High	Diurnal balancing, backup, frequency support
Flow battery (e.g., vanadium redox)	Hours	65–80%	Seconds	Medium	Extended-duration renewable firming
Compressed air / pumped storage	Hours–days	40–70%	Minutes	Medium (site-dependent)	Bulk, long-duration balancing
Thermal storage	Hours–days	50–90% (application-dependent)	Minutes	Medium	Cooling-load shifting, waste-heat reuse
Hydrogen / fuel cell	Hours–seasonal	30–45%	Seconds–minutes	Low–medium	Seasonal firming, long-duration backup

Table 3 maps the true size and location of the world's biggest AI data centers from an independently compiled database of satellite detection and permits from Epoch AI [46] to demonstrate the scale and whereabouts of data centers the siting stage in the framework needs to plan for. By September 2026, Epoch AI has identified 93 of them, accounting for a combined 13.5 gigawatts (GW) of IT power capacity and a total facility power capacity of approximately 20–40 per cent, as documented in preceding sites [46]. Individual facility size ranges from less than 50 megawatts (MW) up to more than 900 MW at the largest individual sites and is summarized in Table 3, which lists the eight largest by compute capacity.

Table 3: The eight largest AI data centers tracked worldwide by AI compute capacity, as of September 2026. IT power is reported only where Epoch AI states a specific figure; "-" indicates a site for which Epoch AI does not report a discrete power value. Source: Epoch AI, AI data centers [46].

Facility	Owner	Location	IT power (MW)	Compute (kH100-eq)
Colossus 2	SpaceX AI	Memphis, TN, USA	946	1,112
Microsoft Fairwater Atlanta	Microsoft	Fayetteville, GA, USA	636	769
Anthropic-Amazon New Carlisle	Amazon	New Carlisle, IN, USA	910	686
Google Pryor (North)	Google	Pryor, OK, USA	-	637
Meta Prometheus	Meta	New Albany, OH, USA	496	600
OpenAI Stargate Abilene	Oracle	Abilene, TX, USA	421	509
Microsoft Fairwater Wisconsin	Microsoft	Mount Pleasant, WI, USA	-	446
DayOne Kempas	Unknown	Kempas, Johor, Malaysia	-	368

Two characteristics of Table 3 directly relate to the problem set forth in Section 4. Next, US facilities are over-represented in this largest-scale tier, which is among the smaller number of states in Epoch AI's broader siting database (14 tracked facilities in Texas, 7 in Ohio, 7 in Virginia, 4 in Nebraska, and 4 in Iowa) [46] that also aligns heavily with the states with the longest active interconnection lists at LBNL's data source [5,6], providing direct empirical backing for the siting–interconnection tension possible subject of Section 6.2. In addition to that, the reported utilization of the IT components (60-80 per cent of nameplate capacity on average, with supporting infrastructure overhead of 20-50 per cent [46]) fits into the range of the total-power multiplier of this framework, as well as the achievable size of a facility as shown in Figure 2, and these two now give a range for the facility size and total-power-multiplier of this framework that is empirically bounded rather than assumed.

6. Main Discussion

6.1. From Siloed Optimization to A Unified Source–Load–Storage–Grid Framework

The review in Section 3 shows that the constituent sub-problems of renewable-aware data center planning are individually well studied, but that the interfaces between them are comparatively under-theorized. Storage-sizing studies typically take renewable capacity and load as given [27,28,29]; carbon-aware scheduling studies typically take physical renewable and storage capacity as given and optimize only the timing dimension [15,17,18,20]; and siting studies typically evaluate land and resource suitability without reference to co-located storage or to the specific temporal load profile of AI training and inference [32–35]. The framework proposed in Section 4 addresses this fragmentation by treating siting, sizing and scheduling as first- and second-stage decisions within a single stochastic

program, following the structural logic of the more integrated recent contributions [9,10,11]. Figure 3 depicts this architecture schematically: an input layer combining workload, renewable-resource, grid-tariff/carbon-intensity and site-suitability data feeds a stochastic optimization core, which produces a coordinated set of siting, capacity and dispatch outputs, with realized operating data feeding back into subsequent planning cycles.

