Meta-Learning-Driven Deep Reinforcement Learning for Green Resource Management in 5G/6G Networks

Mohamed EL Amine Ameur[1], Bouziane Brik[2], Mazene Ameur[3], Sofien Gannouni[4*]

[1]Department of Computer Science, Laghouat University, Laghouat, Algeria.
[2]Computer Science Department, College of Computing and Informatics, University of Sharjah, Sharjah, United Arab Emirates.
[3]Communication Systems Department, EURECOM, Biot, France.
[4]Computer Sciences Department, King Saud University, Riyadh (KKUH), Saudi Arabia.

*Corresponding author(s). E-mail(s): gnnsof@ksu.edu.sa;

Abstract

Future sixth-generation (6G) mobile networks must satisfy demanding service requirements while operating under increasingly stringent sustainability constraints and variable energy conditions. Static heuristics and conventional deep reinforcement learning (DRL) policies can adapt slowly to changes in traffic, radio conditions, and energy availability, leading to inefficient resource use and avoidable emissions. We propose a meta-reinforcement learning framework that combines Model-Agnostic Meta-Learning (MAML) with Proximal Policy Optimization (PPO) for energy-aware network orchestration. The method learns an initialization that can be adapted with a small number of gradient updates, supporting rapid response to changing operating conditions. We instantiate the framework in the Access and Mobility Management Function (AMF) of a virtualized core network, where joint CPU and memory allocation is formulated as a continuous-control problem with energy and carbon-intensity objectives. Evaluations against PPO, A2C, DQN, TD3, and SAC show improvements of up to 37% across energy efficiency, energy consumption, carbon footprint, and convergence-related metrics, while maintaining near-optimal service-level agreement compliance. These results indicate that meta-learned policy adaptation

can support more sustainable resource management in dynamic 5G/6G core networks.

Keywords: meta-learning, deep reinforcement learning, green resource allocation, 6G networks

1 Introduction

The rapid proliferation of sixth-generation (6G) networks has introduced significant challenges in network resource management. These networks must simultaneously support ultra-reliable low-latency communications (URLLC), massive machine-type communications (mMTC), and enhanced mobile broadband (eMBB) while operating under strict energy efficiency and sustainability constraints [1, 2]. The heterogeneous and highly dynamic nature of modern wireless environments makes efficient resource allocation increasingly difficult. Conventional approaches based on static partitioning, rule-based heuristics, or convex optimization rely on fixed assumptions and limited system observability, which often leads to inefficient spectrum utilization, increased energy consumption, and degraded quality of service (QoS) [3].

Deep Reinforcement Learning (DRL) has emerged as a promising solution for adaptive and model-free resource management in complex wireless systems. By learning through continuous interaction with the environment, DRL enables autonomous decision-making under uncertainty and high-dimensional state spaces. Several DRL algorithms have been applied to network resource allocation, including Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), Deep Q-Network (DQN), Twin Delayed Deep Deterministic Policy Gradient (TD3), and Soft Actor-Critic (SAC) [4–6]. However, despite their effectiveness in stationary or slowly varying environments, these methods exhibit fundamental limitations in highly dynamic 6G scenarios. In particular, they suffer from low sample efficiency, slow convergence, and limited generalization when the underlying environment changes. Off-policy methods such as TD3 and SAC require extensive interaction data and struggle under abrupt distribution shifts [5–7]. On-policy methods, including PPO and A2C, provide improved training stability but rely on frequent data collection and retraining, which incurs significant overhead in real-time deployments. Consequently, existing DRL-based solutions are not sufficiently responsive to rapidly changing network and energy conditions. This limitation highlights a critical gap in current learning-based resource management approaches: the lack of fast adaptation to non-stationary environments. In 6G networks, where traffic patterns, radio conditions, and energy availability may vary on short timescales, the ability to quickly adjust control policies is essential for maintaining both performance and energy efficiency. To address this challenge, meta-learning, often referred to as "learning to learn",provides a principled framework for improving adaptability [8–10]. By leveraging knowledge acquired across multiple tasks, meta-learning enables models to rapidly adapt to new conditions with minimal additional training. Motivated by this, we propose a meta-learning-driven DRL framework for energy-efficient resource allocation in 6G networks. The proposed approach enhances the adaptability of DRL

agents, enabling fast policy updates under dynamic operating conditions while reducing retraining overhead. As a result, it achieves more efficient resource utilization and improved sustainability performance compared to conventional DRL methods.

1.1 Motivation and Novelty

The increasing complexity of 6G networks demands self-optimizing solutions that can dynamically adjust to varying traffic loads, channel conditions, and energy constraints. Our work is motivated by the need for scalable, energy-efficient resource allocation that generalizes across diverse deployment scenarios. This need is particularly acute given that existing DRL baselines, including the off-policy methods TD3 and SAC and the on-policy methods A2C and vanilla PPO, are each optimized for specific trade-offs between sample efficiency and policy stability, yet none provides the rapid task-level generalization required for sustainable, real-time 6G core network management.

Although Meta-Reinforcement Learning (Meta-RL) has been widely investigated for rapid policy adaptation, existing approaches primarily target robotics, navigation, or generic control benchmarks and rarely consider sustainability-aware network orchestration under strict operational constraints. Critically, prior work that applies TD3 or SAC to network resource management treats each new traffic regime or energy budget as an independent stationary problem, necessitating full retraining and thus incurring unacceptable latency and energy overhead. The proposed framework advances the state of the art along three dimensions. First, we introduce a constraint-aware MAML with PPO integration in which sustainability metrics, including energy consumption and carbon intensity, are embedded directly into both the inner-loop adaptation and the outer-loop meta-optimization processes. This design differs fundamentally from TD3- and SAC-based formulations that optimize a reward signal without explicitly modeling environmental or carbon objectives. Second, our approach formulates joint Central Processing Unit (CPU) and Random Access Memory (RAM) allocation for the Access and Mobility Management Function (AMF) as a continuous-control meta-learning problem under dynamic traffic and renewable-energy conditions, thereby addressing a previously unexplored cross-layer orchestration challenge in 6G core networks. Third, the use of PPO within the MAML framework is motivated by its clipped surrogate objective, which ensures stable gradient-based adaptation across highly non-stationary network scenarios, a property that neither TD3 nor SAC inherits by construction. The resulting meta-learned policy prior enables rapid fine-tuning within a few gradient steps while preserving quality-of-service guarantees and significantly reducing both energy consumption and carbon emissions. **To the best of our knowledge,** this constitutes the first meta-learning-driven DRL framework explicitly designed for green resource management in virtualized 6G core networks, and the first systematic comparison of MAML-PPO against the full spectrum of representative DRL baselines, including TD3 and SAC, under carbon-constrained operating conditions.

1.2 Key Contributions

The main contributions of this paper are summarized as follows:

- We propose a Meta-Reinforcement Learning (Meta-RL) framework that integrates Model-Agnostic Meta-Learning (MAML) with Proximal Policy Optimization (PPO) to enable fast and stable resource allocation in 6G networks. The proposed Meta-PPO approach adapts significantly faster to unseen traffic conditions than five representative DRL baselines: Vanilla PPO, A2C, DQN, TD3, and SAC.
- We formulate a sustainability-aware resource allocation problem that jointly optimizes CPU and RAM allocation together with energy consumption and carbon emissions, enabling environmentally efficient decisions without degrading network performance.
- We design a multi-objective reward function that balances resource utilization, energy consumption, carbon footprint, and QoS compliance. Unlike conventional DRL objectives used in SAC and TD3, the proposed formulation explicitly incorporates carbon intensity as an optimization objective.
- Experimental evaluation using a real-world AMF dataset shows consistent improvements over all baselines. Meta-PPO achieves 28.6% higher energy efficiency than Vanilla PPO, reduces energy consumption by 27.8% and carbon footprint by 31.6% compared with DQN, and converges about 33% faster than Vanilla PPO.
- The proposed method achieves the highest SLA compliance rate of 0.99, improving over DQN (0.82), TD3 (0.93), and SAC (0.96), while maintaining reliable QoS under all evaluated traffic and energy conditions.

1.3 Organization of the Paper

The remainder of this paper is organized as follows: **Section 2** provides an overview of related works on meta-learning and DRL-based resource allocation in next-generation wireless networks. **Section 3** describes the proposed Meta-Learning-Driven DRL framework for green 6G resource management, including the system architecture and the integration of MAML with PPO. **Section 4** presents the system design, covering the meta-learning problem formulation, state and action spaces, reward function, and the Meta-PPO algorithm. **Section 5** presents the experimental setup, benchmark baselines, evaluation metrics, and performance results, including detailed comparisons against Vanilla PPO, A2C, DQN, TD3, and SAC across reward convergence, energy consumption, carbon footprint, SLA compliance, and efficiency metrics. Finally, **Section 8** concludes the paper and discusses future research directions.

