Environmental Impact of High-Performance Computing: A Comprehensive Analysis of Energy Consumption, Carbon Emissions, and Mitigation Pathways

Chubo Wang *, Jiuyan Lyu, Qingyue Zhao and Jiale Wu

WLSA Shanghai Academy, Shanghai, China

* Corresponding Author Email: chubowang127@gmail.com

Abstract. High-Performance Computing (HPC) is increasingly pivotal for scientific discovery, technolog- ical innovation, and economic competitiveness. However, its rapidly escalating energy consumption and consequent carbon footprint present profound environmental challenges that demand rigorous assessment and proactive mitigation. This study develops and applies an integrated modeling framework to compre- hensively estimate and project the multifaceted environmental impact of global HPC operations. Employing Autoregressive Integrated Moving Average (ARIMA) models, meticulously selected based on the Corrected Akaike Information Criterion (AICc), we forecast future energy consumption trajectories for distinct data center archetypes: hyperscale, cloud, and legacy facilities. The intricate relationships between HPC energy utilization, key macroeconomic indicators (such as Gross Domestic Product, population growth, and rates of technological progress), and resultant CO_2 emissions are systematically analyzed, with causal linkages explored through Granger causality tests. Our projections indicate a persistent upward trend in global HPC energy consumption, although the growing integration of renewable and nuclear energy sources into elec- tricity grids offers a partial, yet insufficient, mitigating effect. The research further extends the modeling to quantitatively assess the potential for carbon emission reductions attributable to accelerated renewable en- ergy deployment, revealing that while achieving a 100% renewable energy supply for HPC remains a distant, multi-decadal prospect, substantial and impactful emission reductions are attainable in the near to medium term through concerted policy and investment. The analysis also critically examines the substantial water footprint of HPC, particularly for cooling large-scale data centers, thereby highlighting significant resource allocation and sustainability concerns, especially in water-stressed regions. Based on these comprehensive empirical findings, a suite of actionable technical and policy recommendations is formulated to guide the transition towards more sustainable HPC development. These include advancements in hardware energy effi- ciency, strategic promotion of renewable energy procurement, enhanced software and workload optimization, and fostering international cooperation on green computing standards. This research underscores the exi- gent need for a holistic, system-level approach to managing the ecological footprint of HPC, ensuring that its profound benefits to society are realized in a manner consistent with global environmental sustainability imperatives and climate change mitigation goals.

Keywords: High-Performance Computing; Environmental Impact Assessment; Energy Consumption Model- ing; Carbon Emission Projections; ARIMA Time Series Analysis; Renewable Energy Transition; Sustainable Computing; Water Footprint.

1. Introduction

High-Performance Computing (HPC) has emerged as an indispensable engine of scientific advancement, industrial inno- vation, and economic development across the globe [2]. Its capacity to perform complex calculations and process vast datasets at extraordinary speeds underpins progress in diverse domains, including fundamental physics, drug discovery, climate science, financial modeling, and the burgeoning field of artificial intelligence. The computational power of con- temporary HPC systems, often exceeding conventional computing capabilities by orders of magnitude, is realized through sophisticated architectures incorporating millions of processor cores and specialized accelerators, housed within large, energy-intensive data center facilities. However, this remarkable computational prowess is inextricably linked to a sub- stantial and rapidly growing

environmental footprint, primarily characterized by significant energy consumption and the resultant greenhouse gas emissions [4]. As the global demand for HPC resources continues its exponential surge—driven by the escalating needs of big data analytics, advanced machine learning models, and increasingly complex scientific simulations—the environmental sustainability of HPC has become a paramount concern for international policymakers, industry leaders, the research community, and civil society [1]. Addressing this challenge is crucial not only for miti- gating climate change but also for ensuring the long-term viability and societal acceptance of HPC as a transformative technology.

The environmental burden attributable to HPC operations extends considerably beyond direct energy use and its associ- ated carbon footprint. The intensive cooling requirements of densely packed electronic components necessitate significant water consumption, potentially straining local freshwater resources, particularly in arid or semi-arid regions where many data centers are sited [6]. Furthermore, the relatively short operational lifecycles of HPC hardware, driven by rapid tech- nological obsolescence, contribute to a growing stream of electronic waste (e-waste), which presents complex challenges related to hazardous material management and resource recovery [3]. The manufacturing processes for specialized HPC components, including semiconductors and interconnects, are themselves resource-intensive, often relying on the extrac- tion of rare earth elements and involving significant embodied energy. While ongoing technological innovations, such as the development of more energy-efficient processor architectures (e.g., ARM-based systems, GPUs, TPUs), advanced liquid cooling techniques, and sophisticated energy management software, offer promising avenues for enhancing opera- tional efficiency, the sheer scale and relentless growth rate of global HPC deployment frequently threaten to outpace these incremental gains. Consequently, a more holistic and systemic approach to HPC sustainability is urgently required.

