

AI-DRIVEN PREDICTIVE AUTO-SCALING FOR ENERGY-EFFICIENT CLOUD DATA CENTRES: A BIBLIOMETRIC ANALYSIS AND CONCEPTUAL FRAMEWORK

Sowmiya M¹, S Asha²

¹Sethu Institute of Technology, Tamil Nadu, India. Email ID: sowmiya2727@gmail.com

²Sethu Institute of Technology, Tamil Nadu, India. Email ID: sasha@sethu.ac.in

Abstract—Rapid growth in cloud computing, artificial intelligence, Internet of Things applications and data-intensive services has intensified concern about the energy consumed by cloud data centres. Dynamic workloads make static provisioning inefficient, while purely reactive auto-scaling can respond only after demand changes have occurred. This study maps the recent knowledge structure of AI-driven resource management and predictive auto-scaling research and develops a conceptual framework linking workload prediction accuracy, resource utilisation efficiency and intelligent auto-scaling decision effectiveness to the energy efficiency of cloud data centres. A bibliometric analysis was conducted using 1,312 Scopus-indexed journal articles covering 2020–2027 and analysed with the Bibliometrix/Biblioshiny environment in R. The corpus spans 337 sources and 4,008 authors, with 33.38% international co-authorship and an average of 18.2 citations per document. Scientific production increased from 131 articles in 2020 to 248 in 2025, while the lower 2026 count reflects an incomplete publication year and the single 2027 record is a future-dated/early-assigned publication. IEEE Access was the dominant source with 172 documents. Keyword and thematic analyses reveal four closely connected streams: cloud computing and quality of service, resource allocation and resource management, machine/deep/reinforcement learning for prediction and control, and green/energy-efficient computing. Highly cited work further demonstrates the importance of resource management, workload prediction, deep reinforcement learning, load balancing and energy-efficient virtual-machine allocation. Based on the bibliometric synthesis, the study proposes an integrated framework in which accurate workload forecasting supports efficient resource utilisation and more effective scaling decisions, which together are expected to improve energy efficiency while maintaining service quality. The paper contributes a structured map of the field, identifies an integration gap across prediction, utilisation, scaling and energy outcomes, and sets out testable propositions for subsequent empirical or experimental validation.

Keywords: Cloud Computing; Cloud Data Centres; Bibliometric Analysis; Workload Prediction; Resource Utilisation; Auto-Scaling; Energy Efficiency; Machine Learning; Reinforcement Learning.

1. Introduction

Cloud computing has become a foundational infrastructure for contemporary digital services because it provides elastic access to computing, storage and networking resources. The same elasticity that makes cloud computing attractive, however, creates a persistent resource-management problem: capacity must be continuously matched to workloads that are time-varying, bursty and often difficult to predict. When resources are provisioned too conservatively, applications may suffer latency, throughput degradation or service-level agreement (SLA) violations; when resources are provisioned too aggressively, servers and virtual machines remain under-utilised while still consuming electrical power. Resource management is therefore not only a performance problem but also an energy-efficiency problem.

Conventional auto-scaling mechanisms commonly rely on threshold-based rules applied to indicators such as CPU utilisation, memory use or request rates. Such mechanisms are relatively simple and transparent, but their reactive nature may introduce a lag between workload change and resource adjustment. Earlier reviews of elastic cloud applications already showed that determining the appropriate amount of resources while satisfying service-level objectives is a central challenge in auto-scaling (Lorido-Botran et al., 2014). More recent work on Kubernetes similarly demonstrates that scaling behaviour depends strongly on the metrics used and on the configuration of the scaling mechanism (Nguyen et al., 2020). These observations have encouraged a shift from purely reactive rules toward predictive and learning-based resource management.

Machine learning has consequently become increasingly important in cloud resource provisioning. Le Duc et al. (2020) show that learning-based methods can support more reliable provisioning across edge-cloud environments, while Chen et al. (2020) demonstrate the relevance of deep learning for high-dimensional and highly variable cloud-workload prediction. Xu et al. (2022) further develop deep neural-network-based multivariate workload prediction, highlighting the growing role of data-driven forecasting in anticipation of future resource demand. In parallel, deep reinforcement learning has been applied to adaptive resource allocation in cloud data centres to jointly respond to dynamic system states and performance-energy objectives (Chen et al., 2022).

Energy-aware resource allocation forms a second major research stream. Zhou et al. (2021), for example, address energy-efficient virtual-machine allocation in cloud data centres and note that energy optimisation should account for fluctuating load and resource provisioning jointly. Load balancing is similarly linked to avoiding overloaded and underloaded hosts and to improving overall cloud performance (Shafiq et al., 2021). Across the broader cloud-edge literature, resource management remains a major challenge because computing environments are heterogeneous, dynamic and increasingly distributed (Hong & Varghese, 2019).

Despite these advances, the literature remains dispersed across several partially overlapping research streams. Workload-prediction studies frequently emphasise forecast accuracy; resource-allocation studies often prioritise utilisation, scheduling and consolidation; auto-scaling studies focus on elasticity, control and QoS; and green-cloud studies prioritise power or energy reduction. What is less clearly integrated is the relationship among (1) the accuracy with which workload demand is anticipated, (2) the efficiency with which provisioned resources are utilised, (3) the effectiveness of intelligent scaling decisions and (4) the resulting energy efficiency of cloud data centres. This fragmentation motivates both a bibliometric mapping of the field and a conceptual synthesis that links these dimensions.

Bibliometric analysis is suitable for this purpose because it can systematically examine large bodies of literature, reveal influential sources and documents, and map intellectual and conceptual structures through citation, co-citation, bibliographic and co-word techniques (Donthu et al., 2021). Bibliometrix provides an integrated R-based environment for such science-mapping analyses (Aria & Cuccurullo, 2017). Accordingly, this study analyses a Scopus corpus of 1,312 articles and uses performance analysis, source analysis, citation analysis, keyword co-occurrence, thematic mapping and collaboration analysis to identify the knowledge structure of AI-driven predictive auto-scaling and energy-efficient cloud computing.

1.1 Research objectives

RO1. To examine the growth, publication structure and influential contributions in research related to AI-driven resource management, auto-scaling and energy-efficient cloud computing.

RO2. To identify the dominant conceptual and thematic structures connecting workload prediction, resource allocation/utilisation, intelligent scaling and energy efficiency.

RO3. To identify research gaps in the integration of workload prediction accuracy, resource utilisation efficiency and intelligent auto-scaling decision effectiveness.

RO4. To develop a conceptual AI-driven predictive auto-scaling framework in which workload prediction accuracy, resource utilisation efficiency and intelligent auto-scaling decision effectiveness are linked to the energy efficiency of cloud data centres.

1.2 Research questions

RQ1. How has research on AI-driven resource management, predictive auto-scaling and energy-efficient cloud computing evolved during the study period?

RQ2. Which sources and documents are most influential in the analysed literature?

RQ3. What major conceptual and thematic clusters structure the research field?

RQ4. How do workload prediction, resource utilisation, intelligent scaling and energy efficiency intersect in the existing knowledge base?

RQ5. What research gap supports an integrated predictive auto-scaling framework for energy-efficient cloud data centres?

