Energy Economics 150 (2025) 108880

Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneeco

AI carbon footprint in China sets to double post-2030 carbon peaking☆

Zhan-Ming Chen [a], Qiyang Xiong [a], Jiahui Duan [b], Jianhong Ma [a], Zhuo Chen [c,d], Shan Guo [e,f,*]

[a] School of Applied Economics, Renmin University of China, Beijing, China
[b] Institute of Energy, Environment and Economy, Tsinghua University, Beijing, China
[c] Institutes of Science and Development, Chinese Academy of Sciences, Beijing, China
[d] School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing, China
[e] School of Public Administration and Policy, Renmin University of China, Beijing, China
[f] Big Data and Responsible Artificial Intelligence for National Governance, Renmin University of China, Beijing, China

ARTICLE INFO

Keywords:
AI data center
Carbon footprint
Architectural carbon modeling
Scenario analysis

ABSTRACT

The unprecedented advancements in artificial intelligence (AI) have significantly increased energy consumption, raising global concerns about AI carbon footprint. As a global AI leader, China stands at a crucial juncture where its AI development intersects with national energy transition and climate strategies, particularly concerning its Carbon Peaking and Carbon Neutrality Goals. This study employs an uncertainty-based Architectural Carbon Modeling Tool and scenario analysis to quantify the energy consumption and carbon footprint of AI data centers in China from 2022 to 2050. Our projections indicate that the electricity consumption of AI data centers will surpass 1000 TWh by 2030, exerting pressure on the power system and driving a significant increase in emissions. The carbon footprint is projected to double after 2030, peaking at 695 Mt. in 2038 and declining to 474 Mt. by 2050, with manufacturing emissions contributing approximately 18 % of the total. Geospatial analysis reveals that under the Business as Usual scenario, energy demand remains concentrated in the eastern provinces, whereas the Advanced Green scenario redistributes AI computing demand to the west, creating new carbon hotspots. Increasing the proportion of internal green electricity is identified as the most effective strategy, with the potential to reduce operational emissions by 42 %. This study comprehensively examines AI's energy implications and provides policy-relevant insights for balancing technological advancements with long-term energy sustainability.

1. Introduction

Artificial intelligence (AI) has rapidly advanced in recent years, attracting substantial investment and driving transformative applications across industries. In this study, AI is defined as computational systems capable of emulating human cognitive functions, such as perception, reasoning, learning, and decision-making, by processing large volumes of data and adapting over time (Chen et al., 2021; Russell and Norvig, 2016). Unlike traditional ICT systems, which operate based on explicit rules and programming, AI systems learn from data, allowing them to adapt dynamically and handle complex, unstructured tasks that rule-based systems struggle with (Wang et al., 2023). AI has also contributed to climate change mitigation and adaptation by enhancing technological effectiveness across multiple sectors (Ding et al., 2024; Tomašev et al., 2020; Vinuesa et al., 2020). Studies show that AI adoption improves carbon factor productivity (Wang et al., 2022), optimizes industrial structure (Liu et al., 2022a), and enhances resource utilization efficiency (Liu et al., 2020). However, as AI becomes more widespread, concerns have emerged about its negative climatic consequences, including increased energy demand and carbon emissions from large-scale training and inference workloads (Schwartz et al., 2020; de Vries, 2023).

The growing adoption of AI, coupled with its substantial energy demand for hardware manufacturing and software operation, poses a potential threat to global Sustainable Development Goals (SDGs), especially SDG 13 for climate action (Jones, 2018; Masanet et al., 2020; Mora et al., 2018). For example, the rapid expansion of AI applications is driving explosive growth in electricity consumption by data centers,

☆ This article is part of a Special issue entitled: 'Energy & AI' published in Energy Economics.

* Corresponding author.
E-mail addresses: chenzhanming@ruc.edu.cn (Z.-M. Chen), qyxiong@ruc.edu.cn (Q. Xiong), duan-jh25@mails.tsinghua.edu.cn (J. Duan), majianhong@ruc.edu.cn (J. Ma), chenzhuo221@mails.ucas.ac.cn (Z. Chen), shan.guo@connect.polyu.hk (S. Guo).

https://doi.org/10.1016/j.eneco.2025.108880
Received 3 December 2024; Received in revised form 22 August 2025; Accepted 29 August 2025
Available online 3 September 2025

which consumed about 460 TWh globally in 2022, accounting for approximately 2 % of global electricity demand (IEA, 2024). This demand is projected to surge to between 620 and 1050 TWh by 2026. Such massive electricity consumption has significantly contributed to the growing carbon footprint (Dhar, 2020; de Vries, 2023).

China, a pivotal player in the global AI race, finds itself at a crucial juncture where its strides in AI development intersect deeply with the country's energy transition and climate strategies, particularly in light of its Carbon Peaking and Carbon Neutrality Goals. China hosts a significant fraction of the world's hyperscale data centers (Xie et al., 2024). From 2016 to 2023, the number of data center racks in China grew at a compound annual growth rate of over 30 % (China Academy of Information and Communications Technology (CAICT), 2022), surpassing 8.3 million standard racks and accounting for 26 % of global capacity by 2023 (China Academy of Information and Communications Technology (CAICT), 2024a, 2024b). The electricity consumption for data center operations in 2023 totaled 150 TWh, accounting for 1.6 % of the country's electricity consumption (China Academy of Information and Communications Technology (CAICT), 2024c). China's coal-dominated energy structure makes AI energy demand a substantial source of carbon emissions (Zhuo et al., 2022). In 2021, the operational emissions from the electricity demand of all data centers in China reached 135 Mt., accounting for about 1.3 % of the country's total emissions (Ni et al., 2024).

In the context of China's Carbon Peaking and Carbon Neutrality Goals, the AI industry is accelerating its low-carbon transition through both internal and external measures. Internal measures aim to reduce the carbon emissions of AI data centers by improving energy efficiency, such as reducing power usage efficiency (PUE) and increasing computing efficiency. External measures focus on reducing carbon emissions by substituting fossil fuels with non-fossil energy sources (Ma et al., 2024). China has mandated that state-owned data centers incorporate at least 5 % renewable energy by 2023, with a full transition to renewable energy by 2030 (Ministry of Finance et al., 2023). One strategy to achieve this goal involves relocating data centers to regions with abundant renewable energy resources. Accordingly, while the energy consumption of AI data centers and related emissions reduction measures have gained increasing attention, there is still a lack of comprehensive studies that quantify the overall carbon footprint of AI data centers, which underscores the need for more thorough assessments of AI's environmental impacts and the strategies required to mitigate them.

In this paper, we delve into the carbon footprint trajectory of AI data centers in China from 2022 to 2050 by utilizing the uncertainty-based Architectural Carbon Modeling Tool and scenario analysis to quantify manufacturing and operational emissions. The BAU scenario reveals that the carbon footprint will rise sharply from 82 Mt. in 2022 to a peak of 695 Mt. in 2038, with manufacturing emissions contributing about 18 % of the total. The eastern coastal provinces are the primary drivers of the initial increase and eventual reduction in carbon footprint. In contrast, the PRO scenario, which assumes moderate energy efficiency improvement and energy mix adjustment, results in a lower and earlier peak. In the MAX scenario, approximately 7219 Mt. of carbon emissions will be reduced from 2022 to 2050 compared to the BAU scenario, owing to the aggressive technological advancements, the deep power system decarbonization, and the rapid East–West Computing Resources Transmission Project (EWCRT) promotion. A comparison of techniques shows that power system decarbonization is the most effective measure, potentially reducing 42 % of overall emissions. Given AI's high energy demands, which pose a challenge to global climate efforts, our study highlights the urgent need to assess the long-term environmental impacts of AI. We also propose actionable strategies to mitigate AI emissions, supporting responsible large-scale deployments that align technological progress with the SDGs.

The research has three contributions: First, we provide a comprehensive assessment of the carbon footprint of AI data centers, incorporating both operational and embodied emissions, filling the gap in research that often focuses solely on operational electricity consumption. Second, through scenario analysis, we address the lack of long-term carbon footprint projections for AI. The results reveal the nonlinear relationship between the growth of digital infrastructure and its environmental impact, and offer critical insights into the future trajectory of AI's emissions through 2050. Third, we evaluate the impact of various parameter improvements on carbon emissions, providing actionable strategies for designing low-carbon AI infrastructure.

The research is structured as follows: Section 2 reviews existing literature on AI's environmental impacts. Section 3 presents the methodologies and data sources. Section 4 discusses the results. Section 5 concludes with key findings and policy recommendations.

2. Literature review

2.1. Paradox of AI's role in energy consumption

AI now plays a pivotal role in economies, not only integrating into the energy system but also reshaping a wide range of industries. While some view AI as a driver of energy efficiency and economic growth, others warn against its escalating energy demands and potential environmental repercussions (Qin et al., 2024; Zhang and Zeng, 2024).

Those who view AI as a "blessing" often emphasize its impact on enhancing energy efficiency through smart grids, predictive maintenance, and instantaneous demand response (Bai et al., 2025; Liu et al., 2022b). These advances support the development of energy systems and contribute to the green economy (Liu et al., 2021). At the city level, several studies have shown that AI can enhance the resilience of energy systems (Jiang and Yu, 2025) and improve green total factor productivity (Zhao et al., 2022), which makes progress in green economic growth systematically. At the industrial level, AI-driven technologies are primarily used to enhance energy efficiency during the production process by optimizing resource utilization and facilitating informed decision-making. Applications such as sales forecasting (Sohrabpour et al., 2021), optimization solutions (Ukoba et al., 2024), virtual financial services (Aysan et al., 2024), and industrial robots (Zhang et al., 2024) have become increasingly widespread and mature. Although the term "AI" appears across many studies, it often refers to different things, such as technological innovations and AI investments (Song et al., 2024), industrial robots (Zhao et al., 2024), corporate report references (Liu et al., 2025; Zhong et al., 2025), and patent data (Nepal et al., 2025). Despite these differences, a common thread is that AI is consistently linked to improvements in energy efficiency and broader sustainability goals across sectors.

However, the energy consumption associated with AI cannot be ignored. Training and operating AI models require substantial computing power, which increases energy demands of computational infrastructures such as AI data centers (Dauvergne, 2022; Schwartz et al., 2020). The computing power needed for deep learning research has been doubling at an alarming rate every few months(Schwartz et al., 2020). This exponential growth in computational requirements not only raises concerns about the environmental impact but also imposes considerable financial burdens on energy providers (Ahmad et al., 2021). Nonetheless, a range of mitigation strategies is being actively explored. Innovations in cooling systems, power supply enhancements, and server virtualization have led to significant reductions in PUE, thereby improving the overall energy efficiency of AI data centers (Lei and Masanet, 2020). Another key solution involves enhancing computing efficiency, where researchers are developing more computationally efficient algorithms that optimize performance while minimizing energy consumption during training and inference phases

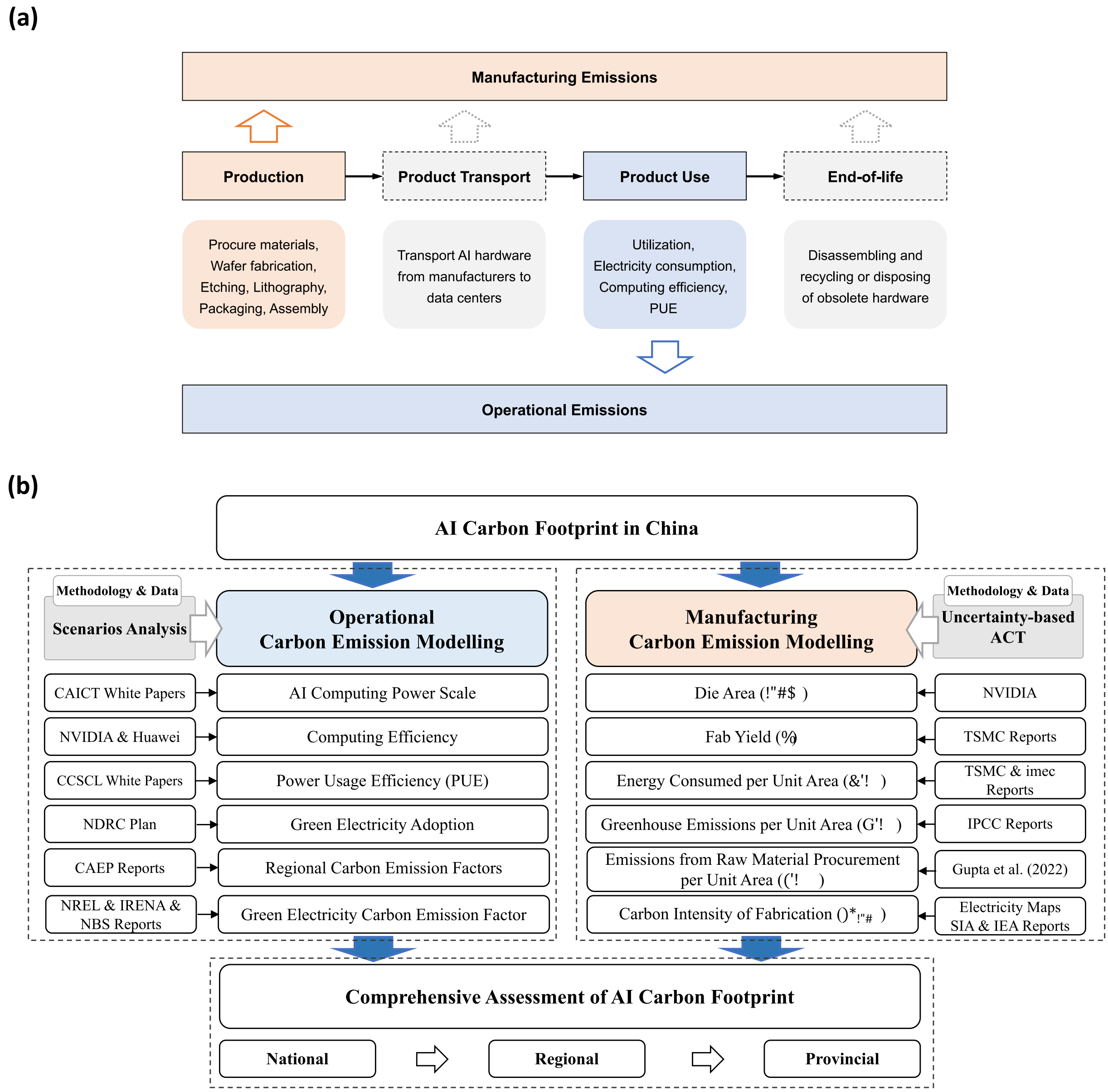

Fig. 1. Modeling Framework. (a) Lifecycle breakdown of AI data center carbon emissions. (b) AI carbon footprint modeling framework. The life cycle stages include Production, Transport, Use, and End-of-life processing. Operational emissions are based on use. Manufacturing emissions are from production phase, while transport and disposal are excluded from the focus of the study due to their minor impact.

(Hoefler et al., 2021). For instance, DeepSeek-V3 demonstrates superior performance while requiring significantly fewer training hours.[1] Additionally, transitioning to renewable energy sources for powering AI infrastructures can substantially lower the carbon footprint, as non-fossil energy sources like wind and solar provide cleaner alternatives (Lau et al., 2023; Ma et al., 2024).

Understanding the full impact of AI on environmental and energy demands requires a broader, long-term perspective rather than isolated estimates. Some studies have proposed frameworks for tracking energy consumption (Henderson et al., 2020) and global GHG emissions (Kaack et al., 2022) linked to specific AI tasks or technologies like machine learning. They often focus on limited aspects and differ in what they include under the label of "AI." A comprehensive energy framework that defines what and how AI energy consumption should be measured is still lacking.

