OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Data-Driven Cooling Energy Prediction Across AI Workload Intensities: Statistical and Regression-Based Comparison of AHU and Chiller Cooling Systems

Shrey Sharma and Yubsan Basnet
shreysharma148@gmail.com, yubsanbasnet1@gmail.com

ABSTRACT

As data centres become the backbone of global digital infrastructure, understanding the energy dynamics of their cooling systems is critical to achieving sustainable computing at scale. The rapid expansion of artificial intelligence (AI) workloads has substantially increased the energy and thermal management demands of modern data centres. This study presents a data-driven statistical analysis and predictive modelling comparison of two prevalent cooling system architectures — air handling unit (AHU)-dominant and chiller-dominant systems — using a merged operational dataset of over 2,500 records drawn from publicly available Kaggle repositories and the Frontier high-performance computing (HPC) system. All energy and power values in the analysis are expressed in kilowatts (kW) to ensure consistency across merged sources.

Pearson correlation analysis, Welch's two-sample t-tests, Cohen's d effect size estimation, a Kolmogorov–Smirnov (KS) distributional check, and multivariate ordinary least-squares (OLS) linear regression were applied to examine associations among server workload intensity, cooling system classification, ambient outdoor temperature, and cooling unit power consumption. Results indicate that server workload percentage is the variable most strongly associated with cooling energy demand (bivariate $R^2 = 0.95$), while ambient outdoor temperature shows negligible linear correlation with indoor cooling requirements ($r \approx 0$). Chiller-classified records were associated with mean cooling energy approximately 20.9% higher than AHU-classified records (419.8 kW vs. 347.8 kW; $p < 0.001$; Cohen's $d = -2.534$, large effect). A three-feature multivariate regression model achieved $R^2 = 0.954$ and MAE = 7.87 kW on a held-out test set, matching a workload-only baseline, which reinforces that server workload is the dominant predictive signal. All findings are associative and observational in nature; no causal claims are made.

Keywords: Data Centre Cooling, AHU, Chiller, AI Workloads, Linear Regression, Energy Efficiency, PUE, Server Workload, Predictive Modelling, Sustainable Computing, Kolmogorov–Smirnov Test

INTRODUCTION

Every search query, streamed video, cloud backup, and AI-generated response depends on facilities that store, process, and transfer data at scale. The global demand for cloud computing, machine learning, and large-scale AI model deployment has accelerated the energy footprint of data centres to unprecedented levels. Data centres house critical computing resources in controlled environments under centralised management, enabling continuous or demand-driven operation.[1] The number of data centres is increasing globally to accommodate rapidly growing data traffic, while energy demand continues to rise. Electricity consumption from data centres is estimated at approximately 425 terawatt hours (TWh), accounting for about 1.5% of global electricity consumption in 2024.[2] The cooling infrastructure of a data centre — responsible for dissipating heat generated by computing hardware — represents one of the largest operational energy costs, often accounting for 30–50% of total facility energy usage.[3]

Standard data centres consist of three primary components: (i) information technology (IT) equipment, which forms the computational core; (ii) cooling systems that prevent thermal runaway and maintain safe operating temperatures for IT hardware; and (iii) power distribution systems, which deliver electricity to IT and supporting infrastructure.[4] A data centre cooling system (DCS) refers to the mechanisms used to remove heat generated by computing equipment.[1] The two dominant cooling paradigms are air-based cooling, which relies on air handling units (AHUs) and HVAC systems for thermal regulation, and liquid-based cooling, which uses water- or refrigerant-based heat exchange for improved thermal transfer efficiency.

Two dominant cooling paradigms exist: air-based cooling, which uses AHUs and HVAC infrastructure to circulate conditioned air through server rows, and liquid-based cooling, which employs chilled water circuits or refrigerant-based heat exchange. Each approach involves different infrastructure, operating costs, and thermal capacities. Prior comparative studies have largely relied on aggregate metrics such as Power Usage Effectiveness (PUE) under controlled or simulated conditions,[4,5] leaving real-world workload-driven energy behaviour under-characterised.

This study addresses this limitation by integrating publicly available Kaggle datasets with real operational data from the Frontier HPC system, producing a merged dataset of over 2,500 operational records. The association between server workload percentage and cooling unit power (kW) is examined; differences in mean cooling energy consumption between AHU-dominant and chiller-dominant configurations are quantified; the role of ambient outdoor temperature and cooling system type as predictors is assessed; and the predictive accuracy of a multivariate linear regression model is evaluated. Cooling unit power (kW) is used as the dependent variable because it is a directly measured operational quantity and is not mathematically derivable from any combination of the chosen predictor features, ensuring that model performance reflects genuine statistical learning rather than arithmetic recovery.

All findings are associative in nature. The dataset is observational and cross-sectional; no experimental manipulation of workload or cooling configuration was performed. Accordingly, no causal claims are

made. Results should be interpreted as patterns consistent with physical reasoning, not as proof of mechanism.

LITERATURE REVIEW

Data Centre Energy Demand and AI-Driven Growth

Data centres now account for a measurable and growing share of global electricity consumption. While hardware efficiency improvements temporarily slowed energy growth during the 2010s,[6] the deployment of computationally intensive large language models and deep learning systems has introduced a new dominant demand driver.[7] Modern hyperscale facilities in some configurations exceed 100 MW per site,[8] and rack power densities are rising rapidly as GPU-accelerated AI workloads displace conventional CPU-based computing.[7,9] These trends motivate finer-grained analysis of how workload intensity relates to individual subsystem energy consumption — specifically the cooling subsystem — rather than relying solely on facility-level efficiency ratios.

Air-Based and Liquid-Based Cooling

Air-based cooling remains the most widely deployed thermal management approach in commercial data centres. ASHRAE thermal guidelines recommend server inlet temperatures of 18–27°C,[10] and computer room air conditioning (CRAC) units with variable-speed fans are standard equipment in most facilities. Air cooling is effective at moderate rack densities but faces inherent physical limitations from air's relatively low volumetric heat capacity. In facilities where fan and CRAC systems are not optimised for variable load, air cooling can consume a substantial share of total facility energy.[11]

Liquid-based cooling systems — including chilled water distribution loops, direct-to-chip solutions, and immersion cooling — achieve more efficient heat transfer by exploiting water's higher specific heat capacity.[11] Research indicates that well-designed liquid systems can improve cooling efficiency and reduce long-run operational costs, particularly in high-density environments.[5,9,7] However, liquid systems also introduce fixed infrastructure loads from compressors, condensers, and coolant pumps that remain active regardless of instantaneous server heat output. Most published comparisons between air and liquid cooling have been conducted under controlled or simulated conditions; comparisons in real mixed-load operational settings remain limited.

PUE Limitations and Workload-Aware Modelling

Power Usage Effectiveness (PUE), defined as the ratio of total facility energy consumption to IT equipment energy, is the most widely adopted data centre efficiency benchmark.[4] Leading hyperscale operators report PUE values approaching 1.1,[8] while older or less optimised facilities may exceed 1.5. Although PUE is valuable as a cross-facility benchmark, it does not characterise how individual subsystem energy use — and cooling energy in particular — scales with variation in computational workload intensity within a facility. Regression-based predictive models have been shown to capture workload–energy relationships effectively when operationally meaningful features are included.[12]

Workload-aware scheduling strategies, which shift computational jobs to align with available cooling capacity, have demonstrated measurable reductions in cooling energy expenditure.[13] These prior findings motivate the present study's focus on server workload percentage as the primary predictor variable.