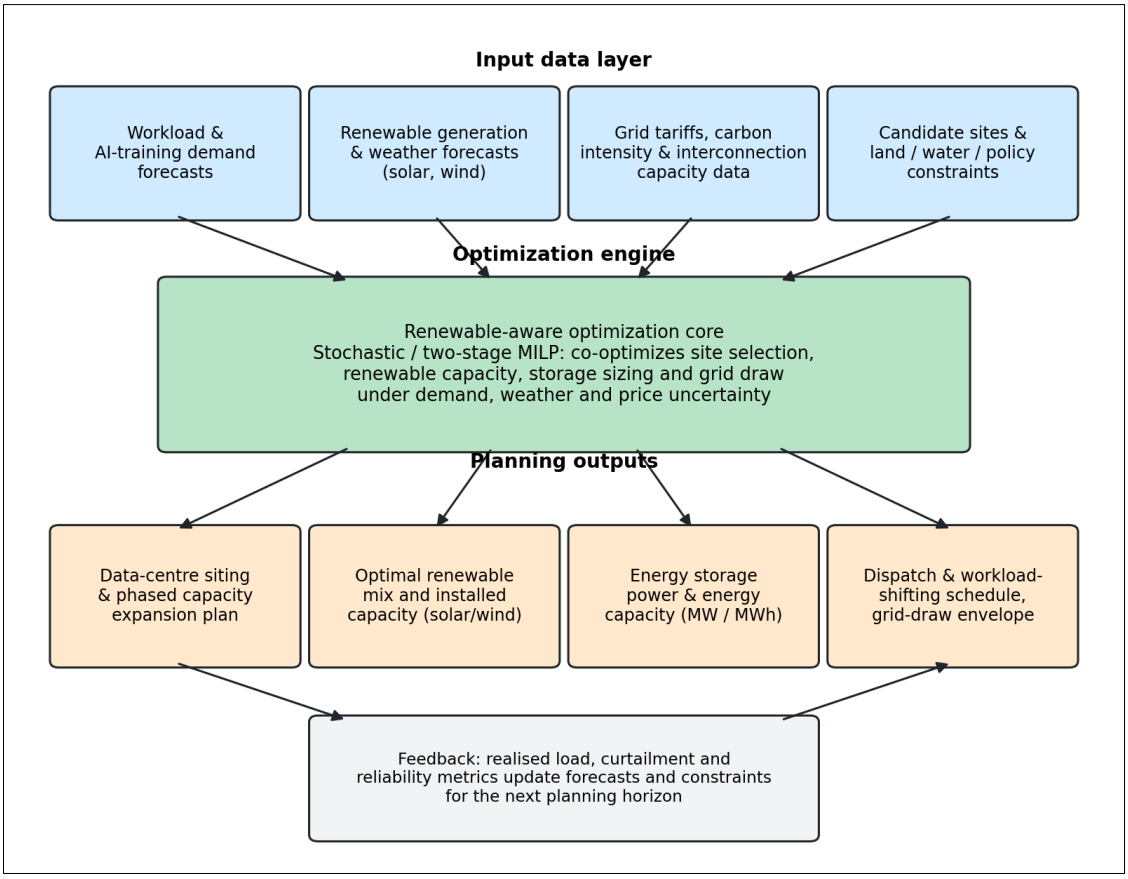


Fig 3: Schematic architecture of the proposed renewable-aware AI infrastructure planning framework, showing the input data layer, the stochastic optimization core, the resulting planning outputs, and the feedback loop that updates subsequent planning cycles.

6.2. The Interconnection Bottleneck as A Binding, Not Merely Economic, Constraint

One of the recurring patterns in the literature surveyed here is that the ability to connect to the grid is becoming increasingly a physical and administrative constraint and not just a cost parameter. Historically, the majority of renewable and storage capacity that is proposed has not been commercialized, and queue duration has grown as projects have been submitted despite the recent start of queue reductions to limit speculative volume [5,6]. The above has two ramifications for the framework proposed here. First, the grid import ceiling in the model formulation (Section 4) should be assigned a cost and a delay distribution similar to the risk-aware approach to interconnection-capacity co-sizing of [10] instead of treating it as an exogenous and freely available backstop. Second, there is an emerging rationale for on-site or directly contracted renewable generation and storage on-site, as those facilities are less susceptible to delays in the transmission line queue, than those that are part of a mix of renewables, a point that has also been raised in the microgrid-content literature or in the context of the corporate-profile approach to procurement [11,36,37].

