Modeling the Environmental Footprint of High-Performance Computing: Energy Consumption, Carbon Emissions, and Mitigation Pathways

Yuran Huang *, Ziyao Peng

Beijing National Day School, Beijing, China

* Corresponding Author Email: 21870876@qq.com

Abstract. High-Performance Computing (HPC) is a critical enabler for advancements in artificial intelligence, data science, and complex scientific simulations, driving innovation across numerous sectors. However, its exponential proliferation and escalating computational demands have led to a substantial increase in global energy consumption and associated carbon emissions, posing significant and growing sustainability challenges. This paper presents a comprehensive analytical framework to model and assess HPC's environmental footprint, with a primary focus on direct energy consumption, resultant carbon emissions, and ancillary impacts such as water usage. We develop and evaluate a suite of predictive models for global HPC energy demand, employing time-series analysis techniques including Autoregressive Integrated Moving Average (ARIMA) for its robustness in handling trended data, and Long Short-Term Memory (LSTM) networks for their capacity to capture complex non-linear patterns. Model efficacy is rigorously evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The intricate relationship between energy consumption, the composition of the electricity generation mix, and consequent carbon emissions is investigated using Generalized Linear Models (GLM), with model selection guided by the Akaike Information Criterion (AIC) to balance model fit and complexity. Furthermore, this study projects the potential impact of increasing renewable energy penetration on emission reduction trajectories and models HPC-related water consumption, a frequently overlooked but critical resource impact. Our findings consistently indicate escalating energy demands for HPC, highlighting the complex interplay of technological efficiency gains, workload growth, and energy sourcing strategies that collectively determine the sector's carbon emissions. The study underscores the pressing necessity for multi-faceted mitigation strategies, encompassing advancements in hardware and software energy efficiency, accelerated adoption of renewable energy sources, innovative cooling technologies, and supportive policy interventions, to foster the sustainable development and deployment of HPC infrastructure globally.

Keywords: High-Performance Computing; Environmental Footprint; Energy Consumption Modeling; Car- bon Emissions Analysis; Time-Series Forecasting; ARIMA; LSTM Networks; Generalized Linear Models; Renewable Energy Transition; Sustainable Computing; Water Usage.

1. Introduction

High-Performance Computing (HPC) has unequivocally emerged as an indispensable technological cornerstone, fueling innovation and discovery across a diverse spectrum of scientific, industrial, and societal domains. Its applications range from accelerating breakthroughs in artificial intelligence (AI) and machine learning, to enabling high-fidelity climate modeling, advancing genomic research, and processing vast datasets in modern big data analytics [1]. The computational prowess offered by contemporary HPC systems allows researchers and engineers to tackle complex, multifaceted problems that were previously computationally intractable, thereby pushing the boundaries of knowledge and technological capability. However, this escalating computational capacity is intrinsically and increasingly linked to substantial, and often rapidly growing, energy consumption. Data centers, which form the physical backbone of HPC infrastructure, are recognized as among the most energy-intensive facilities globally. Their operation contributes significantly to worldwide electricity demand and, consequently, to anthropogenic greenhouse gas (GHG) emissions, particularly when the electricity supply is predominantly derived from fossil fuel-based energy

sources [2]. The exponential surge in demand for HPC resources thus presents a critical environmental paradox: a technology that is pivotal for addressing complex global challenges, including aspects of climate change mitigation and adaptation, itself contributes materially to environmental degradation and resource strain. This paradox necessitates urgent and comprehensive investigation into sustainable HPC practices.

The environmental footprint of HPC is a multi-dimensional issue that extends beyond direct energy consumption and associated carbon emissions. It encompasses significant water usage, primarily for the cooling systems required to maintain optimal operating temperatures for high-density electronic equipment. Furthermore, the rapid obsolescence cycles and continuous upgrading of HPC hardware contribute to a growing stream of electronic waste (e-waste), which poses considerable challenges for responsible disposal and material recovery. The manufacturing of sophisticated HPC com- ponents also involves the extraction and processing of rare earth elements and other finite resources, leading to habitat disruption and further energy expenditure upstream in the supply chain [3]. A holistic approach to addressing these multifaceted impacts is therefore crucial for ensuring the long-term sustainable development, deployment, and operation of HPC technologies on a global scale.