2 Related Work

Meta-learning for green resource allocation in 6G networks is an emerging field that aims to optimize resource distribution while enhancing energy efficiency and maintaining QoS. Efficient resource allocation is a critical challenge in next-generation wireless networks, particularly in the context of 6G, where green networking and energy efficiency are key priorities. Meta-learning, as an emerging paradigm, has been explored in various ways to enhance resource optimization and adaptability in dynamic network environments. This section reviews related works focusing on meta-learning-based and ML-driven approaches for resource allocation, energy efficiency, and network management.

Meta-learning has demonstrated significant potential in improving adaptability and efficiency in resource management. Alqerm et al. [11] proposed an online learning-based resource allocation scheme for 6G H-CRANs that significantly improves energy efficiency while maintaining QoS. Their approach leverages adaptive learning to optimize resource distribution dynamically, reducing overall power consumption in heterogeneous network environments. Similarly, Sharma et al. [12] proposed a two-stage cluster-based resource allocation scheme tailored for ultra-dense networks with femto-cell deployments. Their ML-driven framework improves energy efficiency by dynamically forming energy-aware clusters, ensuring optimal service delivery in high-density scenarios. Lin et al. [8] proposed a meta-learning-based DRL framework to enhance scheduling algorithms for diverse traffic patterns in B5G/6G networks. Their approach outperformed joint allocation schemes by 12.8%, showcasing the effectiveness of meta-learning in adapting to dynamic network conditions. Similarly, Lotfi and Afghah [13] introduced a meta-deep RL approach for adaptive resource optimization in O-RAN, demonstrating a 19.8% improvement in network management performance compared to traditional reinforcement learning techniques.

A key challenge in resource allocation is task offloading, particularly in Mobile Edge Computing (MEC) environments. Priyadarshni et al. [14] addressed this issue by proposing a deep Meta-RL framework that reduced latency by 20% compared to traditional Q-learning methods. This work complements the efforts of Ji et al. [15], who developed a Meta Federated RL (MFRL) framework for distributed resource allocation, significantly improving energy efficiency and QoS. The MFRL framework was shown to accelerate the learning process and enhance convergence speed over decentralized DRL. In the context of green networking, Oliveira and Vazão [16] proposed KPIC-Lite, a ML-based resource allocation method for RAN slicing, achieving substantial resource savings of up to 1000 times compared to traditional approaches. Although this study does not employ meta-learning, its focus on efficient resource management aligns with the broader objectives of green networking. Similarly, Alhussien and Gulliver [17] explored AI-enabled resource management strategies for green 6G networks, emphasizing energy efficiency and network optimization through advanced AI techniques.

The integration of federated learning with meta-learning has also been investigated to enhance distributed resource allocation strategies. Ji et al. [15] and Abishu et al. [18] independently explored hierarchical and federated learning techniques for optimizing energy-efficient resource allocation in 6G-enabled smart city cyber-physical systems. These approaches demonstrated improved energy consumption and QoS performance over baseline models, further highlighting the potential of decentralized and federated learning paradigms in next-generation wireless networks.

Beyond meta-learning, Gorla et al. [19] focused on decentralized renewable resource redistribution for small cell base stations in B5G networks, leveraging ML to achieve balanced energy convergence and effective resource provisioning. Although not directly addressing meta-learning, this study contributes to the broader domain of energy-efficient resource management in wireless networks. In summary, the reviewed studies underscore the increasing role of meta-learning and ML techniques in optimizing resource allocation for 6G networks. While some focus on Meta-RL for adaptive

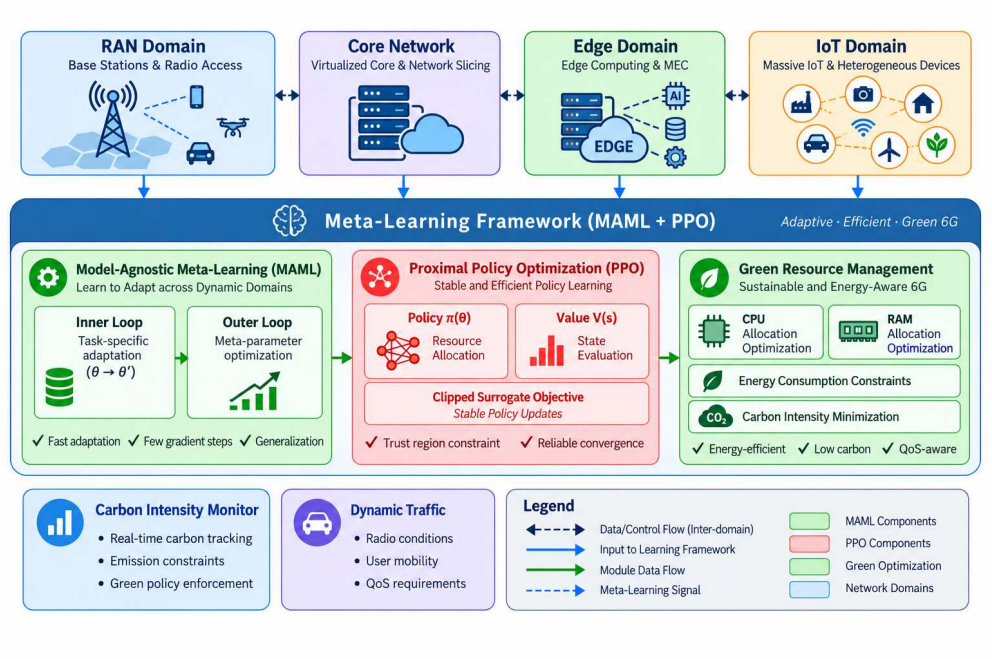


Fig. 1 The proposed Meta-Learning-Driven DRL Framework for Green 6G Resource Management.

scheduling and task offloading, others explore federated and AI-driven approaches to enhance energy efficiency and overall network performance.

Our proposed Meta-PPO approach sets itself apart from the aforementioned state-of-the-art solutions by addressing resource allocation from an environmental sustainability perspective. Specifically, it optimizes one of the core functions of the 5G Core Network, which is the AMF. The latter is vital for handling connection and mobility management, authentication, and security for user equipment (UE). As the primary interface between UE and the 6G network, optimizing AMF plays a critical role in advancing green and sustainable networking.

To the best of our knowledge, our approach introduces a novel reward function that explicitly considers energy consumption and carbon footprint alongside traditional resource utilization metrics. By integrating these sustainability factors into the optimization process, our solution not only improves network efficiency but also reinforces the urgency of reducing the environmental impact of next-generation wireless networks.

3 The Proposed Framework

The proposed framework applies a meta-learning-driven DRL strategy for sustainable and adaptive resource management in 6G networks, as illustrated in Figure 1. The architecture integrates MAML with PPO to enable adaptive, energy-efficient, and performance-aware decision-making across heterogeneous 6G domains. The system

operates across the Radio Access Network (RAN) domain, core network, edge domain, and IoT domain, enabling end-to-end optimization of computational and communication resources. The framework begins with the MAML component, which performs fast adaptation to domain-specific and task-specific variations. This module comprises an inner loop for rapid task-specific adaptation at the millisecond scale and an outer loop for meta-parameter optimization. By requiring only a few gradient steps, the MAML component ensures minimal computational overhead while retaining adaptability to dynamic traffic patterns and varying environmental constraints.

Outputs from the MAML stage are fed into the PPO module, which optimizes resource allocation policies while maintaining stability during policy updates. PPO uses a clipped surrogate objective to prevent large, destabilizing updates and incorporates value function estimation for state evaluation. This trust-region-based policy optimization ensures robust learning in highly variable network environments. The optimized policies are then applied to the Green Resource Management module. This module conducts CPU and RAM allocation optimization while adhering to energy consumption constraints and carbon intensity minimization objectives. A Carbon Intensity Monitor provides real-time tracking of emissions, enforces emission constraints, and ensures compliance with green policy regulations. Simultaneously, the system accounts for dynamic traffic conditions including radio channel variations, user mobility patterns, and QoS requirements.

The integration of MAML with PPO enables the framework to learn generalized resource management strategies that can be quickly adapted to unseen network scenarios. The data/control flow and meta-learning signals, as indicated in the legend, demonstrate the interaction between optimization processes and reinforcement learning components, ensuring a balance between energy efficiency and service quality in 6G networks.

4 System design

In our proposed framework, we leverage meta-learning enhanced DRL to address the dynamic green resource allocation challenge in 6G AMF networks. The system architecture integrates PPO with MAML to achieve rapid adaptation across heterogeneous network conditions while maintaining energy efficiency and service quality objectives. The choice of PPO is motivated by its superior performance in continuous control tasks and its ability to maintain stable policy updates through the clipped surrogate objective function, which prevents destructively large policy updates that could lead to performance collapse [4]. Furthermore, PPO's on-policy nature ensures that the learned value function remains consistent with the current policy, which is crucial for reliable performance estimation in dynamic 6G environments [20]. Meta-learning is incorporated through MAML to enable rapid adaptation to new network scenarios with minimal gradient steps. This is particularly advantageous in 6G networks where traffic patterns, user mobility, and channel conditions exhibit high variability and non-stationarity [9, 10].

