

AI and the Global Energy Challenge: Electricity Demand, Thermal Loads, and Environmental Scenarios

José Genaro González-Hernández^{1,2}, Mario Gómez-García¹, José Federico Chong-Flores¹, Samuel Mar-Barón¹

¹TecNM Campus Instituto Tecnológico de Ciudad Madero, México

²Universidad Tecnológica de Altamira, México

Abstract :

It is well known that in many control applications, the hysteresis can be taken advantage of to provide an easy and accurate solution for control purposes. In this paper, the automation of a hysteresis window controlling system is presented. Different solutions, such as classical digital control, block diagram functions, ladder diagrams, and on-off relays' modules, were implemented by engineering college students, showing students' interest and establishing a solid foundation for their technical training and understanding of automatic control techniques.

The rapid expansion of artificial intelligence (AI) is driving increased investment in computing infrastructure, resulting in higher electricity demand, greater thermal dissipation, and heightened environmental pressures. This study presents a reproducible, scenario-based framework for estimating global data center electricity consumption through 2040. International Energy Agency (IEA) estimates of 415 TWh in 2024 and a central projection of approximately 945 TWh in 2030 serve as contextual benchmarks. For illustrative calibration, a 2025 reference demand of 485 TWh and a rounded 2030 anchor of 950 TWh are employed. Three post-2030 annual growth scenarios (3%, 6%, and 10%) are combined with hypothetical declines in electricity carbon intensity (5%, 3%, and 1%). The projected 2040 electricity demands are 1,277, 1,701, and 2,464 TWh, respectively, with corresponding illustrative operational emissions of 268, 439, and 780 Mt CO₂ per year. Sensitivity analysis and a logistic alternative highlight the strong dependence of outcomes on growth and saturation assumptions. The study further derives IT thermal load from facility electricity use and power usage effectiveness (PUE), examines cooling electricity and water constraints, and identifies methodological limitations. The results represent conditional scenarios for all data centers, rather than direct measurements or forecasts of AI-specific electricity consumption, and are intended to inform energy planning and guide future empirical research.

Keywords: Artificial intelligence; data centers; energy forecasting; carbon emissions; cooling; thermal management; sensitivity analysis.

Date of Submission: 11-10-2026

Date of Acceptance: 23-10-2026

I. INTRODUCTION

The emergence of generative AI, high-performance computing, and widespread inference services has introduced new challenges for electricity planning. Data centers integrate computing processors, storage, networking, power conditioning, and cooling systems, frequently located in geographically concentrated clusters. While efficiency improvements have historically offset substantial growth in computing workloads [1], the increasing prevalence of AI-intensive computing has the potential to significantly alter both equipment power density and total electricity demand [2], [3], [4].

According to the International Energy Agency (IEA), data centers consumed approximately 415 TWh of electricity in 2024, with projections indicating an increase to around 945 TWh by 2030 in its 2025 base case [5]. The IEA notes that data center electricity demand should not be equated solely with AI-related consumption, and that local network constraints may be more significant than global electricity shares imply [5]. Additionally, energy use in data centers results in direct heat rejection, with climate implications that depend on the timing and location of electricity supply [6], [7]. IEA electricity market assessments offer further context for system-level demand planning [8], [9], and its global energy outlook addresses broader energy-system transitions [10].

This research addresses the following questions: (i) How sensitive is projected 2040 data center electricity demand to post-2030 growth assumptions? (ii) To what extent can changes in electricity carbon intensity offset demand growth? (iii) What thermal loads are implied by these demand trajectories? The primary contribution is a transparent and reproducible multi-scenario calculation, rather than an empirical prediction. The analytical horizon spans 2025 to 2040.