This research aims to quantify and model the key environmental impacts of HPC, with a primary focus on energy consumption and associated carbon emissions. Our objectives are: (1) To develop and evaluate predictive models for future HPC energy consumption at both full capacity and average utilization rates. (2) To analyze the relationship between electricity consumption, the energy generation mix, and resulting carbon emissions. (3) To project the potential of renewable energy sources in mitigating HPC's carbon footprint. (4) To model the water consumption associated with HPC operations. (5) To propose actionable recommendations and policy guidelines for promoting sustainable HPC practices.

This study contributes by providing a multi-model forecasting framework for HPC's environmental impact and by synthesizing these findings into strategic pathways for mitigation. The paper is structured as follows: Section 2 details the methodology, including data sources and the mathematical formulation of the models employed. Section 3 presents the results of our analysis. Section 4 discusses the implications of these findings, limitations of the study, and avenues for future research. Finally, Section 5 concludes with key takeaways and a call for concerted action.

2. Methodology

This study employs a multi-faceted quantitative research design, systematically integrating time-series forecasting techniques, advanced statistical modeling, and comprehensive scenario analysis to thoroughly assess the diverse environmental footprint of High-Performance Computing.

2.1. Data Sources and Preprocessing

Data for this study were compiled from a wide array of publicly accessible and reputable sources. These included detailed industry reports on data center energy consumption trends and operational characteristics, statistical information from lists such as the TOP500 supercomputer sites providing insights into the power demands of leading HPC installations, and extensive databases from national and international energy agencies (e.g., International Energy Agency (IEA), U.S. Energy Information Administration (EIA)). These agencies provided critical data on electricity generation mixes, carbon intensity factors for different energy sources, and overall energy market dynamics. Furthermore, the academic literature was extensively reviewed for studies on HPC Power Usage Effectiveness (PUE) evolution, water usage coefficients for various cooling technologies, and lifecycle assessment data. The historical data collected typically spanned from the early 2000s to the most recent available year (approximately 2022), providing a sufficient temporal range for robust time-series analysis.

Prior to model implementation, data underwent standard preprocessing steps. These included handling missing values through appropriate imputation techniques (e.g., mean, median, or time-series

specific methods like interpolation), outlier detection and treatment, and data transformation where necessary (e.g., logarithmic transformations to stabilize variance). For the time-series models like ARIMA, stationarity tests (e.g., Augmented Dickey-Fuller test) were conducted, and differencing was applied as needed. For neural network models like LSTMs, data were typically normalized or standardized (e.g., to a 0-1 range or zero mean and unit variance) to improve training stability and performance.

2.2. Modeling HPC Energy Consumption

Time-series models were systematically employed to forecast the future trajectory of energy consumption attributable to HPC systems globally. This involved analyzing historical trends and projecting them forward under various modeling assumptions.

2.2.1. Power Usage Effectiveness (PUE).

Power Usage Effectiveness (PUE), a widely adopted industry metric for quantifying data center energy efficiency, is defined as the ratio of total facility energy to IT equipment energy [4]:

$$\text{PUE} = \frac{\text{Total Facility Energy Consumption}}{\text{IT Equipment Energy Consumption}} \tag{1}$$

A PUE of 1.0 represents perfect efficiency (i.e., all power is used by IT equipment). Historical trends in PUE were meticulously analyzed to understand the rate of improvement in data center operational efficiency over time, and these trends were incorporated into overall energy consumption projections.

2.2.2. Time-Series Models.

Let Y_t denote the time series variable of interest (e.g., aggregate HPC energy consumption, average PUE) observed at discrete time *t*. An *Autoregressive (AR)* model of order p, denoted AR(p), assumes that the current value of the series can be expressed as a linear combination of its past *p* values. It is particularly useful for capturing inertia and momentum in a time series. The model is formulated as:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \epsilon_t \tag{2}$$