This paper confronts the pressing need for a comprehensive, empirically grounded understanding and robust quantitative assessment of the multifaceted environmental impact of HPC. While existing research has often illuminated specific aspects of this issue—such as the Power Usage Effectiveness (PUE) of individual data centers, the carbon intensity of regional electricity grids, or the energy efficiency of particular algorithms—a fully integrated, global perspective that systematically projects future impacts and evaluates the efficacy of various mitigation strategies remains comparatively underdeveloped. The primary research objectives of this study are therefore articulated as follows:

1. To meticulously estimate the current global energy consumption attributable to HPC systems and to project its future trajectory under a range of plausible operational scenarios and across different data center archetypes (traditional, cloud, and hyperscale). This involves developing robust forecasting models that account for historical trends and anticipated growth drivers.

2. To quantify the associated carbon footprint of global HPC operations, considering regional variations in energy mixes, and to analyze the potential role of transitioning to lower-carbon energy sources, particularly renewable energy, in mitigating these emissions. This includes assessing the feasibility and timeline of such a transition.

3. To develop, articulate, and quantitatively evaluate a coherent set of actionable technical and policy recommendations designed to foster more environmentally sustainable HPC development pathways. This involves a critical appraisal of both supply-side (energy generation, hardware efficiency) and demand-side (workload management, algorithmic optimization) interventions.

By synergistically employing advanced time series analysis techniques, econometric modeling approaches, and scenario- based projections, this research endeavors to provide a scientifically rigorous and policy-relevant framework for compre- hensively assessing the environmental consequences of HPC. The insights and findings generated are intended to inform evidence-based policy decisions at national and international levels, guide the formulation of industry best practices and sustainability standards, and stimulate further interdisciplinary research into the critical domain of green computing tech- nologies and sustainable digital infrastructures.

2. Theoretical Framework and Definitions

2.1. Key Definitions

A clear definitional basis is essential for the quantitative analysis undertaken in this study. The following key terms are central to our framework:

2.1.1. Energy Consumption (E)

This refers to the total electrical energy consumed by the entirety of an HPC system, encompassing not only the IT equipment (servers, storage, networking) but also all supporting infrastructure, including cooling systems (chillers, CRAC units, pumps), power distribution units (PDUs), uninterruptible power supplies (UPS), and lighting. Energy consumption is typically measured in kilowatt-hours (kWh) or, for larger scales, terawatt-hours (TWh) on an annual basis. It is a primary determinant of operational cost and environmental impact.

2.1.2. Carbon Dioxide (CO2) Emissions (C).

These are the greenhouse gas emissions, primarily carbon dioxide, re- sulting directly or indirectly from the energy consumed by HPC operations. Direct emissions are rare (e.g., from on-site backup generators), so the focus is on indirect emissions (Scope 2) associated with purchased electricity. Emissions are calculated as the product of energy consumed and the carbon intensity of the specific energy sources utilized in the elec-tricity generation mix: $C = E \times CI_{mix}$, where CI_{mix} represents the weighted average carbon intensity of the electricity supplied to the HPC facility, expressed in units such as kg CO_2 per kWh or MWh.

2.1.3. Power Usage Effectiveness (PUE).

PUE is the industry-standard metric for characterizing the energy efficiency of a data center's infrastructure. It is formally defined as the ratio of the total energy consumed by the entire data center facility ($E_{facility}$) to the energy delivered specifically to the IT equipment (E_{IT}):

$$\mathrm{PUE} = \frac{E_{\mathrm{facility}}}{E_{\mathrm{IT}}} \tag{1}$$

A PUE value of 1.0 would signify perfect efficiency, where all energy entering the facility is used by the IT equipment, with no losses or consumption by supporting infrastructure. In practice, PUE values are always greater than 1.0, with lower values indicating higher infrastructure efficiency.

2.1.4. Autoregressive Integrated Moving Average (ARIMA) Model.

ARIMA models constitute a widely utilized class of statistical models for analyzing and forecasting univariate time series data that exhibit temporal dependencies and non- stationarity. An ARIMA(p,d,q) model structure is characterized by three key parameters: p, the order of the autoregressive (AR) component, which specifies the number of lagged observations included in the model; d, the degree of differencing required to make the time series stationary; and q, the order of the moving average (MA) component, which indicates the number of lagged forecast errors in the prediction equation. The general mathematical representation of an ARIMA model, often expressed using the backshift operator B (where $BY_t = Y_{t-1}$), is:

$$\Phi(\mathrm{B})(1 - \mathrm{B})\mathrm{d}\mathrm{Yt} = \mathrm{c} + \Theta(\mathrm{B})\epsilon\mathrm{t} \tag{2}$$

Here, $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the autoregressive polynomial in B, $\Theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ is the moving average polynomial in B, Y_t is the observed value of the time series at time t, c is a constant term (drift), and ϵ_t is a white noise error term, assumed to be independently and identically distributed with a mean of zero and constant variance ($\epsilon_t \sim WN(0, \sigma^2)$).

2.2. Model Assumptions

The modeling framework and subsequent analyses presented in this paper are predicated upon several key assumptions, the validity and implications of which are critical to interpreting the results:

(1) Continuation of Historical Trends for Energy Demand: It is assumed that, in the near to medium term, the aggre- gate energy demand for HPC services will continue to evolve along trajectories informed by historical growth pat- terns. This assumption implies that, absent truly disruptive technological breakthroughs or radical shifts in demand drivers (e.g., unforeseen applications or widespread adoption of fundamentally different computing paradigms), past trends provide a reasonable basis for baseline projections. This allows for the establishment of a "business-as- usual" or reference scenario against which mitigation efforts can be assessed.

