

Article

Communication-Aware Decentralised Multi-Agent Reinforcement Learning Framework for UAV-Based Wildfire Suppression: Challenges Under Realistic Communication Constraints

Samuel Cartwright, Maxime Collignon, Adolfo Perrusquía *  and Antonios Tsourdos 

Faculty of Engineering and Applied Sciences, Cranfield University, College Road, Bedford MK43 0AL, UK; samuel.cartwright@cranfield.ac.uk (S.C.); maxime.collignon.250@cranfield.ac.uk (M.C.); antonios.tsourdos@cranfield.ac.uk (A.T.)

* Correspondence: adolfo.perrusquia-guzman@cranfield.ac.uk

Highlights

What are the main findings?

- Wind-driven wildfire dynamics and decentralised learning make UAV-based suppression significantly more challenging than in simplified centralised settings.
- Communication-constrained selective experience sharing improves learning, but containment remains limited due to sparse rewards, partial observability, and coordination failures.

What is the implication of the main finding?

- Realistic wildfire modelling is essential for evaluating UAV swarm learning methods, as simplified environments can overestimate coordination performance.
- Future decentralised MARL systems require better reward design, situational awareness, and communication mechanisms to achieve reliable wildfire containment.

Abstract

The increasing frequency and intensity of wildfires has created an urgent demand for scalable and autonomous wildfire response systems. While recent advances in multi-agent reinforcement learning (MARL) have demonstrated promise for collaborative uncrewed aerial vehicle (UAV)-based wildfire suppression, most existing approaches rely on simplified fire propagation dynamics and highly centralised learning architectures that are difficult to deploy in realistic operational settings. This paper presents a decentralised MARL framework for wildfire suppression that combines stochastic wildfire propagation, wind-driven spread dynamics, and communication-aware multi-agent coordination. The proposed framework extends an existing probabilistic wildfire environment through the incorporation of wind speed and directional effects, producing highly asymmetric and stochastic wildfire behaviour that more closely resembles real wildfire propagation. A decentralised Deep Q-Network (DQN) architecture is then introduced in which UAV agents learn independently through individual replay buffers. To mitigate the sparse-learning challenges introduced by decentralisation, selective experience sharing based on the SUPER algorithm is incorporated, enabling agents to exchange only high-value experiences under realistic communication constraints. Experimental results demonstrate that selective communication significantly improves containment performance and learning efficiency while preserving decentralised execution. The work highlights both the feasibility and challenges of realistic UAV swarm coordination for wildfire suppression, particularly the



Academic Editors: Brian Gullett, Pantelis Velanas and Katerina Margariti

Received: 22 June 2026

Revised: 25 July 2026

Accepted: 28 July 2026

Published: 29 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

trade-offs between communication bandwidth, environmental stochasticity, and collaborative performance.

Keywords: decentralised Deep Q-Network; wildfires; fire suppression; communication degradation; uncrewed aerial vehicles (UAVs)

1. Introduction

Wildfires are becoming increasingly destructive across large parts of the world. Climate change is driving a sustained rise in global temperatures [1–3], intensifying the frequency and severity of extreme weather events such as storms, heatwaves, and floods [4–8]. Among these events, wildfires have emerged as one of the most destructive and recurrent natural disasters in recent decades [9,10]. Recent wildfire seasons in regions such as California, Canada, Southern Europe, and Australia have caused substantial ecological damage, economic losses, and significant threats to human life [11–13]. Beyond their environmental impact, wildfire response operations require considerable financial resources and expose firefighters to highly hazardous operational conditions [14].

Despite significant advances in wildfire prediction, monitoring, and early detection, effective suppression remains largely dependent on human intervention. Conventional firefighting operations are constrained by deployment time, terrain accessibility, visibility conditions, and the inherent risks faced by emergency responders. As climate change continues to increase both the frequency and intensity of wildfire events, there is a growing need for autonomous technologies capable of supporting suppression operations while reducing risks to human personnel. Recent advances in robotics and artificial intelligence have positioned unmanned aerial vehicles (UAVs) as a promising technology for wildfire management due to their rapid deployment, high mobility, and ability to provide real-time situational awareness [15]. However, effective wildfire suppression and search-and-rescue (SAR) operations frequently require the coordinated deployment of multiple UAVs operating as a swarm. Such systems must address several open challenges, including collective perception, decentralized decision-making, communication robustness, and adaptive planning in highly dynamic environments [16–18].

Recent developments in reinforcement learning have significantly advanced autonomous multi-UAV coordination. In particular, Multi-Agent Reinforcement Learning (MARL) enables teams of UAVs to learn cooperative behaviours through repeated interaction with simulated environments, making it a promising framework for autonomous wildfire suppression [19–21]. Existing studies have demonstrated that MARL can successfully coordinate UAV teams for disaster response and wildfire management tasks, highlighting its potential to improve suppression efficiency while reducing the operational burden on human responders [22,23]. Nevertheless, several important challenges remain before such systems can be deployed in realistic wildfire scenarios.

First, many reinforcement learning environments employ simplified wildfire propagation models that fail to capture the strong influence of wind on fire spread [24,25]. As a result, wildfire behaviour is often represented by unrealistically symmetric propagation patterns that differ substantially from those observed in real incidents. Second, many MARL approaches adopt centralized learning architectures [26] based on shared policies, shared replay buffers, or globally available information. Although these architectures often achieve strong performance during simulation, they assume persistent communication infrastructure [27], introduce single points of failure, and scale poorly to large distributed UAV swarms operating in communication-constrained environments. Finally, relatively

little attention has been devoted to understanding how realistic communication constraints influence decentralised cooperative learning and wildfire suppression performance, despite communication being one of the principal challenges for operational multi-UAV systems.

To address these limitations, this paper proposes a communication-aware decentralised MARL framework for cooperative UAV wildfire suppression. The proposed approach extends a stochastic wildfire environment by incorporating wind-driven fire propagation while replacing centralized learning with independent UAV policies that selectively exchange high-priority experiences using the SUPER (Selective Sharing of Valuable Experiences) communication mechanism, which allows neighbouring UAVs to exchange only the most informative experiences rather than complete replay buffers. This enables systematic investigation of the trade-offs between communication constraints, decentralised coordination, and wildfire suppression performance under more realistic operational conditions.

This work addresses both of these limitations simultaneously through the following contributions

1. A stochastic wildfire propagation environment is extended through the incorporation of wind-driven spread dynamics, producing highly directional and asymmetric wildfire behaviours that more closely resemble real wildfire propagation.
2. A decentralised multi-agent reinforcement learning framework is developed for UAV-based wildfire suppression, replacing the centralised learning paradigm of previous work with independent Dueling Deep Q-Network (DQN) agents and individual replay buffers.
3. A communication-aware learning mechanism based on the SUPER selective experience sharing algorithm is introduced, enabling UAVs to exchange only high-value experiences under realistic communication constraints.

The remainder of the paper is structured as follows. Section 2 provides the wildfire environment modelling. Section 3 introduces the proposed methodology and both the wildfire propagation environment and wind model. Section 4 presents the decentralised reinforcement learning framework and communication architecture. Section 5 presents results and analysis. Finally, Section 6 concludes the paper and discusses future research directions.

2. Related Work

A wide range of technologies have been proposed to support responders during natural disasters [28]. In wildfire management, conventional suppression strategies rely heavily on aerial firefighting platforms such as Canadair aircraft, while more recent developments have explored autonomous robotic systems, UAVs, and emergency eVTOL vehicles capable of rapidly reaching affected areas [29,30]. Among these technologies, UAV swarms have emerged as a particularly promising solution due to their flexibility, scalability, and ability to operate in hazardous environments. However, deploying autonomous UAV swarms in operational wildfire scenarios remains challenging due to the dynamic nature of wildfire propagation, limited communication capabilities, and the need for coordinated decision-making among multiple autonomous agents. According to [16], four major research areas remain critical for operational deployment: perception and planning, communication, human-robot interaction, and cost-function trade-offs. Among these, perception, planning, and communication constitute the principal technical challenges for autonomous wildfire suppression, where rapidly evolving environmental conditions require robust decentralised coordination under partial observability.

Several approaches have been proposed to address multi-agent coordination in search-and-rescue (SAR) and disaster response applications. Monte Carlo Tree Search methods [31] have been applied to UAV fleets for SAR missions [32], while reinforcement learning and MARL techniques have demonstrated strong performance in cooperative search and

rescue tasks [33]. Examples include maritime SAR systems using reinforcement learning for trajectory optimization [34,35], adversarial MARL frameworks for victim discovery in obstacle-rich environments [36], heterogeneous-agent SAR architectures capable of communication-fault-tolerant operation [37], and decentralized exploration strategies based on distributed actor-critic methods with recurrent neural networks to handle imperfect communication and partial observability [38]. MARL has also been applied to emergency response coordination in urban disaster scenarios, demonstrating significant improvements in victim identification and responder deployment efficiency [39]. Collectively, these studies demonstrate that MARL provides an effective framework for coordinating distributed UAV teams operating under uncertainty and partial observability. Nevertheless, most existing approaches have been developed for generic SAR or disaster-response scenarios rather than highly dynamic wildfire environments, and typically assume either reliable communication or do not explicitly investigate the impact of communication constraints on cooperative performance.

In parallel with advances in autonomous coordination, wildfire spread modelling has evolved from deterministic empirical models towards stochastic and probabilistic approaches capable of capturing the inherent uncertainty of wildfire behaviour. While physics-based wildfire simulators provide highly detailed representations of combustion and atmospheric interactions, their computational complexity makes them impractical for reinforcement learning, which requires millions of simulation episodes during training. Consequently, stochastic cellular automata have become an attractive alternative by providing a practical compromise between computational efficiency and behavioural realism. Among the most influential environmental factors identified within the wildfire literature is wind, which strongly affects propagation direction and spread velocity through mechanisms such as ember transport and directional heat transfer. Recent studies have combined wildfire modelling with reinforcement learning to develop autonomous suppression and monitoring systems. For example, Q-learning combined with Value Decomposition Networks (VDN) [40] has been used to train UAV swarms capable of monitoring wildfire propagation, outperforming conventional monitoring approaches [41]. Similarly, Multi-Agent Deep Q-Networks (MADQN) have been applied to wildfire suppression tasks, enabling UAV fleets to coordinate retardant deployment and outperform heuristic-based suppression strategies [42]. These studies clearly demonstrate the potential of MARL for autonomous wildfire management. However, existing approaches generally rely on simplified wildfire propagation models, neglect realistic environmental factors such as dynamic wind conditions, or adopt centralized learning architectures with unrestricted information sharing. Consequently, the influence of communication constraints on decentralised wildfire suppression remains largely unexplored.