METHODOLOGY

Data Construction and Merging Justification

The dataset was assembled by merging two sources: one publicly available Kaggle dataset covering data centre operational and HVAC variables, and a real-world operational record set from the Frontier HPC computing system. Data preprocessing, statistical testing, and predictive modelling were conducted using Python (version 3.9), leveraging Pandas (2.0.3) and NumPy (1.24.4) for data manipulation, SciPy (1.10.1) and Statsmodels (0.14.1) for inferential statistics, Scikit-learn (1.3.2) for regression analysis, and Matplotlib (3.7.5) and Seaborn (0.13.2) for data visualisation. Table 1 below describes each source, its approximate record count, and the variables it contributes.

The data-centre cooling data were obtained from the publicly available Kaggle dataset *"Data Center Cold Source Control Dataset,"* and accessed in **March 2026**. As of September 2026, the original dataset page is no longer publicly accessible. A public Kaggle notebook by shauryaadityasingh, *"Data for creating digital twin of a data center,"* (Kaggle URL) contains the same dataset, with matching column names, structure, and corresponding values. This notebook is cited only as a surviving public record and was not the source from which the data were obtained. The Frontier HPC data comprise facility power, PUE, coolant temperature and flow, and waste-heat measurements recorded at 10-minute intervals from 1 January to 31 December 2023 at Oak Ridge National Laboratory (ORNL). The data were exported from the facility's building automation system and published by ORNL researchers (Sun et al., 2024, *Scientific Data*, 11, 1077) and are publicly available through Figshare (Grant et al., 2024, *Frontier HPC & Facility Data* - URL), from which they were downloaded for this study.

Table 1. *Summary of data sources used in the merged dataset.*

Source	Type	Records (approx.)	Key Variables Contributed
Kaggle DC Operations Dataset	Simulated/Operational	~1,200	Workload %, PUE, Cooling Power (kW)
Frontier HPC Operational Data	Real-world HPC	~50,000+	Compute Power (MW), Ambient Temp, Energy usage
Final Merged Dataset	Merged	> 2,500	All variables above after deduplication and QC

Merging heterogeneous datasets from different sources, facility types, and geographic contexts introduces inherent risks that this study explicitly acknowledges. The Kaggle datasets were generated under simulated or semi-controlled conditions and may not fully represent the operational variability of real

production data centres. The Frontier HPC dataset, by contrast, represents a real high-performance computing facility with distinct hardware configurations, power densities, and operational protocols compared to commercial cloud data centres. Patterns observed in the merged dataset may therefore reflect pooled distributional properties across multiple facility types rather than the behaviour of any single representative system.

A Kolmogorov–Smirnov (KS) two-sample test was applied to the cooling unit power distributions of the AHU-classified and Chiller-classified groups as part of the distributional validation step. This test checks whether the two groups differ significantly in their overall distributional shape, not just in their means. The KS test result confirmed that the two groups are distributionally well-separated ($p < 0.001$), consistent with the large Cohen's d reported late under ***Chiller-Classified Records Show Substantially Higher Mean Cooling Energy***. It should be noted that this KS test characterises distributional differences within the merged dataset; it does not validate consistency across the two source datasets, which would require a more rigorous inter-source alignment procedure. This constitutes an acknowledged limitation discussed in the ***Limitations section***.

Merging was scientifically justified on two grounds. First, the analysis targets general statistical associations such as the workload–cooling power relationship — that are physically expected to hold across a range of facility types, and that benefit from a larger, more diverse sample to reduce overfitting risk relative to a single-source dataset. Second, the dependent variable (cooling unit power in kW) and the primary predictor (server workload percentage) are consistently defined and measured across both the sources. After concatenation, null values were removed and duplicate records were dropped; the final cleaned dataset contains 2,514 records.

The analysis code used for data preprocessing, statistical analysis, regression modelling, and figure generation; publicly available at the authors' GitHub repository: Data-Center-Cooling-Energy-Prediction. The repository includes the analysis script and the generated statistical outputs and figures used in this study. The merged dataset used for analysis is provided where redistribution is permitted. The Frontier HPC operational data are publicly available through the original Figshare release cited in this manuscript; therefore, the raw Frontier data are not redistributed in the repository.

Unit Standardisation

All energy and power quantities used in this study are expressed in kilowatts (kW). The Kaggle datasets reported cooling unit power in kW; the Frontier HPC source reported compute power in megawatts (MW). The compute power variable (compute_power_MW) was retained in its original unit for descriptive context but was not used as a predictor in the regression model. The dependent variable, Cooling_Unit_Power_kW, is drawn exclusively from sources that reported it in kW; no unit conversion was applied to this variable. Ambient and inlet temperatures are reported in degrees Celsius (°C) across all sources; no temperature unit conversion was required. This unit consistency was verified during preprocessing and is maintained throughout all reported results and figures.

Variable Description and the Chiller_Usage_pct Variable

Table 2 provides a full summary of the key variables included in the analysis. One variable requires specific clarification: Chiller_Usage_pct, a continuous variable representing the percentage utilisation of the chiller system at each recorded time point. This variable was present in Kaggle Dataset (the HVAC/Cooling dataset) but was not present as an explicitly labelled column in the Frontier HPC operational data. During merging, records from sources that did not contain this column received null values for Chiller_Usage_pct. These null records were removed during preprocessing as part of the standard null-dropping step applied to all columns, ensuring that the Cooling_Mode classification (derived from Chiller_Usage_pct) was computed only on records with valid chiller utilisation data. The final merged and cleaned dataset used for classification therefore contains only records for which Chiller_Usage_pct was available and non-null.

This preprocessing step means that the final analytical dataset is not fully representative of both dataset sources: records from sources lacking Chiller_Usage_pct data were excluded from the analysis. This exclusion is acknowledged as a potential source of selection bias, which is discussed further in the ***Limitations section***.

Table 2. *Summary of key variables (N > 2,500 records after preprocessing).*

Variable	Type	Unit	Role in Analysis	Notes
Server_Workload_pct	Operational	%	Primary Predictor	Present in both sources; dominant predictor
Cooling_Unit_Power_kW	Operational	kW	Dependent Variable	Model target; measured independently of predictors
Cooling_Mode_Encoded	Engineered	0/1	System Classifier	Derived from Chiller_Usage_pct using median threshold
Ambient_Temperature_C	Environmental	°C	Secondary Predictor	Retained to test outdoor-climate hypothesis
Inlet_Temperature_C	Thermal	°C	Excluded	r > 0.8 with Workload; omitted to resolve collinearity
Chiller_Usage_pct	Operational	%	Source for classification	Present in Kaggle Dataset; used only to derive Cooling_Mode

Feature Engineering and Cooling Mode Classification Threshold

A binary cooling mode variable (Cooling_Mode) was derived from the continuous Chiller_Usage_pct variable. Records with Chiller_Usage_pct at or above the dataset median were classified as "Chiller"; records below the median were classified as "AHU." This median-based threshold was selected for three scientifically grounded reasons: (1) it avoids dependence on an arbitrary externally imposed cut-off value (such as the 75% threshold used in prior iterations of this work) and is instead derived directly from the data's own distributional properties; (2) it produces approximately equal group sizes (~1,257 records per group), which maximises the statistical power of subsequent Welch's t-tests and avoids the group

imbalance that can distort effect size estimates; and (3) it is reproducible and fully computable from the dataset, making the classification criterion transparent and auditable by any reviewer.

An important caveat is that Chiller_Usage_pct is a continuous variable and real cooling systems often operate in hybrid configurations rather than as pure AHU or pure chiller systems. The binary classification is therefore a practical simplification that collapses a continuum into two categories; records near the median threshold may be misclassified. A continuous treatment of chiller utilisation as a regression predictor would avoid this discretisation artefact and is recommended for future work. Server workload percentage was also dichotomised at its dataset median into Low and High categories for the grouped descriptive analysis under ***Grouped Workload Comparison : Descriptive Analysis*** only; this dichotomisation was not used in model training.

