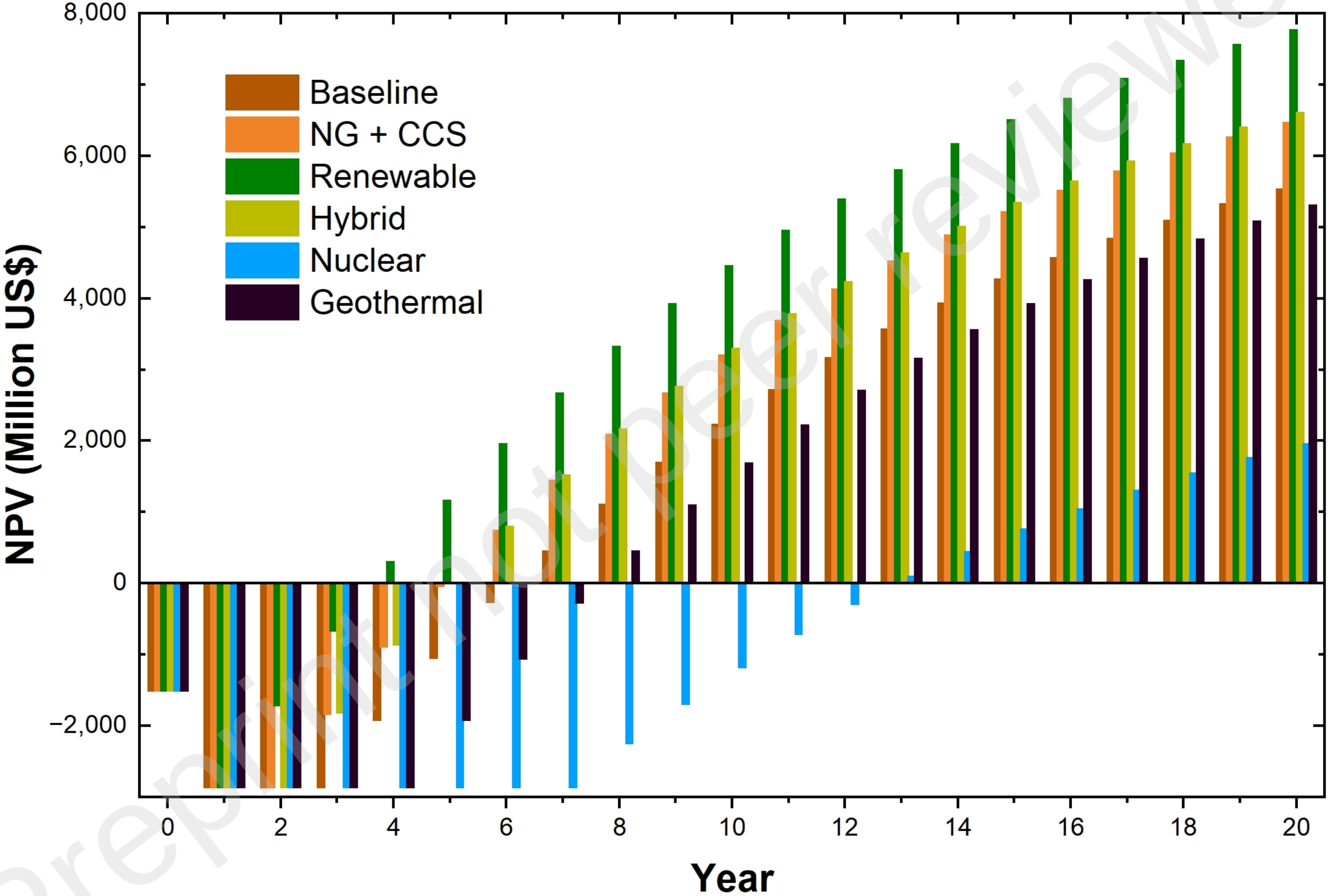
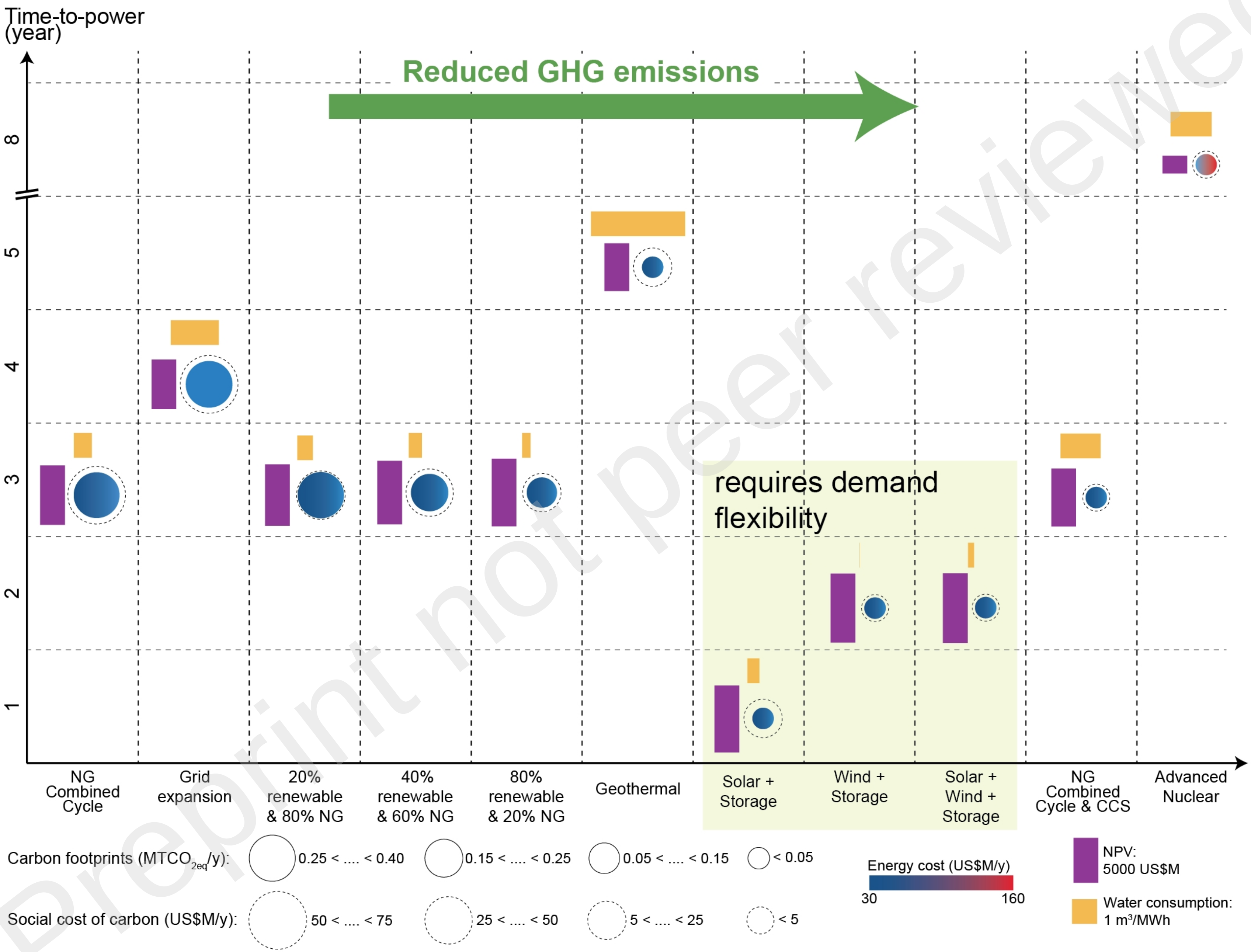
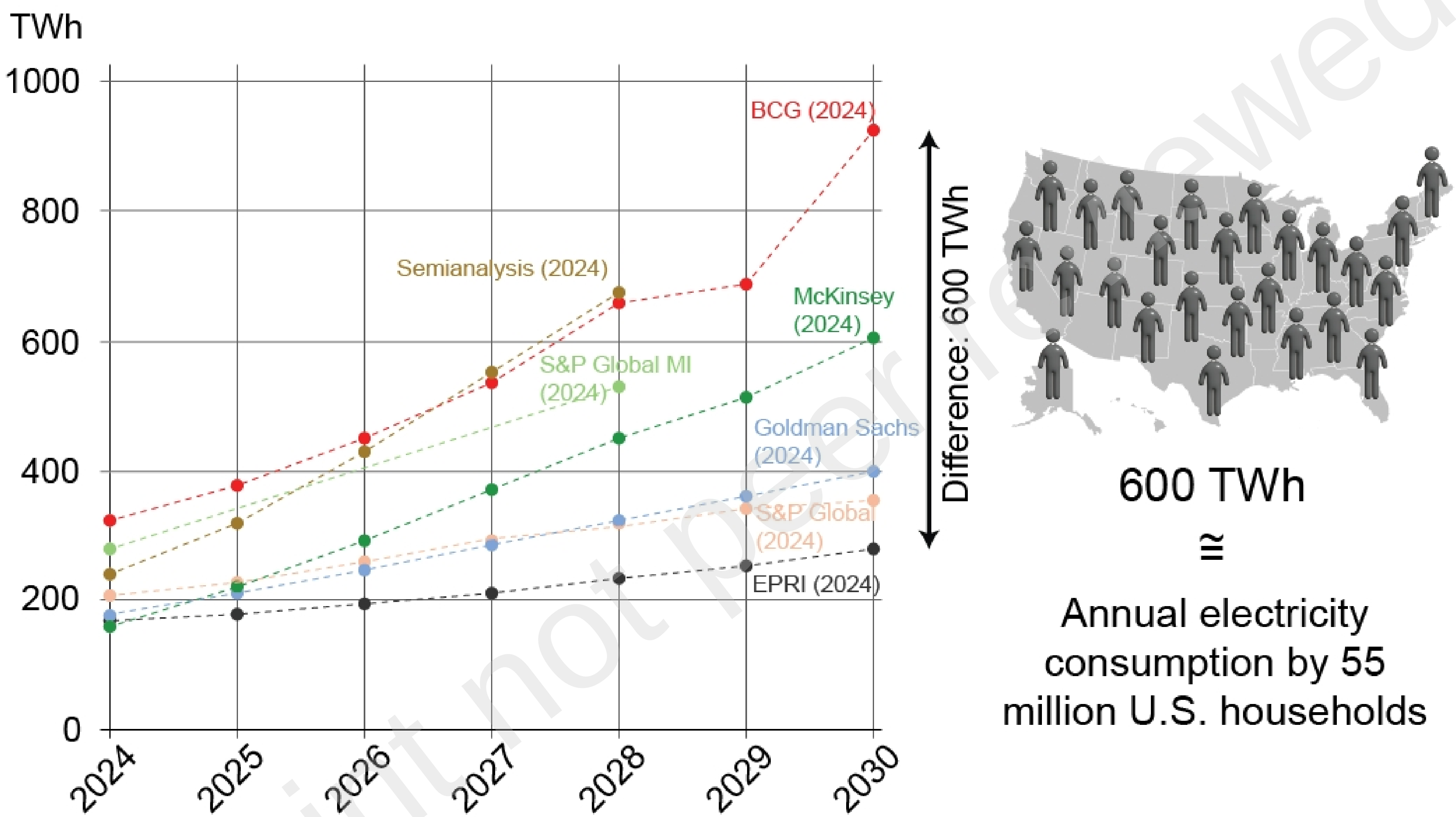