(2) Short-Term Stability and Gradual Long-Term Evolution of Energy Mixes: The composition of regional and national electricity generation portfolios—specifically, the proportional contributions of fossil fuels (coal, natural gas, oil), renewable energy sources (solar, wind, hydro, geothermal), and nuclear power—is assumed to exhibit rel- ative stability in the immediate short term. However, for longer-term projections, the models incorporate scenarios that reflect gradual, policy-driven, or market-induced shifts towards decarbonization, acknowledging the dynamic nature of energy systems.

(3) Incremental Improvement in Power Usage Effectiveness (PUE): While ongoing efforts to enhance data center infrastructure efficiency are acknowledged, PUE values are projected to improve at a slow and incremental pace. This assumption reflects the fact that many modern facilities have already achieved relatively low PUEs, and further substantial reductions become increasingly challenging and costly as they approach the theoretical limit of 1.0. Therefore, transformative leaps in PUE are not assumed in the baseline scenarios.

The stationarity of time series data, a prerequisite for ARIMA modeling, is rigorously assessed using statistical tests such as the Augmented Dickey-Fuller (ADF) test. The ADF test evaluates the null hypothesis (H_0) that a unit root exists in the time series (implying non-stationarity) against the alternative hypothesis (H_1) that the series is stationary (or trend- stationary). Non-stationary series are subjected to differencing until stationarity is achieved.

3. Methodology

3.1. Data Sources and Scope of Analysis

The empirical foundation of this research rests upon a comprehensive compilation of data pertaining to global HPC energy consumption, detailed characteristics of various data center archetypes (including traditional on-premise facilities, enterprise cloud deployments, and large-scale hyperscale centers), historical and projected PUE trends, and regional as well as national electricity generation mixes. These data were meticulously gathered from a diverse array of authoritative sources, including peer-reviewed academic literature, technical reports from leading industry organizations (such as the Uptime Institute, ASHRAE, and the International Energy Agency - IEA), publications from governmental bodies (notably the U.S. Energy Information Administration - EIA for energy statistics and carbon intensities), and proprietary datasets where available. The geographical scope of the analysis is global, aiming to capture the worldwide footprint of HPC. However, regional disaggregation is performed where sufficient data granularity permits, allowing for a more nuanced understanding of geographical variations in energy consumption patterns and carbon intensities. The temporal scope primarily covers the period from 2015, allowing for the observation of recent trends, with projections extending up to 2030 to provide medium-term insights relevant for policy and strategic planning.

3.2. Time Series Forecasting of Energy Consumption

To project future HPC energy consumption, Autoregressive Integrated Moving Average (ARIMA) models were systemat- ically developed and applied to historical time series data for each identified data center category. The robust Box-Jenkins methodology, a widely accepted iterative approach for time series model building, was rigorously followed. This method- ology encompasses three principal stages:

(1) Model Identification: This initial stage involves a thorough examination of the statistical properties of the his- torical energy consumption time series (Y_t). Key tools include the visual inspection of time series plots, and the analysis of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) correlograms. These diagnostics help to infer the appropriate orders (p, d, q) for the ARIMA model. The stationarity of the series, a criti- cal assumption for ARIMA modeling, was assessed using formal statistical tests like the Augmented Dickey-Fuller (ADF) test. If non-stationarity was detected (i.e., presence of a unit root), the series was subjected to an appropriate degree of differencing (d) until stationarity was achieved.

(2) Parameter Estimation: Once a tentative ARIMA(p,d,q) model structure was identified, the model parameters—namely the autoregressive coefficients (ϕ_i), moving average coefficients (θ_j), and any constant term—were estimated from

the data. This estimation was typically performed using robust numerical optimization techniques, such as the max- imum likelihood estimation (MLE) method, which seeks to find parameter values that maximize the probability of observing the given data.

(3) Diagnostic Checking: After parameter estimation, the adequacy of the fitted model was critically evaluated. This involved a comprehensive analysis of the model residuals (the differences between observed values and model predictions). Ideally, the residuals should behave like a white noise process—i.e., be uncorrelated with zero mean and constant variance. Statistical tests, such as the Ljung-Box Q-statistic, were employed to test for any remaining autocorrelation in the residuals. If the model was found to be inadequate, the identification and estimation steps were revisited in an iterative fashion.

The selection among several plausible candidate ARIMA specifications was guided by information criteria that balance model fit (likelihood) against model complexity (number of parameters). Specifically, the Corrected Akaike Information Criterion (AICc) was utilized:

$$C_{red,t} = E_{\text{total},t} \times \left(\text{Share}_{RE,final,t} - \text{Share}_{RE,initial,t}\right) \times CI_{\text{displaced},t} \quad (3)$$

Where $\hat{L}$ denotes the maximized value of the likelihood function for the model, k is the total number of estimated param- eters in the model (including the variance of the error term), and n is the effective sample size. Models with lower AICc values are preferred.

