nature sustainability

Analysis

https://doi.org/10.1038/s41893-025-01681-y

Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA

Received: 3 January 2025

Accepted: 7 October 2025

Published online: 10 November 2025

Check for updates

Tianqi Xiao[1], Francesco Fuso Nerini[2,3,4,5], H. Damon Matthews[6], Massimo Tavoni[4,5] & Fengqi You[1,7,8,9]

The rapidly increasing demand for generative artificial intelligence (AI) models requires extensive server installation with sustainability implications in terms of the compound energy–water–climate impacts. Here we show that the deployment of AI servers across the United States could generate an annual water footprint ranging from 731 to 1,125 million m^3 and additional annual carbon emissions from 24 to 44 Mt CO_2-equivalent between 2024 and 2030, depending on the scale of expansion. Other factors, such as industry efficiency initiatives, grid decarbonization rates and the spatial distribution of server locations within the United States, drive deep uncertainties in the estimated water and carbon footprints. We show that the AI server industry is unlikely to meet its net-zero aspirations by 2030 without substantial reliance on highly uncertain carbon offset and water restoration mechanisms. Although best practices may reduce emissions and water footprints by up to 73% and 86%, respectively, their effectiveness is constrained by current energy infrastructure limitations. These findings underscore the urgency of accelerating the energy transition and point to the need for AI companies to harness the clean energy potential of Midwestern states. Coordinating efforts of private actors and regulatory interventions would ensure the competitive and sustainable development of the AI sector.

The accelerating deployment of artificial intelligence (AI) servers is being fuelled by the growing demand for generative AI applications, ignited by milestones such as the release of ChatGPT in 2022[1]. Projections signal even greater impact, exemplified by the recent Blackwell platform, which analysts herald as a new Moore's Law era[2]. Pronouncements from influential AI industry figures[3] on both demand and supply sides are seen as transformative shifts for the entire data-centre industry. Whereas AI has been employed in various fields to advance sustainability[4,5], the remarkable energy requirements of AI itself raise concerns regarding not only energy provisioning challenges[6] but also water scarcity and climate change issues stemming from the energy–water–climate nexus of AI data centres[1,7,8]. However, the holistic energy–water–climate implications of AI computing are largely unknown, constrained by untransparent industry reports and limited data.

The climate impact of AI servers will stem primarily from their operations (Scopes 1 and 2)[9] and supply-chain activities (Scope 3),

[1]College of Engineering, Cornell University, Ithaca, NY, USA. [2]Unit of Energy Systems Analysis (dESA), KTH Royal Institute of Technology, Stockholm, Sweden. [3]Environmental Change Institute, Oxford University, Oxford, UK. [4]RFF-CMCC European Institute on Economics and the Environment, Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Milan, Italy. [5]Politecnico di Milano, Milan, Italy. [6]Concordia University, Montreal, Quebec, Canada. [7]Cornell AI for Sustainability Initiative, Cornell University, Ithaca, NY, USA. [8]Cornell Atkinson Center for Sustainability, Cornell University, Ithaca, NY, USA. [9]Cornell University AI for Science Institute, Ithaca, NY, USA. e-mail: fengqi.you@cornell.edu

including manufacturing and end-of-life treatment[10,11]. The Scope 2 emissions from indirect energy purchases are expected to constitute a substantial portion and indicate a high dependency on the increase of AI server energy consumption. According to the International Energy Agency[12], 0.6% of global total carbon emissions comes from the data centres and data transmission networks due to their electricity consumption. The industry energy consumption could double by 2026, motivated by AI and other sectors[13], threatening decarbonization targets under the Paris Agreement[14,15], which include a 53% reduction in data-centre emissions by 2030 and net-zero goals for the AI sector. Further, growing AI server energy consumption implies an increasing water footprint through Scope 1 (direct cooling) and Scope 2 (indirect procurement) water use[8,16]. Centralized installation in water-stressed regions may perturb local water balance and threaten supply for millions[17,18].

Previous research has developed bottom-up and top-down methods to assess energy–water–carbon outcomes of servers[19–21], but these approaches face challenges with the rise of AI servers. Top-down approaches based on activity indices, such as data traffic and computing instances, fail to accurately represent AI-driven workloads[22,23]. Detailed bottom-up approaches suffer from limited data availability and lack of industry insight[7]. Assumptions valid for traditional collocation centres often fail for AI data centres, which differ in installation and operation[24]. Recent studies have explored computing task-based analyses to better quantify AI-related energy and resource consumption, providing insights but lacking systematic policy guidance[25,26]. Two notable contributions are the 2024 data-centre report by Lawrence Berkeley National Laboratory[27] and the 2025 Energy and AI Report by the International Energy Agency[28], which present scenarios estimating US and global data-centre energy use under highly uncertain AI growth. Although use of confidential commercial data may limit reproducibility, they offer important benchmarks. Our study extends this foundation by (1) developing an open-source, bottleneck-based modelling approach with comprehensive uncertainty analysis; (2) systematically assessing energy, water and carbon impacts, incorporating dynamic interactions with local energy systems; and (3) proposing actionable mitigation strategies for potential net-zero trajectories across different AI server deployment scenarios.

Here we analyse the combined energy–water–climate impact of operational AI servers in the United States between 2024 and 2030, balancing importance and future uncertainties and addressing a series of fundamental questions. (1) What are the magnitude and spatiotemporal distributions of energy consumption, water footprint and climate impact from AI server deployment? We address this using temporal projection models and regional frameworks, assuming deployment mirrors current large-scale AI data-centre patterns. (2) What are the prospects for near-term net-zero pathways? We evaluate this by analysing best- and worst-case scenarios of key drivers, including industry efficiency improvements, server location distribution and grid decarbonization. The results of these analyses are presented in the following sections.

AI servers' environmental impact in the United States

This section provides a base depiction of the AI servers' energy, water, and carbon impacts in the United States, emphasizing the spatiotemporal characteristics of the system. Figure 1a shows the projected cumulative AI server installations in the United States from 2024 to 2030 under five scenarios: low demand, low power, mid-case, high application and high demand. The mid-case serves as the base scenario, while the low and high demand set projection bounds. Low power assumes server efficiency gains, and high application accounts for increased adoption driven by efficiency. Figure 1b illustrates the state-level allocation of AI servers, showing power usage effectiveness (PUE), projected grid carbon intensity (carbon emissions per unit of electricity), water usage effectiveness (WUE) and projected grid water intensity (water footprint per unit of electricity) for each state. Southern states such as Florida exhibit higher PUE and WUE than northern states such as Washington, reflecting climate impacts. Moreover, the grid factors demonstrate notable sensitivity to location, implying the importance of the local grid for the AI servers' environmental impacts. Figure 1c–g shows energy, water and carbon results. Figure 1c illustrates AI server energy predominates over infrastructure energy. Figure 1d indicates indirect water footprint contributes 71% of total, with direct use at 29%. Annual estimates of energy consumption, water footprint and carbon emissions of AI servers from 2024 to 2030 under each scenario are presented in Fig. 1e–g. Even the lowest scenarios outline a considerable increase in the energy, water and carbon footprints of AI servers. The highest scenario yields the highest environmental impact, largely surpassing previous forecasts for the entire US data-centre market[18,29], underscoring the environmental risks of unchecked AI server expansion.

The United States is selected as the research region because of its dominant position in the global AI market. The research period is selected as 2024–2030, generated from a trade-off between importance and uncertainty. Projection of the AI server accumulative capacity in the United States is the initial step, which was generated on the basis of the forecast of AI chip manufacture capacity, AI server specifications and AI server adoption pattern. In addition, we assume the AI data centre will align with current AI company large-scale data-centre allocation ratios, presented in detail in Supplementary Fig. 5. The PUE and WUE values are derived by a hybrid statistical and thermodynamics-based model[30]. Supplementary Table 3 lists all model inputs, and the applied values are calculated as the average of the best and worst practices. In addition, projected grid factors are calculated on the basis of the Regional Energy Deployment System (ReEDS) model[31] by involving the projected data-centre load data and adopting regulations such as the Inflation Reduction Act. The step-by-step calculation process of the projection is summarized in Methods and sections 1–4 of Supplementary Information. Important uncertainties, such as PUE and WUE estimation, technology advancements, the spatial distribution of AI server allocation and grid development patterns, and sensitivity analysis of key parameters will be discussed in the following sections.