Baseline
NG + CCS
Renewable
Hybrid
Nuclear
Geothermal
NPV (Million US$)
8,000
6,000
4,000
2,000
0
−2,000
0
2
4
6
8
10
12
14
16
18
20
Year



Time-to-power (year)
Reduced GHG emissions
requires demand flexibility
8
5
4
3
2
1
NG Combined Cycle
Grid expansion
20% renewable & 80% NG
40% renewable & 60% NG
80% renewable & 20% NG
Geothermal
Solar + Storage
Wind + Storage
Solar + Wind + Storage
NG Combined Cycle & CCS
Advanced Nuclear
Carbon footprints ($MTCO_{2eq}$/y):
0.25 < .... < 0.40
0.15 < .... < 0.25
0.05 < .... < 0.15
< 0.05
Social cost of carbon (US$M/y):
50 < .... < 75
25 < .... < 50
5 < .... < 25
< 5
Energy cost (US$M/y)
30
160
NPV: 5000 US$M
Water consumption: 1 m^3/MWh


TWh
1000
800
600
400
200
0
2024
2025
2026
2027
2028
2029
2030
BCG (2024)
Semianalysis (2024)
McKinsey (2024)
S&P Global MI (2024)
Goldman Sachs (2024)
S&P Global (2024)
EPRI (2024)
Difference: 600 TWh
600 TWh
≅
Annual electricity consumption by 55 million U.S. households

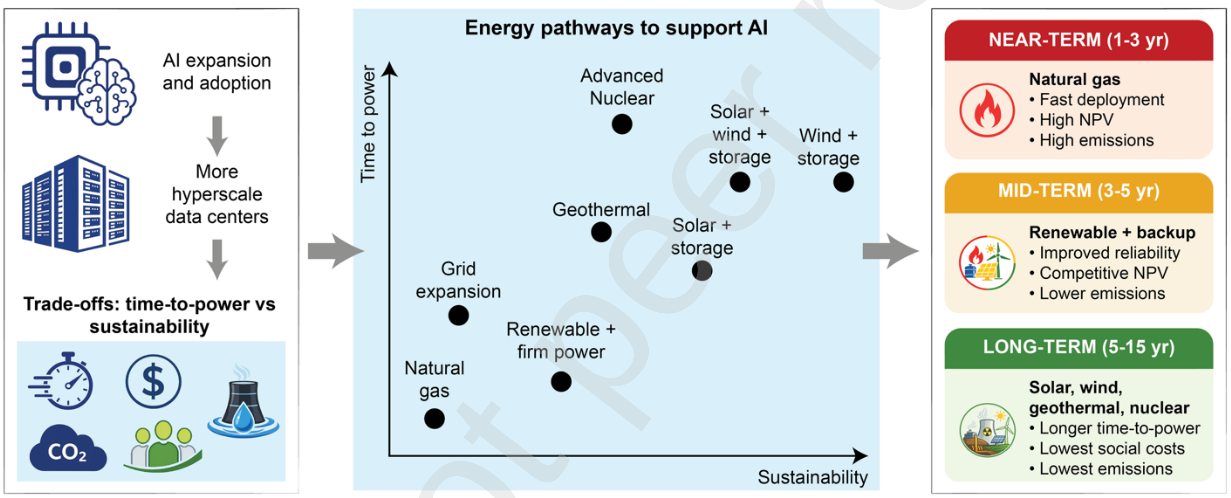
Time-to-Power versus Sustainability: Energy Tradeoffs for the AI Economy