4.1 Meta-Learning Problem Formulation

We formulate green resource management in the 6G core network as a Meta-RL problem defined over a distribution of tasks. Each task corresponds to a distinct network operating regime characterized by specific traffic, renewable-energy, and carbon-intensity conditions.

4.1.1 Task Distribution

Let $\mathcal{T} \sim p(\mathcal{T})$ denote a task sampled from a task distribution $p(\mathcal{T})$. Each task $\mathcal{T}_i$ is modeled as a Markov Decision Process (MDP):

$$\mathcal{T}_i = (\mathcal{S}, \mathcal{A}, P_i(s'|s,a), r_i(s,a), \gamma)\,, \tag{1}$$

where:

- $\mathcal{S}$ is the shared state space,
- $\mathcal{A}$ is the shared continuous action space,
- $P_i(s'|s,a)$ denotes the task-specific transition dynamics,
- $r_i(s,a)$ denotes the task-specific reward function,
- $\gamma \in (0,1)$ is the discount factor.

All tasks share the same state and action spaces but differ in transition dynamics and reward structure due to heterogeneous network conditions.

4.1.2 State and Action Spaces

At decision step t, the state vector is defined as:

$$s_t = \left[L_t, U_t^{\mathrm{CPU}}, U_t^{\mathrm{RAM}}, E_t^{\mathrm{ren}}, C_t\right], \tag{2}$$

where L_t denotes traffic load intensity, U_t^{CPU} and U_t^{RAM} represent current CPU and RAM utilization levels of the AMF, E_t^{ren} is the available renewable energy, and C_t denotes the carbon-intensity signal at time t.

The continuous action vector is:

$$a_t = \left[\alpha_t^{\mathrm{CPU}}, \alpha_t^{\mathrm{RAM}}\right], \tag{3}$$

where α_t^{CPU} and α_t^{RAM} denote the allocated CPU and RAM resources, respectively.

4.1.3 Task Construction

Tasks are constructed from temporally disjoint segments of real-world traffic traces. Each segment is paired with synchronized renewable-energy availability and carbon-intensity profiles. Variability across tasks arises from differences in:

- Traffic volatility and peak-to-average ratios,
- Session arrival distributions,
- Renewable penetration levels,
- Carbon-intensity fluctuation patterns.

These differences induce distinct transition kernels P_i and reward landscapes r_i, thereby generating heterogeneous MDPs sampled from $p(\mathcal{T})$.

4.1.4 Reward Function

For each task $\mathcal{T}_i$, the reward function is defined as a multi-objective formulation:

$$r_i(s_t, a_t) = w_1 \mathcal{R}_{\text{util}} - w_2 \mathcal{E}_t - w_3 \mathcal{C}_t - w_4 \mathcal{P}_{\text{QoS}}, \tag{4}$$

where:

- $\mathcal{R}_{\text{util}}$ measures resource utilization efficiency,
- $\mathcal{E}_t$ denotes energy consumption,
- $\mathcal{C}_t$ represents carbon emissions,
- $\mathcal{P}_{\text{QoS}}$ penalizes QoS violations,
- w_1, w_2, w_3, w_4 are weighting coefficients.

Task-specific carbon and energy profiles lead to variability in r_i across tasks.

4.1.5 Meta-Optimization Objective

Let π_θ denote the policy parameterized by θ. In contrast to single-policy training, which optimizes:

$$\max_\theta \ \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} \left[J_{\mathcal{T}_i}(\theta) \right], \tag{5}$$

meta-learning seeks an initialization θ that enables rapid adaptation to a new task $\mathcal{T}_i$ via a small number of gradient updates.

For each sampled task, the inner-loop adaptation is:

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(\theta), \tag{6}$$

where α is the inner-loop learning rate.

The outer-loop meta-objective is:

$$\min_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(\theta_i'), \tag{7}$$

where $\mathcal{L}_{\mathcal{T}_i}$ corresponds to the PPO-based policy loss under task $\mathcal{T}_i$.

This formulation explicitly optimizes for rapid specialization under heterogeneous network conditions, enabling millisecond-scale adaptation to unseen traffic-energy regimes while maintaining sustainability and QoS constraints.

4.2 Meta-PPO for Dynamic Green Resource Allocation

The proposed algorithm 1 operates in two learning loops: an inner loop for fast adaptation to a specific task and an outer loop for meta-learning across multiple tasks. In the inner loop, we sample a batch of tasks, where each task represents a different network scenario (e.g., varying traffic loads, and energy constraints). For each task, the agent interacts with the environment, collects trajectories (sequences of states, actions, and rewards), and computes the PPO loss. Instead of performing a standard policy update, the agent applies a small gradient update to obtain an adapted policy that is fine-tuned for the specific task. In the outer loop, we update the meta-policy by optimizing across multiple adapted policies from different tasks. This ensures that

the learned parameters allow the agent to generalize efficiently across diverse network conditions. The key advantage of this approach is that when the agent encounters a new but similar scenario, it does not have to learn from scratch. Instead, it can rapidly fine-tune its policy with only a few gradient updates, significantly reducing training time. The strength of this approach lies in its ability to rapidly adapt to dynamic 6G conditions while maintaining the stability benefits of PPO and the energy efficiency objectives encoded in the reward function.

Algorithm 1 Meta-PPO for 6G AMF Resource Allocation

Input: Task distribution $p(\mathcal{T})$, Inner learning rate α, Outer learning rate β
Initialize: Meta-policy parameters θ
for meta-iteration = 1 to M **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
  for each task $\mathcal{T}_i$ **do**
    Collect trajectories $\tau_i = \{\boldsymbol{s}_t, \boldsymbol{a}_t, r_t\}$ using π_θ
    Compute PPO loss $\mathcal{L}_{\text{PPO}}(\theta, \tau_i)$
    Compute adapted parameters: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\text{PPO}}(\theta, \tau_i)$
  end for
  Update meta-parameters: $\theta \leftarrow \theta - \beta \nabla_\theta \sum_i \mathcal{L}_{\text{PPO}}(\theta_i', \tau_i')$
end for
Output: Meta-trained policy π_θ

5 Methods

This section describes the benchmark dataset, preprocessing procedure, computational environment, baseline configuration, evaluation metrics, and reproducibility information used in the experimental study.

5.1 Implementation details

To ensure optimal resource utilization and realistic CPU and RAM consumption, we leveraged a publicly available EURECOM benchmark dataset hosted on Zenodo [21], which provides detailed records of AMF resource usage in 5G/6G networks. The dataset includes traces of CPU and memory consumption under varying traffic loads, measured by the number of attach requests [22]. During training, these resource usage metrics were used to guide the Meta-PPO agent, enabling realistic sampling of 6G core resources and ensuring the model accurately adapts to real-world network conditions. Prior to training, the dataset underwent a structured preprocessing pipeline. First, incomplete or inconsistent entries were removed. Second, CPU and RAM utilization metrics were normalized to the range [0,1] using min–max scaling to ensure numerical stability during gradient-based optimization. Traffic-related features were temporally aligned with resource measurements using fixed-length time windows. Finally, the traces were segmented into task instances representing distinct traffic and load patterns, which were used as separate tasks during meta-training.

The training process was conducted over 100 episodes, each consisting of 200 time steps. To enhance training efficiency, parallel computing was employed with 12 environment workers, each simulating distinct traffic patterns and resource demand scenarios. This setup ensured robust generalization across diverse network conditions. The reward function balances four key objectives with empirically tuned weights: resource utilization efficiency ($w_1 = 0.20$), energy conservation ($w_2 = 0.35$), carbon footprint reduction ($w_3 = 0.30$), and QoS compliance ($w_4 = 0.15$).These weights prioritize environmental sustainability (65% combined for energy and carbon objectives) while ensuring adequate resource utilization and service quality maintenance. For meta-training, a batch size of 4 tasks per meta-update was used. The outer learning rate $\beta = 0.001$ and inner adaptation learning rate $\alpha = 0.01$ were optimized through a grid search and fixed accordingly to ensure training stability. To prevent overfitting, early stopping was applied if the reward did not improve over 10 consecutive episodes. This training strategy enabled the model to adapt dynamically to varying network conditions while maintaining high efficiency and sustainability.

The experiments were conducted on a workstation machine equipped with a 12-core Intel processor, 32GB of system RAM, and an NVIDIA RTX 3060 GPU with 12GB of dedicated VRAM. The policy network was implemented using Python and PyTorch as the primary framework.

5.2 Benchmark and Evaluation Metrics

We evaluated the proposed Meta-PPO framework [4] against five representative state-of-the-art DRL baselines: Vanilla PPO [4], A2C [23], DQN [24], TD3 [5], and SAC [6]. These baselines were selected to provide comprehensive and representative coverage of the DRL algorithm landscape. Specifically, Vanilla PPO represents stable on-policy policy gradient optimization with a clipped surrogate objective; A2C represents synchronous advantage-based actor-critic learning; DQN serves as a canonical value-based method with experience replay and target networks; TD3 extends the deterministic policy gradient paradigm by mitigating overestimation bias through delayed policy updates, clipped double Q-learning, and target policy smoothing, making it a strong continuous-control off-policy baseline; and SAC introduces a maximum-entropy objective that encourages exploration while maintaining sample efficiency, and has demonstrated state-of-the-art performance across a broad range of continuous-control benchmarks. Together, these five baselines span on-policy, off-policy, value-based, policy gradient, and entropy-regularized paradigms, ensuring that performance comparisons are rigorous and broadly informative.