II. THEORETICAL BACKGROUND

2.1. Facility electricity and computational workloads

Facility electricity comprises IT electricity, cooling electricity, and auxiliary consumption. Power usage effectiveness is defined as total facility electricity divided by IT electricity [11]. PUE measures overhead but does not capture energy per AI task, model quality, embodied emissions, or water consumption. Training and inference have different utilization patterns; high-volume inference can materially increase aggregate energy consumption [2], [3], [12].

$$PUE = \frac{E_{IT} + E_{cool} + E_{aux}}{E_{IT}} = \frac{E_{fac}}{E_{IT}} \quad (1)$$

Variable	Description	Unit
E_{fac}	Total electrical energy consumed by the data center facility	kWh, MWh, or TWh
E_{IT}	Electrical energy consumed by information technology equipment, including servers, GPUs, processors, storage, and networking devices	kWh, MWh, or TWh
E_{cool}	Electrical energy consumed by cooling systems, including chillers, pumps, fans, and associated thermal management equipment	kWh, MWh, or TWh
E_{aux}	Electrical energy consumed by auxiliary systems, including lighting and electrical conversion losses	kWh, MWh, or TWh
PUE	Power Usage Effectiveness, defined as the ratio of total facility energy consumption to IT equipment energy consumption	Dimensionless

2.2. Heat generation and cooling

At a steady operating point, nearly all electrical power supplied to IT hardware becomes thermal energy in the local environment. For a coolant stream, extracted heat depends on mass flow, specific heat, and temperature rise. Cooling system electricity is distinct from IT heat output and cannot be inferred without a coefficient of performance, pump power, or other system characteristics [13].

$$\dot{Q}_{IT} \approx P_{IT} \quad (2)$$

$$\dot{Q}_{coolant} = \dot{m}c_p(T_{out} - T_{in}) \quad (3)$$

Variable	Description	Unit
$\dot{Q}_{IT}$	Rate of heat dissipation generated by IT equipment during operation	W or kW
P_{IT}	Electrical power consumed by IT equipment	W or kW
$\dot{Q}_{coolant}$	Rate of thermal energy removed by the cooling fluid	W or kW
$\dot{m}$	Mass flow rate of the cooling fluid	kg/s
c_p	Specific heat capacity of the cooling fluid at constant pressure	J/(kg·K)
T_{out}	Cooling fluid temperature at the outlet of the heat removal system	°C or K
T_{in}	Cooling fluid temperature at the inlet of the heat removal system	°C or K

2.3. Operational electricity-related emissions depend on electricity consumed and the carbon intensity of supply.

Lifecycle assessments also require hardware production, buildings, backup systems, and equipment replacement [6], [7], [14]. Water use depends on cooling technology, climate, and electricity generation; electricity demand alone is not a valid proxy for site-level water withdrawals or consumption [15]. Corporate sustainability reports provide examples of operational disclosure, although their organizational boundaries differ from this global scenario model [16], [17].

$$C_{Mt}(t) = E_{TWh}(t) I_{kgCO_2/kWh}(t) \quad (4)$$

Variable	Description	Unit
$C_{Mt}(t)$	Annual carbon dioxide emissions associated with electricity consumption	Mt CO ₂ /year
$E_{TWh}(t)$	Annual electrical energy consumption	TWh/year
$I_{kgCO_2/kWh}(t)$	Carbon intensity of the electricity supply, expressed as carbon dioxide emissions per unit of electricity consumed	kg CO ₂ /kWh
t	Year of evaluation	Calendar year

III. DEVELOPMENT

3.1. Benchmark selection and transparency

The central reference series uses $E(2025) = 485$ TWh and $E(2030) = 950$ TWh as rounded, illustrative anchors consistent with the scale of recent IEA outlooks [5]. The 2025 anchor should be treated as a model input, not an independently verified observation in this paper. The 2030 anchor is a forecast benchmark rather than a measured value. The implied 2025-2030 compound annual growth rate is $\left(\frac{950}{485}\right)^{0.2} - 1 \approx 14.39\%$. No regression is claimed from these two points. The reference-year calibration combines a 2025 estimate with a rounded 2030 outlook; it is not an empirical regression or an uncertainty interval.