Where *c* is a constant term, ϕ_i are the autoregressive model parameters, and ϵ_t is a white noise error term. A Moving *Average (MA)* model of order q, denoted MA(q), posits that the current value is dependent on a linear combination of past *q* error terms. This structure is effective for modeling shocks or unexpected events whose effects linger. The MA(q) model is:

$$Y_t = \mu + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{3}$$

Where μ is the mean of the series, θ_j are the moving average parameters. These two components are combined in *Autoregressive Moving Average (ARMA(p,q))* models. For non-stationary time series, which often exhibit trends or seasonality, *Autoregressive Integrated Moving Average (ARIMA(p,d,q))* models are employed. Here, *d* represents the degree of differencing applied to the series to achieve stationarity. *Seasonal* ARIMA *(SARIMA(p,d,q) (P, D, Q)* $_m$*)* models further extend ARIMA to explicitly account for seasonality with period *m*. *Holt-Winters Exponential Smoothing* is another classical method particularly well-suited for forecasting data that exhibit both trend and seasonality, offering additive and multiplicative variants. *Long Short-Term Memory (LSTM) Networks*, a specialized type of Recurrent Neural Network (RNN), were chosen for their proven ability to learn and capture complex long-term dependencies and non-linear patterns often present in energy consumption data, which may be missed by linear statistical models. An LSTM unit's architecture, comprising a cell state and gating mechanisms (forget gate f_t, input gate i_t, output gate o_t), is governed by:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\widetilde{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (6)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad (7)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t \odot tanh(C_t) \quad (9)$$

Where x_t is the input vector at time t, h_t is the hidden state vector (output), C_t is the cell state vector, W terms denote weight matrices, b terms are bias vectors, σ is the logistic sigmoid activation function, and $\odot$ represents element-wise multiplication.

2.2.3. Model Evaluation and Selection.

The predictive performance of the developed time-series models was rigorously assessed using standard error metrics. The *Root Mean Squared Error (RMSE)* provides a measure of the standard deviation of the prediction errors:

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\big(Y_k - \widehat{Y_k}\big)^2} \quad (10)$$

The *Mean Absolute Error (MAE)* measures the average magnitude of the errors:

Where Y_k are the actual observed values and $\hat{Y}k$ are the values predicted by the model over N observations in the test set. For selecting the optimal orders (p,d,q) for ARIMA models, the *Akaike* Information *Criterion (AIC)* was employed, which balances model fit against model complexity to prevent overfitting:

$$\mathrm{AIC} = 2k - 2\ln(L) \quad (11)$$

Where k is the number of estimated parameters in the model and L is the maximized value of the likelihood function for the model. Lower AIC values indicate a preferred model.

2.3. Modeling Carbon Emissions

2.3.1. Carbon Intensity.

The carbon intensity (CI) of electricity generation, representing the amount of CO_2 emitted per unit of electricity generated, is a critical determinant of the emissions footprint of electricity consumption. It is defined as:

$$\frac{\text{Total } CO_2 \text{ emissions from electricity generation at time } t}{\text{Total electricity generated at time } t} \quad (12)$$

Projected carbon emissions specifically from HPC operations ($E_{HPC,\,CO2}$) were then estimated by multiplying the projected HPC energy consumption by the relevant carbon intensity, adjusted for any direct renewable energy sourcing:

$$E_{HPC,CO2,t} = \text{Energy}_{HPC,t} \times \text{GridCI}_t \times \left(1 - \text{RenewableShare}_{HPC,t}\right) + \text{Energy}_{HPC,t} \times \text{DirectRenewableCI}_t \times \text{RenewableShare}_{HPC,t} \quad (13)$$

Where GridCIt is the carbon intensity of the grid electricity consumed, RenewableShareHPC, t is the proportion of HPC energy sourced directly from dedicated renewables (e.g., onsite solar, PPAs), and DirectRenewableCI$_t$ is the (often near- zero operational) carbon intensity of these direct renewable sources.