Mohammadreza Heidari[1,2*], Michael Lawrence Brough[2], Edward H. Sargent[3], Jennifer B. Dunn[2]

[1]Trienens Institute of Sustainability and Energy, Northwestern University, Evanston, IL, 60208 USA
[2]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, 60208 USA
[3]Department of Chemistry, Northwestern University, Evanston, IL, 60208 USA
[3]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA

*Corresponding author email address: heidari@northwestern.edu | heidari.environment@gmail.com

Graphical Abstract



Abstract

Artificial Intelligence (AI) is growing at an unprecedented pace and reshaping the U.S. energy system, creating critical challenges for energy access, system resilience, and environmental sustainability. In this Perspective, we discuss potential pathways to power AI and meet its unique demands without compromising long-term sustainability. We outline a three-horizon framework for powering hyperscale data centers: near-term bridge solutions that meet urgent demand, mid-term hybrid systems that integrate renewables with firm generation, and longer-term solutions that require investments in low-carbon energy and storage. The framework illustrates tradeoffs among time-to-power, costs, and environmental effects across multiple electricity supply pathways, including grid electricity, solar and wind with storage, advanced nuclear, natural gas, and hydrogen. We argue that time-to-power is the key parameter driving the net present value (NPV) of hyperscale data centers, with the minimum for nuclear-powered centers and the maximum for renewables-powered centers. We also discuss that natural gas-based solutions can provide high NPV due to rapid deployment, but emit more than 12 times as much as renewables do unless equipped with carbon capture technologies. The analysis recommends expanding renewables and improving storage technologies and demand flexibility, which offer lower

greenhouse gas emissions and water consumption, shorter time-to-power delays, and more competitive economic performance than natural gas-based solutions, the industry's go-to near-term technology.

Keywords: Artificial Intelligence, Energy, Data Center, Time-to-power, Sustainability

Highlights:

- Time-to-power strongly influences the economic value of data centers
- Natural gas offers speed but increases GHG emissions and the social costs of carbon
- Hybrid systems can balance reliability, cost, and sustainability
- Renewable-powered data centers can increase data center economic performance
- A multi-horizon framework guides AI energy infrastructure planning

1. Introduction

Electricity demand from data centers is projected to increase rapidly over the next decade as artificial intelligence expands across the global economy. As Table 1 shows, the International Energy Agency (IEA) projects that global electricity demand from data centers will increase significantly, rising from 460 TWh in 2024 to over 1,000 TWh by 2030 and to 2,000 TWh by 2035.[1] Other models offer different projections; for instance, McKinsey & Company anticipates an increase from 480 TWh in 2023 to between 1,500 TWh and 2,600 TWh by 2030.[2] Regionally, the United States is expected to have the highest demand, followed by China and Europe.[3] According to the IEA, by 2030, electricity demand for data centers is projected to increase by 240 TWh in the U.S., 175 TWh in China, 45 TWh in Europe, and 15 TWh in Japan.[3] Additionally, demand in the U.S. is projected to reach between 320 TWh and 580 TWh by 2028, according to the Berkeley Lab report[4], and to reach 450 TWh by 2028, 606 TWh by 2030, and 1,670 TWh by 2035, according to various McKinsey reports.[5] Bloomberg analysis suggests a figure of 680 TWh by 2035.[6] Despite considerable uncertainty, all projections indicate a substantial increase in electricity demand, with the US expected to experience among the largest increases, at least double by 2028 and fourfold by 2035, necessitating substantial energy resources to support the expansion of data centers.

Table 1. Projections of global data center electricity demand

Source	Region	Baseline year	Projection year	Demand (TWh)
IEA [1]	Global	2024	2035	790-2,000
McKinsey [2]	Global	2023	2030	1,500-2,600
Berkeley Lab [4]	US	2023	2028	320-580
McKinsey [5]	US	-	2028	~ 450
			2030	~ 606
			2035	~ 1,670
BloombergNEF [6]	US	-	2035	~ 680
DOE [7]	US		2030	~ 438
EPRI [8]	US		2030	235-510
EUDCA [9]	Europe		2030	5-18

In the past, improvements in server efficiency, along with the consolidation of computing into the cloud and data centers, helped offset the increasing demand for digital services.[10] However, this trend has changed recently, with a notable increase in demand of 100 TWh between 2018 and 2023.[4] A much higher adaptation rate of generative AI compared to the internet and personal computers will significantly increase energy consumption in the coming years.[11] Data centers are expected to demand an additional 1,500 TWh by 2035.[5] Today, energy-hungry hyperscale data centers that serve approximately six companies (Google, Meta, Amazon Web Services, Microsoft, Oracle, and XAI) are driving up electricity demand. This rapid growth is occurring amid existing concerns about the resilience and sustainability of the electricity grid. [12,13] Adding such a substantial demand to an already stressed electricity grid could worsen these challenges and risk delaying the affordable, reliable, and continuous power supply that data centers, along with communities and other industries, require.

Even today, when the U.S. is at the very early stages of grid expansion to meet data center power demand, communities and other industries are feeling the pinch of rising electricity bills and occasional power outages.[14–16] Their concerns have the attention of policymakers, who must balance the need to fuel AI to compete internationally in technology development with the need to provide electricity to all customers reliably and with minimal cost increases. Together, these challenges have prompted analysis and debate over which power sources are most suitable for build-out to meet this growing demand. The energy systems modeling literature has primarily examined how increasing data center electricity demand may affect the grid. Some of these studies examined the resulting strain on system reliability,[7] the ability for data centers to have flexible load or bring their own generation capacity,[17,18] and the generation mix necessary to meet incremental load in the near term.[8] Current gaps in the literature include limited differentiation and quantitative discussion between various types of on-site generation at data centers, and the integration of on-site battery storage with renewable energy generation at those facilities. Additionally, literature has not considered the social cost of carbon or explicitly addressed the costs and timelines associated with expanding transmission infrastructure.