Higher energy and water usage efficiencies

Over the past decade, efficiency gains in the data-centre industry have stabilized environmental costs despite a doubling of computing instances[20]. This section examines the existing potential for further improvements through system optimization and technology adoption. Figure 2a,b illustrates the achievable PUE and WUE values of AI data centres in each state. The best-practice scenario suggests notable reduction potential: over 7% PUE reduction and over 85% WUE reduction, despite the high efficiency of AI data centres compared with the industry averages (PUE 1.58, WUE 1.8). The worst-practice scenario underscores the risk of neglecting efficiency efforts. The effects of achievable PUE and WUE values are further depicted in Fig. 2c,d, showcasing the corresponding influences on the energy, water and carbon footprints of AI servers. PUE reduction yields over 7% reductions in total energy consumption and carbon emissions. WUE reduction efforts result in over a 29% reduction in the total water footprint. Moreover, it is evident that PUE and WUE efforts can align with each other, as observed in WUE improvement results. Figure 2e delves into the potential impact of adopting advanced technologies within AI data centres, focusing on advanced liquid cooling (ALC) and server utilization optimization (SUO) adoption. The results show that the best ALC adoption can reduce about 1.7% of energy consumption, 2.4% of water footprint and 1.6% of carbon emissions of AI servers by 2030. For SUO, the best-case scenario, representing total adoption by 2030, results in a 5.5% reduction in all footprint values, while the worst-case scenario, representing frozen adoption, leads to a 7.3% increase by 2030. The energy–water–carbon

https://doi.org/10.1038/s41893-025-01681-y

Fig. 1 | Projections of energy, water and carbon footprints of the installed AI servers from 2024 to 2030. Each scenario is denoted by a different colour. **a**, The projected accumulative capacity of AI servers in the United States from 2024 to 2030 under different scenarios. **b**, Spatial allocation data for each state, accompanied by corresponding metrics, which include PUE, WUE, grid carbon factor ($kgCO_2$-equivalent kWh^{-1}), and grid water factor (L kWh^{-1}) data. The metric values are calculated as the average value from 2024 to 2030. **c**, The ratios between AI servers and infrastructure energy consumption. **d**, The ratios of the indirect and direct water footprint of AI servers. **e**–**g**, The energy consumption (**e**), water footprint (**f**) and carbon emissions (**g**) of AI servers from 2024 to 2030 under different scenarios. The red dashed lines in **e**–**g** denote the forecast footprint of the US data centres, based on previous literature[18,29].

https://doi.org/10.1038/s41893-025-01681-y

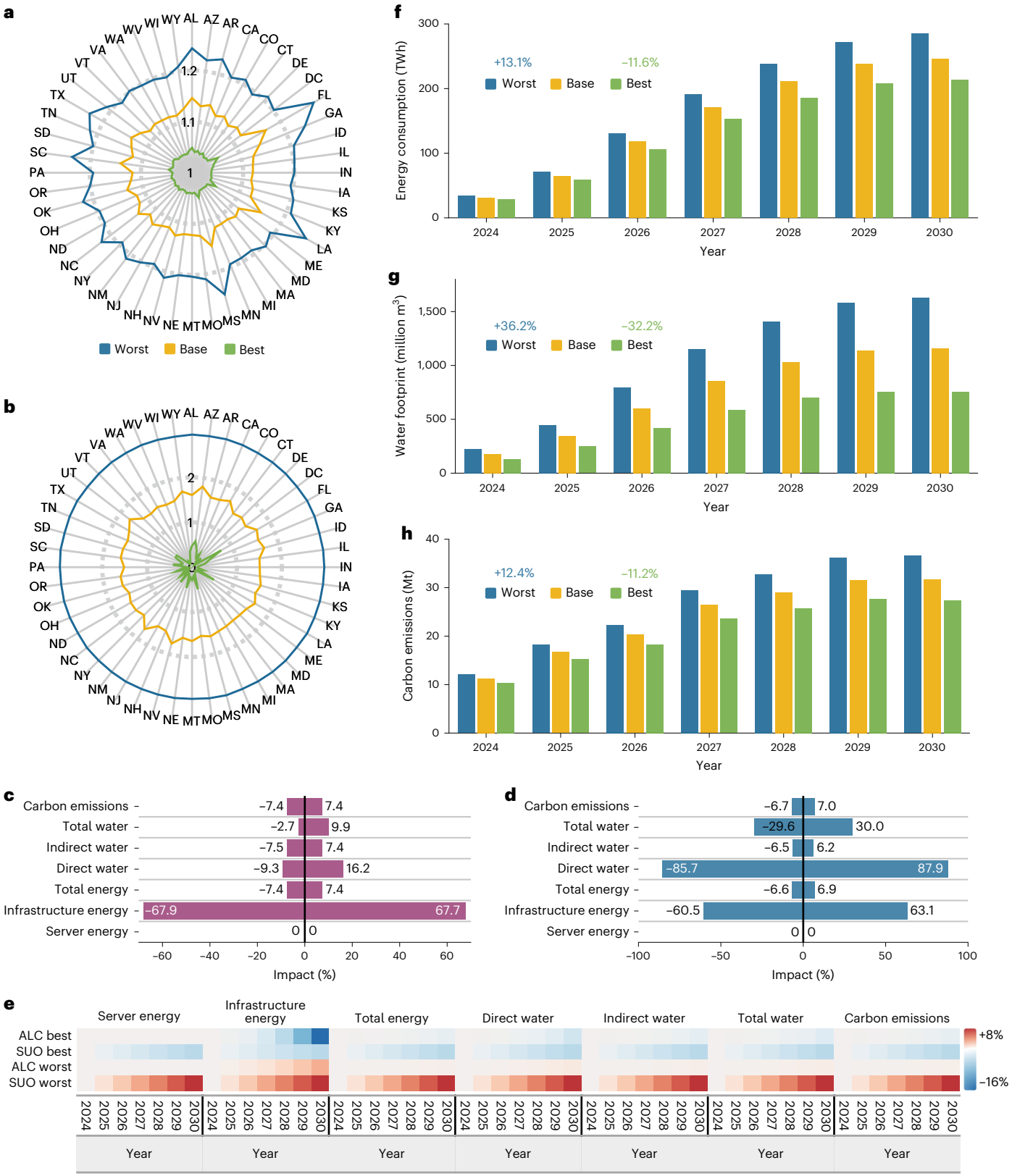

Fig. 2 | Assessment of industry efforts aimed at reducing the environmental impact of AI servers. a, Comparison of the best, base and worst practices regarding AI data-centre PUE. **b**, Comparison of the best, base and worst practices regarding AI data-centre WUE. **c**, Analysis of the impact of PUE practices on energy consumption, water footprint and carbon emissions. **d**, Analysis of the impact of WUE practices on energy consumption, water footprint and carbon emissions. **e**, Assessment of the effect of ALC and SUO adoption on energy consumption, water footprint and carbon emissions. **f**–**h**, The energy consumption (**f**), water footprint (**g**) and carbon emissions (**h**) of AI servers from 2024 to 2030 following the mid-case scenario through the worst, base and best practices of all considered industry efforts.

impacts of AI servers from 2024 to 2030 following the mid-case scenario through the worst, base and best industry practices are further presented in Fig. 2f–h. The maximum reductions of energy, water and carbon drawn from the existing potential are about 12%, 32% and 11%, respectively. These findings underscore the considerable impact of industry efforts on the environmental cost of AI servers.

Base values of PUE and WUE are calculated by averaging the best and worst practices due to the lack of cooling specifications. The worst and best values are derived by solving the corresponding optimization problem with constrained operational parameters. Specifically, the best PUE is achieved mainly by extending the free cooling period through input air set-points adjustment and enhancing facility energy efficiency. WUE is improved by reducing windage and concentration water loss, adopting air-side economizers and enabling more free cooling. The model is based on previous works[30], with relevant parameters detailed in section 3 of Supplementary Information. The best, base and worst practices for ALC and SUO adoption are developed from existing studies and market reports, with assumptions and calculations outlined in Methods. ALC adoption is evaluated through the increased immersion cooling in AI data centres, while SUO improvements focus on raising the active server ratio. Importantly, the advancement in AI hardware could reduce energy, water and carbon footprints by improving energy efficiency, as seen in Nvidia's future chip structures Blackwell and Rubin. However, these gains may be offset by the rebound effect. The uncertainties of the hardware evolution and market dynamics are discussed in the definition of the scenarios, as detailed in section 1 of Supplementary Information.

Influences of AI server spatial distribution

The location of data centres critically shapes their environmental impact[18,32]. This section presents a detailed projection of how AI server spatial distribution may influence the environmental consequences of rapid US expansion. Figure 3a,b presents the top 25%, 50% and 75% locations with the lowest projected water footprint and carbon emissions per unit of server energy. Figure 3c presents the locations with combined lowest water and carbon factors. These allocation strategies are conducted on a grid balancing-area level. Specifically, the ReEDS model is deployed to calculate the grid factors of each balancing area under different projection scenarios. Figure 3d shows that Texas plays a vital role by possessing the most balancing areas with the top 25%, 50% and 75% water and carbon factors. The other western and southern states, including Montana, Louisiana, Idaho and New Mexico, also take a large portion of areas under the combined strategies due to widely leveraged local renewables. West Coast states, such as California, Oregon and Washington, as well as New England states, are suitable for carbon reduction but also lead to higher water footprints. The primary driver behind is the adoption of hydropower, which consumes large volumes of water through evaporation, as detailed in Supplementary Fig. 6. With more hydropower applied in the grid, the Scope 2 water usage of AI servers could dramatically increase when installed in such locations. Figure 3e–g depicts the total energy consumption, water footprint and carbon emissions changes of AI servers under the 25%, 50% and 75% spatial distribution scenarios. The base scenario is applied to all simulations to identify solely the influences of spatial distribution. Installing AI servers in the top water-oriented locations results in large reductions of both water and carbon footprints. Conversely, the carbon-oriented strategies lead to higher water footprints due to the hydropower application on the West Coast and New England states. Moreover, Fig. 3f shows that the water and carbon footprints can be concurrently reduced following a combined allocations strategy.