2.3.2. Generalized Linear Model (GLM).

To gain deeper insights into the macroeconomic factors influencing total carbon emissions from the electricity sector (which, in turn, shapes the $GridCI_t$ relevant to HPC), a Generalized Linear Model (GLM) was employed. GLMs extend ordinary linear regression by allowing for response variables that have error distribution models other than a normal distribution, and by allowing for a link function between the linear predictor and the mean of the response variable. The general form is: $g(E[Y]) = \mathbf{X}\boldsymbol{\beta}$, where Y is the response variable (e.g., national or global carbon emissions from electricity), $g(\cdot)$ is the link function (identity function for a Gaussian response), $E[Y]$ is the expected value of Y, $\mathbf{X}$ is the matrix of predictor variables (such as total electricity consumption, share of fossil fuels in the generation mix, share of renewable energy, lagged consumption values to capture temporal dependencies, and interaction terms), and $\boldsymbol{\beta}$ is the vector of coefficients to be estimated. The specific model used in this study was:

$$Carbon\ Emission_t = \beta_0 + \beta_1 * TotalElec_t + \beta_2 * FossilShare_t + \beta_3 * RenewableShare_t + \beta_4 * \text{LaggedTotalElec}_{t-1} + \beta_5(\text{FossilShare}_t \times \text{RenewableShare}_t) + \epsilon_t \quad (14)$$

Where ϵ_t represents the error term. The statistical significance of the estimated coefficients β_i was assessed using p-values, and model diagnostics were performed to check assumptions. This model helps to disentangle the relative contributions of energy demand versus changes in the carbon intensity of the energy supply.

2.4. Modeling Water Consumption

HPC water consumption, primarily driven by the cooling requirements of data centers (e.g., for chillers, cooling towers, and direct liquid cooling systems), was modeled using time-series techniques analogous to those applied for energy consumption (e.g., ARIMA, Holt-Winters). Historical data on water usage effectiveness (WUE) and total water withdrawal by data centers or representative HPC facilities were utilized where available. Projections considered factors such as HPC energy growth, improvements in cooling technology efficiency, and shifts towards less water-intensive cooling methods (e.g., free air cooling in suitable climates, or advanced liquid cooling).

2.5. Scenario Analysis for Renewable Energy Penetration

To explore the potential of decarbonization strategies, scenario analyses were conducted. These scenarios investigated the impact of varying levels of renewable energy penetration into the general electricity grid mix (affecting $GridCI_t$) and increased direct sourcing of renewable energy by HPC facilities (affecting $RenewableShare_{HPC,t}$) on future carbon emissions attributable to HPC. Scenarios ranged from business-as-usual (BAU) trends to ambitious decarbonization pathwaysaligned with international climate targets (e.g., those inspired by IPCC reports or national NDCs). This involved adjusting the CI_t and $RenewableShare_{HPC,t}$ parameters to simulate different future energy landscapes.

3. Results

3.1. HPC Energy Consumption Forecasts

3.1.1. Full Capacity Energy Consumption.

The application of various time-series models to historical HPC energy consumption data (e.g., spanning 2015-2022) for forecasting up to 2027 revealed a consistent projection of continued growth, albeit with differing magnitudes across models. ARIMA models, such as ARIMA (2,1,1), often captured strong upward trends, predicting substantial increases in energy demand if historical growth rates persist. MA (1) models, tending to smooth out short-term fluctuations, typically offered more conservative, stabilized forecasts. LSTM networks, when trained on sufficiently long and complex datasets, demonstrated a strong capacity to model non-linear growth patterns and often yielded RMSE and MAE values competitive with, or superior to, traditional statistical methods, particularly where

underlying dynamics were intricate. The variation in forecasts underscored the sensitivity of projections to model choice and the inherent uncertainties in extrapolating complex trends.

3.1.2. Utilization Rate (PUE) Trends and Forecasts.

Analysis of historical PUE data for data centers (e.g., from 2007 to 2024) generally confirmed a significant downward trend over the past decade, reflecting widespread adoption of energy efficiency measures and improved operational practices. Forecasts for PUE (e.g., up to 2034) using the suite of time- series models indicated diverse potential future trajectories. The Holt-Winters exponential smoothing model, adept at capturing sustained trends, often projected a continued, albeit potentially slowing, decline in PUE, suggesting ongoing improvements towards optimal energy efficiency (approaching PUE 1.0). In contrast, models like AR (2) sometimes indicated a stabilization or plateauing of PUE values, implying that further substantial gains in this specific metric might become increasingly challenging to achieve without transformative technological breakthroughs.