3.3. Carbon Emission Estimation and Projection

The total carbon emissions ($C_{total,t}$) originating from HPC energy consumption at time t were systematically estimated by integrating the forecasted energy consumption with data on the carbon intensity of electricity generation. The following formulation was applied:

$$C_{\text{total},t} = \sum_j E_{j,t} \sum_i \left(a_{i,j,t} \cdot CI_i\right) \quad (4)$$

In this equation, $E_{j,t}$ represents the projected electrical energy consumed by data center type j (hyperscale, cloud, tradi- tional) at time t. The term $a_{i,j,t}$ signifies the share or proportion of energy source i (e.g., coal, natural gas, nuclear, solar, wind) in the specific electricity mix supplying data center type j at time t. This acknowledges that different data centers may draw power from grids with varying generation portfolios. Finally, CI_i is the specific carbon intensity factor for energy source i, typically expressed in units such as kilograms of CO_2 equivalent per Megawatt-hour (kg CO_2e/MWh) or per British Thermal Unit (BTU). These carbon intensity factors were carefully sourced from

authoritative databases like those provided by the U.S. Energy Information Administration (EIA) and reports from the Intergovernmental Panel on Climate Change (IPCC), ensuring consistency and comparability. Projections of future carbon emissions inherently depend on scenarios for both future energy consumption ($E_{j,t}$) and future energy mix evolution ($a_{i,j,t}$).

3.4. Analysis of Broader Environmental and Resource Impacts

Beyond direct energy consumption and carbon emissions, the study sought to understand the wider environmental and resource implications of HPC growth. The impact of burgeoning HPC infrastructure on broader community-level energy demand across various sectors (residential, commercial, industrial, transportation, and overall electric power systems) was investigated using multiple linear regression models. These models took the general form:

$$Y_{s,t} = \beta_{0,s} + \beta_{1,s}X_t + \sum_k \gamma_{k,s}Z_{k,t} + \epsilon_{s,t} \quad (5)$$

Here, $Y_{s,t}$ is the energy demand of a specific sector s at time t. X_t typically represents a time trend (e.g., year) to cap- ture secular changes. $Z_{k,t}$ are other relevant explanatory variables, which could include measures of local HPC capacity growth, regional Gross Domestic Product (GDP), population changes, or specific policy variables. The coefficients $\beta_{0,s}$, $\beta_{1,s}$, and $\gamma_{k,s}$ is estimated econometrically, and $\epsilon_{s,t}$ is the error term. Furthermore, the significant water usage (W_t) associated with data center cooling, particularly for large hyperscale facilities and major technology companies, was ex- plicitly modeled. ARIMA techniques, similar to those used for energy consumption forecasting, were applied to historical water usage data to project future demand and thereby assess potential implications for local water resources and regional water stress.

3.5. Modeling the Impact of Renewable Energy Penetration

A key focus of the research was to quantify the potential of increased renewable energy (RE) penetration in mitigating HPC-related carbon emissions. The reduction in carbon emissions ($C_{red,t}$) at time t resulting from a shift towards RE was modeled conceptually as:

$$\text{Cred,t} = \text{Etotal,t} \times (\text{ShareRE,final,t} - \text{ShareRE,initial,t}) \times \text{CIdisplaced,t} \quad (6)$$

In this formulation, $\text{Share}_{RE,initial,t}$ and $\text{Share}_{RE,final,t}$ represents the initial (baseline) and final (scenario-based) shares of renewable energy in the relevant electricity generation mix at time t. $E_{\text{total},t}$ is the total electricity consumed (e.g., by HPC). $CI_{\text{displaced},t}$ is the crucial factor representing the average carbon intensity of the conventional energy sources (typically fossil fuels) that are displaced by the increased generation from renewables. Future renewable energy shares were projected using two approaches: extrapolation of historical trends, often fitted using polynomial (e.g., quadratic) regression models of the form $S_{RE,t} = \alpha + \beta_1 t + \beta_2 t^2 + \nu_t$; and scenario-based analyses assuming specific annual growth rates (g_{RE}) for renewable energy capacity or generation. These scenarios allowed for an exploration of the sensitivity of emission reductions to different paces of RE deployment.

4. Results

4.1. HPC Energy Consumption: Current Scope and Projected Trajectories

The initial assessment of HPC energy consumption involved establishing the potential scale based on maximum power capacities of different data center archetypes. Traditional data centers, with typical capacities around 2 MW, could the- oretically consume up to 17.5 TWh annually if operated at full capacity year-round. Cloud facilities, often in the 50 MW range, have a corresponding maximum annual consumption of up to 438 TWh, while hyperscale data centers, with capacities frequently reaching or exceeding 200 MW, could theoretically demand up to 876 TWh per year. These

figures, however, represent an extreme upper bound. A more realistic estimation, applying a conservative average utilization rate of 12% (derived from industry observations), reduces these potential annual consumptions to approximately 2.1 TWh for traditional, 52.6 TWh for cloud, and 105.1 TWh for hyperscale facilities, respectively. These adjusted figures provide a more grounded understanding of the operational energy demands.

