Universiteit Leiden

Ecosystem Simulation for Environmental Policies using Reinforcement Learning

Alejandro Lopez Ruiz (s4075315)

Supervisors:
Aske Plaat & Thomas Moerland

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 12, 2026

Abstract

This thesis investigates whether tabular Q-learning can serve as a useful modeling approach for predator-prey ecosystem simulation in the context of environmental policy design, and compares an RL-based multi-agent simulator against classical Lotka-Volterra and Rosenzweig-MacArthur ordinary differential equation baselines. An alternating training protocol is used to produce learned predator and prey policies, which are evaluated across 50 random seeds and 2000-step simulations. The results reveal a structural trade-off: policies that produce ecologically plausible distributed predation behavior consistently fail to maintain population-level coexistence, while the policy achieving the strongest aggregate coexistence metrics does so through an MDP exploitation strategy driven by the geometry of the grid boundaries rather than by ecological balance. Fitting ODE models directly to the RL population trajectories shows that this best-performing checkpoint operates in a qualitatively different dynamical regime: the fitted predator equilibrium is approximately 120 times higher than the value predicted by the RL-calibrated baseline under uniform-mixing assumptions, and the oscillations visible in the RL trajectories are stochastic demographic noise rather than deterministic ecological cycles. The overall finding is that tabular Q-learning can approximate some classical predator-prey signatures at the aggregate level, but cannot combine ecological realism, robustness, and stable coexistence within a single learned solution. Q-learning in the presented setup is therefore not yet sufficient for policy-oriented ecological modeling, but provides a complementary diagnostic framework for exposing structural failure modes that aggregate models cannot represent.

Contents

1 Introduction

1.1 Context and motivation

Environmental policy creation increasingly depends on decision-support tools that can organize ecological knowledge, structure uncertainty, and assist experts in evaluating alternative courses of action. In this context, ecological decision support systems (DSSs) have been used to support environmental management by combining scientific models and data to help expert judgment guide planning and intervention [10]. Traditionally, these systems have assisted with tasks such as resource management, conservation planning, environmental monitoring, and impact assessment; they provided a structured way to connect ecological understanding with a more practical decision-making process [10, 35]. Their value lies not in replacing human expertise, but in improving the consistency, transparency, and scalability of environmental decisions in settings where many interacting variables must be considered.

However, the need for stronger and more flexible DSSs has grown as environmental problems have become increasingly complex. Contemporary ecological challenges are rarely isolated or static; instead, they emerge from tightly coupled human-environment systems shaped by habitat loss, biodiversity decline, land-use change, climate change, and long-term human pressure. As a result, environmental policy must often be designed under uncertainty, limited observability, and even changing ecological dynamics. These are conditions that make simple or rigid support tools progressively less adequate [35, 13, 7]. In other words, policy-makers need systems that perform better at capturing dynamic interactions, anticipate indirect effects, and remain useful even when ecological relationships are nonlinear, context-dependent, or only partially understood [13, 11, 4].

This challenge has motivated increasing interest in the use of artificial intelligence within ecological decision support. AI methods can process large and heterogeneous datasets, identify hidden patterns, and support adaptive decision-making in complex environments [28]. In biodiversity conservation and environmental management, AI has already been proposed to improve prioritization, monitoring, and long-term planning, especially in problems where uncertainty and scale make purely manual analysis difficult [28, 19]. Among AI methods, **reinforcement learning** is especially relevant because it is explicitly designed for sequential decision-making under uncertainty. Rather than only fitting patterns from static data, reinforcement learning allows an agent to learn policies through interaction with an environment, making it a particularly promising framework for settings in which decisions influence future ecological states over time [19, 6]. For this reason, reinforcement learning is increasingly being considered as a useful foundation for next-generation environmental DSSs, especially in domains where adaptation, feedback, and long-term consequences are central to the problem itself [23, 24, 28].

1.2 Problem statement

Ecological systems are inherently difficult to model; their behavior emerges from many interacting processes that are often nonlinear, context-dependent, and variable over time. Population growth, species interactions, resource competition, and environmental feedback, among others, do not usually evolve in simple proportional ways. Instead, small changes in one part of the system can generate disproportionately large or delayed effects elsewhere [13, 7, 11, 33].