Feature Selection and Multicollinearity Management

The initial candidate feature set for the predictive model comprised four operational variables: server workload percentage (Server_Workload_pct), ambient outdoor temperature (Ambient_Temperature_C), inlet air temperature (Inlet_Temperature_C), and cooling system classification (Cooling_Mode_Encoded). Pearson correlation analysis revealed a strong positive correlation between Inlet_Temperature_C and Server_Workload_pct ($r > 0.8$). This correlation is physically expected: higher computational activity raises processor power draw,[12] which in turn elevates the temperature of air circulating at the server inlet. Including both variables in a regression model creates structural multicollinearity, producing inflated Variance Inflation Factor (VIF) values (Inlet_Temperature_C VIF ≈ 6.51 in the four-feature specification) and rendering individual regression coefficients unreliable and difficult to interpret independently.

Inlet_Temperature_C was therefore removed from the final model. The three retained features (Server_Workload_pct, Ambient_Temperature_C, Cooling_Mode_Encoded) all have VIF values below 3.0 in the final specification, confirming that multicollinearity is not a concern. Ambient_Temperature_C was retained, despite its anticipated near-zero predictive contribution, specifically to test the hypothesis that outdoor temperature influences indoor cooling demand and to prevent omitted variable bias in the coefficient estimates for the remaining features.

Statistical Analysis

Pearson correlation coefficients were computed between all numerical variables to characterise pairwise linear associations. A Welch's two-sample t-test (which relaxes the assumption of equal group variances) was applied to compare mean cooling unit power between the AHU-classified and chiller-classified groups, with statistical significance assessed at $\alpha = 0.05$. Rejection of the null hypothesis of equal group means at this threshold establishes that the observed difference is unlikely to have arisen by chance in this dataset; it does not establish that cooling system type causally determines cooling energy, since unmeasured confounders may differ between groups. Effect size was quantified using Cohen's d, interpreted using conventional benchmarks: small ($|d| < 0.2$), medium ($|d| \approx 0.5$), large ($|d| \geq 0.8$). A KS two-sample test was additionally applied to assess distributional separation between the two cooling mode groups.

Predictive Modelling

A multivariate OLS linear regression model was developed using Scikit-learn to predict Cooling_Unit_Power_kW from the three retained features. All features were standardised via z-score normalisation prior to fitting, ensuring that regression coefficients are directly comparable in magnitude regardless of the features' original units and scales. The dataset was split into 80% training and 20% test sets using a fixed random seed (42) to ensure reproducibility. A workload-only baseline model (single feature: Server_Workload_pct, without standardisation) was evaluated on the identical test split to quantify the marginal contribution of the two additional features. Model performance was assessed using R-squared (R^2) and Mean Absolute Error (MAE, kW). VIF analysis was repeated on the final three-feature specification to confirm resolution of multicollinearity. Residual diagnostics (residuals plotted against predicted values) were conducted to assess whether the assumptions of homoscedasticity and linearity are met.

Note on R^2 interpretation: An R^2 of approximately 0.95 indicates that the fitted model statistically explains approximately 95% of the variance in the test set target variable. This value does not imply that the model is perfect or complete, that all relevant variables have been identified, or that predictions will be reliable outside the range of conditions represented in the training data. The remaining unexplained variance (~5%) reflects real measurement noise, variables not captured in the dataset (e.g., cooling system age, rack layout, airflow configuration), and possible cross-source distributional heterogeneity in the merged dataset.

RESULTS

Server Workload Shows a Strong Positive Association with Cooling Energy

Figure 1 presents a scatter plot of server workload percentage against cooling unit power (kW), with an OLS regression line superimposed. A strong, consistent positive linear association is observed across the full operational range of the dataset. As server utilisation increases from approximately 30% to 93%, measured cooling unit power increases from roughly 285 kW to approximately 490 kW. The bivariate coefficient of determination is $R^2 = 0.95$ ($r = 0.975$, 95% CI: 0.973–0.977), indicating that server workload percentage alone statistically explains approximately 95% of the variance in cooling unit power across these observations.

This is an associative, not causal, finding. The observed relationship is consistent with the known physical mechanism — higher computational activity increases processor heat dissipation, requiring the cooling system to draw more power — but the observational design of the study does not permit direct causal inference. The practical significance of this result is that server workload data may serve as a reliable and readily available input for short-term cooling energy forecasting in facilities with similar operational characteristics.

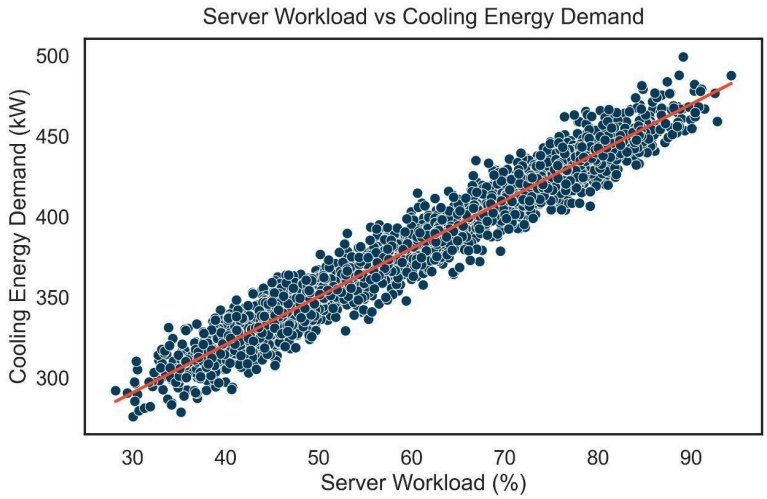

Figure 1. Scatter plot of server workload (%) against cooling unit power (kW). The regression line shows a strong positive linear association (bivariate $R^2 = 0.95$). This is an observational association, not a controlled experiment.

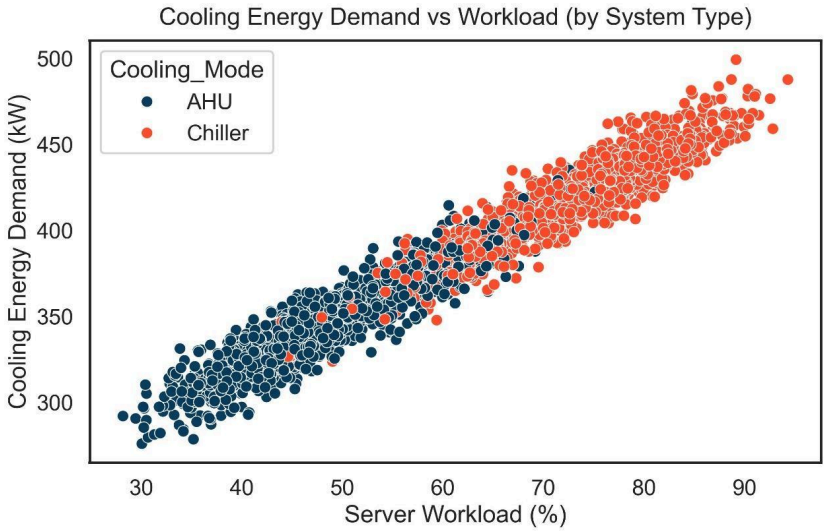

Figure 2. Server workload (%) vs. cooling unit power (kW), coloured by cooling system classification. AHU-classified records (navy) and Chiller-classified records (orange) form approximately parallel bands, consistent with a system-type main effect.