2.2. Research methodology

As AI continues to develop across algorithms, software, and hardware, so does its environmental and economic impact. Some empirical studies have analyzed real-world data to evaluate the impact brought by AI (Guo et al., 2025; Li et al., 2023b). However, many of these studies have explored AI's impact using various methods from the city or industrial level, rather than providing a systematic analysis of large-scale AI systems.

The life cycle perspective has emerged as a valuable tool for addressing these gaps, particularly in environmental and sustainability research (Ayres, 1995). For example, the carbon footprint of digital technologies and services has been assessed from a full life cycle perspective (Cheng et al., 2023; Guo et al., 2020; Li et al., 2023c; Xie et al., 2024; Zhou et al., 2019). The results suggest that the digital economy is becoming a significant contributor to the carbon footprint,

[1] Liu A, Feng B, Xue B, et al. DeepSeek-V3 technical report. URL: https://arxiv.org/abs/2412.19437 (accessed 5.19.25).

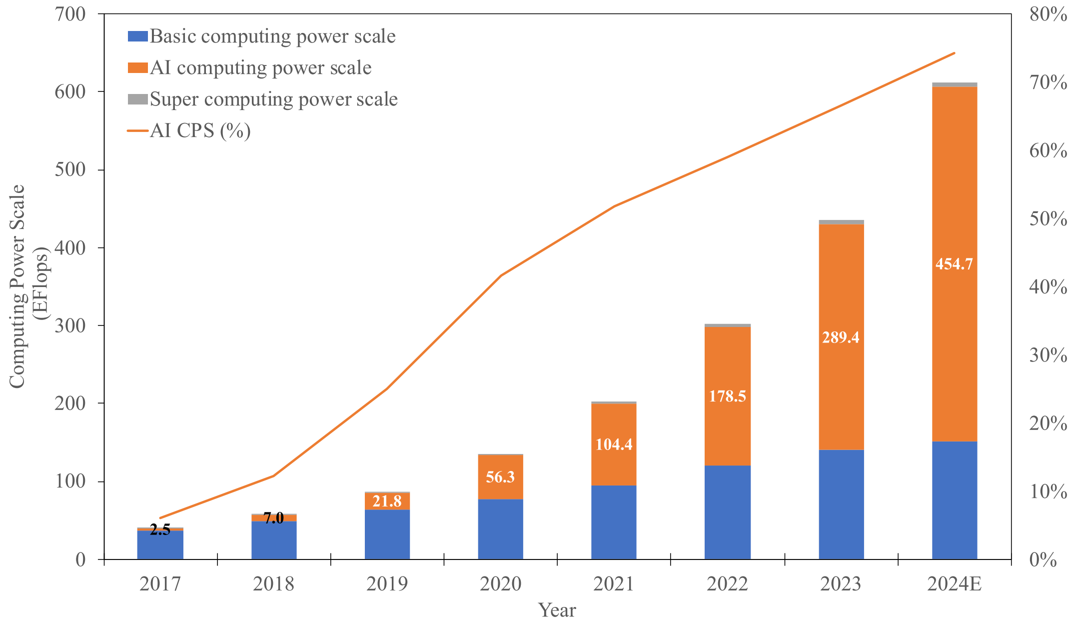

Fig. 2. Annual computing power scale in China, 2017–2024. Data is sourced from the White Paper on China Computing Power Index (China Academy of Information and Communications Technology (CAICT), 2025). Basic computing power reflects the total capacity of general-purpose servers based on six years of shipment data. AI computing power is estimated from AI server shipments over the same period. Supercomputing power is drawn from the global top 500 and China's top 100 high-performance systems, supplemented by manufacturer data. The values for 2024 are projections based on shipment trends.

with both operational and manufacturing energy demand playing crucial roles. Investigations have also been conducted to analyze the carbon emissions resulting from AI's operational electricity consumption (Patterson et al., 2022; de Vries, 2023). However, there is still no comprehensive analysis that estimates and forecasts the total carbon footprint, encompassing both manufacturing and operational emissions of AI.

To sum up, while many studies highlight the environmental benefits of AI, such as improving energy efficiency and enabling green technological innovation, far less attention has been given to its environmental costs, particularly those associated with large-scale infrastructure. Meanwhile, existing research on AI emissions primarily focuses on electricity use during operation (Shehabi et al., 2024), often neglecting the embodied emissions associated with hardware production. Although life cycle assessments have often been used to estimate emissions in the digital economy, few studies have systematically quantified the full life cycle emissions of AI data centers over time and across regions. To fill this gap, we employ an uncertainty-based ACT framework and scenario analysis to estimate manufacturing and operational carbon emissions of AI data centers in China from 2022 to 2050. The results provide critical insights for reducing AI's carbon footprint and guiding responsible large-scale deployment in line with the Sustainable Development Goals.

3. Methods and data sources

To analyze the carbon footprint of AI data centers in China, it is essential to evaluate both operational and manufacturing emissions. This requires a life-cycle analysis that includes production, transportation, usage and end-of-life processing, as illustrated in Fig. 1. Operational emissions refer to the daily operations of data centers, including energy consumed by servers, hardware, and the cooling systems required for optimal functioning. Manufacturing emissions arise from the production, transportation, and disposal of materials and equipment used in AI systems. While transportation and disposal are minor contributors and remain relatively constant across system generations (Gupta et al., 2022), this study focuses on the production phase. We denote the carbon footprint, operational emissions, and manufacturing emissions as C_{total}, C_o and C_m, respectively. The carbon footprint can be calculated using the following equation:

$$C_{total} = C_o + C_m \tag{1}$$

3.1. Operational carbon emission modeling

Operational emissions mainly originate from the electricity consumption of AI data centers, which is primarily driven by three factors: (1) AI computing power scale, (2) the computing efficiency of AI chips, and (3) the PUE of data centers. This study accounts for the heterogeneity in the share of green electricity utilized within data centers and regional carbon emission factors when converting electricity consumption into carbon emissions. Inspired by the bottom-up method of prior studies (Xie et al., 2024), we adopt a top-down allocation framework to estimate operational carbon emissions at both national and provincial levels. The national total is first estimated based on aggregate AI computing power demand and chip efficiency, and then disaggregated across provinces using regional parameters. This allows us to incorporate regional heterogeneity in PUE, green electricity adoption, and grid carbon intensity. The operational electricity consumption and associated carbon emissions for each province are estimated using the following formulas:

$$E_i = (CPS_i/CE) \times U_i \times T \tag{2}$$

$$C_o = \sum_{i=1}^{n} [E_i \times G_i \times F_G + E_i \times (1 - G_i) \times F_i] \tag{3}$$

where E_i denotes the total electricity consumption of AI data centers in province i; C_o represents the operational emissions; CPS_i is AI computing power scale in province i; CE refers to the computing efficiency of AI chips, assumed uniform across provinces; U_i is the average PUE of province i; T denotes the annual operating hours, assumed to be 8760 h per year; G_i is the share of green electricity within data centers; F_G and F_i represent the emission factors for green electricity and the regional power grid, respectively.

3.1.1. Artificial intelligence computing power scale

In recent years, China's data centers have experienced significant

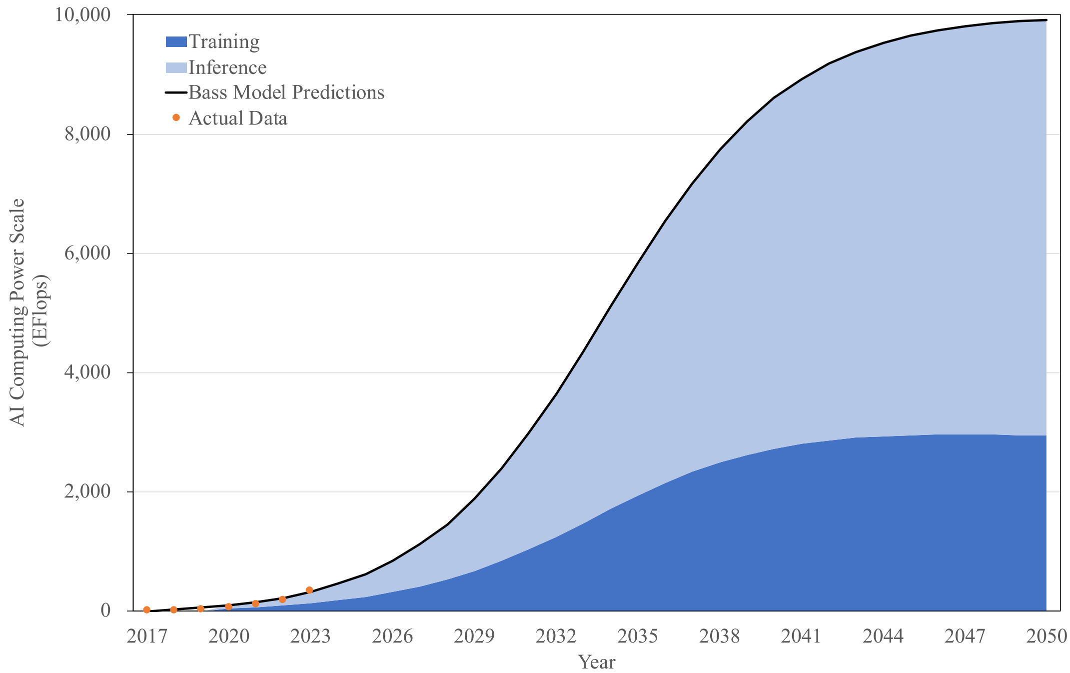

Fig. 3. Forecast of AI computing power scale in China, 2017–2050. The AI computing power scale, calibrated using CAICT data from 2017 to 2022, is projected to forecast trends through 2050 using the Bass Diffusion Model. Inference workloads, based on IDC data, accounted for 55.5 % of AI server usage in 2022 and are expected to reach 62.2 % by 2026 and 70.3 % by 2050, using regression analysis.

expansion, with a clear trend toward clustering. As documented in the White Paper on China Computing Power Index (China Academy of Information and Communications Technology (CAICT), 2023), national computing power has grown rapidly across multiple segments. Fig. 2 highlights the annual increase in computing power scale in China from 2017 to 2024, with AI computing power exhibiting an impressive average annual growth rate of over 100 %.

In the context of China's rapidly evolving technology landscape, the Bass Diffusion Model offers a theoretically grounded and widely validated method for forecasting the expansion of AI computing power. Compared to linear or exponential approaches, it more accurately reflects the S-shaped adoption curve by modeling the interplay between early adopters (innovation effect) and later adopters (imitation effect). This makes it particularly applicable to emerging technologies like AI. The model has been successfully applied to predict the diffusion of emerging technologies, such as broadband Internet(Bacha et al., 2024; Turk and Trkman, 2012), electric vehicles(Bitencourt et al., 2021; Massiani and Gohs, 2015), and ICT adoption (Ntwoku et al., 2017). Given the relatively short history of AI computing power metrics, the Bass model enables a forward-looking estimation based on limited early-stage data and an assumed saturation potential.

The Bass Diffusion Model is given by

$$n(t) = \frac{dN(t)}{dt} = p[M - N(t)] + q\frac{N(t)}{M}[M - N(t)] \tag{4}$$

where $n(t)$ represents the probability density function of AI computing power adoption at time t, $N(t)$ is the cumulative distribution function of adoption at time interval t, and M is the total potential market capacity at the end of diffusion. Moreover, the coefficient of external influence p reflects the impact of external factors (e.g., government initiatives, global technological advancements, and industry-specific policies that encourage the adoption of AI technologies), and the coefficient of internal influence q reflects the impact of internal factors (e.g., competitive pressures, peer influence within the tech community, and the intrinsic benefits perceived by early adopters).

The analytical solution maps out the adoption curve:

$$N(t) = M\frac{1 - \exp(-(p+q)t)}{1 - (q/p) \bullet \exp(-(p+q)t)} \tag{5}$$

which reveals that the AI computing power in China is expected to follow an S-shaped trajectory, characteristic of new technology assimilation. The Bass Diffusion Model predicts initial rapid growth due to favorable policies and significant investments in AI infrastructure, reaching a peak as the market approaches saturation.

The model is calibrated using historical data on AI computing power in China from 2017 to 2022. To fit the annual data, the continuous model is adapted into a discrete form:

$$n(t_i) = (p + (q/M) \bullet N(t_{i-1}))(M - N(t_{i-1})) \tag{6}$$

where $n(t_i)$ is the increment of AI computing power scale in year t_i, $N(t_{i-1})$ is the cumulative AI computing power adoption by year t_{i-1}.

For the estimation of coefficients M, p and q, we utilized a non-linear least squares estimation method. This method allows for a refined approximation of the parameters that govern the diffusion process described by the Bass model. The mathematical representation of this estimation can be expressed in the following equations:

$$n(t_i) = pM + (q-p)N(t_{i-1}) - (q/M) \bullet N^2(t_{i-1}) \tag{7}$$

$$X(i) = M\left(\frac{1 - \exp(-(p+q)t_i)}{1 + (q/p)^{-(p+q)t_i}} - \frac{1 - \exp(-(p+q)t_{i-1})}{1 + (q/p)^{-(p+q)t_{i-1}}}\right) + \mu_i \tag{8}$$

where μ_i is an additive error term, and $X(i)$ is the increment function derived from the analytical solution of the Bass model. By employing iterative algorithms such as Newton's Method, the parameters M, p and q are meticulously tuned to fit the observed data. The results of the coefficients are as follows: the Market Potential (M) is estimated at 9916.6 EFlops, the Coefficient of External Influence (p) is 0.002, and the

Table 1
Nvidia Server GPUs Theoretical Data.

GPU	TFlops	Watts	Year	GFlops/W
Tesla K10	5	225	2012	20
Tesla K20x	4	235	2012	17
Tesla K40	5	235	2013	21
Tesla K80	8	300	2014	27
Tesla M40	7	250	2015	27
Tesla M60	10	300	2015	32
Tesla P100	11	300	2016	35
Tesla V100	16	300	2018	52
A100	20	400	2020	49
A30	10	165	2021	62
H100	67	700	2022	96
H800	51	700	2023	73
H200	67	700	2024	96

Note: Data for NVIDIA server GPUs released before 2021 are sourced from the appendix (Desislavov et al., 2023). Data for recent GPUs (H100, H800, H200) are retrieved from the NVIDIA Data Center GPU Resource Center.
NVIDIA Data Center GPU Resource Center. Retrieved from https://resources.nvidia.com/l/en-us-gpu?ncid=no-ncid

Coefficient of Internal Influence (q) is 0.294.

The result of the simulation is shown in Fig. 3.

The trajectory of AI computing power from 2017 through 2050 exhibits an S-shaped curve typical of technological adoption. The initial growth rate is modest, followed by a sharp acceleration as adoption increases, which eventually reaches a peak and then gradually declines toward market saturation. By 2050, AI computing power in China is projected to reach approximately 9917 exaflop operations per second (EFlops), marking a significant advancement and representing more than a 56-fold increase from the levels recorded in 2022. This substantial growth highlights a prolonged phase of intensive development, expected to continue over the coming decades until the technology matures and stabilizes near market saturation.

Data from IDC reveals that AI server inference accounted for 57.6 % of AI workloads in China in 2022, with this share projected to rise to 62.2 % by 2026. Time-series analysis predicts that by 2050, inference workloads will comprise approximately 70.3 % of the total AI computing power. This shift reflects a strategic transition in data center operations, moving from a focus on training tasks to inference activities.

