IEEE Open Journal of the
Communications Society

Received 20 April 2026; accepted 11 May 2026. Date of publication 13 May 2026; date of current version 9 June 2026.
Digital Object Identifier 10.1109/OJCOMS.2026.3693148

Federated Hierarchical Reinforcement Learning for Resilient Spectrum Sharing in 6G Non-Terrestrial Networks

SADIA MAVERICK KHAF (Graduate Student Member, IEEE) AND GEORGES KADDOUM (Senior Member, IEEE)

École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada

CORRESPONDING AUTHOR: S. KHAF (sadiakhaf@ieee.org)

This work is supported in part by FRQNT under Grant 305094 and in part by the Canada Research Chair Program Tier-II Entitled "Towards a Novel and Intelligent Framework for the Next Generations of IoT Networks".

ABSTRACT High-altitude platform stations (HAPS) offer a pragmatic backbone for extending 6G non-terrestrial networks (NTNs) into dense urban cores, sparsely connected remote regions, and disaster-stricken areas. In these environments, cognitive radio networks (CRNs) must arbitrate spectrum access without trusting all agents or relying on intact terrestrial backhaul. Accordingly, in the present study, we propose a federated hierarchical reinforcement learning (F-HRL) framework where secondary users (SUs) cooperatively access primary users' (PUs) spectrum, ground next-generation Node Bs (gNBs) and uncrewed aerial vehicles (UAVs) enforce trust-aware scheduling, while HAPS coordinators perform Byzantine-resilient aggregation using satellite backhaul. Area-specific reward shaping allows the same policy stack to prioritize spectral efficiency in urban deployments, coverage in remote regions, and rapid recovery after infrastructure loss. We benchmark F-HRL against non-cognitive baselines, simple heuristics, and the hierarchical deep reinforcement learning framework across the following three key scenarios over 5 000 decision steps: nominal spectrum sharing, disaster recovery with a gNB outage, and adversarial environments with malicious SUs. Experimental results demonstrate that F-HRL achieves a nearly 8-fold higher spectrum efficiency than the non-cognitive operation, maintains 86% SU quality-of-service (QoS) satisfaction with a Jain fairness index of 0.950, and sustains over 95% of nominal throughput during infrastructure failures. Under attack conditions with 25% compromised agents, trust-weighted aggregation successfully isolates malicious updates while preserving legitimate throughput. These findings demonstrate that HAPS-centric F-HRL provides adaptive, secure, and resilient spectrum management for future NTN deployments.

INDEX TERMS Cognitive radio, disaster recovery, federated learning, high altitude platform station, non-terrestrial networks, reinforcement learning, transfer learning, spectrum sharing.

I. INTRODUCTION

THE evolution towards 6G wireless networks requires unprecedented capabilities such as ultra-low latency on the order of microseconds [1], ultra-high-speed and massive connectivity, as well as major gains in energy efficiency as compared to 5G [2]. To achieve these goals, it is essential to ensure seamless integration between terrestrial and non-terrestrial segments, forming multi-layered non-terrestrial networks (NTNs) that enhance service continuity, resilience, and resource utilization [3], [4].

Key performance improvements are measured using Key Performance Indicators (KPIs), while broader societal benefits are quantified through Key Value Indicators (KVIs). Meeting these metrics requires addressing the following three interdependent challenges: (i) sustaining licensed services in dense urban cells; (ii) extending coverage to remote or under-served regions; and (iii) restoring connectivity when terrestrial infrastructure is compromised [5].

However, these requirements are difficult to meet with conventional cognitive radio approaches that assume always-on terrestrial backhaul and complete channel state information. When next-generation Node Bs (gNBs) are disabled, sensing reports are withheld and malicious secondary users (SUs) can falsify observations, which may lead to centralized coordination collapses [6], [7], [8]. High-altitude platform stations (HAPS) operating in the lower stratosphere offer a compelling backbone for resilient cognition [3], [9], [10]. A single HAPS can supervise multiple clusters of gNBs, uncrewed aerial vehicles (UAVs), and SUs while coordinating with neighbouring clusters through satellite backhaul [11]. However, disaster-struck urban cores, sparse remote cells, and mixed civilian–public-safety traffic inflict significant heterogeneity, thus requiring adaptive, privacy-preserving decision making beyond static policies.

Furthermore, deterministic coordination becomes brittle under infrastructure loss, privacy constraints, or adversarial interference, with centralized cognitive radio networks (CRNs) struggling to maintain reliable service when base stations fail or when sensing is falsified [12], [13]. This suggests that first responders and mission-critical Internet-of-Things (IoT) installations require autonomous, decentralized, and learning-driven spectrum management where both decision making and coordination dynamically evolve under uncertainty [14], [15].

Adaptive resource and spectrum management in dynamic NTNs increasingly relies on machine learning, especially the combination of Federated Learning (FL) and Reinforcement Learning (RL) [3], [4], [16], [17]. Recent hierarchical multi-agent federated DRL was proposed for TN-NTN resource allocation and task offloading with auction-based incentives [18], while unified distributed ML architectures (FSTL/GFSTL) target split/federated/transfer learning pipelines for 6G intelligent transportation data, rather than control policies [19]. Furthermore, GAIL-powered federated inverse RL was previously applied to multi-satellite beamforming and spectrum allocation using a matching-game formulation with expert demonstrations [20]. Accordingly, in this paper, we propose a federated hierarchical reinforcement learning (F-HRL) framework designed to address the non-stationary, heterogeneous, and security-critical nature of 6G NTNs. This framework extends capabilities of traditional FL–RL models by integrating multi-tier coordination, trust-aware learning, and transfer-based adaptation.

Critically, F-HRL differs from existing hierarchical DRL and federated RL approaches in the following three respects: (i) it decomposes spectrum management across four distinct tiers (SU, gNB, UAV, and HAPS) with tier-specific RL formulations and area-specific reward shaping, rather than a flat or two-tier hierarchy; (ii) it integrates sensing-layer majority-vote fusion with federated divergence-based trust weighting, whereas prior FL-based methods assumed honest participants or treat trust in isolation; and (iii) it incorporates context-aware transfer learning for zero-shot UAV deployment during disasters, which is lacking in the previously proposed federated or hierarchical RL frameworks.

To overcome these limitations, in this study, we propose an F-HRL system that combines Multi-Agent Reinforcement Learning (MARL) for distributed policy learning, Federated Learning (FL) for privacy-preserving coordination, and Transfer Learning (TL) for rapid adaptation in new or emergency deployments. The proposed system is shown in Fig. 1. Section IV details how each tier executes tailored learning policies coordinated through low-Earth orbit (LEO) satellite backhaul.

The primary contributions of this work are as follows:

- **Cross-tier cognitive spectrum management**: We developed a four-tier cognitive F-HRL framework spanning SUs, gNBs, UAVs, and HAPS that adapts to heterogeneous quality-of-service (QoS) objectives and constraints. SUs learn opportunistic access while gNBs enforce primary-user (PU) priority, unlike prior hierarchical approaches that treated PUs and SUs equivalently.
- **Federated coordination via HAPS aggregation**: We integrated periodic model aggregation at the HAPS tier to stabilize multi-agent learning across regions, addressing non-stationarity in MARL.
- **Disaster-resilient UAV failover**: We developed a disaster recovery framework that leveraged context-aware similarity-based selective model sharing for UAV deployment upon gNB outage to support emergency relief scenarios.
- **Trust-aware adversarial detection**: We introduced a Byzantine-robust aggregation and trust-weighted cooperative sensing that combines majority-vote sensing consistency with federated weight-divergence filtering to withstand falsified reports, model poisoning, and malicious agents attacks without sharing raw data.

Collectively, these contributions address persistent gaps where existing FL- and RL-based solutions confine learning to a single tier, ignore disaster-induced outages, or lack defenses against adversaries. By contrast, the proposed approach enables rapid deployment in emergency scenarios while maintaining security, privacy, and performance guarantees essential for critical communication infrastructure.

The remainder of this paper is organized as follows. Section II reviews related research in CRNs, FL, and HAPS-assisted communications. In Section III, we present the system model, network architecture, and problem formulation. Section IV details the proposed F-HRL framework, including security and TL mechanisms. Section V discusses experimental validation and performance analysis. Finally, Section VI concludes the paper with an outline of future research directions.

II. RELATED WORK

To motivate F-HRL, we group previous research efforts into three key themes related to the critical gaps addressed by our unified architecture.

A. MULTI-TIER HIERARCHICAL COORDINATION

Recent studies demonstrated the benefits of combining FL and DRL for adaptive spectrum access, resource management, and joint optimization across complex wireless architectures. Integrating massive antenna arrays (e.g., in HAPS) and managing large, heterogeneous networks frequently leads to a "curse of dimensionality" for optimization algorithms [9], [21]. HRL approaches tackle this complexity by introducing temporal abstraction via options and macros,