Table 1. Variables, values, and status used to support the project development

Item	Value	Status
2024 electricity	415 TWh	IEA historical estimate [5]
2025 anchor	485 TWh	Illustrative calibration input
2030 anchor	950 TWh	Rounded projection benchmark [5]
Post-2030 growth	3%, 6%, 10%	Scenario assumptions
2030 carbon intensity	0.35 kg CO ₂ /kWh	Illustrative assumption
Annual intensity decline	5%, 3%, 1%	Scenario assumptions
Facility PUE	1.25	Illustrative assumption

3.2. Piecewise exponential scenarios

$$E_s(t) = 485 \left(\frac{950}{485}\right)^{0.2(t-2025)}, 2025 \leq t \leq 2030 \quad (5)$$

$$E_s(t) = 950(1 + g_s)^{t-2030}, 2030 \leq t \leq 2040 \quad (6)$$

Variable or constant	Description	Unit
$E_s(t)$	Projected annual electricity demand under scenario s at year t	TWh/year
g_s	Assumed annual growth rate of electricity demand for scenario s	Dimensionless annual fraction
s	Index identifying the scenario: high efficiency, moderate growth, or accelerated expansion	Dimensionless
t	Projection year	Calendar year

The high-efficiency, moderate-growth, and accelerated-expansion cases respectively set $g_s = 0.03, 0.06, \text{ and } 0.1$. These rates are not estimated probabilities; the labels describe combinations of utilization, adoption, and efficiency. Because the model aggregates all data center workloads, its output must not be described as the energy footprint of AI alone.

3.3. Operational emissions and thermal loads

$$I_s(t) = 0.35(1 - d_s)^{t-2030}, t \geq 2030 \quad (7)$$

Variable or constant	Description	Unit
$I_s(t)$	Projected carbon intensity of electricity under scenario s at year t	kg CO ₂ /kWh
d_s	Assumed annual reduction rate of electricity carbon intensity for scenario s	Dimensionless annual fraction
s	Index identifying the decarbonization scenario	Dimensionless
t	Year of evaluation	Calendar year

$$C_{s(t)} = E_{s(t)} I_{s(t)} \quad (8)$$

Variable or constant	Description	Unit
$C_s(t)$	Projected annual electricity-related carbon dioxide emissions under scenario s	Mt CO ₂ /year
$E_s(t)$	Projected annual electricity consumption under scenario s	TWh/year

$$\bar{P}_{fac,GW} = \frac{1000E_{TWh}}{8760} \quad (9)$$

$$\bar{P}_{IT} = \frac{\bar{P}_{fac,GW}}{PUE} \quad (10)$$

Variable or constant	Description	Unit
$\bar{P}_{\text{fac,GW}}$	Average electrical power demand of the data center facilities	GW
E_{TWh}	Annual electricity consumption	TWh/year
$\bar{P}_{\text{IT}}$	Average electrical power consumed by IT equipment	GW
PUE	Power Usage Effectiveness of the data center infrastructure	Dimensionless

Here $d_s = 0.05, 0.03, \text{and } 0.01$ respectively. Carbon intensity trajectories are hypothetical and are not estimates of the electricity supply contracted by individual operators. Average thermal loads are calculated for IT equipment, not peak site cooling loads. For an idealized cooling system with coefficient of performance COP, compressor electricity could be approximated by $E_{\text{coll}} = \frac{Q_{\text{removed}}}{\text{COP}}$, provided compatible boundaries and energy units; no COP is assumed in the primary results.