3.1.2. *Computing efficiency of AI chips*

Computing efficiency is a key metric used to evaluate how effectively an AI chip utilizes electrical power for computation. It is defined as the ratio of a chip's computing power to its electrical power consumption, expressed in terms of floating-point operations per second per watt (FLOPS/W). This metric is crucial for assessing the energy efficiency of AI chips, as it directly impacts power consumption, operational costs, and environmental sustainability. A higher computing efficiency indicates better utilization of electrical power, contributing to lower energy costs.

The computing efficiency of AI chips is calculated as follows:

$$CE = CP/P_{IT} \tag{9}$$

where CE is the computing efficiency, measured in FLOPS/W; CP denotes the computing power, measured in FLOPS; P_{IT} signifies the electrical power consumption of IT equipment, measured in W.

Nvidia holds a dominant position in the GPU market, controlling approximately 80 % of the global GPU semiconductor market in 2023.[2] As AI models become more complex and data volumes continue to expand, the role of GPUs in providing the required computational power while optimizing energy efficiency has become increasingly critical. Table 1 summarizes the theoretical efficiencies of various Nvidia server GPUs since 2012.

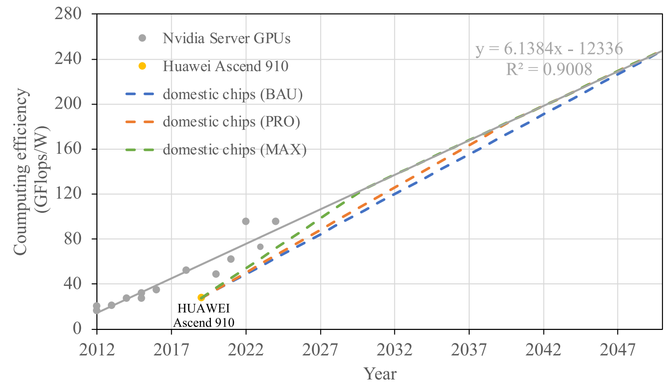

Fig. 4. Computing Efficiency of AI Server GPUs. The gray dots represent global advanced AI chips, with Nvidia Server GPUs as the benchmark. The yellow dot denotes Huawei Ascend 910, a key domestic AI chip. The gray solid line represents the fitted trend for global advanced AI chips, illustrating the trajectory of cutting-edge computing efficiency, with the equation $y = 6.1384x - 12336$ and an R^2 value of 0.9008. The three dashed lines depict the projected computing efficiency trajectories of domestic AI chips under three scenarios, reflecting varying levels of technological advancement. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Huawei's Ascend 910 chip, launched in 2019, has become a key player in China's AI data centers due to its adequate performance and high cost-effectiveness. Huawei has made large-scale deployments in the western hubs of the EWCRT project, such as Guizhou and Inner Mongolia, to support the AI computing power needs of domestic universities and enterprise clients. Meanwhile, U.S. regulators implemented rules in 2023 that barred Nvidia from selling its advanced chips to Chinese customers, spurring a strong push for domestic AI chip development among Chinese manufacturers. As a result, it is crucial to consider the technology gap and localization rate of domestic chips in the scenario analysis.

While there have been significant improvements in GPU efficiency over the years, achieving higher computing efficiency is becoming increasingly challenging. To begin with, we employ the regression model to predict the ideal computing efficiency of global advanced AI chips, which provides a benchmark for the development of China's domestic AI chips. In the scenario analysis, we hypothesize different levels of localization rates for AI chips in China, and the timeline for domestic chips to catch up with global leading standards (Fig. 4). We define three scenarios based on different localization rates and technological advancement:

Business as Usual (BAU): Computing efficiency is expected to reach a 50 % localization rate by 2050, with domestic chips achieving global leading standards by the same year.

Green Development (PRO): Computing efficiency is expected to reach a 70 % localization rate by 2050, with domestic chips achieving global leading standards by 2040.

Advanced Green (MAX): Computing efficiency is expected to reach a 90 % localization rate by 2050, with domestic chips achieving global leading standards by 2030.

The localization rate and the computing efficiency of both domestic and global leading chips determine the computing efficiency of AI chips in any given year. The formula used to calculate the computing efficiency is:

$$CE_t = L_t \times CE_{D,t} + (1 - L_t) \times CE_{G,t}$$

[2] Nguyen, J., 2024. What you need to know about Nvidia and the AI chip arms race. Marketplace. URL: https://www.marketplace.org/2024/03/08/what-you-need-to-know-about-nvidia-and-the-ai-chip-arms-race/ (accessed 6.22.24)

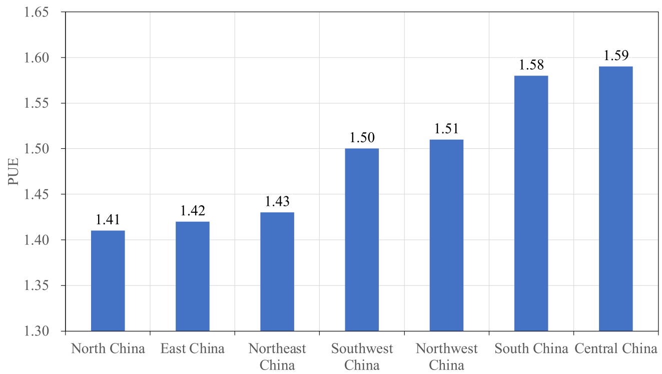

Fig. 5. Average PUE for seven regions in China in 2021. Data is sourced from the White Paper on China Data Center Industry Development (China Communications Services Corporation Limited, 2023).

where CE_t is the computing efficiency in year t; L stands for the localization rate of domestic chips; CE_D refers to the computing efficiency of domestic chips; CE_G denotes the computing efficiency of imported global advanced chips.

3.1.3. Power usage efficiency (PUE)

PUE is a key metric for assessing the energy efficiency of a data center, which is calculated as the ratio of the total energy consumed by the data center to the energy consumed solely by its IT equipment. A PUE value greater than 1 indicates that a portion of the energy is used for non-IT functions such as cooling and power distribution, while a value closer to 1 reflects higher efficiency, as more energy is dedicated directly to computing tasks. The formula for PUE is as follows:

$$U = E_{total}/E_{IT} \tag{10}$$

where E_{total} is the total energy consumption of the data center; E_{IT} is the energy consumption of its IT equipment.

Fig. 5 illustrates the average PUE values for seven regions in China in 2021, based on data from the White Paper on China Data Center Industry Development (China Communications Services Corporation Limited, 2023).

The average PUE of data centers in China was 1.49 in 2021, reflecting a steady improvement from 1.6 in 2019 (China Communications Services Corporation Limited, 2023). This progress is largely attributed to advancements in thermal management systems and reductions in power distribution losses. North and East China demonstrate relatively high efficiency, with an average PUE close to 1.40. In contrast, Central and South China exhibit relatively higher PUE, largely due to geographical conditions and lower rack utilization rates.

3.2. Scenarios analysis

This section outlines the methodology used to estimate operational emissions from AI data centers in China, based on a 3 × 3 scenario matrix covering the period from 2022 to 2050 (Fig. 6). The scenario design adopts a dual-dimensional approach, focusing on the two most influential factors in determining operational carbon emissions of AI: distribution of AI computing power scale and sustainable technological progress. The distribution of AI computing power scale determines the regional energy mix and carbon intensity of AI data centers. Given the significant variations in energy sources across China, the geographical allocation of AI workloads plays a crucial role in shaping emissions outcomes. Sustainable technological progress captures the efficiency gains in AI computing hardware, PUE, and the transition to green electricity, all of which directly impact data center energy consumption.

The horizontal axis classifies scenarios according to the scale of computing power deployment: Regional Equilibrium, Centralized Growth, and Strategic Distribution. The vertical axis incorporates technological parameters, including AI chip computing efficiency, PUE, the share of green electricity, and regional power grid carbon emission factors. It is divided into three scenarios, each reflecting a distinct level of sustainable technological advancement: Moderate Progress, Accelerated Transition, and Net-zero Commitment.

3.2.1. Distribution of AI computing power scale

In the Regional Equilibrium scenario, the distribution of AI computing power across provinces remains unchanged from the 2022 baseline, as reported in the White Paper on China Computing Power Index (China Academy of Information and Communications Technology (CAICT), 2023). This scenario reflects a market-driven trajectory in which AI infrastructure expansion follows historical trends without major policy interventions (Fig. 7). This scenario assumes that AI computing demand grows proportionally across all regions, maintaining the existing regional imbalance in computing power deployment. As a result, eastern provinces continue to dominate AI computation.

The Centralized Growth scenario represents a market-driven and policy-supported evolution of AI computing power, where AI demand leads to the formation of highly efficient, large-scale data center clusters in designated hubs. This scenario prioritizes the development of eight national-level data center hubs, as identified in the EWCRT project (Ministry of Science and Technology, 2022). These hubs are expected to achieve economies of scale, improving operational efficiency and reducing the overall energy footprint of AI computing. By 2050, these eight hubs are expected to account for over 90 % of China's total AI computing power, enabling optimized AI workload distribution that matches the growing demand for AI services.

In contrast, the Strategic Distribution scenario envisions a major shift in AI computing infrastructure toward the western hubs, driven by strategic policy interventions that incentivize data center relocation to areas with higher renewable energy availability. By 2050, western hubs are projected to constitute more than 50 % of the national AI computing power, highlighting a strategic focus on leveraging the western region's abundant resources and favorable conditions.[3]

3.2.2. Sustainable technological Progress

The sustainable technological progress scenarios consider four key factors that directly influence the operational emissions of AI data centers: computing efficiency, PUE, green electricity adoption, and regional carbon emission factors. Each of these factors is shaped by both technological advancements and evolving policy frameworks.

Computing Efficiency. The scenario analysis for AI chip computing efficiency is driven by both domestic technological advancements and policy-driven incentives for semiconductor localization. As of 2022, the localization rate for AI chips in China stood at 15 %,[4] with domestic alternatives lagging behind top international chips by approximately 1.5× in computing efficiency, reflecting a 2–3 year technological gap. However, rapid technological advancements in domestic AI chip design are narrowing the performance gap with leading global competitors. Under the three scenarios analyzed, we predict that domestic chips will achieve the global leading standards by 2050, 2040, and 2030, respectively. Moreover, the escalating trade frictions have underscored the urgent need for domestic alternatives, with significant growth in the adoption of domestic AI chips anticipated in the future. Provincial computing power construction plans indicate localization requirements

[3] National Development and Reform Commission, "Opinions on Deepening the Implementation of the 'East-to-West Data Transfer' Project and Accelerating the Construction of the National Integrated Computing Power Network."

[4] International Data Corporation (IDC)

! ! !		**Distribution of AI Computing Power Scale**		
		Regional Equilibrium	**Centralized Growth**	**Strategic Distribution**
Sustainable Technological Progress	**Moderate Progress**	**Business as Usual (BAU):** AI computing power across each province remains at 2022 levels. Computing efficiency will reach a 50% localization rate by 2050, with domestic chips achieving global leading standards. PUE < 1.25 in eastern hubs, PUE < 1.2 in China's western hubs, and PUE < 1.3 in other data centers by 2035. The proportion of green electricity in eastern hubs will increase to 30% by 2030, achievable five years earlier in western hubs. Upper Bound Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)	**S2:** AI computing power in eight hubs will represent over 90% by 2050. Computing efficiency will reach a 50% localization rate by 2050, with domestic chips achieving global leading standards. PUE < 1.25 in eastern hubs, PUE < 1.2 in China's western hubs, and PUE < 1.3 in other data centers by 2035. The proportion of green electricity in eastern hubs will increase to 30% by 2030, achievable five years earlier in western hubs. Upper Bound Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)	**S3:** AI computing power in Western hubs will exceed 50% by 2050. Computing efficiency will reach a 50% localization rate by 2050, with domestic chips achieving global leading standards. PUE < 1.25 in eastern hubs, PUE < 1.2 in China's western hubs, and PUE < 1.3 in other data centers by 2035. The proportion of green electricity in eastern hubs will increase to 30% by 2030, achievable five years earlier in western hubs. Upper Bound Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)
	Accelerated Transition	**S4:** AI computing power across each province remains at 2022 levels. Computing efficiency will reach a 70% localization rate by 2050, with domestic chips achieving global leading standards by 2040. PUE < 1.25 in eastern hubs, PUE < 1.2 in China's western hubs, and PUE < 1.3 in other data centers by 2035. The proportion of green electricity in eastern hubs will increase to 50% by 2040, achievable five years earlier in western hubs. Intermediate Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)	**Green Development (PRO):** AI computing power in eight hubs will represent over 90% by 2050. Computing efficiency will reach a 70% localization rate by 2050, with domestic chips achieving global leading standards by 2040. PUE < 1.25 in eastern hubs, PUE < 1.2 in China's western hubs, and PUE < 1.3 in other data centers by 2035. The proportion of green electricity in eastern hubs will increase to 50% by 2040, achievable five years earlier in western hubs. Intermediate Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)	**S6:** AI computing power in Western hubs will exceed 50% by 2050. Computing efficiency will reach a 70% localization rate by 2050, with domestic chips achieving global leading standards by 2040. PUE < 1.25 in eastern hubs, PUE < 1.2 in China's western hubs, and PUE < 1.3 in other data centers by 2035. The proportion of green electricity in eastern hubs will increase to 50% by 2040, achievable five years earlier in western hubs. Intermediate Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)
	Net-zero Commitment	**S7:** AI computing power across each province remains at 2022 levels. Computing efficiency will reach a 90% localization rate by 2050, with domestic chips achieving global leading standards by 2030. PUE for all data center hubs will fall to 1.1 after 2030. The proportion of green electricity in eastern and western hubs will gradually increase to 100% by 2040. Lower Bound Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)	**S8:** AI computing power in eight hubs will represent over 90% by 2050. Computing efficiency will reach a 90% localization rate by 2050, with domestic chips achieving global leading standards by 2030. PUE for all data center hubs will fall to 1.1 after 2030. The proportion of green electricity in eastern and western hubs will gradually increase to 100% by 2040. Lower Bound Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)	**Advanced Green (MAX):** AI computing power in Western hubs will exceed 50% by 2050. Computing efficiency will reach a 90% localization rate by 2050, with domestic chips achieving global leading standards by 2030. PUE for all data center hubs will fall to 1.1 after 2030. The proportion of green electricity in eastern and western hubs will gradually increase to 100% by 2040. Lower Bound Scenario in China Regional Power Grids Carbon Dioxide Emission Factors (2023)

Fig. 6. Scenario Settings for Operational Emissions of AI Data Centers in China. The analysis assumes a constant annual working time of 8760 h and a green electricity carbon emission factor of 0.024–0.028 tCO_2/MWh across all scenarios. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
National Renewable Energy Laboratory (NREL), International Renewable Energy Agency (IRENA), and National Bureau of Statistics (NBS)

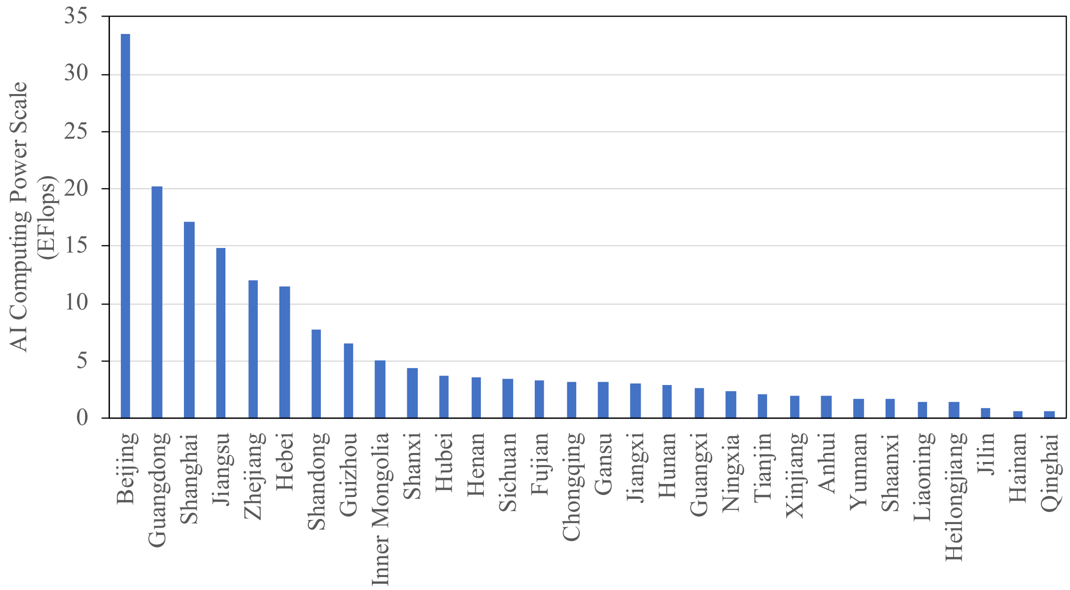

Fig. 7. Distribution of provincial computing power in 2022. Data is sourced from the White Paper on China Computing Power Index (China Academy of Information and Communications Technology (CAICT), 2023). The four data center hubs in China's western region are the Inner Mongolia Hub (including the Helinger Cluster), Gansu Hub (including the Qingyang Cluster), Ningxia Hub (including the Zhongwei Cluster), and Guizhou Hub (including the Guian Cluster). The four data center hubs in China's eastern region are the Beijing-Tianjin-Hebei Hub (including the Zhangjiakou Cluster), Yangtze River Delta Hub (including the Wuhu Cluster and Demonstration Zone of the Yangtze River Delta Cluster), Greater Bay Hub (including the Shaoguan Cluster), and Chengdu-Chongqing Hub (including the Tianfu Cluster and Chongqing Cluster).

ranging from 50 % to 90 %,[5] providing the basis for scenario assumptions.