Among the existing literature, the wildfire propagation framework developed by [43] is the work most closely related to the present study, as it combines stochastic wildfire propagation with MARL-based UAV suppression. However, the original framework assumes isotropic wildfire propagation and a centralized learning architecture based on shared policies and replay buffers, preventing the investigation of communication-constrained decentralised coordination. The present work addresses these limitations by extending the wildfire environment to incorporate wind-driven propagation effects, producing more realistic directional fire behaviour, while replacing the centralized learning structure with a decentralized architecture in which individual UAVs learn independently and selectively exchange valuable experiences through the SUPER communication mechanism. This enables systematic investigation of wildfire suppression under realistic communication constraints while preserving the benefits of cooperative learning.

Overall, recent research has demonstrated significant progress in autonomous wildfire management through advances in wildfire modelling, multi-agent reinforcement learning, and cooperative UAV coordination. Nevertheless, three important challenges remain: achieving a suitable balance between wildfire model realism and computational efficiency, enabling scalable decentralised coordination under partial observability, and understanding how communication constraints influence cooperative suppression performance. These open challenges motivate the communication-aware decentralised MARL framework proposed in this paper.

3. Methodology

Figure 1 presents the overall architecture and operational workflow of the proposed communication-aware decentralised MARL framework for wildfire suppression. At the beginning of each episode, a stochastic wildfire evolves over a heterogeneous environment according to the proposed propagation model, which incorporates vegetation-dependent ignition, wind-driven spread, and thermal field generation. UAV agents are randomly deployed along the environment boundary and, at each decision step, independently acquire only locally observable information, including thermal measurements, fire-related guidance information, neighbouring UAV proximity indicators, environmental conditions, and positional context. Based on these observations, each UAV selects a movement action using its own Dueling DQN policy [44], interacts with the environment through wildfire suppression, receives a reward, and stores the resulting experience in its local replay buffer. Following the UAV actions, the wildfire state and thermal field are updated, producing the observations for the next decision step. During training, neighbouring UAVs within communication range improve learning efficiency by selectively exchanging only highly informative experiences identified through the SUPER algorithm, rather than complete replay buffers. During execution, however, no experience sharing takes place, and each UAV operates fully autonomously using only its local observations and independently learned policy. This design enables the investigation of the trade-offs between decentralised learning, communication constraints, and cooperative wildfire containment performance.

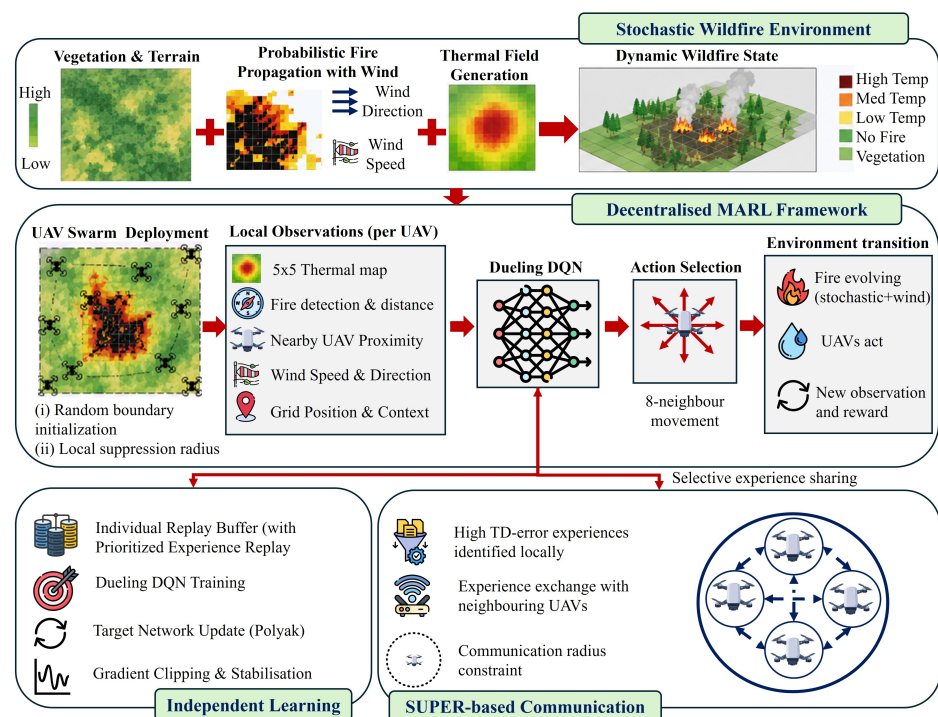


Figure 1. High-level architecture of the proposed methodology.

3.1. Wildfire Environment Modelling

Existing operational wildfire simulators such as the Rothermel [45] and McArthur models [46] have been widely adopted for wildfire prediction by incorporating detailed physical processes including fuel properties, terrain slope, and combustion dynamics. However, their usage for training dynamic agents is limited and prohibitive due to their high computation demands and the impossibility to modify the underlying Markov process. Therefore, in this paper we adopt a completely different environment that allows to model agent-environment interactions while preserving the key fire propagation characteristics necessary for evaluating decentralised coordination algorithms.

The proposed wildfire environment extends our previous framework [43]. While the vegetation-dependent stochastic propagation and thermal diffusion models are retained, the present work introduces wind-dependent directional fire propagation and time-varying wind conditions. Furthermore, the agent observation model is redesigned to support decentralised learning by removing globally available information, enlarging the local thermal observation window, and replacing the shared-policy architecture with independent policies that exchange only selected experiences during training. For completeness, the principal components of the wildfire model are briefly summarised in the following sections.

3.2. Vegetation Density Modelling

The forest environment is represented as a two-dimensional grid $G \in \mathbb{N}_N^2$ of size $N \times N$, where each cell corresponds to a local forest region. In order to capture the heterogeneous nature of real vegetation distributions, each cell (i, j) is assigned a vegetation density value $d_{ij} \in [0, 1]$. Rather than generating these densities independently, spatial correlation is introduced by applying a Gaussian smoothing filter with standard deviation σ_d to an initially uniform random field. This produces contiguous regions of high and low vegetation density, creating realistic fuel variations across the environment.

The vegetation density directly influences wildfire propagation behaviour. Regions with higher density values are more susceptible to ignition and support faster fire transmission, while sparse regions act as natural barriers that can slow or fragment propagation fronts. In this work, the smoothing parameter is set to $\sigma_d = 10$, which was found to produce realistic large-scale spatial heterogeneity while maintaining sufficient variability within the environment.

3.3. Stochastic Fire Propagation

Wildfire spread is modelled as a stochastic local interaction process operating over the four-neighbour topology of the grid. Each cell X exists in one of three possible states,

$$s_X^k \in \{\mathcal{H}, \mathcal{F}, \mathcal{D}\} = \{\text{Healthy}, \text{Burning}, \text{Burned}\}, \quad (1)$$

where k denotes the current simulation timestep. At each timestep, burning cells may ignite neighbouring healthy cells according to a probabilistic transition model governed by local vegetation density. The baseline ignition probability is defined as

$$\gamma_X = \alpha d_X, \quad (2)$$

where α is the global propagation parameter and d_X denotes the vegetation density of the neighbouring cell. Consequently, densely vegetated regions exhibit higher ignition likelihoods and support faster wildfire expansion.

Once ignited, a burning cell remains active until transitioning to the burned state with probability β . Burned cells become permanently inert and may no longer participate in the

propagation process. Together, these local interactions define a Markovian propagation process whose global wildfire behaviour emerges from the collective evolution of all cells within the grid.

In the original framework, parameter values of $\alpha = 0.25$ and $\beta = 0.1$ were shown to produce fragmented and irregular fire fronts representative of heterogeneous forest landscapes [43]. These values provide sufficient propagation aggressiveness while preventing unrealistically uniform spread patterns.

3.4. Thermal Field Generation

In addition to wildfire propagation, the environment maintains a temperature field designed to emulate thermal observations captured by UAV-mounted infrared sensors. This temperature map evolves dynamically alongside the wildfire states and provides the primary sensory input for the reinforcement learning agents.

At each timestep, burning cells are assigned a maximum temperature $T_{\max}$, while heat diffuses spatially across neighbouring cells through Gaussian smoothing. Simultaneously, previously heated regions cool gradually over time according to a temporal decay factor $c \in [0, 1]$. The resulting thermal dynamics are expressed as

$$T^{k+1} = c \left(T_{\max} \mathbb{1}_{\mathcal{F}}(s^{k+1}) + \mathbb{1}_{\mathcal{H} \cup \mathcal{D}}(s^{k+1}) \odot T^k \right) * w_{\sigma_T}, \quad (3)$$

where $\mathbb{1}_{\mathcal{F}}(s)$ denotes the binary mask of burning cells, $\odot$ represents the Hadamard product, and $*$ denotes two-dimensional convolution with Gaussian kernel w_{σ_T} . The parameters $T_{\max} = 100$, $\sigma_T = 1$, and $c = 0.965$ were selected to generate smooth but informative thermal gradients surrounding active fire fronts. Although simplified relative to full physical heat-transfer models, this approach reproduces realistic temperature anomalies while remaining computationally efficient for large-scale reinforcement learning training. The resulting temperature field provides agents with indirect information about nearby fire activity, allowing UAVs to infer propagation direction and fire intensity without direct access to global wildfire states.

3.5. Wind Modelling

To improve realism, wind is introduced as a primary propagation influence. Wind speed and direction directly alter the probability that a burning cell ignites neighbouring cells, introducing asymmetric spread patterns and significantly increasing environmental stochasticity. The ignition probability (2) is denoted as $\gamma_{X_{\text{initial}}}$, and it is modified according to wind influence

$$\gamma_{X_{\text{final}}} = \text{clip}(\gamma_{X_{\text{initial}}} (1 + c w_s \cos \theta), 0, 1) \quad (4)$$

where c controls the strength of the wind effect, w_s denotes wind speed, and θ represents the angular difference between wind direction and neighbour direction. The clipping operation guarantees that the ignition probability always remains within the valid interval $[0, 1]$, preserving its probabilistic interpretation under all wind conditions.