Traditional ecological modeling approaches have consistently provided an essential foundation for

understanding these systems, mainly through equation-based methods such as ordinary differential equation models, logistic growth models, and predator-prey systems. These approaches are valuable because they are mathematically interpretable, analytically tractable, and often useful for capturing broad population-level tendencies. However, they rely heavily on assumptions that, in real ecosystems, are often only partially satisfied, which limits the extent to which classical models can fully represent complex ecological behavior [13, 8, 4].

These limitations become particularly important when those ecological models are used to support policy design. Environmental policy decisions are rarely made in stable and fully observable contexts; therefore, their usefulness as a decision-support tool is reduced. In such cases, policy recommendations may become overly sensitive to hidden assumptions, may fail to anticipate indirect or delayed consequences, or may inadequately reflect the adaptive nature of ecological systems themselves [13, 11, 4].

This motivates the exploration of more flexible modeling approaches that can better accommodate stochasticity, partial observability, and adaptive ecological dynamics. In particular, models that learn from interactions with the environment rather than depending entirely on pre-specified analytical structure may offer a useful complement to classical ecological methods [33, 9, 7].

1.3 Research objective

The objective of this thesis is to investigate whether reinforcement learning can serve as a useful modeling approach for ecosystem simulation in the context of environmental policy design. More specifically, the thesis aims to evaluate whether an RL-based predator-prey simulation can complement or improve upon classical ecological models. Traditional ecological approaches, particularly equation-based models such as ordinary differential equation (ODE) systems, remain highly valuable due to interpretability and analytical clarity, but struggle to capture the full complexity of real ecological systems [13, 7]. Reinforcement learning, by contrast, offers a framework in which agents learn through interactions with the environment, making it especially relevant for systems in which behavior, feedback, and environmental variability are central components [6, 24].

Accordingly, this thesis focuses on the development and evaluation of an RL-based ecosystem simulator centered on predator-prey dynamics. The study is designed around a comparison between two modeling paradigms: classical ecological baseline models based on ODEs, and a grid-based RL environment in which predator and prey agents learn and interact under partially observed conditions. Through this comparison, the thesis seeks to determine not only whether reinforcement learning can reproduce meaningful ecological dynamics, but also whether it may provide a stronger basis for policy-oriented ecological simulation [28, 19, 24].

In this sense, this thesis's contribution is dual. First, it develops a computational framework for studying predator-prey ecosystems through reinforcement learning. Second, it evaluates that framework against established ecological modeling approaches to assess its scientific and practical relevance for environmental decision support. Rather than treating reinforcement learning as a replacement for classical ecological theory, the thesis explores whether it can act as a flexible complementary tool for simulating ecological systems whose dynamics are difficult to specify fully in advance [10, 28, 24].

1.4 Research question and sub-questions

Based on the motivation and objective of this thesis, the central research question is:

How effective are reinforcement learning methods, compared to classical ecological models, in supporting ecosystem policy design?

This question reflects the main purpose of the study: to evaluate whether reinforcement learning can provide a useful and scientifically meaningful alternative or complement to equation-based ecological modeling in a policy-relevant context. The comparison is not only concerned with predictive performance, but also with the ability of the models to represent adaptive ecological behavior, remain robust under environmental variability, and generate insights that are relevant for conservation-oriented decision-making [28, 13, 24].

To operationalize this broader question, the thesis addresses the following sub-questions:

1. **How accurately can Q-learning agents reproduce population dynamics in a predator–prey ecosystem compared to classical models?**
2. **How robust are the RL-derived population policies under noisy or perturbed environmental conditions, and how do they align with conservation objectives such as species survival?**

Together, these sub-questions structure the empirical part of the thesis. The first focuses on whether RL-based simulation can replicate meaningful ecological dynamics relative to established baseline models, while the second examines whether the learned behaviors remain useful when the system is subjected to uncertainty or perturbation. This distinction is important because, in ecological policy settings, a model must not only produce plausible dynamics under idealized conditions, but also remain informative when the environment becomes more variable and less predictable.

1.5 Scope of the thesis

This thesis examines the use of reinforcement learning for ecosystem simulation in a policy-relevant ecological setting. The work includes the design of a grid-based simulation environment, the implementation of Q-learning agents, and a comparison between this RL-based approach and classical ecological models based on ODEs. Within this scope, the thesis evaluates whether RL can reproduce meaningful ecological dynamics and provide useful insights for decision-support [28, 13].