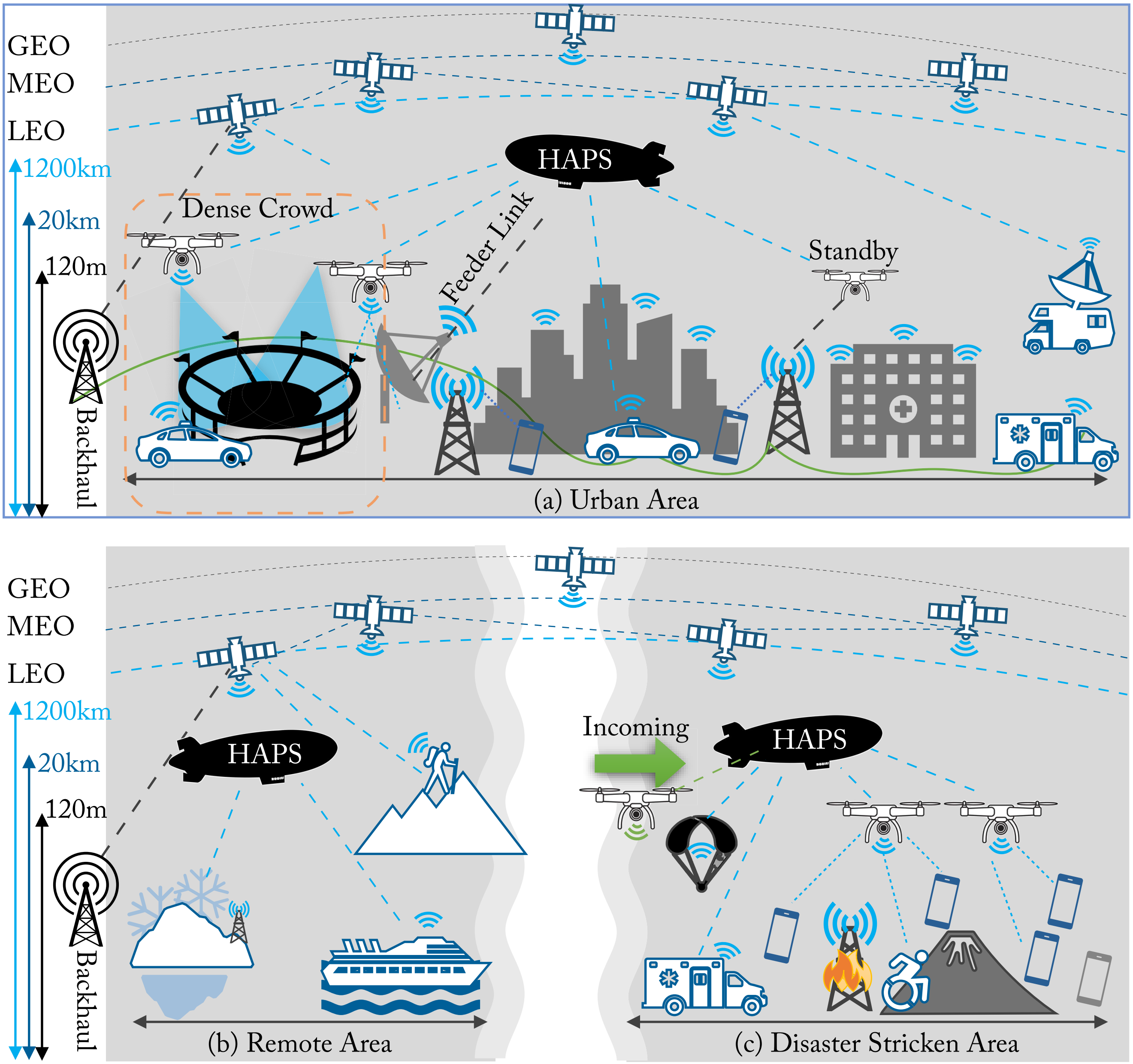

FIGURE 1. **HAPS-assisted multi-tier NTN architecture in three representative scenarios: (a) Urban area with dense crowd and standby UAVs, (b) Remote area with extended coverage via HAPS, and (c) Disaster-stricken area where HAPS and UAVs restore connectivity to PUs and SUs.**

modeling large problems as Semi-Markov Decision Processes (SMDPs). For resource scheduling in HAPS-integrated networks, HDRL reduces complexity by delegating tasks, enabling fast execution (e.g., 50× faster than exhaustive search), while maintaining up to 95% of the spectral efficiency achieved by exhaustive search. HDRL was also used for resource scheduling, UE association, and channel assignment in HAPS systems.

Previous architectural surveys and design studies argued that NTN deployments must explicitly balance cross-tier coordination, communication overhead, and energy objectives [22], [23]. Recent FL frameworks for satellite and LEO scenarios demonstrated viability of federated coordination under constrained backhaul [10] and energy-aware client participation [24]. In terrestrial IoT settings, reinforcement learning was paired with FL to improve routing fairness and convergence [25], thus reinforcing the need for multi-agent formulations embraced in the present study.

Furthermore, techniques such as Federated Deep Deterministic Policy Gradient (F-DDPG) were previously proposed for satellite Reconfigurable Intelligent Surface (RIS)-enhanced Integrated Sensing and Communication (ISAC) systems [26], yielding significant gains (e.g., up to

76.8% improvement in spectral efficiency) as compared to non-federated baselines. Federated DRL algorithms were reported to perform comparably to centralized learning while achieving faster execution times [16], [27].

While cognitive radio techniques have been extensively studied under practical imperfections [28] and hardware constraints [29], applying DRL approaches such as Deep Q-Networks (DQNs) to HAPS systems for dynamic resource allocation remains limited by the "curse of dimensionality," arising from large coverage areas and massive antenna arrays. To mitigate this challenge, complexity-reduction strategies such as HRL and HDRL were explored [3], [30], decomposing large action spaces into smaller sub-tasks, or *options* [17], [21].

Recent HDRL frameworks employing PPO-based policies across multiple tiers have demonstrated effective hierarchical coordination for joint UE association, channel assignment, and resource scheduling in HAPS-integrated networks [3]. Similarly, hierarchical multi-agent federated DRL was previously proposed for TN-NTN resource allocation and task offloading with auction-based incentives [18], while unified distributed ML architectures (FSTL/GFSTL) focus on split/federated/transfer learning pipelines for 6G intelligent transportation data rather than control policies [19]. Despite these advances, existing HDRL approaches do not explicitly address adversarial robustness, distributed trust, or coordinated disaster recovery across heterogeneous NTN layers.

It is worth contrasting F-HRL with two additional categories of approaches. First, multi-agent PPO (MA-PPO) methods employ independent PPO agents with shared centralized critics, but lack federated aggregation for cross-region knowledge transfer and trust-weighted Byzantine resilience. Second, optimization-based methods such as the Whale Optimization Algorithm (WOA) can effectively solve per-step allocation problems, but are inherently unable to adapt to non-stationary PU traffic or infrastructure failures without re-optimization. Both categories are evaluated as baselines in Section V.

B. RESILIENCE AND RAPID DEPLOYMENT

Mission-critical applications require the network to maintain connectivity despite mobility, extreme congestion, or infrastructure loss. DRL and FL are critical for managing dynamic and mobile environments, particularly in multi-tiered NTNs involving UAVs and ground stations.

Providing flexible coverage extension and serving as temporary gNBs, UAVs are essential in 5G. Due to the high mobility of UAVs, fast handover management is crucial. Dynamic Spectrum Sharing (DSS) based on FL and DRL was previously proposed for emergency CRNs to deal with spectrum shortages after disasters [31]. Federated UAV coordination was also explored for dense IoT deployments, combining hypergraph models with FRL to stabilize throughput [32], while over-the-air transfer learning across UAV swarms was reported to improve modulation recognition yet assumes trusted agents [33].

While DRL/FL is applied to emergency scenarios, previous studies generally assumed static infrastructure or reliable ad-hoc relaying, thus failing to investigate the challenge of sequential infrastructure failure such as multiple gNB outages—and the subsequent need for rapid system reconfiguration. Furthermore, implementing DL algorithms in real-time CRNs faces the challenge that most modern DL systems are trained offline, which results in the lack of verified generalization ability in complicated and dynamic physical channels.

Furthermore, previous HDRL frameworks addressed cross-tier coordination by imposing constraints–i.e., higher tiers make decisions that constrain local tiers, ensuring coordinated spectrum use and managing dynamic movement [34].

Current federated and hierarchical frameworks do not integrate explicit mechanisms for rapid redeployment and knowledge transfer, i.e., transfer learning, across non-stationary topologies. The lack of reliable terrestrial infrastructure in disaster zones prevents the use of traditional centralized retraining. F-HRL addresses this concern by using pre-trained base models across representative conditions (urban, remote, disaster) and using context-similarity scoring to rapidly fine-tune the model upon deployment or failure detection. This mechanism allows UAV-assisted coordination to quickly restore over 85% of pre-failure throughput after sequential gNB outages [31], [35].

C. TRUST, ROBUSTNESS, AND SECURITY IN FEDERATED RL

Distributed learning in critical infrastructure requires resilience against adversarial behavior–a concern magnified in CRNs where agents opportunistically access shared spectrum. Unless tensor-based detectors and trust scoring are incorporated, cooperative sensing remains vulnerable to falsified observations and hidden terminal effects [6], [7]. While vertical FL reduces data leakage by partitioning features across participants [36] and deep recurrent methods handle partial observability in heterogeneous networks [37], these approaches typically assume honest participants.

Although recent efforts deployed FRL to mitigate jamming and noisy gradients [38], larger-scale surveys stressed that existing countermeasures treat isolated threat vectors [4]. Byzantine-resilient aggregation and communication-efficient FRL were proposed to tolerate poisoned updates and data heterogeneity [39], [40], while NTN security studies advocated for coupling spectrum control with anomaly detection [41]. These works rarely integrated sensing trust with hierarchical recovery, thus reinforcing the need for the unified trust-weighted aggregation that underpins our F-HRL stack.

Based on a literature review conducted through the joint lens of coordination depth, adaptation strategy, and resilience features, in Table 1, we compare representative studies across these facets. The catalogue spans contributions published in 2020–2025 covering satellite RIS, UAV swarms, vehicular CRNs, and HAPS-assisted deployments. This perspective highlights that most prior efforts optimized a single tier or a

TABLE 1. Representative Studies Grouped by Coordination Scope, Adaptation Strategy, and Resilience Mechanisms.

Study	Coordination scope	Adaptation strategy	Resilience / trust handling
Pala et al. [26]	Satellite–RIS link (single tier)	Federated DDPG for RIS beamforming; spectral-efficiency maximization	No explicit failover or security model; assumes benign participants and intact infrastructure.
Ahmadzadeh et al. [16]	RIS-aided CRN (single tier)	Over-the-air FL with DRL for robust RIS-aided cognitive radio	Addresses channel distortion during aggregation, but omits adversarial, trust, or disaster considerations.
Yang et al. [31]	UAV-assisted emergency CRN (dual tier)	Federated multi-agent Q-learning for post-disaster spectrum sharing	Restores traffic after a single outage, but lacks trust modelling and sequential failure analysis.
Hu et al. [34]	UAV relay + gNB hierarchy	Hierarchical actor–critic with cache-aware scheduling	Ensures QoS under mobility, yet depends on centralized coordination, without federated robustness.
Tehrani et al. [2]	Distributed gNB cluster	Federated DRL for coordinated NextG control loops	Boosts distributed control, yet assumes stable backhaul and trusted participants.
Ning et al. [12]	Cooperative CRN sensing	Centralized RL with cooperative feedback loops	Handles sensing uncertainty, but no federation or hierarchical fallback paths.
Farajzadeh et al. [23]	NTN architectural survey	Blueprint for federated orchestration across satellite and HAPS tiers	Highlights scalability/energy trade-offs, but leaves disaster adaptation open.
Naous et al. [22]	NTN survey (multi-tier)	Categorizes RL enablers for satellite/HAPS integration	Frames open problems, but does not deliver a deployed stack.
Sharma et al. [38]	gNB with cooperative SUs	Federated D3QN for anti-jamming power control	Mitigates jamming, but omits disaster recovery and multi-tier orchestration.
Zhou et al. [33]	UAV swarm for V2X monitoring	Over-the-air federated transfer learning for modulation recognition across mobile relays	Tackles mobility-induced drift, yet assumes trusted agents and no spectrum failover plan.
Zhang et al. [36]	Cooperative CRN sensors (single tier)	Vertical FL coordinating heterogeneous feature sets	Preserves sensing privacy, but leaves Byzantine behaviour and cross-tier recovery unaddressed.
Chen et al. [42]	Distributed SU cluster	Lightweight FL-based cooperative sensing with reliability weighting	Improves detection accuracy; resilience is limited to statistical weighting without dynamic trust feedback.
Khaf et al. [11]	HAPS-assisted terrestrial/UAV users	Capability-aware cooperation with partial hierarchy	Enhances coverage after single outages; lacks federated training and sustained multi-tier trust governance.
Le and Kaddoum [35]	Vehicular CRN (multi-agent SUs)	LSTM-based channel access with multi-agent RL	Addresses vehicular mobility, but independent agents, no federated aggregation or coordinated failover.
Yang et al. [32]	UAV-enabled IoT (dense)	FRL with hypergraph scheduling for dense UAV clusters	Improves throughput, but assumes reliable backhaul and honest agents.
Malik et al. [25]	CR-enabled IoT mesh	RL-based routing for fairness and load balancing	Focuses on terrestrial IoT without NTN layers or trust governance.
Getu et al. [6]	MIMO cognitive sensing	Tensor-based detection to mitigate hidden terminals	Enhances sensing fidelity, yet lacks federated coordination.
Zuo et al. [39]	FL aggregation layer	Byzantine-robust aggregation for heterogeneous clients	Addresses poisoning, but not spectrum context or multi-tier failover.
Krouka et al. [40]	Multi-agent FRL cluster	Communication-efficient consensus via ADMM	Improves communication cost, yet assumes trustworthy sensing inputs.
Das et al. [4]	Survey across FRL deployments	Taxonomy of privacy- and security-aware FL primitives	Identifies need for trust-weighted aggregation and cross-layer resilience, but offers no integrated implementation.
Seid et al. [18]	TN-NTN multi-tier	FeDRL/MAFDRL with auction-based incentives for resource allocation and task offloading	Targets incentive-driven offloading; no PU/SU spectrum sharing, trust weighting, or disaster recovery.
Naseh et al. [19]	TN-NTN ITS hierarchy	FSTL/GFSTL split–federated–transfer learning architecture for ITS data	Learning framework only; no RL spectrum policy, PU/SU coordination, or adversarial resilience.
Hassan et al. [20]	6G multi-satellite spectrum	GAIL-based federated inverse RL for beamforming and spectrum allocation with matching-game association	Focuses on satellite beamforming/SE, not CRN PU/SU coordination or disaster recovery.