There are three key factors to consider when evaluating different power sources. The first factor is cost. While a low levelized cost of power may be a substantial determinant of which technologies are selected to build out the grid, another key factor is time-to-power. It usually takes several years to obtain permits and construct a power plant facility or to expand the electricity grid to provide a reliable energy source for data centers. These dynamics can force developers to pursue short-term bridge-to-power strategies, such as relying on firm fossil-fuel energy sources, including delaying their retirements[7], that may undermine long-term sustainability goals. The final factor is greenhouse gas (GHG) emissions, which vary widely across power-generation options. Electric power generation already accounts for 25% of GHG emissions in the U.S.[19] Building out more generation capacity could cause that number to rise and worsen the already substantial effects of a changing climate on the U.S. It is reported that the U.S. spends nearly 1 trillion dollars per year on recovery from climate change-related disasters. [20]

We evaluate these three factors for multiple technologies that can contribute to grid expansion. With this analysis, we explore 1) near-term bridge-to-power solutions to meet urgent demand, 2) mid-term hybrid systems that integrate renewable energy into firm energy sources, and 3) longer-term investments in renewable and nuclear energy infrastructure. This type of analysis can identify the advantages and trade-offs of different grid expansion approaches more rapidly than energy systems models. Energy system models, however, are essential to understanding regional and detailed temporal effects and to identifying optimized strategies. Disadvantages of energy systems models are the time and data required to expand them to characterize the growing demand associated with data centers. In this Perspective, we provide a comparative assessment that offers timely guidance as the modeling community develops new methods to address sudden, spatially concentrated load additions to the grid.

This Perspective argues that expanding electricity consumption to power digital economies merits close attention to social and environmental factors beyond in addition to costs. It calls for framing of data centers as first-order drivers of the energy transition that can have substantial social and environmental effects. We therefore propose a three-stage framework that encompasses environmental, social, and economic perspectives. We highlight the need for system-level decision-grade metrics that account for GHG intensity, water consumption, levelized cost of energy (LCOE), time-to-power, and the social cost of carbon to guide the choice of energy provision for data centers. These metrics can inform approaches to scaling data centers and AI infrastructure to meet growing demand while limiting adverse social and environmental effects.

2. Methodology

2.1. Goal and scope definition

The goal of this study was to evaluate the environmental and economic benefits of powering hyperscale data centers. The functional unit is a typical hyperscale data center with a 100 MW capacity, operating for 20 years, and powered by various energy sources. The system boundary is cradle-to-gate and includes lifecycle greenhouse gas (GHG) emissions and water consumption of fuels.

2.2. Model Development

We developed a model that includes the following parameters: cost, time-to-power, water demand, GHG emissions, and the social cost of carbon for powering a typical 100-MW hyperscale data center. Linking time-to-power (deployment timelines) with the net present value (NPV) of hyperscale data center operations allows us to quantify the economic value of rapid power deployment (Eq.1). NPV also examines the importance of energy costs and their share relative to the potential economic benefits of a 100 MW hyperscale data center. GHG emissions, water consumption, and social costs were calculated using Eq. 2. This model evaluates various energy supply options, including electricity grid expansion, natural gas combined-cycle plants, natural gas combined-cycle plants with carbon capture and storage (CCS), advanced nuclear, and multiple combinations of renewable technologies. Specifically, we examined solar-and-storage, wind-and-storage, solar-and-wind-and-storage, geothermal, and hybrid solutions that integrate renewables (solar, wind, and storage) with firm energy sources such as natural gas.

$$NPV = -C_0 + \sum_{t=T_{TP}}^{T_{TP}+L} \frac{(R_t - O_t - E_t)}{(1+r)^n} \qquad \text{Eq. 1}$$

where:

- NPV: Net Present Value ($)
- C_0: data center capital cost ($)

- T_{TP}: time-to-power (year)
- L: operating lifetime (year)
- R_t: annual revenue ($)
- O_t: annual non-energy operating cost ($)
- E_t: annual energy cost $)
- r: discount rate

$$X_i = E_{DC} \times XI_i \quad \text{Eq. 2}$$

where:

- X_i: annual greenhouse gas emissions ($kgCO_{2eq}/y$), water consumption (m^3/y), social cost ($/y), or energy cost ($/y) for electricity supply in energy pathway i
- E_{DC}: annual electricity demand of the data center (MWh/y)
- XI_i: Impact intensity of electricity source i for greenhouse gas emissions ($kgCO_{2eq}/MWh$), water consumption (m^3/MWh), social cost ($/metric $tonCO_{2eq}$), or energy cost ($/MWh)

2.3. Life cycle inventory

GHG emissions and water consumption from electricity generation, and the levelized cost of energy for various energy sources, are calculated from the literature and presented in Tables S1-S3 of the Supplementary Information (SI). The social cost of carbon is calculated using the EPA factor ($191/metric $tonCO_{2eq}$)[21]. Based on expert-derived insights, we assume a 2-year construction period for a 100 MW hyperscale data center, with a capital cost of $3 billion, annual operational costs of $50 million (excluding energy), projected annual revenue of $1.5 billion, and a 20-year lifespan (Table S2).

3. Results and Discussion

Figure 1 illustrates the economic performance of operating a typical hyperscale data center with a 100 MW capacity over its 20-year lifespan, considering various energy pathways. In the baseline scenario, which represents the electricity grid expansion, the total net present value (NPV) is estimated at approximately $5.5 billion. The advanced nuclear pathway showed the lowest economic performance, with an NPV that is 60% lower than the baseline due to slower deployment. In contrast, several other scenarios demonstrate competitive or even greater economic value than the baseline, highlighting the importance of rapid deployment. The NPV for the geothermal pathway is about 5% lower than the baseline, while solar, wind, and storage options have the potential to increase NPV by 40% relative to grid expansion.

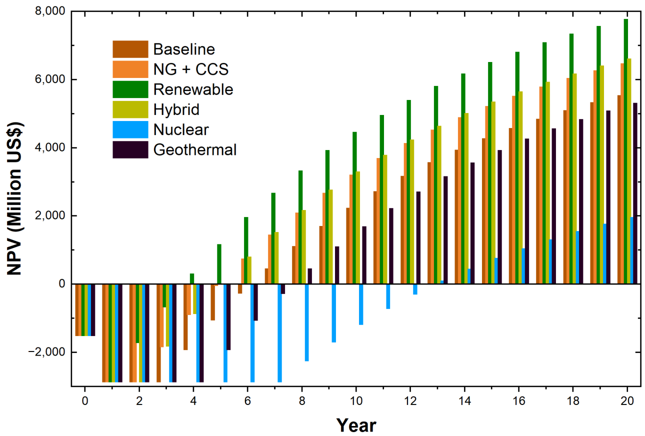


Fig. 1. Net present value of operating a typical hyperscale data center (100 MW) powered by various energy sources.

Figure 2 shows time-to-power vs. sustainability performance of powering a 100 MW data center for 20 years in terms of environmental impacts such as GHG emissions and water consumption, economic performance (i.e., total NPVs), and social impacts (social cost of carbon). The vertical axis shows the estimated time required for permitting, construction, and deployment of different power technologies. Purple bars represent the total NPV and circular markers indicate GHG emissions associated with electricity supply from each option, expressed in million metric tons of CO_2 equivalent ($MtCO_2eq$). NG combined-cycle power plants are recognized as the most emission-intensive energy option, producing 11% more emissions than the grid expansion scenario. In contrast, all other energy pathways have lower emissions than grid expansion. Among these, advanced nuclear technology is projected to have the lowest emissions, being 98% lower, followed by solar, wind, and geothermal technologies, which emit between 90% and 92% less. Circle size represents the corresponding social cost of carbon, with larger circles indicating greater societal damages from emissions. Circle colors represent energy costs, with advanced nuclear identified as the most expensive option. Yellow bars indicate water consumption for each energy pathway, with geothermal having the highest (3.78 m^3/MWh) and wind the lowest (0.004 m^3/MWh) consumption. The geothermal scenario requires nearly twice as much water as the grid expansion scenario. In contrast, other energy pathways, such as advanced nuclear, NG combined cycle with CCS, NG combined cycle, and solar energy, use less water, reducing demand by approximately 15%, 17%, 63%, and 75%, respectively.