At the same time, the thesis does not aim to produce a fully realistic ecological model of natural systems, nor does it aim to deliver direct policy recommendations for specific environmental cases. The simulator is intended as a research-oriented approximation rather than as a complete ecological forecasting or management tool. Likewise, the work does not attempt to cover all possible ecological relationships, species types, or environmental processes, but instead restricts itself to a predator-prey setting to keep the study analytically and computationally manageable. Although the broader reinforcement learning literature includes more advanced multi-agent methods, this thesis is primarily centered on tabular Q-learning as a first comparative framework.

These boundaries are intentional and appropriate to the aims of the study. As a bachelor's thesis, the project is designed as a focused comparative investigation rather than a complete ecological

decision-support platform. Its purpose is to test whether reinforcement learning can serve as a meaningful complementary modeling approach in a controlled ecosystem setting, and to evaluate this potential against established classical baselines. By limiting the study to a clearly defined interaction structure, selected benchmark systems, and a manageable RL implementation, the thesis aims to maintain both methodological clarity and practical feasibility while still contributing to the broader discussion on AI-supported ecological modeling and environmental policy design [10, 28].

1.6 Thesis structure

Section 2 presents the theoretical background and related work relevant to the study. It introduces classical ecological modeling approaches, including predator-prey and population-dynamics models, discusses their main limitations in complex ecological settings, and reviews the foundations of reinforcement learning and its use in ecological and environmental applications. **Section 3** describes the methodology of the thesis, including the overall comparative framework, the classical ODE baselines, the reinforcement learning setup, the case-study orientation, and the evaluation criteria used in the analysis.

Section 4 describes the experimental setup, including the simulation settings, training configurations, benchmark scenarios, and comparison procedures. **Section 5** reports the results of the experiment, focusing on the behavior of the RL-based ecosystem simulation, its comparison with classical ecological models, and its robustness under different conditions. Finally, **Section 6** discusses the implications of the findings, the main limitations of the study, and the extent to which reinforcement learning can contribute to ecological modeling and policy-oriented simulation. **Section 7** concludes the thesis by summarizing the main results, answering the research question, and outlining directions for future work.

2 Background and Related Work

2.1 Classical ecological modeling

2.1.1 Population dynamics

Population ecology is concerned with understanding how and why the size of biological populations changes over time. At its core, the field studies processes such as birth, death, competition, and resource limitation, and seeks to describe how these processes shape the growth, decline, or stabilization of populations. The development of population ecology was closely tied to the rise of theoretical and mathematical ecology, in which ecological systems began to be represented through formal quantitative models rather than only descriptive biological observation [18].

One of the foundational ideas in this development was that population growth cannot continue indefinitely. Early simple models of growth often assumed that the rate of increase of a population was proportional to its current size, which yields exponential growth. In its simplest continuous form, this can be written as:

$$\frac{dN}{dt} = rN \tag{1}$$

where N denotes population size and r is the intrinsic growth rate. This equation describes unrestricted growth, meaning that the larger the population becomes, the more rapidly it grows. Although useful at first approximation, such a model is ecologically unrealistic in most natural systems.

The introduction of natural constraints led to one of the most influential ideas in population ecology: the logistic growth model. Rather than assuming unlimited exponential increase, the logistic model incorporates the idea that growth slows as the population approaches an environmental limit. This limit is known as the carrying capacity, commonly denoted by K, and represents the maximum population size that the environment can sustain under given conditions [17, 5]. The logistic model is typically written as:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \tag{2}$$

This equation introduces density dependence into ecological modeling. When N is small relative to K, the term $1 - \frac{N}{K}$ is close to 1, and the population grows approximately exponentially. As N approaches K, the term becomes smaller, slowing growth until the net rate of change tends towards zero. In this sense, the logistic model captures the idea that population dynamics depend not only on intrinsic reproductive potential, but also on environmental limits and feedback from population density itself [17, 5].

As population ecology developed further, these basic models became the starting point for more elaborate frameworks. Researchers extended growth equations to account for more realistic ecological settings, including delayed feedback, migration, structured populations, and interactions between multiple species. In this sense, the logistic model was not the endpoint of classical ecological modeling, but rather the beginning of a broader tradition in which ecological processes were expressed through differential equations and analyzed as dynamical systems [18, 7]. Modern studies continue to show that density-dependent growth remains a useful conceptual and mathematical tool, even though real ecological systems often depart from the simplest assumptions of constant carrying capacity or homogeneous population structure [5].