To project the potential risks and benefits behind future AI server deployment, Fig. 3h,i illustrates the current water scarcity and renewable energy potential of each state. The top ten states grappling with severe water scarcity issues are California, Nevada, Arizona, Utah, Washington, New Mexico, Colorado, Wyoming, Oregon and Montana, which are concentrated primarily in the western United States. Large adoption of hydropower leads to increased unit energy–water footprint in several states such as Arizona, California, Nevada, New Mexico and Utah, exacerbating the water scarcity issue. In addition, Texas, New Mexico, Kansas, Arizona, California, Colorado, Nevada, Nebraska, Oklahoma and South Dakota are identified as the top ten states with abundant renewable energy potentials. Combined with the water- and carbon- oriented strategies, Texas, Montana, Nebraska and South Dakota, situated in the Midwestern United States, emerge as optimal candidates for AI server installation, considering both water scarcity concerns and future decarbonization efforts.

Influence of renewable energy penetration within grid

This section analyses how future grid development will affect AI server environmental impacts. Figure 4a,b illustrates the modifications of grid carbon and water factors of each state under low renewable energy cost (LRC, best-practice) and high renewable energy cost (HRC, worst-practice) scenarios, compared with the base scenario. These scenarios, extracted from the predefined cases of the ReEDS model[31,33], represent the highest and lowest levels of renewable energy penetration, respectively. The resulting total carbon emissions and water footprint of AI servers are depicted in Fig. 4c,d. The HRC scenario indicates a 20% increase in carbon emissions alongside a 2.0% increase in water footprint, while the LRC scenario suggests an over 15% reduction in carbon emissions accompanied by a 2.5% of water footprint reduction. The carbon emissions of AI servers are shown to be heavily influenced by the grid decarbonization pattern, indicating both considerable reduction potential and associated risks.

Figure 4e details changes in AI server carbon emissions and water footprints by state under LRC and HRC scenarios. The development pattern, including renewable energy penetration levels and sources, notably impacts the resulting footprints of each state. States such as Georgia, Nevada, North Carolina and Tennessee show marked sensitivity to renewable cost scenarios. In addition, the Pacific states, including California, Oregon and Washington, which achieved a low grid carbon factor, slow their hydropower adoption pace under the LRC scenario, avoiding exacerbating their water scarcity issues with additional AI server installations. The wind and solar resources, which have no water consumption while generating, are further adopted in the LRC scenario to reduce carbon emissions, meanwhile alleviating the severe water scarcity challenges in these states. The presented results suggest that the impacts of grid decarbonization patterns on the environmental costs of AI servers are evident not only over time but also in the spatial variations observed among different US states. These results connect to the importance of ambitious green electricity policies in US states, which potentially could further reduce carbon emissions in our scenarios.

Pathways to net-zero carbon and water goals

Building on the preceding analysis of environmental costs, efficiency, spatial distribution and grid decarbonization, this section evaluates water and carbon net-zero pathways for US AI servers. Figure 5a presents the pathways from 2024 to 2030 to achieve net-zero carbon emissions and a net-zero water footprint of AI servers under the mid-case scenario. Figure 5b presents residual emissions and water footprints across different scenarios and practices. The 2030 target aligns with major AI data-centre operator goals[14,34,35]. The top and bottom 25% of locations are used to create the best and worst cases for spatial distribution. The best and worst grid development practices are modelled under LRC and HRC scenarios, respectively. These practices form the upper and lower bounds, although better solutions can be obtained through extra policies and actions beyond current considerations.

The best and worst distribution patterns of AI server deployment lead to a 49% reduction and a 90% increase in carbon emissions, and

https://doi.org/10.1038/s41893-025-01681-y

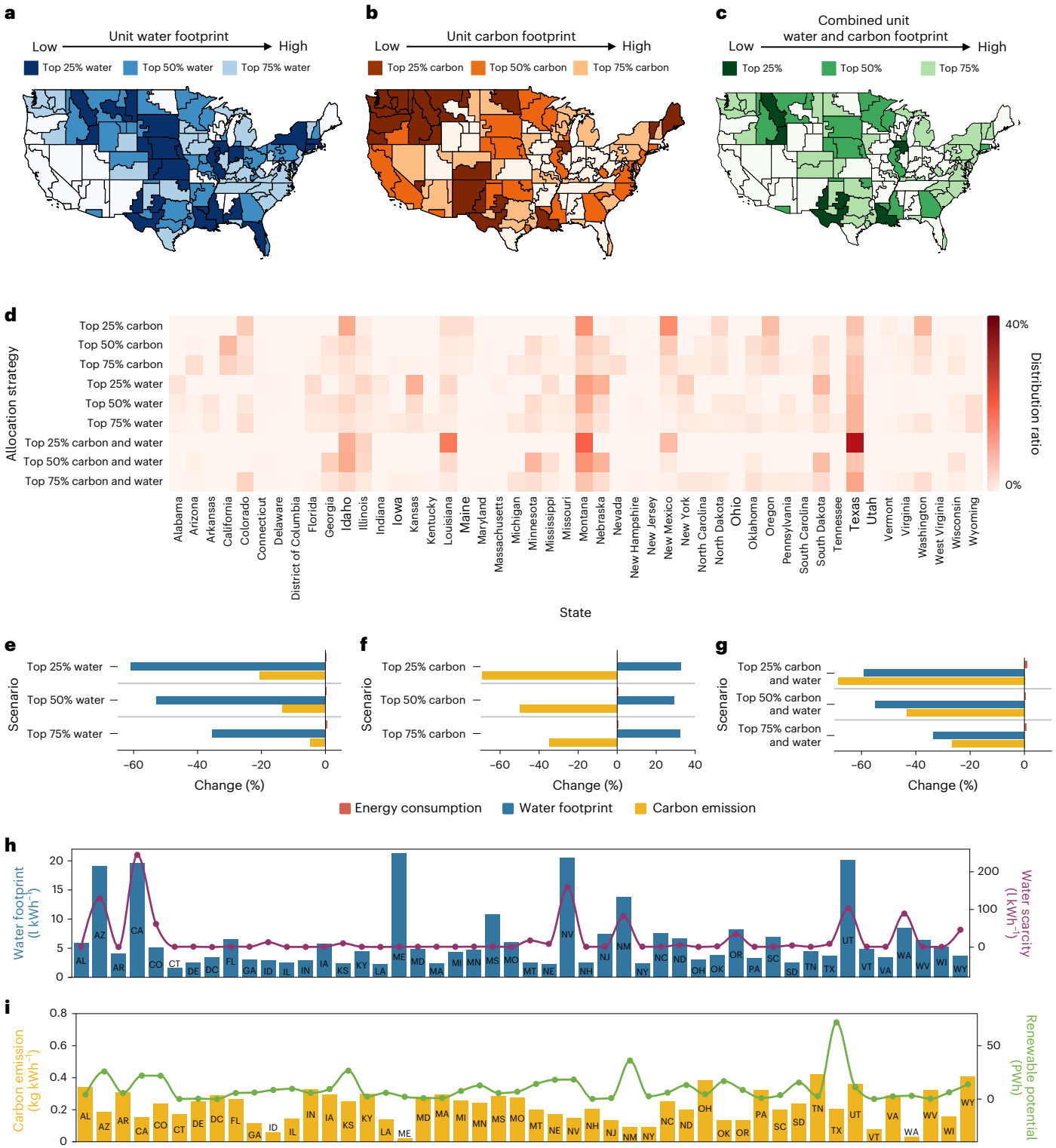

Fig. 3 | Impact of spatial distribution on water footprint and carbon emissions of AI servers. a, The top 25%, 50% and 75% of locations with the lowest projected water footprint per unit energy (2024–2030). **b**, The top 25%, 50% and 75% of locations with the lowest projected carbon emissions per unit energy. **c**, Locations ranked in the top 25%, 50% and 75% for both lowest water and carbon per unit energy. **d**, State area ratios under water- and carbon-oriented allocation strategies. **e**, Changes in total energy–water–carbon footprints for the 25%, 50% and 75% water-oriented scenarios. **f**, Changes in total water–carbon footprints for the 25%, 50% and 75% carbon-oriented scenarios. **g**, Changes in total energy–water–carbon footprints for the 25%, 50% and 75% combined water- and carbon-oriented scenarios. **h**, Current water footprint and water scarcity per unit server energy by state. **i**, Current carbon emissions per unit server energy and renewable energy potential by state.