PUE. To mitigate carbon emissions from data centers, the local government in China has introduced a series of regulatory mandates and financial incentives to improve PUE, such as those in Shanghai,[6] Shenzhen[7] and Guizhou.[8] Under the Moderate Progress and Accelerated Transition scenarios, PUE targets are set based on provincial plans and national objectives.[9] By 2035, data centers in eastern hubs are required to achieve a PUE of less than 1.25, those in western hubs less than 1.2, and others less than 1.3. The Net-zero Commitment scenario envisions even more aggressive technological advancements, including the widespread adoption of liquid cooling, AI-driven energy management, and modular data center architectures, which collectively drive PUE across all hubs to $\leq$1.1 by 2030 (Masanet et al., 2020). These improvements are reinforced by government-led energy efficiency certification programs to achieve best-in-class PUE performance.

Green Electricity Adoption. Increasing the share of green electricity is also a key strategy for reducing carbon emissions within data centers. According to national objectives,[10] the utilization rate of renewable energy will increase by 10 % annually before the end of 2025, mandating that new AI data centers meet progressively stricter renewable energy procurement requirements. Under the Moderate Progress scenario, the adoption of green electricity in data centers is projected to grow steadily. By 2025, the proportion of green electricity in western hubs is expected to reach 30 %, with eastern hubs meeting the target by 2030. The Accelerated Transition scenario targets a 50 % adoption of green electricity by 2035 in western hubs and by 2040 in eastern hubs. Under the Net-zero Commitment scenario, green electricity adoption accelerates significantly, with both western and eastern hubs expected to achieve 100 % reliance on green electricity by 2040. This assumption aligns with China's net-zero carbon emissions policy, which aims for all digital infrastructure to be fully powered by clean energy in the future (Qiu et al., 2021).

Regional Carbon Emission Factors. The regional carbon emission factors for the three scenarios are derived from predictions provided by the Chinese Academy of Environmental Planning, which assesses China's provincial power grids from 2020 to 2035 (Chinese Academy of Environmental Planning, 2023). The report examines three scenarios: high-speed renewable energy development (upper bound), policy-driven renewable energy growth (lower bound), and an intermediate scenario that represents an optimized trajectory derived from 13,000 simulation runs, balancing economic feasibility with environmental goals. These scenarios offer a comprehensive range of carbon emission factors, enabling a robust analysis of potential future outcomes under varying levels of renewable energy adoption and technological advancements. Accordingly, the report's predictions form the basis for our scenario classification of regional carbon emission factors.

3.3. Manufacturing carbon emission modeling

3.3.1. Uncertainty-based architectural carbon modeling tool

Given the efficiency improvements and increasing hardware complexity over the past decade, the carbon footprint for computing systems has shifted from operational emissions to manufacturing emissions, which are owed to the production and packaging phase of processors, memory, and storage components (Gupta et al., 2022). For data

[5] The General Office of the People's Government of Inner Mongolia Autonomous Region, "Work Plan for Promoting the High-Quality Development of Digital Economy in Inner Mongolia Autonomous Region (2023-2025)." Shandong Province, "Action Plan for the Construction of Integrated Computing Power Network in Shandong Province (2022–2025)."

[6] Shanghai Municipal Economic and Information Commission, "Guidelines for the Unified Scheduling of Computing Power Resources in Shanghai."

[7] Shenzhen Municipal Bureau of Industry and Information Technology, "Action Plan for High-Quality Development of Computing Power Infrastructure in Shenzhen (2024-2025)."

[8] Guizhou Provincial People's Government Office, "Implementation Opinions on Accelerating the Construction of the 'East Data West Computing' Project and the National Integrated Computing Network National Hub Node (Guizhou)."

[9] National Development and Reform Commission, "Special Action Plan for the Green and Low-Carbon Development of Data Centers."

[10] National Development and Reform Commission, "Special Action Plan for the Green and Low-Carbon Development of Data Centers."

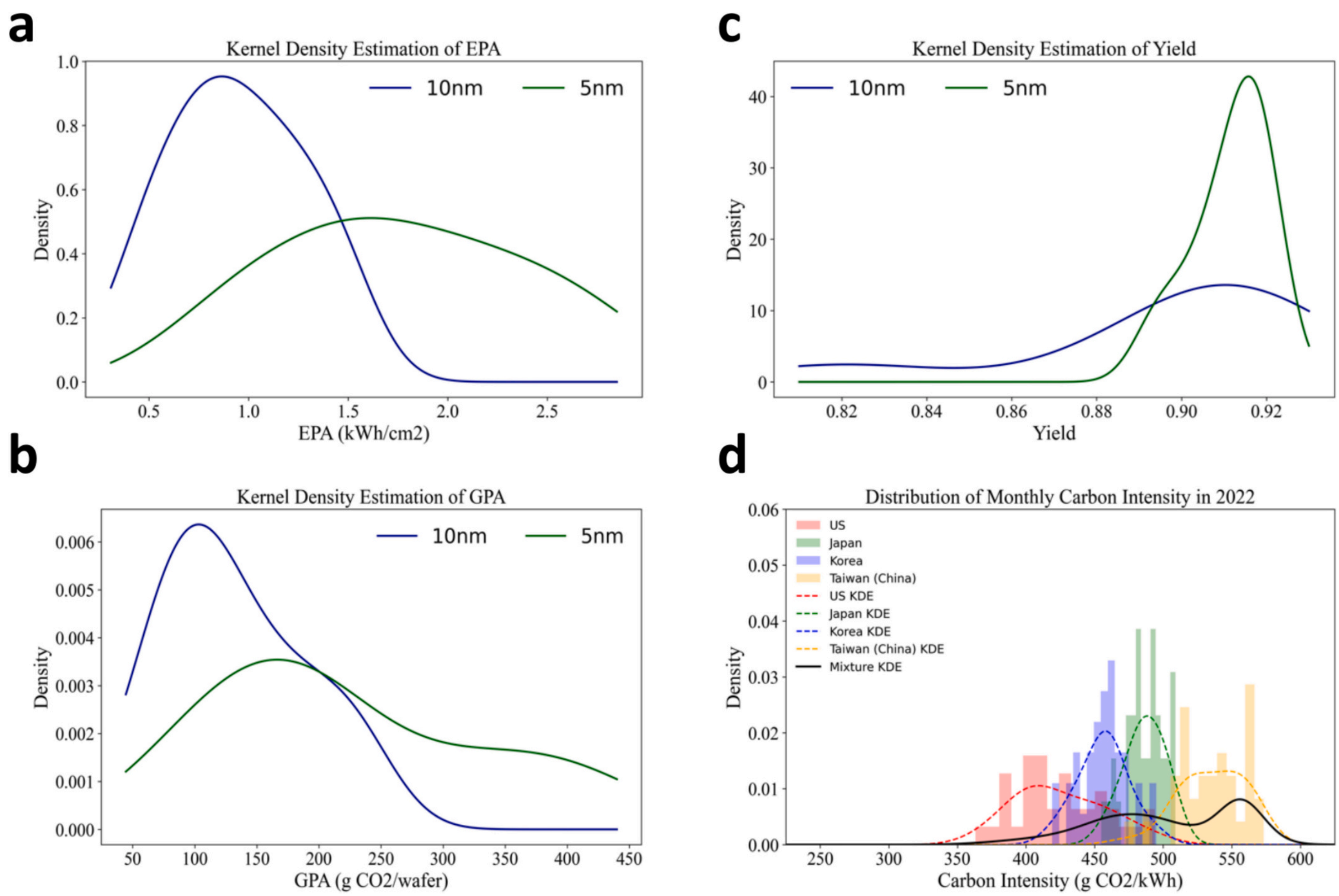

Fig. 8. Kernel Density Estimation of Key Parameters. a, EPA for 5 nm and 10 nm Nodes. b, GPA for 5 nm and 10 nm Nodes. c, Yield for 5 nm and 10 nm Nodes. d, Weighted Monthly Carbon Intensity by Production Region (2022).

center servers, the manufacturing emissions account for approximately 10 % to 60 % of the total carbon footprint as reported by the server vendors (Ji et al., 2024). In China, the Development Research Center of the State Council estimates that approximately 90 % of data center lifecycle emissions originate from operations, while construction accounts for around 10 %.[11]

Given the growing importance of hardware carbon footprints, various tools have been developed to estimate emissions across product life cycles. Among them, LCA tools are widely adopted in industry but are not well-suited for early-stage, fine-grained hardware modeling. Common LCA approaches, such as economic input-output (EIO-LCA) models and database-based tools, face limitations such as coarse granularity, reliance on outdated process assumptions, and a lack of component-level resolution. EIO-LCA estimates emissions based on average economic cost-to-carbon factors, whereas database-driven tools often utilize legacy data (e.g., 45 nm nodes) and provide only platform-level totals. Additionally, product environmental reports from industry typically present aggregated emissions across broad lifecycle stages without disaggregating by chip type or fabrication node. As a result, these tools offer limited value for data-driven, architecture-aware carbon optimization in modern AI hardware systems.

In this work, building on prior works of ACT (Gupta et al., 2022; Ji et al., 2024), we categorize manufacturing emissions into two main sources: the production and the packaging processes of electronic components. The production process involves creating electronic components from raw materials, encompassing stages such as wafer fabrication, doping, etching, and lithography. The packaging process entails assembling these components into functional chips and circuit boards. This includes dicing the wafer into individual chips, attaching them to substrates, wire bonding for electrical connections, and encapsulating them in protective materials. The formula for calculating manufacturing emissions is as follows:

$$C_m = C_{chip} + C_{mem} + C_{package} \tag{11}$$

$$C_{chip} = CPA \times Area = \frac{1}{Y}\left(CI_{fab} \times EPA + GPA + MPA\right) \times Area \tag{12}$$

$$C_{mem} = CPC \times Capacity \tag{13}$$

C_m denotes the total manufacturing emission, comprising carbon emissions from both the production and packaging processes. Production-related emissions consist of emissions from AI chips (C_{chip}, such as GPUs and CPUs) and memory components (C_{mem}, including DRAM and SSD). The packaging-related emissions from these electronic components are represented by $C_{package}$.

Modeling production emissions for various types of electronic components requires distinct approaches. For functional circuit chips, such as CPUs and GPUs, production emissions are calculated based on the die area (*Area*) and carbon emissions per unit area (*CPA*). CPA is determined by several semiconductor fabrication parameters, including the fab yield (Y), energy consumed per unit area produced (*EPA*), emissions per unit area from chemicals burned during hardware production (*GPA*, including gases such as perfluorocarbons (PFCs) and other high-global-

[11] Development Research Center of the State Council. Retrieved from https://www.drc.gov.cn/DocView.aspx?chnid=379&docid=2907711&leafid=1338&utm_source=chatgpt.com

warming-potential process gases), and emissions from raw material procurement for fabrication (*MPA*). The energy consumed during production is converted into carbon emissions using the carbon intensity of fabrication facilities (CI_{fab}). For memory and storage systems, such as DRAM and SSDs, production emissions are estimated based on the device capacity (*Capacity*) and the carbon-per-capacity factor (*CPC*).

The inherent uncertainties in hardware and software characteristics (e.g., spatial, temporal, process-driven, and system-driven) underscore the need for an uncertainty-based approach to address the wide variations in embodied carbon emission estimates (Eggleston et al., 2006). For example, spatial uncertainties stem from regional variations in grid carbon intensity and manufacturing practices; temporal uncertainties reflect changes in process efficiency and policy over time; process-driven uncertainties involve fab-level differences in gas usage, materials, and yield; and system-driven uncertainties arise from hardware design, packaging, and software co-optimization.

To account for real-world variability in chip manufacturing, we extend the ACT model into a probabilistic framework by replacing fixed parameter values with probability distributions. These distributions are generated using kernel density estimation (KDE), which smooths limited observed data into continuous curves. We then apply Monte Carlo simulations to repeatedly sample from these distributions and model a wide range of fabrication scenarios. By combining input parameters through outer addition or multiplication, we derive time-varying distributions of manufacturing carbon emissions that capture multiple sources of uncertainty and reflect the evolving nature of semiconductor production.

We augment the original ACT framework by modeling *EPA*, *GPA*, *Y*, and CI_{fab} probabilistically, assuming that the annual decrease rate of *MPA* aligns with the average yearly reduction in final energy carbon intensity as projected in the IEA's Stated Policies Scenario (IEA, 2019). Moreover, while the number of metal layers impacts EPA and GPA and introduces uncertainty, we assume they are fixed per process node due to the lack of layer-level data (Boakes et al., 2023). It should also be noted that our estimation of manufacturing carbon emissions focuses exclusively on the direct environmental impacts of chip fabrication processes. Indirect emissions (i.e., from the construction and maintenance of semiconductor facilities, as well as the production of specialized manufacturing equipment) are beyond the scope of this model. Consequently, the reported values may underestimate the full lifecycle carbon burden and should be regarded as a lower bound.

3.3.2. Parameter-level uncertainty characterization

In this study, we estimate manufacturing carbon emissions using the NVIDIA DGX H100 server as a representative example of AI infrastructure widely deployed in modern data centers. The system incorporates a 10 nm CPU and a 5 nm GPU, along with memory and storage components. Detailed specifications of its hardware configuration are provided in Table S1 of the supplementary information. In this section, we characterize the uncertainties associated with key fabrication-level parameters that influence the manufacturing emissions of functional circuit chips.