3.4. Sensitivity and saturation alternative

$$\frac{\delta E(2040)}{\delta g} = 10(950)(1 + g)^9 \quad (11)$$

$$\varepsilon_g = \frac{g \delta E}{E \delta g} = \frac{10g}{1 + g} \quad (12)$$

Variable	Description	Unit
ε_g	Elasticity of projected electricity demand with respect to the annual growth rate	Dimensionless
g	Assumed annual electricity demand growth rate	Dimensionless annual fraction
E	Projected annual electricity consumption in 2040	TWh/year

At $g=6\%$, the elasticity ε_g is approximately 0.566. A logistic model provides an alternative in which demand growth gradually slows toward an assumed ceiling K :

$$E_{\log(t)} = \frac{K}{1 + \left(\frac{K}{950} - 1\right) e^{-r(t-2030)}} \quad (13)$$

Variable or constant	Description	Unit
$E_{\log(t)}$	Projected annual electricity demand obtained using the logistic growth model	TWh/year
K	Assumed asymptotic saturation level of annual electricity demand	TWh/year
r	Intrinsic growth rate parameter of the logistic model	year ⁻¹

For demonstration only, $K=2,500$ TWh and $r=0.12$ year⁻¹. These parameters are not empirically fitted and should not be interpreted as a physical upper bound. The logistic comparison illustrates structural model uncertainty rather than superior predictive performance.

IV. RESULTS

4.1. Electricity demand trajectories

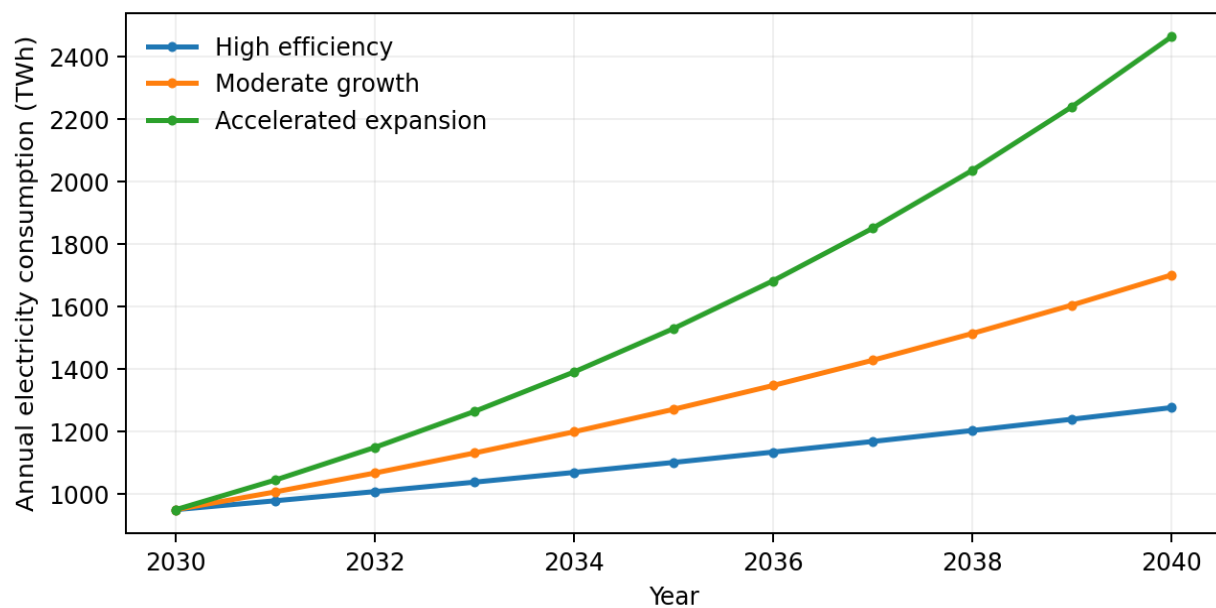


Figure 1. Illustrative global data center electricity scenarios, 2030–2040.

Table 2. High efficiency, moderate growth and accelerated (Twh) by year.