2. Conceptual Background

2.1 Workload prediction accuracy

Workload prediction accuracy refers to the extent to which a forecasting model can estimate near-future demand with sufficiently low error to support proactive resource decisions. In cloud environments, workload is multidimensional and can involve request rate, CPU demand, memory pressure, storage I/O, network traffic and application-specific indicators. Chen et al. (2020) emphasise the difficulty of forecasting high-dimensional and highly variable cloud workloads, while Xu et al. (2022) show the value of multivariate deep-learning approaches for workload prediction. From an energy perspective, the relevance of accuracy is operational rather than merely statistical: over-prediction can trigger unnecessary capacity and energy use, whereas under-prediction can produce late scale-out actions, overload and SLA degradation. Consequently, prediction accuracy is treated in this study as the first antecedent of energy-efficient predictive auto-scaling.

2.2 Resource utilisation efficiency

Resource utilisation efficiency captures how effectively provisioned computing resources are converted into useful workload processing while minimising idle or stranded capacity. Cloud resource-management research repeatedly focuses on allocation, provisioning, consolidation, scheduling and load balancing. Hong and Varghese (2019) identify resource management as a core challenge in distributed cloud-edge environments, and Shafiq et al. (2021) highlight the importance of balancing tasks to avoid overloaded or underloaded infrastructure. Energy-efficient VM allocation research further suggests that the number and loading of active physical machines are directly relevant to energy consumption (Zhou et al., 2021). Resource utilisation efficiency therefore acts as the operational bridge between resource demand and infrastructure energy use.

2.3 Intelligent auto-scaling decision effectiveness

Intelligent auto-scaling decision effectiveness refers to the quality of decisions concerning when scaling should occur, in which direction it should occur, and how much capacity should be added, removed or reconfigured. Traditional threshold-based scaling largely reacts to observed utilisation, whereas intelligent mechanisms may combine workload forecasts, system state, cost, SLA constraints and learned policies. Nguyen et al. (2020) show that Kubernetes horizontal pod auto-scaling behaviour depends on monitored metrics and configuration, while Chen et al. (2022) demonstrate how actor-critic deep reinforcement learning can adapt allocation decisions to dynamic cloud states. Effective scaling is therefore multidimensional: it requires timely, stable and proportionate decisions that avoid both excessive capacity and performance degradation.

2.4 Energy efficiency of cloud data centres

Energy efficiency of cloud data centres is the outcome construct in the proposed framework. At the resource-management level, it refers to accomplishing a required amount of computing work or service quality with lower energy consumption. Prior studies use measures such as total energy consumed, energy per completed task, power consumption, number of active physical machines, energy-delay trade-offs and related energy-aware objectives. Zhou et al. (2021) illustrate the connection between VM allocation and data-centre energy efficiency, whereas Chen et al. (2022) explicitly frame adaptive resource allocation in relation to both latency and energy efficiency. In the present framework, energy efficiency is considered jointly with service quality so that energy reduction is not achieved by simply withholding resources and degrading application performance.

2.5 Proposed constructs

Construct	Role	Conceptual meaning	Indicative measures for later validation
Workload Prediction Accuracy	IV1	Accuracy with which near-future workload demand is forecast for proactive scaling.	MAE, RMSE, MAPE, forecast bias, prediction horizon accuracy
Resource Utilisation Efficiency	IV2	Extent to which allocated CPU, memory, VM/container and other	Average utilisation, idle-resource ratio,

		resources are productively used with low idle capacity.	consolidation efficiency, resource wastage
Intelligent Auto-Scaling Decision Effectiveness	IV3	Timeliness, correctness and stability of scale-up/down or scale-out/in decisions under dynamic workload.	Scaling latency, unnecessary scaling actions, oscillation rate, SLA-aware decision quality
Energy Efficiency of Cloud Data Centres	DV	Computing/service output achieved relative to energy consumed, while maintaining required service quality.	Total energy, energy/task, power, active hosts, energy-delay product, SLA-compliant energy saving

3. Research Methodology

3.1 Research design

This study adopts a bibliometric research design to quantitatively map the development, influence, conceptual structure and thematic organisation of the literature. Bibliometric analysis is appropriate when a field contains a sufficiently large body of publications and the research aim is to examine patterns of scientific production and relationships among documents, sources, authors and concepts (Donthu et al., 2021). The analysis was implemented using Bibliometrix/Biblioshiny in R, an open-source environment designed for comprehensive science mapping (Aria & Cuccurullo, 2017). The bibliometric stage is used to identify and synthesise the structure of the field; it does not by itself establish causal effects among the proposed variables. The causal relationships in the conceptual framework are therefore presented as propositions for subsequent empirical or experimental validation.

3.2 Database and search strategy

Scopus was selected as the bibliographic database because the study focuses on computer science and engineering literature and requires structured bibliographic fields including titles, abstracts, keywords, affiliations, references and citation counts. The search was designed around four concept blocks: cloud-computing context, resource scaling/provisioning, AI-based prediction or learning, and energy/sustainability outcomes. The search structure used for the study was:

TITLE-ABS-KEY(("cloud computing" OR "cloud data center*" OR "cloud data centre*" OR "cloud infrastructure" OR Kubernetes OR container*) AND ("auto-scaling" OR autoscaling OR "predictive scaling" OR "resource provisioning" OR "resource allocation" OR "resource management") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "reinforcement learning" OR "deep reinforcement learning" OR LSTM OR "workload prediction" OR "workload forecasting") AND ("energy efficiency" OR "energy consumption" OR "power consumption" OR "energy-aware" OR sustainability OR "green computing" OR "green cloud computing"))

The exact query should be cross-checked against the saved Scopus search history before submission because search syntax is part of reproducible reporting. The exported dataset supplied for this manuscript contains 1,312 records.

3.3 Screening and eligibility

The Scopus screening settings limited the publication years to 2020–2027, the subject areas to Computer Science and Engineering, the document type to Article, and the publication stage to Final. The Open Access filter was set to “All open access,” limiting the exported corpus to records categorised by Scopus as open access. The final exported corpus comprised 1,312 articles. The Scopus interface reported 1,186 records tagged within Computer Science and 689 within Engineering; these subject-area counts overlap because a publication may belong to more than one subject classification and therefore should not be summed. Because the supplied screening record does not preserve an earlier unfiltered retrieval count or a manual title/abstract exclusion count, this manuscript does not invent a PRISMA-style initial total. Instead, it transparently reports the 1,312-record analytical corpus after the confirmed database filters.

Filter	Applied criterion	Observed Scopus count/interpretation
Year	2020–2027	Final corpus spans 2020–2027; 2027 contains one future-dated record.
Subject area	Computer Science; Engineering	1,186 Computer Science and 689 Engineering tags; categories overlap.
Document type	Article	1,312
Publication stage	Final	1,312
Open access	All open access	1,312
Final exported corpus	Scopus CSV	1,312 records

3.4 Data analysis

The analysis combined performance analysis and science mapping. Performance indicators included annual scientific production, number of sources, authorship, average citations and the most relevant publication outlets. Science-mapping outputs included keyword co-occurrence, thematic mapping, three-field relationships, author productivity patterns, Bradford source concentration and international collaboration. Co-word analysis was used to reveal conceptual proximity among frequently occurring terms, while the thematic map positioned themes according to centrality and density. These techniques are consistent with recommended bibliometric practice in which performance analysis describes productivity and impact, while science mapping examines relationships and the intellectual or conceptual structure of a field (Donthu et al., 2021).