2.1.2 Predator-prey models

While population dynamics models describe how a single population changes over time, many ecological systems cannot be understood by studying species in isolation. In nature, populations are embedded in networks of interaction in which the growth or decline of one species depends on the presence of others. Among these interactions, the predator-prey relationship is one of the most central and widely studied in ecology. Predator-prey models aim to capture how the consumption of prey by predators influences the dynamics of both populations, often generating oscillatory or cyclic behavior over time [29, 18].

A useful way to represent these models is to extend the single-population growth framework into an interacting two-population system. Let N denote the prey population and P the predator. A general prey-predator interaction can be expressed as:

$$\frac{dN}{dt} = f(N) - g(N, P) \tag{3}$$

$$\frac{dP}{dt} = h(N, P) - m(P) \tag{4}$$

Here, $f(N)$ represents prey growth in the absence of predation, $g(N, P)$ represents prey loss due to predator interaction, $h(N, P)$ is predator growth supported by prey consumption, and $m(P)$ represents predator mortality. This formulation is deliberately general, but it shows the key conceptual shift from single-species population ecology to interaction-based: population change is no longer determined only by internal reproduction and resource limitation, but also by cross-species dependence.

Predator-prey models became important not only because they explained ecological cycles, but also because they provided a mathematically tractable way to study regulation, instability, coexistence, and extinction. In many cases, they serve as a minimal framework for analyzing how local interactions can produce system-level ecological behavior [29]. Over time, the basic predator-prey framework has been extended to include more realistic assumptions, such as resource limits, migration, nonlinear response functions, stochasticity, and even role reversal or context-dependent interaction structure, all of which reflect the fact that real ecological systems often deviate from the simplest assumptions of early models [15, 29].

2.1.3 Lotka–Volterra and related ODE approaches

The most influential classical formulation of predator-prey dynamics is the **Lotka-Volterra model**, which became one of the foundation systems of theoretical ecology. Building on the broader development of population ecology, Lotka-Volterra equations provided a compact mathematical description of how two interacting populations can influence one another over time. Their importance lies not only in their historical role, but also in the fact that they established the basic equation-based framework through which predator-prey interactions have often been analyzed ever since [18, 29].

In its standard form, the Lotka-Volterra predator-prey system assumes that the prey population grows exponentially in the absence of predators, while the predator population declines exponentially in the absence of prey. The interaction between the two populations is then modeled through a bilinear term representing encounters between predator and prey. The classical equations are:

$$\frac{dN}{dt} = \alpha N - \beta NP \tag{5}$$

$$\frac{dP}{dt} = \delta\beta NP - \gamma P \tag{6}$$

where N denotes prey population, P predator population, α the prey intrinsic growth rate, β the predation rate coefficient, δ the predator conversion efficiency (the fraction of consumed prey biomass converted into predator growth), and γ the predator mortality rate. Together, these equations describe a simple feedback structure: prey increase supports predator growth, while predator pressure reduces prey abundance.

One of the main reasons for the enduring influence of the Lotka-Volterra model is that it mathematically captures cyclical predator-prey dynamics. It provides a clear baseline for understanding how interaction alone can generate oscillatory behavior, and it helped establish the idea that

ecological systems can be studied as dynamical systems governed by coupled differential equations [18, 29].

For this reason, many related ODE approaches have extended the Lotka-Volterra framework to incorporate more realistic ecological behavior. One common modification is to replace unrestricted prey growth with logistic growth, thereby introducing carrying capacity into the predator-prey system:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) - \beta NP \tag{7}$$

$$\frac{dP}{dt} = \delta NP - \gamma P \tag{8}$$

This extension is important because it connects predator-prey dynamics to the density-dependent reasoning introduced earlier in population ecology. By including the carrying capacity K, the prey population is no longer assumed to grow without bound, which makes the model more ecologically plausible in many settings [5, 29]. Other extensions have further modified the framework to include migration, alternative functional responses, heterogeneous interaction structure, or more complex stability conditions, reflecting the fact that classical ecological modeling has continuously evolved beyond the simplest two-equation system [32, 1].

These related ODE approaches remain highly relevant because they provide interpretable and widely accepted ecological baselines. They allow researchers to study equilibrium behavior, oscillations, and stability in a transparent mathematical framework, and they continue to serve as a reference point for more complex ecological models.