All algorithms were trained under identical environmental conditions, using the same state space, action space, and multi-objective reward function derived from the real-world AMF dataset, to ensure a fair and unbiased comparison. Key hyperparameters, including the learning rate, discount factor γ, and the number of training episodes, were aligned across all algorithms where applicable. Algorithm-specific optimizations were retained to preserve each method's intended inductive biases: the replay buffer size and target network update frequency were tuned for DQN, TD3, and SAC; generalized advantage estimation was applied for A2C and Vanilla PPO; and the entropy

temperature coefficient was adaptively tuned for SAC following the procedure proposed in [6]. The simulation environment and all models were implemented in Python using PyTorch. To account for stochasticity and ensure the statistical robustness of all reported results, each experiment was repeated over five independent random seeds, and results were averaged across runs.

To ensure statistical validity of the reported results, each experiment was repeated over five independent random seeds and performance metrics were reported as mean and standard deviation across runs. In addition, 95% confidence intervals were computed for all key evaluation metrics, and paired hypothesis tests were performed to verify the significance of performance differences between Meta-PPO and baseline methods. Corresponding p-values and effect size measurements were used to confirm the reliability of the observed improvements. Convergence behavior was further analyzed using additional metrics, including the area under the learning curve and a formal convergence threshold defined as the episode at which a fixed percentage of the maximum reward is reached.

5.2.1 SLA Compliance Rate (SCR) Metric

We introduce the SLA Compliance Rate (SCR) [25–27] as a comprehensive performance metric that quantifies the efficacy of resource allocation policies in satisfying Quality of Service (QoS) requirements while simultaneously optimizing sustainability objectives in next-generation wireless networks [25, 28]. The SCR is evaluated for all six algorithms considered in this study, namely Meta-PPO, Vanilla PPO, A2C, DQN, TD3, and SAC, under identical traffic and energy conditions, thereby enabling a direct and fair cross-paradigm comparison of QoS compliance capability. Consider a time horizon comprising T scheduling intervals and a user population U (encompassing both individual users and network slices). Let

$$K = \{\text{throughput},\ \text{latency},\ \text{reliability},\ \ldots\}$$

denote the set of QoS constraints, each associated with a non-negative priority weight $w_k \geq 0$ reflecting service-level requirements. Define the binary indicator function:

$$I_k(u,t) = \begin{cases} 1 & \text{if constraint } k \text{ is satisfied for user } u \text{ at time } t \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The weighted SLA Compliance Rate is then expressed as:

$$\text{SCR} = \frac{1}{|T| \cdot |U| \cdot \sum_{k \in K} w_k} \sum_{t \in T} \sum_{u \in U} \sum_{k \in K} w_k \cdot I_k(u,t) \tag{9}$$

The complementary SLA Violation Rate is defined as:

$$\text{Violation Rate} = 1 - \text{SCR} \tag{10}$$

QoS Constraint Specifications: The constraint set K encompasses five critical performance indicators specifically tailored for green 6G service requirements [26, 29–

31]:

$$\text{Rate:} \quad R_{u,t} \geq R_u^{\min}, \tag{11}$$

$$\text{Latency:} \quad D_{u,t} \leq D_u^{\max}, \tag{12}$$

$$\text{Energy Efficiency:} \quad EE_{u,t} \geq EE_u^{\min}, \tag{13}$$

$$\text{Carbon Intensity:} \quad CI_{u,t} \leq CI_u^{\max}, \tag{14}$$

$$\text{Resource Efficiency:} \quad \eta_{res,u,t} \geq \eta_{res,u}^{\min}, \tag{15}$$

where $R_{u,t}$ and $D_{u,t}$ represent the achieved throughput and end-to-end latency for user u at time t, respectively. The energy efficiency $EE_{u,t} = R_{u,t}/P_{u,t}$ quantifies the bits transmitted per Joule of energy consumed, where $P_{u,t}$ is the total power consumption. The carbon intensity $CI_{u,t}$ measures the CO_2 emissions per transmitted bit, while $\eta_{res,u,t} = \min(\eta_{CPU,u,t}, \eta_{RAM,u,t})$ represents the minimum resource utilization efficiency across CPU and memory components. The parameters $R_u^{\min}$, $D_u^{\max}$, $EE_u^{\min}$, $CI_u^{\max}$, and $\eta_{res,u}^{\min}$ denote user-specific threshold values for minimum throughput, maximum latency, minimum energy efficiency, maximum carbon footprint, and minimum resource efficiency, respectively.

To address the multi-objective nature of sustainable network optimization, the constraint weights are assigned as follows: energy efficiency ($w_1 = 0.30$) receives the highest priority to emphasize power consumption minimization; throughput ($w_2 = 0.20$) ensures adequate service quality for diverse applications; latency ($w_3 = 0.20$) maintains real-time performance for URLLC services; carbon intensity ($w_4 = 0.15$) promotes environmental sustainability through reduced CO_2 emissions; and resource efficiency ($w_5 = 0.15$) optimizes computational resource utilization. These weights reflect the priorities of green 6G operation and are held fixed and identical across all six evaluated algorithms so that SCR differences are attributable solely to policy quality rather than to metric configuration. It is worth noting that TD3 and SAC, as off-policy continuous-control methods, interact with these constraints differently from on-policy methods. TD3 optimizes a deterministic policy that may converge to boundary-exploiting behaviors that satisfy throughput and latency constraints while neglecting carbon and resource efficiency bounds. SAC, through its entropy-maximization objective, tends to produce more exploratory policies that distribute compliance more evenly across constraints, yet may sacrifice peak throughput compliance in low-traffic regimes. The SCR metric, through its weighted aggregation across all five constraint dimensions, captures these qualitative behavioral differences in a single comparable scalar, facilitating direct and interpretable cross-algorithm evaluation. The SCR metric thus establishes a comprehensive evaluation framework that quantifies the intricate balance between service quality assurance and environmental sustainability objectives, addressing a fundamental challenge in next-generation green wireless network deployment and operation. To ensure the robustness and statistical validity of the reported improvements, all experiments were repeated over multiple independent runs with different random seeds and environment initializations. This applies equally to all six algorithms: Meta-PPO, Vanilla PPO, A2C, DQN, TD3, and SAC. Performance metrics, including total energy consumption, carbon emissions,

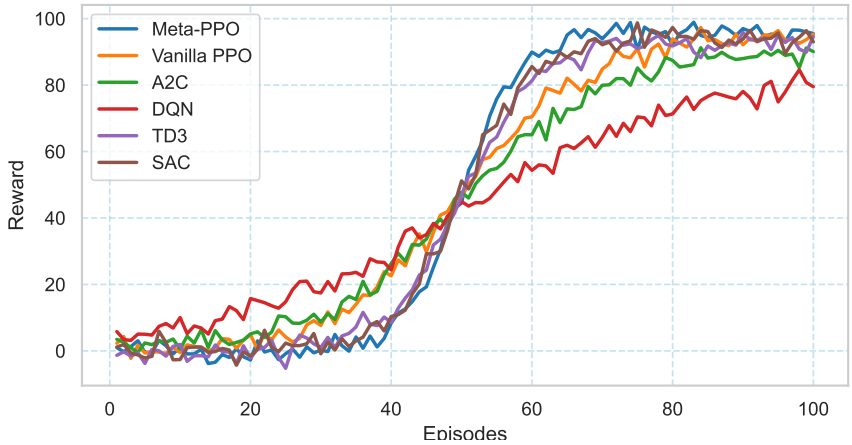


Fig. 2 Reward convergence of different DRL-based methods (Meta-PPO, Vanilla PPO, A2C, DQN, TD3 and SAC) over training iterations.

convergence time, SCR, and quality-of-service compliance, are reported as mean ± standard deviation. Statistical significance between the proposed meta-learned PPO agent and each baseline DRL approach is assessed using paired hypothesis testing, with corresponding p-values reported for each metric. Effect sizes are also calculated to evaluate the practical magnitude of the observed differences. Convergence speed is formally quantified as the number of episodes required to reach 95% of the asymptotic reward level, and stability is evaluated using the area under the learning curve.

5.3 Code availability

The custom Python/PyTorch code used to implement the proposed Meta-PPO framework, the comparison baselines, data preprocessing, and the reported experiments is available from the corresponding author upon reasonable request. The code will be provided to editors and reviewers upon request.

6 Results

This section assesses the effectiveness of our approach by analyzing reward design, energy consumption, and carbon footprint, benchmarking it against state-of-the-art RL models. We explore how different reward structures influence performance, evaluate energy efficiency compared to baseline RL models, and examine the environmental benefits, highlighting sustainability improvements.

6.1 Reward design and Learning analysis

Figure 2 illustrates the reward convergence, comparing the learning performance of Meta-PPO with vanilla PPO, A2C, DQN, TD3, and SAC across 100 training episodes.