TABLE 2. Summary of Literature Gaps and F-HRL Contributions.

Theme	Representative gaps	F-HRL Contributions
NTN FL (Complexity & Cross-Tier)	Single-agent DRL methods struggle with the exponentially large action spaces and state spaces in massive MIMO HAPS/NTNs. Centralized approaches incur significant computational bottlenecks and communication overhead.	Employs HAPS as a coordinating layer (Tier 4) for cross-tier adaptation (SUs, gNBs, UAVs), thus mitigating the curse of dimensionality via policy decomposition.
FL CRN (Resilience & Topology)	Available FL assumes static infrastructure or relies on constant cooperation. Current frameworks lack specific mechanisms for managing sequential infrastructure failures (disaster recovery) and adapting policies in non-stationary topologies.	Manages multiple sequential gNB outages. Achieves rapid adaptation and self-configuration using context-aware TL to fine-tune pre-trained models for emergency deployment.
RL CRN (Trust & Security)	Available DRL solutions handle threats (jamming or poisoning) in isolation. Secure model sharing remains an open challenge. Trust mechanisms are minimally integrated with sensing capabilities.	Implements robust trust-weighted aggregation and observation-consistency checks to suppress outlier updates, thus ensuring resilience against Byzantine nodes and sensing data falsification (SDF).
Scalability & Overhead	Previous FL-RL methods do not analyze communication overhead or transfer learning search cost in dense NTN deployments. Scalability to hundreds of agents remains unquantified.	Provides quantitative complexity analysis with concrete deployment numbers (Section IV-D), showing quadratic context cost is negligible for $N_h < 1000$ and identifying gradient compression for denser deployments.

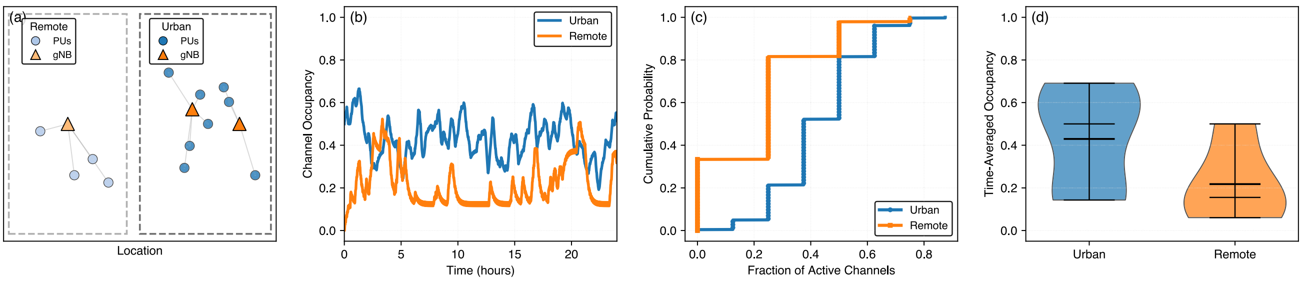

FIGURE 2. PU traffic statistics across urban and remote areas: (a) PU–gNB topology, (b) diurnal occupancy, (c) CDF of active-channel fraction, (d) per-PU time-averaged occupancy.

single threat model, thus leaving cross-tier failover and trust governance largely unexplored.

In Table 2, we summarize the principal literature gaps and highlight how the proposed approach addresses each dimension.

In contrast to the above literature, our F-HRL framework unifies cross-tier coordination, trust-aware aggregation, and disaster recovery within a single NTN-compatible architecture.

III. SYSTEM MODEL

We consider a heterogeneous wireless system integrating HAPSs, terrestrial gNBs, UAVs, and clustered user equipments (UEs) (see Fig. 1). HAPSs are deployed at altitudes of 20 km–50 km and act as regional spectrum coordinators and FL aggregators. Inter-HAPS coordination relies on low-Earth orbit (LEO) satellite backhaul, while medium Earth orbit and geostationary satellite layers are excluded due to latency exceeding the slot-level decision framework discussed in this paper.

UEs are divided into the following two sets: the set of PUs denoted by $\mathcal{U}_P$ and the set of SUs denoted by $\mathcal{U}_S$. Thus, the total UE set is $\mathcal{U} = \mathcal{U}_P \cup \mathcal{U}_S$ with $\mathcal{U}_P \cap \mathcal{U}_S = \emptyset$. PU activity is exogenous and bursty, generated by a mix of Markov-modulated Poisson process (MMPP) traffic, heavy-tailed Pareto on/off processes, and fixed traces. Fig. 2 summarizes PU traffic across the urban and remote areas: (a) the PU–gNB topology, (b) diurnal occupancy with urban maintaining 40–50% utilization and remote 15–25% with evening peaks, (c) a right-shifted CDF indicating consistently higher urban utilization relative to sparser remote activity, and (d) per-PU heterogeneity captured via a violin plot where the thin line denotes the median, the bold line the mean, and the width the density. Urban PUs show a broader occupancy variability, while remote PUs are more concentrated at low-to-moderate usage, highlighting the challenge for SUs to adapt to spatially heterogeneous traffic. PUs consume their allocations without adaptation, while SUs operate as cognitive users that sense occupancy and opportunistically transmit when no PU activity is detected.

A. SPECTRUM AND CHANNEL MODEL

Let $\mathcal{U}_{S,b}$ denote the SUs associated with gNB b, $\mathcal{K}_b$ the channel set at gNB b, $\mathcal{U}_d$ the UEs served by UAV d, $\mathcal{B}_h$ the gNBs coordinated by HAPS h, and $\mathcal{D}_h$ the UAVs managed by HAPS h. The global gNB and UAV sets are $\mathcal{B} = \cup_h \mathcal{B}_h$ and $\mathcal{D} = \cup_h \mathcal{D}_h$, respectively.

The network operates over 6 channels spanning 3GPP bands n77, n78, n79, and n258. Channel resource block capacities range from 50 to 113 RBs depending on bandwidth allocation (20–100 MHz); gNB scheduling capacities range from 106 to 273 RBs. Propagation follows 3GPP TR 38.901/38.811:

- Urban links: UMi LoS pathloss,

$$\mathrm{PL}_{\mathrm{UMi}} = 32.4 + 21\log_{10}(f_c[\mathrm{GHz}]) + 20\log_{10}(d_{3\mathrm{D}}[\mathrm{m}]) \tag{1}$$

- Remote links: RMa pathloss,

$$\mathrm{PL}_{\mathrm{RMa}} = 32.4 + 20\log_{10}(f_c[\mathrm{GHz}]) + 23.3\log_{10}(d_{3\mathrm{D}}[\mathrm{m}]) \tag{2}$$

Small-scale fading follows Rician distribution for HAPS–UE links with elevation-dependent K-factors, and Rayleigh distribution for terrestrial links under non-line-of-sight conditions. Shadowing is log-normal with standard deviation 4 dB for HAPS–UE links and 8 dB for terrestrial links. UAVs fly at the altitudes of 100 m–300 m and use the same channel model as gNBs.

The channel gain from UE u to its serving node on channel k is

$$g_{u,k} = \mathrm{PL}(d_{3\mathrm{D}})^{-1} \cdot |h_{u,k}|^2 \cdot 10^{-\chi/10}, \tag{3}$$

where $h_{u,k}$ captures small-scale fading (Rician or Rayleigh as above) and χ is log-normal shadowing with standard deviation 4 dB for HAPS links and 8 dB for terrestrial links. The SINR at UE u on channel k is then

$$\gamma_{u,k} = \frac{p_u\, g_{u,k}}{N_0 B_k + \sum_{j\neq u} p_j\, g_{j,k}}, \tag{4}$$

where N_0 is the noise power spectral density and the sum captures co-channel interference from other active transmitters. The achievable rate follows as $R_{u,k} = B_k \log_2(1 + \gamma_{u,k})$.

Each SU performs energy detection to sense channel occupancy, and collaborative SUs improve their accuracy using majority vote [15].

1) MODELING ASSUMPTIONS AND LIMITATIONS

The present model adopts several simplifying assumptions that we explicitly acknowledge below. First, the HAPS is treated as quasi-stationary, since stratospheric station-keeping maintains its position within a 400 m radius, which is negligible relative to its 20 km altitude. Second, while UAV positions update per decision step, intra-step Doppler effects are not modeled; this is justified for low-altitude UAVs at 10–30 m/s in sub-6 GHz bands where the coherence time (in the order of milliseconds) significantly exceeds the slot duration. Third, signaling overhead for handover is not explicitly modeled, but is quantified in terms of communication cost in Section IV. Extending the framework to incorporate detailed channel dynamics, including Doppler effects and mobility-induced handover overhead, is left for future work.

B. PROBLEM FORMULATION

We model the network as a partially observable Markov decision process (POMDP). The global state is as follows:

$$\mathbf{s} = \left\{ (\bar{\lambda}_b,\ \bar{\gamma}_b,\ I_b)_{b\in\mathcal{B}},\ \ (E_d)_{d\in\mathcal{D}} \right\}, \tag{5}$$

where $\bar{\lambda}_b$ is the aggregate spectrum demand at gNB b, $\bar{\gamma}_b$ is the average SINR at gNB b, I_b is the local interference indicator at gNB b, and E_d is the residual energy of UAV d.

Each tier $i \in \{h, b, d, u\}$ (HAPS, gNB, UAV, and SU, respectively) observes only a partial, and thus, imperfect view of $\mathbf{s}$:

$$\mathbf{o}_i^t = \mathbf{s}_i^t + \boldsymbol{\epsilon}_i^t, \tag{6}$$

where $\mathbf{s}_i^t \subset \mathbf{s}^t$ denotes the state components visible to tier i, and $\boldsymbol{\epsilon}_i^t \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I})$ captures measurement and sensing errors (e.g., imperfect CSI).

We decompose the overall spectrum management problem into four tier-specific subproblems as outlined below.