EPA (Energy per Unit Area) represents the energy consumed during semiconductor manufacturing per unit area and is influenced by temporal variations in process energy efficiencies. Using annual efficiency improvement data from TSMC (Taiwan Semiconductor Manufacturing Company) and imec (Interuniversity Microelectronics Centre), we generated two EPA distributions: one based on the original ACT framework (0.8–3.5 kWh/cm^2) (Gupta et al., 2022) and another using updated imec data (1.56–3.77 kWh/cm^2) (Boakes et al., 2023). These distributions were constructed by normalizing process node energy efficiencies to their initial values and adjusting EPA values by yearly efficiency improvements at the time of mass production.

As shown in Fig. 8a, the 10 nm EPA distribution (blue) is narrower and peaks between 0.5 and 1.5 kWh/cm^2, indicating relatively stable and lower energy intensity. In contrast, the 5 nm EPA distribution (green) is broader and shifts toward higher values (1.0–2.0 kWh/cm^2), reflecting increased energy intensity and greater process variability at more advanced nodes. This divergence arises from the higher complexity of smaller technology nodes, which require denser component integration, more advanced materials, and energy-intensive fabrication techniques. These factors contribute to both elevated average energy consumption and wider uncertainty during manufacturing. Similar trends have been reported in prior studies (Bhagavathula et al., 2024), which found that chips manufactured at smaller nodes consistently exhibit higher EPA and GPA values than those produced at larger nodes.

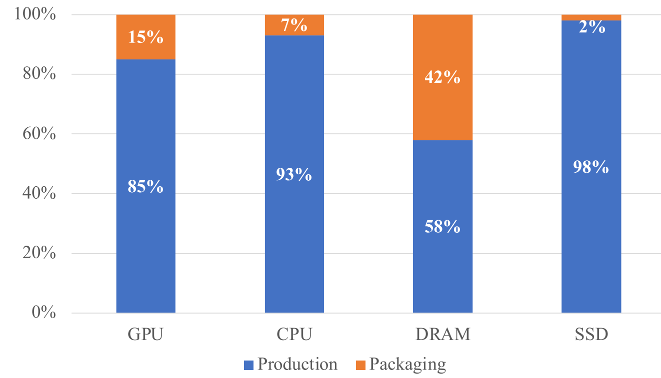

Fig. 9. Proportion of Carbon Emissions from Production and Packaging Phases.

GPA (Greenhouse Gas Emissions per Wafer) refers to non-energy emissions from semiconductor fabrication, primarily caused by process gases. Its uncertainty arises from errors in emission factors and temporal changes in gas composition and abatement efficiency. Following the approach of Bhagavathula et al. (2024), we constructed GPA distributions using 95 % confidence intervals of Tier 2 emission factors from 2006 to 2023, with deterministic resampling where data were unavailable. KDE was applied to capture both process-level and temporal variability.

As shown in Fig. 8b, the 10 nm GPA distribution is narrower and peaks between 100 and 150 g CO_2/wafer, indicating lower and more consistent process emissions. In contrast, the 5 nm GPA distribution is broader and shifts toward higher values (150–250 g CO_2/wafer). This difference is mainly due to the increased use of special process gases and more frequent chemical treatments required to build smaller and denser chip structures, which shows a similar pattern to the EPA distributions.

Yield, defined as the proportion of defect-free dies on a wafer, is influenced by temporal shifts in defect density during manufacturing. Using TSMC's defect density data (defects/cm^2) for two process nodes[12] and the Poisson yield model, we calculated yield values over time. KDE was applied to generate probability density functions for resampling.

As shown in Fig. 8c, both the 10 nm and 5 nm process nodes show peak yield values between 0.90 and 0.92, indicating high production efficiency. However, the 5 nm yield distribution is noticeably narrower, suggesting more consistent and stable manufacturing outcomes. This may be partly due to the adoption of EUV (Extreme Ultraviolet) lithography at the 5 nm node, which reduces the number of manufacturing steps. Fewer steps mean fewer chances for defects, leading to better and more uniform yields across production batches.[13]

CI_{fab} represents the carbon intensity associated with the energy consumed during semiconductor manufacturing. Uncertainty in carbon

[12] TSMC Logic Technology. Retrieved from https://www.tsmc.com/english/dedicatedFoundry/technology/logic

[13] AnandTech. Retrieved from https://www.anandtech.com/show/16028/better-yield-on-5nm-than-7nm-tsmc-update-on-defect-rates-for-n5

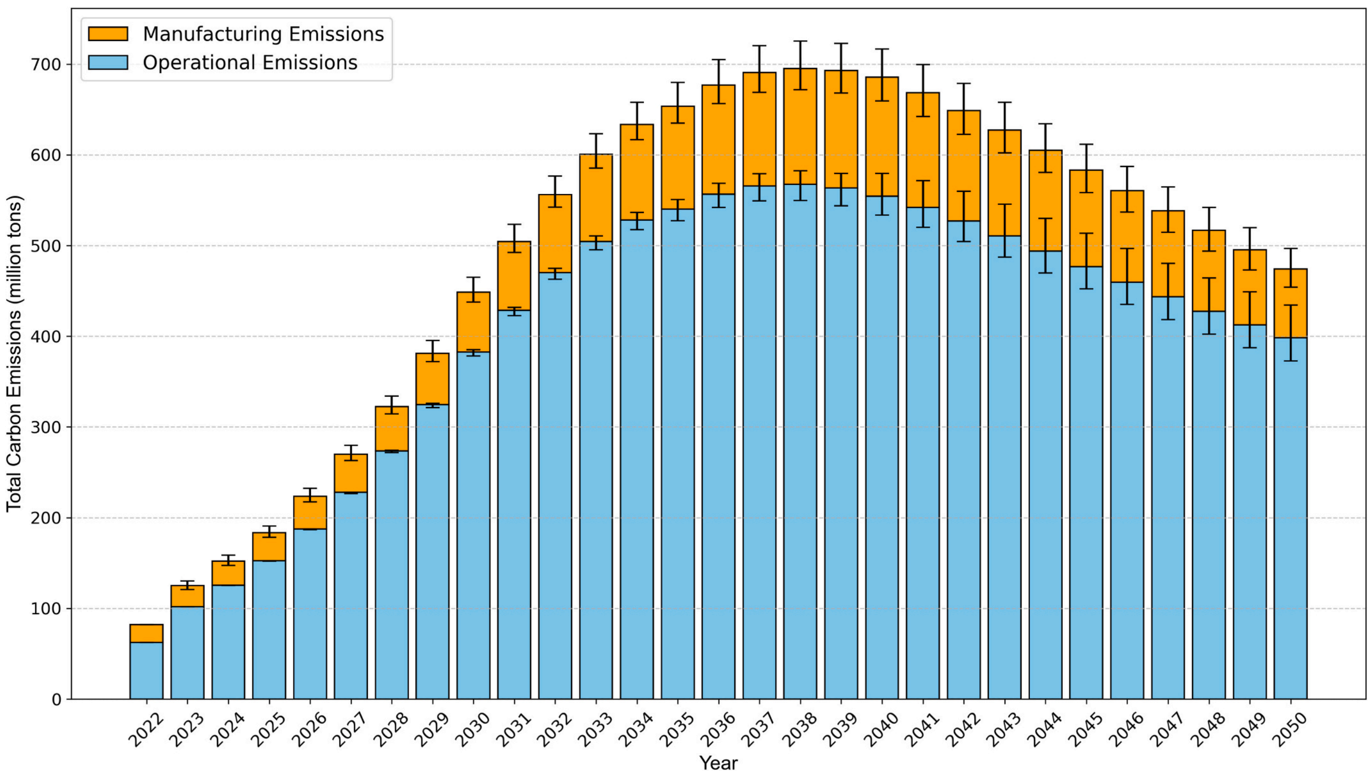

Fig. 10. Carbon Footprint of AI Data Centers in China (2022–2050). Operational emissions represent baseline results for the Business As Usual (BAU) scenario, with error bars indicating the upper and lower bounds due to variations in the distribution of AI computing power scale under the assumption of moderate technology progress. Manufacturing emissions display the mean values based on bootstrapping samples ($n = 10{,}000$) each year, with error bars representing the maximum and minimum values in the annual distribution.

intensity arises from three key factors: (1) seasonal fluctuations in carbon intensity due to variations in energy demand, with data sourced from Electricity Maps[14]; (2) geographic shifts in wafer production capacity across regions, using data from the Semiconductor Industry Association (SIA) (Varadarajan et al., 2024); and (3) future reductions in electricity generation carbon intensity projected by the IEA (IEA, 2020). To address these uncertainties, we first generated KDE curves for the carbon intensity of major fabrication regions based on historical data. These regional KDEs were then aggregated using weights derived from global wafer fabrication capacity to create global carbon intensity distribution for each year.

Fig. 8d illustrates the global carbon intensity distribution for 2022, which incorporates the weighted contributions of major semiconductor-producing regions. Northeast Asia collectively contributes over half of worldwide semiconductor fabrication capacity (Varadarajan et al., 2024). These shares were used as weights to blend the regional KDE curves into a single global distribution. The use of international data reflects the reality that modern AI chips are fabricated in multiple countries and regions. As such, accounting for geographic heterogeneity in grid emissions is critical to producing a realistic estimate of CI_{fab}. A dynamic visualization of carbon intensity from 2022 to 2050 is included in the appendix.

3.3.3. *Carbon emissions from the packaging phase*

The packaging process of electronic components is a significant contributor to overall manufacturing emissions (Li et al., 2023a). Fig. 9 illustrates the proportion of carbon emissions from the production and packaging phases of various hardware components. For example, while the majority of SSD emissions originate from the production phase, packaging accounts for 42 % of DRAM emissions. This higher proportion is attributed to the smaller size of DRAM chips and the need for precise, delicate packaging to protect them from external factors such as temperature fluctuations, humidity, and electrostatic discharge. In contrast, SSDs require less intricate packaging, resulting in a lower contribution from packaging to their total manufacturing emissions. Similarly, packaging accounts for approximately 10 % of the total manufacturing emissions of GPUs and CPUs. This smaller share is likely due to the more complex production processes of GPUs and CPUs, which involve advanced lithography and larger die areas. These production complexities result in the production phase dominating their overall carbon emissions.

4. Results and discussions

4.1. *Comprehensive assessment of AI carbon footprint*

We assess the AI carbon footprint from 2022 to 2050, categorized into operational and manufacturing emissions. Fig. 10 illustrates the temporal trends in these two emission categories. The findings reveal a substantial increase in the AI carbon footprint, rising from 82 Mt. in 2022 to a peak of 695 Mt. in 2038. This increase underscores a significant challenge: while China aims to peak emissions by 2030, the rapid rise of emissions post-2030 highlights the need for major reductions in other sectors to accommodate the growing energy demands of AI. In other words, the expanding emissions from the AI sector will compete for allowances within China's emission cap, potentially creating ripple effects across industries, especially those that are less energy-intensive. After 2040, total emissions are expected to decrease significantly, reaching 474 Mt. by 2050, as data center efficiency improves and clean electricity sources become more prevalent.

[14] Electricity Maps. Retrieved from https://portal.electricitymaps.com/datasets

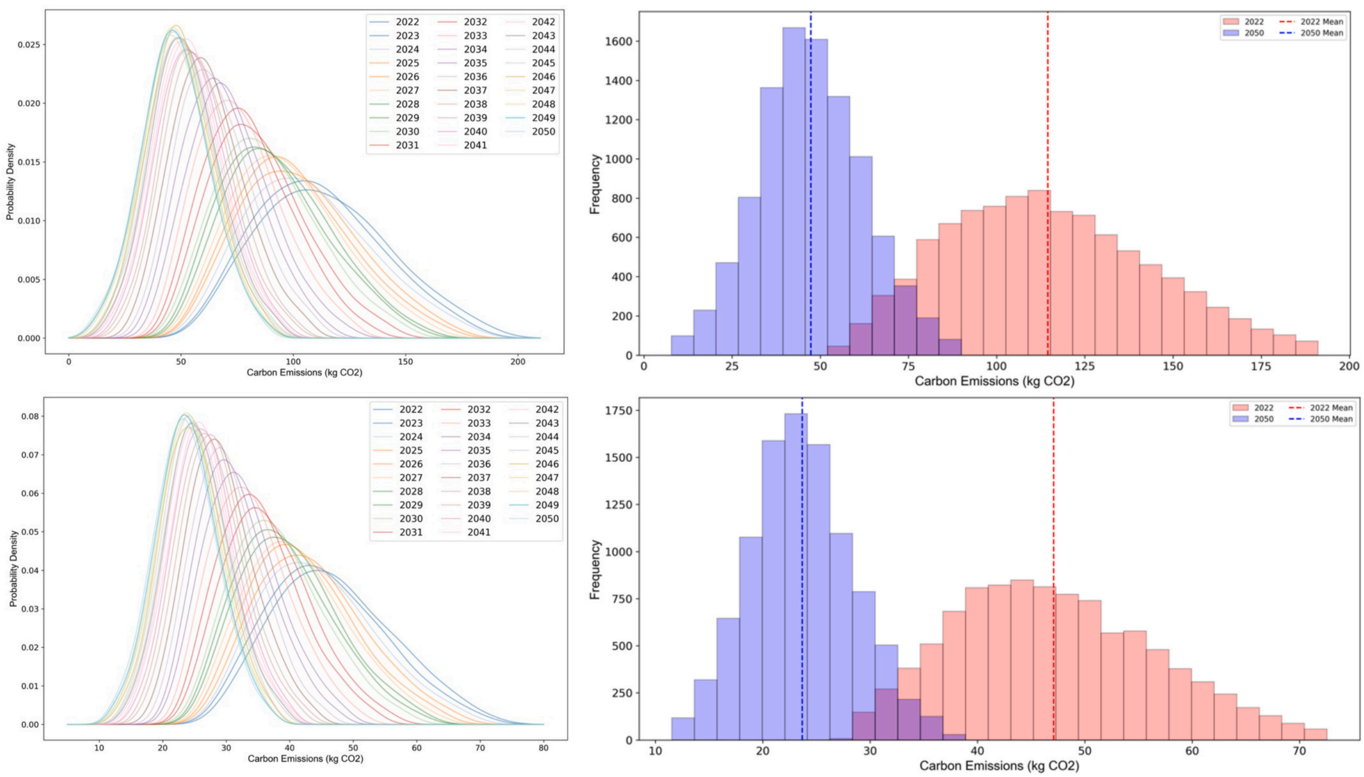

Fig. 11. Probability Distribution Simulations of Chip Manufacturing Emissions. This figure presents the probability distribution of manufacturing emissions, with the first row representing GPUs and the second row representing CPUs. The first column shows the trend of manufacturing emissions distribution from 2022 to 2050. The second column provides a specific comparison of the distribution of manufacturing emissions between 2022 and 2050. For GPUs, the mean emissions in 2022 and 2050 are 114.8 kg CO_2 (SD: 28.7, Range: 56.5–192.7) and 47.5 kg CO_2 (SD: 14.8, Range: 8.6–90.0), respectively. For CPUs, the mean emissions in 2022 and 2050 were 47.0 kg CO_2 (SD: 9.5, Range: 28.0–73.5) and 23.8 kg CO_2 (SD: 4.9, Range: 11.4–37.6), respectively.

Manufacturing emissions constitute a significant share of the carbon footprint, with an average contribution of 18 %. These emissions are projected to peak at 131 Mt. in 2040, followed by a sharp decline to 76 Mt. by 2050. The increasing demand for more advanced chips will intensify the need for cleaner manufacturing processes. If these processes remain carbon-intensive, costs may rise, potentially creating inefficiencies that could be passed on to consumers, impacting the overall economic competitiveness of AI products.