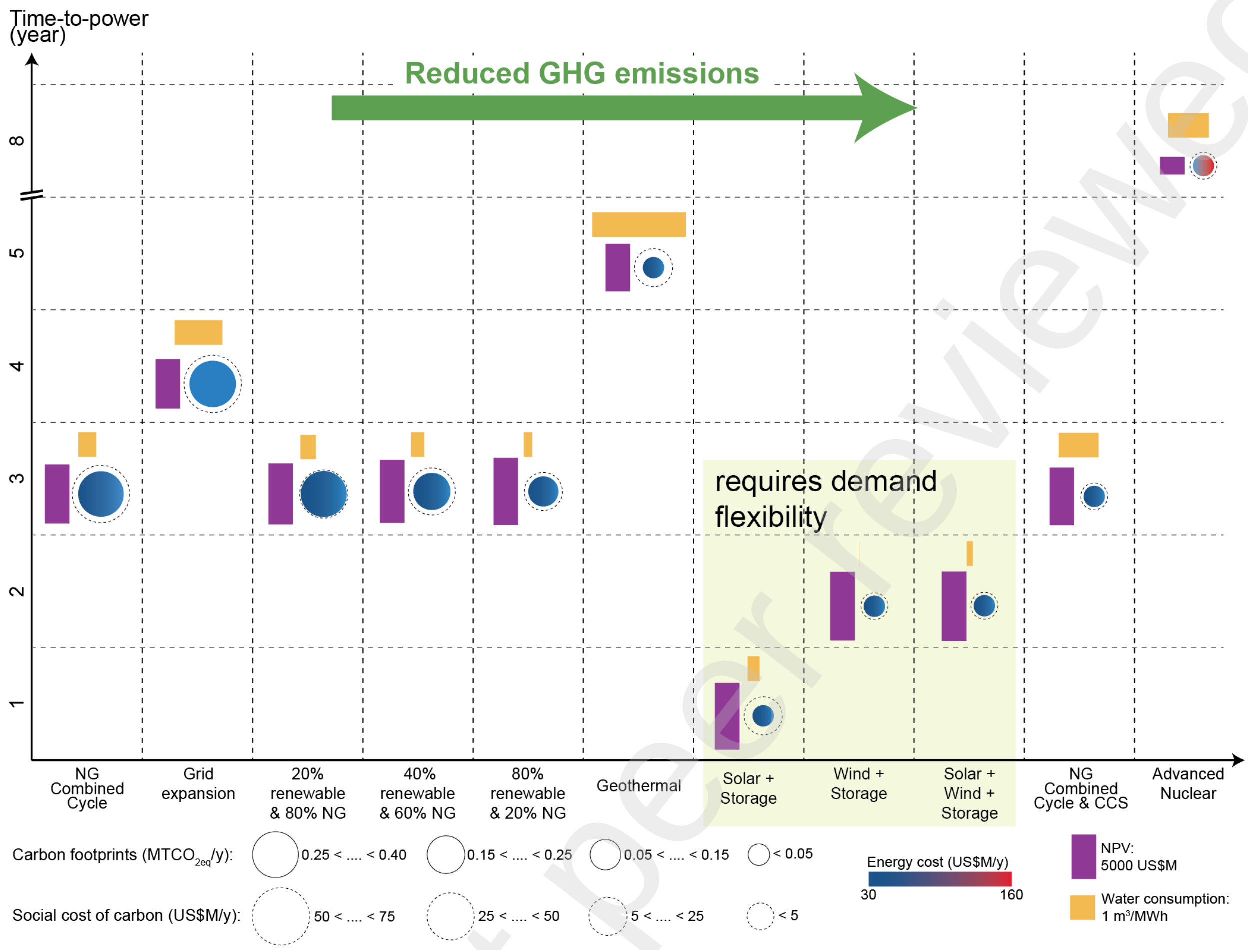


Fig. 2. Trade-offs between time-to-power, economic performance (net present value - NPV), GHG emissions, water consumption, and social impacts (social cost of carbon) across energy supply options for a typical 100 MW hyperscale data center.

Near-term bridge solutions

Natural gas-based technologies, such as NG combined-cycle plants, although GHG-intensive, are increasingly viewed as near-term solutions with short time-to-power that can meet growing electricity demands over the next five years. [11] Time-to-power is a crucial criterion for hyperscalers and investors, as it directly affects data center revenue. Natural gas-based technologies can deliver power within three years. In contrast, low-emitting nuclear energy takes six to 10 years to become operational. A shorter timeline for natural gas solutions can triple economic benefits relative to nuclear over the lifespan of hyperscale data centers. However, the GHG emissions and associated social costs of data centers that use natural gas-driven electricity sources are considerably higher than those of data centers that use lower-emitting electricity sources, unless they are equipped with carbon capture and storage technologies (Fig. 2). For example, constructing natural gas combined cycle plants typically takes about three years. With this approach, hyperscalers could generate around $6.5 billion over 20 years from running a typical 100 MW hyperscale data center, but at the cost of emitting over six million metric tons of CO_2 equivalent ($MTCO_{2eq}$). This level of emissions is nearly 20-fold higher than that of data centers powered by renewable electricity sources with comparable construction times and economic benefits. The social impacts of GHG emissions from natural gas-powered data centers could approach $1.2 billion. These GHG emissions and the social cost of carbon can be significantly mitigated through carbon capture and storage, making them comparable to those of lower-emitting electricity sources. Additionally, natural gas solutions consume more water than solar and wind but less than geothermal and grid expansion, highlighting the complexity of trade-offs between time-to-power and sustainability.

Mid-term hybrid solutions

Hybrid energy systems that integrate renewable energy with firm sources, such as natural gas, offer benefits such as reduced emissions and enhanced reliability. And if these hybrid energy systems are strategically combined with operational strategies, such as workload optimization, they can effectively meet surging power demand while reducing environmental and social trade-offs. Managing computation can reduce energy consumption during peak loads or at specific times of day. This approach can help manage energy consumption and optimize power sources, especially for solar and wind energy. These renewable sources have significant economic potential due to readily available technology and shorter construction timelines. Additionally, they demonstrate strong environmental performance, with GHG emissions and water consumption being the lowest among all energy options.

An interesting possibility is to blend natural gas and solar-and-wind-and-storage solutions for data centers. With lower environmental and social impacts from these renewable energy resources and greater firmness from NG-based technologies, hybrid systems can be a suitable choice, balancing firmness and environmental concerns. For example, providing 20% of power from renewable energy resources and 80% from NG can reduce the GHG emissions and water consumption of near-term solutions by around 19% and 13%, respectively. Hybrid approaches to providing power can be even more beneficial when computing workloads are optimized. This can be achieved through more efficient algorithms and by scheduling latency-tolerant tasks at times when sufficient renewable power is available. Such optimizations give data center developers greater flexibility to incorporate more renewable energy resources. Increasing the share of renewable energy from 20% to 40% can reduce GHG emissions and the associated social costs by 37%. If the share is increased to 80%, these emissions can be reduced by 74%. Additionally, water consumption can be decreased by 27% with a 40% renewable share and by 53% with an 80% share. Importantly, these shifts can increase net present values compared to near-term energy solutions.