https://doi.org/10.1038/s41893-025-01681-y

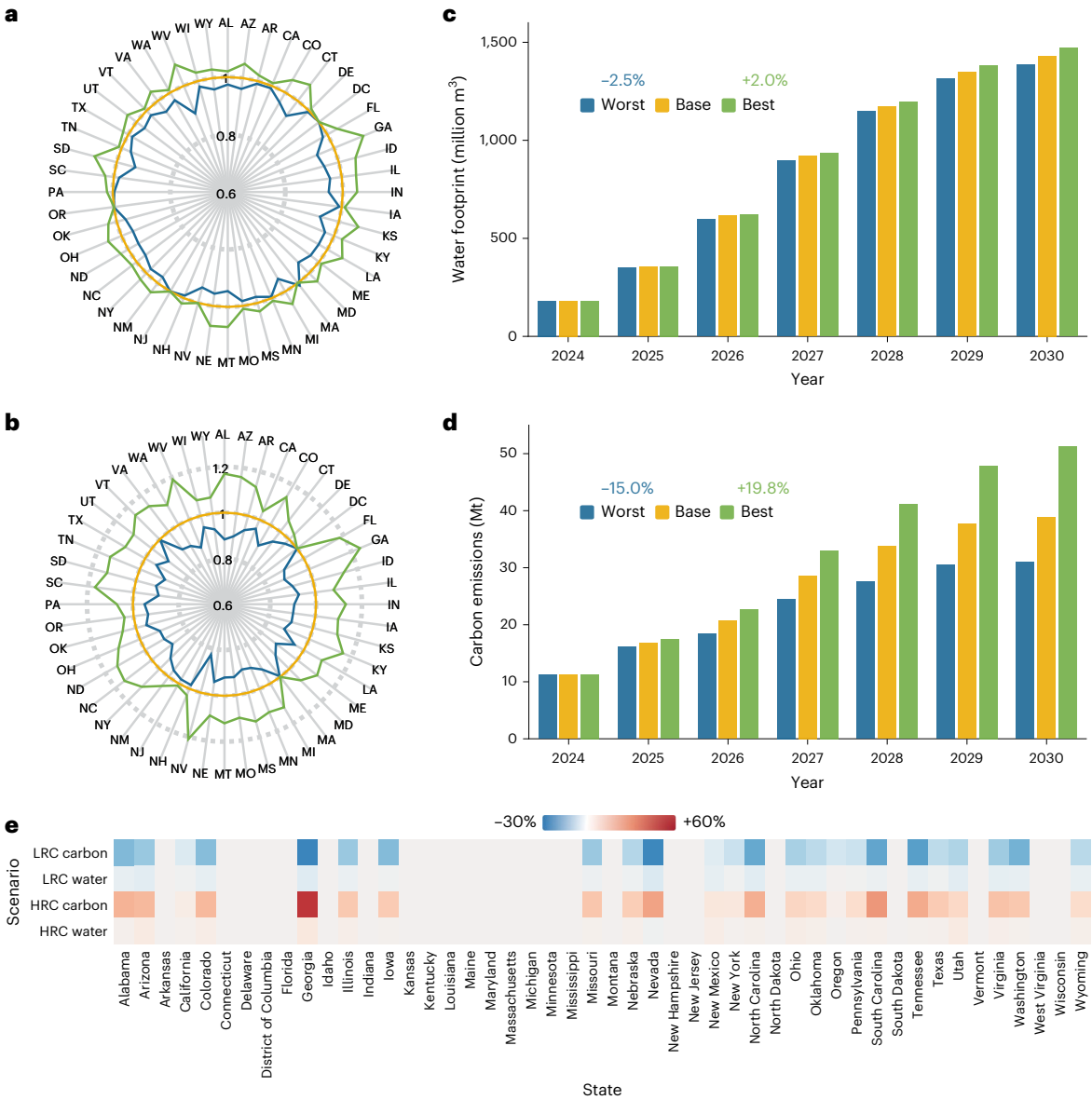

Fig. 4 | Assessment of the impact of grid renewable energy penetration on water footprint and carbon emissions of AI servers. a, Modifications of the grid water factor under the best and worst scenarios compared with the base scenario. **b**, Modifications of the grid carbon factor under the best and worst scenarios compared with the base scenario. **c**,**d**, The water footprint (**c**) and carbon emissions (**d**) of AI servers from 2024 to 2030 following the mid-case scenario under best, base and worst scenarios, respectively. **e**, The changes of AI server carbon and water footprints of each state under the best (LRC) and worst (HRC) scenarios.

a 52% reduction and a 354% increase in water footprints, respectively. Optimistic grid decarbonization provides a further 13% reduction in carbon, while the worst case results in a 23% increase. Industry efficiency efforts are critical for water sustainability, offering over 32% reduction in water footprint under best practices. Notably, combined best practices cut residual emissions and water footprints by 73% and 86%, respectively, indicating a feasible pathway to net zero. Under the mid-case, best industry, spatial and grid scenarios each reduces over 21 Mt, 25 Mt and 92 Mt of carbon from a base of 186 Mt due to AI server installation. The best practices produce about 11 Mt residual carbon emissions by 2030, requiring 28 GW of wind or 43 GW of solar to fully offset[36]. With over 13 GW of AI company renewables already claimed[37], residual emissions could remain manageable with further expansion. However, achieving this best-case scenario could be extremely challenging, particularly due to the facility constraints in deploying AI servers at optimal locations and the difficulty in reaching ideal energy and water efficiency levels within AI data centres. Conversely, the worst practices pose a risk of unachievable net-zero pathways, signifying 71 Mt annual residual carbon emissions and over 5,224 million m^3 annual residual water usage by 2030 under the low-demand scenario. Such an environmental cost is nearly impossible to be fully compensated during a short period.

Projected net-zero pathways could be further influenced by other uncertainties. Figure 6 presents a sensitivity analysis of key factors: server lifetime, AI server manufacturing capacity, US allocation ratio, server idle and maximum power ratios, and training/inference distributions. Low, applied and high values of each parameter are listed in the figure; left and right halves present the effects of lower and higher values, respectively. Changes in these factors result in up to 40% variations in energy consumption, which are closely mirrored by corresponding shifts in water footprint and carbon emissions. Unmodelled uncertainties can largely be reflected by modifying these simulation

https://doi.org/10.1038/s41893-025-01681-y

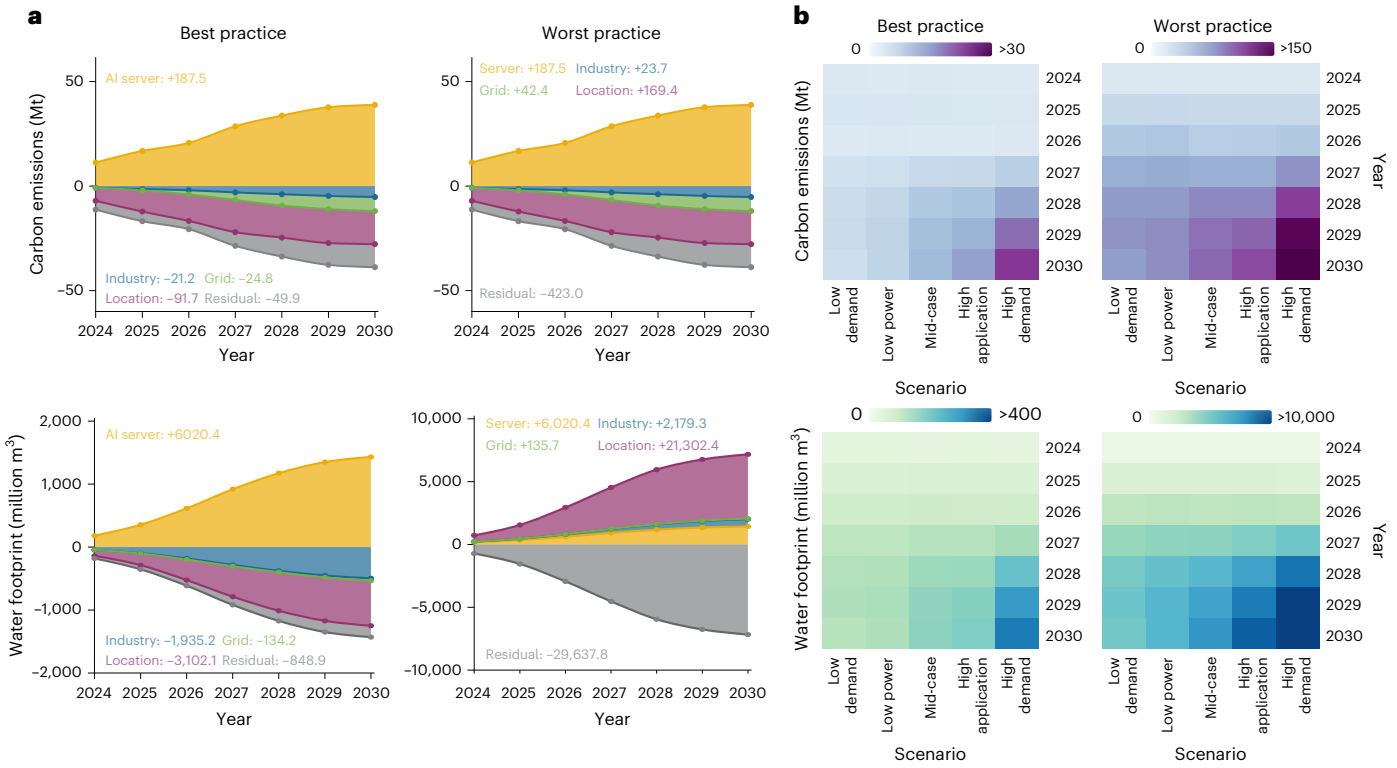

Fig. 5 | The pathways towards achieving net-zero carbon emissions and water footprints for US AI servers. a, The contributions of each influential factor to the water footprint and carbon emissions of AI servers with best and worst practices under the mid-case scenario. The total contributions of each factor from 2024 to 2030 are also listed. The curves above the 0 level indicate the increase of carbon emissions and water footprint, while the curves below represent the reductions. The grey area represents the residual footprints that need to be reduced. **b**, Presentation of the capacity of residual carbon emissions and water footprints to attain net-zero targets under different temporal scenarios for each year spanning from 2024 to 2030, under both best- and worst-practice scenarios. Specifically, the top and bottom 25% of locations are used to create the best and worst cases for spatial distribution. The best grid development practice is modelled under the LRC scenario, and the worst is modelled under the HRC scenario.

factors. For example, innovations such as DeepSeek can lead to low power requirements for AI computing tasks, which then may result in a rebound effect incorporating more AI applications. This statement may imply larger chip-on-wafer-on-substrate (CoWoS) manufacturing capacity due to expanding applications as well as longer lifetime and smaller training/job ratio due to more inference-based usage. As Fig. 6 indicates, the study's key conclusions could remain robust unless future uncertainties greatly exceed modelled ranges. The modelling approach used in this study enables future revisions with more available data. The sensitivity analysis aims to capture the potential uncertainties inherent to this problem, offering insights into the magnitude of their influence on projection.