4. Results

4.1 Main characteristics of the bibliometric corpus

The dataset contains 1,312 documents published across 337 sources between 2020 and 2027. A total of 4,008 authors contributed to the corpus, and 71 authors were associated with single-authored documents. International co-authorship accounted for 33.38% of the corpus and the mean number of co-authors per document was 3.91. The database contained 3,196 author keywords and 48,664 cited references. The average document age was 2.72 years and the average citation count was 18.2 per document. These indicators show a large, recent and strongly collaborative research domain.

Indicator	Value
Timespan	2020–2027
Sources	337
Documents	1,312
Authors	4,008
Authors of single-authored documents	71
International co-authorship	33.38%
Co-authors per document	3.91
Author keywords (DE)	3,196
References	48,664
Document average age	2.72 years
Average citations per document	18.2

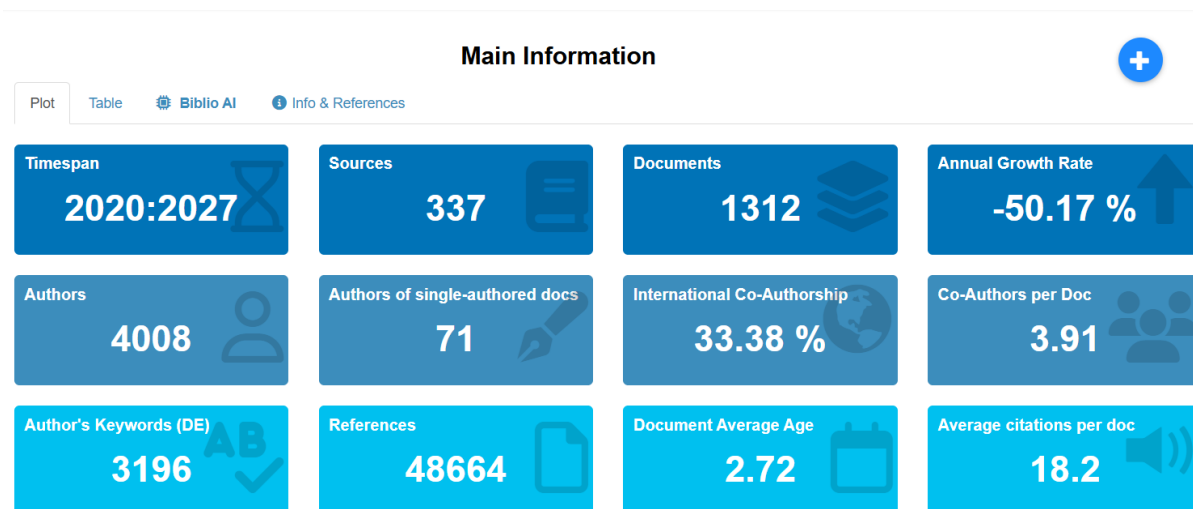


Figure 1. Main bibliometric information generated in Bibliometrix/Biblioshiny.

4.2 Annual scientific production

Annual production increased from 131 articles in 2020 to 141 in 2021 and then rose more sharply to 192 in 2022. Output continued to grow to 207 articles in 2023, 218 in 2024 and a peak of 248 in 2025. The dataset contains 174 articles assigned to 2026 and one article assigned to 2027. Because the database was exported during 2026, the apparent decline after 2025 should not be interpreted as evidence that research interest is contracting. The 2026 publication year is incomplete, while the 2027 record reflects advance publication metadata. For the same reason, the Bibliometrix annual growth-rate value of -50.17% is not substantively meaningful as a field-level trend when the final years are incomplete.

Year	Articles
2020	131
2021	141
2022	192
2023	207
2024	218
2025	248
2026	174
2027	1

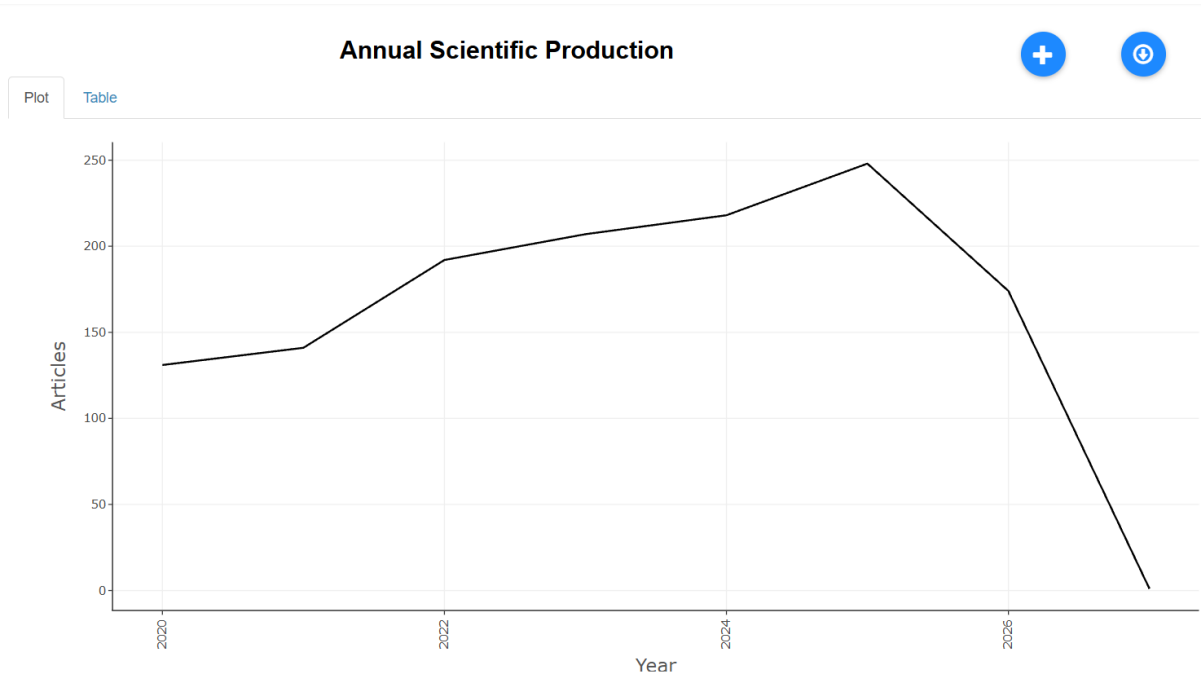


Figure 2. Annual scientific production, 2020–2027.

4.3 Most relevant sources and source concentration

IEEE Access is the most prolific outlet in the dataset with 172 documents, substantially ahead of the International Journal of Advanced Computer Science and Applications (49), Journal of Cloud Computing (46), Computers, Materials & Continua (41), Applied Sciences (33), Sensors (31), Electronics (26), Future Internet (22), IEEE Transactions on Cloud Computing (20) and Cluster Computing (18). The distribution shows that the field is multidisciplinary but anchored in computer science, cloud computing and engineering outlets. Bradford analysis identifies a small core of nine sources, followed by a larger second zone of 57 sources and a long tail of 271 sources, illustrating the concentration of publication output in a limited set of journals alongside broad dispersal across the wider literature.