Longer-term solutions

In contrast, nuclear and renewable energy resources are typically viewed as longer-term solutions. The main challenges associated with nuclear power include the lengthy construction timelines required to generate energy for data centers and concerns about radioactive waste management. Additionally, although some reports suggest that nuclear power has one of the lowest mortality rates among electricity sources, estimated at roughly 0.03 deaths per TWh, compared with approximately 24 deaths for coal, 2.8 deaths for natural gas, 0.04 deaths for wind, and 0.02 deaths for solar,[22] the risk of catastrophic accidents and disasters continues to influence public perception of nuclear energy and may delay its deployment. Meanwhile, utilities are already extending the life of existing nuclear plants and even bringing previously retired units back online to meet rising demand. [23,24] In contrast, renewable energy sources such as geothermal energy provide firm power with lower GHG emissions, although deployment is still slow and water consumption is almost double that of advanced nuclear technologies. Conversely, other renewable technologies, such as solar and wind, are commercially available, have fewer long-term risks and environmental impacts, and are widely accepted, but they still require storage solutions to deliver reliable, consistent power for hyperscale data centers. Continued investment in R&D of storage technologies is essential for renewables to meet their potential.

Regarding economic benefits and sustainability performance, nuclear energy exhibits the lowest carbon emissions and high-water consumption, estimated at 0.16 $MTCO_{2eq}$ and 23 million m^3, for powering a 100 MW data center for 20 years, but the lowest total net present value among all scenarios due to a longer time-to-power, typically eight years to become operational for data centers. Interestingly, the total net present value of a 100 MW data center powered by on-site renewable energy (solar, wind, and storage) is higher compared to that of near-term solutions, demonstrating the potential of renewable resources to power data centers profitably. In addition, the annual GHG emissions (0.027 to 0.034 $MTCO_{2eq}$) from powering data centers with renewables are comparable to those from nuclear-powered centers and are around 90% lower than those of near-term bridge solutions. As a result, the social costs of carbon are seven to 40 times less than when data centers are powered by fossil fuels like natural gas.

The changing nature of data center electricity demand

While the introduction highlights the wide range of demand projections, the nature of data center electricity demand is also changing in ways that make forecasting increasingly difficult. In previous decades, efficiency largely offset the growth of computing demand, keeping overall data center demand relatively stable. Annual data center energy consumption in the U.S. remained relatively constant at 60-70 TWh between 2005 and 2018; however, this trend has recently changed.[11] In 2019, data centers accounted for only 3% of the total gross power demand in the U.S., with the remaining demand coming from industry (24%) and buildings (73%). [2] By 2023, the power demand from data centers had increased to 5%. [2] Projections suggest that this demand will continue to grow, reaching 13% by 2030 and 25% by 2035. [2] This increase is expected to surpass the power demands of the entire industrial sector, be three times higher than those of hydrogen and synthetic fuels, and be four times greater than those of transportation.[2]

Several features distinguish current and past growth in electricity demand. First, the speed of growth today is unprecedented. Comparing the first two years of generative AI adoption with legacy industries and even with previous digital technologies reveals much sharper growth than in the past. The adoption rate of generative AI was 40% between 2022 and 2024, compared with around 20% for the internet between 1995 and 1997 and 10% for personal computers between 1981 and 1983. [11] A profound mismatch exists between the time scales required to build utility-scale renewable energy and industrial facilities (multiple years) and the time required to create demand for AI algorithms as new models and platforms are developed (months). Second, only a few firms account for a significant share of this growth. This not only shapes the geographical distribution of electricity demand but also creates asymmetries in bargaining power with utilities and local regulators. Third, the growth is not evenly distributed across the US. For example, the mid-Atlantic states have been gaining attention for building hyperscale data centers, despite decades of pressure from power supply shortages.[25,26]

The uncertainty surrounding AI-related growth is as important as the growth itself. The combinations of growth speed and scale are making traditional forecasting tools increasingly inadequate. Load growth, which could traditionally be projected with reasonable uncertainty over 5- to 10-year horizons, now requires extra attention, as the margin of error may be significant. Projected data center load growth varies significantly in 2030, reaching 600 TWh, equivalent to the annual electricity consumption of 55 million US households (Fig. 3). These uncertainties can have multidimensional effects locally and nationally. Utilities, developers, and policymakers are forced to make rapid, billion-dollar decisions under deep uncertainty, and these decisions can shape market dynamics, electricity prices, and decarbonization roadmaps over the next few decades.

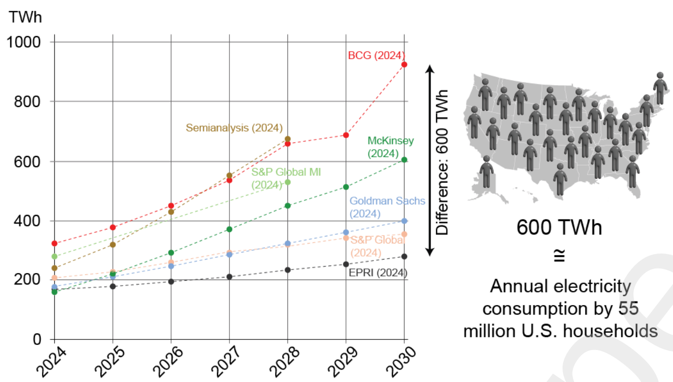


Figure 3 | Wide uncertainty in projections of data center electricity demand. Electricity demand projections are compiled from Kamia et al. [27].

Energy supply strategies

On-site power plants can be beneficial for powering hyperscale data centers, helping overcome grid congestion and delays while enhancing reliability. On-site natural gas power plants have gained attention due to the availability and modularity of natural gas. However, GHG emissions remain a significant concern unless natural gas power plants are equipped with carbon capture technologies. In contrast to natural gas, solar photovoltaic systems and energy storage can supply clean power; however, the intermittency and land constraints force on-site PV systems to serve as a supplement to grid power rather than a replacement for it. Moreover, the large-scale deployment of renewable energy, along with supply chain limitations and the need for independence from foreign resources, will require significant investment. Despite these challenges, emerging hybrid microgrids, which combine solar PV and wind energy supported by battery storage, have gained attention as a solution to address intermittency issues. When properly synchronized with computing load management, these systems can be a reliable and sustainable source of energy. In addition to solar and wind, low-carbon power options such as nuclear energy, hydrogen fuel cells, and geothermal energy can be integrated into energy systems to provide firm, sustainable power. These sources offer unique pathways to power data centers, but over a longer timeframe of five to 15 years.[11] As discussed earlier, coupling renewable energy with natural gas is a possible option to quickly fill energy supply gaps while the world transitions to more sustainable alternatives.