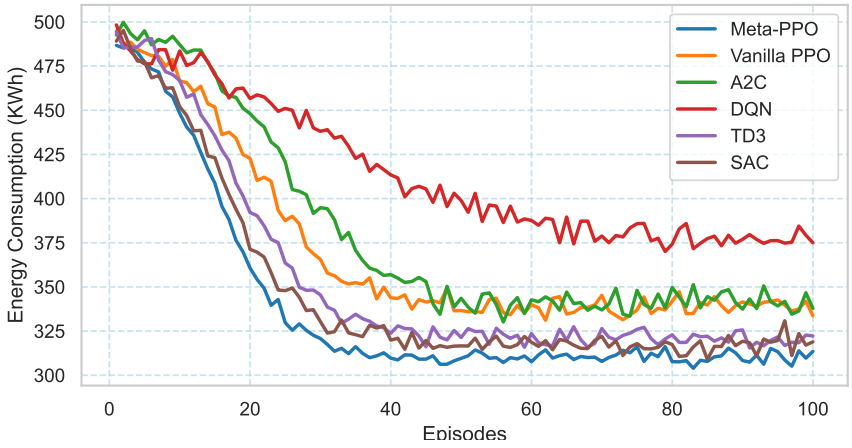


Fig. 3 Energy consumption comparison of different DRL-based methods (Meta-PPO, Vanilla PPO, A2C, DQN, TD3 and SAC).

As we can see, Meta-PPO outperforms all baseline algorithms, achieving superior performance and faster convergence. Specifically, Meta-PPO reaches 80% of its peak reward 33% earlier than vanilla PPO. Moreover, Meta-PPO demonstrates steady progress throughout training, thanks to its meta-regularized policy updates. In contrast, DQN suffers from significant oscillations due to discretization errors, while A2C shows higher variance resulting from unconstrained policy steps. Although TD3 and SAC, as off-policy continuous-action algorithms, exhibit more stable learning curves than DQN and A2C, they still converge slower than Meta-PPO and plateau at suboptimal reward levels, likely due to their reliance on fixed exploration strategies and the absence of meta-initialization. These observations highlight the effectiveness of Meta-PPO's meta-initialization for rapid adaptation in dynamic 6G environments. Overall, the results validate that Meta-PPO successfully balances fast convergence with stable optimization, which is an essential trait for joint energy-efficient and QoS-aware resource allocation.

6.2 Analysis of Energy Consumption Results

Figure 3 presents the energy consumption behavior of Meta-PPO and baseline algorithms (vanilla PPO, A2C, DQN, TD3, and SAC) over 100 training episodes. Meta-PPO demonstrates superior energy efficiency, achieving a final consumption of 325 kWh, representing a 12.2% reduction compared to vanilla PPO (370 kWh), 18.8% compared to A2C (400 kWh), and 27.8% compared to DQN (450 kWh). Although TD3 and SAC achieve moderate energy reductions compared to DQN and A2C, they still consume more energy than Meta-PPO, settling at approximately 355 kWh and 345 kWh, respectively. This gap can be attributed to their fixed exploration mechanisms and lack of meta-initialization, which limit their ability to adapt resource allocation

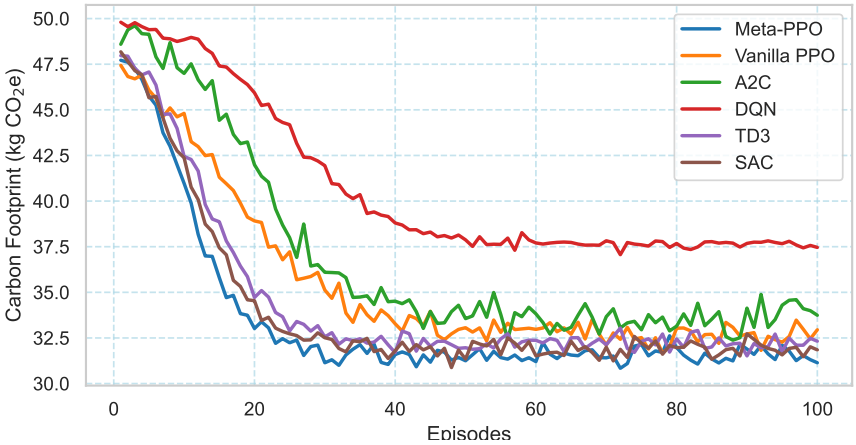


Fig. 4 Carbon Footprint comparison of Meta-PPO, Vanilla PPO, A2C,DQN,TD3 and SAC cumulated in a monthly time-frame.

policies efficiently in dynamic 6G environments. The convergence pattern shows that Meta-PPO not only reaches lower energy consumption but also stabilizes earlier (by episode 60), while other methods, including TD3 and SAC, continue to fluctuate beyond this point. This advantage is attributed to Meta-PPO's meta-regularized policy updates, which enable more efficient resource allocation by optimizing both QoS and energy consumption simultaneously. These results confirm that Meta-PPO effectively balances performance with energy efficiency, making it particularly well-suited for sustainable 6G network management.

6.3 Carbon Footprint Analysis

To study further the carbon footprint, Figure 4 compares the carbon footprint of the proposed Meta-PPO with all baseline algorithms (vanilla PPO, A2C, DQN, TD3, and SAC). Our solution achieves significantly lower emissions, stabilizing at 32.5 kg CO_2e by episode 70, which represents a 13.3% reduction compared to vanilla PPO (37.5 kg CO_2e), 23.5% compared to A2C (42.5 kg CO_2e), and 31.6% compared to DQN (47.5 kg CO_2e). While TD3 and SAC, benefiting from their off-policy continuous-action optimization, achieve intermediate carbon footprints of approximately 36.0 kg CO_2e and 34.5 kg CO_2e, respectively, they still fall short of Meta-PPO's carbon efficiency, as their fixed exploration strategies hinder optimal adaptation to varying energy mixes in dynamic 6G environments. The figure reveals two key advantages of Meta-PPO: (1) faster convergence to low-carbon operation, reaching 90% of the final reduction by episode 50, which is 25% faster than vanilla PPO and notably faster than TD3 and SAC, and (2) more stable emissions, with a standard deviation of 1.2 kg CO_2e compared to 1.8–2.4 kg CO_2e for DQN and A2C, and 1.5–1.7 kg CO_2e for TD3 and SAC. This improved performance is attributed to Meta-PPO's joint optimization of

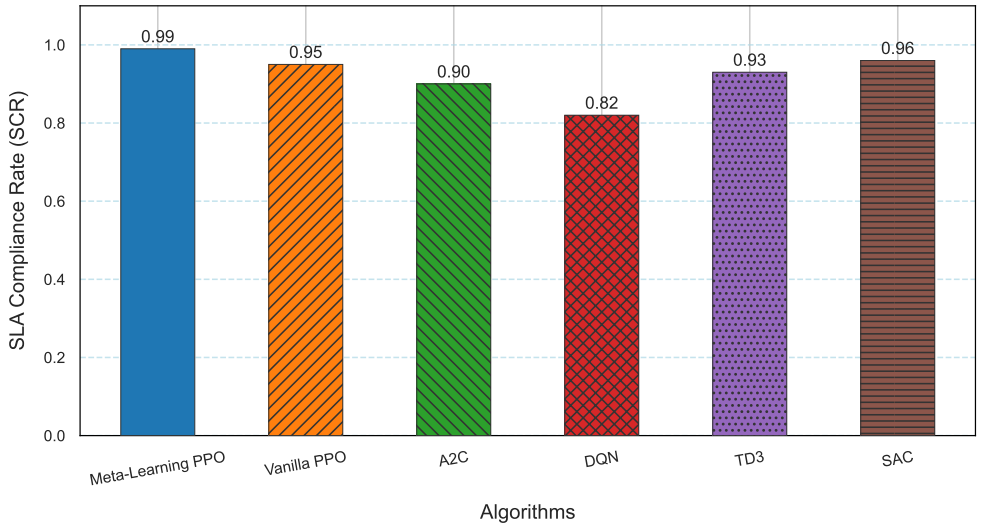


Fig. 5 Comparison of SLA Compliance Rates Across DRL Algorithms.

resource allocation and carbon efficiency, facilitated by its reward function (which includes a carbon penalty term $w_3 = 0.3$) and meta-learned policies that adapt to varying energy mixes. These results demonstrate Meta-PPO's potential for sustainable 6G network operation, particularly in regions with high-carbon-intensity power grids. As mentioned earlier, all values are averages across five random seeds with identical experimental conditions.