1) (S1) SU OPPORTUNISTIC ACCESS AND POWER CONTROL

Each SU u chooses transmission channel k_u and power p_u to maximize its achievable rate $R_{u,k_u} = R_{u,k}\big|_{k=k_u}$ as shown below.

$$\max_{k_u, p_u}\ R_{u,k_u} = B_{k_u} \log_2(1 + \gamma_{u,k_u}) \tag{7}$$

subject to k_u being vacant, $0 \le p_u \le P^{\max}$, and $\gamma_{u,k_u} \ge \Gamma_{\min}$.

2) (S2) GNB LOCAL SCHEDULING

Each gNB b schedules SUs on residual capacity after satisfying PU demand:

$$\max_{x_{u,k}} \sum_{u\in\mathcal{U}_{S,b}} \sum_{k\in\mathcal{K}_b} x_{u,k} R_{u,k} \tag{8}$$

subject to PU-priority allocation, $\sum_u x_{u,k} \le 1$, $I_k^{\mathrm{PU}} \le I_k^{\max}$, and $x_{u,k} \in \{0, 1\}$.

3) (S3) UAV PLACEMENT AND COVERAGE

Each UAV d determines position $\mathbf{p}_d$ to maximize coverage of displaced UEs:

$$\max_{\mathbf{p}_d} \sum_{u\in\mathcal{U}_d} w_u \mathbb{I}\{\mathrm{cov}_d(u)\} R_{u,k_u} \tag{9}$$

subject to activation $a_d \in \{0, 1\}$, $E_d \ge E_{\min}$, and backhaul constraints $C_d^{\mathrm{BH}} \ge C_{\min}$.

4) (S4) HAPS BACKHAUL ROUTING AND RECOVERY COORDINATION

Each HAPS routes admitted loads from gNBs and UAVs subject to backhaul capacity:

$$\max_{x_{b,h}, x_{d,h}, a_d} \sum_{b\in\mathcal{B}_h} x_{b,h}\, \bar{\lambda}_b + \sum_{d\in\mathcal{D}_h} x_{d,h}\, \Lambda_d \tag{10}$$

subject to

$$\sum_{b\in\mathcal{B}_h} x_{b,h}\, \bar{\lambda}_b + \sum_{d\in\mathcal{D}_h} x_{d,h}\, \Lambda_d \le C_h^{\mathrm{BH}}, \tag{11}$$

$$a_d \in \{0, 1\}, \quad x_{d,h} \le a_d, \quad x_{b,h}, x_{d,h} \in \{0, 1\}. \tag{12}$$

5) COUPLING CONDITIONS

Cross-tier consistency requires (i) each SU associated with exactly one node; (ii) admitted SU and UAV loads respecting HAPS backhaul; (iii) interference $I_{u,k}$ including co-channel transmissions across gNBs and UAVs, and (iv) PU protection $I_k^{\mathrm{PU}} \leq I_k^{\max}$.

6) CONSTRAINT ENFORCEMENT IN THE RL FRAMEWORK

The coupling constraints are enforced within the RL framework as follows. The HAPS backhaul capacity constraint (S4) is enforced via action masking: the HAPS-tier RL agent's discrete action space is restricted so that only routing decisions satisfying the backhaul capacity limit are presented to the policy. The binary routing and activation variables in the same subproblem are inherently satisfied by the discrete action space. For PU-priority constraints (S2), the gNB scheduler allocates PU demand first and offers only residual capacity to SUs. For S1, the SU action space enforces channel vacancy and power bounds via masking/clipping. For S3, UAV activation and routing binaries are enforced by discrete actions, with backhaul feasibility enforced by the HAPS routing mask. As a safety net, post-hoc constraint violation detection triggers a reward penalty when any constraint is violated during execution, reinforcing thus feasibility through learning.

7) REWARD SHAPING

Since S1–S4 are intractable to solve jointly, each tier optimizes a shaped reward that combines throughput, QoS satisfaction, collision penalties, and regional fairness. Urban regions emphasize fairness, remote regions prioritize coverage, and disaster regions increase the PU-protection weight. The penalty weights are held fixed throughout training. Table 4 summarizes the key simulation parameters.

In summary, the hierarchical structure ensures that PUs remain protected, SUs opportunistically exploit spectrum under constraints, UAVs provide resilience under failures, and, finally, HAPS orchestrates regional backhaul feasibility.

IV. PROPOSED SOLUTION

We design an F-HRL framework that coordinates spectrum access for UEs, terrestrial gNBs, UAV relays, and HAPS coordinators. The key objective is to maximize spectrum utilization and service resilience without centralizing raw observations, while concurrently maintaining explicit protection for licensed PUs.

A. HIERARCHICAL MARL WITH FEDERATED LEARNING

Each tier executes a learning component tailored to its sensing capability and control scope as follows:

- **SUs** run lightweight models that perform channel selection and power control using local measurements such as occupancy and SINR estimates.
- **gNBs** host RL-based schedulers that allocate RBs among associated SUs subject to PU-protection and congestion constraints, and maintain a centralized critic for CTDE training of local SUs.
- **UAVs** employ RL-based policies to choose positions and activation status, trading off coverage and energy consumption.

TABLE 3. Key Notations.

Symbol	Meaning
Network Sets	
$\mathcal{U}_P, \mathcal{U}_S$	Sets of PUs and SUs
$\mathcal{U}_{S,b}$	SUs associated with gNB b
$\mathcal{K}_b$	Channel set at gNB b
$\mathcal{U}_d$	UEs served by UAV d
$\mathcal{B}, \mathcal{D}$	Global sets of gNBs and UAVs
$\mathcal{B}_h, \mathcal{D}_h$	gNBs and UAVs coordinated by HAPS h
$\mathcal{R}_h$	Agent set participating in HAPS h federation round
N_h	Number of agents in HAPS region h
Decision Variables	
k_u, p_u	Channel and power for SU u
$x_{u,k}$	Scheduling indicator for SU u on channel k
$\mathbf{p}_d$	UAV position
$x_{b,h}, x_{d,h}$	Routing indicators via HAPS h
a_d	UAV activation binary
Channel and Signal	
$g_{u,k}$	Channel gain of UE u on channel k
N_0	Noise power spectral density
B_k	Bandwidth of channel k
$\gamma_{u,k}$	SINR of UE u on channel k
$R_{u,k}$	Achievable rate of UE u on channel k
System Parameters	
$\bar{\lambda}_b$	Spectrum demand at gNB b
Λ_d	Traffic load from UAV d
w_u	Priority weight for UE u
E_d	Residual energy of UAV d
C_d^{BH}	UAV backhaul capacity
C_h^{BH}	HAPS backhaul capacity
$\mathbf{o}_i^t$	Noisy observation received by tier-i agent at step t
$P^{\max}$	Maximum SU transmit power
I_b	Interference indicator at gNB b
I_k^{PU}	Interference observed at PU channel k
$I_{u,k}$	Aggregate co-channel interference at UE u on channel k
$E_{\min}$	Minimum energy threshold for UAVs
$\Gamma_{\min}$	Minimum SINR threshold
$I_k^{\max}$	Maximum interference threshold on channel k
$\bar{\gamma}_b$	Average SINR at gNB b
σ_e^2	Estimation error variance
$C_{\min}$	Minimum backhaul capacity
$\mathrm{cov}_d(u)$	Coverage indicator: UAV d covers UE u
τ	Federated aggregation period in steps

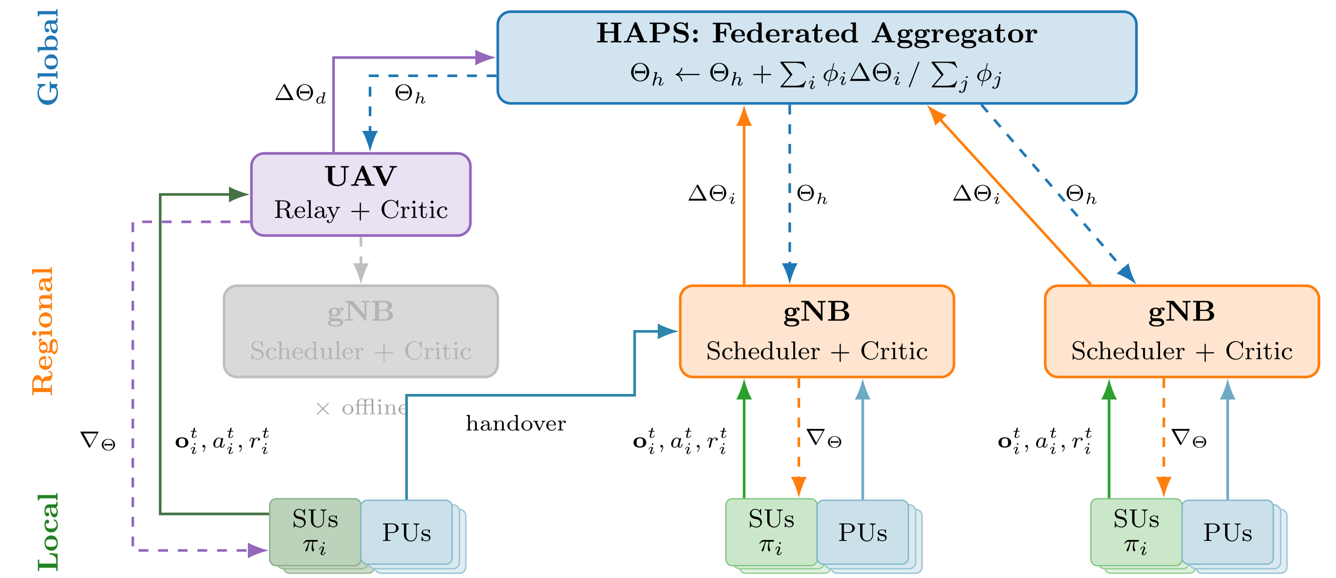

FIGURE 3. F-HRL learning architecture with disaster recovery. gNB_0 is offline; PUs and SUs are served by nearest available gNB and UAVs.

- **HAPS** nodes act as federated aggregators: aggregating regional deltas, they apply trust-aware weighting and dispatch adapted policies for redeployment or disaster response.

During operation, each agent i maps its partial observation $\mathbf{o}_i^t$ to action a_i^t through policy π_i, while the HAPS tier maintains global parameters Θ_h shared across regions.

Fig. 3 illustrates the architecture.

B. TWO-TIMESCALE CONVERGENCE

To mitigate non-stationarity in MARL, F-HRL employs two-timescale learning: fast local policy updates and slower HAPS-level aggregation every τ steps, yielding a quasi-stationary reference model [43], [44].

Each agent $i \in \mathcal{R}_h$ performs local updates

$$\Theta_i^{t+1} = \Theta_i^t - \alpha_t \nabla_{\Theta_i} \mathcal{L}_i(\Theta_i^t), \tag{13}$$

while the HAPS aggregator updates every τ steps as

$$\Theta_h^{r+1} = \Theta_h^r + \beta_r \frac{\sum_{i\in\mathcal{R}_h} \phi_i^{(r+1)}(\Theta_i^{(r)} - \Theta_h^{(r)})}{\sum_{j\in\mathcal{R}_h} \phi_j^{(r+1)}}. \tag{14}$$

Assume diminishing step sizes with $\sum_t \alpha_t = \sum_r \beta_r = \infty$, $\sum_t \alpha_t^2, \sum_r \beta_r^2 < \infty$, and $\beta_r/\alpha_{r\tau} \to 0$, bounded iterates, and Lipschitz losses $\mathcal{L}_i$.