Year	High efficiency	Moderate growth	Accelerated
2030	950	950	950
2035	1,101	1,271	1,530
2040	1,277	1,701	2,464

The difference between the high-efficiency and accelerated cases in 2040 is approximately 1,187 TWh/year. This spread is driven by assumptions and should not be interpreted as a confidence interval.

4.2. Operational carbon emissions

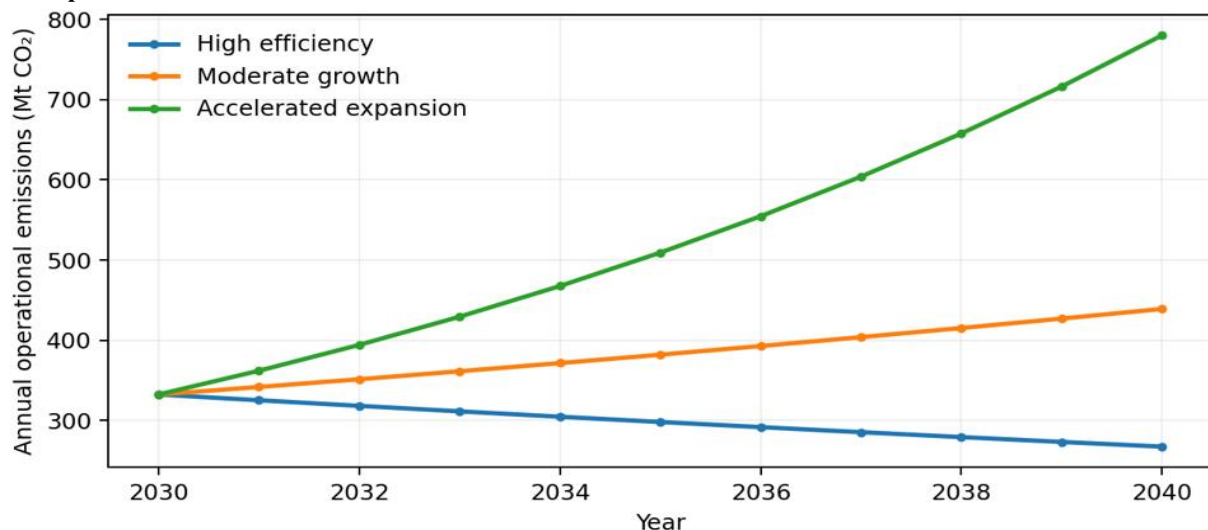


Figure 2. Hypothetical operational electricity-related emissions; lifecycle emissions excluded.

Table 3. High efficiency, moderate growth and accelerated (Mt) by year.

Year	High efficiency (Mt)	Moderate growth (Mt)	Accelerated (Mt)
2030	333	333	333
2035	298	382	509
2040	268	439	780

In the high-efficiency case, carbon intensity falls faster than electricity use grows, producing declining modeled emissions. In the accelerated case, electricity demand rises faster than the assumed carbon intensity decreases. Results are scenario arithmetic, not measured global data center emissions.