Rank	Source	Documents
1	IEEE Access	172
2	International Journal of Advanced Computer Science and Applications	49
3	Journal of Cloud Computing	46
4	Computers, Materials and Continua	41
5	Applied Sciences (Switzerland)	33
6	Sensors	31
7	Electronics (Switzerland)	26
8	Future Internet	22
9	IEEE Transactions on Cloud Computing	20
10	Cluster Computing	18

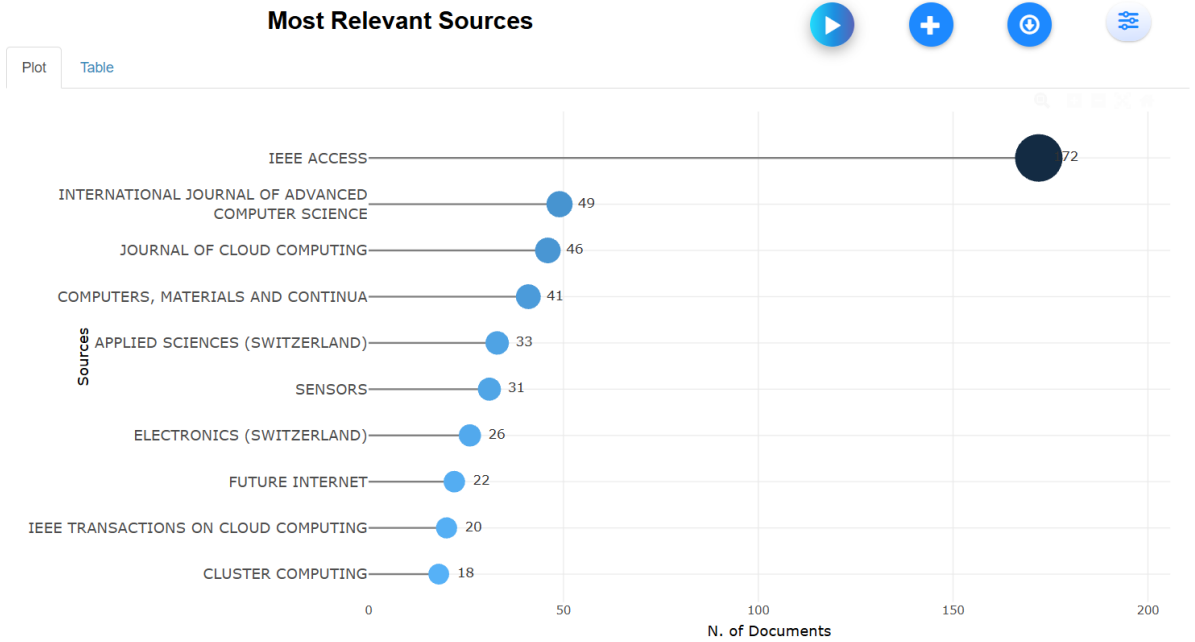


Figure 3. Most relevant publication sources.

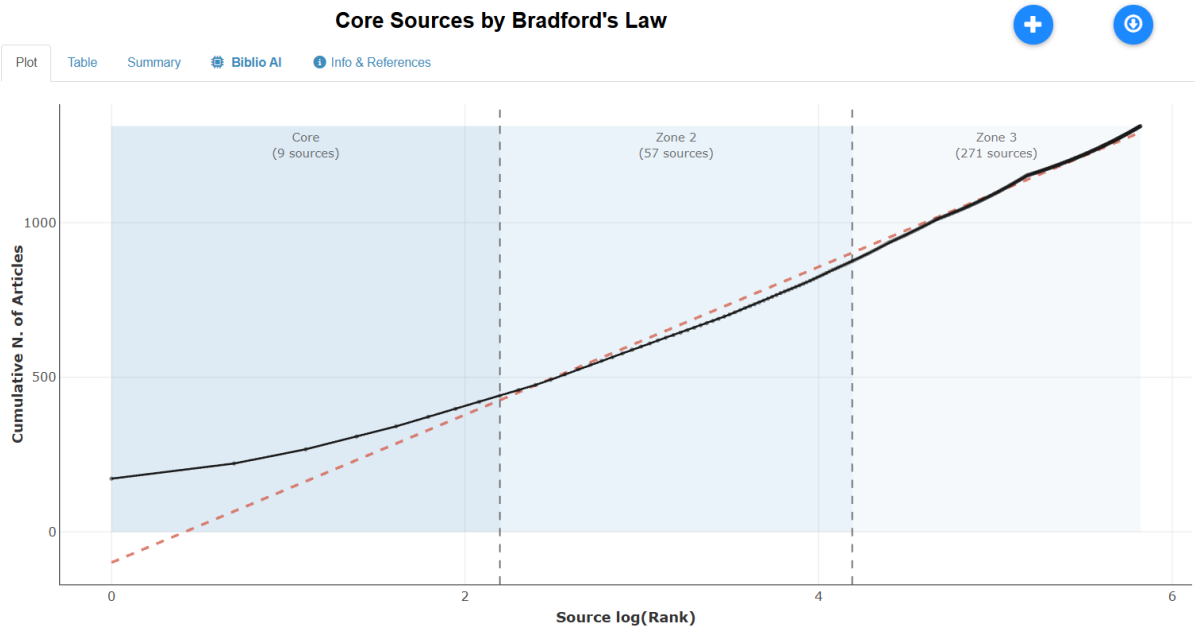


Figure 4. Core sources according to Bradford's Law.

4.4 Most globally cited documents

The citation structure is led by Hong and Varghese's resource-management survey, which accumulated 433 citations in the exported Scopus dataset. Other highly cited works include studies on deep-reinforcement-learning-based resource allocation and task offloading, fog-computing architectures, energy-efficient virtual-machine allocation, machine-learning-based resource provisioning, deep-learning workload prediction and cloud load balancing. This pattern is important for the conceptual framework because the most influential documents do not belong to a single technical

subfield; instead, they collectively connect resource management, prediction, allocation, scheduling and energy considerations.