2.2 Limitations of classical ecological models

Classical ecological models have played a central role in the development of theoretical ecology as they provide interpretable and mathematically structured representations of population and ecosystem dynamics. By reducing ecological systems to a manageable set of variables and relationships, they make it possible to analyze growth, interaction, equilibrium, and stability. However, this same abstraction also creates limitations. Ecological systems are not only complex but also often shaped by feedback mechanisms, uncertainty, and context-specific interactions that are difficult to capture fully within simplified equation-based frameworks. As a result, classical models are highly valuable as conceptual and analytical tools, but their descriptive and predictive power may weaken when they are applied to real systems whose behavior departs from the assumptions on which these models are built [13, 8, 7].

One major limitation is that ecological interactions are frequently **nonlinear**. In real ecosystems, the effect of one variable on another is rarely constant across all conditions. Species interactions can also vary over time, meaning that ecological relationships are often dynamic rather than stationary. This poses a challenge for classical models that rely on fixed functional forms and stable parameter assumptions. Although such formulations are mathematically convenient, they may fail to capture the changing structure of ecological interactions observed in practice [11, 33, 9].

A second limitation concerns **stochasticity and uncertainty**. Many classical ecological models are deterministic, meaning that the same initial conditions always produce the same trajectory. Natural ecosystems, however, are influenced by random fluctuations in weather, resource distribution,

disease, migration, and demographic events. These stochastic components can substantially alter population outcomes, especially over long time horizons or in systems with sensitive interaction structures. This is particularly problematic when the model is used to support ecological inference or policy-relevant decision-making, since environmental decisions often must be made precisely under uncertain and changing conditions [13, 33, 11].

Classical ecological models also struggle with hidden assumptions, heterogeneity, and context dependence. Many equation-based formulations implicitly assume homogeneous populations, simplified interaction mechanisms, or fixed causal structure. What appears to be a general ecological law in one setting may fail in another because the underlying structure of the system has changed. Banitz et al. argue that model-derived explanations are often constrained by assumptions that remain hidden beneath the formal structure of the model, which limits how confidently such explanations can be generalized beyond the conditions under which the model was developed [4].

These issues become even more significant when ecological models are considered as tools for environmental policy support. Policy design requires models that remain informative under incomplete information, uncertainty, and perturbation, and that can help decision-makers reason about long-term consequences rather than only idealized system behavior. If the model cannot represent shifting dynamics, stochastic influences, or adaptive ecological responses, then its usefulness for decision support is reduced. This does not mean that classical models should be discarded; rather, it means that their strength in interpretability and analytical clarity must be weighed against the fact that they may not adequately represent all the features that matter in policy-relevant ecological systems [13, 4, 11].

For this reason, the limitations of classical ecological models motivate exploring better-suited complementary approaches to dynamic, uncertain, and partially observable environments. In particular, there is increasing interest in modeling frameworks that can adapt to feedback from the system itself rather than relying entirely on prespecified equations and fixed interaction structures. This is one of the main motivations for considering reinforcement learning in ecological simulation: not because classical ecological models lack value, but because more adaptive approaches may represent certain ecological features more flexibly in settings where complexity and uncertainty are central to the problem.

2.3 Reinforcement learning foundations

2.3.1 Markov Decision Processes

A large part of reinforcement learning is built on the formalism of the Markov Decision Process (MDP), which provides a mathematical framework for sequential decision-making under uncertainty. In contrast to static prediction problems, where a model maps inputs directly to outputs, reinforcement learning considers settings in which an agent repeatedly interacts with an environment over time. At each step, the agent observes the current situation, selects an action, receives a reward, and then transitions to a new situation. The objective is not simply to maximize immediate reward, but to learn a sequence of decisions that yield high cumulative reward over time. This makes the MDP framework particularly suitable for problems in which actions have delayed consequences and the quality of a decision depends on how it affects future system states.

Formally, an MDP is typically defined as a tuple.

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \tag{9}$$

where $\mathcal{S}$ is the set of possible states, $\mathcal{A}$ is the set of possible actions, P is the transition function, R is the reward function, and $\gamma \in [0, 1]$ is the discount factor. The state space $\mathcal{S}$ represents all situations the agent may encounter, while the action space $\mathcal{A}$ contains all actions available to the agent. The transition function specifies how the environment evolves after an action is taken, and the reward function quantifies the immediate utility of a state-action transition. The discount factor determines how much future rewards are valued relative to immediate rewards.