Under these conditions, Θ_h^r evolves on a slower timescale and is quasi-static over each federated interval. The local iterates track the ODE $\dot{\Theta}_i = -\nabla_{\Theta_i}\mathcal{L}_i(\Theta_i; \Theta_h^r)$ and converge to $\Theta_i^\star(\Theta_h^r)$. Substitution into the slow update yields a reduced ODE for Θ_h, whose convergence follows from standard two-timescale stochastic approximation; the detailed proof can be found in Borkar [43] and Leslie [44].

Remark: In practice, F-HRL uses fixed step sizes, rather than diminishing schedules. Then, the iterates converge to an ϵ-neighborhood of the equilibrium, with radius scaling as $\epsilon = \mathcal{O}(\alpha_{\text{loc}} + \beta_r^2/\alpha_{\text{loc}})$; equivalently, $\limsup_{r\to\infty} \mathbb{E}[\|\Theta_h^r - \Theta_h^\star\|^2] \le C_1\alpha_{\text{loc}} + C_2\beta_r^2/\alpha_{\text{loc}}$ for some constants $C_1, C_2 > 0$. This means that smaller local steps and stronger timescale separation yield tighter neighborhoods.

Algorithm 1 F-HRL Global Round

Input: Base model Θ_{base}, trust parameters $(\rho, \sigma_{\text{trust}})$, aggregation period τ
Output: Updated policies $\{\Theta_i\}$ and Θ_h

```
for r ← 1 to R do
    foreach agent i do
        Θ_i^(r) ← Θ_h^(r) − α_loc ∇_Θ L_i(Θ_h^(r))
        ΔΘ_i^(r) ← Θ_i^(r) − Θ_h^(r)
    end
    Δ̃^(r) ← median({ΔΘ_i^(r)}_{i∈R_h})
    φ_i^(r+1) ←
        ρ φ_i^(r) + (1 − ρ) exp(−‖ΔΘ_i^(r) − Δ̃^(r)‖² / 2σ²_trust)
    Θ_h^(r+1) ← Θ_h^(r) + Σ_i φ_i^(r+1) ΔΘ_i^(r) / Σ_j φ_j^(r+1)
    Broadcast Θ_h^(r+1) to all agents
end
```

C. FEDERATED AGGREGATION AND ADAPTATION

Every τ steps, HAPS h aggregates updates from $\mathcal{R}_h$ using trust weights (Algorithm 1). Each agent's delta $\Delta\Theta_i^{(r)} = \Theta_i^{(r)} - \Theta_h^{(r)}$ is weighted by

$$\phi_i^{(r+1)} = \rho\phi_i^{(r)} + (1-\rho)\exp\left(-\frac{\|\Delta\Theta_i^{(r)} - \tilde{\Delta}^{(r)}\|^2}{2\sigma_{\text{trust}}^2}\right), \tag{15}$$

TABLE 4. Simulation and RL Hyperparameter Settings.

Parameter	Value
Network topology	
HAPS altitude	20 km
Terrestrial gNBs / UAV relays	3 / 3 standby
PUs / SUs	12 / 12
Spectrum & Physical layer	
Active channels	6 (20–100 MHz each)
Channel / gNB capacity	50–113 / 106–273 RBs
PU / SU transmit power	2 W / 0.5–1.0 W
Noise PSD (N_0)	−174 dBm/Hz
RL hyperparameters	
SU DQN: α / γ / ε-decay	0.003 / 0.99 / 0.995
SU DQN: hidden / batch / buffer	64 / 32 / 200
gNB/HAPS DQN: α / buffer	0.001 / 10000
Learning & Federation	
Horizon	5000 steps
FL aggregation (τ) / Trust ρ	50 steps / 0.9
Disaster window	Steps 2500–3500

where $\tilde{\Delta}^{(r)}$ is the coordinate-wise median, ρ is trust momentum, and σ^2_{trust} penalizes divergence from consensus.

To enable rapid redeployment without full retraining, HAPS construct context vectors over the preceding τ steps as follows:

$$\mathbf{c}_i = \left[\bar{\mathbf{o}}_i, \bar{r}_i\right], \tag{16}$$

where $\bar{\mathbf{o}}_i$ and $\bar{r}_i$ are the temporal means of agent i's local observations and rewards, respectively. HAPS computes cosine similarities $\text{sim}(\mathbf{c}_i, \mathbf{c}_j)$. If the best match exceeds threshold η, the stored model $\Theta_{j^\star}$ is transferred; otherwise, the base model is adapted via

$$\Theta_{\text{adapted}} = \Theta_{\text{base}} - \alpha_{\text{adapt}} \nabla_{\Theta} \mathcal{L}_i(\Theta_{\text{base}}), \tag{17}$$

$$\alpha_{\text{adapt}} = \alpha_0 \max_j \text{sim}(\mathbf{c}_i, \mathbf{c}_j), \tag{18}$$

where $\mathcal{L}_i$ is the local RL training loss for agent i.

D. COMPLEXITY ANALYSIS

Let $N_h = |\mathcal{R}_h|$ denote the number of agents in region h, and let $|\mathcal{A}_i|$ and $|\Theta_i|$ denote the action-space cardinality and parameter dimension of agent $i \in \mathcal{R}_h$. A flat controller requires searching $\prod_i |\mathcal{A}_i|$, which grows exponentially with N_h. In contrast, the hierarchical decomposition reduces this complexity to $\sum_i |\mathcal{A}_i|$, scaling linearly with the number of agents.

Per-step inference: Each agent performs a forward pass with complexity $\mathcal{O}(|\Theta_i|)$, yielding an aggregate regional cost of $\mathcal{O}\big(\sum_{i\in\mathcal{R}_h} |\Theta_i|\big)$ across the HAPS, $|\mathcal{B}_h|$ gNBs, $|\mathcal{D}_h|$ UAVs, and SUs. Under CTDE training, each gNB additionally evaluates a centralized critic of size $|\Theta_{\text{C}}|$, incurring an extra $\mathcal{O}(|\mathcal{B}_h|\,|\Theta_{\text{C}}|)$.

Federated round (Algorithm 1): Every τ steps, each agent computes a gradient update on $\mathcal{L}_i(\Theta_i)$ using a mini-batch of size M, with cost $\mathcal{O}(M|\Theta_i|)$. The HAPS aggregator computes the coordinate-wise median $\tilde{\Delta}^{(r)}$ and the trust-weighted aggregation, each at $\mathcal{O}(N_h|\Theta|)$, and evaluates pairwise context similarities $\text{sim}(\mathbf{c}_i, \mathbf{c}_j)$ at $\mathcal{O}(N_h^2 d_{\text{ctx}})$. Letting $|\Theta| = \max_i |\Theta_i|$, the per-round complexity is

$$\mathcal{O}\big(N_h M|\Theta| + N_h^2 d_{\text{ctx}}\big). \tag{19}$$

If a HAPS manages hundreds of agents, this can become expensive. In such cases, we propose using locality-sensitive hashing to reduce similarity search to $\mathcal{O}(N_h \log N_h d_{\text{ctx}})$. Since $d_{\text{ctx}} \ll |\Theta|$ (typically $d_{\text{ctx}} \approx 10$ vs. $|\Theta| \approx 10^4$), the gradient computation dominates.

Communication: Each federated round exchanges $\mathcal{O}(N_h|\Theta|)$ parameters for uplink model deltas $\Delta\Theta_i^{(r)}$ and downlink broadcasts of $\Theta_h^{(r+1)}$. Inter-HAPS synchronization over LEO backhaul every T_{sync} rounds introduces an amortized overhead of $\mathcal{O}((H-1)S_{\text{sync}}/T_{\text{sync}})$, where H denotes the number of HAPS regions and S_{sync} the compressed summary size. For the deployment in Section V ($N_h = 18$, $|\Theta| = 10^4$), the uplink+downlink exchange is about $2N_h|\Theta|$ parameters, i.e., approximately 3.6×10^5 values (~1.4 MB with 32-bit floats) per round; with $\tau = 50$ this is ~0.12 s on a 100 Mbps backhaul. For a denser region ($N_h = 200$, $|\Theta| = 10^4$), the exchange is about 4×10^6 values (~16 MB), motivating gradient compression.

E. SECURITY AND TRUST MANAGEMENT

The proposed F-HRL framework achieves resilience against adversarial behavior through complementary mechanisms operating at the sensing and federated layers.

Sensing layer: Cooperative SUs mitigate spectrum-sensing data falsification (SDF) by sharing local occupancy decisions and applying majority voting prior to channel access [15]. This mechanism tolerates up to $\lfloor(|\mathcal{U}_{S,b}| - 1)/2\rfloor$ falsified reports per gNB, preventing any single malicious SU from dominating local decisions.

Federated layer: Trust-weighted aggregation (15) provides Byzantine resilience during model fusion. The coordinate-wise median $\tilde{\Delta}^{(r)}$ and Gaussian trust kernel in ϕ_i suppress updates that deviate from regional consensus. Furthermore, HAPS tracks each agent's spectrum-access behavior (request patterns, grant efficiency, and adaptivity) as an auxiliary diagnostic metric for anomaly monitoring and post-hoc analysis; this behavioral metric is not part of the active aggregation weight in Eq. (15). Empirically, malicious SUs converge to lower scores than benign SUs (Section V).

V. PERFORMANCE EVALUATION

We benchmarked the proposed F-HRL stack against a non-cognitive baseline, a simpler cognitive baseline, a standard 5G heuristic baseline, and the HDRL implementation of [3] across the following key NTN operating scenarios: nominal spectrum sharing, disaster recovery, and malicious SU

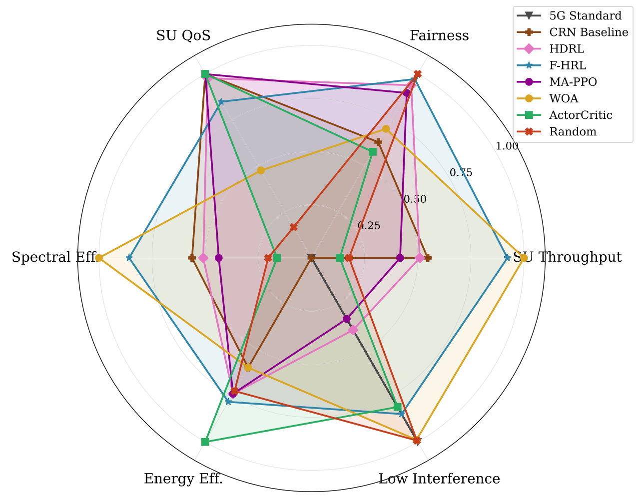

FIGURE 4. Multi-metric policy comparison.

TABLE 5. Key Performance Indicators Across Scenarios.