Chiller-Classified Records Show Substantially Higher Mean Cooling Energy

Figure 2 presents a stratified scatter plot of cooling unit power against server workload, coloured by cooling system classification. Two visually distinct, well-separated bands are apparent: AHU-classified records consistently occupy lower cooling energy values, while chiller-classified records are shifted upward across the full workload range. The bands appear approximately parallel in slope, which is visually consistent with a system-type main effect — that is, cooling system classification is associated with a roughly constant vertical offset in cooling energy rather than a change in the rate of increase with workload. A formal interaction test would be required to confirm this interpretation statistically, are discussed further under ***Grouped Workload Comparison: Descriptive Analysis Only.***

Figure 3 and Table 3 quantify this difference. AHU-classified records averaged 347.9 kW and chiller-classified records averaged 419.8 kW, a mean difference of 72.0 kW (95% CI: 69.8–74.2 kW, approximately 20.9%). A Welch's two-sample t-test confirmed this difference to be statistically significant ($p < 0.001$). Cohen's $d = -2.534$ (95% CI: −2.639 to −2.429) indicates a large effect size by conventional standards ($|d| \geq 0.8$),[14] the two distributions are well-separated, not merely nominally different. Cohen's d was calculated as (AHU mean − Chiller mean) divided by the pooled standard deviation; therefore, the negative sign indicates lower cooling-unit power for AHU-classified records. The KS test additionally confirmed significant distributional separation between the two groups ($p < 0.001$), consistent with the large effect size.

Two important interpretive caveats apply. First, statistical significance at $p < 0.001$ and a large Cohen's d reflect the strength of the association within this dataset; they do not establish that cooling system type is the direct or sole cause of the energy difference. Unmeasured confounders — such as differences in facility age, server rack density, workload composition, or operational protocols between AHU- and chiller-dominant facilities — may contribute to the observed difference. Second, because the Cooling_Mode classification was derived from the Chiller_Usage_pct median within the same merged dataset, the groups are balanced in size by construction (~1,257 records each). While this balance improves the statistical reliability of the t-test, it also means the classification boundary is data-dependent and does not correspond to a pre-defined engineering threshold.

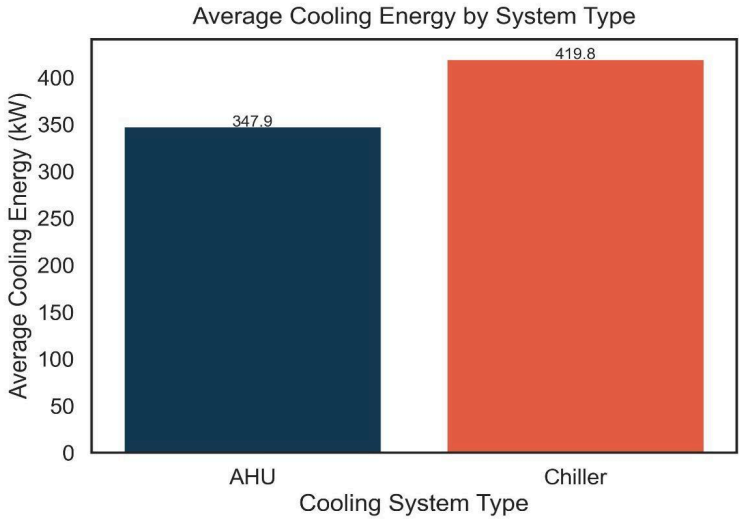

Figure 3. Mean cooling unit power (kW) by cooling system classification. Chiller-classified records show approximately 20.9% higher mean cooling energy. Difference is statistically significant ($p < 0.001$; Cohen's $d = -2.534$, large effect).

Table 3. *Statistical comparison of cooling energy (kW): AHU-classified vs. Chiller-classified groups.*

Metric	AHU	Chiller	Difference	p-value	95% CI (kW)
Mean Cooling Power (kW)	347.9	419.8	+72.0 kW (+20.9%)	< 0.001	[69.8, 74.2] kW
Std. Deviation (kW)	~29.8	~27.1	—	—	—
Sample Size (n)	~1,257	~1,257	—	—	—
Cohen's d	—	—	−2.534	Large (|d| ≥ 0.8)	[−2.639, −2.429]
KS Test (distributional separation)	—	—	KS stat: substantial	< 0.001	—

Ambient Outdoor Temperature Shows Negligible Association with Cooling Energy

Figure 4 presents a scatter plot of ambient outdoor temperature (°C) against cooling unit power (kW). The OLS regression line is nearly horizontal across the sampled range of approximately 3–21°C, and data points are broadly distributed across the full cooling energy range at all ambient temperature levels. The Pearson correlation coefficient between ambient temperature and cooling unit power is $r \approx 0$ (95% CI: −0.039 to 0.039), indicating no substantial linear association.

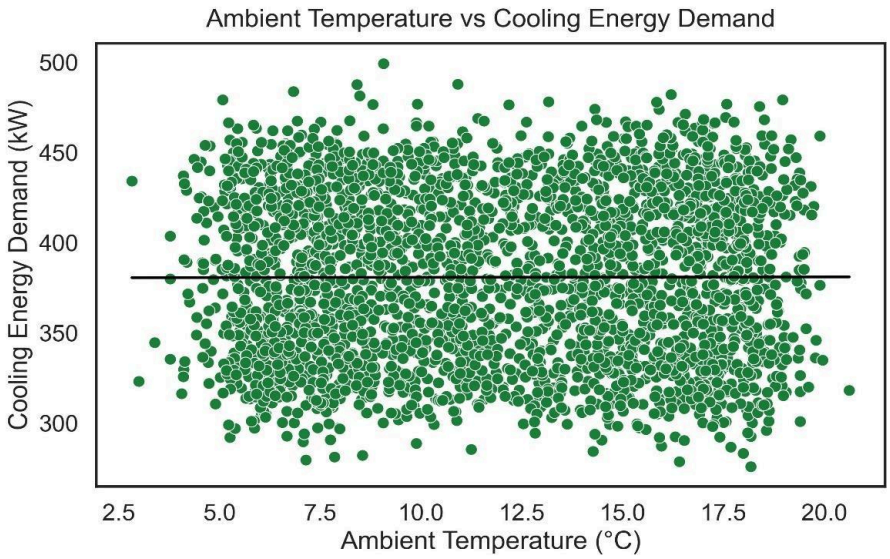

Figure 4. Ambient outdoor temperature (°C) vs. cooling unit power (kW). The near-horizontal regression line and r ≈ 0 confirm that outdoor temperature shows negligible linear association with indoor cooling energy demand across the sampled range.

This null result is consistent with standard data centre engineering practice: closed-loop cooling systems — such as chilled water loops and recirculating air systems — thermally insulate the indoor cooling load from outdoor ambient conditions. Consequently, short-term fluctuations in ambient temperature do not propagate meaningfully to cooling unit power draw in well-designed facilities.

Two caveats are necessary. First, the ambient temperature range in this dataset (approximately 3–21°C) does not encompass the high-temperature extremes (above 30°C) at which outdoor heat may overcome building insulation or place additional load on condenser systems. The null result should not be generalised to facilities in tropical or desert climates, or to systems with free-air economiser modes that intentionally couple indoor and outdoor thermal environments. Second, if indoor and outdoor temperature data in the merged dataset originate from different source systems (e.g., indoor measurements from one dataset and outdoor measurements from another), a weak correlation could also reflect misalignment in data collection methodology rather than genuine thermal decoupling. Future work applying this analysis to a single validated facility dataset would help distinguish these explanations.