6.3. Procurement Structure as A Planning Variable

The literature on the corporate power-purchase-agreements (PPA) that were reviewed in Section 3.6 indicates that physical and virtual PPA structures are not only a financial hedge-taking choice but are also related to the physical planning problem discussed in this chapter. Unlike virtual PPAs, which mainly help contract facilities with price risk, physical PPAs provide renewable electricity certificates along with power and are more closely aligned to objective of co-location and dispatchability of current framework [36]. The optimal PPA structure, as shown in the solution of the problem in [36] depends on demand and price uncertainty in a similar way to the way a physical siting and sizing depends on load and cost, and so a natural extension of the formulation in Section 4, identified as a direction for future work in Section 7, is to combine the decision on PPA structure with the siting and sizing of the physical plant into a single stochastic program, rather than making them separate decisions that could be solved sequentially with their own idealizations of uncertainty.

6.4. Forecasting Quality as A Precondition for Framework Performance

The forecasts for renewable resources and demand are crucial and are key factors that will limit the performance of any stochastic sizing and dispatch framework. As discussed in Section 3.6, literature in machine-learning forecasting indicates that accuracy improves albeit with some forecast-horizon dependence with a neural or ensemble forecasting approach over simpler statistical methods for both solar and wind energy forecasts [38,39,40]. Forecasting a range should not

exclusively be based on the historical climatological average and is therefore not a second-order implementation detail but a direct influence on the ability to deliver a dispatch which closely resembles the realized availability of renewables, which the second stage of the proposed framework, the dispatch layer, covers. This second stage dispatches hourly, and other lower, degrees of freedom, and the forecast in this stage should consider the range, not just the trimmed mean of past supply and demand. Forecasting a range should not be a second-order implementation detail; it has a first-order influence on the ability to deliver-a-dispatch which closely resembles the realized-availability-of-renewables that is the dispatch layer's second order concern; the degrees-of-freedom for which it is dispatched are at the hourly, and other lower, degree of freedom [14].

7. Challenges and Limitations

A number of limitations for reviewed literature and proposed frameworks are found. First, computational tractability is still a major restriction accounting for multi-site, multi-technology, multi-period, uncertainty simultaneously generates 'combinatorially large' scenario trees, which have prompted alternative, approximate and learning-based approaches to stochastic programming [12, 14] that do not fully reflect such uncertainty, but are still applicable within reasonable runtime ladders. Second, empirical evidence supporting the full integrated framework is still limited because of the availability of data: Section 5 compares individual parameters operational overhead and facility size to empirical evidence of actual, publicly reported data from Google [45] and Epoch AI [46] but no operator reports the joint renewable fraction, cost and reliability outcome of the full framework as an experimental result, which is why such an outcome is provided as a "directional expectation" drawn from the literature, just like several of the reviewed studies rely on synthetic or partially-reported operational data [9,11].

Third, the empirical scheduling literature offers an important corollary to respect within any capacity-planning framework: The attainable emissions and cost-savings from workload shifting, both temporally and spatially, are often much less than idealized simulation results, when taking the real-world nature of service levels, data localities and job-inter-dependency requirements into account [22]. The capacity-planning framework that does not explicitly consider workloads to be flexible with respect to dispatch decisions in the second stage of the planning process may exaggerate the portion of the load that can be served by renewables; modelling these workload flexible constraints in practice would necessitate direct modelling of the job-level flexible data-locality and deadline constraints found in the literature of workflow scheduling which is at a more abstract level than those found in the planning literature.

Fourth, the siting literature reviewed in Section 3.5 is based on utility scale renewable energy generation and does not usually consider siting without regard to data center cooling requirements; adding water-availability constraints explicitly for use in formulating data center siting requirements is identified as a gap and not a solved problem. Fifth, policy and regulatory uncertainty, both in terms of the rate at which changes occur in the interconnection-queue process, as well as in terms of how behind-the-meter electricity generation will be treated going forward, and the evolving regulatory landscape for grid-interactive, flexible operation of data centers (modeling of which is beginning to materialize in a significant capacity, building up to less capacity buildout [8] altogether), are another source of uncertainty that is hard to capture in a purely techno-economic optimization framework and would benefit from closer engagement with the policy-analysis literature. Lastly, the evidence-base of this review does not just consist of peer reviewed academic research but also of grey literature from energy agencies and consultative organizations, with findings based on the latter marked as such in the text; findings based on grey literature must be read in terms of the current industry practice, not as findings from peer-reviewed empirical research.