4.3. Sensitivity and structural uncertainty

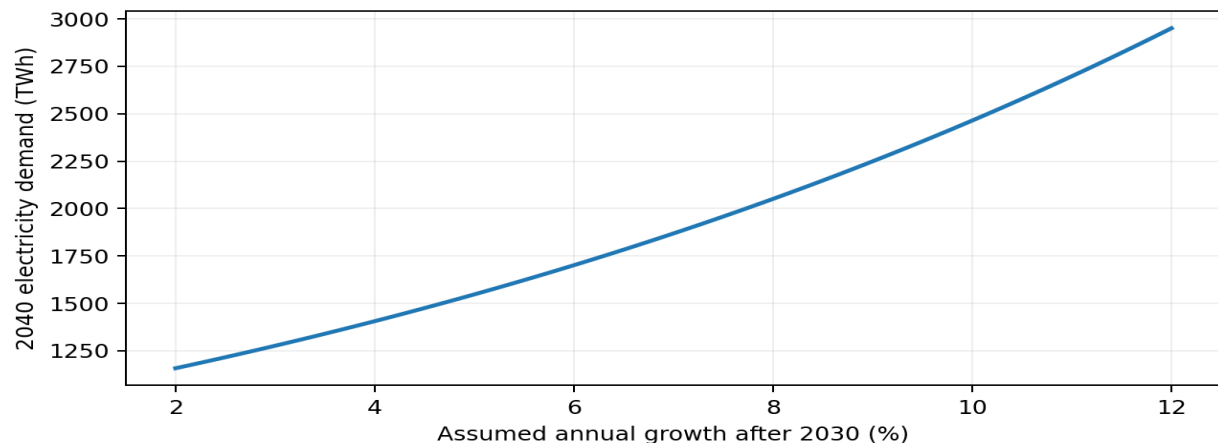


Figure 3. Sensitivity of projected 2040 demand to the assumed 2030–2040 growth rate.

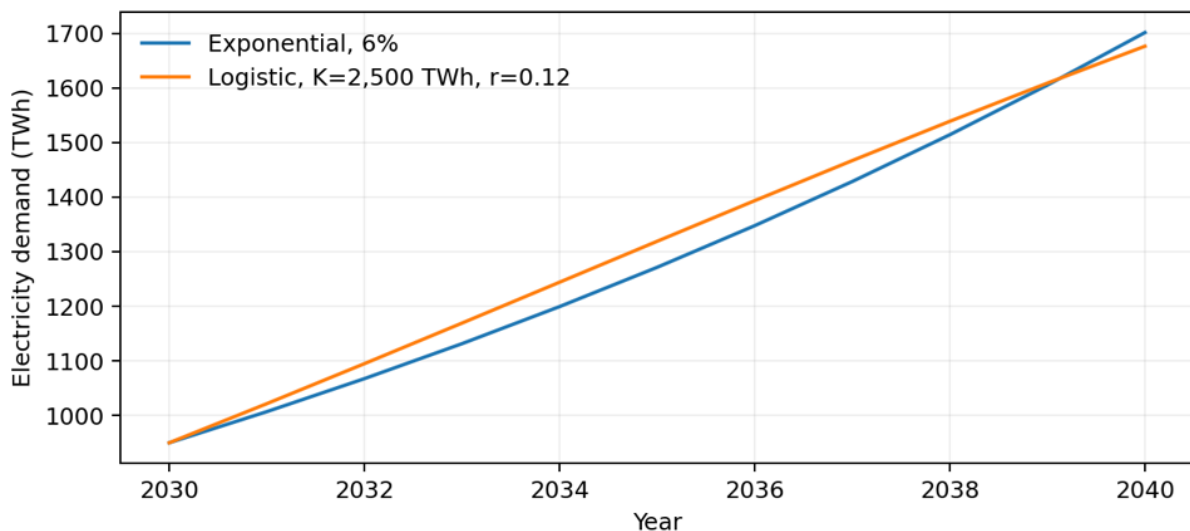


Figure 4. Exponential versus illustrative logistic trajectories; logistic parameters are not fitted.

With $K=2,500$ TWh and $r=0.12$, the illustrative logistic model yields 1,676 TWh in 2040, compared with 1,701 TWh under the 6% exponential case. The models diverge despite sharing the 2030 anchor, showing that assumptions about saturation materially affect long-horizon projections. Numerically, the unrounded logistic and exponential results are 1,676.25 and 1,701.31 TWh, respectively.

4.4. Average thermal power

2040 scenario	Facility mean load (GW)	IT mean heat (GW), PUE=1.25
High efficiency	145.7	116.6
Moderate growth	194.2	155.4
Accelerated expansion	281.3	225.0

The moderate-growth scenario implies approximately 194 GW of average global facility electrical load and 155 GW of average IT heat dissipation in 2040. This is a global aggregate and cannot be converted into individual facility chiller sizing without local hourly profiles, heat recovery, and cooling system data. The IT heat estimate assumes near-complete conversion of IT electricity into local heat; it excludes the detailed thermodynamic performance of cooling equipment.