The angular difference is computed as

$$\theta = \min(|w_d - n_d|, 2\pi - |w_d - n_d|), \quad (5)$$

where w_d is the wind direction and n_d is the neighbour direction. This formulation allows wind to both enhance and suppress local propagation probabilities depending on directional alignment. As a result, fires develop highly directional spread patterns rather than approximately circular expansion. Wind speed values are sampled from a normal

distribution derived from empirical wildfire measurements, while wind direction is sampled uniformly between 0 and 2π . Additionally, wind conditions evolve slightly during each simulation episode through small stochastic perturbations applied at fixed intervals. Although simplified relative to full atmospheric modelling, this approach significantly improves propagation realism while maintaining computational tractability.

The resulting environment produces a wide variety of wildfire behaviours (see Figure 2). Fires may propagate narrowly along dominant wind directions, spread asymmetrically across one side of the environment, or rapidly reach environment boundaries through directional acceleration. These behaviours create substantially more difficult learning conditions for reinforcement learning agents, since valuable experiences become both rarer and less spatially predictable.

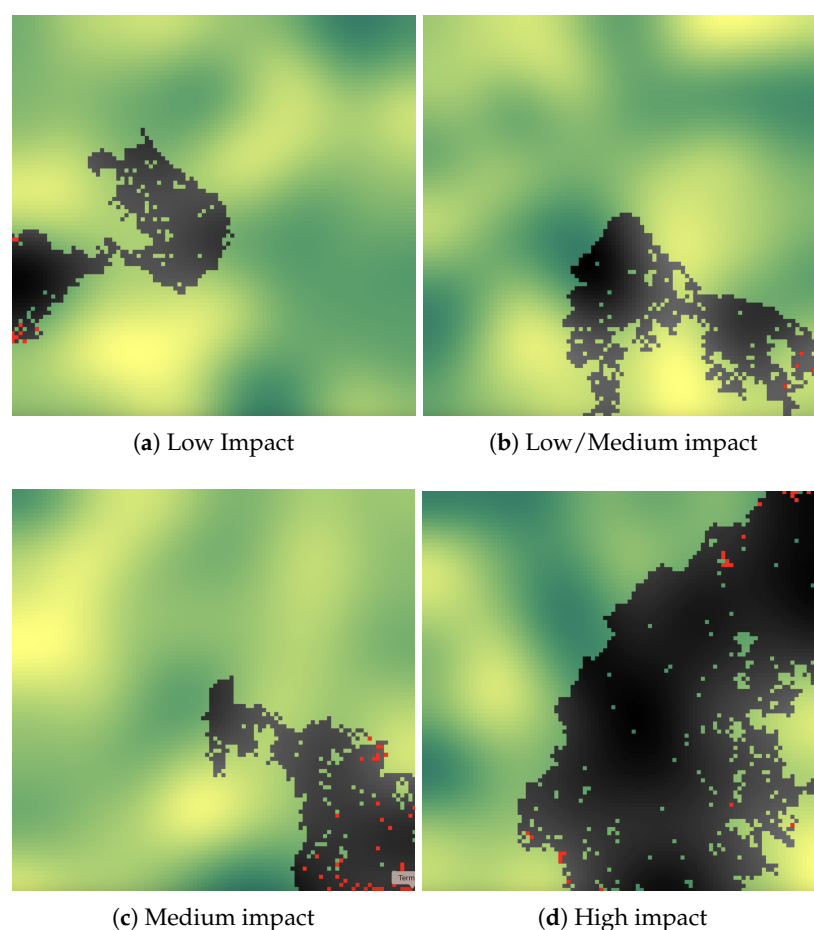


Figure 2. Representative wildfire propagation patterns generated under different wind conditions. Red cells denote actively burning vegetation, while black cells denote burned or extinguished regions. The four examples illustrate increasing wind influence on wildfire behaviour, ranging from low to high impact. Each subfigure corresponds to a different combination of wind speed and direction, producing progressively stronger directional propagation and increasingly asymmetric fire fronts.

The final hyperparameters used in the wildfire environment modelling are summarized in Table 1.

Table 1. Wildfire environment parameters used throughout the experiments.

Parameter	Value	Description
Grid size	100×100	Forest environment dimensions
σ_d	10	Vegetation density smoothing parameter
α	0.25	Fire propagation coefficient
β	0.10	Burnout probability
T_{max}	100	Maximum burning-cell temperature
σ_T	1	Thermal diffusion parameter
c	0.965	Thermal decay coefficient
Wind speed mean	5.19 m/s	Wind distribution mean
Wind speed std. dev.	2.0, m/s	Wind distribution standard deviation
Wind direction	$[0, 2\pi)$	Uniformly sampled direction

4. Decentralised MARL for Wildfire Suppression

The primary objective of this work is not simply to improve suppression performance within simulation, but to move toward a more realistic formulation of autonomous wildfire response using UAV swarms. While the original framework introduced in [43] demonstrated that reinforcement learning agents could learn collaborative suppression behaviours, the training paradigm relied heavily on centralised assumptions. Agents shared a common replay buffer, possessed access to globally available information, and effectively learned a single shared policy. Although this simplified training considerably, it introduced several limitations from the perspective of real-world deployment. In practice, UAV swarms operating over wildfire environments cannot assume persistent communication with a central controller, nor can they rely on unrestricted access to global state information. Wildfires often span extremely large regions, communication bandwidth is inherently constrained, and partial communication failures are likely during operation. A fully centralised architecture therefore introduces both scalability limitations and a critical single point of failure.

To address these issues, the present work reformulates the suppression problem as a decentralised multi-agent reinforcement learning task. Each UAV agent learns independently through its own replay buffer and policy network, while communication between agents is modelled explicitly through constrained selective experience sharing. Importantly, this transition toward decentralisation substantially increases the difficulty of the learning problem. The combination of independent learning, sparse rewards, and the newly introduced wind-driven wildfire dynamics creates an environment in which valuable experiences become rare and difficult to discover during early training. Consequently, the decentralised framework requires significant modifications not only to the communication structure, but also to the state representation, reward design, and performance evaluation methodology.

4.1. Environmental Complexity and Learning Challenges

The introduction of wind-driven wildfire propagation fundamentally alters the reinforcement learning problem faced by the UAV agents. In the previous vegetation-only propagation model, fires expanded relatively uniformly around ignition regions, making active fire fronts comparatively easy to locate. In contrast, the enhanced wildfire model produces highly asymmetric and stochastic propagation patterns. Depending on wind conditions, fires may propagate narrowly across one region of the environment, spread rapidly toward a boundary, or fragment into discontinuous fronts. This increased environmental stochasticity creates several major learning challenges. First, the probability that an agent randomly encounters active fire regions during early exploration decreases significantly. Valuable experiences therefore become sparse, particularly during the initial stages of

training when policies remain largely random. Second, the relationship between actions and long-term outcomes becomes more difficult to infer, since wildfire propagation is now strongly influenced by exogenous environmental dynamics. Finally, the increased variability of fire behaviour complicates performance evaluation itself. A low burn percentage, for example, does not necessarily indicate successful suppression, since weak or directionally constrained fires may naturally extinguish without meaningful agent intervention.

For this reason, the primary objective of the agents is reformulated from simply minimising burned area toward achieving wildfire containment. Containment is defined as preventing the wildfire from reaching the environment boundary, thereby simulating successful suppression before large-scale escape occurs. This distinction becomes particularly important in stochastic environments where propagation intensity varies significantly between episodes.

4.2. Action Space and Local Observations

The UAVs are intentionally modelled at the mission-planning level rather than at the flight-control level. Each movement action represents the successful execution of a high-level navigation command between neighbouring grid cells rather than the physical trajectory followed by the aircraft. Consequently, low-level flight dynamics, including velocity limits, turning-rate constraints, inertia, aerodynamic effects, and wind-induced disturbances acting on the UAV, are not explicitly modelled. This abstraction allows the work to isolate the effects of decentralised decision-making, communication constraints, and wildfire evolution without introducing additional complexity associated with vehicle dynamics. Although the wildfire environment becomes substantially more complex, the action space remains intentionally simple in order to isolate the effects of decentralised coordination and communication. At each timestep, every UAV agent may move to one of the eight neighbouring cells surrounding its current position. Agents are initially deployed randomly along the environment boundary to emulate realistic aerial deployment conditions.

Wildfire suppression itself occurs automatically whenever a burning cell lies within the UAV extinguishing radius. When a UAV suppresses a burning cell, the cell is immediately removed from the active fire front and therefore cannot ignite neighbouring cells during subsequent wildfire propagation. Because the temperature field is generated from the active burning cells, suppression also affects future thermal observations by removing the corresponding heat source. Consequently, UAV actions influence both the wildfire evolution and the thermal environment through their effect on the set of actively burning cells. Consequently, the problem remains fully discrete in both action selection and environmental transitions. While simplified, this formulation allows the work to focus specifically on coordination and learning dynamics without introducing additional complexity arising from continuous control. Nevertheless, the framework naturally supports future extensions toward hybrid action spaces. In particular, a more realistic suppression model could incorporate explicit control over retardant deployment, allowing agents to activate or deactivate extinguishing systems while operating under finite suppressant capacity constraints. Such an extension would introduce resource management considerations alongside spatial coordination which is topic of our future work.