3.2. Carbon Emissions Analysis

3.2.1. Relationship between Energy Consumption and Carbon Emissions.

A longitudinal analysis of historical data (e.g., global electricity generation and CO2 emissions from 1971 to 2011) unequivocally revealed a strong, positive, and statistically significant correlation between total electricity generation and associated carbon emissions. This historical linkage underscores the carbon-intensive nature of the global energy system over past decades. Periods of accelerated growth in global electricity demand, particularly evident from the mid-1990s onwards, corresponded with commensurately sharper increases in CO2 emissions. However, a subtle potential decoupling effect was observed in the later years of this historical dataset (e.g., around 2009-2011), where the rate of emissions growth appeared to slow relative to the continued increase in electricity generation. This nascent trend could be attributed to the early-stage impacts of energy efficiency improvements and the initial, albeit limited, penetration of cleaner energy sources into the global energy mix.

3.2.2. Energy Mix Composition.

Analysis of net summer generating capacity by energy source (based on representative recent datasets) highlighted the persistent dominance of fossil fuels, with coal, for instance, constituting a significant portion of installed capacity (e.g., approximately 180 GW in sample data). Concurrently, these data also reflected a no- table and accelerating growth in the installed capacity of renewable energy sources, particularly wind power (e.g., around 147 GW) and solar photovoltaics (PV) (e.g., around 89 GW), signaling an ongoing but gradual energy transition.

3.2.3. Forecast of World Electricity Generation by Source.

Auto ARIMA models, with order selection optimized using the Akaike Information Criterion (AIC), were employed to forecast global electricity generation categorized by primary energy source (fossil fuels versus renewables including nuclear) up to a medium-term horizon (e.g., 2030). Representative optimal models, such as ARIMA(5,2,0)[ct] for the renewables + nuclear category (with a sample AIC $\approx$ 450.08) and ARIMA(0,1,0)[c] for fossil fuels (sample AIC $\approx$ 541.20), projected a continued overall increase in total global electricity generation. Within this growing demand, the share of electricity generated from renewable and nuclear sources was anticipated to increase substantially, potentially reaching approximately 45% of the total by 2030. Nevertheless, this also implies a continued, significant reliance on fossil fuels to meet a large portion of global electricity needs in the coming decade, underscoring the challenge of rapid decarbonization.

3.3. Impact of Renewable Energy on Carbon Reduction (GLM Results)

The Generalized Linear Model (GLM), as specified in Equation 15, applied historical national or global data to understand drivers of carbon emissions from the electricity sector, yielded several key insights (based on representative coefficient values and significance levels from typical analyses):

The coefficient associated with Fossil Fuel Share in electricity generation (e.g., $\beta_2 \approx 0.0008$) was consistently positive and typically statistically significant or marginally significant (e.g., p-value ≈ 0.086). This finding aligns with a priori expectations, confirming that a higher proportion of fossil fuels in the energy mix leads to increased CO2 emissions.

The coefficient for Renewable Energy Share (e.g., $\beta 3 \approx 0.0010$) in some model configurations yielded a paradox- ically positive sign, though often lacking statistical significance (e.g., p-value ≈ 0.098). This counterintuitive result, when observed, does not imply that renewables increase emissions per se. Rather, it may reflect complex systemic interactions, such as periods where rapid growth in overall energy demand (even if partially met by new renewables) still leads to increased utilization of existing fossil fuel capacity, or it could be an artifact of multicollinearity with other predictors or aggregation effects in the data. It highlights that simply adding renewables may not be sufficient if overall demand and fossil fuel use are not simultaneously managed.

Variables such as Total Electricity Consumption and the Interaction Term (FossilShare × RenewableShare) exhibited varying levels of significance across different datasets and model specifications, sometimes failing to reach statistical significance, indicating nuanced relationships that may require more sophisticated modeling or disaggregated data to fully elucidate.

The intercept term (e.g., $\beta_0 \approx 3.478 \times 10^{-10}$) was often statistically significant (e.g., p-value ≈ 0.025) but typically of a magnitude that rendered it practically negligible in interpreting emission levels. Standard model diagnostic checks, including analysis of residuals (e.g., Residuals vs. Fitted values plots, Q-Q plots), were performed to assess the adequacy of the GLM assumptions.