V. FINDINGS AND COMMENTS

First, projections of 2040 demand are highly sensitive to assumptions regarding the rate of ongoing infrastructure expansion. Employing a scenario framework is more robust than extrapolating from a single, brief historical trend [5]. Second, operational emissions do not necessarily increase in direct proportion to electricity consumption; emissions are determined by both demand and grid carbon intensity, which can vary significantly by location and over time [6], [7]. Third, rising rack densities make cooling system design a critical engineering constraint. Direct-to-chip and other liquid-cooling technologies facilitate heat extraction at high power densities, yet electricity for pumping, water consumption, refrigerant use, and equipment manufacturing remain significant considerations [13], [18], [15]. The International Renewable Energy Agency's energy-transition assessment provides an additional perspective on decarbonization pathways [19], while the International Energy Agency's global energy review offers further context regarding electricity systems [20].

Fourth, relying solely on facility power usage effectiveness (PUE) can obscure workload-level efficiency; improved facility overhead does not ensure reduced total energy consumption if artificial intelligence activity grows more rapidly than efficiency improvements. Fifth, factors such as grid interconnection, peak demand, reliability, and regional concentration may impose more significant constraints than global annual energy totals [5], [9]. Comprehensive environmental assessments should incorporate both lifecycle emissions and water stress [14], [15]. Industry surveys provide documentation of operational efficiency and reliability practices [11], [18].

Limitations: The study has no original meter readings, no AI-only workload decomposition, no regional grid dispatch simulation, no probabilistic parameter estimation, and no empirical validation of the logistic ceiling. Scenario intensities are assumed, not observed. Reported TWh values are rounded. The models are reproducible but should be recalibrated as verified annual data become available.

VI. CONCLUSIONS

A transparent scenario model was constructed to project global data center electricity demand through 2040, driven by the rapid expansion of AI infrastructure. For 2040, projected electricity consumption is 1,277 TWh under high efficiency, 1,701 TWh under moderate growth, and 2,464 TWh under accelerated expansion scenarios. Assuming reductions in carbon intensity, the corresponding annual operational emissions are approximately 268, 439, and 780 Mt CO₂, respectively. With a power usage effectiveness (PUE) of 1.25, the moderate scenario indicates an average global IT heat load of approximately 155 GW.

These estimates represent conditional projections for all data centers, rather than forecasts limited to AI-specific facilities. The results highlight the necessity for integrated advancements in computing efficiency, clean electricity adoption, thermal management, and grid infrastructure planning. Future research should calibrate regional demand using metered data, quantify the proportion of AI workloads, model cooling system coefficient of performance (COP) and water consumption, and validate both exponential and saturation models with updated empirical observations.

An independent recalculation of Equations (4)-(11) confirms that the tabulated 2035 and 2040 electricity and carbon results are reproduced within the specified rounding precision. For 2030, the emissions value is 332.5 Mt CO₂ (950 TWh × 0.35 kg CO₂/kWh), which is reported as 333 Mt using half-up rounding. In 2040, the precise modeled electricity values are 1,276.72, 1,701.31, and 2,464.06 TWh, with corresponding emissions of 267.55, 439.10, and 779.96 Mt CO₂.

The local sensitivity at $g = 0.06$ is calculated as $\partial E(2040)/\partial g = 10 \times 950 \times 1.06^9 \approx 16,053$ TWh per unit of g , equating to approximately 160.5 TWh for a one-percentage-point change under a local linear approximation. The dimensionless elasticity is $10 \times 0.06 / 1.06 \approx 0.566$. These values represent deterministic sensitivities rather than probabilistic uncertainty bounds.