Rank	Document	Year	Focus	Scopus citations
1	Hong & Varghese: Resource management in fog/Edge computing: A survey on architectures, infrastructure, and algorithms	2020	Resource management	433
2	Alfakih et al.: Task offloading and resource allocation for mobile edge computing by deep reinforcement learning based on SARSA	2020	DRL/resource allocation	330
3	Sabireen & Neelanarayanan: A Review on Fog Computing: Architecture, Fog with IoT, Algorithms and Research Challenges	2021	Fog architecture	252
4	Zhou et al.: AFED-EF: An Energy-Efficient VM Allocation Algorithm for IoT Applications in a Cloud Data Center	2021	Energy-efficient VM allocation	233
5	Le Duc et al.: Machine learning methods for reliable resource provisioning in edge-cloud computing: A survey	2020	ML resource provisioning	215
6	Chen et al.: Towards Accurate Prediction for High-Dimensional and Highly-Variable Cloud Workloads with Deep Learning	2020	Workload prediction	214
7	Shafiq et al.: A Load Balancing Algorithm for the Data Centres to Optimize Cloud Computing Applications	2021	Load balancing	196
8	Xu et al.: esDNN: Deep Neural Network Based Multivariate Workload Prediction in Cloud Computing Environments	2022	Workload prediction	170
9	Chen et al.: Adaptive and Efficient Resource Allocation in Cloud Datacenters Using Actor-Critic Deep Reinforcement Learning	2022	DRL/resource allocation	164
10	Nguyen et al.: Horizontal Pod Autoscaling in Kubernetes for Elastic Container Orchestration	2020	Auto-scaling	163

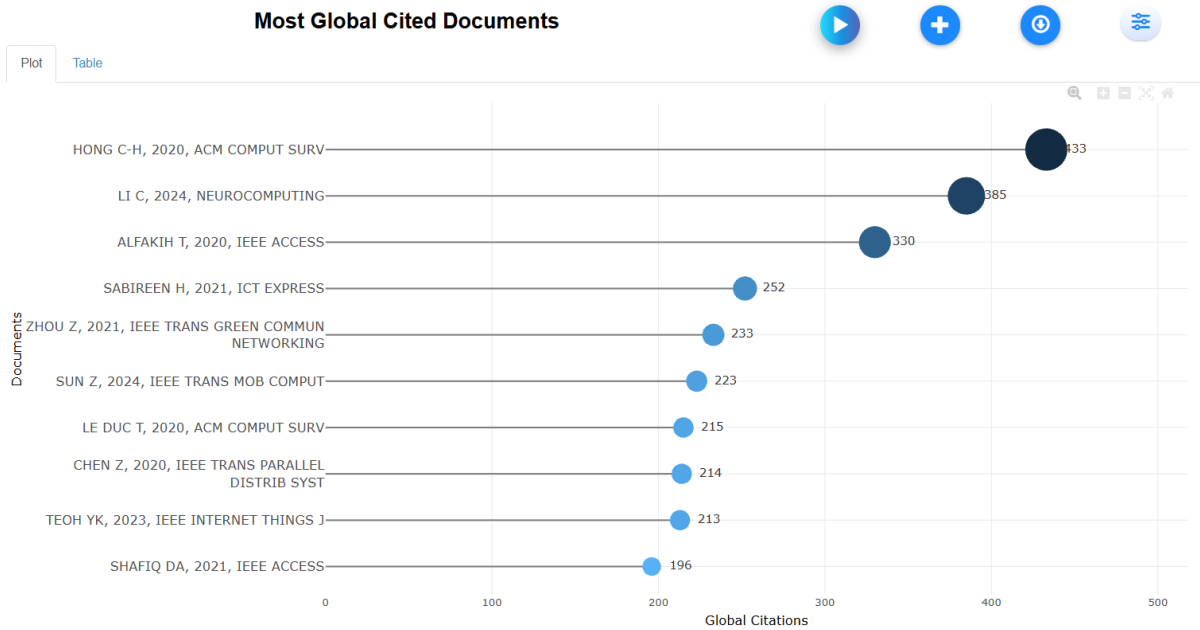


Figure 5. Most globally cited documents in the exported corpus.

4.5 Author productivity

The author-productivity curve is strongly right-skewed: most authors contribute only one document, while a much smaller number publish repeatedly in the field. The observed distribution is consistent with the general form expected under Lotka-type scientific productivity patterns, where a small group of prolific authors accounts for multiple publications. Because the supplied output does not report the fitted Lotka exponent and goodness-of-fit statistics, no numerical claim about exact conformity is made here.

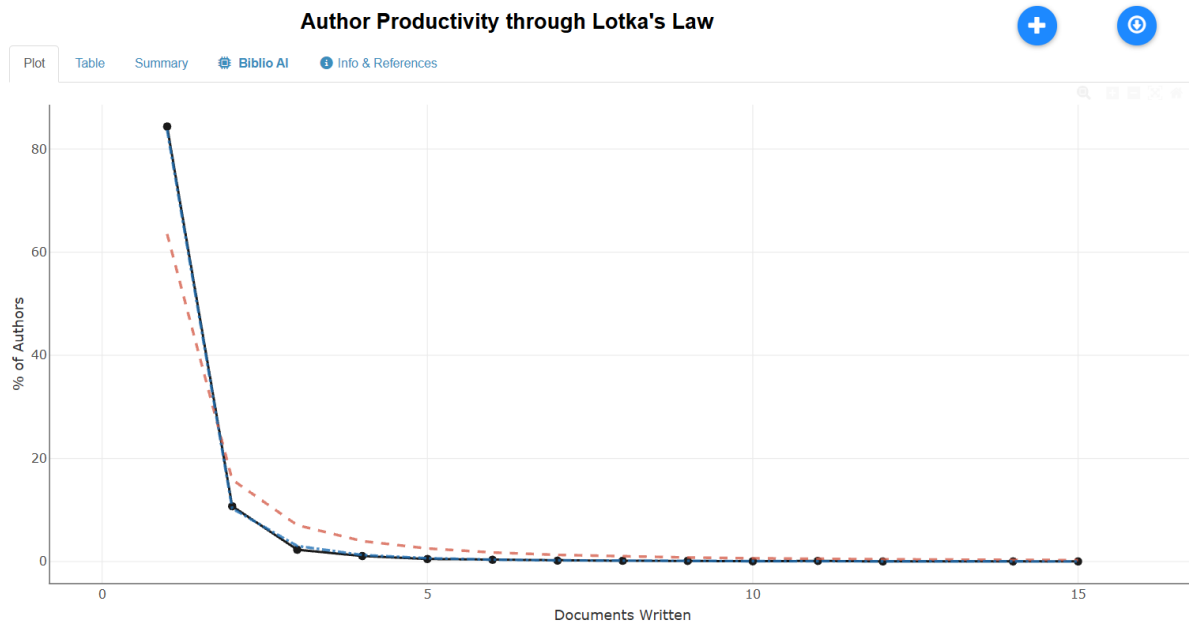


Figure 6. Author productivity pattern compared with Lotka's Law.