A more substantial modification concerns the information available to each UAV agent. Since the goal of the work is to approximate realistic decentralised execution, agents are intentionally restricted to locally observable information and operationally plausible telemetry. Unlike the original centralised framework, UAVs no longer receive direct information concerning the precise positions of neighbouring agents or globally accessible fire states. Instead, each UAV receives a compact local observation vector consisting of several complementary information sources:

1. A 5×5 logarithmic temperature map : This is a small grid of temperature readings centred around the UAV position. This temperature field is derived directly from the wildfire thermal diffusion model and provides indirect information regarding nearby fire intensity and propagation direction. Compared to the original 3×3 representation used in earlier work, the larger observation window significantly improves the agent's ability to infer local fire structure and directional spread.
2. Fire information: In addition to thermal observations, each UAV receives limited fire guidance information consisting of the initial relative direction and distance toward the ignition region. This guidance signal is included primarily to stabilise early-stage learning by increasing the probability that agents encounter active fire fronts before meaningful exploration policies emerge.
3. Proximity Information: Local coordination awareness is incorporated through a lightweight proximity representation. Rather than providing exact neighbouring UAV positions, agents receive a bounded one-hot proximity indicator encoding whether other UAVs exist within a limited communication radius. This maintains partial observability while still allowing the agents to infer local swarm density and avoid excessive clustering.
4. Wind speed and direction: Environmental context is further provided through direct access to wind speed and wind direction values. Importantly, however, the agents are not given access to the internal propagation weighting parameters governing wildfire spread. As a result, UAVs must learn the relationship between wind conditions and wildfire evolution implicitly through interaction with the environment rather than through privileged environmental knowledge.
5. Grid location and temporal/fire information: Agents receive coarse positional and temporal information. Grid coordinates emulate GPS-like localisation capabilities, while temporal indicators such as current episode progress and global fire intensity provide contextual awareness regarding the current suppression scenario.

Collectively, these observations define a partially observable decision-making problem in which UAVs must infer large-scale wildfire behaviour using only local thermal information, limited coordination awareness, and environmental context. This formulation provides a more realistic approximation of wildfire behaviour than the previous isotropic propagation model by incorporating directional wind effects and time-varying wind conditions.

4.3. Reward Function

The reward function was redesigned to support decentralised learning in the more stochastic wildfire environment. In the original setting, fire propagation was comparatively regular, and agent performance could be encouraged primarily through local extinguishing rewards. After introducing wind-driven spread, however, this formulation becomes insufficient. Fires may propagate rapidly in narrow directions, fragment into disconnected fronts, or reach the boundary before agents have encountered meaningful thermal information. Therefore, the reward must provide immediate local learning signals while still encouraging the global objective of containment.

For each UAV agent i , the total reward at timestep k is defined as

$$r_i^k = r_{\text{ext},i}^k + r_{\text{temp},i}^k + r_{\text{sep},i}^k + r_{\text{guide},i}^k + r_{\text{cont},i}^k \quad (6)$$

where the terms respectively correspond to extinguishing performance, thermal-gradient following, inter-agent separation, early fire guidance, and containment-oriented shaping.

The principal reward is assigned when an agent extinguishes burning cells within its suppression radius. Let e_i^k denote the number of cells extinguished by agent i at timestep k , and let N_a be the number of UAVs. The extinguishing reward is given by

$$r_{\text{ext},i}^k = \lambda_{\text{ext}} \frac{e_i^k}{N_a} - \lambda_{\text{idle}} \mathbb{1}(e_i^k = 0), \quad (7)$$

where $\lambda_{\text{ext}} > 0$ rewards direct suppression and $\lambda_{\text{idle}} > 0$ discourages actions that do not contribute to extinguishing. This term provides the most direct local signal because it links the agent's action to an observable reduction in active fire cells.

However, relying only on extinguishing rewards leads to sparse learning, particularly at the start of training when UAVs rarely encounter the fire. A thermal shaping term is therefore introduced to encourage agents to move toward hotter regions of the environment. Let T_i^k be the temperature at the position of agent i at timestep k . The thermal reward is defined as

$$r_{\text{temp},i}^k = \lambda_{\text{temp}} \max\left(0, \frac{T_i^{k+1} - T_i^k}{T_{\text{max}}}\right), \quad (8)$$

with an upper bound applied in implementation to avoid excessive reward magnitudes. This term encourages movement toward the fire front without giving the agent direct access to the global fire state.

To prevent UAVs from collapsing onto the same local region, a separation penalty is included. If $\mathcal{N}_i^k$ denotes the set of neighbouring UAVs within a minimum safety radius r_{safe} , then

$$r_{\text{sep},i}^k = -\lambda_{\text{sep}} |\mathcal{N}_i^k|. \quad (9)$$

This term approximates a collision-avoidance requirement and also promotes wider spatial coverage of the fire front.

During the first part of each episode, an additional guidance term is used to reduce the probability of agents wandering for long periods without encountering the fire. Let d_i^k be the distance from UAV i to the initial ignition region. For the first K_g timesteps,

$$r_{\text{guide},i}^k = \begin{cases} \lambda_g^+, & d_i^{k+1} < d_i^k, \\ -\lambda_g^-, & d_i^{k+1} > d_i^k, \\ 0, & \text{otherwise,} \end{cases} \quad k \leq K_g. \quad (10)$$

After K_g , this term is removed:

$$r_{\text{guide},i}^k = 0, \quad k > K_g. \quad (11)$$

This term is not intended to represent a permanent control objective. Instead, it approximates an initial deployment instruction that would be available in a real wildfire response mission, where UAVs would normally be directed toward the estimated ignition or active fire region before autonomous suppression begins.

Finally, because the overall objective is containment rather than merely collecting local extinguishing rewards, a containment-oriented shaping term is introduced. Let B^k be the set of burning cells at timestep k , and let ∂G denote the grid boundary. The distance from the active fire front to the environment boundary is

$$d_{\partial G}^k = \min_{x \in B^k} \text{dist}(x, \partial G). \quad (12)$$

A penalty is applied when the active fire front moves closer to the boundary:

$$r_{\text{cont},i}^k = -\lambda_b \max(0, d_{\partial G}^k - d_{\partial G}^{k+1}). \quad (13)$$

At the terminal timestep, an additional global containment bonus or escape penalty is distributed equally among all agents:

$$r_{\text{terminal},i} = \frac{1}{N_a} \begin{cases} +\lambda_C, & \text{if the fire is contained,} \\ -\lambda_E, & \text{if the fire reaches } \partial G. \end{cases} \quad (14)$$

The final reward used for learning is therefore

$$\tilde{r}_i^k = \text{clip}\left(r_i^k + r_{\text{terminal},i}, r_{\min}, r_{\max}\right), \quad (15)$$

where reward clipping is applied to stabilise Q-value estimation.

This design deliberately combines local and global signals. The extinguishing and temperature terms make learning possible under partial observability, since they are directly related to information available to each UAV. The separation term improves swarm coverage and prevents redundant agent behaviour. The temporary guidance term improves early exploration without permanently biasing the learned policy. The containment term aligns individual learning with the mission-level objective of preventing the fire from escaping the operational area.

In the experiments, the reward weights were selected to preserve extinguishing as the dominant immediate reward while preventing agents from exploiting local fire-following behaviour without contributing to containment. The values used were

$$\lambda_{\text{ext}} = 3, \quad \lambda_{\text{idle}} = \frac{0.05}{N_a}, \quad \lambda_{\text{temp}} = 2, \quad \lambda_{\text{sep}} = 0.05, \quad \lambda_g^+ = \frac{0.5}{N_a}, \quad \lambda_g^- = \frac{0.2}{N_a}.$$

The reward was clipped to the interval $[-5, 10]$ at each timestep to avoid unstable Q-value growth during training.

4.4. Enhanced DQN Architecture

Transitioning from centralised learning to decentralised multi-agent learning introduces several challenges that significantly increase training difficulty. Independent replay buffers reduce the frequency with which informative experiences are encountered, while the introduction of wind-driven wildfire propagation increases environmental stochasticity and non-stationarity. To address these challenges, the baseline DQN architecture is extended through specialised feature extraction, a dueling value estimation structure, and several stabilisation mechanisms. Rather than processing the observation vector directly, the state representation is first decomposed into three independent feature extraction streams corresponding to: (i) local thermal observations, (ii) positional and fire-related information, and (iii) environmental variables associated with wind conditions. Each stream is processed independently before being fused into a common latent representation used for action selection.

This decomposition serves two important purposes. First, it enables specialised processing of different observation modalities. Second, it prevents the thermal observations from dominating the learning process. Without preprocessing, thermal information represents 25 of the 42 input features available to each UAV. Following feature extraction, the relative contribution of thermal observations is reduced to 24 of the 62 features passed to the DQN, producing a more balanced state representation and ensuring that positional, fire-state, and environmental information contribute more effectively to decision making.

The thermal stream is processed using lightweight convolution-inspired feature extraction layers designed to capture local spatial relationships within the 5×5 temperature map. Such processing is particularly important because thermal observations are inherently noisy and contain spatial patterns associated with fire-front geometry and propagation direction. To improve robustness, this stream incorporates normalisation layers and dropout regularisation, whereas the remaining streams employ standard fully connected processing layers. The outputs of the three feature extraction branches are concatenated and passed to the decision-making network. The resulting DQN consists of multiple fully connected layers followed by a Dueling DQN structure, where the network separates into independent value and advantage streams before recombination. The final output layer contains eight Q-values corresponding to the eight possible movement actions available to each UAV within the Moore neighbourhood. The dueling architecture improves learning efficiency in sparse-reward environments by explicitly separating estimation of the state value from the relative advantage of individual actions. This distinction is particularly beneficial in wild-fire suppression tasks, where multiple actions may yield similar short-term rewards despite having significantly different long-term consequences for containment performance.

Several additional stabilisation techniques are incorporated into the learning process. A soft target network update strategy based on Polyak averaging replaces conventional hard target updates. Rather than periodically replacing the target network parameters, Polyak averaging gradually updates the target network, reducing oscillations in estimated Q-values and improving convergence stability. Gradient clipping is additionally applied during backpropagation to prevent excessively large parameter updates arising from rare but highly informative experiences. The framework also incorporates Prioritised Experience Replay (PER) [47]. Each transition stored within the replay buffer is assigned a priority based on its temporal-difference error, enabling experiences associated with larger prediction errors to be sampled more frequently during training. This mechanism is particularly valuable in the wildfire environment, where many collected experiences correspond to routine exploration far from active fire fronts and contribute little useful learning information. By prioritising informative transitions, the agents focus training on experiences associated with fire discovery, suppression, and containment. Together, these architectural modifications improve representation learning, increase sample efficiency, and provide the stability required for effective decentralised learning in a highly stochastic wildfire environment. The final hyperparameters used in the Dueling DQN architecture are summarised in Table 2.

Table 2. Observation dimensions and feature extraction structure of the proposed Dueling DQN architecture.