3.4. HPC Water Consumption Forecasts

Time-series analysis of available historical water usage data attributable to data centers and HPC facilities (e.g., from 2012 to 2021) generally indicated an overall increasing trend in water consumption. Forecasts extending to a future horizon (e.g., 2026) using different models showed a range of possibilities: AR (2) models often projected a substantial and continued increase in water demand, reflecting historical growth momentum. ARMA, ARIMA, and SARIMA models typically suggested more moderate, though still positive, growth trajectories. MA (1) models tended to predict relative stability or only slight increases. In contrast, the Holt-Winters model, in some instances, forecasted an eventual decrease in water usage after an initial period of increase, potentially reflecting assumptions about the widespread adoption of highly water-efficient cooling technologies in the longer term. This divergence highlights the uncertainty in water footprint projections, heavily dependent on technological choices and climatic conditions influencing cooling needs.

4. Discussion

The collective findings of this multifaceted study compellingly underscore the escalating environmental footprint associated with the global proliferation of High-Performance Computing. Our energy consumption forecasts, derived from a diverse suite of time-series models, consistently point towards a continued and significant rise in HPC energy demand in the coming years. While the precise magnitude of this increase varies depending on the specific model assumptions and underlying data characteristics, the overarching trajectory is one of sustained growth. The observed historical decline in Power Usage Effectiveness (PUE) across data centers is a positive development, indicating tangible progress in enhancing operational energy efficiency at the facility level. However, a critical observation is that these efficiency gains, while important, are frequently outpaced by the sheer exponential growth in computational demand and the increasing scale and density of HPC deployments. Consequently, the net effect is often an overall increase in absolute energy consumption by the HPC sector.

The robust historical correlation identified between aggregate electricity consumption and carbon emissions serves as a stark reminder of the deeply entrenched carbon-intensive nature of current global and many regional energy systems. Our projections suggest that while the share of renewable

energy sources in the global electricity generation mix is anticipated to increase significantly by 2030, potentially reaching around 45%, fossil fuels will likely continue to constitute a substantial component of energy supply in the medium term. This persistence of fossil fuel reliance, coupled with rising HPC energy demand, implies that without aggressive and targeted interventions, the carbon emissions attributable to the HPC sector are poised to continue their upward trajectory, posing a considerable challenge to achieving broader climate mitigation goals.

The results from the Generalized Linear Model (GLM) concerning the specific impact of renewable energy share on aggregate carbon emissions warrant careful and nuanced interpretation. The instances where the coefficient for renewable energy share appeared non-significant or, paradoxically, positive do not inherently contradict the established scientific consensus that renewable energy technologies possess significantly lower lifecycle greenhouse gas emissions compared to fossil fuels. Rather, such findings may reflect the intricate complexities of large-scale energy systems. For example, during periods of rapid overall energy demand growth, the addition of new renewable capacity might not be sufficient to displace existing fossil fuel generation or meet all new demand, leading to a situation where both renewable generation and fossil-fuel-based emissions increase concurrently at an aggregated level. Furthermore, multicollinearity among predictor variables, limitations in data aggregation, or specific model misspecifications could also contribute to such counterintuitive statistical outcomes [5]. Nonetheless, the consistent positive and often significant coefficient for fossil fuel share unequivocally reaffirms its role as a primary driver of carbon emissions.

Beyond energy and carbon, the projected increase in water consumption associated with HPC operations, particularly for cooling purposes, presents another significant environmental concern [6]. This issue is especially critical in water scarce regions, where the establishment and operation of large-scale data centers can exert considerable pressure on local freshwater resources and potentially exacerbate existing water stress conditions. The choice of cooling technology and local climatic factors are key determinants of this water footprint.