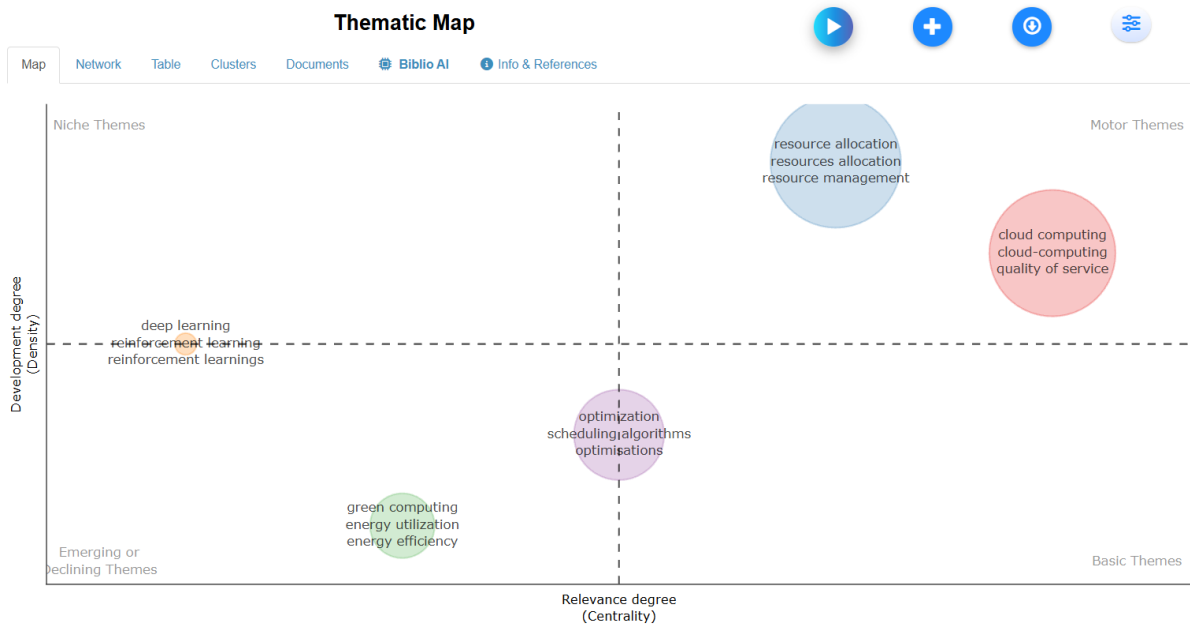


Figure 8. Thematic map of the literature.

4.8 Three-field relationships and international collaboration

The three-field plot demonstrates dense relationships between influential cited references, active authors and recurring keywords. Although individual links are numerous, the visual pattern confirms that resource allocation, cloud computing, quality of service and related resource-management terms are connected to multiple authors and intellectual foundations rather than being confined to a single school of research. The country-collaboration map likewise shows extensive cross-border collaboration, consistent with the 33.38% international co-authorship rate. The visible collaboration links across North America, Europe and Asia indicate that research on cloud and edge resource management is internationally distributed.

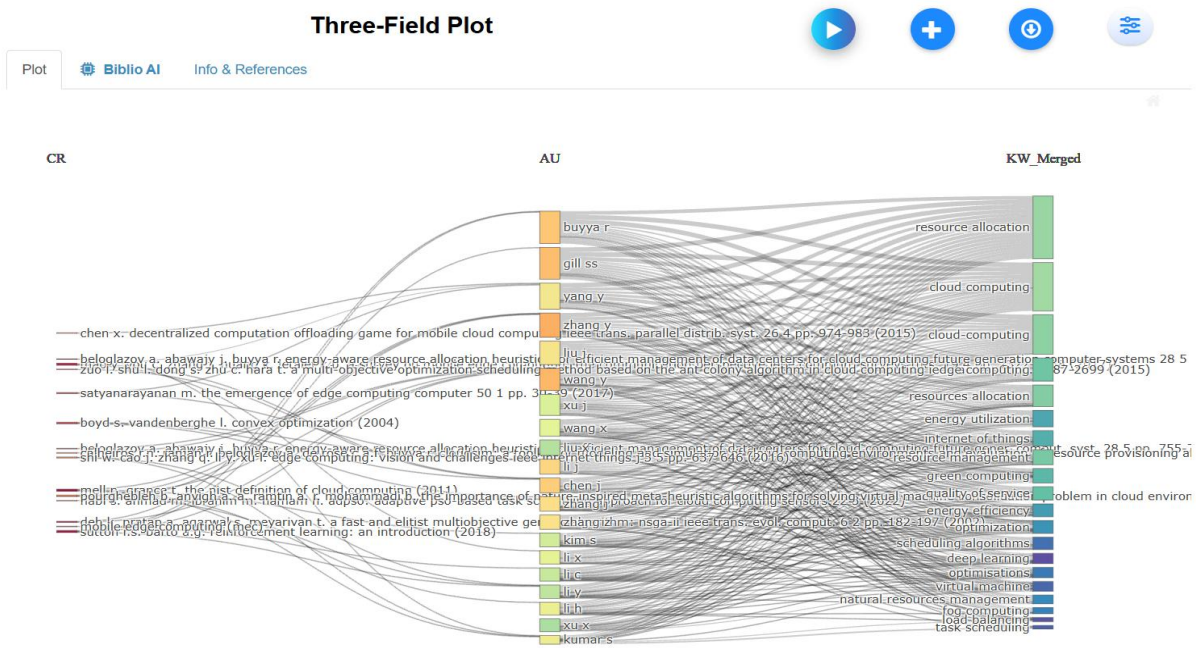


Figure 9. Three-field plot linking cited references, authors and keywords.

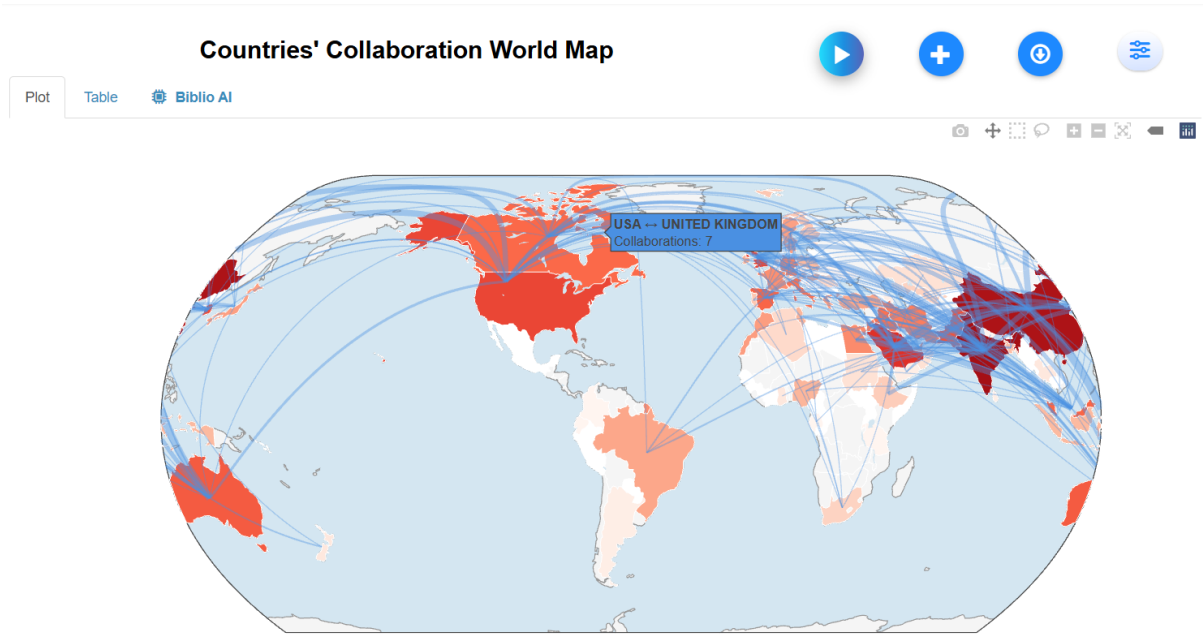


Figure 10. Country collaboration network.