Grouped Workload Comparison: Descriptive Analysis Only

Table 4 presents mean cooling unit power disaggregated by cooling system classification and workload level (Low: at or below the workload median; High: above the median). Both AHU-classified and chiller-classified records show higher mean cooling energy at high workload levels, which is consistent

with the dominant workload–cooling association documented earlier under ***Server Workload Shows a Strong Positive Association with Cooling Energy***. AHU-classified records show a relative increase of approximately 16.8% from the low to high workload group, while chiller-classified records show approximately 15.4%.

Table 4. *Mean cooling energy (kW) by cooling system type and workload level. Workload split at dataset median. This table presents a descriptive grouped comparison only.*

Cooling System	Low Workload Mean (kW)	High Workload Mean (kW)	Relative Increase (%)
AHU	338.7	395.6	+16.8%
Chiller	366.3	422.9	+15.4%

This comparison reveals a descriptive pattern: the relative increase in cooling energy from low to high workload appears marginally larger for AHU-classified records (16.8%) than for chiller-classified records (15.4%). This study's code additionally computed the Pearson correlation between a workload–cooling mode interaction term (Server_Workload_pct × Cooling_Mode_Encoded) and Cooling_Unit_Power_kW as an exploratory check. While this correlation provides directional evidence, it does not constitute a formal interaction test. A proper test — specifically a two-way ANOVA with a cooling system × workload level interaction term, or a regression model including the product term with a significance test of its coefficient — was not performed in this study. The parallel bands visible in Figure 2 are visually more consistent with a system-type main effect than with a moderated workload–cooling relationship. This grouped analysis is therefore presented strictly as a descriptive, exploratory observation and should not be interpreted as evidence of a statistically confirmed interaction effect.

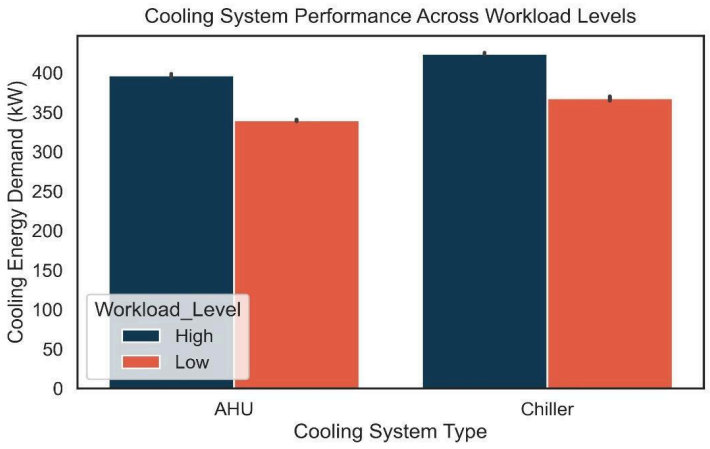

Figure 5. Mean cooling energy demand (kW) by cooling system type and workload level (Low/High split at workload median). Both system types show higher cooling energy under high workload. This is a descriptive grouped comparison only; no formal interaction test was performed.

Feature Correlation Matrix

Figure 6 presents the Pearson correlation matrix across key numerical variables in the dataset. Three structural patterns are evident. First, Server_Workload_pct, Inlet_Temperature_C, and Cooling_Unit_Power_kW form a densely intercorrelated cluster ($r > 0.8$ between each pair). This intercorrelation reflects their shared physical basis: higher computational activity simultaneously increases server heat generation, elevates inlet air temperature, and increases cooling unit power draw. This structural collinearity was the primary justification for excluding Inlet_Temperature_C from the predictive model *(see **Feature Selection and Multicollinearity Management**, above)*. Second, Cooling_Mode_Encoded shows moderate positive correlation with Cooling_Unit_Power_kW ($r \approx 0.79$, 95% CI: 0.775–0.804), consistent with the system-type energy offset documented under ***Chiller-Classified Records Show Substantially Higher Mean Cooling Energy**, above*. Third, Ambient_Temperature_C generally shows weak linear associations with the other variables, with the strongest observed correlation being a small positive association with total_waste_heat_MW ($r = 0.13$).

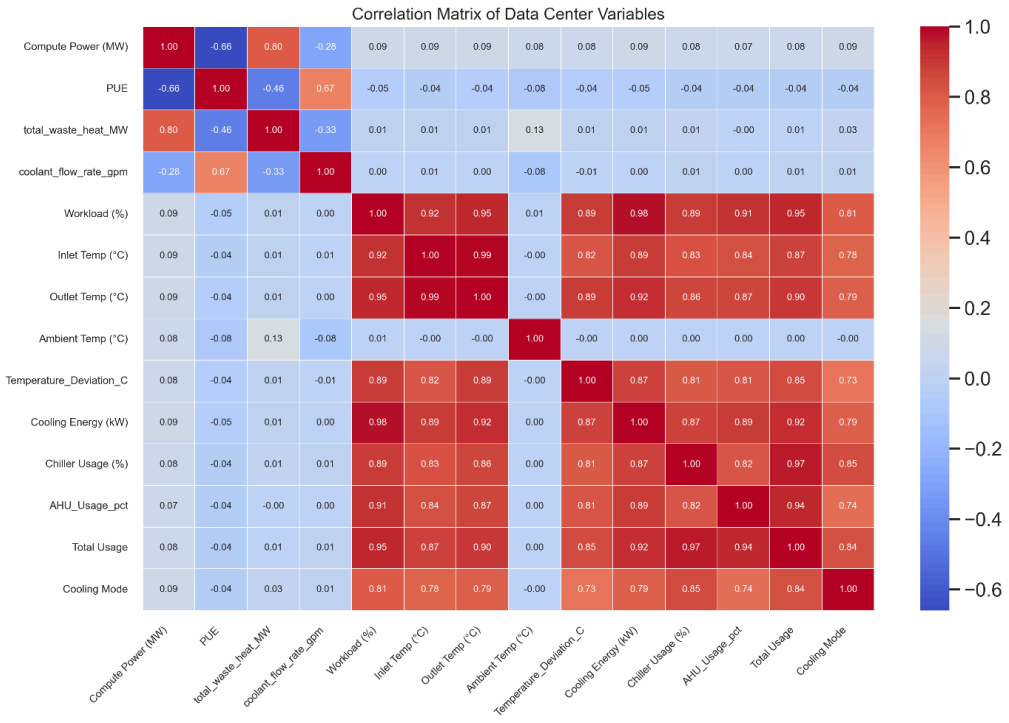

Figure 6. Pearson correlation matrix of key dataset variables. The workload cluster (Workload %, Inlet Temp, Cooling Energy) shows $r > 0.8$ intercorrelations, motivating the exclusion of Inlet Temp from the model. Ambient Temperature shows near-zero correlations with most variables with a small exception of a small positive association with total_waste_heat_MW. Several variables shown in Figure 6 are included for exploratory correlation context and were not used as model features. Table 5 defines these.

Table 5.

Variable	Definition	Unit
Compute_Power_MW	Compute power draw of the HPC/server system	MW
PUE	Power Usage Effectiveness — total facility power ÷ IT equipment power	Ratio
total_waste_heat_MW	Total heat rejected by the cooling system	MW
coolant_flow_rate_gpm	Liquid coolant flow rate through the chiller loop	GPM
Outlet_Temp	Air temperature at server/rack outlet	°C
Temperature_Deviation_C	Deviation from the facility's target temperature setpoint	°C
AHU_Usage_pct	Percentage utilisation of the air handling unit	%
Total_Usage	Combined AHU + chiller utilisation	%

Predictive Model Performance

Table 6 presents model performance on the held-out 20% test set. The final three-feature multivariate model achieved $R^2 = 0.954$ and MAE = 7.87 kW. The workload-only baseline model achieved $R^2 = 0.954$ on the identical test split. The marginal contribution of the two additional features (Ambient_Temperature_C and Cooling_Mode_Encoded) is $\Delta R^2 \approx 0.000$, meaning that ambient temperature and cooling system classification together provide no meaningful additional predictive accuracy beyond workload alone, within this dataset.