Component	Dimension	Description
Thermal observation map	25	5×5 local temperature grid
Fire-state features	4	Direction and distance guidance
Agent proximity features	1	Nearby UAV indicator
Wind features	2	Wind speed and direction
Position features	10	Local mission context
Total raw observations	42	Input dimension
Thermal feature extractor output	24	Learned thermal representation
Other feature streams	38	Non-thermal processed features
Fused latent representation	62	Combined feature vector
Output actions	8	Moore neighbourhood movement actions

4.5. Selective Experience Sharing and Communication Constraints

In this work, communication constraints are modelled at the information-exchange level rather than at the physical communication layer. UAVs exchange only a limited number of high-priority experiences instead of complete replay buffers or network parameters, and such exchanges are only permitted when two UAVs are within a predefined communication range. Communication latency, packet loss, interference, and channel-capacity limitations are beyond the scope of this study.

Although decentralised learning improves robustness and deployment realism, fully independent learning significantly slows convergence in sparse-reward environments. To alleviate this limitation while preserving decentralised execution, the proposed framework incorporates selective experience sharing using the SUPER (Selective Sharing of Valuable Experiences) algorithm [48]. SUPER extends the concept of Prioritised Experience Replay beyond individual agents. Rather than sharing complete replay buffers or transmitting global state information, agents exchange only highly informative experiences [49]. Similar to PER, each experience is assigned a priority according to its temporal-difference error. Experiences associated with large prediction errors correspond to situations where the observed outcome differs substantially from the agent's expectation and are therefore considered particularly valuable for learning.

At predefined sharing intervals, each UAV identifies a subset of high-priority experiences from its local replay buffer. In the baseline implementation, agents collectively share up to 100 high-value experiences every 10 training episodes. These experiences are inserted into neighbouring replay buffers, where they can subsequently be sampled through PER and used to improve policy updates. By selectively sharing only the most informative experiences, the communication overhead remains significantly lower than approaches based on full replay-buffer exchange while still providing meaningful learning benefits. To improve realism, communication is additionally constrained by spatial proximity. Wildfire response operations often occur over large geographical regions where communication range, bandwidth limitations, environmental interference, and energy constraints restrict information exchange. Consequently, experience sharing is permitted only when two UAVs are located within a predefined communication radius. When a communication link is established, a sharing pair is formed and each UAV exchanges 10 high-priority experiences with its neighbour. These experiences are added to the recipient's replay buffer and become available for future training updates. The communication radius is defined in grid cells and may be varied experimentally to investigate the effect of communication constraints on learning performance. In the present work, unrestricted communication, 20-cell communication radius, and 10-cell communication radius configurations are evaluated. The resulting communication model occupies an intermediate position between fully centralised and fully decentralised learning. During training, agents benefit from peer-assisted learning through selective sharing of valuable experiences while maintaining independent replay buffers and policy networks. During execution, however, no communication is required and each UAV operates entirely autonomously using only locally observable information. Importantly, communication opportunities emerge naturally from the evolving spatial configuration of the swarm. Consequently, the effectiveness of information exchange depends not only on the quality of shared experiences but also on the probability of UAV encounters within communication range. This formulation enables investigation of the relationship between communication availability, replay diversity, learning efficiency, and emergent coordination behaviour in realistic wildfire suppression scenarios. For example, during a large wildfire mission, UAVs monitoring different regions of the fire may temporarily lose connectivity because of their geographical separation. Once two UAVs enter communication range,

they exchange only a small number of high-priority experiences before continuing their missions independently.

4.6. Hyperparameters and Training Configuration

The proposed decentralised MARL framework requires several hyperparameters related to replay memory, exploration, optimisation, episode duration, and communication. These values were selected to balance training stability, sample efficiency, and exploration in the sparse and highly stochastic wildfire environment. Each UAV maintains an independent replay buffer with a maximum capacity of 30,000 experiences. Since useful suppression experiences are relatively rare, learning is delayed until each buffer contains at least 3000 samples, corresponding to 10% of the total buffer capacity. Training updates are performed using mini-batches of 128 experiences sampled from PER.

Each model is trained for 5000 episodes. Since agents typically observe approximately 200–300 environment steps per episode, this provides more than 10^6 interaction steps during training. Episodes are terminated after a maximum of 600 steps if the fire has not already been extinguished or escaped the environment. The discount factor is set to $\gamma = 0.99$, encouraging agents to account for the long-term effect of current suppression actions on future wildfire spread. The learning rate is initialised as 3×10^{-4} and decayed during training using a multiplicative factor of 0.9 every 20,000 training steps. Exploration is controlled using an ϵ -greedy policy. The exploration rate is initialised at $\epsilon = 0.9$ and annealed to a minimum value of $\epsilon = 0.02$ over the first 1500 episodes. This relatively high initial exploration rate is necessary because agents are deployed at the environment boundary and may otherwise fail to encounter informative fire-related experiences during early training. Selective experience sharing is applied at fixed intervals during training. In the unrestricted communication case, agents share up to 100 high-priority experiences every 10 episodes. For radius-limited communication, UAVs may exchange experiences only when located within the specified communication radius. When a valid communication pair is formed, each UAV shares 10 high-priority experiences with the other agent. The communication radii evaluated in this work are unrestricted sharing, 20 cells, and 10 cells. Table 3 summarises the main hyperparameters used in the training process.

Table 3. Main training and communication hyperparameters.

Hyperparameter	Value
Replay buffer size per agent	30,000 experiences
Minimum replay size before training	3000 experiences
Mini-batch size	128
Training episodes	5000
Typical episode length	200–300 steps
Maximum episode length	600 steps
Discount factor	$\gamma = 0.99$
Initial learning rate	3×10^{-4}
Learning-rate decay	0.9 every 20,000 steps
Initial exploration rate	$\epsilon = 0.9$
Minimum exploration rate	$\epsilon = 0.02$
Exploration decay period	1500 episodes
Action space	8 neighbouring movements
Replay sampling	Prioritised Experience Replay
Target update	Polyak averaging
Communication interval	Every 10 episodes
Unrestricted sharing limit	100 experiences
Pairwise radius-limited exchange	10 experiences per UAV
Communication radii tested	Unrestricted, 20 cells, 10 cells

5. Results

This section evaluates the performance of decentralised multi-agent reinforcement learning for wildfire suppression under different communication constraints. The analysis focuses not only on containment rates, but also on the coordination failures and behavioural limitations that emerge when agents operate with partial observability and restricted information exchange. Three communication configurations were compared: unrestricted sharing, a 20-cell sharing radius, and a 10-cell sharing radius. Each model was trained for 5000 episodes, and the best-performing model from training was subsequently evaluated over 1000 test episodes. The environment remained highly stochastic because fire propagation depended on probabilistic spread dynamics and wind influence. As a result, performance should be interpreted not only as a measure of learned suppression ability, but also as evidence of how robust each decentralised policy is to uncertain wildfire behaviour.

5.1. Training Performance

Table 4 reports the best training metrics obtained for each communication configuration. The 10-cell sharing model achieved the highest rolling average reward and the highest rolling average episode score. The unrestricted sharing model produced the lowest overall training score, despite having access to the greatest amount of shared experience.

Table 4. Training performance for different communication configurations.

Communication Configuration	Highest Rolling Reward	Score at Max Reward	Highest Rolling Score
Unrestricted sharing	28.806	12.114	17.108
10-cell sharing radius	34.254	19.365	22.293
20-cell sharing radius	30.508	13.091	17.716

The result is notable because unrestricted sharing does not lead to the strongest learning outcome. Instead, the best training performance occurs under the most restrictive communication setting. This suggests that, in this environment, communication constraints may improve learning by limiting the repeated circulation of similar experiences and increasing the diversity of information encountered by each agent during training.

5.2. Containment Performance and Suppression Effectiveness

The primary evaluation metrics were fire containment and suppression effectiveness. A wildfire was considered contained if it did not reach any boundary cell of the 100×100 environment. Figures 3–5 show the statistical results for the unrestricted sharing, 10-cell sharing radius, and 20-cell sharing radius cases. Table 5 summarises the test results over 1000 independent episodes.

Table 5. Containment performance over 1000 evaluation episodes.

Communication Configuration	Contained (%)	Escaped (%)
Unrestricted sharing	26.5	73.5
10-cell sharing radius	41.1	58.9
20-cell sharing radius	33.3	66.7

The containment results follow the same trend observed during training. The 10-cell communication radius achieves the highest containment rate, containing 41.1% of fires. The 20-cell model also improves upon unrestricted sharing, although less substantially. These results indicate that communication constraints do not necessarily degrade decen-

tralised MARL performance. In this case, constrained communication improves both training and evaluation outcomes.

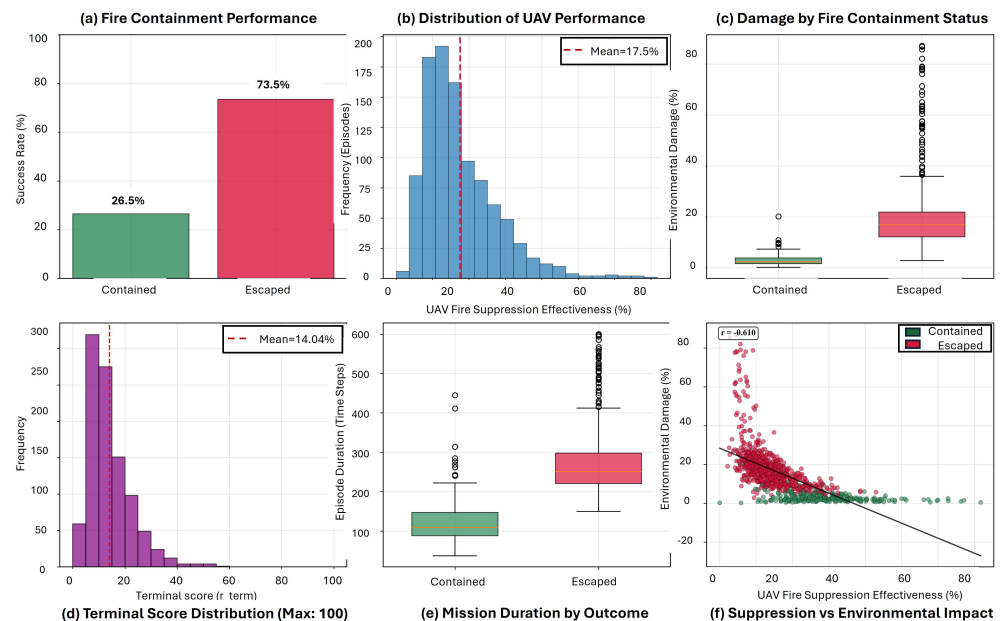


Figure 3. Containment performance under unrestricted sharing.