5. Discussion

5.1 Evolution from reactive scaling toward prediction and adaptive control

The bibliometric evidence suggests that the field is moving beyond basic elasticity toward a combination of forecasting, adaptive allocation and learning-based control. Classical auto-scaling literature framed elasticity as the problem of matching leased resources to demand while meeting service-level objectives (Lorido-Botran et al., 2014). In the present corpus, the prominence of deep learning, reinforcement learning, workload prediction and resource allocation indicates a more data-driven phase. Chen et al. (2020) and Xu et al. (2022) represent the forecasting side of this transition, while Chen et al. (2022) represents the control side through reinforcement-learning-based allocation. The transition is conceptually important because predictive scaling requires both an anticipatory signal and a policy capable of acting on that signal.

5.2 Workload prediction accuracy as an energy-relevant capability

Prediction accuracy is frequently evaluated as a modelling outcome, yet its operational value lies in the downstream decisions it enables. A forecast that anticipates a burst before saturation can allow capacity to be activated earlier; conversely, a forecast of declining demand can permit resources to be released before extended idle periods occur. However, higher predictive accuracy does not automatically guarantee lower energy consumption. The forecast must be translated into resource actions with appropriate safety margins, prediction horizons and service constraints. This explains why workload prediction should be integrated with resource utilisation and scaling-decision effectiveness rather than treated as an isolated optimisation problem.

5.3 Resource utilisation efficiency as the operational mechanism

The strong centrality of resource allocation and resource management in both the co-occurrence network and thematic map indicates that utilisation remains the operational core of the field. Energy efficiency depends on how many physical and virtual resources remain active and how efficiently workloads are packed and scheduled across them. Zhou et al. (2021) explicitly connect VM allocation to cloud-data-centre energy consumption, while Shafiq et al. (2021) show the importance of balancing workload to avoid problematic utilisation states. In the proposed model, utilisation efficiency is therefore not merely another performance metric; it is the mechanism through which resource provisioning becomes energetically consequential.

5.4 Intelligent scaling decision effectiveness as a control capability

Auto-scaling effectiveness requires more than triggering an action whenever utilisation crosses a threshold. Decisions must be timely, proportionate and stable. Excessive sensitivity can cause oscillation and repeated scaling, while conservative policies can react too late. Kubernetes research demonstrates the practical importance of metric selection and scaling configuration (Nguyen et al., 2020). Reinforcement-learning approaches seek to improve this control layer by learning policies from environmental feedback rather than relying entirely on manually specified thresholds (Chen et al., 2022). For energy-efficient data centres, an effective intelligent controller should therefore balance service quality, resource cost and energy consequences simultaneously.

5.5 Research gap

The central research gap is one of integration. The bibliometric structure shows substantial work on resource allocation, cloud computing, QoS, machine learning and energy efficiency, but these topics are not equally central and are often studied as separate optimisation problems. Prediction studies commonly focus on forecast error; resource-management studies focus on utilisation or scheduling; auto-scaling studies focus on elasticity and SLA compliance; and green-cloud studies focus on energy-aware placement or consolidation. Comparatively less attention is directed to a single predictive control framework in which forecast accuracy, resource-utilisation efficiency and the quality of scaling decisions are explicitly connected to energy efficiency while preserving service quality. The thematic position of green computing/energy efficiency away from the motor themes further indicates an opportunity to make energy outcomes more central to AI-driven resource-management research.

Literature stream	Typical emphasis	Observed limitation	Framework response
Workload prediction	Forecasting demand; minimising prediction error	Accuracy may be assessed without tracing energy consequences of forecast-driven decisions	Treat workload prediction accuracy as IV1 linked to energy efficiency
Resource allocation/utilisation	Placement, scheduling, load balancing, consolidation	Often optimises utilisation/performance without explicitly integrating predictive scaling	Treat resource utilisation efficiency as IV2
Auto-scaling/intelligent control	Elasticity, thresholds, RL policies, SLA/QoS	Decision quality can be separated from forecasting and data-centre energy outcomes	Treat intelligent auto-scaling decision effectiveness as IV3
Green/energy-aware cloud	Power reduction, energy-aware VM allocation, active-host reduction	Energy is frequently optimised in specialised models rather than as the outcome of an integrated predictive controller	Treat energy efficiency as the DV and integrate all three antecedents

6. Proposed Conceptual Framework

Based on the bibliometric synthesis, this study proposes that energy-efficient cloud auto-scaling depends on three complementary capabilities. First, workload prediction accuracy determines how reliably future resource demand can be anticipated. Second, resource utilisation efficiency determines whether allocated infrastructure is productively used rather than left idle or fragmented. Third, intelligent auto-scaling decision effectiveness determines whether the system converts forecast and utilisation information into timely and proportionate scaling actions. The dependent variable is the energy efficiency of cloud data centres, assessed jointly with service quality so that reductions in energy use do not simply result from under-provisioning.

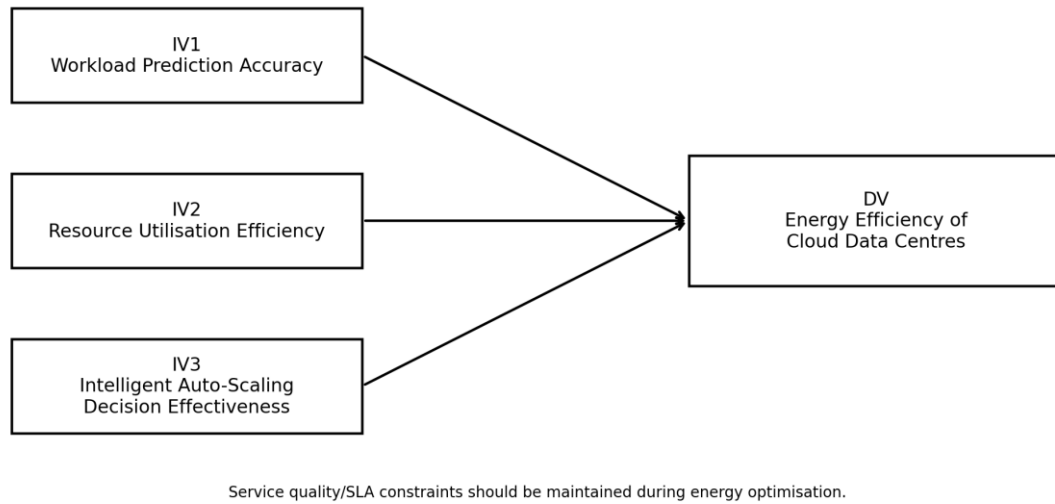


Figure 11. Proposed conceptual framework.

6.1 Propositions for subsequent validation

P1. Higher workload prediction accuracy is positively associated with the energy efficiency of cloud data centres.

P2. Higher resource utilisation efficiency is positively associated with the energy efficiency of cloud data centres.

P3. Greater intelligent auto-scaling decision effectiveness is positively associated with the energy efficiency of cloud data centres.

These propositions should be validated using simulation, experimental cloud traces, a cloud testbed or a combination of these approaches. A later technical evaluation can compare the proposed predictive framework against threshold-based and reactive baselines using energy consumption, SLA violations, response time, resource utilisation, scaling frequency and forecasting error. Such validation would directly address the fourth applied objective of developing and evaluating the framework.