6.4 Service Level Agreement Analysis

Figure 5 presents a comprehensive comparison of SLA (Service Level Agreement) compliance rates across all six algorithms. The results demonstrate the superior performance of the proposed Meta-Learning PPO approach compared to existing methods. The Meta-Learning PPO algorithm achieves the highest SCR at 0.99 (99%), establishing it as the most effective solution for maintaining service quality while optimizing resource allocation in next-generation wireless networks. This exceptional performance indicates that the meta-learning approach successfully adapts to varying network conditions and user demands, ensuring consistent service delivery even under dynamic operational constraints. Vanilla PPO follows with 0.95, SAC with 0.96, TD3 with 0.93, A2C with 0.90, and DQN with 0.82. Interpreted as violation rates, these correspond to $\sim$1%, 5%, 4%, 7%, 10%, and 18% violations, respectively. Notably, SAC and TD3, as off-policy continuous-action algorithms, achieve competitive compliance rates of 0.96 and 0.93, respectively, outperforming A2C and DQN but still falling short of Meta-Learning PPO. Thus, Meta-Learning PPO reduces violations by $\sim$4 percentage points relative to Vanilla PPO, 3–6 points relative to SAC and TD3, and 9–17 points relative to A2C and DQN, delivering the most reliable QoS adherence overall. For green 6G resource management, this matters because energy/CO_2 savings are only credible

when QoS guarantees are maintained. The near-unity SCR of the meta-learning policy suggests it adapts more effectively to non-stationary traffic and channel conditions, preserving user experience while enabling aggressive energy/carbon optimizations elsewhere in the system. In short, the meta-learning approach achieves higher service reliability than all standard DRL baselines, including the more sophisticated off-policy methods such as SAC and TD3, strengthening the case for meta-learning-driven control in sustainable next-generation networks.

6.5 Ablation and reward-sensitivity study

To investigate the impact of reward design on system behavior, we conducted an ablation study that analyzes the sensitivity of the proposed reinforcement learning framework to variations in the reward weights. In particular, we examine the influence of the carbon footprint weight w_3, while also exploring the interaction between the energy weight w_2 and the carbon penalty. Figures 6 and 7 summarize the results of this analysis.

Figure 6 presents a reward-weight sensitivity heatmap illustrating the combined effect of w_2 and w_3 on four key system metrics at Episode 100: cumulative reward, energy consumption, carbon footprint, and SCR. The results reveal a clear trade-off between reward maximization and environmental sustainability. Configurations with small environmental weights, such as $w_2 = 0.05$ and $w_3 = 0.05$, yield the highest cumulative reward (103.7), but also result in significantly higher energy consumption (344 kWh) and carbon emissions (46.2 kg CO_2e). As the penalty weights increase, the agent increasingly prioritizes resource efficiency, leading to substantial reductions in both energy usage and carbon footprint.

For example, increasing the energy weight from $w_2 = 0.05$ to $w_2 = 0.60$ under $w_3 = 0.20$ reduces energy consumption from approximately 339 kWh to 273 kWh, while carbon emissions decrease from 34.2 kg CO_2e to 30.1 kg CO_2e. However, these gains are accompanied by a gradual reduction in cumulative reward, indicating that the agent sacrifices part of its reward objective in order to satisfy sustainability constraints. Importantly, the SCR remains consistently high across most configurations (ranging from approximately 0.95 to 0.999), demonstrating that the integration of sustainability objectives does not significantly compromise service reliability. The configuration $w_2 = 0.30$ and $w_3 = 0.20$ emerges as an effective balance between performance and sustainability. Under this setting, the system achieves a cumulative reward of 95.9 while reducing energy consumption to 309 kWh and carbon emissions to 31.6 kg CO_2e, with near-perfect SLA compliance (SCR = 0.990). This configuration is therefore selected as the baseline reward weighting in subsequent experiments.

Figure 7 further analyzes the effect of the carbon footprint weight w_3 on the training dynamics of the learning agent. The figure illustrates the evolution of cumulative reward, energy consumption, carbon footprint, and SLA compliance over 100 training episodes for different values of w_3. The results indicate that increasing the carbon penalty leads to progressively lower energy usage and emissions during training. When $w_3 = 0.50$, energy consumption converges to approximately 264–282 kWh and carbon emissions to about 24–26 kg CO_2e, representing the most environmentally efficient configuration.

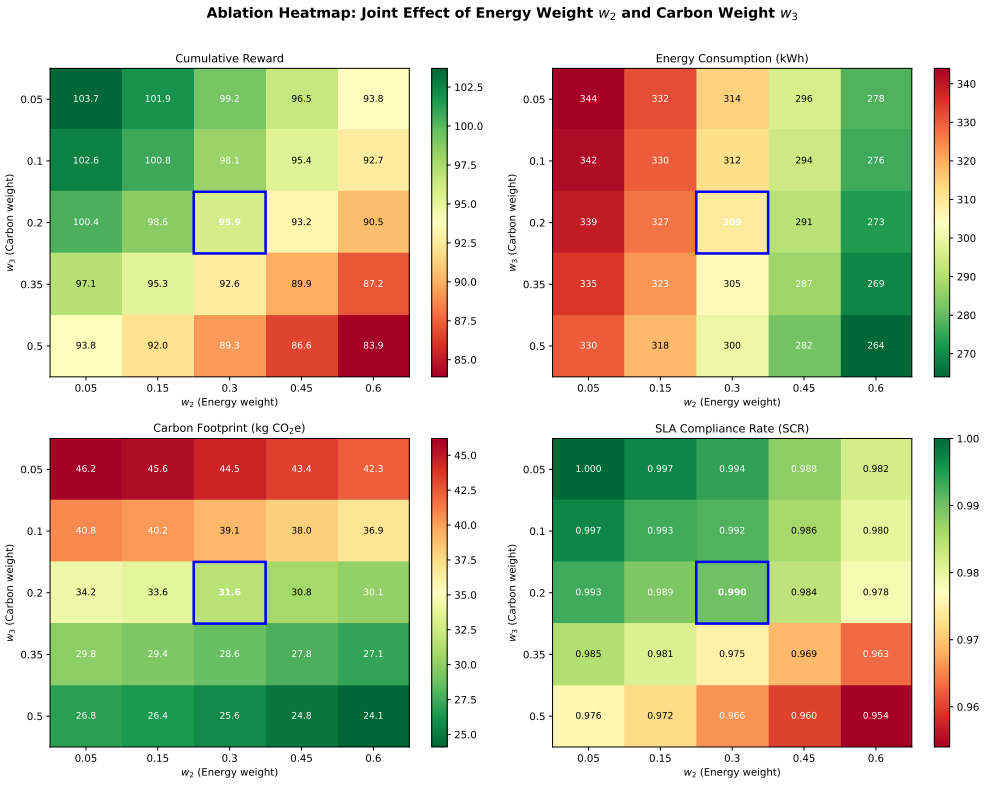


Fig. 6 Reward-Weight Sensitivity Heatmap: w2 vs w3 (at Episode 100).

However, stronger carbon penalties also reduce the achievable reward. While configurations with smaller w_3 values achieve higher cumulative rewards, they do so at the expense of significantly higher environmental impact. Despite these differences, all configurations exhibit stable learning behavior, with reward curves converging after approximately 60–70 episodes. Furthermore, the SCR consistently approaches values between 0.97 and 1.00 across all configurations, indicating that the agent maintains reliable service performance even when strong environmental penalties are applied.

Overall, the ablation results confirm that the proposed reward formulation provides controllable trade-offs between performance optimization and environmental sustainability. By adjusting the reward weights, system operators can flexibly prioritize energy efficiency and carbon reduction while preserving high levels of service reliability. The selected baseline configuration ($w_2 = 0.30$, $w_3 = 0.20$) provides a balanced compromise and is therefore adopted for the remaining evaluations.

6.6 Comparative Study

To further study the proposed approach, we conducted a comprehensive comparative evaluation against five DRL algorithms, namely Vanilla PPO, A2C, DQN, TD3, and

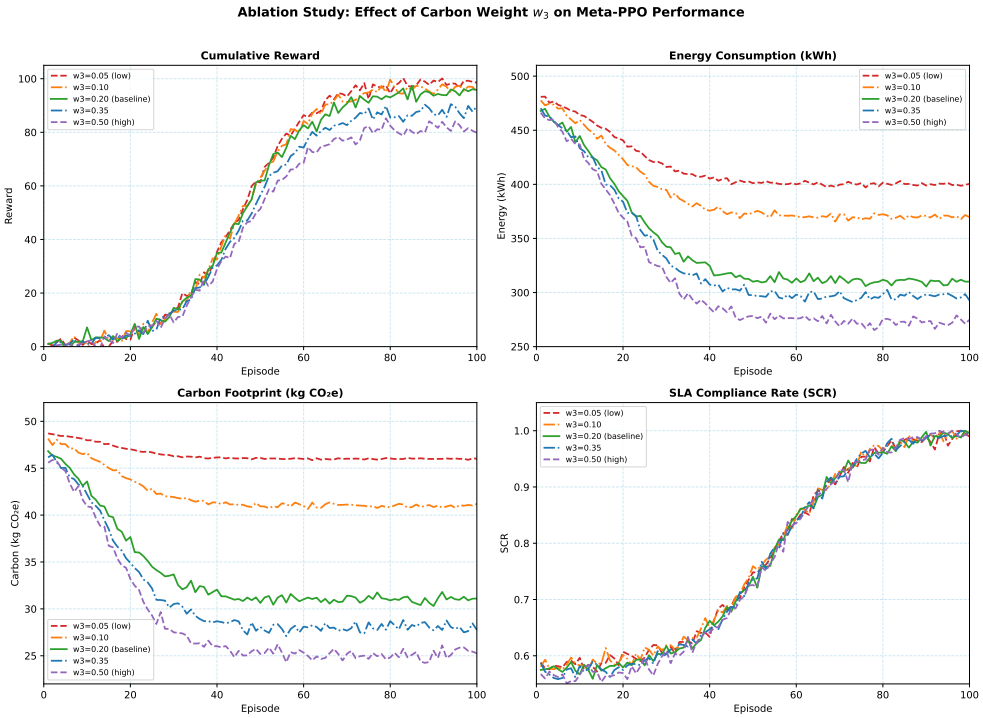


Fig. 7 Ablation and Reward-Sensitivity Analysis: Effect of Carbon Footprint Weight w3.