Scenario	Spec. Eff.	SU Tput (Mbps)	Col. (/step)	Fair. (Jain)
Non-cognitive	0.111	0.0	0.00	1.000
F-HRL (nominal)	0.871	167.5	5.13	0.950
F-HRL (attack)	0.748	128.9	8.46	0.887
F-HRL (disaster)	0.964	164.4	4.76	0.954

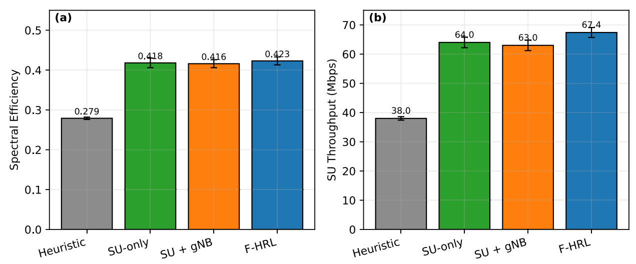

FIGURE 5. Hierarchy ablation study: (a) Spectral efficiency, (b) SU throughput, across four hierarchy configurations.

attacks. Unless otherwise stated, the results were averaged over 5000 decision steps with synchronized random seeds.

A. NETWORK CONFIGURATION

Unless stated otherwise, we simulated a single HAPS at 20 km coordinating three terrestrial gNBs and three standby UAVs. A total of 12 PUs and 12 SUs were distributed across urban and remote regions, contending over six channels (20–100 MHz bandwidth) allocated across the gNBs. Propagation, channel models, and transmit power levels followed Section III. Table 4 summarized the common configuration used in all experiments.

SU agents employed a range of RL algorithms with state vectors capturing sensed PU occupancy, and RB allocation trends among collaborative SUs. gNB and UAV policies enforced PU priority and protected against malicious SUs. HAPS performed federated aggregation with robust weighting every 50 steps and redistributed channels among active UAVs and gNBs as needed.

B. ATTACK MODEL

Malicious SUs falsified sensing with probability 0.8 and always required maximum RBs at full transmit power. 25% of SUs were modeled as malicious agents for the trust evaluation experiments.

Fig. 4 provided a comprehensive multi-metric comparison across all evaluated policies. Six normalized KPIs were displayed: SU throughput, Jain fairness index, QoS satisfaction ratio, spectral efficiency, energy efficiency (i.e., throughput per unit transmit energy), and interference footprint (defined as the inverse of collision rate weighted by transmit power). As can be seen in the results, F-HRL attained the highest throughput, the highest fairness score (0.950), and the lowest interference among learning-based approaches. HDRL achieved 97.9% QoS satisfaction, but lower throughput at 94.7 Mbps, while the CRN baseline exhibited the highest interference due to its fixed full-power transmissions. The 5G Standard policy, which excluded SU access entirely, achieved trivially zero interference, but forfeited all secondary throughput. These results confirmed that federated coordination enabled SUs to more effectively balance throughput, fairness, and interference than single-tier or non-cognitive alternatives. Importantly, the F-HRL architecture is algorithm-agnostic: the hierarchical coordination and federated aggregation layers are independent of the per-tier RL algorithm, so DQN, PPO, actor–critic, or other methods can be substituted at any tier without altering the overall framework.

C. KPIS AND KVIS

Table 5 showed the mean indicators across the 5000-step evaluation window. Spectrum efficiency reported aggregate RB utilization as a fraction of available capacity. Disaster and attack scenarios additionally logged retention and recovery dynamics, referenced herein in dedicated subsections.

Fig. 5 presented an ablation study comparing four hierarchy configurations: (i) a heuristic baseline with fixed small RB requests (5 RBs), random channel selection, and no learning at any tier, (ii) SU-only DQN where only SUs learned while gNB and HAPS used fixed policies, (iii) SU+gNB DQN adding gNB-level learning, and (iv) F-HRL with learning at all tiers. The heuristic baseline remained the lowest (0.279 spectral efficiency, 38.0 Mbps SU throughput). Relative to SU-only learning (0.418 spectral efficiency, 64 Mbps), adding the gNB tier (SU+gNB: 0.416, 63 Mbps) showed no statistically significant degradation across runs. Furthermore, adding the HAPS tier (full F-HRL: 0.423, 67.4 Mbps) preserved efficiency/throughput and slightly improved both metrics, indicating that hierarchy scales coordination without sacrificing core performance. Fig. 5 includes error bars across 10 random seeds to show

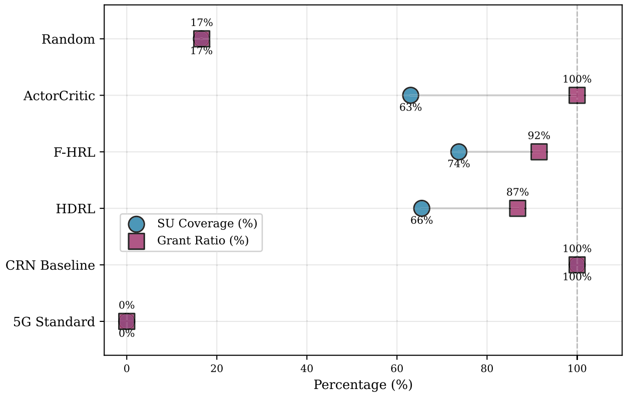

FIGURE 6. SU coverage and grant fulfillment.

TABLE 6. Policy Comparison Across Key Metrics.

Policy	SU Throughput (Mbps)	SU QoS (%)	Delay (steps)
5G Standard	0.0	0.0	0.0000
CRN Baseline	97.1	100.0	0.0000
HDRL	94.7	97.9	0.0000
F-HRL	167.5	86.2	0.6000
MA-PPO	73.5	99.9	0.0000
WOA	177.5	47.5	0.4000
ActorCritic	29.9	100.0	0.0000
Random	31.1	17.0	0.0000

the run-to-run variance. All configurations maintained perfect PU protection (1.0). Multi-seed runs across 10 random seeds exhibited negligible variance.

1) SU COVERAGE AND GRANT FULFILLMENT

Fig. 6 further examined SU access performance through two complementary metrics: coverage (i.e., the percentage of SUs receiving grants per step), and fulfillment (i.e., the ratio of granted to requested RBs). F-HRL achieved 74% coverage and 92% fulfillment, outperforming HDRL at 66% and 87%, respectively, by learning adaptive request strategies. The CRN baseline attained perfect scores on both metrics because its fixed, conservative requests were always satisfiable; however, this came at the cost of reduced throughput (see Fig. 4). Random access yielded only 17% on both metrics, confirming that uncoordinated spectrum contention severely limited SU performance.

The convergence curve showed that learning remained stable across all tiers, with variance across Monte Carlo running below 0.01. This stability was a direct result of federated aggregation and trust-weighted updates, which dampened the effect of outlier agents. Importantly, the convergence rate was faster than in single-tier RL baselines reported in previous research [3], [34], thus validating the benefit of hierarchical learning. However, the persistent contention post-convergence suggested that further improvements in cross-tier reward shaping or adaptive aggregation frequency may be needed.

2) RESILIENCE TO INFRASTRUCTURE OUTAGES

Fig. 7 showed disaster recovery performance when gNB0 failed at step 2 500. Fig. 7(a) depicted UE reassociation dynamics: prior to failure, gNB0 served the majority of UEs; upon failure, PUs and SUs tried to connect to the next available node and service degraded. HAPS recognized gNB0 failure and activated a UAV, while UEs continued trying to get service from next available nodes while the UAV was in transit. gNB1 experienced a high load and kept HAPS informed of the UE demand. HAPS detected the overload and activated another UAV to meet the demand and redistributed the available channels. Once gNB0 was restored, both UAVs returned to standby. Fig. 7(b) decomposed throughput into PU and SU contributions, demonstrating that PU traffic was prioritized throughout the recovery phase. During the disaster window, PU throughput retention reached 99.1% while SU retention amounted to 95.1%, confirming that the scheduling policy enforced PU protection even under constrained relay capacity. The results validated the system's resilience: HAPS-coordinated UAV deployment restored PU service within 50 steps, matching best-case recovery times in recent literature.

3) ADVERSARIAL ROBUSTNESS AND TRUST MANAGEMENT

Three of the twelve SUs, i.e., 25%, were configured as greedy attackers that falsified spectrum sensing reports and required maximum RBs at full transmit power. Fig. 8 presents the resulting score distribution after 100 episodes of HAPS-coordinated federated aggregation. Benign SUs converged to a median score of 0.75 (IQR 0.70–0.85), while malicious SUs were suppressed to a median of 0.34 (IQR 0.28–0.39), yielding a separation of 0.40. The upper whisker of malicious scores reached 0.44, which is below the minimum benign score of 0.455, giving the optimal threshold 0.45, with 100% detection accuracy (and 0 false positives and 0 false negatives). At a lower threshold $\tau = 0.30$, one malicious agent remained undetected, allowing that attacker to influence aggregation, yielding 91.7% accuracy. At a higher threshold $\tau = 0.50$, one benign agent was misclassified. These results confirm that the F-HRL trust-weighted aggregation mechanism reliably distinguishes adversarial from benign agents.

4) POLICY BENCHMARKING AND SU FAIRNESS

To contextualize the hierarchical ablation in Fig. 5, we compared F-HRL against a broader range of RL algorithms and baseline policies. Table 6 and Fig. 9 provide comprehensive policy benchmarking. As can be seen in the results, F-HRL achieved the highest SU throughput among learning-based methods (167.5 Mbps) and outperformed other RL alternatives, while HDRL attained 94.7 Mbps and the CRN Baseline demonstrated conservative yet reliable performance at 97.1 Mbps. Two additional baselines were evaluated:

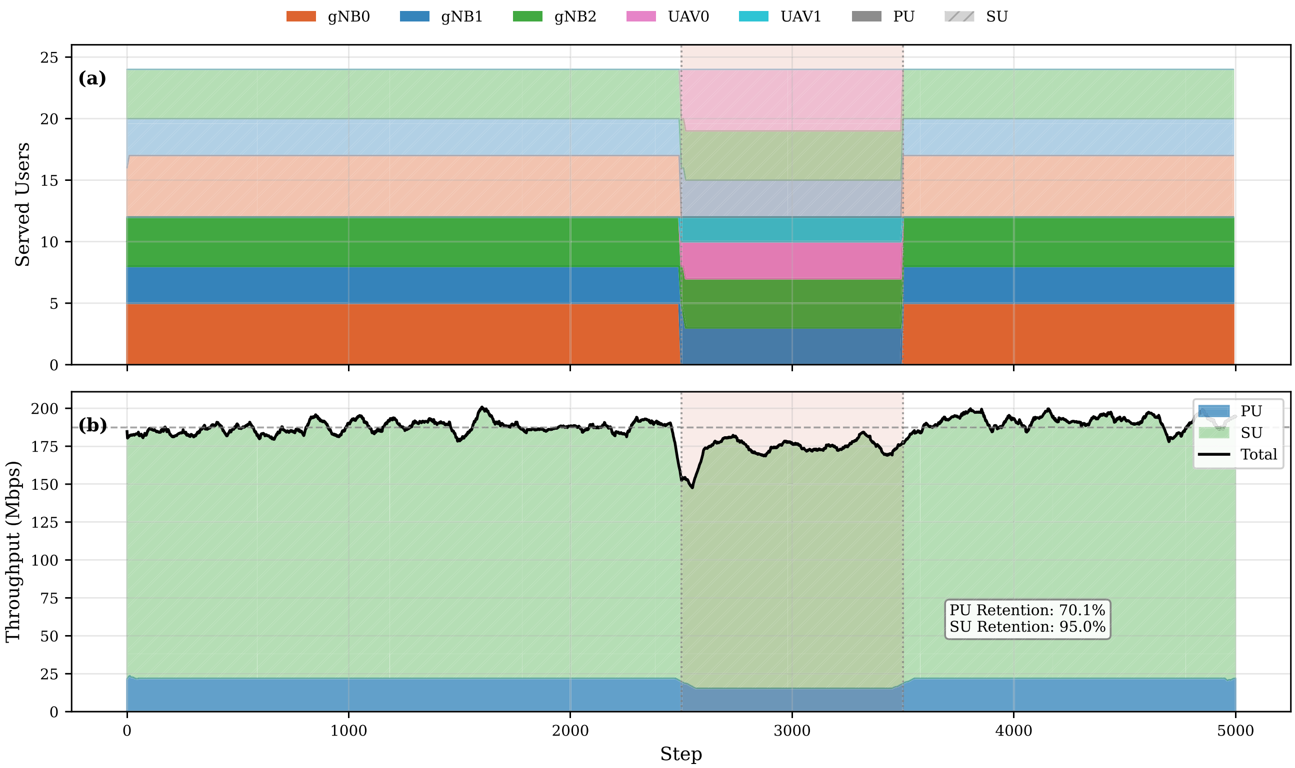

FIGURE 7. Disaster recovery: (a) UE reassociation dynamics during gNB0 outage, (b) PU and SU throughput retention.