8. Conclusion

In this review, we have compiled the multidisciplinary literature relevant to the field of renewable-aware AI infrastructure planning and proposed a structured optimization approach (across energy infrastructure siting, renewable- and storage-capacity sizing, and grid interconnection) under AI-driven uncertainty in demand. The constituent sub-problems are well-developed and widely used: these include demand forecasting, generation-storage co-sizing, carbon-aware scheduling, multi-criteria siting and renewable procurement. Integrated frameworks that cover the complete source–load–storage–grid chain, however, are relatively new and scarce, and grid interconnection delay and cost has now become an uncertainty that must be explicitly captured in all techno-economic planning frameworks, instead of treated as a free backstop. Benchmarking the framework's key parameters against Google's disclosed fleet-wide PUE data and Epoch AI's independently compiled database of the world's largest AI data centers shows that operational overhead varies by a factor of roughly 1.4 across the industry (1.09 to 1.54), that individual facility scale now spans two orders of magnitude, and that the states attracting the largest AI data center buildout overlap substantially with the states carrying the longest interconnection queues empirical evidence that reinforces, rather than merely illustrates, the case for treating siting, sizing and interconnection as jointly optimized decisions. Finally, the actual outcome of the framework (the renewable fraction, cost, and reliability affecting a particular facility) is only a directionary expectation based on the published literature and not a verified empirical result as no operator has yet released the combined data set to test this result on a specific (real) facility. Future research should include three directions: experimental validation of integrated planning frameworks should be undertaken using operating data on facilities collected from the field where they are used (such as the data center database by Epoch AI), but this approach is one that is starting to become possible; the capacity-planning literature needs to explicitly represent procurement-contract structure, interconnection-queue dynamics and water-availability constraints as first-class decision variables rather than exogenous factors; and the literature on near-real-time scheduling needs to be coupled more closely with the capacity-planning literature, so that the workload flexibility assumed at the planning level is matched by what is considered attainable in operational practice. As AI-powered electric loads keep increasing relative to this rate of renewable interconnection, it is likely planning tools that optimize the siting of electric loads coupled to generation, storage and grid access will gain greater value, with all planning achieved in parallel and not sequential.