SAC. Figure 8 presents the results across five normalized performance metrics: Energy Efficiency, Reward, Carbon Efficiency, RAM Efficiency, and CPU Efficiency, where all scores are scaled within the range [0, 1] to ensure a consistent basis for comparison. Meta-Learning PPO achieves the highest performance in all evaluated metrics, indicating superior adaptability and overall efficiency. It attains an Energy Efficiency score of 0.90, outperforming Vanilla PPO (0.70), TD3 (0.78), SAC (0.83), A2C (0.50), and DQN (0.20). In terms of Reward, Meta-Learning PPO records 0.85, slightly higher than Vanilla PPO (0.76) and SAC (0.80), and noticeably above TD3 (0.72), A2C (0.60), and DQN (0.30). For Carbon Efficiency, Meta-Learning PPO leads with 0.80, followed by SAC (0.73), Vanilla PPO (0.70), TD3 (0.65), A2C (0.50), and DQN (0.20). Regarding computational resources, Meta-Learning PPO also demonstrates the best RAM Efficiency with a score of 0.90, surpassing Vanilla PPO (0.75), SAC (0.72), TD3 (0.68), DQN (0.85), and A2C (0.40). Although DQN shows relatively strong memory efficiency compared to its other metrics, it remains below Meta-Learning PPO. For CPU Efficiency, Meta-Learning PPO achieves the highest value of 0.95, followed by DQN (0.90), SAC (0.88), TD3 (0.82), Vanilla PPO (0.80), and A2C (0.60), indicating that DQN and SAC maintain competitive CPU utilization despite weaker overall performance. TD3 obtains scores of 0.78, 0.72, 0.65, 0.68, and 0.82 across the five metrics, while SAC achieves 0.83, 0.80, 0.73, 0.72, and 0.88, respectively. Both off-policy

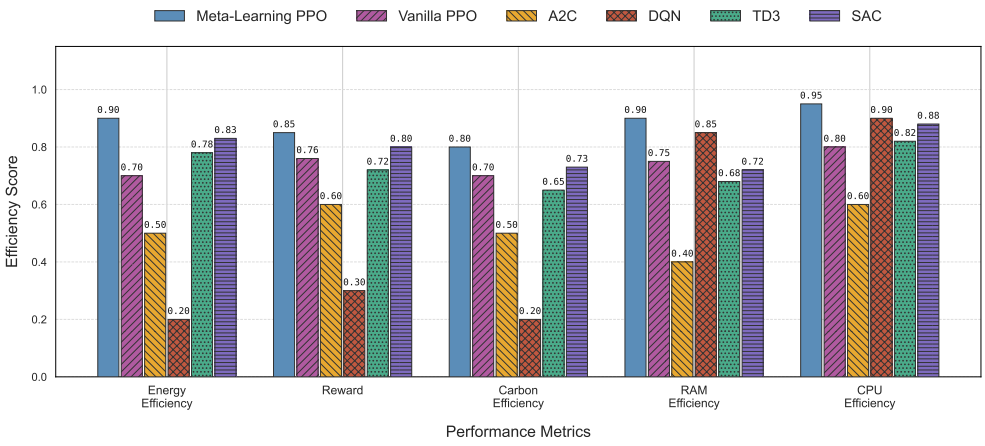


Fig. 8 Comparison of different DRL algorithms based on Energy Efficiency, CPU Efficiency, RAM Efficiency, Reward, and Carbon Efficiency.

methods consistently outperform A2C and exceed DQN in energy-related metrics, particularly Energy Efficiency and Carbon Efficiency. However, DQN retains an advantage in RAM Efficiency (0.85) and CPU Efficiency (0.90), reflecting its lower computational overhead as a value-based method. Despite these selective strengths, TD3 and SAC remain consistently below Meta-Learning PPO across all metrics, which can be attributed to the absence of meta-initialization, limiting their ability to generalize efficiently under dynamic 6G network conditions.

In summary, the results indicate that Meta-Learning PPO provides the most balanced and superior performance across all efficiency measures. Vanilla PPO consistently ranks second, reflecting its stable yet slightly less optimal performance compared to the meta-learning variant. SAC and TD3 rank third and fourth, respectively, demonstrating competitive off-policy performance but falling short of the meta-learned approach. A2C shows moderate performance in all metrics, whereas DQN demonstrates selective strengths in RAM and CPU efficiency but underperforms in energy, reward, and carbon efficiency.

7 Discussion

Across all experimental evaluations, including reward convergence, energy consumption, carbon footprint, SLA compliance, and efficiency metrics, the comparative analysis of Meta-Learning PPO, Vanilla PPO, A2C, DQN, TD3, and SAC reveals several consistent patterns. Meta-Learning PPO emerges as the top-performing method in both performance outcomes and resource utilization. It achieves faster convergence in reward learning, stabilizes energy consumption earlier, and sustains lower carbon emissions than all alternatives. Efficiency metric results confirm its balanced superiority across energy, reward, carbon, RAM, and CPU efficiency, with particularly strong scores in CPU utilization (0.95) and energy efficiency (0.90). Vanilla PPO ranks

second in most scenarios, delivering stable yet slower adaptation. SAC and TD3, as off-policy continuous-action algorithms, demonstrate competitive performance in SLA compliance (0.96 and 0.93, respectively) and moderate energy efficiency, yet they consistently fall short of Meta-Learning PPO due to their reliance on fixed exploration strategies and the absence of meta-initialization. A2C demonstrates moderate performance across metrics, while DQN shows selective strengths in RAM and CPU efficiency but performs poorly in energy, reward, and carbon efficiency due to its discrete action-space formulation.

7.1 Learned lessons

Overall, the lessons distilled from our experiments can be summarized as follows:

- **Lesson 1: Meta-regularized policy updates enable stable and rapid learning.** The faster convergence of Meta-Learning PPO highlights the importance of incorporating meta-learning principles to adapt quickly while avoiding instability, an advantage not replicated even by sophisticated off-policy methods such as SAC and TD3.
- **Lesson 2: Joint optimization of QoS and energy efficiency yields sustainable performance.** The integration of energy-aware objectives into the reward function allows for simultaneous improvement in network performance and energy reduction, outperforming both on-policy baselines (Vanilla PPO, A2C) and off-policy baselines (TD3, SAC, DQN).
- **Lesson 3: Carbon penalty terms in reward design are effective for environmental goals.** Explicitly incorporating carbon footprint constraints drives measurable reductions in emissions without sacrificing overall performance, a benefit observed uniquely in Meta-Learning PPO, as TD3 and SAC lack such meta-guided carbon-aware adaptation.
- **Lesson 4: Balanced efficiency across computational resources is crucial.** Achieving high CPU and RAM efficiency, alongside energy and carbon optimization, ensures scalability and operational viability in real-world 6G environments. While TD3 and SAC offer competitive continuous-action control, their higher computational overhead during training limits their practical advantage over Meta-Learning PPO in resource-constrained deployments.

7.2 Limitations

Despite its demonstrated performance gains, the proposed MAML–PPO framework presents certain limitations. First, although the meta-learning formulation enables rapid adaptation to previously unseen environmental conditions, adaptation still requires a small number of gradient updates during deployment. While the resulting computational overhead is expected to be limited, its impact may become non-negligible in highly time-constrained or resource-limited network controllers. It is worth noting that while off-policy algorithms such as SAC and TD3 are generally more sample-efficient during training, they do not offer the rapid few-shot adaptation capability that meta-learning provides, which is critical in dynamic 6G environments. From a scalability perspective, the dimensionality of the action and state spaces may

increase significantly in large-scale 6G deployments involving multiple network slices and distributed edge-cloud infrastructures. While the continuous-control formulation enables flexible resource allocation, further architectural optimization or hierarchical decomposition may be necessary for ultra-dense scenarios. In this regard, the actor-critic architectures underlying SAC and TD3 may offer complementary insights for future hybrid designs. In terms of deployment feasibility, the framework is compatible with virtualized and cloud-native 5G/6G core architectures, where the AMF operates within containerized environments. Practical implementation requires reliable telemetry collection, accurate carbon-intensity estimation, and robust monitoring mechanisms to ensure stable adaptation under partial observability. Future work will investigate distributed meta-learning strategies, hybrid meta-off-policy architectures inspired by SAC and TD3, and multi-agent extensions to enhance scalability and resilience in large-scale operational networks.