The ARIMA model forecasts, developed from historical energy consumption data spanning 2015-2021, revealed dis- tinct and informative future trajectories for these data center categories, extending to 2030. The key findings were:

Hyperscale Data Centers: The optimal model identified was an ARIMA (1,1,0), characterized by the equation $\Delta Y_t = 0.9665 Y_{t-1} + \epsilon_t$. This model structure, coupled with the positive coefficient, indicated a statistically significant and persistent increasing trend in energy consumption for this category. Hyperscale facilities are thus projected to be the primary drivers of future growth in HPC energy demand.

Cloud (Non-Hyperscale) Data Centers: For this segment, an ARIMA (0,1,0) model, essentially a random walk with drift ($\Delta Y_t = c + \epsilon_t$), was found to be most appropriate. This suggests that the energy consumption of non- hyperscale cloud facilities is projected to remain relatively stable or exhibit only modest growth in the forecast period, without a strong underlying autoregressive trend.

Traditional Data Centers: Similar to hyperscale, an ARIMA (1,1,0) model ($\Delta Y_t = 0.9860 Y_{t-1} + \epsilon_t$) provided the best fit. However, the context of declining market share for older, less efficient traditional facilities suggests this model captures a managed decline or consolidation phase, where the aggregate energy consumption of this category
is expected to decrease over time.

A summary of the key parameter estimates and model fit statistics for these ARIMA models is presented in Table 1. These projections collectively point towards a significant structural transformation within the HPC landscape, characterized by a consolidation of computational workloads into larger, more centralized, and often more energy-efficient (on a per- computation basis) hyperscale facilities. Nevertheless, the absolute growth in energy demand from the hyperscale sector is a major concern.

Table 1. Summary of Python ARIMA Model Output Parameters for Energy Consumption by Data Center Type.

Parameter	Hyperscale	Cloud (non-hyperscale)	Traditional
Optimal ARIMA Order	(1,1,0)	(0,1,0)	(1,1,0)
AR (1) Coefficient (approx.)	0.9665	Not Applicable	0.9860
Error Variance (σ^2)	6.6920	19.7109	3.6517
AIC Value	35.153	36.914	32.381

4.2. Power Usage Effectiveness (PUE) Trends and Implications

The global average Power Usage Effectiveness (PUE) for data centers has exhibited a consistent, albeit gradual, downward trend over the period from 2007 to 2023. Linear regression analysis of historical PUE data ($PUE_t = \hat{\alpha} + \hat{\beta} t$) yielded a statistically significant negative coefficient for the time trend ($\hat{\beta} < 0$), confirming ongoing improvements in data center infrastructure efficiency. This positive development is attributable to a combination of factors, including advancements in cooling technologies, more efficient power distribution systems, optimized facility design, and the increasing market share of newer, large-scale hyperscale data centers, which generally achieve lower (better) PUE values than older, smaller facilities. This trend of improving PUE is expected to continue, which will help to somewhat mitigate the growth in total facility energy consumption relative to the growth in IT equipment energy demand. However, the rate of PUE improvement appears to be slowing as the industry approaches practical limits of efficiency for conventional air-cooled facilities, suggesting that further substantial gains may require more transformative technological shifts.

4.3. Projected Carbon Emissions from Global HPC Operations

The projected carbon emissions from the global HPC sector were derived by combining the forecasted energy consumption for each data center type with scenarios for the evolving carbon intensity of electricity generation mixes in key regions. Utilizing fuel-specific carbon intensity data from the EIA (e.g., coal at approximately 95.55 kg CO_2/MBtu; diesel at 74.14 kg CO_2/MBtu; natural gas at 52.91 kg CO_2/MBtu), the analysis indicates that, despite efficiency improvements, the total carbon emissions attributable to HPC are projected to continue rising globally through 2030 under baseline energy mix scenarios. The magnitude of this increase is highly sensitive to the pace of decarbonization of electricity grids. Regions that achieve faster transitions away from coal and towards natural gas and, more significantly, renewable energy sources and nuclear power, will see a correspondingly slower growth, or even a potential stabilization or reduction, in HPC-related emissions. Conversely, HPC expansion in regions heavily reliant on carbon-intensive electricity will exacerbate global emissions.

4.4. Broader Energy System Interactions and Resource Impacts

The deployment and operation of large-scale HPC infrastructure were found to have discernible impacts on broader energy systems and resource availability at the community and regional levels. Multiple linear regression models analyzing sec- toral energy demand indicated that the growth of HPC facilities, particularly hyperscale centers, is significantly correlated with an increase in overall electric power demand within the host communities or regions. Projections to 2030 suggest that this could place considerable strain on local electricity generation and transmission infrastructure if not carefully planned. The impacts on direct energy demand in other sectors (residential, commercial, industrial) were more varied and less direct, likely mediated through economic effects. A particularly salient finding relates to the substantial water usage associated with cooling data centers. ARIMA modeling of water consumption for representative large technology companies indicated that future water demand for HPC cooling is projected to reach extremely high levels. For instance, a single major global technology entity's data center operations might consume nearly 500 billion liters of water annually by 2030. Such significant water withdrawals raise profound concerns about potential conflicts with other water users (agricultural, municipal, ecological) and could exacerbate water stress in already arid or water-scarce regions where many data centers are increasingly being located. This underscores the critical need for water-efficient cooling technologies and integrated water resource management in the context of HPC development.