In addition to energy supply, backup power systems need attention. Backup power systems are necessary to improve the reliability of hyperscale data centers and reduce the potential costs associated with service outages, hardware failures, and data corruption. Traditionally, data center facilities have relied on diesel generators and uninterruptible energy supply (UPS) systems to provide power for several minutes to several hours in an outage. [28] However, using diesel generators can be limited by air quality regulations, noise restrictions, limited runtime allowances (often 30-50 hours per year in non-attainment zones), and long equipment lead times (1-2 years). [28–30] Similarly, UPS systems face multiple challenges. Scalability, complex safety management, and degradation, which increase lifecycle costs, can hinder the use of UPS for hyperscale data centers. To address these challenges, more flexible, long-duration backup options with lower emissions are needed. Some potential solutions include renewables and battery storage, natural-gas fuel cells, natural-gas-fired reciprocating engines, modular gas turbines, and hybrid-gas configurations, which can offer faster ramping, lower emissions than diesel, and extended runtime. Using backup power systems can also support emerging DC-Flex strategies, which can temporarily offset grid demand during peak hours, providing grid services and improved reliability at data centers.

Technology innovation and supply chain risks

Technological innovations can also play a critical role in reducing the growth of AI workloads, energy demand, and the associated carbon and social footprints. These innovations encompass advancements in energy efficiency and production, hardware and IT infrastructure, cooling systems, maintenance practices, autonomous operations, edge computing, and supportive policies. Additionally, developing microgrids and hybrid systems that integrate renewable energy sources, storage, and backup generators can create resilient, flexible power architectures.

Hyperscale data centers are the most efficient type of data centers.[11] Approximately 85% of electricity consumption is used for servers, storage, and networking, while the remaining 15% is allocated for cooling and other infrastructure. [11] For thermal management, expanding the use of innovative cooling solutions can be highly beneficial. These include microfluid cooling, in which the coolant routes through hot regions inside the chip through microchannels that are directly integrated into semiconductor dies; solid-state cooling, in which electrochemicals are used to pump heat without relying on conventional refrigerants; temperature chaining, in which multi-stage liquid cooling circuits not only manage heat but also harvest for reuse; and deep-earth cooling integration, which use natural subsurface temperature gradients to draw heat away from data centers. However, multiple concerns should be addressed before using these innovations at scale. For example, for microfluidic liquid flow, manufacturability and embedding microchannels in chips without defecting; for solid-state cooling, material fatigue and durability under cyclic loading; for temperature chaining, compatibility with existing racks, complexity in plumbing, control, and balancing fluid circuits, and avoiding mixing or cross contamination; and for deep-earth cooling, high upfront drilling cost, geologic risks, and maintenance due to ground perturbation make these technologies vulnerable and hard to be at scale in near future.

Supply chain limitations can affect the deployment timeline of data centers. AI supply chains are complex, globally distributed, and materially intensive, spanning high-purity quartz for semiconductor fabrication, critical minerals for GPUs and batteries, and steel and concrete for large-scale construction. IEA has reported a 20% delay in planned data centers due to equipment and material shortages.[11] These material supply chain constraints have increased equipment lead times, resulting in project delays. For example, generator lead times increased by about 60% from 2019 to 2023. This was around 120% for uninterruptible power supplies, 300% for medium-voltage transformers, 150% for medium-voltage switchgear, and 200% for power distribution units over the same period. [5] Current gas turbines' lead time is in the order of a couple of years, creating additional uncertainty for developers seeking firm power to meet rapid AI load growth.

Environmental impacts

If managed poorly, the rapid expansion of AI infrastructure risks amplifying environmental and social burdens across the energy-water-material nexus. If the steep rise in electricity demand is not effectively coupled with renewable energy resources, water consumption, GHG emissions, and local air pollution will rise. AI growth is water-intensive. A typical 100 MW hyperscale data center requires 0.5-5 million liters of water per day for cooling servers. This is comparable to the daily water needs of a small city with a population of 15,000. Additionally, indirect water consumption for powering a typical hyperscale data center is on the order of a million liters per day. Providing this amount of water, particularly in regions with water shortages, can cause irreversible impacts on local ecosystems and natural water reservoirs. Land use change is another environmental concern. A typical hyperscale data center requires 80-100 hectares of space for buildings, cooling systems, and circulation, and, depending on the energy source (onsite or grid), may require even more space, exceeding 500 hectares. This is equal to the size of Central Park in New York City, or 500 soccer fields.

Environmental Justice

Despite widespread recognition of the importance of equity in energy transitions, multiple barriers can continue to hinder climate equity in the context of large-scale infrastructure, such as data centers. First, the impacts of locating AI infrastructure may be unevenly distributed due to the potential to build data centers in low-income or marginalized communities, which often have cheaper land, weaker political influence, and more lenient zoning regulations. These can create disproportionate exposure to air pollution, noise pollution, and water stress, while the economic benefits of data centers, such as high-skill employment, infrastructure investment, and tax revenues, may bypass affected communities. Procedural inequities may also intensify these gaps if community voices are excluded from the permitting process. Additionally, a lack of transparency in environmental assessment reports can limit stakeholders' ability to hold developers accountable for energy sourcing, water withdrawal, and local emissions. Finally, many low-income or marginalized communities typically do not have access to the organizational capacity, funding, and technical experts to meaningfully engage in decision-making processes, which leaves them structurally disadvantaged. These concerns highlight the risk of intensifying existing climate inequities in the absence of responsible development of AI infrastructures.

4. Conclusion

AI-driven electricity demand is expanding at an unprecedented pace, creating uncertainties in forecasting and pushing developers toward fast but carbon-intensive energy sources such as natural gas. Renewables, including solar and wind, offer lower emissions. Data centers using natural gas and renewables for electricity have comparable net present value. Still, renewables' intermittency requires stronger storage, supportive policy, and hybrid near-term strategies to balance firmness, costs, time-to-power, environmental impacts, and social effects.

There is an urgent need to develop comprehensive energy system models that evaluate not only data center energy demand and supply strategies but also their impacts on local environments, including air quality and water scarcity. Innovative strategies such as colocation and load shifting should be explored in depth, as they can reduce the need to expand generation capacity. For example, load shifting for latency-tolerant computing during peak hours can reduce peak demand and, when combined with backup options, support the electricity grid and increase its resiliency. These energy systems models should also enhance the understanding of how data centers and power supply systems explicitly designed for AI computing affect local and global resources. Energy systems models can inform governments, industry, and environmental advocacy organizations about the potential adverse effects of supplying ever-growing amounts of electricity to data centers and cost-effective strategies to mitigate them.

There is also a growing need for models specifically designed to understand how to integrate renewable energy into data center energy systems, including their potential benefits and drawbacks. These models should evaluate a variety of scenarios, including the development of effective backup options to support renewable energy, the identification of resource bottlenecks, the assessment of environmental impacts, the analysis of community effects, the exploration of manufacturing supply chains, the evaluation of material consumption, and the consideration of workforce development. Additionally, these models should account for GHG emissions across scopes 1, 2, and 3 throughout the supply chain. This focus becomes increasingly important as national policies promote domestic manufacturing.