Operational emissions exhibit a similar trend, peaking at 568 Mt. in 2038, and decreasing to 398 Mt. by 2050. However, the reduction rate for operational emissions is slower, with operational emissions projected to be 70 % of their peak level by 2050, while manufacturing emissions are expected to decrease to 58 %. The slower decline is primarily due to the high carbon intensity of China's energy mix, with a large portion of electricity still generated from coal-fired power plants. In contrast, chip manufacturing emissions are distributed globally. The regions like the United States and South Korea, with cleaner energy sources and more efficient technologies, are expected to achieve net-zero emissions sooner. Given that energy costs account for over half of the total cost of data centers,[15] the slower reduction in operational emissions could result in higher operational costs for AI data centers in China. This, in turn, may affect the cost-effectiveness of AI infrastructure and potentially undermine the global competitiveness of Chinese AI companies.

4.2. Stochastic results of the manufacturing carbon emissions

The adoption of emerging technologies and the progress of national net-zero initiatives involve significant uncertainties, which can lead to unintended consequences in the manufacturing phases. Using deterministic carbon models may cause estimation results to deviate from reality. To address these complexities, we refer to the prior works (Bhagavathula et al., 2024; Gupta et al., 2022) and utilize a novel probabilistic framework that accounts for inherent uncertainties in hardware characteristics, providing distribution-based outputs of embodied carbon emissions. We extend existing models by incorporating expected future changes in national chip production capacities, the declining trend in carbon intensity of fabrication over time, and variations in emissions from procuring raw materials for fabrication manufacturing, using data from reports by the International Energy Agency (IEA) and the Semiconductor Industry Association (SIA). In this study, we simulate the manufacturing emissions of Nvidia H100 systems using Kernel Density Estimation (KDE), characterizing the uncertainty of carbon intensity of fabrication (CI), energy-per-area (EPA), gas-per-area (GPA), material emissions-per-area (MPA) and fab yield. Fig. 11 shows the probability distribution of chip manufacturing emissions from 10,000 Monte Carlo simulation samples. Both GPU and CPU manufacturing emissions exhibit similar trends, showing a gradual leftward shift and a more concentrated distribution from 2022 to 2050. This trend reflects the stabilization of manufacturing technologies and a decrease in the carbon intensity of fabrication. By 2050, both the mean and standard deviation of the emission distribution will be lower than in 2022, indicating improvements in manufacturing efficiency and the adoption of cleaner energy sources.

4.3. Scenario analysis of the operational carbon emissions

Illustrated in Fig. 12 are the operational carbon emissions from AI data centers in China under a 3 × 3 scenario matrix from 2022 to 2050.

[15] China Business Industry Research Institute, "2025 China Data Center Industry Market Outlook and Forecast Report."

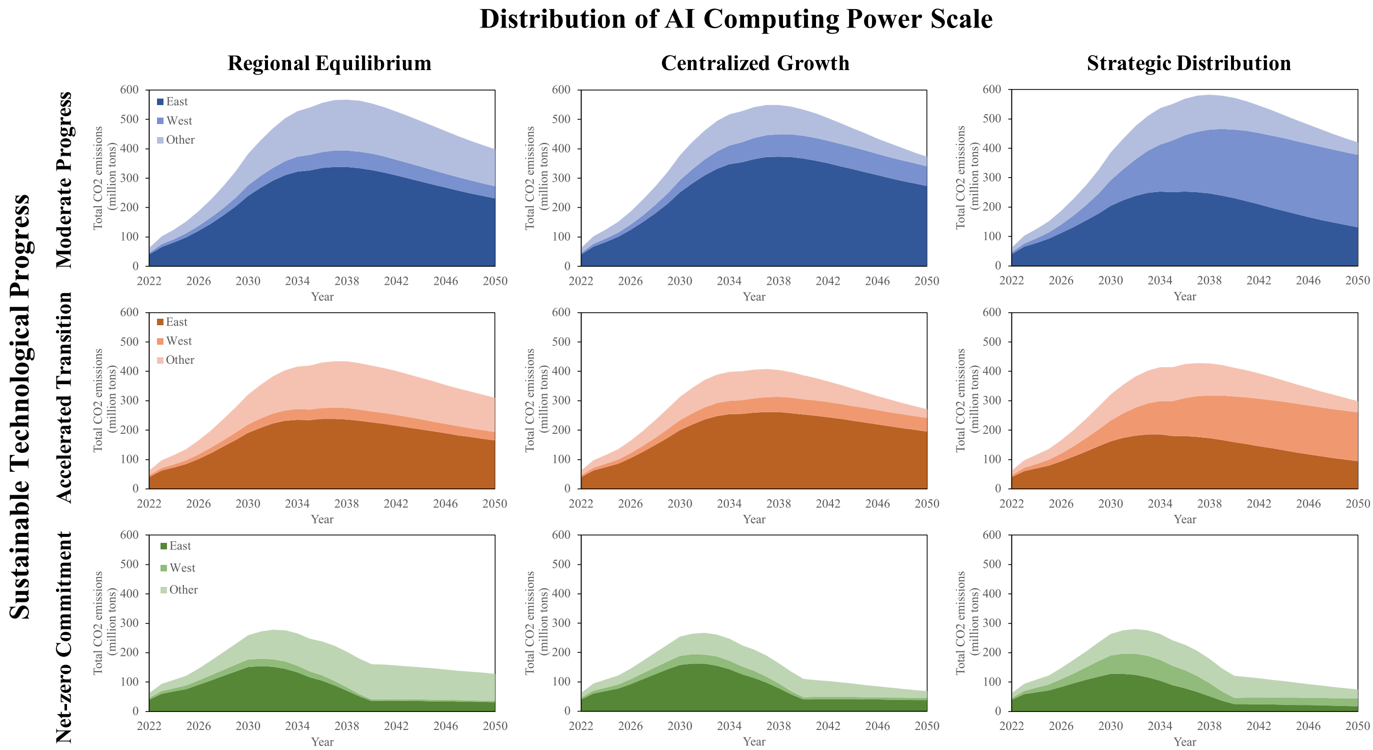

Fig. 12. Operational Carbon Emissions of AI Data Centers in China (2022–2050). The horizontal axis represents three scenarios of AI computing power distribution: Regional Equilibrium, Centralized Growth, and Strategic Distribution. The vertical axis represents three scenarios for sustainable technological progress: Moderate Progress, Accelerated Transition, and Net-zero Commitment. Working hours per year and the carbon emission factor of green electricity remain constant across all scenarios. Operational emissions in each subplot are categorized into three regions: eastern hubs, western hubs, and other provinces. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please see Supplementary Fig. S1 for detailed breakdowns of electricity consumption. Despite different assumptions about the distribution of computing power and technological advancements, operational emissions consistently show an initial rise followed by a decline. All nine scenarios reach their peak emissions after 2030.

From a regional perspective (horizontal comparison), the operational carbon emissions across the scenarios exhibit relatively stable trends over time. However, there is a notable shift in the regional distribution of emissions. Particularly, the emissions of Western hubs are on the rise, especially under the Strategic Distribution scenarios (S3, S6, S8), where it is assumed that AI computing power in Western hubs will exceed 50 % by 2050. This shift indicates the redistribution of computing resources, with significant computing loads being moved to the western regions, resulting in higher carbon emissions in these areas over time. The horizontal comparison results indicate that, under constant technological assumptions, this regional shift in AI computing power only transfers operational carbon emissions to the western regions, without resulting in an overall reduction (in fact, an increase) in the national carbon footprint. This is because the current PUE in the western regions is higher than that in North and East China (see Section 3.1.3). Additionally, the limited local power consumption capacity and inflexibility of the local power grid result in continued reliance on thermal power generation for peak load balancing, thereby keeping the grid's emission factor relatively high.

From a technological progress perspective (vertical comparison), there is a clear downward trend in carbon emissions as the technological assumptions become increasingly stringent. Although the peak time for the Accelerated Transition (S4, S5, S6) scenarios is similar to that of the BAU scenario, their peak emissions are 25 % lower. Furthermore, the Net-zero Commitment scenarios (S7, S8, S9) exhibit the lowest carbon emissions, peaking as early as 2032, five years ahead of the other scenarios. This is attributed to ambitious targets for the adoption of green electricity, where 100 % green energy usage is achieved in both eastern and western hubs by 2040, leading to a stabilization in carbon emissions after 2040.

Regarding the diagonal of the chart, the scenarios representing Business as Usual (BAU), Green Development (PRO), and Advanced Green (MAX) exhibit different emission patterns. The BAU and PRO scenarios exhibit a rapid escalation in carbon emissions, peaking at approximately 568 Mt. and 408 Mt., respectively, by 2038. Between 2022 and 2050, the PRO scenario is projected to reduce total carbon emissions by approximately 3084 Mt. compared to the BAU scenario. This significant reduction is primarily attributed to a lower carbon emission factor and higher adoption of green energy strategies within the PRO scenario. In the PRO scenario, emissions in the eastern hubs are substantially lower compared to the BAU scenario, while emissions in the western hubs show a modest increase due to the ongoing infrastructural and technological development in those regions.

The MAX scenario incorporates further enhancements, including significant improvements in PUE and a higher rate of adoption of green electricity. Our findings indicate that from 2022 to 2050, the MAX scenario can reduce carbon emissions by 7219 Mt. compared to the BAU scenario, demonstrating a strong potential for emissions reduction. However, the rapid development of western hubs in this scenario leads to higher carbon emissions in these regions before 2037. This is due to the rapid relocation of AI computing power to the western regions, outpacing the rate of technological advancements in these areas. The early stage of western hub development contributes to higher regional emissions, as improvements in energy infrastructure and the adoption of green electricity have not yet fully materialized. This highlights a critical trade-off: while the MAX scenario benefits from long-term emissions reductions, its short-term impact on the regional carbon footprint, particularly in the western regions, requires careful consideration for national carbon reduction goals.

Table 2
Operational Emissions from AI Data Centers in Each Province of China (in Million Tons).

	Year	Guizhou	InnerMongolia	Gansu	Ningxia	Beijing	Guangdong	Shanghai	Jiangsu	Zhejiang	Hebei	Sichuan	Chongqing	Tianjin	Shandong	Shanxi
BAU	**2025**	3.49	5.38	2.39	2.25	33.18	11.19	11.43	15.48	8.29	14.49	0.59	1.91	2.23	7.56	5.08
	2030	6.95	15.53	6.47	6.12	78.78	28.14	29.08	35.67	21.33	36.91	1.20	4.04	5.02	19.76	14.68
	2035	10.17	25.87	7.44	9.23	82.62	46.32	46.00	52.93	30.92	53.00	1.37	5.80	7.39	28.20	24.53
	2040	9.30	29.43	7.05	9.78	72.57	50.55	49.59	53.41	32.41	55.09	1.20	5.54	7.33	28.75	27.96
	2045	7.51	26.92	5.70	8.59	56.74	45.13	43.96	45.13	28.03	47.38	0.96	4.57	6.17	24.51	25.59
	2050	6.09	23.44	4.53	7.30	44.72	38.75	37.58	37.16	23.42	39.44	0.78	3.74	5.11	20.35	22.29
PRO	**2025**	3.51	5.37	2.43	2.37	29.75	10.06	8.81	13.56	7.61	12.73	0.58	1.75	2.23	6.78	4.48
	2030	6.75	14.65	6.48	6.52	66.66	23.60	19.65	30.18	18.52	30.95	1.31	3.60	5.24	15.25	10.92
	2035	8.97	21.71	7.44	8.47	65.93	35.45	30.74	41.10	25.50	42.06	1.68	5.00	8.14	18.08	14.99
	2040	8.84	26.11	7.92	9.54	54.92	37.43	34.20	40.90	26.50	43.47	1.75	4.89	8.65	14.65	13.38
	2045	7.78	25.73	7.21	9.16	44.54	34.39	32.44	36.12	24.05	39.37	1.63	4.40	8.23	9.45	9.20
	2050	6.80	24.06	6.37	8.55	36.26	30.42	29.25	30.99	21.03	34.37	1.50	3.93	7.61	5.23	5.31
MAX	**2025**	4.75	7.48	3.17	4.31	24.12	9.86	7.99	10.86	6.29	10.24	0.51	0.99	1.77	6.08	4.34
	2030	10.29	24.79	10.90	15.94	40.88	18.00	14.04	18.69	11.90	19.32	0.84	1.75	2.86	14.26	11.06
	2035	9.31	29.88	10.07	17.10	19.82	14.73	12.06	14.71	9.28	14.55	0.75	1.35	2.22	15.93	14.31
	2040	5.91	5.44	4.87	4.62	7.18	4.34	3.66	3.19	2.57	2.45	0.75	0.69	0.45	13.37	13.29
	2045	6.82	6.36	5.80	5.56	6.32	3.82	3.23	2.81	2.27	2.16	0.66	0.61	0.40	8.92	9.40
	2050	7.16	6.75	6.24	6.02	5.14	3.11	2.62	2.28	1.84	1.76	0.53	0.49	0.32	5.28	5.76

	Year	Hubei	Henan	Fujian	Jiangxi	Hunan	Guangxi	Xinjiang	Anhui	Yunnan	Shaanxi	Liaoning	Heilongjiang	Jilin	Qinghai	Hainan
BAU	**2025**	2.06	4.32	2.02	2.96	2.11	1.55	2.77	2.65	0.34	1.90	1.69	1.63	0.94	0.08	0.41
	2030	5.94	10.32	5.50	7.64	5.59	4.59	8.00	7.64	0.75	5.45	4.21	4.29	2.16	0.12	0.85
	2035	9.24	14.62	8.80	10.96	8.57	7.41	11.67	12.14	0.61	8.47	5.61	6.78	2.41	0.10	0.93
	2040	10.23	14.33	9.59	11.18	9.05	8.41	12.60	13.57	0.48	9.38	5.43	7.26	2.09	0.08	0.77
	2045	9.16	11.83	8.55	9.45	7.92	7.68	11.05	12.20	0.36	8.35	4.49	6.36	1.62	0.07	0.58
	2050	7.83	9.55	7.34	7.73	6.72	6.68	9.28	10.45	0.28	7.07	3.66	5.37	1.27	0.06	0.44
PRO	**2025**	1.85	3.84	1.78	2.59	1.91	1.32	2.46	2.43	0.31	1.70	1.36	1.44	0.82	0.06	0.37
	2030	3.76	7.74	3.98	5.88	4.25	3.23	6.00	5.50	0.49	4.15	2.87	3.24	1.55	0.08	0.63
	2035	4.65	8.99	5.25	6.98	5.23	4.35	7.22	7.26	0.34	5.32	3.30	4.18	1.21	0.06	0.49
	2040	3.67	6.91	4.44	5.71	4.27	3.86	6.12	6.20	0.20	4.64	2.57	3.53	0.75	0.04	0.29
	2045	2.34	4.28	2.95	3.67	2.78	2.64	4.03	4.12	0.12	3.11	1.62	2.33	0.42	0.02	0.16
	2050	1.28	2.29	1.67	2.01	1.56	1.52	2.25	2.32	0.06	1.75	0.88	1.30	0.21	0.01	0.08
MAX	**2025**	1.82	3.09	1.69	2.15	2.01	1.32	1.83	2.13	0.26	1.33	1.13	1.30	0.74	0.05	0.31
	2030	3.77	6.63	4.04	5.07	4.58	3.40	4.66	4.91	0.38	3.39	2.38	3.03	1.31	0.08	0.49
	2035	4.14	7.26	5.07	5.81	4.93	3.97	5.90	6.24	0.28	4.18	2.84	3.62	1.10	0.06	0.41
	2040	3.36	5.98	4.52	5.09	4.14	3.65	5.52	5.50	0.18	3.87	2.35	3.13	0.72	0.04	0.27
	2045	2.20	3.95	3.12	3.49	2.76	2.57	3.92	3.78	0.11	2.73	1.55	2.13	0.42	0.03	0.16
	2050	1.29	2.32	1.89	2.09	1.63	1.57	2.41	2.27	0.06	1.67	0.91	1.27	0.23	0.02	0.09

Note: Provinces marked in blue are in the western hub, and provinces marked in red are in the eastern hub.