4.5. The Role and Potential of Renewable Energy in Emission Mitigation

The quantitative modeling of increased renewable energy (RE) penetration in electricity generation mixes (as per Equation 6) demonstrated its substantial potential for mitigating carbon emissions from the HPC sector. Illustratively, for the United States, a hypothetical increase in the RE share of the electricity mix from a baseline of 8.79% to 15%, assuming this RE displaces electricity generated from average U.S. fossil fuel sources, could result in an annual CO_2 emission reduction of approximately 8.35×10^{10} kilograms. This highlights the direct and impactful leverage of decarbonizing the electricity supply. However, the analysis also tempered expectations regarding the timeline for a full transition. A quadratic regres-sion model fitted to historical global renewable energy penetration trends ($S_{RE,t} = 0.0071t^2 - 28.382t + 28342$, where t is the year) suggested that achieving a 100% renewable energy supply globally is a very long-term endeavor, potentially not realized until well into the 22nd century or even later under current trajectories. This underscores the multi-decadal nature of the energy transition challenge. Scenario analyses exploring accelerated annual growth rates for renewable en- ergy (e.g., sustained growth of 5%, 10%, or 15% per year in RE generation) clearly demonstrated that more aggressive deployment policies lead to progressively steeper and earlier declines in projected CO_2 emissions from the energy sector, thereby emphasizing the critical importance of proactive policy drivers and substantial investments in accelerating the

renewable energy transition to make a meaningful impact on HPC's carbon footprint in the coming decades.

5. Discussion

The empirical findings of this comprehensive study paint a nuanced picture of the environmental impact of High-Performance Computing. A critical dichotomy emerges: HPC is undeniably an indispensable catalyst for scientific progress, techno- logical innovation, and economic competitiveness; yet, its operational demands, particularly its escalating energy con- sumption and associated carbon emissions, constitute a substantial and growing environmental burden that cannot be ignored. The projected structural shift within the data center industry towards larger, more centralized hyperscale facili- ties is a complex development. While these advanced facilities often boast superior energy efficiency (lower PUEs) and greater economies of scale in resource management compared to older, smaller traditional data centers, their immense individual power requirements mean they will increasingly concentrate energy demand in specific geographical locations. This concentration poses significant challenges for local electricity grid stability, capacity planning, and the sustainable management of local natural resources such as water.

The observed global trend of gradual improvement in Power Usage Effectiveness (PUE) is a welcome development, reflecting concerted industry efforts and technological advancements in data center design and operation. However, the impact of these PUE improvements on overall energy consumption is often partially, if not entirely, offset by the relentless Jevons paradox-like growth in computational demand and data processing volumes. Therefore, a narrow focus on PUE as the sole metric of data center sustainability is insufficient. A more holistic and effective strategy must necessarily encompass a broader array of interventions, including the aggressive decarbonization of energy sources supplying HPC facilities, radical improvements in the energy efficiency of IT hardware at the component and system levels (performance per watt), and sophisticated software and workload optimization techniques to minimize energy waste during computation and idle periods.

The analysis of projected carbon emissions robustly underscores the paramount importance of the electricity generation mix. Regions and countries that continue to rely heavily on fossil fuels, particularly coal, for their electricity supply will find that their expanding HPC sectors contribute disproportionately and increasingly to global greenhouse gas emissions. Conversely, the transition to renewable energy sources (solar, wind, geothermal, etc.) and other low-carbon alternatives like nuclear power is identified as the most critical lever for achieving long-term environmental sustainability in the HPC domain. Nevertheless, our projections soberly indicate that this energy transition will be a protracted and challenging process, likely spanning many decades, unless it is substantially accelerated by decisive policy interventions, significant technological breakthroughs (particularly in energy storage and grid modernization), and massive global investments. The scenario analyses conducted in this study clearly demonstrate that more ambitious renewable energy deployment targets and faster growth rates directly translate into more significant and rapid reductions in CO_2 emissions from the HPC sector, thereby highlighting the urgency of action.

The substantial water footprint associated with cooling large-scale data centers, as starkly illustrated by the projections for major technology companies, raises serious and pressing concerns regarding resource equity, local ecological impacts, and the potential for exacerbating water stress in vulnerable regions. This finding necessitates a paradigm shift towards water-stewardship in data center design and operation, prioritizing the development and deployment of innovative, water-efficient cooling technologies (such as advanced direct liquid cooling, closed-loop systems, or water-free cooling solutions like free air cooling in suitable climates) and mandating careful hydrogeological assessments and integrated water resource management plans as part of the site selection and permitting process for new HPC facilities.

The methodological reliance of this study on ARIMA models for time series forecasting provides a robust framework for short-to-medium-term projections, particularly when historical trends are

relatively stable and well-defined. However, it is important to acknowledge the inherent limitations of such models. They are primarily extrapolative and may be less adept at capturing the impact of sudden structural breaks, unforeseen disruptive technological innovations, or radical policy shifts that could fundamentally alter future trajectories. Future research endeavors could beneficially complement this work by incorporating more dynamic and behaviorally rich modeling techniques, such as system dynamics modeling or agent-based models, which are better suited to exploring complex feedback loops, policy sensitivities, and the emergent behavior of socio-technical systems. Furthermore, while this study aimed for a global scope, the analyses are inevitably constrained by data availability and granularity, which may mask significant regional and national variations. More localized and context-specific studies are therefore warranted to provide tailored insights for specific geographies.