Another important point that requires careful attention is energy-efficiency improvements. These can be achieved through enhancements in server architecture, improvements in cooling system efficiency, and reductions in electricity transmission loss. Most studies have focused on Power Usage Effectiveness (PUE), often neglecting cooling systems and transmission losses. PUE measures the percentage of a data center's total energy consumption used by IT equipment and is always greater than one; lower numbers indicate greater efficiency. PUE values are generally higher for smaller data centers, reportedly around 1.91, while mid-sized and colocation data centers have PUE values of about 1.68. In contrast, hyperscale and AI-specialized data centers have PUE values of 1.22 and 1.14, respectively. [4] These figures highlight potential opportunities to implement energy-efficiency strategies across data centers of varying sizes. For example, the high PUE in small data centers underscores the importance of reducing energy consumption beyond IT equipment, such as cooling systems. For hyperscale and AI-specialized data centers, lower PUE numbers might seem less significant from an energy-efficiency perspective; however, even a slight percentage reduction in PUE can yield substantial energy savings given the scale of energy consumption in these facilities. Therefore, trade-off studies are essential for exploring energy-saving benefits across various types of data centers. These analyses can provide policymakers, investors, and developers with valuable insights, helping them prioritize resources to maximize the benefits of energy efficiency improvements.

CRediT

M. Heidari: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Validation, Visualization, Writing – original draft, Supervision. **M. Brough:** Investigation, Data curation, Writing – review and editing. **E. Sargent:** Conceptualization, Methodology, Validation, Writing – review and editing. **J. Dunn:** Conceptualization, Methodology, Validation, Writing – review and editing

Acknowledgement

The authors acknowledge financial support from the Paula M. Trienens Institute for Sustainability and Energy at Northwestern University, as well as insights and input from industry experts, including, but not limited to, McKinsey & Company, Lazard, Nvidia, EdgeConneX, ComEd, and Constellation. Any errors in this study are solely the responsibility of the authors.

References

[1] Energy supply for AI. Paris: International Energy Agency (IEA); 2025.
[2] Data center power in the US 2025.
[3] Energy demand from AI. Paris: International Energy Agency (IEA); 2025.
[4] Shehabi A, Smith S, Sartor D, Brown R, Herrlin M, Koomey J, et al. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California.; 2024. https://doi.org/10.2172/1372902.
[5] How data centers and the energy sector can sate AI's hunger for power. McKinsey & Company; 2024.
[6] Power for AI: Easier Said Than Built. BloombergNEF; 2025.
[7] Resource Adequacy Report: Evaluating the Reliability and Security of the United States Electric Grid. US Department of Energy; 2025.
[8] Powering Data Centers: U.S. Energy System and Emissions Impacts of Growing Loads. Electric Power Research Institute (EPRI); 2024.
[9] State of European Data Centers. The European Data Center Association (EUDCA), 2025.
[10] Kamiya G, Coroamă VC. Data Centre Energy Use: Critical Review of Models and Results.
[11] Energy and AI. International Energy Agency; 2025.
[12] Gorin A, Zanchi R, Dyson M. How Data Centers Can Set the Stage for Larger Loads to Come | Rocky Mountain Institute (RMI) https://rmi.org/how-data-centers-can-set-the-stage-for-larger-loads-to-come/?utm_source (accessed May 11, 2025).
[13] Electricity Markets, Grid Security and Resilience | U.S. GAO https://www.gao.gov/electricity-markets-grid-security-and-resilience?utm_source (accessed May 11, 2025).
[14] Big Tech's A.I. Data Centers Are Driving Up Electricity Bills for Everyone - The New York Times 2025. https://www.nytimes.com/2025/08/14/business/energy-environment/ai-data-centers-electricity-costs.html (accessed May 11, 2025).

[15] Reed R. How data centers may lead to higher electricity bills - Harvard Law School | Harvard Law School, https://hls.harvard.edu/today/how-data-centers-may-lead-to-higher-electricity-bills/ (accessed May 11, 2025).
[16] Giangreco L. AI Use And Data Centers Are Causing ComEd Bills To Spike — And It Will Likely Get Worse, https://blockclubchicago.org/2025/08/27/ai-use-and-data-centers-are-causing-comed-bills-to-spike-and-it-will-likely-get-worse/ (accessed May 11, 2025).
[17] Brancucci C, Cutler D, Elgqvist, E, Jenkins J, Case N, Burger S, et al. Flexible Data Centers: A Faster, More Affordable Path to Power. Camus, encoord, and Princeton ZERO Lab; 2025.
[18] Knittel CR, Senga JRL, Wang S. Flexible Data Centers and the Grid: Lower Costs, Higher Emissions?
[19] Sources of Greenhouse Gas Emissions | US EPA https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions (accessed May 11, 2025).
[20] Roston E. US Spending on Climate Damage Nears $1 Trillion Per Year 2025. https://www.insurancejournal.com/news/national/2025/06/17/828169.htm?utm_source (accessed May 11, 2025).
[21] EPA Report on the Social Cost of Greenhouse Gases_ Estimates Incorporating Recent Scientific Advances
[22] Ritchie H, Roser M. Safest Form of Energy
[23] US eyes 7% more nuclear power through restarts, upgrades | Reuters https://www.reuters.com/business/energy/us-eyes-7-more-nuclear-power-through-restarts-upgrades-2025-08-07/ (accessed December 2, 2026).
[24] Big tech is bringing nuclear power back to life https://www.economist.com/business/2024/10/09/big-tech-is-bringing-nuclear-power-back-to-life?utm_medium=cpc.adword.pd&utm_source=google&ppccampaignID=17210591673&ppcadID=&utm_campaign=a.22brand_pmax&utm_content=conversion.direct-response.anonymous&gclsrc=aw.ds&gad_source=1&gad_campaignid=17210596221&gbraid=0AAAAADBuq3KVpZ7w4XAdaD7tLV6WnrZ6D&gclid=CjwKCAiAkbbMBhB2EiwANbxtbZMNB37QbLEMcMeX-y1P9WuJWJ5f9XCL9PWv5Yz8h6V4n_FQ0zS17hoCpagQAvD_BwE (accessed December 2, 2026).
[25] Roberts B. Data Center Infrastructure in the United States, 2025 (Map). NREL; 2025.
[26] Energy Secretary Strengthens Mid-Atlantic Grid Reliability Heading into Winter Months | Department of Energy 2025. https://www.energy.gov/articles/energy-secretary-strengthens-mid-atlantic-grid-reliability-heading-winter-months#:~:text=As%20outlined%20in%20the%20North,voiced%20concerns%20about%20resource%20adequacy. (accessed November 18, 2025).
[27] Kamiya G, Coroamă VC. Data Centre Energy Use: Critical Review of Models and Results (IEA-4e). 2025.
[28] Takci MT, Qadrdan M, Summers J, Gustafsson J. Data centres as a source of flexibility for power systems. Energy Reports 2025;13:3661–71. https://doi.org/10.1016/j.egyr.2025.03.020.
[29] Supply Chain Stability & Lead Times in Data Centers: 2025 Benchmark Guide for Critical Equipment 2025. https://www.caeled.com/blog/data-center-lighting/supply-chain-stability-lead-times-in-data-centers-2025-benchmark-guide-for-critical-equipment/#:~:text=Generator%20Lead%20Times,weeks%2C%20causing%20major%20build%20delays. (accessed March 12, 2025).
[30] Fact Sheet and Frequently Asked Questions_ Use of Backup Generators 2025.