Discussion

Investment in AI servers is accelerating, as seen in projects such as the $500 billion Stargate[3]. While AI advancement is a key priority, our study highlights its environmental impact. To mitigate these risks, we identify that concentrating AI server deployment in Midwestern states—Texas, Montana, Nebraska and South Dakota—is optimal, given their abundant renewables, low water scarcity and favourable projected unit water and carbon intensities. These states possess substantial untapped wind and solar resources, enabling robust green power portfolios and reducing competition with other sectors. Their lower water stress also helps ease public concerns and reduces the need for costly water-saving measures. However, several implementation challenges must be acknowledged. Texas, crucial to the optimal strategy, may need to support an additional 74–178 TWh of AI server demand, possibly exceeding its current total renewable generation of 139 TWh (ref. 38). This scale-up would require substantial investment in new renewable capacity and transmission infrastructure, which is already constrained by existing congestion[39]. Meanwhile, Montana, Nebraska and South Dakota currently host minimal data-centre capacity, indicating that most of their existing internet infrastructure may support only residential or standard industrial applications. This raises potential connectivity and security concerns for high-performance AI services. Enabling AI-grade infrastructure in these regions will require broadband and security upgrades, with associated emissions and capital costs increasing the expense of the optimal strategy. These challenges underscore the complexity of sustainable siting decisions and highlight the need for strategic coordination to ensure that AI investments support both technological leadership and long-term sustainability goals.

Expanding AI servers in the suggested regions may be further obstructed by public health concerns and other impacts on local sustainable development. Operational demands and construction of supporting facilities could generate air and water pollution through fossil fuel use, substantial water consumption and large-scale construction and transportation. To address these issues while managing economic costs, we recommend that AI companies engage in public–private partnerships with local governments. Such partnerships could alleviate governmental budget pressures, create local jobs and mitigate health concerns by funding green power upgrades and monitoring systems, such as tracking $PM_{2.5}$ (particulate matter with a diameter of 2.5 micrometres or less) levels, through combined AI investment and regulation. For the governments, implementing tax exemptions on a related facility could incentivize a win–win development process, and it is essential to ensure transparent accountability within public–private partnerships to prevent privatized gains at public cost.

https://doi.org/10.1038/s41893-025-01681-y

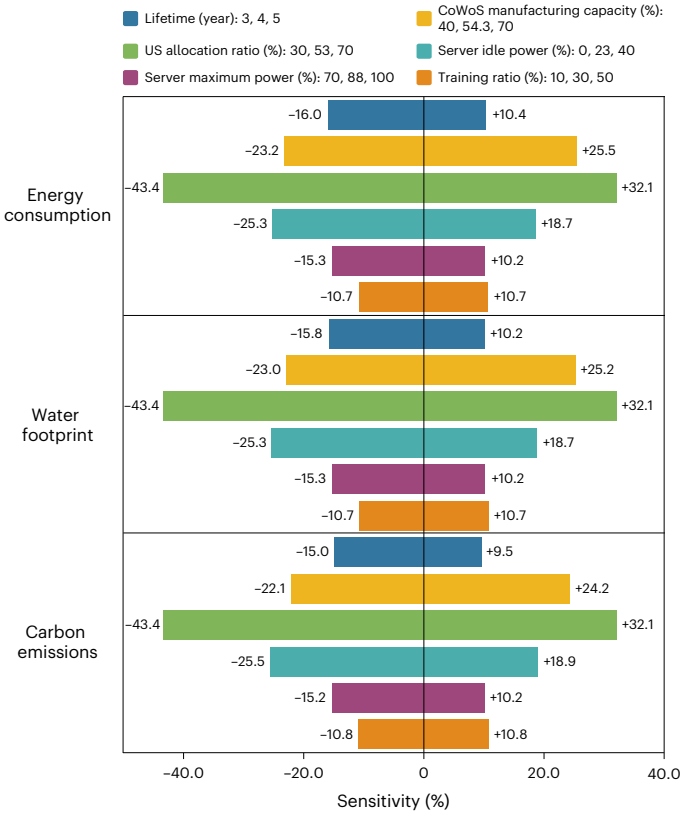

Fig. 6 | Sensitivity of AI server energy–water–carbon impacts to key modeling assumptions. Sensitivity analysis for AI server energy consumption, water footprint and carbon emissions considering uncertainties of server lifetime, manufacturing capacities for AI servers, US allocation ratio, server idle power ratio, server maximum power ratio and training/inference distributions. The listed numbers for each parameter from left to right represent the considered low, applied and high values.

Beyond environmental benefits, these measures could stimulate local economic growth, as evidenced in Virginia[40]. This discussion outlines policy recommendations for both industry and legislative stakeholders that balance economic and environmental impacts. They also emphasize energy market adaptations as AI-driven hyperscale loads challenge utilities' 'duty to serve'. Requirements to connect large data centres could lead to server collocation with energy generation, but as shown in this study, these choices should not crowd out renewable investments to decarbonize the economy. Finally, the effectiveness of the recommended policy may be influenced by broader political and economic factors beyond the scope of this study.

The challenges associated with the optimal spatial distribution of AI servers introduce substantial uncertainties to the projected reductions of 73% and 86% in carbon and water footprints, respectively, under best-practice scenarios. These challenges are unlikely to be offset by improvements in efficiency or decarbonization efforts beyond what has already been considered. Our best-case scenario for industry efficiency approaches the physical limits of AI data centres. Moreover, projections from the US Energy Information Administration offer limited support for additional grid decarbonization compared with the considered best case[41]. These constraints suggest that, without additional interventions, AI data centres are likely to generate substantial environmental impacts in the coming years. Consequently, AI companies are expected to continue investing in the offset mechanisms, including power purchase agreement, carbon removal and water restoration, reaching an unprecedented level of reliance on them to accomplish their net-zero aspirations by 2030. Nevertheless, severe concerns could be raised regarding the complexity of securing long-term contracts and providing convincing reduction credits[9,42], especially bearing in mind the considerable capacity that would be needed. We argue that AI companies should shift to a more transparent approach by closely cooperating with third-party verification groups, service providers and governmental agencies. Such collaboration could reduce uncertainties through better integration of social resources and serve as a model for other electrified sectors.

Emerging high-efficiency technologies in hardware and software, exemplified by DeepSeek[43], may fundamentally transform AI server supply and demand. As reflected in the study's five scenarios, these innovations may cause deviations of up to 393 million m^3 in water footprints and 20 $MtCO_2$-equivalent in emissions between minimum and maximum impact cases, underscoring the need for tailored managements. While efficiency gains may reduce cost per computing task, they risk a rebound effect, where lower costs increase total application volume. This dynamic, as reflected by the high-application scenario, may amplify total demand and complicate AI's environmental trajectory. To address these uncertainties, we recommend government agencies work with industry to establish real-time monitoring systems, enabling timely alerts and proactive measures before considerable environmental impacts occur. Moreover, the potential increase in total computing jobs poses both challenges and opportunities, calling for ongoing enhancements in energy and water efficiency through system optimization and adoption of strategies such as SUO and ALC to manage the added workload complexity and flexibility. Therefore, we also suggest the data-centre industry establish AI-specific benchmarks for energy, water and carbon performance, which is crucial for continuous operational efficiency gains.

Methods

The methodology framework of this study aims to achieve two goals: (1) draft the energy–water–climate impacts of AI servers in the United States from 2024 to 2030 to handle the massive concerns about AI developments, and (2) identify the best and worst practices of each influencing factor to scheme the net-zero pathways for realizing water and climate targets set for 2030. Compared with many previous climate pathway studies, which often extend predictions to 2050 for better integrating climate goals, this study focuses on the period from 2024 to 2030 due to the great uncertainties surrounding the future of AI applications and hardware development. For assessing these uncertainties, scenario-based projections are first constructed to obtain potential capacity-increasing patterns of AI servers. Technology dynamics, such as SUO and ALC adoption, are defined with best, base and worst scenarios, and a similar method is employed to capture the impact of grid decarbonization and spatial distribution. The utilized models and data required during the calculation process are illustrated in the following sections. More details on model assumptions and data generation are provided in sections 1–4 of Supplementary Information.