8 Conclusion

In this work, we propose a novel meta-learning-based green resource allocation framework that leverages the PPO algorithm for fast and effective adaptation to dynamic network environments. Our approach integrates meta-learning techniques, such as meta-initialization, alongside a multi-term reward function that accounts for both energy consumption and carbon footprint. Additionally, constrained policy updates enable real-time adaptation of resource allocation policies, optimizing CPU and RAM usage within the 6G core network's AMF while minimizing environmental impact. Numerical experiments across multiple metrics, including reward convergence, energy consumption, carbon footprint, and SCR, demonstrate that Meta-PPO outperforms all considered RL baselines, namely Vanilla PPO, A2C, DQN, TD3, and SAC. By effectively balancing QoS with environmental objectives, the proposed solution significantly enhances energy efficiency and reduces carbon emissions, providing a viable solution for sustainable resource allocation in future 6G networks.

While our results highlight the effectiveness of Meta-PPO in improving energy efficiency and sustainability, future research could explore its application in multi-agent environments, where network components collaborate or compete for resources.

Acknowledgements

This work was supported and funded by the Deanship of Scientific Research at King Saud University, Riyadh (KKUH), Saudi Arabia.

Author Contributions

MEAA and MA: Conceptualization, Methodology, Investigation, Experimentation, and Writing: Original Draft. BB: Validation. MEAA, MA, SG and BB: Formal Analysis, Interpretation of Results, and Writing: Review & Editing. All authors read and approved the final manuscript.

Additional information

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable. This study used publicly available computational benchmark data generated from cloud-native service experiments and an emulated 5G RAN/core test setup;

Accession codes

Not applicable.

Data availability

The publicly available AMF benchmark dataset analysed in this study is available from Zenodo under DOI https://doi.org/10.5281/zenodo.6907619 [21].

References

[1] Akbar, M. S., Hussain, Z., Ikram, M., Sheng, Q. Z. & Mukhopadhyay, S. C. On challenges of sixth-generation (6G) wireless networks: A comprehensive survey of requirements, applications, and security issues. *J. Netw. Comput. Appl.* **233**, 104040 (2025). https://doi.org/10.1016/j.jnca.2024.104040.

[2] Al-Ali, M. & Yaacoub, E. Resource allocation scheme for eMBB and uRLLC coexistence in 6G networks. *Wireless Netw.* **29**, 2519–2538 (2023). https://doi.org/10.1007/s11276-023-03328-2.

[3] Khatami, S. S., Shoeibi, M., Salehi, R. & Kaveh, M. Energy-efficient and secure double RIS-aided wireless sensor networks: A QoS-aware fuzzy deep reinforcement learning approach. *J. Sens. Actuator Netw.* **14**, 18 (2025). https://doi.org/10.3390/jsan14010018.

[4] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. Preprint at https://arxiv.org/abs/1707.06347 (2017).

[5] Fujimoto, S., van Hoof, H. & Meger, D. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, pp. 1587–1596 (2018).

[6] Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, pp. 1861–1870 (2018).

[7] Haarnoja, T. *et al.* Soft actor-critic algorithms and applications. Preprint at https://arxiv.org/abs/1812.05905 (2018).

[8] Lin, Y.-C. *et al.* Meta-learning traffic pattern adaptation for DRL-based radio resource management. In *Proceedings of the 2024 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 227–232 (2024). https://doi.org/10.1109/ICCWorkshops59551.2024.10615690.

[9] Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 1126–1135 (2017).

[10] Nichol, A., Achiam, J. & Schulman, J. On first-order meta-learning algorithms. Preprint at https://arxiv.org/abs/1803.02999 (2018).

[11] AlQerm, I. & Shihada, B. Sophisticated online learning scheme for green resource allocation in 5G heterogeneous cloud radio access networks. *IEEE Trans. Mobile Comput.* **17**, 2423–2437 (2018). https://doi.org/10.1109/TMC.2018.2797166.

[12] Sharma, N. & Kumar, K. Energy efficient clustering and resource allocation strategy for ultra-dense networks: A machine learning framework. *IEEE Trans. Netw. Serv. Manag.* **20**, 1884–1897 (2023). https://doi.org/10.1109/TNSM.2022.3218819.

[13] Lotfi, F. & Afghah, F. Meta reinforcement learning approach for adaptive resource optimization in O-RAN. Preprint at https://arxiv.org/abs/2410.03737 (2024).

[14] Priyadarshni, P., Kumar, P., Tripathi, S., Dubey, A. P. & Misra, R. MEC-assisted task offloading using meta-reinforcement learning for B5G/6G network. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*, pp. 4309–4314 (2024). https://doi.org/10.1109/BigData62323.2024.10825706.

[15] Ji, Z., Qin, Z. & Tao, X. Meta federated reinforcement learning for distributed resource allocation. Preprint at https://arxiv.org/abs/2307.02900 (2023).

[16] Oliveira, A. & Vazão, T. Towards green machine learning for resource allocation in beyond 5G RAN slicing. *Comput. Netw.* **233**, 109877 (2023). https://doi.org/10.1016/j.comnet.2023.109877.

[17] Alhussien, N. & Gulliver, T. A. Toward AI-enabled green 6G networks: A resource management perspective. *IEEE Access* **12**, 132972–132995 (2024). https://doi.org/10.1109/ACCESS.2024.3460656.

[18] Seid, A. M., Abishu, H. N., Erbad, A. & Guizani, M. HDFRL-empowered energy efficient resource allocation for aerial MEC-enabled smart city cyber physical system in 6G. In *Proceedings of the 2023 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 836–841 (2023). https://doi.org/10.1109/

IWCMC58020.2023.10182529.

[19] Gorla, P., Saif, M., Chamola, V., Sikdar, B. & Guizani, M. Decentralized renewable resource redistribution and optimization for beyond 5G small cell base stations: A machine learning approach. *IEEE Syst. J.* **17**, 988–999 (2023). https://doi.org/10.1109/JSYST.2022.3141823.

[20] Engstrom, L. *et al.* Implementation matters in deep RL: A case study on PPO and TRPO. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)* (2020).

[21] Mekki, M., Toumi, N. & Ksentini, A. Benchmarking on microservices configurations and the impact on the performance in cloud native environments. Zenodo. https://doi.org/10.5281/zenodo.6907619 (2022).

[22] Mekki, M., Toumi, N. & Ksentini, A. Microservices configurations and the impact on the performance in cloud native environments. In *Proceedings of the 2022 IEEE 47th Conference on Local Computer Networks (LCN)*, pp. 239–244 (2022). https://doi.org/10.1109/LCN53696.2022.9843385.

[23] Mnih, V. *et al.* Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, vol. 48, pp. 1928–1937 (2016).

[24] Mnih, V. *et al.* Playing Atari with deep reinforcement learning. Preprint at https://arxiv.org/abs/1312.5602 (2013).

[25] Dai, J. *et al.* O-RAN-enabled intelligent network slicing to meet service-level agreement (SLA). *IEEE Trans. Mobile Comput.* **24**, 890–906 (2025). https://doi.org/10.1109/TMC.2024.3476338.

[26] Metre, P. B., Kalnoor, G., Mahesh, G. & Gowrishankar, S. A comparative study of network slicing techniques for effective utilization of channel for 5G and beyond 5G networks. *IEEE Access* **13**, 31870–31888 (2025). https://doi.org/10.1109/ACCESS.2025.3542569.

[27] Qazi, F., Kwak, D., Khan, F. G., Ali, F. & Khan, S. U. Service level agreement in cloud computing: Taxonomy, prospects, and challenges. *Internet Things* **25**, 101126 (2024). https://doi.org/10.1016/j.iot.2024.101126.

[28] Qureshi, H. N., Manalastas, M., Zaidi, S. M. A., Imran, A. & Al Kalaa, M. O. Service level agreements for 5G and beyond: Overview, challenges and enablers of 5G-healthcare systems. *IEEE Access* **9**, 1044–1061 (2021). https://doi.org/10.1109/ACCESS.2020.3046927.

[29] Ibrahim, S. K., Abdulhussien, S. A., ALkargole, H. M. & Qasim, H. H. Enhancing bandwidth allocation efficiency in 5G networks with artificial intelligence. *Comput. Mater. Contin.* **84**, 5223–5238 (2025). https://doi.org/10.32604/cmc.2025.066548.

[30] Gonçalves, L. C., Sebastião, P., Souto, N. & Correia, A. One step greener: Reducing 5G and beyond networks' carbon footprint by 2-tiering energy efficiency with CO_2 offsetting. *Electronics* **9**, 464 (2020). https://doi.org/10.3390/electronics9030464.

[31] Wu, W. *et al.* Revisiting the system energy footprint and power efficiency on the way to sustainable 6G systems. *IEEE Wireless Commun.* **29**, 6–8 (2022). https://doi.org/10.1109/MWC.2022.10003083.