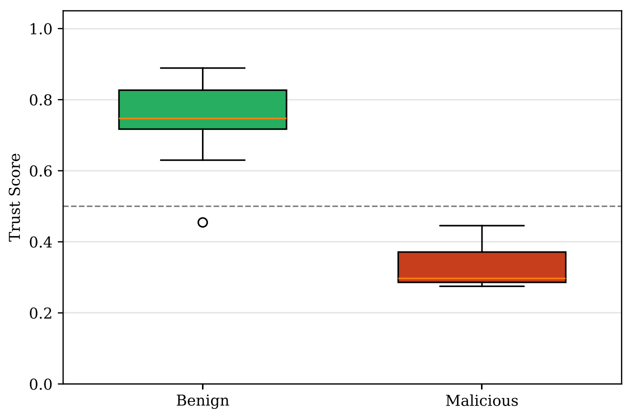

FIGURE 8. Trust score distribution.

MA-PPO achieved 73.5 Mbps, while WOA achieved the highest SU throughput at 177.5 Mbps, but with a lower SU QoS (47.5%) and a higher delay (0.4000), indicating a poorer QoS compliance despite centralized optimization. Delay reports average SU request-to-grant backlog-fulfillment latency (in scheduler steps); zero values indicate no delayed backlog fulfillment events in the logged runs.

D. DISCUSSION

Table 7 provided a per-region breakdown of F-HRL performance. Urban-1, which was found to have the highest capacity of 273 RBs and served five SUs, achieved 100% PU QoS and 91.0% SU QoS. Urban-2 attained the best SU QoS at 95.7%, which can be attributed to its lower contention ratio of 3 SUs sharing 200 RBs. The remote region, despite its limited 106 RB budget and four SUs, still reached 89.5% SU QoS, thus confirming that F-HRL adapted resource allocation to heterogeneous regional capacities.

TABLE 7. Regional Performance Breakdown for F-HRL.

Region	gNB	RBs	SUs	PU QoS (%)	SU QoS (%)
Urban-1	gNB0	273	5	100.0	91.0
Urban-2	gNB1	200	3	96.9	95.7
Remote	gNB2	106	4	100.0	89.5

We verified the convergence neighborhood prediction from Proposition 1 by running the F-HRL framework under three learning rate configurations ($\alpha \in \{0.003, 0.001, 0.0003\}$), as shown in Fig. 10. As α decreased from 0.003 to 0.0003, the steady-state parameter drift $\|\Theta_h^r - \Theta_h^{r-1}\|_2$ decreased from 0.125 to 0.116, confirming that in line with our expectation, smaller step sizes produce tighter convergence neighborhoods.

Overall, F-HRL maximized spectrum utilization and per-SU throughput, and traded off energy efficiency and QoS fulfillment according to regional priorities. Disaster recovery rapidly restored PU service and efficiently allowed SU access,

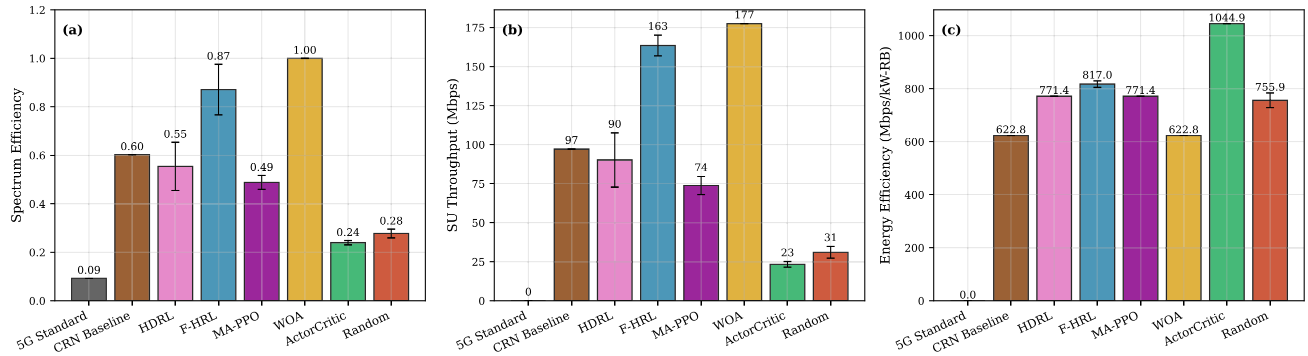

FIGURE 9. Policy comparison: (a) Spectral efficiency, (b) SU throughput, (c) Energy efficiency across all evaluated policies.

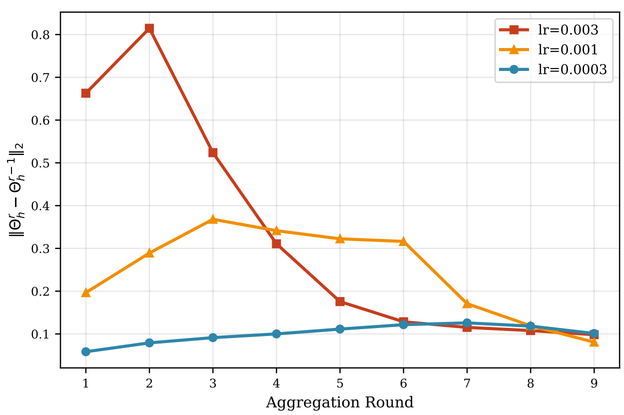

FIGURE 10. Convergence neighborhood: Parameter drift per aggregation round for three learning rate configurations.

while maintaining PU protection and priority. These insights demonstrated that F-HRL is highly adaptable to heterogeneous environments and provided a solid foundation for 6G NTN CRNs.

VI. CONCLUSION

In this paper, we presented a federated hierarchical reinforcement learning framework for HAPS-assisted CRNs operating across urban, remote, and disaster-stricken environments. The proposed architecture couples lightweight SU-level cooperative learning, trust-aware gNB/UAV coordination, and Byzantine-resilient aggregation at the HAPS tier, thereby enabling spectrum access decisions without centralizing raw observations. Simulation results demonstrated that the proposed framework consistently improved QoS across heterogeneous environments while maintaining PU protection and rapid post-failure recovery.

Major findings are briefly summarized below:

- **High spectrum utilization with hierarchical learning:** F-HRL achieved 87.1% spectrum efficiency as compared to 11.1% for the non-cognitive baseline, demonstrating a nearly 8-fold improvement while maintaining sustained per-SU throughput and high fairness.
- **Robust disaster recovery:** When a gNB failed, UAV relays quickly restored PU service with 99.1% retention and maintained 95.1% of original SU throughput, demonstrating effective resilience despite constrained relay capacity.
- **Adversarial robustness:** Trust-aware aggregation successfully isolated compromised agents and maintained legitimate throughput near nominal levels during attacks with 25% malicious SUs.
- **Policy comparison validates federated approach:** F-HRL achieved 74% SU coverage with 92% grant fulfillment, thus outperforming HDRL and maintaining the highest fairness score among learning-based approaches.

A brief agenda for future work is outlined below. First, context-aware selective model transfer should be explored to analyze the impact of regional transferability. Second, broader adversarial test suites such as coordinated power jamming should be incorporated to stress-test the trust pipeline. Third, extending the framework to account for detailed channel dynamics, including Doppler effects from UAV mobility and handover signaling overhead, would improve fidelity of the physical layer model. Fourth, investigating gradient compression techniques and hierarchical aggregation trees for deployments exceeding 1000 agents would address scalability requirements of dense satellite constellations.

Despite these open directions, the present results demonstrate that the proposed HAPS-centric F-HRL pipeline provides a practical template for resilient, privacy-aware spectrum coordination in emerging non-terrestrial networks.

REFERENCES

[1] F. Lotfi, O. Semiari, and W. Saad, "Semantic-aware collaborative deep reinforcement learning over wireless cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2022, pp. 5256–5261.

[2] P. Tehrani, F. Restuccia, and M. Levorato, "Federated deep reinforcement learning for the distributed control of NextG wireless networks," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Dec. 2021, pp. 248–253.

[3] M. Umer, M. A. Mohsin, A. A. Nasir, H. Abou-Zeid, and S. A. Hassan, "Intelligent spectrum sharing in integrated TN-NTNs: A hierarchical deep reinforcement learning approach," *IEEE Wireless Commun.*, vol. 32, no. 3, pp. 64–71, Jun. 2025.

[4] S. K. Das, R. Mudi, M. S. Rahman, K. M. Rabie, and X. Li, "Federated reinforcement learning for wireless networks: Fundamentals, challenges and future research trends," *IEEE Open J. Veh. Technol.*, vol. 5, pp. 1400–1440, 2024.