Data description and discussion

This section provides a comprehensive overview of the data used in this study. Historical DGX (Nvidia's high-performance AI server line) parameters were sourced from official documentation, and future scenarios were projected on the basis of historical configurations and current industry forecasts. To attain the units of AI servers, we collected the most updated industrial report data for projecting the future manufacturing capacity of CoWoS technology, which is the bottleneck for top-tier AI server production. The data resources of the preceding process have been introduced and validated in section 1 of Supplementary Information. AI server electricity usage was assessed using recent experimental data on maximum power[44,45], idle power[44,46] and utilization rate[46–49], derived from existing AI server systems. PUE and WUE values for AI data centres across different locations were calculated using

operational data from previous studies[30,50] and industrial resources[51], combined with the collected average climate data for each state[52]. The allocation ratios of AI servers to each state were determined on the basis of configurations of existing and planned AI data centres, which are collected from reports of major AI companies in the United States, as data resources detailed in section 2 of Supplementary Information. In addition, projections for grid carbon and water factors were derived from the ReEDS model[31], using its default scenario data[33]. All datasets employed in this study are publicly available, with most originating from well-established sources. A key uncertainty lies in estimating the number of manufactured AI server units, as official supply-chain reports remain largely opaque. To maintain transparency and ensure reproducibility, we rely on the best available industry reports rather than commercial sources such as International Data Cooperation data[53], which are not granted for open access and would limit future validation despite their potential to provide better estimates. The validations of applied data are further detailed in sections 1 and 4 of Supplementary Information.

AI server power capacity projections

The energy consumption of AI servers is projected to be driven predominantly by top-tier models designed for large-scale generative AI computing[6,7]. This trend is attributed to their substantial power requirements and the increasing number of units being deployed. In this study, we estimate the power capacity of these high-performance AI servers by examining a critical manufacturing bottleneck: the CoWoS process[54]. This process, which is controlled nearly exclusively by the Taiwan Semiconductor Manufacturing Company, serves as a key determinant of the manufacturing capacity for AI servers in recent years[55]. Our analysis uses forecast data and projection assumptions of the CoWoS process to estimate total production capacity. Other factors are integral to translating this capacity into the power capacity of AI servers: the CoWoS size of AI chips, which determines how many chips can be produced by each wafer; the rated power of future AI servers, which reflects the power demand per unit; and the adoption patterns of AI servers, which dictate the mix of various server types over time. The values of these factors are derived mainly from the DGX systems produced by Nvidia, which is the dominant product for the top-tier AI server markets[56].

Considering the influencing factors for the total AI server capacity shipments and existing uncertainties, we generate distinct scenarios as follows:

- Mid-case scenario: the CoWoS capacity is projected to slightly increase after 2026, consistent with the growth rate in 2023. Under this scenario, AI servers' rated power is expected to have a linear relationship with the anticipated die size increase while adoption patterns remain aligned with current trajectory.
- Low-demand scenario: characterized by lower CoWoS capacity growth and lower AI server rated power compared with the mid-case scenario, this reflects a scenario of lower overall demand for AI servers.
- Low-power scenario: maintains the same assumptions as the mid-case scenario but with lower AI server rated power, representing efficiency gains in AI hardware and software development.
- High-application scenario: assumes lower AI server rated power alongside high CoWoS capacity, capturing the potential rebound effect where efficiency gains drive increased AI workload deployment.
- High-demand scenario: features higher CoWoS capacity expansion, higher AI server rated power and higher adoption of new servers compared with the mid-case scenario, reflecting a scenario of strong AI server demand.

Based on the assumptions and scenarios outlined, the annual projections for top-tier AI server shipments and their average rated power are calculated as follows:

$$N_{\mathrm{AI}} = \frac{C_{\mathrm{CoWoS}} \times R_{\mathrm{Nvidia}} \times \sum_i R_i n_i}{N_{\mathrm{GPU}}}$$
$$\bar{P}_{\mathrm{AI}} = \sum_i R_i P_i \quad (1)$$

where N_{AI} and $\bar{P}_{\mathrm{AI}}$ represent the annually projected shipments and average rated power of the top-tier AI servers. R_{AI} is the ratio of CoWoS capacity allocated to top-tier AI servers and is set as 40% for 2022, 40.7% for 2023, 48.5% for 2024 and 54.3% for 2025, according to industry reports[57,58]. For years beyond 2025, this ratio is assumed to remain constant at the 2025 value due to a lack of further data. The sensitivity analysis regarding this value is provided in Fig. 6. C_{CoWoS} is the projected CoWoS capacity within each scenario. N_{GPU} is the number of graphic processor units (GPUs) per server and is set as 8, reflecting the configuration of most commonly used AI server systems[59]. In addition, R_i, n_i and P_i represent the projected adoption ratio, units yield per CoWoS wafer and rated power of the ith type of chip at each year, respectively. The details of the projections and related data resources are provided in section 1 of Supplementary Information, Supplementary Figs. 1–4 and Supplementary Table 1.

AI server electricity usage calculation

The applied AI server electricity usage model is a utilization-based approach initially derived from CPU (central processing unit)-dominant servers[60] and can be written as the following:

$$P_{\mathrm{server}} = (P_{\mathrm{max}} - P_{\mathrm{idle}})u + P_{\mathrm{idle}} \quad (2)$$

The preceding model assumes the total server power has a linear relationship with the processor utilization rate u. While this relationship has been well validated for CPU machines, its application to GPU utilization is less established except for a few cases[61]. However, several recent studies have shown a strong correlation between GPU utilization and overall server power consumption when dealing with AI workloads[44,46], indicating that GPUs are the dominant contributors to energy use in AI servers[45]. Although systematic experimental validation specific to GPUs is still limited, the consistency of findings across various case studies supports the assumption that the linear relationship applies here as well. The maximum power P_{max} and idle power P_{idle} are generated on the basis of the recent DGX system experimental results, and their values are set as 23% and 88% of the server rated power, respectively[44,46]. The sensitivity analysis was conducted to quantify the uncertainty, as shown in Supplementary Fig. 6. Moreover, the GPU processor utilization u is calculated as the following:

$$u = u_{\mathrm{active}} \times r_{\mathrm{active}} \quad (3)$$

where u_{active} and r_{active} represent the average processor utilization of active GPUs and the ratio of active GPUs to total GPUs, respectively. Note that the u_{active} and r_{active} commonly have higher values during training compared with inference[62]. Specifically, we use currently available AI traces, including Philly trace[47], Helios trace[46], PAI trace[48] and Acme trace[49], to determine the r_{active} for training and inference tasks. These traces provide comprehensive analyses on the relationship between GPU utilization rate and job characteristics. Based on the data provided in these works, the r_{active} is set as 50% and 90% for inference and training, respectively. Moreover, the u_{active} values are further determined on the basis of recent experimental studies[44]. The values are set as 50% and 80% for inference and training, respectively. Therefore, the processor utilization rates for inference and training in this work are set as 25% and 72%, respectively. Following the previous works[63,64], our base estimations assume 30% of computing capacity for training and

70% for inference. A detailed sensitivity analysis on the impact of these utilization rate settings is provided in Fig. 6.

Assessment of the environmental footprints of AI servers

This study employs a state-level allocation method to evaluate the energy, water and carbon footprints of AI servers. To capture the current and future distributions of AI server capacity, we compiled data of current and in-construction large-scale data centres belonging to major purchasers of top-tier AI servers, including Google, Meta, Microsoft, AWS, XAI and Tesla. The analysis incorporates the location, building area and construction year of each data centre to calculate the state-level distribution of server capacity by annually aggregating the total building area for each state. On the basis of our calculations, no major changes in spatial distribution are projected between 2024 and 2030, even with the anticipated addition of new data centres. Therefore, we assume the current spatial distribution will remain constant from 2024 to 2030 to account for uncertainties in directly integrating the projected contributions of in-construction data centres. Further details on the methodology and spatial distribution results are provided in section 2 of Supplementary Information.

For each state, the actual energy consumption can be derived from the server electricity usage and the PUE value of AI data centres. Meanwhile, the water footprint and carbon emissions should be analysed across three scopes. Scope 1 encompasses the on-site water footprint, calculated on the basis of on-site WUE (shortened as WUE in this work) and on-site carbon emissions (typically negligible for data centres[18]). Scope 2 includes off-site water footprint and carbon emissions, which are contingent on the local grid power supply portfolio. Scope 3, representing embodied water footprint and carbon emissions during facility manufacturing, lies beyond the spatial scope of this study. A regional PUE and WUE model, following the idea in previous research[30,50], is applied to estimate the PUE and WUE values of AI data centres in different states. This hybrid model integrates thermodynamics and statistical data to generate estimations on the basis of local climate data. Specifically, we collected the average climate data of each state between 2024 and 2030 from an existing climate model[52], which is then employed in calculating the PUE and WUE values of each state. Considering that the specific cooling settings for AI data centres are unknown, the base values are calculated by averaging the worst and best cases. The model parameters are detailed in Supplementary Table 2. Subsequently, the Scope 2 water footprint and carbon emissions are calculated on the basis of the grid water and carbon factors derived from the ReEDS model[31]. This approach also allows us to incorporate the projected data-centre load data, which can further interact with the grid system through services such as demand response. The validation of the ReEDS model results by using current high-resolution data is presented in Supplementary Figs. 7 and 8, and the related discussion is presented in section 4 of Supplementary Information. Optimization and analytical techniques are employed to determine optimal parameters during the simulation to generate the best and worst practices concerning industrial efficiency efforts, spatial distributions and grid decarbonization. Moreover, the water scarcity and remaining renewable energy potential data of each state are computed on the basis of the calculated environmental cost and standard data from previous literature[18,65]. The preceding calculation process depends mainly on previously established approaches, and its integration into our framework is further discussed in sections 3 and 4 of Supplementary Information.