6. Policy Recommendations and Future Research

Based on the comprehensive quantitative analysis and the ensuing discussion of its implications, a coherent set of tech- nical and policy recommendations emerges. These recommendations are designed to guide stakeholders in fostering a more environmentally sustainable High-Performance Computing ecosystem, balancing the imperative for technological advancement with the critical need for ecological stewardship.

6.1. Technical Recommendations for Enhanced Sustainability

6.1.1. Advancements in Hardware and System Energy Efficiency.

Sustained and intensified investment in research and development (R&D) is crucial for achieving breakthrough improvements in the energy efficiency of all HPC hard- ware components, including processors (CPUs, GPUs, AI accelerators), memory systems, high-speed interconnects, and storage solutions. The overarching goal should be to significantly enhance performance per watt. This necessitates ex- ploring novel semiconductor materials, innovative chip architectures (e.g., neuromorphic, quantum-inspired), advanced packaging technologies, and system-level co-design approaches. Furthermore, radical innovations in data center cooling technologies—such as widespread adoption of direct-to-chip liquid cooling, two-phase immersion cooling, or intelligent free cooling systems optimized for specific climates—are essential to drastically reduce the energy consumed by cooling infrastructure, which often constitutes a substantial portion of the non-IT facility power budget.

6.1.2. Optimization of Software, Workloads, and Resource Management.

The development and widespread deploy- ment of sophisticated, energy-aware software tools are paramount. This includes intelligent job schedulers that can optimize task placement and execution across heterogeneous resources to minimize overall energy consumption and re- duce idle power states. Advanced resource management systems should dynamically match computational resources to fluctuating workload demands, leveraging techniques like fine-grained Dynamic Voltage and Frequency Scaling (DVFS), predictive workload consolidation, and power-capping strategies. Furthermore, promoting the development and use of energy-efficient programming languages, compilers, and runtime systems can contribute significantly to reducing the energy footprint of HPC applications.

6.1.3. Promotion of Algorithmic and Application-Level Efficiency.

A focus on algorithmic efficiency can yield sub- stantial energy savings, often with minimal or no performance degradation. Research and educational initiatives should promote the design and adoption of algorithms with lower intrinsic computational complexity, reduced data movement requirements (as data transfer is often a major energy bottleneck), and optimized numerical precision where full preci- sion is not strictly necessary. Application-specific co-design, where algorithms,

software, and hardware are developed in concert for specific problem domains, can unlock further energy efficiency gains.

6.2. Policy Recommendations for a Sustainable HPC Framework

6.2.1. Strategic Promotion and Integration of Renewable Energy

Governments and regulatory bodies must play a proactive role in accelerating the transition of HPC facilities to renewable energy sources. This can be achieved through a portfolio of policy instruments, including targeted financial incentives (e.g., tax credits, grants, feed-in tariffs for on- site renewable generation), the establishment of green energy tariffs by utilities, streamlining permitting processes for corporate Power Purchase Agreements (PPAs) for off-site renewables, and implementing robust carbon pricing mecha- nisms (e.g., carbon taxes or emissions trading schemes) that make fossil fuel-based electricity less economically attractive. Mandating minimum renewable energy procurement levels for large data centers could also be considered.

6.2.2. Dedicated Support for Green Technology Innovation and Deployment

Public funding agencies and private sector investors should significantly increase support for R&D initiatives focused on developing and commercializing next-generation low-energy HPC hardware, sustainable manufacturing processes for IT components (reducing embodied energy and hazardous materials), and circular economy models for the entire lifecycle of IT equipment. This includes promoting design for disassembly, facilitating component reuse and remanufacturing, and investing in advanced e-waste recycling infrastructure and technologies.

6.2.3. Enhancement of Standardization, Reporting, and Transparency

The development and enforcement of rigorous international standards for energy efficiency reporting (e.g., standardized PUE measurement protocols, metrics for IT equipment efficiency like SERT) and comprehensive carbon footprint accounting for HPC facilities are urgently needed. Mandatory public disclosure of key environmental performance indicators—including total energy consumption, PUE, water usage effectiveness (WUE), and Scope 1, 2, and 3 carbon emissions—would significantly increase transparency, enable benchmarking, and hold operators accountable for their environmental performance.

6.2.4. Integration of HPC Needs into Energy and Resource Planning

Regional and national energy infrastructure planning processes, as well as water resource management strategies, must explicitly consider and proactively address the substantial and growing demands of large-scale data centers. This requires integrated planning to ensure that the development of new HPC facilities is aligned with grid capacity, renewable energy availability, and sustainable water resource allocation, thereby preventing undue strain on local infrastructure and ecosystems and avoiding conflicts with other essential service users.