Table 6. *Predictive model performance on 20% held-out test set (random seed = 42). Target variable: Cooling_Unit_Power_kW.*

Model	Features Used	R^2 (Test)	MAE (kW)
Workload-Only Baseline	Server_Workload_pct only	0.954	—
Multivariate Model (final)	Workload + Ambient Temp + Cooling Mode	0.954	7.87
Marginal Improvement	—	$\Delta \approx 0.000$	—

The R^2 of 0.954 should be interpreted with appropriate precision. It indicates that the model statistically accounts for approximately 95.4% of the variance in cooling unit power on the test set, given the assumptions of the linear specification. It does not indicate that the model is complete, that all operationally relevant variables have been identified, or that predictions will generalise reliably to facilities, hardware generations, or workload profiles outside the range represented in the merged training data. The remaining unexplained variance (~4.6%) represents genuine real-world complexity: measurement noise, unmodelled variables such as cooling system age and airflow configuration, and the cross-source heterogeneity inherent in the merged dataset.

Although both models round to $R^2 = 0.954$ at three decimal places, their unrounded test-set R^2 values were 0.95362 for the workload-only baseline and 0.95358 for the full multivariate model, giving a difference of −0.000043 (full model minus workload-only baseline). Thus, the apparent equality reflects rounding rather than exact model equivalence. The very small difference indicates that, within this dataset, adding Ambient_Temperature_C and Cooling_Mode_Encoded produces virtually no change in held-out predictive performance beyond Server_Workload_pct alone. This finding is consistent with the hierarchical regression results reported under ***Model Performance and the Sufficiency of Workload****, below.*

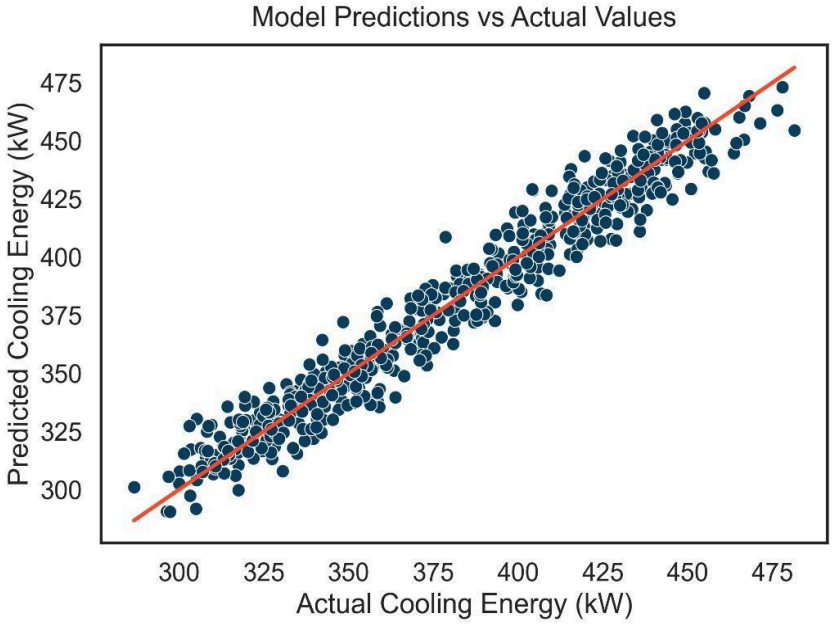

Figure 7. Actual vs. predicted cooling unit power (kW) on the held-out test set. Data points align closely with the 45-degree reference line, with no systematic deviation across the full prediction range. $R^2 = 0.954$, MAE = 7.87 kW.

Residual Diagnostics

Figure 8 presents the residual plot of the final multivariate model. Residuals are approximately symmetric and randomly distributed around zero across the full range of predicted values (approximately 285–490 kW), with the large majority falling within ±25 kW. No fan-shaped spread, systematic curvature, or clustering pattern is evident. A small number of outliers at the prediction extremes do not follow any consistent direction. These diagnostics support the appropriateness of the linear regression specification for this dataset and are consistent with the model assumptions of homoscedasticity and zero-mean residuals.

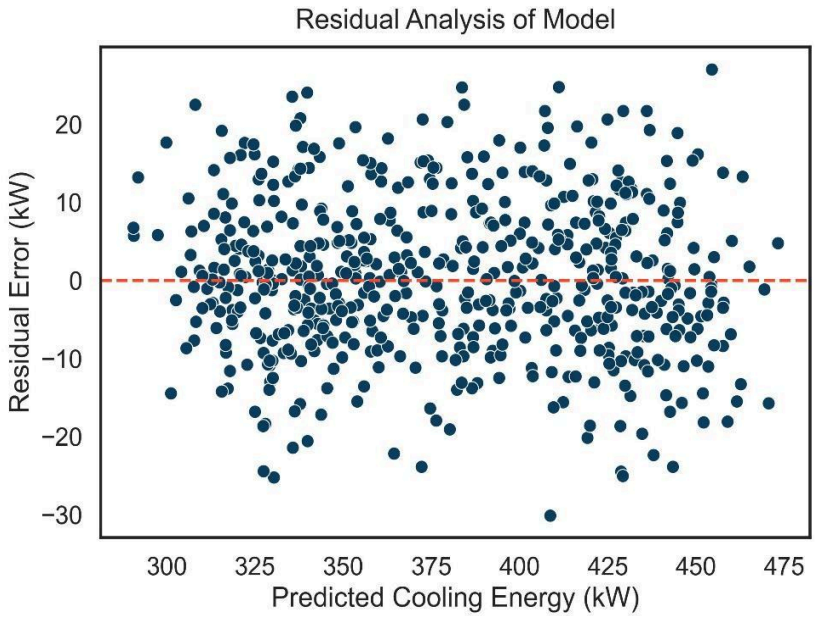

Figure 8. Residual plot of the final multivariate regression model. Residuals are randomly distributed around zero (dashed line) across the full predicted value range, supporting the assumptions of linearity and homoscedasticity.

Multicollinearity Analysis and Feature Coefficients

Table 6 presents standardised regression coefficients, 95% confidence intervals, and VIF scores for the final model, alongside the VIF value previously observed for the excluded Inlet_Temperature_C. In the final three-feature specification, all retained features have VIF values below 3.0, indicating low multicollinearity among the predictors[15]. The standardised coefficient for Server_Workload_pct (β = 0.976, 95% CI: 0.962–0.989) was substantially larger in magnitude than those of Cooling_Mode_Encoded (β = −0.0005, 95% CI: −0.014–0.013) and Ambient_Temperature_C (β = −0.0029, 95% CI: −0.011–0.005). This indicates that server workload accounts for the strongest linear association with cooling-unit power within the fitted model. The coefficients for the latter two features were close to zero, with confidence intervals spanning zero, indicating little evidence of an independent linear association with cooling-unit power after accounting for the other predictors.