The logistics-based alternative serves as a hypothetical structural comparison rather than a fitted forecast. Observational confidence intervals, validation errors, or statistical significance cannot be determined without further historical data and parameter estimation. The reported electricity-related emissions do not include embodied carbon, backup fuel, or cooling water impacts. Additionally, cooling electrical demand cannot be derived from PUE alone without explicit assumptions regarding equipment and operational conditions.

REFERENCES

- [1] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020, <https://doi.org/10.1126/science.aba3758>
- [2] A. de Vries, "The growing energy footprint of artificial intelligence," *Joule*, vol. 7, no. 10, pp. 2191–2194, 2023, <https://doi.org/10.1016/j.joule.2023.09.004>
- [3] A. S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: Watts driving the cost of AI deployment?," in *Proc. ACM FAccT*, 2024, pp. 85–99, <https://doi.org/10.1145/3630106.3658542>
- [4] D. Patterson et al., "Carbon emissions and large neural network training," arXiv:2104.10350, 2021. <https://arxiv.org/abs/2104.10350>
- [5] International Energy Agency, Energy and AI. Paris, France: IEA, 2025. <https://www.iea.org/reports/energy-and-ai>
- [6] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of BLOOM, a 176B parameter language model," arXiv:2211.02001, 2022. <https://arxiv.org/abs/2211.02001>
- [7] R. Verdecchia, J. Sallou, and L. Cruz, "A systematic review of Green AI," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 4, Art. no. e1507, 2023. <https://doi.org/10.1002/widm.1507>
- [8] International Energy Agency, Electricity 2025. Paris, France: IEA, 2025. <https://www.iea.org/reports/electricity-2025>
- [9] International Energy Agency, Electricity 2024. Paris, France: IEA, 2024. <https://www.iea.org/reports/electricity-2024>
- [10] International Energy Agency, World Energy Outlook 2024. Paris, France: IEA, 2024. <https://www.iea.org/reports/world-energy-outlook-2024>
- [11] D. Donnellan et al., "Uptime Institute Global Data Center Survey 2024," Uptime Institute, Jul. 2024. <https://intelligence.uptimeinstitute.com/resource/uptime-institute-global-data-center-survey-2024>
- [12] S. Samsi et al., "From words to watts: Benchmarking the energy costs of large language model inference," arXiv:2310.03003, 2023. <https://arxiv.org/abs/2310.03003>
- [13] ASHRAE, 2021 ASHRAE Handbook—Fundamentals. Atlanta, GA, USA: ASHRAE, 2021. <https://www.ashrae.org/technical-resources/publications-library/ashrae-handbook/description-2021-ashrae-handbook-fundamentals>
- [14] A. Shehabi et al., 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, 2024, doi: 10.71468/P1WC7Q. <https://doi.org/10.71468/P1WC7Q>
- [15] P. Li, J. Yang, M. A. Islam, and S. Ren, "Making AI less thirsty: Uncovering and addressing the secret water footprint of AI models," arXiv:2304.03271, 2023. <https://arxiv.org/abs/2304.03271>
- [16] Google, 2025 Environmental Report, Jun. 2025. <https://sustainability.google/reports/google-2025-environmental-report/>
- [17] Microsoft, 2025 Environmental Sustainability Report, 2025. [Online]. Available: <https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report>
- [18] D. Donnellan et al., "Uptime Institute Global Data Center Survey 2023," Uptime Institute, Jul. 2023. <https://intelligence.uptimeinstitute.com/index.php/resource/uptime-institute-global-data-center-survey-2023>
- [19] International Renewable Energy Agency, World Energy Transitions Outlook 2024: 1.5°C Pathway. Abu Dhabi, UAE: IRENA, Nov. 2024. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2024/Nov/IRENA_World_energy_transitions_outlook_2024.pdf
- [20] International Energy Agency, Global Energy Review 2025. Paris, France: IEA, 2025. <https://www.iea.org/reports/global-energy-review-2025>