Uncertainties and limitations

There are substantial uncertainties inherent in projecting the evolution of AI servers. Our analysis presents a range of scenarios based on current data to evaluate the impacts of data-centre operational efficiency, spatial distribution and grid development. However, several key uncertainties remain an unmodelled field for this work. For a better understanding of our study and to outline future research directions, these uncertainties are categorized as follows:

- Model and algorithm innovations: the model and algorithm breakthroughs in the AI industry could fundamentally alter computing requirements.
- Supply-chain uncertainties: the complex production process of AI servers may reveal new bottlenecks beyond the current CoWoS technology, leading to varying expansion patterns.
- Hardware and facility evolutions: continued improvements in AI computing hardware and data-centre efficiency may substantially affect the environmental impact of these servers.
- Out-of-scope factors: there are other major contributors that are out of the scope of this study and would be critical to the process, such as market forces and geopolitical influences.

The impacts of these factors are multifaceted and challenging to model with existing data. For example, while the recent release of DeepSeek has been interpreted as reducing the energy demands of AI servers, it may also trigger a rebound effect by spurring increased AI computing activity, ultimately resulting in higher overall energy, water and carbon footprints[43]. However, no fresh data have become available to simulate this complex process, based on our best knowledge when drafting. To further assess the influence of unpredictable uncertainties, we conducted a sensitivity analysis on key factors, including manufacturing capacities for AI servers, US allocation ratios, server lifetimes, idle and maximum power ratios and training/inference distributions. As shown in Fig. 6, our findings suggest that the key conclusions of this study are expected to remain robust as long as the impact of future uncertainties does not notably exceed the ranges considered. Given the highly dynamic nature of AI evolution, our modelling approach allows for future revisions as more data become available on potential shifts in industry trends.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data and material used in the analysis are available in our GitHub repository: https://github.com/PEESEgroup/US-AI-Server-Analysis. Source data are provided with this paper.

Code availability

All codes used in the analysis are available in our GitHub repository: https://github.com/PEESEgroup/US-AI-Server-Analysis

References

1. Guan, L. Reaching carbon neutrality requires energy-efficient training of AI. *Nature* **626**, 33–33 (2024).
2. NVIDIA Blackwell platform arrives to power a new era of computing. *Nvidia Newsroom* https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing (2024).
3. Announcing the Stargate Project. *OpenAI* https://openai.com/index/announcing-the-stargate-project/ (2025).
4. Kaack, L. H. et al. Aligning artificial intelligence with climate change mitigation. *Nat. Clim. Change* **12**, 518–527 (2022).
5. Vinuesa, R. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nat. Commun.* **11**, 233 (2020).
6. de Vries, A. The growing energy footprint of artificial intelligence. *Joule* **7**, 2191–2194 (2023).
7. Crawford, K. Generative AI's environmental costs are soaring—and mostly secret. *Nature* **626**, 693 (2024).

8. Krämer, K. AI & robotics briefing: data centres' huge 'water footprint' becomes clear amid AI boom. *Nature Briefing* https://www.nature.com/articles/d41586-023-03768-y (2024).
9. Cao, Z., Zhou, X., Hu, H., Wang, Z. & Wen, Y. Toward a systematic survey for carbon neutral data centers. *IEEE Commun. Surv. Tutor.* **24**, 895–936 (2022).
10. Malmodin, J., Lundén, D., Moberg, Å, Andersson, G. & Nilsson, M. Life cycle assessment of ICT: carbon footprint and operational electricity use from the operator, national, and subscriber perspective in Sweden. *J. Ind. Ecol.* **18**, 829–845 (2014).
11. Alissa, H. et al. Using life cycle assessment to drive innovation for sustainable cool clouds. *Nature* **641**, 331–338 (2025).
12. Data centres and data transmission networks. *IEA* https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks (2024).
13. *Electricity 2024: Analysis and Forecast to 2026* (IEA, 2024); https://www.iea.org/reports/electricity-2024
14. Walsh, N. How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability. *Microsoft Azure Blog* https://azure.microsoft.com/en-us/blog (2022).
15. Guidance for ICT companies setting science based targets. *SBTi* https://sciencebasedtargets.org/sectors/ict (2020).
16. Farfan, J. & Lohrmann, A. Gone with the clouds: estimating the electricity and water footprint of digital data services in Europe. *Energy Convers. Manage.* **290**, 117225 (2023).
17. Mytton, D. Data centre water consumption. *npj Clean Water* **4**, 11 (2021).
18. Siddik, M. A. B., Shehabi, A. & Marston, L. The environmental footprint of data centers in the United States. *Environ. Res. Lett.* **16**, 064017 (2021).
19. Shehabi, A. et al. *United States Data Center Energy Usage Report* (Berkeley Lab, 2016); https://eta.lbl.gov/publications/united-states-data-center-energy
20. Masanet, E., Shehabi, A., Lei, N., Smith, S. & Koomey, J. Recalibrating global data center energy-use estimates. *Science* **367**, 984–986 (2020).
21. Malmodin, J., Lövehagen, N., Bergmark, P. & Lundén, D. ICT sector electricity consumption and greenhouse gas emissions—2020 outcome. *Telecommun. Policy* **48**, 102701 (2024).
22. Andrae, A. S. Comparison of several simplistic high-level approaches for estimating the global energy and electricity use of ICT networks and data centers. *Int. J. Green Technol.* **5**, 50–63 (2019).
23. Belkhir, L. & Elmeligi, A. Assessing ICT global emissions footprint: trends to 2040 & recommendations. *J. Clean. Prod.* **177**, 448–463 (2018).
24. Cao, Z. et al. Data center sustainability: revisits and outlooks. *IEEE Trans. Sustain. Comput.* **9**, 236–248 (2023).
25. Luccioni, A. S., Viguier, S. & Ligozat, A.-L. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *J. Mach. Learn. Res.* **24**, 11990–12004 (2023).
26. Tomlinson, B., Black, R. W., Patterson, D. J. & Torrance, A. W. The carbon emissions of writing and illustrating are lower for AI than for humans. *Sci. Rep.* **14**, 3732 (2024).
27. Shehabi, A. et al. *2024 United States Data Center Energy Usage Report* (Berkeley Lab, 2024); https://eta.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report
28. *Energy and AI* (IEA, 2025); https://www.iea.org/reports/energy-and-ai
29. Investing in the rising data center economy. *McKinsey & Company* https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/investing-in-the-rising-data-center-economy (2023).
30. Lei, N. & Masanet, E. Climate- and technology-specific PUE and WUE estimations for US data centers using a hybrid statistical and thermodynamics-based approach. *Resour. Conserv. Recycl.* **182**, 106323 (2022).
31. Ho, J. et al. *Regional Energy Deployment System (ReEDS) Model Documentation* Version 2020 (National Renewable Energy Lab, 2021); https://docs.nrel.gov/docs/fy21osti/78195.pdf
32. Siddik, M. A. B., Shehabi, A., Rao, P. & Marston, L. T. Spatially and temporally detailed water and carbon footprints of US electricity generation and use. *Water Resour. Res.* **60**, e2024WR038350 (2024).
33. 2023 electricity ATB technologies and data overview. *NREL* https://atb.nrel.gov/electricity/2023/index (2023).
34. *Meta Sustainbility: Connecting to a Better Reality* (Meta, 2024); https://sustainability.atmeta.com/
35. *Innovating Across Our Operations and Supply Chain* (Google Sustainability, 2024); (https://sustainability.google/operating-sustainably/
36. Ritchie, H., Roser, M. & Rosado, P. Renewable energy. *Our World in Data* https://ourworldindata.org/renewable-energy (2020).
37. Companies with the largest operating capacity of clean power in the United States as of end of 2022. *Statista* https://www.statista.com/statistics/1375798/clean-power-operating-capacity-by-company-us (2023).
38. Lyu, C., Fleckenstein, S. & Nerod, Z. *Texas Energy Policy Landscape and Analysis Report* (Center for Local, State, and Urban Policy, 2024); https://closup.umich.edu/sites/closup/files/2024-06/closup-wp-64-Texas-Energy-Policy-Landscape-Analysis.pdf
39. Millstein, D. et al. Solar and wind grid system value in the United States: the effect of transmission congestion, generation profiles, and curtailment. *Joule* **5**, 1749–1775 (2021).
40. *Panel Discussion: Data Centers in Virginia* (Virginia Senate Finance Committee, 2023); https://sfac.virginia.gov/pdf/retreat/2023%20Tysons/13.%20Datacenters%20Panel.pdf
41. *Annual Energy Outlook 2023* (EIA, 2023); https://www.eia.gov/outlooks/aeo/
42. Bjørn, A., Lloyd, S. M., Brander, M. & Matthews, H. D. Renewable energy certificates threaten the integrity of corporate science-based targets. *Nat. Clim. Change* **12**, 539–546 (2022).
43. Pipe, A. & Rattner, N. How DeepSeek's lower-power, less-data model stacks up. *Wall Street Journal* (16 February 2025); https://www.wsj.com/tech/ai/deepseek-ai-how-it-works-725cb464
44. Patel, P. et al. Characterizing power management opportunities for LLMs in the cloud. In *Proc. 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* Vol. 3, 207–222 (Association for Computing Machinery, 2024).
45. Dodge, J. et al. Measuring the carbon intensity of AI in cloud instances. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* 1877–1894 (Association for Computing Machinery, 2022).
46. Hu, Q., Sun, P., Yan, S., Wen, Y. & Zhang, T. Characterization and prediction of deep learning workloads in large-scale GPU datacenters. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '21)*, 1-15 (2021). https://doi.org/10.1145/3458817.3476223
47. Jeon, M. et al. Analysis of large-scale multi-tenant GPU clusters for DNN training workloads. In *2019 USENIX Annual Technical Conference (USENIX ATC 19)*, 947–960 (2019). https://www.usenix.org/conference/atc19/presentation/jeon
48. Weng, Q. et al. Beware of Fragmentation: scheduling GPU-sharing workloads with fragmentation gradient descent. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, 995–1008 (2023). https://www.usenix.org/conference/atc23/presentation/weng
49. Hu, Q. et al. Characterization of large language model development in the datacenter. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, 709–729 (2024). https://www.usenix.org/conference/nsdi24/presentation/hu