Table 7. *Standardised regression coefficients, 95% confidence intervals and VIF scores for the final model. All retained features have VIF < 3.0.*

Feature	Std. β	VIF	Interpretation	95% CI
Server_Workload_pct	0.976	2.8022	Primary predictor; acceptable VIF	0.962–0.989
Cooling_Mode_Encoded	−0.00048	2.8025	Near-zero association; acceptable VIF	−0.014–0.013

Ambient_ Temperature_C	−0.00289	1.000	Near-zero association; CI includes zero	−0.011–0.005
Inlet_Temperature _C (excluded)	N/A	6.51	Excluded due to multicollinearity	N/A

DISCUSSION

Workload Dominance and Real-World Interpretation

The central finding of this study — that server workload percentage alone statistically explains approximately 95% of the variance in cooling unit power — has clear practical significance. In a data centre, the physical chain of events is well-understood: when servers execute more computational operations, processors and memory modules draw more electrical power; that power is dissipated as heat; and the cooling system must remove that heat by drawing additional power in proportion to the thermal load. The linear and near-proportional nature of this relationship, confirmed by the tight residual distribution in Figure 8, suggests that the cooling systems represented in this dataset operate in a regime where their power consumption is approximately proportional to server heat output[3] — consistent with variable-speed fan or proportional pump control strategies.

The practical implication is significant: in facilities with similar operational profiles, server workload data — which is typically logged in real time by infrastructure management systems — may be sufficient to forecast cooling energy demand with high accuracy (~7.87 kW MAE) without requiring additional environmental sensors. Workload-aware pre-conditioning, in which cooling capacity is adjusted in anticipation of scheduled computational jobs, becomes a tractable strategy when this linear relationship holds. This finding is consistent with prior research demonstrating that workload-aware scheduling can yield measurable energy savings.[13]

AHU vs. Chiller: Interpretation of the Energy Difference

The 20.9% higher mean cooling energy associated with chiller-classified records (72.0 kW mean difference; Cohen's d = −2.534) is a large and statistically robust observation within this dataset. The directional result is consistent with engineering expectations: chiller systems maintain active compressor,[5] condenser, and pump loads that are approximately constant regardless of the server heat load they are removing, whereas AHU systems using variable-speed fans can scale their energy consumption more proportionally to instantaneous demand. The chiller's fixed infrastructure overhead would be expected to manifest as a baseline upward shift in cooling energy — precisely the pattern observed in Figure 2.

However, this study cannot establish that cooling system architecture is the direct or sole cause of this energy difference. The median-based classification groups records by the level of chiller utilisation, which is likely correlated with other facility characteristics not captured in the dataset, such as server rack density, facility age, or workload type. Additionally, the classification boundary is constructed from the

data itself, meaning the two groups may not cleanly represent distinct engineering paradigms. The finding is therefore reported as a robust statistical association that is directionally meaningful, whilst explicitly noting that causal attribution is not supported by the observational study design.

Ambient Temperature: A Validated Null Finding

The near-zero correlation between ambient outdoor temperature and cooling unit power ($r \approx 0$) is consistent with established data centre engineering principles. Well-insulated, closed-loop cooling systems are specifically designed to decouple indoor thermal management from outdoor climate variation. This null result serves as a positive validation of the dataset's internal consistency: if indoor cooling demand were strongly correlated with outdoor temperature in a standard data centre, it would suggest either a design flaw in the cooling system or a data collection error. The absence of this correlation is therefore reassuring rather than unexpected.

The result has a practical implication: facility operators cannot rely on seasonal or diurnal ambient temperature variation to passively reduce cooling energy costs in standard closed-loop systems. Cooling efficiency improvements in such facilities must come from internal operational changes — such as workload scheduling, airflow optimisation, or cooling set-point adjustment — rather than from favourable outdoor conditions. Facilities equipped with free-air economiser systems, which intentionally couple indoor and outdoor thermal environments,[16] would be expected to show a stronger ambient temperature effect and should be studied separately.

Model Performance and the Sufficiency of Workload

The hierarchical regression provides additional context for the contrast between the large unadjusted cooling-mode difference and the near-zero adjusted regression coefficient. Cooling Mode alone explained 61.63% of the variance in cooling-unit power ($R^2 = 0.61632$). Adding Ambient_Temperature_C produced virtually no additional explanatory power ($\Delta R^2 = 0.000034$, $p = 0.614$), with R^2 increasing to 0.61636. In contrast, adding Server_Workload_pct increased R^2 by 0.33485 ($p < 0.001$), resulting in $R^2 = 0.95120$ for the full hierarchical model.

Table 8. Hierarchical regression results showing the incremental contribution of Ambient_Temperature_C and Server_Workload_pct to the prediction of Cooling_Unit_Power_kW.

Model	Predictors	R^2	ΔR^2	F-change	p-change
1	Cooling Mode	0.61632	—	—	—
2	Cooling Mode + Ambient Temperature	0.61636	0.000034	0.255	0.614
3	Cooling Mode + Ambient Temperature + Server Workload	0.95120	0.33485	19,577.210	<0.001

These results indicate that the large cooling-mode difference observed in the unadjusted comparison overlaps substantially with variation captured by server workload. After accounting for Ambient_Temperature_C and Server_Workload_pct, the standardized coefficient for Cooling_Mode_Encoded was near zero ($\beta = -0.00048$, 95% CI: −0.0147 to 0.0127), with $p = 0.889$,

indicating little evidence of a unique linear association with cooling-unit power in the final model. Thus, the large group-level difference observed in the t-test should not be interpreted as evidence that cooling mode explains the same magnitude of variation after workload is controlled.

Distributional Heterogeneity as a Confound

A limitation that warrants explicit discussion is the potential for distributional heterogeneity across the two merged source datasets to artificially inflate R^2 or distort group comparisons. If the Kaggle simulated datasets exhibit tighter, more regular relationships between workload and cooling power than real operational data — as is common in synthetic datasets — the merged dataset's high R^2 may partially reflect the statistical regularity of simulated data rather than the strength of the physical relationship in real facilities. Similarly, if the Frontier HPC source systematically differs in energy range or workload profile from the Kaggle sources, the between-source heterogeneity could add variance that a within-source analysis would not capture. The absence of a formal inter-source distributional alignment test — such as a domain adversarial validation or per-source stratified analysis — is acknowledged as the most significant methodological limitation of this study. Future work should include such validation before reporting predictive performance metrics from merged heterogeneous datasets.

CONCLUSION

This study applied statistical analysis and linear regression modelling to a merged data centre operational dataset of over 2,500 records to characterise associations between operational variables and cooling unit power consumption. The study's principal findings are as follows:

- Server workload percentage is strongly and linearly associated with cooling unit power consumption (bivariate $R^2 = 0.95$). A workload-only linear regression model achieves $R^2 = 0.95362$ and MAE = 7.88 kW on a held-out test set, while the three-feature multivariate model achieves $R^2 = 0.95358$. The small R^2 difference of -0.000043 indicates that the additional features produce little change in overall explained variance. This indicates that, within this dataset, server workload alone captures most of the explained variation in cooling-unit power for short-term forecasting purposes.
- Chiller-classified records are associated with mean cooling energy approximately 20.9% higher than AHU-classified records (419.8 kW vs. 347.9 kW; $p < 0.001$; Cohen's $d = -2.534$, large effect). This association is statistically robust and directionally consistent with the known fixed-overhead energy profile of refrigerant-based chiller infrastructure. Causal attribution is not claimed.
- Ambient outdoor temperature shows negligible linear association with indoor cooling unit power ($r \approx 0$) across the sampled range of 3–21°C, consistent with closed-loop cooling system design principles. This null result validates dataset internal consistency and establishes that ambient temperature variation is not a useful predictor of cooling energy in these facility types.
- Removing Inlet_Temperature_C from the predictive model on structural collinearity grounds ($r > 0.8$ with Server_Workload_pct) reduced VIF values for all retained features below 3.0 without

affecting model performance, confirming that the removal was both statistically appropriate and practically costless.