To enhance the robustness of the results, we conduct a Local Sensitivity Analysis (LSA) to assess the impact of key parameters on electricity consumption and carbon emissions. Detailed results of the sensitivity analysis are provided in the supplementary information.

The geospatial nuances of the operational carbon footprint across China are illustrated in Table 2. In the BAU scenario, the eastern coastal provinces emerge as the primary drivers of both the increase and decrease in carbon emissions, including the Beijing-Tianjin-Hebei region, the Yangtze River Delta, and the Greater Bay Area. This scenario assumes that the regional distribution of AI computing power remains consistent with 2022 levels. As a result, emissions are concentrated along the eastern coastline, which has traditionally hosted the majority of the country's data centers.

Alternatively, in the PRO scenario, assuming regional agglomeration of AI data centers, the carbon emissions converge in eight designated AI hubs. With the partial relocation of computing power and advancements in technology, AI carbon emissions in the eastern coastal provinces are somewhat alleviated.

The northwest regions are identified as carbon hotspots in the MAX scenario, primarily shaped by the EWCRT project. Despite the implementation of advanced green measures, emissions in the western hubs remain high, as the rapid expansion of AI computing power outpaces technological improvements. These trends highlight the dual impact of AI expansion and regional development policies on carbon emissions. While the shift of AI computing power to western hubs does not result in significant reductions in local carbon emissions, the overall efforts to improve PUE and increase green electricity adoption are effective on a national scale.

4.4. *Impact of policy measures on carbon emission reduction*

This section analyzes the marginal carbon reduction effects of various policy measures. Three marginal improvements are considered based on the BAU scenario: increasing the proportion of green electricity, enhancing PUE, and reducing regional carbon emission factors. Each scenario assumes improving the corresponding measure to its optimal level (equivalent to the MAX scenario, as shown in Fig. 6), while maintaining the baseline conditions of all other measures in the BAU scenario.

As shown in Fig. 13, increasing the proportion of green electricity has the most significant carbon reduction effect, projected to reduce emissions by 4961 Mt. CO_2 from 2023 to 2050, resulting in a substantial 42 % reduction compared to the BAU scenario. Enhancing PUE and minimizing regional carbon emission factors have similar effects, contributing to approximately 13–15 % reductions in operational carbon emissions. Additionally, the emission reduction potential of all three strategies increases progressively over time. By 2040, the cumulative reduction potential of these three strategies consistently remains around 90 %. This trend suggests that continued advancements in green energy adoption, PUE optimization, and carbon emission factor management will become increasingly crucial with the expansion of AI computing power.

In summary, while all three strategies contribute to reducing carbon emissions, increasing the proportion of green electricity is the most effective. The growing potential for emission reductions highlights the importance of integrating these improvements to achieve substantial and sustained reductions in carbon emissions from AI data centers.

5. Conclusions and policy recommendations

5.1. *Conclusions*

The rapid advancements in AI have led to a substantial increase in energy consumption, raising global concerns about associated carbon emissions. As a key player in the global AI race, China faces a critical juncture where its AI development aligns with its Carbon Peaking and

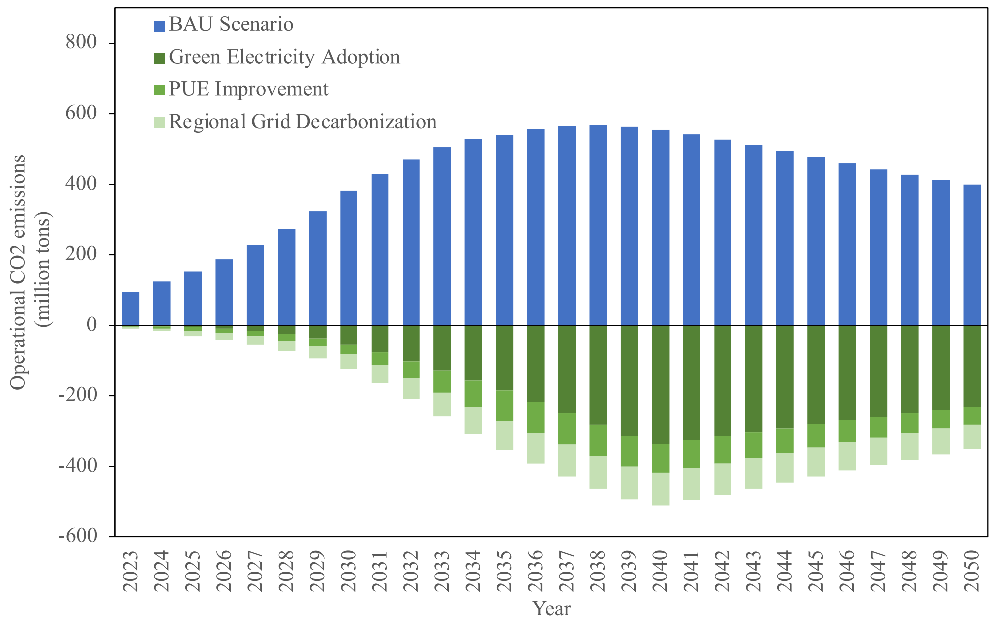

Fig. 13. Marginal Impact of Policy Measures on Carbon Emissions Reduction (2023-2050). The blue bars represent operational emissions under the BAU scenario, and the segmented green bars depict the incremental reductions achieved by three policy measures: increasing internal green electricity usage, improving PUE, and reducing regional carbon emission factors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Carbon Neutrality Goals. In this study, we employ the uncertainty-based Architectural Carbon Modeling Tool and scenario analysis to assess the carbon footprint trajectory of AI data centers in China from 2022 to 2050. Our findings reveal a sharp rise in carbon footprint, increasing from 82 Mt. in 2022 to a peak of 695 Mt. in 2038, before declining to 474 Mt. by 2050, with manufacturing emissions accounting for about 18 %. The projected doubling of emissions by 2030 highlights the urgent need for accelerated reductions in other sectors to accommodate AI growth.

Geospatial analysis further demonstrates that in the BAU scenario, the eastern coastal provinces, such as the Beijing-Tianjin-Hebei region, the Yangtze River Delta, and the Greater Bay Area, dominate both the initial rise and eventual decline in carbon emissions due to the concentration of AI computing power. In the PRO scenario, emissions are concentrated in eight designated AI hubs, whereas the MAX scenario shifts computing demands to the northwest, creating new carbon hotspots. Despite green initiatives, emissions in the western regions remain high initially, due to the rapid expansion of AI outpacing technological improvements. These findings emphasize the dual challenges posed by AI's substantial energy demands and the uneven regional distribution of carbon emissions.

5.2. Policy implications

In light of the study's findings, we propose a set of actionable policy recommendations to help China mitigate the carbon footprint associated with AI data centers.

First, since increasing the proportion of green electricity and improving PUE together account for over 50 % of potential reductions in operational emissions, internal efficiency measures should focus on these two priorities. Specifically, AI data centers should be incentivized to increase their use of renewable energy by investing in on-site renewable generation, such as rooftop solar panels. Moreover, PUE enhancement can be achieved by adopting advanced cooling technologies, such as liquid cooling, and optimizing operational processes to reduce energy waste. AI can play a vital role in this process by analyzing energy consumption patterns and identifying inefficiencies in real-time operations (Stern and Valero, 2021).

Second, as regional power grid decarbonization directly contributes to a 15 % reduction in operational emissions and indirectly mitigates upstream manufacturing emissions, external policy efforts must prioritize power system transformation. Policymakers should facilitate the integration of renewable energy into the national grid by investing in renewable infrastructure, providing tax incentives, and offering subsidies to promote the adoption of clean energy. Moreover, coordinated optimization between AI data centers and power generation facilities, commonly referred to as Compute-Energy Coordination, should be promoted to enhance power system efficiency. For example, AI technologies can be employed to improve the forecasting of weather conditions and electricity demand, enabling more accurate and responsive energy dispatch (Abdalla et al., 2021). Additionally, digital finance can support energy transition by facilitating investments in green infrastructure (Wang et al., 2024). Such coordination optimizes the use of renewable resources, reduces dependence on fossil fuels, and contributes to the development of a more flexible and low-carbon power grid.

Finally, regionally differentiated strategies are essential to align AI development with local energy conditions. Eastern provinces, which currently host a large share of AI workloads, should focus on improving operational efficiency by investing in compact and energy-efficient AI models. In contrast, western regions are witnessing a rapid expansion of AI computing capacity. Accelerating the deployment of renewable energy infrastructure in these areas is crucial to ensure that technological advancements keep pace with the increasing demand. Furthermore, targeted national investments in region-specific green digital infrastructure are particularly important, as capital plays a significant role in influencing consumption-based emissions across regions with different developmental characteristics (Chen et al., 2018; Guo et al., 2025). In this context, it has been suggested that energy transition policies should be tailored to regional characteristics, with the integration of renewable energy and urbanization dynamics driving sustainable resource utilization in different regions (Pan et al., 2024). These collaborative initiatives are essential for aligning AI development with global climate goals (SDG 13) while addressing regional energy inequalities.

5.3. Limitations

This study has several limitations that suggest directions for future research. First, the scenarios are based on simplified assumptions that can not fully capture the dynamic nature of technological innovation, policy shocks or market fluctuations. Future studies could incorporate additional variables to more comprehensively evaluate their potential impact on the AI carbon footprint. Additionally, the findings are specific to China. Comparative studies across different countries and regions would help illuminate how national contexts shape the environmental impacts of AI. Finally, this study does not address the broader socio-economic impacts of transitioning to greener AI data centers, such as economic costs, labor market effects, and regional disparities. Incorporating these dimensions would provide a more holistic understanding of sustainable AI development.

CRediT authorship contribution statement

Zhan-Ming Chen: Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Qiyang Xiong:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. **Jiahui Duan:** Writing – original draft, Visualization, Validation, Software. **Jianhong Ma:** Writing – review & editing, Resources. **Zhuo Chen:** Visualization, Data curation. **Shan Guo:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge financial support from the National Natural Science Foundation of China (Grant No. 72274206, 72004225, and 72574222), the Humanities and Social Science Fund of Ministry of Education of China (Grant No. 24YJAZH036), the Social Science Foundation of Beijing (Grant No. 23GLB029), and Big Data and Responsible Artificial Intelligence for National Governance, Renmin University of China.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eneco.2025.108880.

References

Abdalla, A.N., Nazir, M.S., Tao, H., Cao, S., Ji, R., Jiang, M., Yao, L., 2021. Integration of energy storage system and renewable energy sources based on artificial intelligence: an overview. J Energy Storage 40, 102811. https://doi.org/10.1016/j.est.2021.102811.

Ahmad, T., Zhang, D., Huang, C., Zhang, H., Dai, N., Song, Y., Chen, H., 2021. Artificial intelligence in sustainable energy industry: Status Quo, challenges and opportunities. J. Clean. Prod. 289, 125834. https://doi.org/10.1016/j.jclepro.2021.125834.

Ayres, R.U., 1995. Life cycle analysis: A critique. Resour. Conserv. Recycl., Life Cycle Manag. 14, 199–223. https://doi.org/10.1016/0921-3449(95)00017-D.

Aysan, A.F., Gozgor, G., Nanaeva, Z., 2024. Technological perspectives of metaverse for financial service providers. Technol. Forecast. Soc. Change 202, 123323.

Bacha, R., Gasmi, F., Metevier, S., 2024. Broadband adoption in Algeria and the structural determinants of its pace. Telecommun. Policy 48, 102761. https://doi.org/10.1016/j.telpol.2024.102761.

Bai, C., Yao, D., Xue, Q., 2025. Does artificial intelligence suppress firms' greenwashing behavior? Evidence from robot adoption in China. Energy Econ. 142, 108168. https://doi.org/10.1016/j.eneco.2024.108168.

Bhagavathula, A., Han, L., Gupta, U., 2024. Understanding the Implications of Uncertainty in Embodied Carbon Models for Sustainable Computing. HotCarbon.

Bitencourt, L., Abud, T., Santos, R., Borba, B., 2021. Bass diffusion model adaptation considering public policies to improve electric vehicle sales—a Brazilian case study. Energies 14, 5435. https://doi.org/10.3390/en14175435.

Boakes, L., Garcia Bardon, M., Schellekens, V., Liu, I.-Y., Vanhouche, B., Mirabelli, G., Sebaai, F., Van Winckel, L., Gallagher, E., Rolin, C., Ragnarsson, L.-Å., 2023. Cradle-to-gate life cycle assessment of CMOS logic technologies. In: 2023 International Electron Devices Meeting (IEDM). Presented at the 2023 International Electron Devices Meeting (IEDM), pp. 1–4. https://doi.org/10.1109/IEDM45741.2023.10413725.

Chen, Z.-M., Ohshita, S., Lenzen, M., Wiedmann, T., Jiborn, M., Chen, B., Lester, L., Guan, D., Meng, J., Xu, S., Chen, G., Zheng, X., Xue, J., Alsaedi, A., Hayat, T., Liu, Z., 2018. Consumption-based greenhouse gas emissions accounting with capital stock change highlights dynamics of fast-developing countries. Nat. Commun. 9, 3581. https://doi.org/10.1038/s41467-018-05905-y.

Chen, C., Hu, Y., Karuppiah, M., Kumar, P.M., 2021. Artificial intelligence on economic evaluation of energy efficiency and renewable energy technologies. Sustain Energy Technol Assess 47, 101358. https://doi.org/10.1016/j.seta.2021.101358.

Cheng, Y., Zhang, Y., Wang, J., Jiang, J., 2023. The impact of the urban digital economy on China's carbon intensity: spatial spillover and mediating effect. Resour. Conserv. Recycl. 189, 106762. https://doi.org/10.1016/j.resconrec.2022.106762.

China Academy of Information and Communications Technology (CAICT), 2022. White Paper on Data Center 2022. http://www.caict.ac.cn/english/research/whitepapers/202205/P020220510510013255944.pdf.

China Academy of Information and Communications Technology (CAICT), 2023. White Paper on China Computing Power Index 2023. http://www.caict.ac.cn/kxyj/qwfb/bps/202309/t20230914_461823.htm.

China Academy of Information and Communications Technology (CAICT), 2024a. Green Computing Technology Innovation Research Report 2024. http://www.caict.ac.cn/kxyj/qwfb/ztbg/202403/t20240315_473805.htm.

China Academy of Information and Communications Technology (CAICT), 2024b. China Comprehensive Computing Power Index Report 2024. http://www.caict.ac.cn/kxyj/qwfb/bps/202412/t20241231_649676.htm.

China Academy of Information and Communications Technology (CAICT), 2024c. China Green Computing Power Development Research Report 2024. http://www.caict.ac.cn/sytj/202308/P020230815520237777247.pdf.

China Academy of Information and Communications Technology (CAICT), 2025. Blue Paper on Advanced Computing and Computing Power Index 2024. https://www.caict.ac.cn/kxyj/qwfb/bps/202501/t20250117_651810.htm.