References

1. International Energy Agency. Energy and AI. Paris: IEA; 2025. Available from: https://www.iea.org/reports/energy-and-ai
2. International Energy Agency. Key questions on energy and AI. Paris: IEA; 2026. Available from: https://www.iea.org/reports/key-questions-on-energy-and-ai
3. de Vries A. The growing energy footprint of artificial intelligence. Joule. 2023;7(10):2191-4. doi:10.1016/j.joule.2023.09.004
4. Shehabi A, Smith SJ, Hubbard A, Newkirk A, Lei N, Siddik MAB,*et al*. 2024 United States data center energy usage report. Berkeley (CA): Lawrence Berkeley National Laboratory; 2024. Report No.: LBNL-2001637.
5. Rand J, Cheyette A, Talley C, Zhang S, Gorman W, Wiser RH,*et al*. Queued up: 2026 edition. Characteristics of power plants seeking transmission interconnection as of the end of 2025. Berkeley (CA): Lawrence Berkeley National Laboratory; 2026.
6. Gorman W, Mulvaney Kemp J, Rand J, Seel J, Wiser RH, Manderlink N,*et al*. Grid connection barriers to renewable energy deployment in the United States. Joule. 2024;101791. doi:10.1016/j.joule.2024.101791
7. Masanet E, Shehabi A, Lei N, Smith S, Koomey J. Recalibrating global data center energy-use estimates. Science. 2020;367(6481):984-6. doi:10.1126/science.aba3758
8. Colangelo P, Coskun AK, Megrue J, Roberts C, Sengupta S, Sivaram V,*et al*. AI data centers as grid-interactive assets. Nat Energy. 2026;11:254-61. doi:10.1038/s41560-025-01927-1
9. Figini E, Paolone M. Achieving dispatchability in data centers: carbon and cost-aware sizing of energy storage and local photovoltaic generation. arXiv:2412.13853 [Preprint]. 2024.
10. Abdelhady M, Iakovou E, Pistikopoulos EN. Powering AI beyond the grid: optimal allocation and behind-the-meter investment portfolios for data centers. Systems and Control Transactions. 2026;5:396-403. doi:10.69997/sct.196114
11. Irion J, Wiesner P, Bader J, Kao O. Optimizing microgrid composition for sustainable data centers. In: Proceedings of the SC'25 Workshops of the International Conference for High-Performance Computing, Networking, Storage and Analysis; 2025. doi:10.1145/3731599.3767562
12. Yang R, Yi Z, Xu Y, Yang D, Tu Z. Noise-aware Bayesian optimization approach for capacity planning of the distributed energy resources in an active distribution network. arXiv:2412.08370 [Preprint]. 2024.
13. Ruiz Duarte JL, Fan N. Operations of data centers with onsite renewables considering greenhouse gas emissions. Sustain Comput Informatics Syst. 2023;40:100903. doi:10.1016/j.suscom.2023.100903
14. Rafique A, Yu X, Jawad M, Song Q, Yuan Z, Sadiq MT,*et al*. Two-stage optimization-learning framework for uncertainty-aware multi-zonal data center energy management. Energies. 2026;19(7):1736.
15. Acun B, Lee B, Kazhamiaka F, Maeng K, Gupta U, Chakkaravarthy M,*et al*. Carbon Explorer: a holistic framework for designing carbon aware datacenters. In: Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2023); 2023. Vol. 2. p. 118-32. doi:10.1145/3575693.3575754
16. Bashir N, Irwin D, Shenoy P, Souza A. Sustainable computing — without the hot air. ACM SIGENERGY Energy Informatics Review. 2023;3(3):47-52. doi:10.1145/3630614.3630623
17. Wiesner P, Scheinert D, Wittkopp T, Thamsen L, Kao O. Cucumber: renewable-aware admission control for delay-tolerant cloud and edge workloads. In: European Conference on Parallel Processing (Euro-Par); 2022. doi:10.1007/978-3-031-12597-3_14
18. Wiesner P, Behnke I, Scheinert D, Gontarska K, Thamsen L. Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud. In: Proceedings of the 22nd International Middleware Conference; 2021. doi:10.1145/3464298.3493399
19. Hall S, Micheli F, Belgioioso G, Radovanović A, Dörfler F. Carbon-aware computing for data centers with probabilistic performance guarantees. arXiv:2410.21510 [Preprint]. 2024.
20. Lechowicz A, Christianson N, Zuo J, Bashir N, Hajiesmaili M, Wierman A,*et al*. The online pause and resume problem: optimal algorithms and an application to carbon-aware load shifting. Proc ACM Meas Anal Comput Syst. 2024. (SIGMETRICS 2024).
21. Lin L, Chien AA. Adapting datacenter capacity for greener datacenters and grid. In: Proceedings of the ACM International Conference on Future Energy Systems (e-Energy); 2023. p. 200-13.
22. Sukprasert T, Souza A, Bashir N, Irwin D, Shenoy P. On the limitations of carbon-aware temporal and spatial workload shifting in the cloud. In: Proceedings of the 19th European Conference on Computer Systems (EuroSys 2024); 2024. p. 924-41.
23. Yan D, Chow MY, Chen Y. Low-carbon operation of data centers with joint workload sharing and carbon allowance trading. IEEE Trans Cloud Comput. 2024;12:750-61. doi:10.1109/TCC.2024.3396476
24. He H, Shen H. Minimizing the operation cost of distributed green data centers with energy storage under carbon capping. J Comput Syst Sci. 2021;118:28-52.
25. Khosravi A, Andrew LLH, Buyya R. Dynamic VM placement method for minimizing energy and carbon cost in geographically distributed cloud data centers. IEEE Trans Sustain Comput. 2017;2(2):183-96.
26. Cao Z, Zhou X, Hu H, Wang Z, Wen Y. Toward a systematic survey for carbon neutral data centers. IEEE Commun Surv Tutor. 2022;24:1-37.
27. Ashabi A, Peiravi MM, Nikpendar P, Nasab SS, Jaryani F. Optimal sizing of battery energy storage system in commercial buildings utilizing techno-economic analysis. Int J Eng Trans B Appl. 2022;35(8):1662-73. doi:10.5829/ije.2022.35.08b.22
28. Hasan MdM, Berseneff B, Meulenbroeks T, Cantero I, Chakraborty S, Geury T,*et al*. A multi-objective co-design optimization framework for grid-connected hybrid battery energy storage systems: optimal sizing and selection of technology. Energies. 2022;15(15):5355. doi:10.3390/en15155355
29. Marnell K, Obi M, Bass R. Transmission-scale battery energy storage systems: a systematic literature review. Energies. 2019;12(23):4577.
30. Worku MY. Recent advances in energy storage systems