4.1. Limitations

This comprehensive study, despite its rigorous approach, is subject to several inherent limitations that should be acknowledged: (1) *Data Availability and Granularity*: The analyses primarily rely on publicly available, often aggregated, datasets. Such data may lack the fine-grained geographical, temporal, or technological specificity required for highly precise regional or facility-level forecasts. Access to proprietary data on direct renewable energy procurement by specific HPC facilities and detailed operational parameters is often restricted, limiting the depth of certain analyses. (2) *Model Assumptions and Uncertainty Quantification*: All predictive models, whether statistical or machine learning-based, are founded on historical trends and embody a set of underlying assumptions about the continuity of these trends and relationships. These assumptions may not hold true in the face of disruptive technological advancements (e.g., breakthroughs in quantum computing, radical improvements in semiconductor efficiency, novel cooling technologies) or significant, unforeseen policy shifts. While efforts were made to evaluate model performance, a more formal quantification of forecast uncertainty (e.g., through prediction intervals or ensemble methods) would further enhance the robustness of the projections. (3) *Scope of Environmental Impacts*: While this study provides a focused analysis of energy consumption, associated carbon emissions, and water usage, other important dimensions of HPC's environmental footprint, such as the lifecycle impacts of e-waste generation from hardware obsolescence, and the resource depletion associated with the manufacturing of complex HPC components (including rare earth elements), were not quantitatively modeled in depth. A complete environmental picture would necessitate a more comprehensive lifecycle assessment (LCA) approach [7]. (4) *Complexity of Energy Systems and Policy Dynamics*: The intricate interplay between diverse energy sources, electricity grid stability requirements, dynamic market forces, rates of technological innovation, and the effectiveness of various car- bon reduction policies forms a highly complex system. The statistical models employed, while

informative, necessarily provide a simplified representation of these multifaceted interactions. Accurately predicting the precise impact of future, often politically contingent, policy interventions remain a significant challenge.

4.2. Future Research Directions

To build upon the findings of this study and further advance the understanding and mitigation of HPC's environmental impact, future research should prioritize the following directions: (1) Develop more sophisticated and highly integrated assessment models. These models should aim to endogenously incorporate technological learning curves, the dynamic impacts of specific policy interventions, economic feedback mechanisms (such as carbon pricing effects), and behavioral changes among HPC users and providers. (2) Advocate for and contribute to the establishment of improved global data collection frameworks. This includes developing standardized methodologies for reporting HPC-specific energy consumption, direct renewable energy procurement, Power Usage Effectiveness (PUE), Water Usage Effectiveness (WUE), and other key environmental performance indicators at a more granular, consistent, and internationally comparable level. (3) Conduct comprehensive and detailed Lifecycle Assessments (LCAs) for various HPC architectures, components, and operational models. Such LCAs should meticulously account for the full spectrum of environmental impacts across the entire value chain, including Scope 3 emissions related to upstream supply chains (manufacturing, transport) and downstream end-of-life management (e-waste processing, material recovery). (4) Investigate the broader socio-economic and ethical implications of HPC's environmental footprint. This includes exploring issues of environmental justice (e.g., the siting of data centers in relation to vulnerable communities), regional disparities in access to sustainable HPC resources, and the equitable distribution of environmental burdens and benefits. (5) Explore the transformative potential of emerging technologies, particularly Artificial Intelligence (AI), in optimizing the energy efficiency of HPC workloads (e.g., through intelligent job scheduling, dynamic resource allocation, and AI-accelerated simulations). Research should also focus o advanced grid management techniques that facilitate the seamless integration of variable renewable energy sources to power HPC facilities [8, 9]. (6) Analyze the efficacy and transferability of specific carbon reduction technology pathways and policy instruments in diverse regional and national contexts. For instance, detailed case studies of green data center development initiatives, such as those being pursued in China and other leading HPC nations, can offer valuable lessons and best practices [10].

5. Summary

This research provides a detailed quantitative assessment of the significant and growing environmental footprint associated with High-Performance Computing, with a particular emphasis on its substantial energy consumption, consequent carbon emissions, and considerable water usage. Our comprehensive modeling efforts indicate that, in the absence of concerted and transformative interventions, the environmental burden imposed by the HPC sector will continue to escalate. This trajectory poses a risk of potentially undermining the positive contributions that HPC technology makes towards achieving broader goals of sustainable development and scientific advancement. Effecting a meaningful transition towards a genuinely sustainable HPC ecosystem necessitates a strategically orchestrated, multi-pronged approach that engages all stakeholders.