China Communications Services Corporation Limited, 2023. White Paper on China Data Center Industry Development. https://www.esensoft.com/industry-news/dx-26947.html.

Chinese Academy of Environmental Planning, 2023. China Regional Power Grids Carbon Dioxide Emission Factors (2023). http://www.caep.org.cn/sy/tdftzhyjzx/zxdt/202310/t20231027_1044179.shtml.

Dauvergne, P., 2022. Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. Rev. Int. Polit. Econ. 29, 696–718. https://doi.org/10.1080/09692290.2020.1814381.

de Vries, A., 2023. The growing energy footprint of artificial intelligence. Joule 7, 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004.

Desislavov, R., Martínez-Plumed, F., Hernández-Orallo, J., 2023. Trends in AI inference energy consumption: beyond the performance-vs-parameter laws of deep learning. Sustain. Comput. Inform. Syst. 38, 100857. https://doi.org/10.1016/j.suscom.2023.100857.

Dhar, P., 2020. The carbon impact of artificial intelligence. Nat. Mach. Intell. 2, 423–425. https://doi.org/10.1038/s42256-020-0219-9.

Ding, C., Ke, J., Levine, M., Zhou, N., 2024. Potential of artificial intelligence in reducing energy and carbon emissions of commercial buildings at scale. Nat. Commun. 15, 5916. https://doi.org/10.1038/s41467-024-50088-4.

Eggleston, H.S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K., 2006. 2006 IPCC Guidelines for National Greenhouse Gas Inventories.

Guo, S., Li, Y., Hu, Y., Xue, F., Chen, B., Chen, Z.-M., 2020. Embodied energy in service industry in global cities: a study of six Asian cities. Land Use Policy 91, 104264. https://doi.org/10.1016/j.landusepol.2019.104264.

Guo, Q., Peng, Y., Luo, K., 2025. The impact of artificial intelligence on energy environmental performance: empirical evidence from cities in China. Energy Econ. 141, 108136. https://doi.org/10.1016/j.eneco.2024.108136.

Guo, S., Tian, T., Gong, B., Wan, Y.H., Zhou, J.X., Wu, X.F., 2025. Urban shrinkage and carbon emissions: Demand-side accounting for Chinese cities. Appl. Energy 384, 125501. https://doi.org/10.1016/j.apenergy.2024.125501.

Gupta, U., Elgamal, M., Hills, G., Wei, G.-Y., Lee, H.-H.S., Brooks, D., Wu, C.-J., 2022. ACT: designing sustainable computer systems with an architectural carbon modeling tool. In: Proceedings of the 49th Annual International Symposium on Computer Architecture, ISCA '22. Association for Computing Machinery, New York, NY, USA, pp. 784–799. https://doi.org/10.1145/3470496.3527408.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J., 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. J. Mach. Learn. Res. 21, 1–43.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., Peste, A., 2021. Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks. J. Mach. Learn. Res. 22, 1–124.

IEA, 2019. Final Energy Carbon Intensity in the Stated Policies Scenario 2000–2040. https://www.iea.org/data-and-statistics/charts/final-energy-carbon-intensity-in-the-stated-policies-scenario-2000-2040.

IEA, 2020. Carbon Intensity of Electricity Generation in Selected Regions in the Sustainable Development Scenario, 2000–2040. https://www.iea.org/data-and-statistics/charts/carbon-intensity-of-electricity-generation-in-selected-regions-in-the-sustainable-development-scenario-2000-2040.

IEA, 2024. Electricity 2024. https://www.iea.org/reports/electricity-2024.

Ji, S., Yang, Z., Chen, X., Hu, J., Shi, Y., Jones, A.K., Zhou, P., 2024. Towards Data-center Level Carbon Modeling and Optimization for Deep Learning Inference. arXiv preprint. https://doi.org/10.48550/arXiv.2403.04976.

Jiang, M., Yu, X., 2025. Enhancing the resilience of urban energy systems: the role of artificial intelligence. Energy Econ. 144, 108313. https://doi.org/10.1016/j.eneco.2025.108313.

Jones, N., 2018. How to stop data centres from gobbling up the world's electricity. Nature 561, 163–166. https://doi.org/10.1038/d41586-018-06610-y.

Kaack, L.H., Donti, P.L., Strubell, E., Kamiya, G., Creutzig, F., Rolnick, D., 2022. Aligning artificial intelligence with climate change mitigation. Nat. Clim. Chang. 12, 518–527. https://doi.org/10.1038/s41558-022-01377-7.

Lau, C.K., Gozgor, G., Mahalik, M.K., Patel, G., Li, J., 2023. Introducing a new measure of energy transition: green quality of energy mix and its impact on CO_2 emissions. Energy Econ. 122, 106702. https://doi.org/10.1016/j.eneco.2023.106702.

Lei, N., Masanet, E., 2020. Statistical analysis for predicting location-specific data center PUE and its improvement potential. Energy 201, 117556. https://doi.org/10.1016/j.energy.2020.117556.

Li, B., Basu Roy, R., Wang, D., Samsi, S., Gadepally, V., Tiwari, D., 2023a. Toward sustainable HPC: Carbon footprint estimation and environmental implications of HPC systems. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Presented at the SC '23: International Conference for High Performance Computing, Networking, Storage and Analysis, ACM, Denver CO USA, pp. 1–15. https://doi.org/10.1145/3581784.3607035.

Li, J., Ma, S., Qu, Y., Wang, J., 2023b. The impact of artificial intelligence on firms' energy and resource efficiency: empirical evidence from China. Res. Policy 82, 103507. https://doi.org/10.1016/j.resourpol.2023.103507.

Li, T., Yu, L., Ma, Y., Duan, T., Huang, W., Zhou, Y., Jin, D., Li, Y., Jiang, T., 2023c. Carbon emissions of 5G mobile networks in China. Nat. Sustain. 6, 1620–1631. https://doi.org/10.1038/s41893-023-01206-5.

Liu, J., Chang, H., Forrest, J.Y.-L., Yang, B., 2020. Influence of artificial intelligence on technological innovation: evidence from the panel data of China's manufacturing sectors. Technol. Forecast. Soc. Change 158, 120142. https://doi.org/10.1016/j.techfore.2020.120142.

Liu, L., Yang, K., Fujii, H., Liu, J., 2021. Artificial intelligence and energy intensity in China's industrial sector: effect and transmission channel. Econ. Anal. Policy 70, 276–293. https://doi.org/10.1016/j.eap.2021.03.002.

Liu, J., Liu, L., Qian, Y., Song, S., 2022a. The effect of artificial intelligence on carbon intensity: evidence from China's industrial sector. Socio Econ. Plan. Sci. 83, 101002. https://doi.org/10.1016/j.seps.2020.101002.

Liu, Z., Sun, Y., Xing, C., Liu, J., He, Y., Zhou, Y., Zhang, G., 2022b. Artificial intelligence powered large-scale renewable integrations in multi-energy systems for carbon neutrality transition: challenges and future perspectives. Energy AI 10, 100195. https://doi.org/10.1016/j.egyai.2022.100195.

Liu, Y., Shen, F., Guo, J., Hu, G., Song, Y., 2025. Can artificial intelligence technology improve companies' capacity for green innovation? Evidence from listed companies in China. Energy Econ. 143, 108280. https://doi.org/10.1016/j.eneco.2025.108280.

Ma, J., Wang, N., Chen, Z., Wang, L., Xiong, Q., Chen, P., Zhang, H., Zheng, Y., Chen, Z.-M., 2024. Accounting and decomposition of China's CO_2 emissions 1981–2021. Appl. Energy 375, 124104. https://doi.org/10.1016/j.apenergy.2024.124104.

Masanet, E., Shehabi, A., Lei, N., Smith, S., Koomey, J., 2020. Recalibrating global data center energy-use estimates. Science 367, 984–986. https://doi.org/10.1126/science.aba3758.

Massiani, J., Gohs, A., 2015. The choice of bass model coefficients to forecast diffusion for innovative products: an empirical investigation for new automotive technologies. Res. Transp. Econ. 50, 17–28. https://doi.org/10.1016/j.retrec.2015.06.003.

Ministry of Finance, Ministry of Ecology and Environment, Ministry of Industry and Information Technology, 2023. Green Data Center Government Procurement Demand Standards. https://www.gov.cn/zhengce/zhengceku/2023-04/22/content_5752654.htm (accessed 10.21.24).

Ministry of Science and Technology, 2022. Implementation Plan for Science and Technology Support for Carbon Peaking and Carbon Neutrality (2022-2030). https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/qtwj/qtwj2022/202208/t20220817_181986.html.

Mora, C., Rollins, R.L., Taladay, K., Kantar, M.B., Chock, M.K., Shimada, M., Franklin, E. C., 2018. Bitcoin emissions alone could push global warming above 2°C. Nat. Clim. Chang. 8, 931–933. https://doi.org/10.1038/s41558-018-0321-8.

Nepal, R., Zhao, X., Dong, K., Wang, J., Sharif, A., 2025. Can artificial intelligence technology innovation boost energy resilience? The role of green finance. Energy Econ. 142, 108159. https://doi.org/10.1016/j.eneco.2024.108159.

Ni, W., Hu, X., Du, H., Kang, Y., Ju, Y., Wang, Q., 2024. CO_2 emission-mitigation pathways for China's data centers. Resour. Conserv. Recycl. 202, 107383. https://doi.org/10.1016/j.resconrec.2023.107383.

Ntwoku, H., Negash, S., Meso, P., 2017. ICT adoption in Cameroon SME: application of bass diffusion model. Inf. Technol. Dev. 23, 296–317.

Pan, W.C., Pruseth, S.K., Jose, A., Padhan, H., Gozgor, G., 2024. What drives natural capital in E7 and G7 economies? The roles of energy transition and urbanization. J. Clean. Prod. 476, 143811. https://doi.org/10.1016/j.jclepro.2024.143811.

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D.R., Texier, M., Dean, J., 2022. The carbon footprint of machine learning training will plateau, then shrink. Computer 55, 18–28. https://doi.org/10.1109/MC.2022.3148714.

Qin, M., Hu, W., Qi, X., Chang, T., 2024. Do the benefits outweigh the disadvantages? Exploring the role of artificial intelligence in renewable energy. Energy Econ. 131, 107403. https://doi.org/10.1016/j.eneco.2024.107403.

Qiu, S., Lei, T., Wu, J., Bi, S., 2021. Energy demand and supply planning of China through 2060. Energy 234, 121193. https://doi.org/10.1016/j.energy.2021.121193.

Russell, S.J., Norvig, P., 2016. Artificial intelligence: a modern approach. Pearson.

Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O., 2020. Green AI. Commun. ACM 63, 54–63. https://doi.org/10.1145/3381831.

Shehabi, A., Smith, S., Hubbard, Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D., 2024. 2024 United States Data Center Energy Usage Report. https://escholarship.org/uc/item/32d6m0d1 (accessed 3.26.25).

Sohrabpour, V., Oghazi, P., Toorajipour, R., Nazarpour, A., 2021. Export sales forecasting using artificial intelligence. Technol. Forecast. Soc. Change 163, 120480.

Song, M., Pan, H., Shen, Z., Tamayo-Verleene, K., 2024. Assessing the influence of artificial intelligence on the energy efficiency for sustainable ecological products value. Energy Econ. 131, 107392. https://doi.org/10.1016/j.eneco.2024.107392.

Stern, N., Valero, A., 2021. Innovation, growth and the transition to net-zero emissions. Res. Policy 50, 104293. https://doi.org/10.1016/j.respol.2021.104293.

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D.C.M., Ezer, D., van der Haert, F.C., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., de Wever, W., Teh, Y.W., Khan, M.E., Winne, R.D., Schaul, T., Clopath, C., 2020. AI for social good: unlocking the opportunity for positive impact. Nat. Commun. 11, 2468. https://doi.org/10.1038/s41467-020-15871-z.

Turk, T., Trkman, P., 2012. Bass model estimates for broadband diffusion in European countries. Technol. Forecast. Soc. Change 79, 85–96. https://doi.org/10.1016/j.techfore.2011.06.010.

Ukoba, K., Olatunji, K.O., Adeoye, E., Jen, T.-C., Madyira, D.M., 2024. Optimizing renewable energy systems through artificial intelligence: review and future prospects. Energy Environ. 35, 3833–3879.

Varadarajan, R., Iacob Koch-Weser, C.R., Fitzgerald, J., Singh, J., Thornton, M., Casanova, R., Isaacs, D., 2024. Emerging Resilience in the Semiconductor Supply Chain. https://www.semiconductors.org/wp-content/uploads/2024/05/Report_Emerging-Resilience-in-the-Semiconductor-Supply-Chain.pdf.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M., Fuso Nerini, F., 2020. The role of artificial intelligence in achieving the sustainable development goals. Nat. Commun. 11, 233. https://doi.org/10.1038/s41467-019-14108-y.

Wang, L., Wang, H., Cao, Z., He, Y., Dong, Z., Wang, S., 2022. Can industrial intellectualization reduce carbon emissions? — empirical evidence from the perspective of carbon total factor productivity in China. Technol. Forecast. Soc. Change 184, 121969. https://doi.org/10.1016/j.techfore.2022.121969.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C.P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C.W., Bengio, Y., Zitnik, M., 2023. Scientific discovery in the age of artificial intelligence. Nature 620, 47–60. https://doi.org/10.1038/s41586-023-06221-2.

Wang, Z., Cao, X., Ren, X., Gozgor, G., 2024. Digital finance and the energy transition: evidence from Chinese prefecture-level cities. Glob. Financ. J. 61, 100987. https://doi.org/10.1016/j.gfj.2024.100987.

Xie, X., Han, Y., Tan, H., 2024. Greening China's digital economy: exploring the contribution of the East–West Computing Resources Transmission Project to CO_2 reduction. Humanit. Soc. Sci. Commun. 11, 1–15. https://doi.org/10.1057/s41599-024-02963-0.

Zhang, W., Zeng, M., 2024. Is artificial intelligence a curse or a blessing for enterprise energy intensity? Evidence from China. Energy Econ. 134, 107561. https://doi.org/10.1016/j.eneco.2024.107561.

Zhang, H., Zhou, P., Sun, X., Ni, G., 2024. Disparities in energy efficiency and its determinants in Chinese cities: from the perspective of heterogeneity. Energy 289, 129959. https://doi.org/10.1016/j.energy.2023.129959t.

Zhao, P., Gao, Y., Sun, X., 2022. How does artificial intelligence affect green economic growth?—evidence from China. Sci. Total Environ. 834, 155306. https://doi.org/10.1016/j.scitotenv.2022.155306.

Zhao, Q., Wang, L., Stan, S.-E., Mirza, N., 2024. Can artificial intelligence help accelerate the transition to renewable energy? Energy Econ. 134, 107584. https://doi.org/10.1016/j.eneco.2024.107584.

Zhong, Q., Zhang, Q., Yang, J., 2025. Can artificial intelligence empower energy enterprises to cope with climate policy uncertainty? Energy Econ. 141, 108088. https://doi.org/10.1016/j.eneco.2024.108088.

Zhou, X., Zhou, D., Wang, Q., Su, B., 2019. How information and communication technology drives carbon emissions: a sector-level analysis for China. Energy Econ. 81, 380–392. https://doi.org/10.1016/j.eneco.2019.04.014.

Zhuo, Z., Du, E., Zhang, N., Nielsen, C.P., Lu, X., Xiao, J., Wu, J., Kang, C., 2022. Cost increase in the electricity supply to achieve carbon neutrality in China. Nat. Commun. 13, 3172. https://doi.org/10.1038/s41467-022-30747-0.