These findings are associative and observational. They support workload-aware scheduling and thermal management as the most actionable levers for cooling energy optimisation in AI-loaded data centres. The specific regression model developed here should not be treated as ready for direct operational deployment without validation against independently collected single-facility datasets.

LIMITATIONS

Several limitations qualify the scope and generalisability of the findings reported in this study:

- Dataset heterogeneity and potential distributional inflation: The merged dataset combines Kaggle datasets (which may be fully or partially simulated) with real operational data from the Frontier HPC system. No formal inter-source distributional consistency test was performed. The high observed R^2 may partly reflect the statistical regularity characteristic of simulated datasets rather than the strength of the physical relationship in real facilities. Future work should apply per-source stratified analysis or domain adversarial validation to assess whether the reported associations hold within each source independently.
- Chiller_Usage_pct availability and selection bias: The Chiller_Usage_pct variable was present in Kaggle Dataset but not in the Frontier HPC source. Records lacking this variable were excluded during preprocessing, meaning the final analytical dataset likely underrepresents the Frontier HPC source and may be biased towards the operational characteristics of the Kaggle datasets.
- Binary classification boundary noise: The AHU/Chiller classification uses the dataset median of Chiller_Usage_pct as a threshold. Records near the boundary may be misclassified, introducing noise into group comparisons. A continuous chiller utilisation predictor would avoid this artefact.
- No causal inference: All analyses are observational. Statistical significance and large effect sizes quantify the strength of associations within this dataset; they do not establish that workload or cooling system type directly cause differences in cooling energy. Unmeasured confounders — including facility age, rack density, hardware generation, and operational protocol — may account for some or all of the observed differences.
- Limited ambient temperature range: The dataset's ambient temperature range (approximately 3–21°C) does not include high-temperature conditions above 30°C. The null result for ambient temperature should not be generalised to facilities in hotter climates or to systems with free-air economiser modes.
- Linear model limitations: OLS linear regression assumes linearity, homoscedasticity, and independence of observations. Whilst residual diagnostics support these assumptions within the dataset's range, the model does not represent thermal hysteresis, time-lagged cooling responses, control-system feedback dynamics, or nonlinear behaviour at high-load extremes.
- No temporal structure: The dataset is treated as a cross-sectional sample. Sequential autocorrelation in workload and cooling demand over time — which is characteristic of real operational data centres — is not modelled.

FUTURE WORK

The following extensions are recommended to address the limitations identified in this study and to advance the research agenda:

- Inter-source distributional validation: Before reporting predictive metrics from merged heterogeneous datasets, future studies should apply formal distributional alignment tests (e.g., KS tests per variable across sources, adversarial domain classifiers) to quantify the degree to which the merged dataset's statistical properties reflect real operational variation rather than source-specific simulation artefacts.
- Formal interaction testing: A two-way ANOVA with a cooling system × workload level interaction term, or a regression model including the Cooling_Mode × Server_Workload_pct product term with a significance test of its coefficient, should be applied to test whether the 1.4 percentage-point difference in workload sensitivity between system types constitutes a statistically confirmed moderation effect.
- Temporal and time-of-day features: Incorporating timestamps, time-of-day, and day-of-week features would allow modelling of predictable workload variation patterns — such as diurnal peaks and weekend troughs — and the thermal lag between workload changes and cooling system response. These additions would improve the practical utility of the model for real-time pre-conditioning applications.
- Continuous chiller utilisation as a predictor: Replacing the binary Cooling_Mode variable with the continuous Chiller_Usage_pct as a regression feature would avoid classification boundary artefacts and could reveal a dose–response relationship between chiller utilisation level and cooling energy.
- Nonlinear modelling: Gradient boosting, random forests, or recurrent neural networks may capture nonlinear or temporally structured workload–cooling relationships that are not adequately represented by the linear model, particularly at high-load operational extremes.
- Water Usage Effectiveness: Extending the analysis to include Water Usage Effectiveness (WUE) alongside energy metrics would provide a complementary sustainability perspective, particularly relevant for chiller-based systems that often involve evaporative cooling towers.
- Multi-site validated datasets: Applying the analytical pipeline to independently collected, single-facility datasets from multiple geographic regions and hardware generations would establish whether the associations reported here generalise across data centre types, climate zones, and workload profiles.

ACKNOWLEDGEMENTS

The authors acknowledge the publicly available datasets provided by the Kaggle data science community, which formed the empirical foundation of this research. The authors also recognise the open-source scientific Python ecosystem, including NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn, without which this analysis would not have been possible. No external funding was received for this study. The

authors initiated and completed this research paper during their 11th-grade academic year at Shiksha Deep B.S.S., Nepal.

AUTHOR CONTRIBUTIONS

Shrey Sharma led the conceptualisation and overall design of the study, including definition of the research problem, dataset construction, data collection, preprocessing, feature engineering, statistical analysis, predictive modelling, and generation of all figures and tables. He also contributed to integrating and refining technical content. Yubsan Basnet contributed to the literature review, data collection, methodology design, interpretation of results, and was primarily responsible for manuscript writing including all sections as well as review and editing to improve clarity, structure, and presentation. Both authors critically reviewed the final manuscript and approved it for submission.

REFERENCES AND NOTES

1. Arregoces, M. and Portolani, M. *Data Center Fundamentals.* Indianapolis, IN: Cisco Press, 2003.

2. International Energy Agency. "Data Centres and Data Transmission Networks." *IEA,* 2024. Accessed April 2024. https://www.iea.org/reports/data-centres-and-data-transmission-networks.

3. Ahmed, A., Khan, A. A. and Hussain, M. I. "Data Center Cooling: A Critical Review of Techniques, Challenges, and Energy-Saving Solutions." *Renewable and Sustainable Energy Reviews* 134 (2020).

4. The Green Grid. "PUE: A Comprehensive Examination of the Metric." Beaverton: The Green Grid, 2007.

5. Shehabi, A. et al. *United States Data Center Energy Usage Report.* Lawrence Berkeley National Laboratory, LBNL-1005775, 2016.

6. Masanet, E., Shehabi, A., Lei, N., Smith, S. and Koomey, J. "Recalibrating Global Data Center Energy-Use Estimates." *Science* 367, no. 6481 (2020): 984–986.

7. De Vries, A. "The Growing Energy Footprint of Artificial Intelligence." *Joule* 7, no. 10 (2023): 2191–2194.

8. Meta Platforms Inc. *Data Center Efficiency Report.* Meta Sustainability, 2024.

9. Hewlett Packard Enterprise. "Liquid Cooling for HPC and AI Workloads." HPE Technical White Paper, 2024.

10. ASHRAE. *Thermal Guidelines for Data Processing Environments.* 4th edn. Atlanta, GA: ASHRAE, 2015.

11. Whitehead, B., Andrews, D. and Shah, A. "The Life Cycle Assessment of a UK Data Centre." *International Journal of Life Cycle Assessment* 20, no. 3 (2014): 332–349.

12. Dayarathna, M., Wen, Y. and Fan, R. "Data Center Energy Consumption Modeling: A Survey." *IEEE Communications Surveys and Tutorials* 18, no. 1 (2016): 732–794.

13. Pei, J. et al. "Workload-Aware Scheduling for Energy Efficiency in Data Centers." *IEEE Transactions on Services Computing* 17, no. 2 (2024): 488–501.

14. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

15. O'Brien, R. M. "A Caution Regarding Rules of Thumb for Variance Inflation Factors." *Quality & Quantity* 41, no. 5 (2007): 673–690.

16. Kang, S., Moon, J. M. and Jeong, J. W. "Energy Efficiency of Air-Side Economizers in Data Center Cooling Systems." *Applied Thermal Engineering* 236 (2024): 121558.