The core pillars of such an approach must include: (1) *Continuous Enhancement of Energy Efficiency*: This requires sustained investment in research and development focused on creating more energy-efficient hardware components (including processors, memory modules, and storage systems) and developing sophisticated software solutions (such as energy-aware algorithms, intelligent compilers, and optimized job schedulers). At the facility level, data centers must aggressively pursue and implement operational best practices in areas like advanced cooling systems, efficient power distribution architectures, and dynamic server utilization optimization, with the aim of consistently achieving PUE val- ues approaching the theoretical optimum of 1.0. (2) *Accelerated Adoption of*

Renewable Energy Sources: HPC facilities should strategically prioritize the sourcing of their electricity from clean, renewable sources. This can be achieved through a combination of direct onsite generation (e.g., solar PV installations), long-term Power Purchase Agreements (PPAs) with renewable energy developers, and procurement through credible green tariffs or renewable energy certificates. Concur- rently, governments and regulatory bodies have a crucial role to play in incentivizing broad renewable energy development and accelerating the decarbonization of national and regional electricity grids. (3) *Implementation of Sustainable Water Management Practices*: Given the significant water demands of many cooling systems, data centers must proactively adopt highly water-efficient cooling technologies. Furthermore, they should explore and implement advanced water recycling and reuse systems, particularly when sited in water-stressed or arid regions, to minimize their impact on local freshwater resources. (4) *Promotion of Circular Economy Principles throughout the HPC Lifecycle*: The challenge of e-waste from obsolete HPC hardware must be tackled by embedding circular economy principles into the design, manufacturing, op- eration, and end-of-life phases. This includes designing hardware for enhanced durability, modularity, and recyclability, and developing robust, environmentally sound programs for the collection, refurbishment, and responsible recycling of HPC components. (5) *Fostering Supportive Policies and International Standardization*: The development and widespread adoption of clear, consistent metrics (such as Power Usage Effectiveness (PUE), Water Usage Effectiveness (WUE), and Carbon Usage Effectiveness (CUE) [4]), standardized reporting protocols, and robust policy frameworks are essential for promoting transparency, enabling benchmarking, and ensuring accountability in managing HPC's environmental impact. Market-based mechanisms like carbon pricing and green credit programs can provide powerful financial incentives for adopting sustainable practices. (6) *Systematic Integration of Environmental Considerations into HPC Development and Procurement*: Sustainability criteria should be established as core design principles for all future generations of HPC systems, software, and applications. Procurement processes for HPC resources should also incorporate stringent environ- mental performance requirements, thereby driving demand for greener solutions.

Confronting and mitigating the environmental challenges posed by High-Performance Computing is not merely an oper-ational or technical concern; it represents a profound strategic imperative for the 21st century. By proactively embracing and comprehensively implementing sustainable practices, the global HPC community can ensure that this exceptionally powerful and transformative technology continues to serve as a vital engine for innovation, scientific discovery, and posi- tive societal change, all while respecting planetary boundaries and contributing to a more sustainable and equitable world.

Acknowledgements

The authors wish to express their gratitude to the anonymous reviewers whose insightful comments and constructive feedback significantly contributed to enhancing the quality and clarity of this manuscript. (Optional: This research was supported in part by Grant [Number] from the [Funding Agency Name].)

References

[1] A. Shehabi, S. Smith, R. Masanet, H. Horner, L. Azevedo, E. Mills, Building and Environment 46, 990 (2011).

[2] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, arXiv:2104.10350 (2021).

[3] Y. Liu, Z. Geng, J. Li, Z. Wang, Global Energy Interconnection 3, 272 (2020).

[4] D. Azevedo, C. LOPER, B. FORREST, The Green Grid, White Paper #32: Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric (2010).

[5] T. Bhattacharya, S. Verma, S.K. Singh, P.K. Singh, International Journal of Energy and Environmental Engineering 14, 627 (2023).

[6] T. Renugadevi, M. Sangeetha, R. Senthil Kumar, S. Velvizhi, Sustainability 12, 6383 (2020).

[7] Z. Yin, H. Wang, Environment, Development and Sustainability (2024) doi:10.1007/s10668-023-04246-7.

[8] A. Radovanovic´, G. Koutitas, L. Tassiulas, IEEE Transactions on Power Systems 38, 1270 (2022).

[9] P. Wiesner, S. Salvaneschi, P. Felber, R. B bikers, M. Mezini, Proceedings of the 22nd International Middleware Conference (ACM, New York, NY, USA, 2021) p. 1.

[10] G. Li, Y. Zhang, Q. Wang, Y. Gao, J. Mu, Environmental Research 231, 116248 (2023).