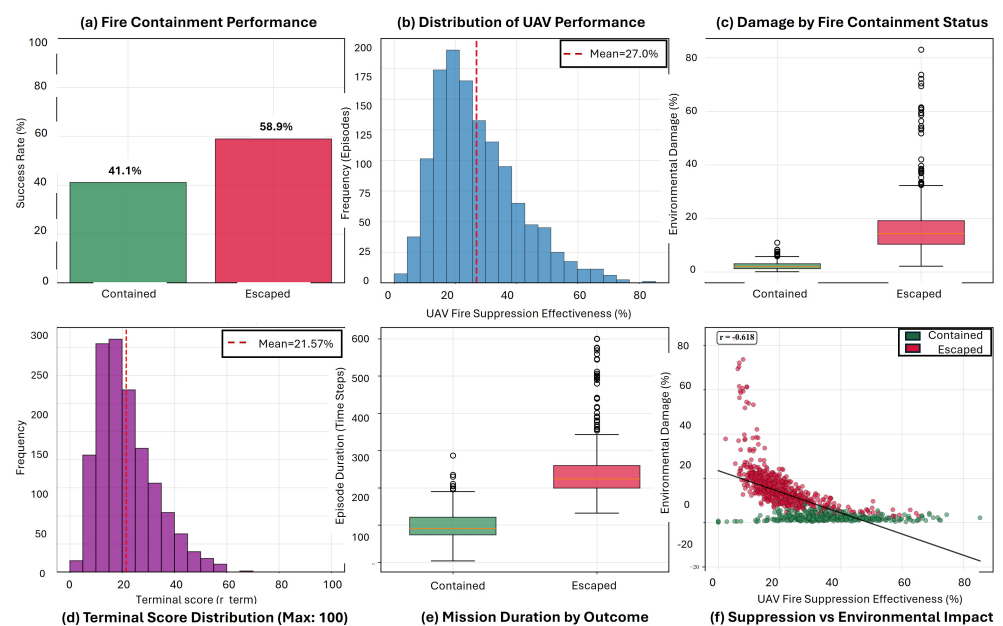


Figure 4. Containment performance under 10 cell sharing radius.

A likely explanation is that unrestricted sharing causes certain agents or trajectories to dominate the replay process. When experiences are repeatedly shared across the whole team, agents may train on a narrow and redundant subset of transitions. Radius-limited sharing, by contrast, distributes experience exchange through local pairwise interactions. This appears to produce more useful variation in the replay buffers and improves policy generalisation.

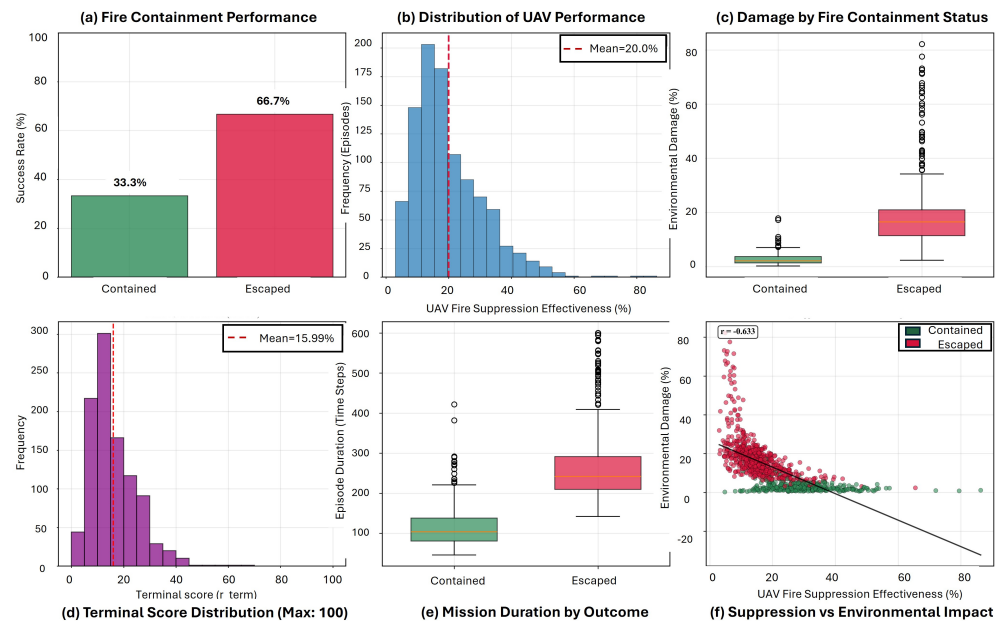


Figure 5. Containment performance under 20 cell sharing radius.

Containment alone does not fully describe agent performance. Some fires may be contained with limited intervention, whereas others require sustained suppression. For this reason, the percentage of burned cells extinguished by the agents was also analysed. The 10-cell model achieved the highest average suppression effectiveness, with a reported mean of approximately 27.0%. The unrestricted and 20-cell models achieved lower mean suppression effectiveness, approximately 17.5% and 20.0%, respectively. This supports the containment results and suggests that the 10-cell model is more effective at actively engaging with fires across the test set.

However, higher suppression effectiveness does not automatically imply better containment strategy. A recurring behaviour observed across models was that agents often chased actively burning cells instead of positioning themselves to prevent future spread. This indicates a reward-alignment problem. The agents are directly rewarded for extinguishing burning cells, but containment requires anticipating the fire front and acting before spread reaches critical boundary regions. The learned behaviour is therefore locally rational but globally suboptimal.

5.3. Observed Coordination Failures

The quantitative results show that the constrained communication model performs best, but they do not fully explain why containment remains limited. Behavioural inspection reveals two major failure modes: reward-driven fire chasing and loss of situational awareness.

5.3.1. Reward-Driven Fire Chasing

Agents frequently followed the active fire front because this behaviour provided immediate extinguishing rewards. In practice, however, this response is often ineffective for containment. A more desirable strategy would require agents to position themselves ahead of the fire, suppress boundary expansion, and coordinate to restrict the fire's growth. Figure 6 illustrates a typical reward-driven behaviour in which one UAV repeatedly follows the active fire front while the remaining UAVs fail to establish an effective containment boundary. Although the leading UAV successfully suppresses local burning cells, the swarm does not coordinate sufficiently to prevent further fire propagation.

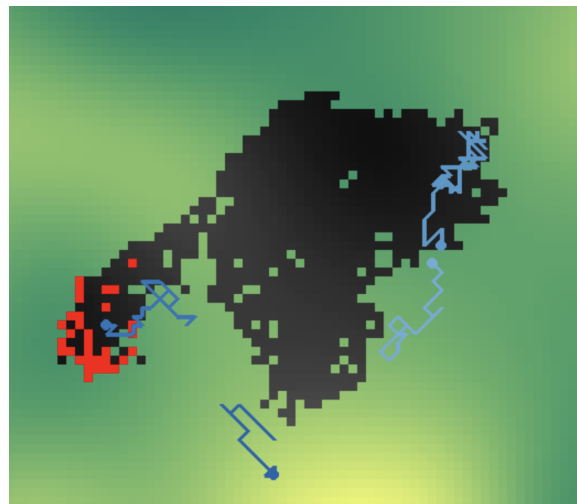


Figure 6. Example of reward-driven fire-chasing behaviour. Red cells denote actively burning vegetation, while black cells denote burned or extinguished regions. The light-blue trajectory corresponds to the highlighted UAV, whereas the blue trajectories represent the remaining UAVs in the swarm. Although the highlighted UAV successfully follows and suppresses sections of the active fire front, the overall team fails to establish an effective containment strategy, illustrating the tendency of agents to optimise local suppression rewards rather than global fire containment.

5.3.2. Loss of Situational Awareness

A second failure mode occurred when agents moved away from the active fire region and lost useful observational information. Since each agent relies on local state information, an agent that loses sight of the fire may no longer receive strong signals indicating where to move next. This can lead to ineffective wandering in areas with little suppression value. Figure 7 illustrates a representative failure case caused by limited local observations. Once several UAVs move away from the active fire front, the available local thermal information becomes insufficient to guide them back towards the wildfire, resulting in inefficient exploration and reduced containment performance.

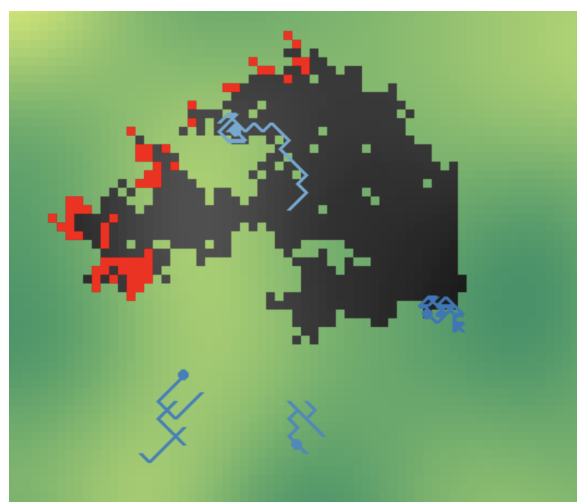


Figure 7. Example of loss of situational awareness under decentralised execution. Red cells denote actively burning vegetation, while black cells denote burned or extinguished regions. The light-blue trajectory highlights the UAV that loses contact with the active fire front, while the blue trajectories correspond to the remaining UAVs. Once local thermal information becomes insufficient, the affected UAV is unable to relocate the wildfire efficiently, resulting in ineffective exploration and reduced cooperative suppression performance.

This behaviour highlights the limitations of decentralised execution under partial observability. Even when agents have learned useful local suppression behaviours, the team can still fail if individual agents become disconnected from the shared task. The result suggests that future systems may require event-triggered communication, limited global alerts, or learned signalling mechanisms to help agents recover situational awareness without requiring continuous full communication.

5.4. Ablation Studies

We further conduct ablation studies to show how wind and the number of agents affect the overall fire suppression performance.

5.4.1. Removing Wind

To evaluate the influence of environmental stochasticity, an additional model was trained and tested without wind effects, using the best-performing 10-cell communication setting. The results are observed in Figure 8.