50. Lei, N. & Masanet, E. Statistical analysis for predicting location-specific data center PUE and its improvement potential. *Energy* **201**, 117556 (2020).
51. 3M specialty fluids. *3M* https://www.3m.com/3M/en_US/p/c/electronics-components/specialty-fluids/ (2024).
52. Meteonorm 8.2 Global Meteorological Database (version 8.2). *Meteotest AG* https://mn8.meteonorm.com (2023).
53. *Worldwide Quarterly AI Infrastructure Tracker* (IDC, 2024); https://www.idc.com/getdoc.jsp?containerId=IDC_P37251
54. TSMC reportedly sensing increased orders again, CoWoS production capacity surges. *TrendForce* https://www.trendforce.com/news/ (2024).
55. TSMC explores radical new chip packaging approach to feed AI boom. *NIKKEI Asia* https://asia.nikkei.com/business/tech/semiconductors/tsmc-explores-radical-new-chip-packaging-approach-to-feed-ai-boom (2024).
56. Shah, A. Nvidia shipped 3.76 million data-center GPUs in 2023, according to study. *HPC Wire* https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study (2024).
57. Kung, F. *See Generative AI's Impact on the AI Server Market to 2025* (TrendForce, 2024); https://files.futurememorystorage.com/proceedings/2024/20240808_BMKT-301-1_KUNG.pdf
58. Nvidia secures 60% of TSMC's doubled CoWoS capacity for 2025. *Digitimes Asia* https://www.digitimes.com/news/a20241122PD200/nvidia-tsmc-capacity-cowos-2025.html (2024).
59. *DGX systems: built for the unique demands of AI* (Nvidia, 2024); https://www.nvidia.com/en-gb/data-center/dgx-systems/
60. Fan, X., Weber, W.-D. & Barroso, L. A. Power provisioning for a warehouse-sized computer. In *Proc. 34th Annual International Symposium on Computer Architecture* Vol. 35, 13–23 (Association for Computing Machinery, 2007).
61. Masanet, E. R., Brown, R. E., Shehabi, A., Koomey, J. G. & Nordman, B. Estimating the energy use and efficiency potential of US data centers. *Proc. IEEE* **99**, 1440–1453 (2011).
62. Ye, Z., et al. Deep learning workload scheduling in GPU datacenters: a survey. *ACM Comput. Surv.* **56**, 146 (2024).
63. Wu, C.-J. et al. Sustainable AI: environmental implications, challenges and opportunities. *Proc. Mach. Learn. Syst.* **4**, 795–813 (2022).
64. Berthelot, A., Caron, E., Jay, M. & Lefèvre, L. Estimating the environmental impact of generative-AI services using an LCA-based methodology. *Procedia CIRP* **122**, 707–712 (2024).
65. Lopez, A., Roberts, B., Heimiller, D., Blair, N. & Porro, G. US renewable energy technical potentials: a GIS-based analysis (NREL, 2012); https://www.osti.gov/servlets/purl/1219777

Acknowledgements

T.X. and F.Y. gratefully acknowledge support from the National Science Foundation (number 1643244) for this work. F.Y. acknowledges the partial support from the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Sciences programme.

Author contributions

T.X. and F.Y. contributed to project conceptualization, methodology and investigation. F.Y. was responsible for funding acquisition, supervision and project administration. T.X. wrote the original draft. T.X., F.Y., F.F.N., H.D.M. and M.T. participated in the review and editing of the paper and have approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41893-025-01681-y.

Correspondence and requests for materials should be addressed to Fengqi You.

Peer review information *Nature Sustainability* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Professor Fengqi You (fengqi.you@cornell.edu) (Cornell University)

Last updated by author(s): Sep 28, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The following software and codes are used for data collection in this study: 1. The climate data from 2024 to 2030: Meteonorm V8.2.0.24079 2. The grid data from 2024 to 2030: Regional Energy Deployment System (ReEDS) Model 2.0 3. The data center location and AI projection information data: Google Chrome 140.0.7339.208
Data analysis	The following software and codes are used for data analysis in this study: 1. Python 3.9.13: numpy: 1.21.5 pyomo: 6.5.0 csv: 1.0 cyipopt: 1.1.0 h5py: 3.12.1 pandas:2.2.3 scipy:1.13.1 2. Microsoft Excel for Microsoft 365 MSO (Version 22405 Build 16.0.17628.20006) 3. Specific codes and data are available in our GitHub repository: https://github.com/PEESEgroup/US-AI-Server-Analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data and material used in the analysis are available in our GitHub repository: https://github.com/PEESEgroup/US-AI-Server-Analysis

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This information has not been collected.
Reporting on race, ethnicity, or other socially relevant groupings	This information has not been collected.
Population characteristics	No human participants are involved in this study.
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study shows that the deployment of AI servers across the U.S. between 2024 and 2030 could contribute to an annual water footprint ranging from 731 to 1125 million cubic meters and additional annual carbon emissions of 24 to 44 Mt carbon dioxide equivalent between 2024 and 2030, depending on the scale of expansion
Research sample	This study is constructed based on several datasets and verified simulation models. All involved dataset are listed below: 1. The climate data of each grid cell: Meteonorm V8.2.0.24079 2. Grid projection data: ReEDS model, https://github.com/NREL/ReEDS-2.0 3. Water use data: World Resources Institute, https://www.wri.org/research/guidance-calculating-water-use-embedded-purchased-electricity 4. Data center estimation model parameter data: (1) Lei, N. & Masanet, E. Climate-and technology-specific PUE and WUE estimations for US data centers using a hybrid statistical and thermodynamics-based approach. Resources, Conservation and Recycling 182, 106323 (2022). (2) https://www.3m.com/3M/en_US/p/c/electronics-components/specialty-fluids/ 5. Data center location data: https://baxtel.com/
Sampling strategy	No sample size calculation method was used in this study. The sample size verification is listed as below: 1. Nvidia chip manufacture data is collected as the total AI server market data, considering it holds over 90% of the AI server market and is the dominant contributor for high-performance AI computing. 2. U.S. is selected as our research area considering its leading position in AI software and hardware development. Detailed information can be found in method section and Supplemental Information
Data collection	All data collection was performed by Tianqi Xiao under the supervision of Professor Fengqi You. The data are collected from the aforementioned resources.
Timing and spatial scale	Data collection was conducted from September 2023 to Dec 2024. The timing and spatial scale of all datasets are listed as below: 1. The Nvidia manufacture data

(1) Timing scale: 2022 - 2030
(2) Spatial scale: global.
2. The climate data of the U.S.
(1) Timing scale: 2024-2030
(2) Spatial scale: U.S.
3. Grid projection data:
(1) Timing scale: 2024-2030
(2) Spatial scale: U.S.
4. Water use data:
(1). Timing scale: 2024-2030
(2). Spatial scale: U.S.
5. Data center estimation model parameter data:
(1). Timing scale: Current
(2). Spatial scale: U.S.

Data exclusions: No data were excluded

Reproducibility: Our results and major findings can be easily reproduced following our reported data and code in https://github.com/PEESEgroup/US-AI-Server-Analysis.

Randomization: This research utilizes open-source data and verified simulation models to generate projections. Since there are no samples involved, randomization is not applicable or necessary for this study.

Blinding: This research utilizes open-source data and verified simulation models to generate projections. Since there are no human participants or real-world measures involved, blinding is not applicable or necessary for this study.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks: Not applicable

Novel plant genotypes: Not applicable

Authentication: Not applicable