[5] A. Anderson et al., “D5.1 Early results on KPIs/KVIs, testing methodologies and benchmarking,” Zenodo, Dec. 2023, doi: 10.5281/zenodo.10473249.
[6] T. M. Getu, W. Ajib, R. Landry, and G. Kaddoum, “A tensor-based spectrum sensing technique for MIMO cognitive radio networks,” in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2019, pp. 1–5.
[7] T. M. Getu, W. Ajib, R. Landry, and G. Kaddoum, “Toward overcoming a hidden terminal problem arising in MIMO cognitive radio networks: A tensor-based spectrum sensing algorithm,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9833–9847, Oct. 2019.
[8] T. X. Quach, H. Tran, E. Uhlemann, G. Kaddoum, and Q. A. Tran, “Power allocation policy and performance analysis of secure and reliable communication in cognitive radio networks,” *Wireless Netw.*, vol. 25, no. 4, pp. 1477–1489, May 2019.
[9] O. Abbasi, G. Kaddoum, and H. Yanikomeroglu, “An ML-assisted OFDM-based hemispherical array antenna with hybrid beamforming for HAPS,” *IEEE Trans. Wireless Commun.*, vol. 24, no. 10, pp. 8908–8922, Oct. 2025.
[10] M. A. Qureshi, E. Lagunas, and G. Kaddoum, “Reinforcement learning for link adaptation and channel selection in LEO satellite cognitive communications,” *IEEE Commun. Lett.*, vol. 27, no. 3, pp. 951–955, Mar. 2023.
[11] S. Khaf, G. Kaddoum, and M. Altamimi, “HAPs-assisted cognitive radio with UE capability-centered cooperation,” in *Proc. GLOBECOM - IEEE Global Commun. Conf.*, Dec. 2024, pp. 1317–1322.
[12] W. Ning, X. Huang, K. Yang, F. Wu, and S. Leng, “Reinforcement learning enabled cooperative spectrum sensing in cognitive radio networks,” *J. Commun. Netw.*, vol. 22, no. 1, pp. 12–22, Feb. 2020.
[13] X. Silvirianti and G. Kaddoum, “Quantum adaptive learning for coverage optimization in LEO satellite network,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montréal, QC, Canada, Jun. 2025, pp. 2382–2387, doi: 10.1109/ICC52391.2025.11161401.
[14] I. Ngomane, M. Velempini, and S. V. Dlamini, “The detection of the spectrum sensing data falsification attack in cognitive radio ad hoc networks,” in *Proc. Conf. Inf. Commun. Technol. Soc. (ICTAS)*, Mar. 2018, pp. 1–5.
[15] S. Khaf, M. T. Alkhodary, and G. Kaddoum, “Partially cooperative scalable spectrum sensing in cognitive radio networks under SDF attacks,” *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8901–8912, Jun. 2022.
[16] M. Ahmadzadeh, S. Pakravan, G. A. Hodtani, M. Zeng, and J.-Y. Chouinard, “Deep reinforcement learning for robust RIS-aided over-the-air federated learning in cognitive radio,” in *Proc. IEEE Middle East Conf. Commun. Netw. (MECOM)*, Nov. 2024, pp. 368–373.
[17] F. Li, B. Shen, J. Guo, K.-Y. Lam, G. Wei, and L. Wang, “Dynamic spectrum access for Internet-of-Things based on federated deep reinforcement learning,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7952–7956, Jul. 2022.
[18] A. M. Seid et al., “A hierarchical MAFDRL-based resource allocation and incentive mechanism for TN-NTN in 6G networks,” *IEEE Trans. Mobile Comput.*, vol. 25, no. 2, pp. 2426–2443, Feb. 2026.
[19] D. Naseh, A. Bozorgchenani, S. S. Shinde, and D. Tarchi, “Unified distributed machine learning for 6G intelligent transportation systems: A hierarchical approach for terrestrial and non-terrestrial networks,” *Network*, vol. 5, no. 3, p. 41, Sep. 2025.
[20] S. S. Hassan et al., “Enhancing spectrum efficiency in 6G satellite networks: A GAIL-powered policy learning via asynchronous federated inverse reinforcement learning,” 2024, *arXiv:2409.18718*.
[21] H. Khoshkbari, G. Kaddoum, and M. Altamimi, “A generalized deep hierarchical reinforcement learning for HAPS resource scheduler,” in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, May 2025, pp. 91–96.
[22] T. Naous, M. Itani, M. Awad, and S. Sharafeddine, “Reinforcement learning in the sky: A survey on enabling intelligence in NTN-based communications,” *IEEE Access*, vol. 11, pp. 19941–19968, 2023.
[23] A. Farajzadeh, A. Yadav, and H. Yanikomeroglu, “Federated learning in NTN: Design, architecture, and challenges,” *IEEE Commun. Mag.*, vol. 63, no. 6, pp. 26–33, Jun. 2025.
[24] M. C. Hlophe, B. T. Maharaj, and M. M. Sande, “Energy-efficient transmissions in federated learning-assisted cognitive radio networks,” in *Proc. IEEE 21st Int. Conf. Commun. Technol. (ICCT)*, Oct. 2021, pp. 216–222.
[25] T. Safdar Malik et al., “RL-IoT: Reinforcement learning-based routing approach for cognitive radio-enabled IoT communications,” *IEEE Internet Things J.*, vol. 10, no. 2, pp. 1836–1847, Jan. 2023.
[26] S. Pala, K. Singh, C.-P. Li, and O. A. Dobre, “Empowering ISAC systems with federated learning: A focus on satellite and RIS-enhanced terrestrial integrated networks,” *IEEE Trans. Wireless Commun.*, vol. 24, no. 1, pp. 810–824, Jan. 2025.
[27] Y. Xu, Y. Yang, Q. Liu, and Z. Li, “Joint energy-efficient resource allocation and transmission duration for cognitive HetNets under imperfect CSI,” *Signal Process.*, vol. 167, Feb. 2020, Art. no. 107309.
[28] S. K. Sharma, T. E. Bogale, S. Chatzinotas, B. Ottersten, L. B. Le, and X. Wang, “Cognitive radio techniques under practical imperfections: A survey,” *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1858–1884, 4th Quart., 2015.
[29] E. K. Hamza, L. S. Mahdy, and A. A. Thabit, “Design and implementation of cognitive radio based on Xilinx FPGA,” *ARPN J. Eng. Appl. Sci.*, vol. 14, no. 4, pp. 892–897, 2019.
[30] K. Lee, “Low-complexity transmit power control for secure communications in wireless-powered cognitive radio networks,” *Sensors*, vol. 21, no. 23, p. 7837, Nov. 2021.
[31] T. Yang, W. Zhang, Y. Bo, J. Sun, and C.-X. Wang, “Dynamic spectrum sharing based on federated learning and multi-agent actor-critic reinforcement learning,” in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2023, pp. 947–952.
[32] F. Yang et al., “A federated reinforcement learning approach for optimizing wireless communication in UAV-enabled IoT network with dense deployments,” *IEEE Internet Things J.*, vol. 11, no. 20, pp. 33953–33966, Oct. 2024.
[33] Q. Zhou, S. Wu, C. Jiang, R. Zhang, and X. Jing, “Over-the-air federated transfer learning over UAV swarm for automatic modulation recognition in V2X radio monitoring,” *IEEE Trans. Veh. Technol.*, vol. 73, no. 3, pp. 3597–3607, Mar. 2024.
[34] F. Hu, Y. Deng, and A. H. Aghvami, “Cooperative multigroup broadcast 360° video delivery network: A hierarchical federated deep reinforcement learning approach,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 4009–4024, Jun. 2022.
[35] T.-D. Le and G. Kaddoum, “LSTM-based channel access scheme for vehicles in cognitive vehicular networks with multi-agent settings,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9132–9143, Sep. 2021.
[36] Y. Zhang, Q. Wu, and M. Shikh-Bahaei, “Vertical federated learning based privacy-preserving cooperative sensing in cognitive radio networks,” in *Proc. IEEE GLOBECOM Workshops*, Dec. 2020, pp. 1–6.
[37] H. Khoshkbari and G. Kaddoum, “Deep recurrent reinforcement learning for partially observable user association in a vertical heterogenous network,” *IEEE Commun. Lett.*, vol. 27, no. 12, pp. 3235–3239, Dec. 2023.
[38] H. Sharma, N. Kumar, and R. Tekchandani, “Mitigating jamming attack in 5G heterogeneous networks: A federated deep reinforcement learning approach,” *IEEE Trans. Veh. Technol.*, vol. 72, no. 2, pp. 2439–2452, Feb. 2023.
[39] S. Zuo et al., “Federated learning resilient to Byzantine attacks and data heterogeneity,” *IEEE Trans. Mobile Comput.*, vol. 24, no. 10, pp. 10729–10742, Oct. 2025.
[40] M. Krouka, A. Elgabli, C. B. Issaid, and M. Bennis, “Communication-efficient and federated multi-agent reinforcement learning,” *IEEE Trans. Cognit. Commun. Netw.*, vol. 8, no. 1, pp. 311–320, Mar. 2022.
[41] F. Alharbi, A. Alhuzali, E. Alalwany, and A. Alfahaid, “Dynamic orbital resource management (DORM) for 6G networks: Enhancing cybersecurity in non-terrestrial systems,” *IEEE Access*, vol. 13, pp. 137033–137047, 2025.
[42] Z. Chen, Y.-Q. Xu, H. Wang, and D. Guo, “Federated learning-based cooperative spectrum sensing in cognitive radio,” *IEEE Commun. Lett.*, vol. 26, no. 2, pp. 330–334, Feb. 2022.
[43] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[44] D. S. Leslie and E. J. Collins, “Individual Q-learning in normal form games,” *SIAM J. Control Optim.*, vol. 44, no. 2, pp. 495–514, Jan. 2005.

SADIA MAVERICK KHAF (Graduate Student Member, IEEE) received the B.E. degree in electrical engineering from the National University of Sciences and Technology, Pakistan, in 2015, and the M.S. degree in electrical and electronics engineering from Bilkent University, Türkiye, in 2018. They are currently pursuing the Ph.D. degree in electrical engineering with École de technologie supérieure, Montréal, Canada. From 2019 to 2020, they were a Lecturer and a Research Associate with the Faculty of Electrical Engineering at the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan. Their research interests include cognitive radio networks, non-terrestrial networks, spectrum sharing, radio resource management, reinforcement learning, federated learning, and resilient beyond-5G/6G wireless networks. They received the Fonds de recherche du Québec - Nature et technologies Doctoral Research Scholarship, the P.E.O. International Peace Scholarship, and the IEEE Canada Foundation Women in Engineering Prize. They currently serve as Vice Regional Student Representative for IEEE Canada, Eastern Region, Chair of the IEEE ÉTS Student Branch, Treasurer of IEEE Women in Engineering (WIE) Montréal, and Co-Founder of the IEEE WIE ÉTS.

GEORGES KADDOUM (Senior Member, IEEE) received the bachelor's degree in electrical engineering from the École Nationale Supérieure de Techniques Avancées, Brest, France, the M.S. degree in telecommunications and signal processing from the Université de Bretagne-Occidentale and Télécom Bretagne, Brest, in 2005, and the Ph.D. degree (Hons.) in signal processing and telecommunications from the National Institute of Applied Sciences, University of Toulouse, Toulouse, France, in 2009. He is currently a Professor and the Research Director of the Resilient Machine Learning Institute, holding the Tier 2 Canada Research Chair with the École de Technologie Supérieure (ÉTS), Canada. He has a prolific publication record with over 300 journal articles and conference papers, two chapters in books, and eight pending patents. His recent research interests include wireless communication networks, tactical communications, resource allocations, and network security. He has received several accolades, including the best paper awards at prestigious conferences, such as IWCMC 2023, PIMRC 2017, and WiMob 2014. Notably, he has been recognized with IEEE Transactions on Communications Exemplary Reviewer Award in 2019, 2017, and 2015, as well as the Research Excellence Award from the Université du Québec in 2018. In 2019 and 2025, he received the Research Excellence Award from ÉTS. His outstanding contributions were further acknowledged with the 2022 IEEE Technical Committee on Scalable Computing Award for Excellence (Middle Career Researcher), the 2023 MITACS Award for Exceptional Leadership, and the Gold Medal Award from Engineers Canada. He has served as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE COMMUNICATIONS LETTERS. He is an Area Editor of IEEE TRANSACTIONS ON MACHINE LEARNING IN COMMUNICATIONS AND NETWORKING and an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS.