RL-Driven Bandwidth Adaptation for Cognitive Weather Radars

Pappas, A.; Yarovoy, A.; Fioranelli, F.; Sardar, S.; Schleiss, M.

DOI
10.1109/RadarConf2559087.2025.11205041

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the 2025 IEEE Radar Conference, RadarConf 2025

RL-driven bandwidth adaptation for cognitive weather radars

Apostolos Pappas, Alexander Yarovoy, Francesco Fioranelli
Microwave Sensing, Signals & Systems (MS3) Group
TU Delft
Delft, The Netherlands
a.pappas-2@tudelft.nl, a.yarovoy@tudelft.nl, f.fioranelli@tudelft.nl

Shafi Sardar, Marc Schleiss
Department of Geoscience & Remote Sensing
TU Delft
Delft, The Netherlands
s.s.a.sardar@tudelft.nl, m.a.schleiss@tudelft.nl

2025 IEEE Radar Conference (RadarConf25) | 979-8-3315-4433-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/RADARCONF2559087.2025.11205041

***Abstract*—The problem of enabling adaptive capabilities in the context of weather radar is considered in this paper. Inspired by the cognitive radar framework, an approach based on Reinforcement Learning (RL) is formulated to deal with the monitoring of multiple storm cells moving near a potential area of interest. The approach aims to dynamically adjust the radar waveform bandwidth, and consequently maximum measurable range and range resolution, in order to provide the best monitoring based on a purposely-defined reward function. The approach is successfully validated with a simulator developed in Python & *StoneSoup*. Results demonstrate that the proposed method outperforms traditional fixed-scan ('sit and spin') strategies commonly used in weather radar operations.**

***Index Terms*—cognitive radar, adaptive radar, weather radar, reinforcement learning**

I. Introduction

Weather radar plays a fundamental role in meteorology, enabling the detection, analysis, and forecasting of precipitation and severe weather events. By transmitting electromagnetic waves and analyzing their reflections from hydrometeors, weather radars can estimate the intensity, distribution, and movement of rain cells among others. Most conventional weather radar systems in use today use a static, pre-defined 'sit-and-spin' scanning strategy, mechanically scanning 360° of their surroundings in azimuth. They rarely adapt any of their operational parameters in response to their environment, which can lead to sub-optimal performance. Cognitive radar approaches introduced a paradigm shift by incorporating environment learning, adaptive sensing, and real-time decision-making, and they could also benefit traditional weather radar.

Even though the cognitive radar paradigm has been used in many applications such as spectrum sensing and sharing [1], [2], resource allocation and task scheduling [3], [4] among others, there are only a handful of studies regarding the use of cognitive weather radar. For example, Manfredi et al. [5], [6], proposed the application of Reinforcement Learning (RL) among other methods, to enable the intelligent focusing of weather radars on ongoing targets via adaptive sector selection or the usage of switching Kalman filters. Their study showed that intelligently selecting the scanned sector size can be used to enhance data quality and resolution, improving the radar's ability to detect, classify, and track meteorological phenomena with greater precision. Therefore, in this work we further assume that 360° mechanically scanning FMCW radars could benefit from the adjustment of their operational bandwidth, as this could enhance range resolution, at the cost of maximum observable range.

This work is funded by the Dutch Research Council NWO under the Open Technology Programme project *SMARTER*.

The emergence of RL in cognitive radar applications has offered significant advantages by enabling adaptive and intelligent decision-making. Unlike traditional rule-based approaches, RL allows the radar system to learn optimal strategies through interaction with its environment, improving performance over time. This adaptability enhances detection accuracy, resource allocation, and resilience to interference. Furthermore, RL-based cognitive radar systems can generalize across varying operational conditions, making them well-suited for complex, uncertain scenarios where traditional methods may be less effective. Besides [6], another case of sector scanning using RL is [7], where its use led to better decision making and overall strategies than human actions.

In this paper, results of the application of RL for the adaptive selection of bandwidth values by a single radar inspired by the cognitive radar paradigm are presented, focusing on the applicability of the perception-action-circle. Two key contributions are considered. The first is a simulation framework tailored for weather radar applications, enabling the evaluation of adaptive sensing strategies in meteorological environments, with an arbitrary number of targets in the scene. The second is a new reward function for bandwidth allocation, designed to balance measurement resolution with coverage while taking into account the relative threat posed by different targets.

The paper is structured as follows. Section II introduces the problem and its key components such as states and actions. In Section III key elements of the environment, the simulator and proposed reward are presented. Initial simulation results are shown in Section IV, followed by a short discussion in Section V and some conclusions.

II. Problem formulation

Accurate definition of system, observation, and action spaces, along with state transitions, is crucial for adaptive sensing and decision-making in cognitive radar. In the following, the key components of our RL framework are presented.

A. System state

Following [8] and [9], the system state can be defined as a combination of target, environment and platform kinematics. Here, storms are modeled as extended targets with a-priori defined shape characteristics included as part of the system state. For simplicity, we only consider fixed radar platforms that do not change position. Let there be N targets in the scene, with $\mathbf{x}_t$ and $\mathbf{y}_t$ denoting vectors containing the location of the center-of-mass (COM) for each storm target at time t. Furthermore, the velocities of each target center can be expressed by vectors $\dot{\mathbf{x}}_t$ and $\dot{\mathbf{y}}_t$, while $\mathbf{a}_t$ contains the area of each target at time instance t. $\mathbf{z}_{\max,t}$ denotes the maximum reflectivity of each target present in the scene. Thus, we define the system state $\mathbf{S}$ for a single radar at instance $t \in \mathcal{T}$ as $\mathbf{S} = \{\mathbf{E}, \mathbf{P}\}$, where $\mathbf{E} = \{\mathbf{E}_t, \mathbf{A}\}$ denotes the environment variables, containing both target information, $\mathbf{E}_t$, and information on so-called Areas of Interest (AoI) in $\mathbf{A}$, the role of which will be described in more detail in Section III-C. $\mathbf{P}$ describes the platform state. More in detail, $\mathbf{E}$ can then be written as:

$$\mathbf{E} = \left\{ \mathbf{E}_t = \left\{ \left[\mathbf{x}_t, \mathbf{y}_t, \dot{\mathbf{x}}_t, \dot{\mathbf{y}}_t, \mathbf{z}_{\max,t}, \mathbf{a}_t \right], \mathbf{A} = \left[x_A, y_A, a_A \right] \right\} \middle| \begin{array}{l} \mathbf{x}_t, \mathbf{y}_t, \dot{\mathbf{x}}_t, \dot{\mathbf{y}}_t \in \mathbb{R}^{(N\times 1)},\ \mathbf{z}_{\max,t}, \mathbf{a}_t \in \mathbb{R}^{+(N\times 1)}, \\ x_A, y_A \in \mathbb{R},\ a_A \in \mathbb{R}^+ \quad t \in \mathcal{T} \end{array} \right\} \tag{1}$$

In this paper, we consider only a case with a single AoI, which means that the contents of $\mathbf{A}$, which are the position (x^A, y^A) and radius α^A, are scalars. The second part of $\mathbf{S}$ contains, as mentioned, the radar platform state.

As we consider only a single fixed radar, the platform state $\mathbf{P}$ is sufficiently described by a vector containing the cartesian coordinates of its position (x_0, y_0), the bandwidth used at t, B_t, and the corresponding range resolution, ΔR_t and maximum range, $R_{max,t}$.

$$\mathbf{P} = \left\{ \mathbf{P}_t = \left[x_0, y_0, B_t, \Delta R_t, R_{max,t} \right] \middle| \begin{array}{l} x_0, y_0 \in \mathbb{R},\ B_t, \Delta R_t, R_{max,t} \in \mathbb{R}_+, \quad t \in \mathcal{T} \end{array} \right\} \tag{2}$$

Note that, as with $\mathbf{E}$, $\mathbf{P}$ can be expanded in future work to contain more radar parameters that may change over time.

B. Observation state

The state formulation $\mathbf{S}$, as expanded in (1)-(2), represents the true state of the system and is not what the radar agent will observe. Therefore, we also need to define the observation space of our radar agent. Since in this work the radar is observing storm cell targets, the measurement model is based on the reconstruction of their reflectivity. This is further clarified in Sections III-A and III-B.

The radar performs measurements of reflectivity at range r and azimuth θ. After completing a full scan, the acquired data are used to calculate the COM of each target, which feeds into the observation state. Due to the range limitations of the considered radar, there is the possibility that not all targets can be seen and tracked, making the observation state a subset of the system state. With this in mind, one can write the observation state as $\mathbf{O} = \{\mathbf{M}, \mathbf{P}\}$, with $\mathbf{P}$ being the same as in (2), and $\mathbf{M}$ representing the measurement state. Considering that the radar may not observe every target at all times, and defining the number of observed targets as N' with $N' = N$ when all targets are observed, $\mathbf{M}$ becomes:

$$\mathbf{M} = \left\{ \mathbf{M}_t = \left\{ \left[\hat{\mathbf{r}}_t, \hat{\theta}_t, \hat{\mathbf{z}}_{\max,t}, \hat{\mathbf{d}}_t \right], \mathbf{A} = \left[x_A, y_A, a_A \right] \right\} \middle| \begin{array}{l} \hat{\mathbf{r}}_t, \hat{\mathbf{z}}_{\max,t}, \hat{\mathbf{d}}_t \in \mathbb{R}^{+(N'\times 1)}, \hat{\theta}_t \in \mathbb{R}^{(N'\times 1)}, \\ x_A, y_A \in \mathbb{R},\ a_A \in \mathbb{R}^+, \quad t \in \mathcal{T} \end{array} \right\} \tag{3}$$

where $\hat{\mathbf{r}}_t$ and $\hat{\theta}_t$ are the measured range and azimuth of the center of mass of each target, and $\hat{\mathbf{z}}_{\max,t}$ the maximum measured reflectivity per target at time t. Additionally, $\hat{\mathbf{d}}_t$ expresses the polar distance of the observed targets from the area of interest as estimated by the radar measurements. The AoI vector $\mathbf{A}$ is defined in the same manner as in (1).

C. Action space

In FMCW radars, the maximum usable range is $R_{max} = (cF_{LPF}T_s)/2B$ where c is the speed of light, F_{LPF} is the Low Pass Filter (LPF) cutoff frequency, T_s the sweep time and B the radar bandwidth. As the radar range resolution is defined by its bandwidth as $\Delta R = \frac{c}{2B}$, the aforementioned formula in reduces to $R_{max} = \Delta R \cdot F_{LPF} \cdot T_s$. Hence, decreasing the bandwidth results in a coarser range resolution and larger maximum range. A coarser range resolution causes the radar signal to be averaged across broader areas, which results in underestimated target reflectivity values. This trade-off is known as the 'range-resolution' dilemma.

To limit the action space of the radar agent, and assuming that weather targets do not move with very high velocities, we define 7 available actions. The agent can modify its bandwidth by $\pm$5 MHz, $\pm$1 MHz, $\pm$0.5 MHz or take no action at all. The min and max possible bandwidth are 5 & 50 MHz respectively.

It shall be noted that another way to control the maximum range, independently of the bandwidth, would be to tune the sweep time, T_s. Increasing the sweep time would increase the maximum range without affecting the range resolution. However, for the sake of simplicity, and to avoid any target distortions caused by increasing T_s, we decided to only tune the bandwidth in this study, leaving multi-parameters adaptation for future work.

D. State transition

In contrast to classical RL problems, the actions taken by our agent (i.e., change the bandwidth) does not have any effect on the transition model or environment. However, it may affect the observations the cognitive weather radar makes

of the existing storm cells in the environment, as presented afterwards in Section III-A. For this work, a near constant velocity model is assumed for the modeled targets.

III. Simulation & Environment description

In this section, details of the simulated targets and radar observations are presented. Note that the storm target models are intentionally kept as simple as possible to reduce computational load. At this stage, the goal is not to achieve high physical fidelity, but rather to ensure computational efficiency. This simplification is crucial for demonstrating the feasibility of reinforcement learning (RL)-based approaches, which require running millions of simulation steps.

A. Target simulation

Every target is generated at a random location with random initial velocity. The most important part of target simulation is the generation of reflectivity values for each cell encapsulated by it. As designing a detailed measurement model, focused on scatterer level simulation, would be computationally expensive for the needs of this study, we instead design the measured reflectivity from each target as a Gaussian. Each Gaussian is mathematically modeled as follows (note that the creation of the Gaussian is time dependent, as the storm cells and their centers of mass move while running the simulated scenarios):

$$z_{i,true}(t) = S_i \cdot e^{-\frac{\mathrm{d}_i(t)^2}{2\lambda_i(t)^2}} \tag{4}$$

Here, S_i the i_{th} storm cell's center strength, $\lambda_i(t)$ is the correlation length which controls the spread of the reflectivity within the pre-defined storm cell shape, and $\mathrm{d}_i(t) = \sqrt{(x - c_x(t))^2 + (y - c_y(t))^2}$ is the euclidean distance of each grid point (x, y) belonging to the simulated grid (in Cartesian coordinates) to the center of the target.

A simple Gaussian as the one presented in (4) is sufficient for the simulation needs of this paper, yet it is oblivious to the resolution cell sizes that it is applied to. To address this, we modify both the peak value $S_i(t)$ and spread $\lambda_i(t)$ to accommodate for not only changing resolutions, but also cell spreading due to range. Thus, we further define $S_i(t)$ as:

$$S_i(t) = \frac{S_{i,peak}}{A_C(r)^\alpha} \tag{5}$$

where $A_C(r) = r \cdot \Delta\phi \cdot \Delta R$ with $\Delta\phi$ and ΔR being the azimuth and range resolutions of the radar, r is the range of the center of the resolution cell at hand, and α is an empirically derived constant that defines the influence of the resolution cell area on the peak of the Gaussian distribution of the reflectivity. For the purposes of this work we define $0 < \alpha < 1$. We assume that the radar measurement of more homogeneous storm targets will be less affected by the different resolutions than slightly more heterogeneous targets. The type of the target is randomly chosen at its creation and it serves as an extra factor of variability in the simulations and consecutive training of radar agents. On the same note, the peak of the reflectivity distribution is also randomly generated among targets.

Similarly to the distribution's peak, we also assume that the range resolution of the radar affects the spread of the distribution in a proportional manner. Hence, we generate the spread of each target as:

$$\lambda_i(t) = \frac{\rho_i}{3} \cdot \Delta R(t)^\beta \tag{6}$$

where ρ_i is the i_{th} target's radius and $0 < \beta < 1$ a constant defining the effect of changing range resolutions on the measured reflectivity spread. With the aforementioned assumptions, the use of a fine range resolutions will lead to distributions with slightly higher peaks and sharper appearance. On the contrary, the use of coarser range resolutions will lead to slightly lower peaks and larger spread.

B. Radar measurement simulation

For simplicity, each target moves in the environment following a constant velocity and linear trajectory. To reduce computational cost, we avoid using a densely populated simulation grid to generate target representations. Instead, we generate discrete detection points located at the centers of grid cells that overlap with the targets. If a target spans N cells, N detection points are created, each positioned at the center of the corresponding cell, preserving the original range and azimuth information. In essence, the simulator produces only as many points as necessary to cover the targets, resulting in significantly faster runtime. Fig. 1 gives an example of this approach where 38 detection points are generated.

It should also be noted that the cell areas corresponding to each detection point are computed and stored to generate the reflectivity values at each point, as explained in Section III-A. By convention, the maximum reflectivity value is attributed to the point representing the center of the target.

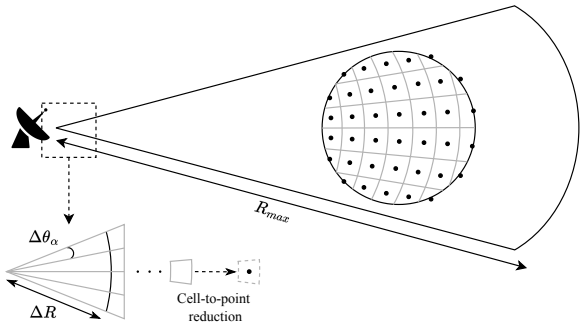

Fig. 1. Schematic representation of the reduction of cell to point representation of extended targets, to ensure faster simulation runtimes. Elements such as the azimuth and range resolutions are also presented in the figure. Points generated at the edges of the circular target are also taken into account.

In the simplified target simulation environment described above, the radar's measurements consist of clusters of reflectivity 'detection' points, corresponding to the various targets in the scene. Using these points, we can calculate the center-of-mass of each target as mentioned in Section II-B. Each COM is then tracked by the radar using an Extended Kalman

Filter implemented in the *StoneSoup* [10] framework, an open-source tracking and state estimation framework designed for data, sensor fusion, and multi-object tracking applications. To relate our simulated measurements to *StoneSoup* terms in order to ensure proper functionality, each target reflectivity point represents a detection for the framework, an idea that is also applied to each COM for tracking purposes. The radar is defined within such framework as a `FixedRadarSensor`, with a pre-defined update time of 60 s corresponding to a whole PPI scan, and can initiate, update and delete tracks. All radar's parameters are adjustable on runtime, however in this work only the bandwidth will be modified.

C. RL Environment description

Vital to every RL application is the definition of the environment the agent has to interact with in order to be trained. The environment defines the observation and action spaces, as well as the reward structures and system, providing a controlled setting for training and evaluating RL algorithms. The reward definition is also crucial to shape policies of RL agents, as it guides their behavior by providing feedback on actions taken. In this sub-section, we describe the definition of the environment created for the training & testing of cognitive radar agents, as well as the approach chosen for their training.

Since in our case the actions taken by the radar agent does not change the environment but only the way the radar perceives it, Areas of Interest (AoI) are introduced in order to make the learning process more challenging and interesting. An AoI can be used to describe anything of interest to a user, such as catchments, settlements, or other assets such as farms or greenhouses, all of which would benefit from better rainfall measurements and warnings about incoming storms.

Carefully chosen weights as a function of the distance to the AoI are introduced to increase the reward corresponding to capturing targets closer to the AoI. Theoretically, this should encourage policies that focus more on targets closer to these regions. Of course, this will not happen in every episode, raising the question of how the cognitive weather radar will deal with cases in which multiple interesting targets need to be monitored simultaneously. The reward for our agent's interactions with its environment is formulated based on its objective, which is to monitor targets in relation to an AoI. The proposed reward does not take into account whether the radar covers the entirety of the targets or not, which can be in some cases impossible due to the maximum range capabilities of the radar and the target size. Instead, even if parts of the targets are measured, the agent is still rewarded for its actions. Hence, the reward $R(t)$ for capturing N' targets at time t is:

$$R(t) = \begin{cases} \sum_{i=0}^{N'-1} T_i(t) \cdot (1 - F_B(i,t)), & \text{if } N' > 0 \\ -10, & \text{if no target is captured} \end{cases} \quad (7)$$

where $T_i(t)$ is a threat indicator calculated based on the measured reflectivity values, weighted relatively to its distance to the AoI. More specifically, when the radar observes multiple targets, the threat indicator of each target is computed as the ratio of its measured maximum linear reflectivity $\hat{z}_i(t)$ over the maximum measured linear reflectivity of the environment, and hence $T_i(t) = \frac{max(\hat{z}_i(t))}{max_i(\hat{z}_i(t))} \cdot d_i(t)$. The second term of $T_i(t)$ denoted by $d_i(t)$ is defined as a factor of the distance of the target i-th from the AoI. The penalty was empirically selected as to greatly punish the agent when it would miss all targets.

In equation (7), $(1 - F_B(i,t))$ is a factor accounting for bandwidth utilization for the i-th target at timestep t. $F_B(i,t)$ is calculated as a metric of distance of the utilized bandwidth value and the optimal (continuous) bandwidth value for the said target. Specifically, the further away from the optimal value the utilized bandwidth is, the more the factor $(1 - F_B(i,t))$ approaches zero and vice versa.

The design of this reward function can be characterized as myopic, as it rewards the agent based on the targets it can observe, rather than on the total reflectivity existing in its possible maximum observable range. The reason why this approach can still be valid is the assumption that there is enough coverage of the area of operation from other, non adaptive radars, and that targets located far away from the AoI, and missed by the radar agent due to its coverage strategy, can still be covered by other systems. Hence, in a stand-alone system, additional care shall be given in order for the radar agent to explore its environment for more targets in a periodical manner. This would require the definition of different rewards and possible exploration patterns.

D. Training of the radar agent

The training of the radar agent was performed using the Proximal Policy Optimization (PPO) [11] algorithm, as implemented in the *Stable Baselines 3* toolbox [12]. PPO is a state-of-the-art (RL) algorithm widely used for training autonomous agents in complex decision-making environments. Unlike traditional policy gradient approaches, PPO incorporates a clipped surrogate objective that constrains policy updates, preventing overly large adjustments that could lead to instability. Additionally, it uses an actor-critic framework, updating the policy from collected trajectories via a clipped objective to ensure stable and consistent learning.

Regarding the environment and agent, Open AI's *gymnasium* [13] was used, which is a standardized toolkit for developing, benchmarking, and comparing reinforcement learning algorithms through a diverse set of pre-built or custom simulation environments. Using the simulator proposed in Sections III-A and III-B, and the principles described in Section III-C, we created an appropriate custom *gymnasium* environment for training radar agent models. As for the scenarios on which the agent is trained, two circular storm targets of 2 km diameter are randomly generated per episode within the maximum possible observable range of the radar with uniformly random velocities for each direction spanning from 1 to 6 m/s, and random maximum reflectivity values between 20 and 45 dBZ.

IV. Results

We employ PPO to train the radar agent using the action space described in II-C. The radar agent was trained for 2 million steps, with learning rate of $3 \cdot 10^{-4}$, 2048 steps to run per update, $\gamma = 0.99$, entropy coefficient of 0 and value function coefficient of 0.5. The custom environment can either run for 300 simulation steps, or as long as the targets are present within the maximum possible observable range.

Presenting and comparing mean reward per episode is usually adequate for proving the advantages of RL methods. Yet, in this application, it can be seen from the reward function itself that the reward of a well trained agent will be surpassed by the simple 'sit-and-spin' mode only in specific cases where the non adaptability of the aforementioned approach generates higher rewards. Hence, simply comparing to the 'sit-and-spin' mode does not necessarily do justice to well trained and performing agents. Such cases which will be analyzed in this Section. To prove the efficacy of the policies learned by the radar agents, we designed four challenging custom test cases:

- **Case 1:** Target 1 is generated at (-30000, 10000) m, moving towards the AoI with initial velocity of (4,0) m/s and with a maximum reflectivity of 40 dBZ. Target 2 is generated at (-20000, -25000) m, with an initial velocity of (3, 0) m/s and with 35 dBZ as maximum reflectivity.
- **Case 2:** Same initial conditions as in case 1 except that the maximum reflectivities of both targets is 40 dBZ.
- **Case 3:** Target 1 is generated using the same initial conditions. Target 2 is now generated at (-20000, 20000) m with initial velocity of (4, 0) m/s. This case was designed to understand the radar's behavior for a case in which target 2 is leading quite substantially the other.
- **Case 4:** In this final case, the two targets pass through the AoI very close to each other. The differences with case 2 is that Target 1 is generated at (-30000, 8500) m with a lower max reflectivity of 35 dBZ.

The aforementioned cases where run and analyzed, producing the results of the plots in Fig. 2 and 3. The blue line in Fig. 2 represents the maximum range of the radar, which is inversely proportional to the used bandwidth. It should first be noted that the radar agent never utilized the full available bandwidth range. This can be traced in the design of the test cases. Since in none of them the target passes right above the radar, there is no need for the agent to use the highest available bandwidth.

Additionally, when examining the agent's action space, we can see that the +5 MHz bandwidth increase occurred only once, during the fourth test case. We explain this behavior by the targets' initial conditions and velocities with respect to the AoI location, which influenced the bandwidth selection and consequently, the maximum range needed during each scan.

An interesting comparison can be drawn between cases 1 and 2. Even though the kinematics of both targets are the same, the difference in reflectivities led to vast differences in terms of the actions taken by the agent. Fig. 2, shows that in the first case, the radar decided to stop monitoring Target 2 to focus on Target 1 which was approaching the AoI. The agent increased the used bandwidth sharply, in order to quickly cover the target adequately. Apart from the distance of the targets from the AoI, the measured maximum reflectivity also appears to be an important factor for decision making. In the first case, the radar decided to focus on Target 1 since this lead to a larger reward than monitoring both. This is certainly not the case in the second test case. There, both targets have identical maximum reflectivities. Even though this selection appears inefficient, or perhaps suboptimal since naturally one would expect the same behavior as in the first test case, this sort of behavior is consistent with the reward function. With this, having very similar reflectivity measurements will make the first factor of $T_i(t)$ in Section III-C become close to 1 for both targets, making $d_i(t)$ and $F_B(t)$ the main contributors.

The radar agent is forced to make important decisions in the third test case as well, where it decides around the 150th step to stop monitoring Target 2 and pursue only the first. There, the second target leads by several kilometers and after it has passed the AoI, the radar affords to change its focus to pursue a good performance regarding Target 2 measurement. It is important to observe the impact of this choice on the received reward as seen in Fig. 3, which might reveal important elements of the learned policy. What can be noticed is that the radar agent selects a strategy that is seemingly sub-optimal reward wise. However, it is important to consider also that Target 1 becomes increasingly important the closer it gets to the AoI, whereas at the same time Target 2 is already 10 km away from it. On the contrary, in the 4th test case where the targets are generated and stay close to each other throughout the experiment due to their identical velocities, the radar chooses to continuously monitor both targets, until they move further away from its maximum observable range. Here, as seen in Fig. 3, the adaptive approach selects actions that produce better rewards than the non-adaptive one for almost all episode steps, until it deteriorates to match its performance due to both targets having moved far away from the radar.

V. Discussion

The presented results prove the feasibility of using RL methods for bandwidth utilization in a cognitive weather radar context. The proposed reward led to agents that can outperform classic 'sit-and-spin' approaches for as much as 66.7% (percentage of increase of mean reward), as shown by the four test cases in Tab. I. However, it is important to keep in mind that, even if the radar agent showed consistent behavior, the designed rewards should match the corresponding user's needs. Typically, several questions may be raised on the explainability of the learned policies, or lack thereof. To address this, future work could explore integrating interpretable models or utilizing more appropriate visualization techniques to highlight decision-making processes which can help clarify how the agent evaluates states and makes choices, thus improving trust and understanding of the model's behavior. Furthermore, as mentioned above, the current design of the reward can be seen as myopic, since the radar receives rewards based on only what it can observe. Beyond this, multiple radar agents could

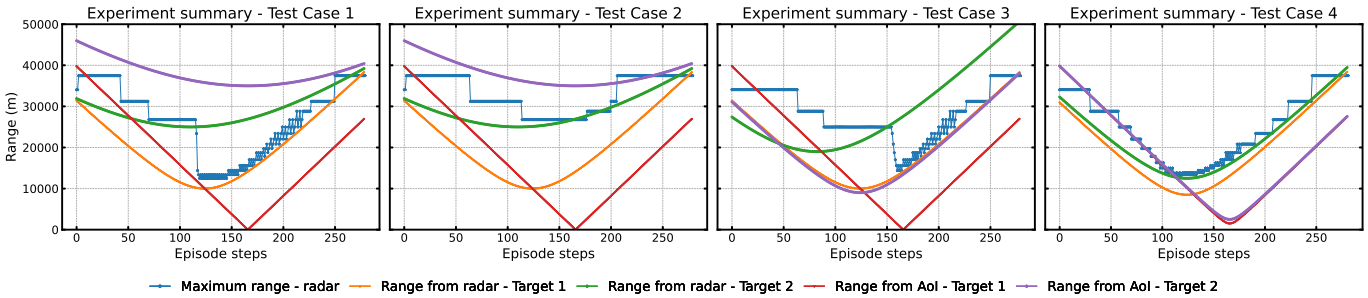

Fig. 2. Experiment summaries for each of the 4 test cases. Each plot contains five elements over simulation steps: the maximum range measurable by the radar (related to used bandwidth), the ranges of each target from the radar and the ranges of each target from the AoI.

Fig. 3. Reward generated by the actions taken by the adaptive radar agent compared to the 'sit-and-spin' approach for each of the test cases.

operate at the same time, sharing useful information regarding target locations and taking more informed actions.

TABLE I
COMPARISON TABLE OF THE MEAN REWARD OF EACH TEST CASE. THE BEST OF EACH CASE IS SHOWN IN BOLD.

Test Case	**Case 1**	**Case 2**	**Case 3**	**Case 4**
Mean Reward (Adaptive)	**0.60**	**0.79**	**0.78**	**0.65**
Mean Reward (Non-Adaptive)	0.44	0.67	0.65	0.39

VI. CONCLUSION

This paper demonstrated the feasibility of RL-driven bandwidth adaptation for cognitive weather radars. It was shown that through adaptive tuning of bandwidth based on target location and distance from a defined AoI, the trained radar agent managed to outperform conventional methods. By dynamically optimizing system behavior, the proposed radar agent enhances responsiveness and efficiency, offering a more effective approach for situationally-aware tracking and monitoring. Although the rewarding system appears to be adequate for this initial feasibility, it leaves room for improvement especially in explainability terms. Moreover, its expansion to the case of multiple cooperating radars and/or AoI in more complex scenarios needs to be studied.

REFERENCES

[1] A. F. Martone, K. D. Sherbondy, J. A. Kovarskiy, B. H. Kirk, R. M. Narayanan, C. E. Thornton, R. M. Buehrer, J. W. Owen, B. Ravenscroft, S. Blunt *et al.*, "Closing the loop on cognitive radar for spectrum sharing," *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 9, pp. 44–55, 2021.

[2] A. Aubry, V. Carotenuto, A. De Maio, and M. A. Govoni, "Multi-snapshot spectrum sensing for cognitive radar via block-sparsity exploitation," *IEEE Transactions on Signal Processing*, vol. 67, no. 6, pp. 1396–1406, 2018.

[3] R. K. McCargar and G. E. Smith, "A reconfigurable resource manager for distributed networked radar," in *2021 IEEE Radar Conference*. IEEE, 2021, pp. 1–6.

[4] A. Charlish, F. Hoffmann, C. Degen, and I. Schlangen, "The development from adaptive to cognitive radar resource management," *IEEE Aerospace and Electronic Systems Magazine*, vol. 35, no. 6.

[5] V. Manfredi, S. Mahadevan, and J. Kurose, "Switching kalman filters for prediction and tracking in an adaptive meteorological sensing network," in *2005 Second Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2005. IEEE SECON 2005.* IEEE, 2005, pp. 197–206.

[6] V. Manfredi and J. Kurose, "Scan strategies for meteorological radars," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[7] F. Smits, A. Huizing, W. van Rossum, and P. Hiemstra, "A cognitive radar network: Architecture and application to multiplatform radar management," in *2008 European Radar Conference*, pp. 312–315.

[8] A. Charlish and F. Hoffmann, "Anticipation in cognitive radar using stochastic control," in *2015 IEEE radar conference (RadarCon)*. IEEE, 2015, pp. 1692–1697.

[9] K. Granström and M. Baum, "A tutorial on multiple extended object tracking," *Authorea Preprints*, 2023.

[10] P. A. Thomas, J. Barr, B. Balaji, and K. White, "An open source framework for tracking and state estimation ('stone soup')," in *Signal Processing, Sensor/Information Fusion, and Target Recognition XXVI*, vol. 10200. SPIE, 2017, pp. 62–71.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[12] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of machine learning research*, vol. 22, no. 268.

[13] OpenAI and F. Foundation, "Gymnasium," 2023. [Online]. Available: https://github.com/Farama-Foundation/Gymnasium