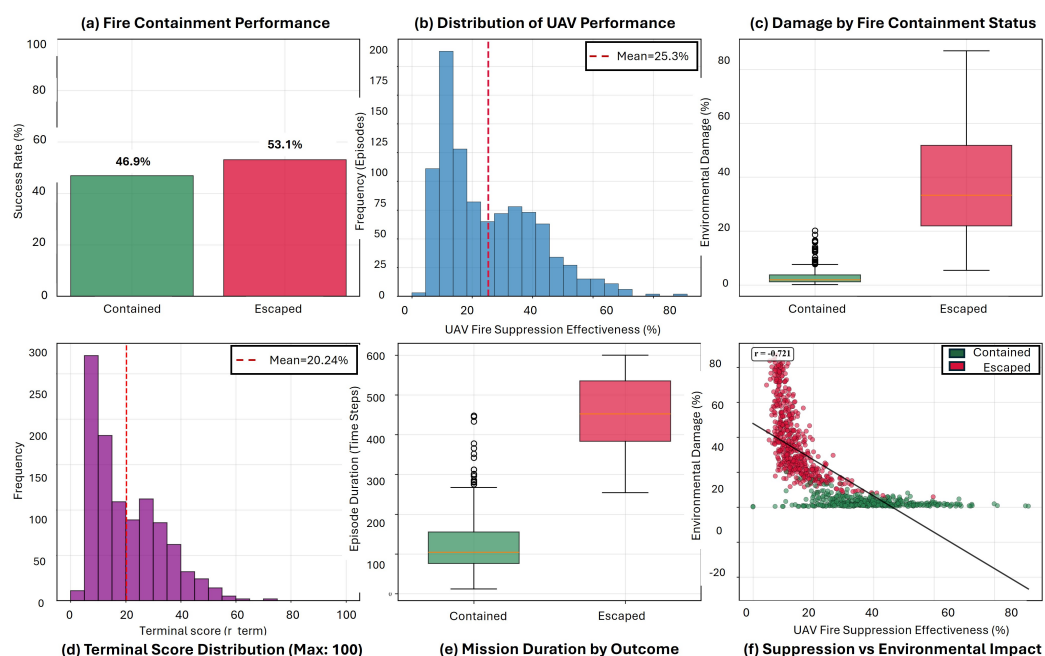


Figure 8. Performance of the 10-cell communication model in the no-wind environment.

Removing wind increased containment performance to 46.9%. This improvement suggests that wind contributes to the difficulty of the task by increasing the unpredictability of fire spread. Nevertheless, the improvement is moderate rather than transformative. The model still fails in more than half of the test episodes, indicating that environmental stochasticity is not the only limitation. Coordination failures, reward misalignment, and partial observability remain major constraints even in the simplified no-wind setting.

The no-wind results also show that agents still struggle with larger fires. Although the absence of wind makes spread more predictable, it can also result in longer episodes and larger burned regions. This reinforces the conclusion that the agents are not yet learning robust containment strategies; they are primarily learning reactive suppression behaviours.

5.4.2. Increasing the Number of Agents

A second ablation examined whether increasing the number of agents improves containment as shown in Table 6. The 10-cell communication model was extended to teams of seven and nine agents.

Table 6. Containment performance as the number of agents increases.

Number of Agents	Containment Rate (%)
5 agents	41.1
7 agents	48.0
9 agents	51.0

Increasing the number of agents improves containment, with the nine-agent model reaching 51.0%. This confirms that greater spatial coverage is beneficial in wildfire suppression. However, the improvement is smaller than might be expected from the increase in available agents. This suggests that simply adding agents does not solve the central coordination problem.

The result is important for communication-constrained decentralised MARL because larger teams also increase the complexity of coordination and experience sharing. In the reported experiments, increasing the number of agents also increased the average number of shared experiences per sharing episode. This may reduce the effectiveness of experience sharing if communication becomes too frequent or redundant. Therefore, scaling the number of agents requires more than increasing team size; it also requires communication protocols that preserve useful diversity while supporting coordinated action.

5.5. Discussions

The results show that moderate communication constraints can improve decentralised MARL performance in wildfire suppression. The 10-cell communication radius produced the strongest training performance, the highest containment rate, and the highest suppression effectiveness among the three main communication configurations. This suggests that restricted local sharing can be more effective than unrestricted information exchange when replay diversity and experience relevance are important. At the same time, the results reveal persistent challenges. Agents often optimise local suppression rewards rather than global containment objectives, leading to fire-chasing behaviour. Agents also lose situational awareness under partial observability, reducing team effectiveness during longer episodes. Ablation studies confirm that reducing environmental stochasticity and increasing team size improve performance, but neither fully resolves the coordination problem. Overall, the findings indicate that the key challenge in communication-constrained wildfire suppression is not merely enabling agents to communicate, but designing communication and reward mechanisms that promote strategic containment, preserve useful experience diversity, and maintain situational awareness during decentralised execution.

Compared with our previous centralised framework [43], the proposed approach replaces the shared policy and replay buffer with independent policies and selective peer-to-peer experience sharing. Furthermore, globally available neighbouring-agent information has been removed from the observation space, allowing each UAV to operate solely from local observations during execution. Although both approaches address cooperative wildfire suppression, the different wildfire environments and learning architectures prevent a direct quantitative comparison of performance.

Although the proposed framework provides a flexible environment for investigating decentralised coordination, it is not intended to prescribe operational deployment parameters. Variables such as UAV flight altitude, sensing footprint, fleet size, and wildfire extent depend on the sensing payload, communication infrastructure, airspace regulations, and mission objectives. The present work therefore focuses on the learning problem itself, while operational parameter optimisation constitutes an important direction for future research.

5.6. Limitations and Future Research Directions

Although the proposed framework represents a substantial step toward realistic decentralised wildfire suppression, several simplifying assumptions remain. The wildfire environment is modelled as a two-dimensional grid and does not explicitly represent terrain elevation, fuel moisture content, atmospheric conditions, or heterogeneous wind fields. Wind is assumed to be globally consistent across the environment, whereas real wildfire behaviour is influenced by highly localised and time-varying meteorological effects. Similarly, UAV agents are assumed to possess accurate localisation and sensing capabilities, while communication within the specified communication radius is considered reliable. Practical deployment would likely involve sensor uncertainty, packet loss, intermittent communication, and platform resource limitations such as battery constraints and finite retardant payloads. Furthermore, the wildfire environment is formulated as an abstract discretised representation for reinforcement learning rather than a geographically referenced operational model. Consequently, the grid resolution does not correspond to a fixed physical ground area, and UAV operating altitude is not explicitly modelled. In practice, the relationship between sensing resolution, flight altitude, sensor footprint, and the number of UAVs required depends on the sensing payload, mission objectives, airspace regulations, and wildfire characteristics. Likewise, the framework does not consider coordination with manned aerial firefighting assets, whose integration would require explicit modelling of airspace management and operational safety constraints. Incorporating these operational considerations represents an important direction for future research.

6. Conclusions

This paper presented a decentralised multi-agent reinforcement learning framework for UAV-based wildfire suppression under realistic communication and environmental constraints. By incorporating wind-driven wildfire propagation and decentralised learning mechanisms, the work moves beyond many of the simplifying assumptions commonly found within existing MARL wildfire studies. The introduction of selective experience sharing proved particularly important within the highly stochastic wildfire environment, significantly improving learning efficiency while preserving decentralised execution. Experimental results demonstrated that communication-aware decentralised MARL can achieve meaningful wildfire containment behaviours despite sparse rewards and complex propagation dynamics. Nevertheless, the work also highlights several remaining challenges. Realistic wildfire environments introduce substantial instability into reinforcement learning processes, particularly when communication limitations and partial observability are considered simultaneously. Future work should therefore investigate more advanced coordination strategies, dynamic weather systems, continuous action spaces, and resource-constrained UAV models. Ultimately, the results suggest that decentralised UAV swarms represent a promising direction for future autonomous wildfire response systems, particularly when communication-aware learning strategies are incorporated into the training framework.

Author Contributions: Conceptualization, S.C. and M.C.; methodology, S.C. and M.C.; software, M.C.; validation, S.C.; formal analysis, S.C. and A.P.; investigation, S.C. and M.C.; resources, A.P. and A.T.; writing—original draft preparation, A.P. and A.T.; visualization, A.P.; supervision, A.P. and A.T.; project administration, A.T.; funding acquisition, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article material; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Uncrewed Aerial Vehicle
MARL	Multi-Agent Reinforcement Learning
DQN	Deep Q-Network
SAR	Search and Rescue
VDN	Value Decomposition Network
MADQN	Multi-Agent Deep Q-Network
PER	Prioritized Experience Replay
SUPER	Selective Sharing of Valuable Experiences