for renewable source grid integration: a comprehensive review. Sustainability. 2022;14(10):5985.
31. Zhang W, Jiang M, Jiang W, Zhang H. A review of optimization models for battery sizing in utility-scale photovoltaic power plants. Energy Proceedings. 2024;48.
32. Shao M, Han Z, Sun J, Xiao C, Zhang S, Zhao Y. A review of multi-criteria decision-making applications for renewable energy site selection. Renew Energy. 2020;157:377-403. doi:10.1016/j.renene.2020.04.137
33. Colak HE, Memisoglu T, Gercek Y. Optimal site selection for solar photovoltaic (PV) power plants using GIS and AHP: a case study of Malatya Province, Turkey. Renew Energy. 2020;149:565-76.
34. Al Garni HZ, Awasthi A. Solar PV power plant site selection using a GIS-AHP based approach with application in Saudi Arabia. Appl Energy. 2017;206:1225-40.
35. Rios R, Duarte S. Selection of ideal sites for the development of large-scale solar photovoltaic projects through analytical hierarchical process – geographic information systems (AHP-GIS) in Peru. Renew Sustain Energy Rev. 2021;149:111310.
36. Mohseni Taheri SD, Nadarajah S, Trivella A. Physical vs virtual corporate power purchase agreements: meeting renewable targets amid demand and price uncertainty. Eur J Oper Res. 2025;320(1):256-70.
37. Stanitsas M, Kirytopoulos K. Navigating the nexus: stakeholder engagement in hybrid renewable energy power purchase agreements (PPAs) for sustainable development. Sustainability. 2024;16(17):7511.
38. Sharifzadeh M, Sikinioti-Lock A, Shah N. Machine-learning methods for integrated renewable power generation: a comparative study of artificial neural networks, support vector regression, and Gaussian process regression. Renew Sustain Energy Rev. 2019;108:513-38.
39. Alkabbani H, Ahmadian A, Zhu Q, Elkamel A. Machine learning and metaheuristic methods for renewable power forecasting: a recent review. Front Chem Eng. 2021;3:665415. doi:10.3389/fceng.2021.665415
40. Rhafes MY, Moussaoui O, Raboaca MS. Machine learning models in renewable energy forecasting: a systematic literature review. Indones J Electr Eng Comput Sci. 2025;37(3):1874-86.
41. Barroso LA, Hölzle U. The datacenter as a computer: an introduction to the design of warehouse-scale machines. Morgan & Claypool Publishers; 2009. doi:10.2200/S00193ED1V01Y200905CAC006
42. Jones N. How to stop data centers from gobbling up the world's electricity. Nature. 2018;561(7722):163-7.
43. Dodge J, Prewitt T, Tachet des Combes R, Odmark E, Schwartz R, Strubell E,*et al*. Measuring the carbon intensity of AI in cloud instances. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022); 2022. p. 1877-94.
44. Souza A, Jasoria S, Chakrabarty B, Bridgwater A, Lundberg A, Skogh F,*et al*. Casper: carbon-aware scheduling and provisioning for distributed web services. In: Proceedings of the 14th International Green and Sustainable Computing Conference; 2023. p. 67-73.
45. Google. Growing the Internet while reducing energy consumption. Google Data Centers; 2026. Available from: https://www.google.com/about/datacenters/efficiency/
46. Epoch AI. AI data centers [Internet]. 2026 [cited 2026 Sep 29]. Available from: https://epoch.ai/data/ai-data-centers

How to Cite This
Akpabla EK. Renewable-aware AI infrastructure planning: an optimization framework for matching data center expansion with renewable generation and energy storage. International Journal of Multidisciplinary Research and Growth Evaluation. 2026;7(5):659-669.

Creative Commons (CC) License
This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution NonCommercialShareAlike 4.0 International (CC BY-NC-SA 4.0) License, which allows others to remix, tweak, and build upon the work noncommercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.