6.3. Priorities for Future Research Directions

To build upon the findings of this study and further advance the understanding and management of HPC's environmental impact, future research should prioritize several key areas:

1. Comprehensive Life Cycle Assessment (LCA) Studies: There is a pressing need for more holistic LCAs of entire HPC systems and data centers. Such assessments should meticulously quantify the environmental impacts across all lifecycle stages—from raw material extraction and component manufacturing (including embodied energy and supply chain emissions) through operational energy and water use, to end-of-life management (disposal, recycling, reuse).

2. Development of Advanced Dynamic Modeling Frameworks: Future modeling efforts should move beyond ex- trapolative time series approaches to incorporate more sophisticated techniques like system dynamics modeling or agent-based modeling. These frameworks can better capture complex feedback loops, simulate the dynamic impacts of various policy interventions and technological

innovations, and explore the emergent behavior of the coupled HPC-energy-environment system under uncertainty.

3. Advancing the Circular Economy for HPC Infrastructure: Dedicated research is required to investigate and promote viable pathways for enhancing the circularity of HPC hardware and infrastructure. This includes ex- ploring innovative business models (e.g., hardware-as-a-service with built-in refurbishment), developing advanced techniques for material recovery and component remanufacturing, and addressing the logistical and economic chal- lenges of creating closed-loop supply chains for HPC equipment.

4. In-depth Analysis of Socio-Economic and Ethical Dimensions: A deeper understanding of the broader socio- economic consequences of large-scale HPC development is needed. This includes rigorous analysis of impacts on local employment, regional economic development, skills requirements, and critical ethical considerations related to resource equity, digital divides, and the equitable distribution of both the benefits and environmental burdens of HPC.

In conclusion, mitigating the substantial and growing environmental impact of High-Performance Computing neces- sitates a globally coordinated, multi-stakeholder commitment involving sustained technological innovation, robust and forward-looking policy frameworks, and a fundamental shift towards sustainable practices throughout the entire HPC value chain. The research presented in this paper provides a quantitative and analytical foundation for these crucial ef- forts, underscoring both the urgency of the challenge and the feasibility of steering the future of HPC towards a more environmentally responsible and ultimately sustainable trajectory.

References

[1] Li, Baolin. Making Machine Learning on HPC Systems Cost-Effective and Carbon-Friendly. Diss. Northeastern University Boston, 2024.

[2] Khosravi, Atefeh, Saurabh Kumar Garg, and Rajkumar Buyya. "Energy and carbon-efficient placement of virtual machines in distributed cloud data centers." Euro-Par 2013 Parallel Processing: 19th International Conference, Aachen, Germany, August 26-30, 2013. Proceedings 19. Springer Berlin Heidelberg, 2013.

[3] Mostafaei, Hasan, et al. "Development of sustainable HPC using rubber powder and waste wire: carbon footprint analysis, mechanical and microstructural properties." European Journal of Environmental and Civil Engineering (2024): 1-22.

[4] Rong, Huigui, et al. "Optimizing energy consumption for data centers." Renewable and Sustainable Energy Reviews 58 (2016): 674-691.

[5] Ji, Shixin, et al. "Towards Data-center Level Carbon Modeling and Optimization for Deep Learning Inference." arXiv preprint arXiv:2403.04976 (2024).

[6] Marra, Osvaldo, et al. "Green Computing and power saving in HPC data centers." CMCC Research Paper 121 (2011).

[7] Zhang, M. (2024, June 21). Data Center Power: A Comprehensive Overview of energy. Dgtl Infra. Retrieved from https://dgtlinfra.com/data-center-power/

[8] Zaloumis, C. (2024, October 9). Are your data centers keeping you from sustain- ability. Sustainability. Retrieved fromhttps://www.ibm.com/think/insights/ are-your-data-centers-keeping-you-from-sustainability#

[9] Liew, K. S., Mahnendran, S., & Department of Mathematics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia. (2002). The performance of AICC as an order selection criterion in ARMA Time Series mod- els. In Pertanika J. Sci. & Technol. (Vols. 10–10, Issue 1, pp. 25-33). Retrieved from https://econwpa.ub. uni-muenchen.de/econ-wp/ge/papers/0307/0307003.pdf

[10] Fuqua School of Business. (n.d.). Introduction to ARIMA models. Retrieved from https://people.duke.edu/~rnau/411arim.htm#arima110

[11] Fleitas, A. G. (2024, August 1). What is Data Center PUE? Defining power usage effectiveness. Retrieved from https://www.datacenterknowledge.com/sustainability/ what-is-data-center-pue-defining-power-usage-effectiveness

[12] Tech, E. (2024, May 2). Exploring the differences: 4 Types of Data Centers EdgeUno. EdgeUno. Retrieved from https://edgeuno.com/exploring-the-differences-4-types-of-data-centers/

[13] U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. (n.d.). Retrieved from https: //www.eia.gov/totalenergy/data/browser/index.php?tbl=T01.03#/?f=A

[14] U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. (n.d.-b). Retrieved from https://www.eia.gov/environment/emissions/co2_vol_mass.php

[15] Renewable energy is having a good year, but challenges loom ahead — IEEFA. (n.d.). Retrieved from https:// ieefa.org/resources/renewable-energy-having-good-year-challenges-loom-ahead