References

1. Tett, S.F.B.; Stott, P.A.; Allen, M.R.; Ingram, W.J.; Mitchell, J.F.B. Causes of twentieth-century temperature change near the Earth's surface. *Nature* **1999**, *399*, 569–572.
2. Paveglio, T.; Brenkert-Smith, H.; Hall, T.; Smith, A. Understanding social impact from wildfires: advancing means for assessment. *Int. J. Wildland Fire* **2015**, *24*, 212–224. <https://doi.org/10.1071/WF14091>.
3. Palaiologou, P.; Ager, A.A.; Nielsen-Pincus, M.; Evers, C.R.; Day, M.A. Social vulnerability to large wildfires in the western USA. *Landsc. Urban Plan.* **2019**, *189*, 99–116.
4. Deo, A.A.; Ganer, D.; Nair, G. Tropical cyclone activity in global warming scenario. *Nat. Hazards* **2011**, *59*, 771–786. <https://doi.org/10.1007/S11069-011-9794-8>.
5. Chapman, S.C.; Watkins, N.W.; Stainforth, D.A. Warming trends in summer heatwaves. *Geophys. Res. Lett.* **2019**, *46*, 1634–1640.
6. Smith, H.J. Increasing heatwave activity. *Science* **2019**, *363*, 596.
7. Hirabayashi, Y.; Mahendran, R.; Koirala, S.; Konoshima, L.; Yamazaki, D.; Watanabe, S.; Kim, H.; Kanae, S. Global flood risk under climate change. *Nat. Clim. Change* **2013**, *3*, 816–821. <https://doi.org/10.1038/NCLIMATE1911>.
8. Gruffeille, C.; Perrusquía, A.; Tsourdos, A.; Guo, W. Disaster area coverage optimisation using reinforcement learning. In *Proceedings of the 2024 International Conference on Unmanned Aircraft Systems (ICUAS)*; IEEE: New York, NY, USA, 2024; pp. 61–67.
9. Liu, Y.; Stanturf, J.A.; Goodrick, S.L. Trends in global wildfire potential in a changing climate. *For. Ecol. Manag.* **2010**, *259*, 685–697. <https://doi.org/10.1016/J.FORECO.2009.09.002>.
10. Kumar, A.; Perrusquía, A.; Al-Rubaye, S.; Guo, W. Wildfire and smoke early detection for drone applications: A light-weight deep learning approach. *Eng. Appl. Artif. Intell.* **2024**, *136*, 108977.
11. Tekalign, W. Impacts of Wildfire and Prescribed Fire on Wildlife and Habitats: A Review. *J. Nat. Sci. Res.* **2016**, *6*, 15–27.
12. Chitale, V.; Behera, M.D. How will forest fires impact the distribution of endemic plants in the Himalayan biodiversity hotspot? *Biodivers. Conserv.* **2019**, *28*, 2259–2273.
13. Krawchuk, M.; Moritz, M.; Parisien, M.; Dorn, J.V.; Hayhoe, K. Global Pyrogeography: the Current and Future Distribution of Wildfire. *PLoS ONE* **2009**, *4*, e5102. <https://doi.org/10.1371/journal.pone.0005102>.
14. Withen, P. Climate Change and Wildland Firefighter Health and Safety. *New Solut. J. Environ. Occup. Health Policy* **2015**, *24*, 577–584. <https://doi.org/10.2190/NS.24.4.i>.
15. Quero, C.O.; Martinez-Carranza, J. Unmanned aerial systems in search and rescue: A global perspective on current challenges and future applications. *Int. J. Disaster Risk Reduct.* **2025**, *118*, 105199.
16. Drew, D.S. Multi-agent systems for search and rescue applications. *Curr. Robot. Rep.* **2021**, *2*, 189–200.
17. Tan, L.; Zhao, H. A multi-UAV rapid post-disaster search and rescue method based on deep reinforcement learning. *Complex Intell. Syst.* **2026**, *12*, 41.
18. Ekechi, C.C.; Elfouly, T.; Alouani, A.; Khattab, T. A survey on UAV control with multi-agent reinforcement learning. *Drones* **2025**, *9*, 484.
19. Cao, X.; Li, M.; Tao, Y.; Lu, P. HMA-SAR: Multi-agent search and rescue for unknown located dynamic targets in completely unknown environments. *IEEE Robot. Autom. Lett.* **2024**, *9*, 5567–5574.
20. Peng, K.; Qi, H.; Ma, T. Graph based multi-agent reinforcement learning with evolutionary population for cooperation. *Neural Netw.* **2025**, *196*, 108437.
21. Zhao, Y.; Luo, H.; Tao, H.; Wang, J.; Zhou, J.; Wu, K. MECOS: Cooperative Multi-UAV-assisted Cross-boundary Maritime Data Collection Leveraging MARL and LLM. *IEEE Internet Things J.* **2025**, *13*, 8561–8577.

22. Fang, H.; Tam, W.C.; Lovreglio, R.; Emon, M.I.S.; Huang, M.X. Development of a tenability-based path planning model for building fire evacuations using reinforcement learning. *J. Build. Eng.* **2026**, *118*, 115132.
23. Khabibulin, R. Reinforcement Learning–Integrated Digital Twin for Proactive Fire Response Decision Support. In *Proceedings of the 2025 7th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*; IEEE: New York, NY, USA, 2025; pp. 655–659.
24. Murray, L.; Castillo, T.; de Diego, I.M.; Weber, R.; Gonzalez-Olabarria, J.R.; Garcia-Gonzalo, J.; Weintraub, A.; Carrasco-Barra, J. Deep reinforcement learning for optimal firebreak placement in forest fire prevention. *Appl. Soft Comput.* **2025**, *175*, 113043.
25. Tapley, A.; Dotter, M.; Doyle, M.; Fennelly, A.; Gandikota, D.; Smith, S.; Threet, M.; Welsh, T. Reinforcement learning for wildfire mitigation in simulated disaster environments. *arXiv* **2023**, arXiv:2311.15925.
26. Luo, Y.; Yang, T. HIG-SAR: A Hierarchical Information-Guided MARL Framework for Multi-UAV Search and Rescue in Unknown Environments. In *Proceedings of the 2026 7th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*; IEEE: New York, NY, USA, 2026; pp. 391–396.
27. Mishra, M.; Poddar, P.; Agrawal, R.; Chen, J.; Tokekar, P.; Sujit, P. Multi-agent deep reinforcement learning for persistent monitoring with sensing, communication, and localization constraints. *IEEE Trans. Autom. Sci. Eng.* **2024**, *22*, 2831–2843.
28. Krichen, M.; Abdalzaher, M.S.; Elwekeil, M.; Fouda, M.M. Managing natural disasters: An analysis of technological advancements, opportunities, and challenges. *Internet Things-Cyber-Phys. Syst.* **2024**, *4*, 99–109.
29. Doo, J. *Unsettled Issues Regarding the Use of eVTOL Aircraft During Natural Disasters*; Technical Report, SAE Technical Paper; SAE International: Warrendale, PA, USA, 2022.
30. El Debeiki, M.; Al-Rubaye, S.; Perrusquía, A.; Conrad, C.; Flores-Campos, J.A. An Advanced Path Planning and UAV Relay System: Enhancing Connectivity in Rural Environments. *Future Internet* **2024**, *16*, 89.
31. Łświechowski, M.; Godlewski, K.; Sawicki, B.; Mańdziuk, J. Monte Carlo tree search: A review of recent modifications and applications. *arXiv* **2021**, arXiv:2103.04931.
32. Baker, C.A.; Ramchurn, S.; Teacy, W.; Jennings, N.R. Planning search and rescue missions for UAV teams. In *ECAI 2016*; IOS Press: Amsterdam, The Netherlands, 2016; pp. 1777–1782.
33. Bildik, E.; Tsourdos, A.; Perrusquía, A.; Inalhan, G. Decoys Deployment for Missile Interception: A Multi-Agent Reinforcement Learning Approach. *Aerospace* **2024**, *11*, 684.
34. Ai, B.; Jia, M.; Xu, H.; Xu, J.; Wen, Z.; Li, B.; Zhang, D. Coverage path planning for maritime search and rescue using reinforcement learning. *Ocean. Eng.* **2021**, *241*, 110098.
35. Panahi, F.H.; Panahi, F.H.; Ohtsuki, T. A reinforcement learning-based fire warning and suppression system using unmanned aerial vehicles. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 1–16.
36. Rahman, A.; Bhattacharya, A.; Ramachandran, T.; Mukherjee, S.; Sharma, H.; Fujimoto, T.; Chatterjee, S. AdverSAR: Adversarial Search and Rescue via Multi-Agent Reinforcement Learning. In *Proceedings of the 2022 IEEE International Symposium on Technologies for Homeland Security (HST)*, Virtual, 14–15 November 2022; pp. 1–7. <https://doi.org/10.1109/HST56032.2022.10025434>.
37. Lei, C.; Wu, S.; Yang, Y.; Xue, J.; Zhang, Q. Maritime Search and Rescue Leveraging Heterogeneous Units: A Multi-Agent Reinforcement Learning Approach. In *Proceedings of the 2023 IEEE/CIC International Conference on Communications in China (ICCC)*, Dalian, China, 10–12 August 2023; pp. 1–6. <https://doi.org/10.1109/ICCC57788.2023.10233604>.
38. Sadhu, V.; Sun, C.; Karimian, A.; Tron, R.; Pompili, D. Aerial-DeepSearch: Distributed Multi-Agent Deep Reinforcement Learning for Search Missions. In *Proceedings of the 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Virtual, 10–13 December 2020; pp. 165–173. <https://doi.org/10.1109/MASS50613.2020.00030>.
39. Yang, Z.; Nguyen, L.; Zhu, J.; Pan, Z.; Li, J.; Jin, F. Coordinating Disaster Emergency Response with Heuristic Reinforcement Learning. In *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Virtual, 7–10 December 2020; pp. 565–572. <https://doi.org/10.1109/ASONAM49781.2020.9381416>.
40. Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W.M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J.Z.; Tuyls, K.; et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv* **2017**, arXiv:1706.05296.
41. Viseras, A.; Meissner, M.; Marchal, J. Wildfire Front Monitoring with Multiple UAVs using Deep Q-Learning. *IEEE Access* **2021**, *13*, 123269–123281. <https://doi.org/10.1109/ACCESS.2021.3055651>.
42. Haksar, R.N.; Schwager, M. Distributed Deep Reinforcement Learning for Fighting Forest Fires with a Network of Aerial Robots. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 1–5 October 2018; pp. 1067–1074. <https://doi.org/10.1109/IROS.2018.8593539>.
43. Collignon, M.; Perrusquía, A.; Tsourdos, A.; Guo, W. Search and rescue operations in wildfires using unmanned aerial vehicles: A multi-agent deep reinforcement learning approach. *Neurocomputing* **2025**, *653*, 131211.
44. Zhao, M. Breaking the Cold-Start Barrier: Reinforcement Learning with Double and Dueling DQNs. *arXiv* **2025**, arXiv:2508.21259.

45. Andrews, P.L. *The Rothermel Surface Fire Spread Model and Associated Developments: A Comprehensive Explanation*; Gen. Tech. Rep. RMRS-GTR-371; US Department of Agriculture, Forest Service, Rocky Mountain Research Station: Fort Collins, CO, USA, 2018; Volume 371, 121p.
46. Noble, I.; Gill, A.; Bary, G. McArthur's fire-danger meters expressed as equations. *Aust. J. Ecol.* **1980**, *5*, 201–203.
47. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized experience replay. *arXiv* **2015**, arXiv:1511.05952.
48. Gerstgrasser, M.; Danino, T.; Keren, S. Selectively sharing experiences improves multi-agent reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 59543–59565.
49. Kumar, R.R.; Varakantham, P. On solving cooperative MARL problems with a few good experiences. *arXiv* **2020**, arXiv:2001.07993.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.