7. Theoretical and Practical Implications

7.1 Theoretical implications

The study contributes to the literature by reframing energy-efficient auto-scaling as an integrated capability rather than a collection of isolated optimisation tasks. The bibliometric results show that resource allocation is already a motor theme, while AI and energy-efficiency themes remain distributed across neighbouring clusters. The proposed framework therefore connects three research layers: predictive intelligence (workload prediction accuracy), operational efficiency (resource utilisation efficiency) and adaptive control (auto-scaling decision effectiveness). This structure creates a clearer basis for cumulative research because future studies can test not only whether a new forecasting or control algorithm improves its immediate technical metric, but also whether that improvement translates into measurable data-centre energy gains.

7.2 Practical implications

For cloud operators, the framework suggests that energy-saving initiatives should not be restricted to server consolidation or power-aware scheduling. Forecasting models need to be evaluated in terms of the decisions they enable, utilisation metrics need to be incorporated into the control loop, and scaling policies should be assessed for timeliness and stability. In practical deployment, an AI-driven controller could ingest historical and real-time workload metrics, predict short-term demand, estimate the resource requirement, observe current utilisation and then select a scaling action subject to

SLA constraints. This integrated sequence can reduce prolonged idle capacity while avoiding the performance penalties associated with late scaling.

8. Future Research Agenda

Future research direction	Recommended focus
Prediction-to-energy linkage	Compare forecasting models not only on MAE/RMSE but on downstream energy and SLA outcomes under identical scaling policies.
Multi-resource forecasting	Move beyond CPU-only prediction by jointly modelling CPU, memory, network and storage demand where application behaviour requires it.
Decision stability	Evaluate oscillation, unnecessary scale events, cooldown behaviour and scaling latency as explicit indicators of decision effectiveness.
Energy-aware reinforcement learning	Design reward functions that jointly represent energy, SLA/QoS, utilisation and scaling cost rather than optimising a single objective.
Container and Kubernetes environments	Extend predictive scaling experiments to pod, node and cluster-level control in container-orchestrated cloud platforms.
Trace diversity and generalisation	Validate models across multiple real-world traces, workload types and data-centre configurations to test transferability.
Carbon-aware extensions	Distinguish energy efficiency from carbon efficiency by incorporating time- and location-varying electricity carbon intensity in future studies.

9. Limitations

This study has several limitations. First, it uses Scopus as the sole bibliographic database; coverage may therefore differ from Web of Science, IEEE Xplore, ACM Digital Library or other sources. Second, the use of Computer Science and Engineering subject filters and a broad cloud/resource-management search strategy means that adjacent edge/fog, IoT and scheduling research is present in the corpus. This breadth is useful for mapping the wider intellectual context but may introduce peripheral articles that are not exclusively about centralised cloud data centres. Third, citation counts favour older publications because they have had more time to accumulate citations. Fourth, the 2026 year is incomplete and the 2027 record is future-dated, so annual-growth statistics that use the terminal years should be interpreted cautiously. Fifth, bibliometric relationships demonstrate intellectual proximity and research prominence; they do not prove causal effects among workload prediction accuracy, resource utilisation efficiency, scaling effectiveness and energy efficiency. The proposed relationships require subsequent experimental or empirical testing.

10. Conclusion

This study mapped 1,312 Scopus-indexed articles on AI-enabled cloud resource management and energy-efficient computing using Bibliometrix/Biblioshiny. The literature expanded strongly from 2020 through 2025 and is distributed across 337 sources, with IEEE Access as the most prolific outlet. Resource allocation and cloud computing occupy central positions in the co-occurrence and thematic structures, while machine learning, deep learning, reinforcement learning and energy-efficiency concepts form closely connected but comparatively more distributed streams. Highly cited documents likewise demonstrate the field's convergence around resource management, workload prediction, adaptive allocation, load balancing and energy-aware VM management.

The main conceptual contribution is the identification of an integration gap. Existing studies often optimise prediction accuracy, utilisation, scaling decisions or energy consumption separately. The proposed framework brings these elements together by positioning workload prediction accuracy, resource utilisation efficiency and intelligent auto-scaling decision effectiveness as three antecedents of cloud-data-centre energy efficiency. This provides a theoretically coherent basis for the next stage of research: the development and experimental evaluation of an AI-driven predictive auto-scaling mechanism that reduces energy use while maintaining service quality.

References

- [1] Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959–975. <https://doi.org/10.1016/j.joi.2017.08.007>
- [2] Chen, Z., Hu, J., Min, G., Luo, C., & El-Ghazawi, T. (2022). Adaptive and efficient resource allocation in cloud datacenters using actor-critic deep reinforcement learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(8), 1911–1923. <https://doi.org/10.1109/TPDS.2021.3132422>
- [3] Chen, Z., Hu, J., Min, G., Zomaya, A. Y., & El-Ghazawi, T. (2020). Towards accurate prediction for high-dimensional and highly-variable cloud workloads with deep learning. *IEEE Transactions on Parallel and Distributed Systems*, 31(4), 923–934. <https://doi.org/10.1109/TPDS.2019.2953745>
- [4] Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- [5] Hong, C.-H., & Varghese, B. (2019). Resource management in fog/edge computing: A survey on architectures, infrastructure, and algorithms. *ACM Computing Surveys*, 52(5), Article 97. <https://doi.org/10.1145/3326066>
- [6] Le Duc, T., Leiva, R. G., Casari, P., & Östberg, P.-O. (2020). Machine learning methods for reliable resource provisioning in edge-cloud computing: A survey. *ACM Computing Surveys*, 52(5). <https://doi.org/10.1145/3341145>
- [7] Lorido-Botran, T., Miguel-Alonso, J., & Lozano, J. A. (2014). A review of auto-scaling techniques for elastic applications in cloud environments. *Journal of Grid Computing*, 12(4), 559–592. <https://doi.org/10.1007/s10723-014-9314-7>
- [8] Nguyen, T.-T., Yeom, Y.-J., Kim, T., Park, D.-H., & Kim, S. (2020). Horizontal pod autoscaling in Kubernetes for elastic container orchestration. *Sensors*, 20(16), 4621. <https://doi.org/10.3390/s20164621>
- [9] Shafiq, D. A., Jhanjhi, N. Z., Abdullah, A., & Alzain, M. A. (2021). A load balancing algorithm for the data centres to optimize cloud computing applications. *IEEE Access*, 9, 41731–41744. <https://doi.org/10.1109/ACCESS.2021.3065308>
- [10] Xu, M., Song, C., Wu, H., Gill, S. S., Ye, K., & Xu, C. (2022). esDNN: Deep neural network based multivariate workload prediction in cloud computing environments. *ACM Transactions on Internet Technology*, 22(3), Article 75. <https://doi.org/10.1145/3524114>
- [11] Zhou, Z., Shojafar, M., Alazab, M., Abawajy, J., & Li, F. (2021). AFED-EF: An energy-efficient VM allocation algorithm for IoT applications in a cloud data center. *IEEE Transactions on Green Communications and Networking*, 5(2), 658–669. <https://doi.org/10.1109/TGCN.2021.3067309>