More precisely, the transition dynamics are defined through conditional probabilities of the form:

$$P(s' \mid s, a) = \Pr(S_{t+1} = s' \mid S_t = s, A_t = a) \tag{10}$$

which describe the probability of moving to state s' after taking action a in state s. This probabilistic formulation is essential because many environments are stochastic: the same action in the same nominal state does not always produce the same outcome. Similarly, the reward function may be written as

$$R(s, a, s') = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s'] \tag{11}$$

so that rewards depend on the state, the chosen action, and optionally the resulting next state. Together, these elements define the dynamics of the decision problem.

A policy specifies the behavior of the agent. In general, a policy is a mapping from states to actions, or more generally, to a probability distribution over actions. A deterministic policy may be written as

$$\pi(s) = a \tag{12}$$

while a stochastic policy is written as

$$\pi(a \mid s) = \Pr(A_t = a \mid S_t = s) \tag{13}$$

Reinforcement learning aims to find a policy that maximizes the expected cumulative reward. This cumulative reward is usually expressed as the **return**

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{14}$$

which is the discounted sum of the future rewards from time step t onward. Discounting serves two purposes: it makes the infinite-horizon sum finite when rewards continue indefinitely, and it encodes a preference for rewards obtained sooner rather than later.

The defining assumption of an MDP is the Markov property, which states that the future depends on the present state and action, but not on the full history beyond what is already encoded in the present state. Formally, this means

$$\Pr(S_{t+1}, R_{t+1} \mid S_t, A_t, S_{t-1}, A_{t-1}, \ldots, S_0, A_0) = \Pr(S_{t+1}, R_{t+1} \mid S_t, A_t) \tag{15}$$

This assumption is central because it allows sequential decision problems to be represented compactly: if the current state contains all relevant information, then an optimal decision can be chosen based on that state alone. In practice, however, many real systems are only approximately Markovian. The quality of the MDP formulation, therefore, depends on whether the chosen state representation captures enough information about the environment to make future outcomes predictable in this restricted sense.

In reinforcement learning, several important quantities are defined relative to a policy π. The state-value function gives the expected return from a state under that policy:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t \mid S_t = s] \tag{16}$$

Similarly, the action-value function gives the expected return from taking action a in state s and then following policy π:

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}[G_t \mid S_t = s, A_t = a] \tag{17}$$

These functions are fundamental because they provide a quantitative basis for comparing different actions and policies.

2.3.2 Q-learning

Among the classical reinforcement learning algorithms, Q-learning is one of the most important and widely used model-free methods. Its significance lies in the fact that it allows an agent to learn an effective decision policy directly from experience, without requiring prior knowledge of the transition probabilities or reward dynamics of the environment. Instead of solving the MDP analytically from a complete system model, Q-learning estimates the values of taking a given action in a given state through repeated interaction with the environment [36].

More precisely, Q-learning seeks to approximate the optimal action-value function $Q^*(s, a)$, which represents the maximum expected return obtainable by taking action a in state s and thereafter following the best possible policy. This function satisfies the **Bellman optimality equation**:

$$Q^*(s, a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} Q^*(S_{t+1}, a') \mid S_t = s, A_t = a\right] \tag{18}$$

This equation expresses a recursive relationship: the optimal value of a state-action pair is equal to the immediate reward plus the discounted value of the best action available in the next state. In principle, if $Q^*(s, a)$ were known exactly, then an optimal policy could be obtained by selecting the action with the highest value in each state:

$$\pi^*(s) = \arg\max_{a} Q^*(s, a) \tag{19}$$

The challenge is that $Q^*(s, a)$ is generally unknown in advance and must therefore be learned from interaction. Q-learning addresses this by iteratively updating a table or function approximation of action values based on sampled transitions. Given a transition (s, a, r, s'), the Q-learning update rule is

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right] \tag{20}$$

where $\alpha \in (0, 1]$ is the learning rate, r is the reward observed after taking action a in state s, and γ is the discount factor. The term inside the bracket is knows as the temporal-difference (TD) error:

$$\delta_t = r + \gamma \max_{a'} Q(s', a') - Q(s, a) \tag{21}$$

This error measures the discrepancy between the current value estimate and a one-step bootstrap target formed from the observed reward and the estimated value of the next state. Intuitively, if the observed transition suggests that the chosen action was better than expected, the Q-value is increased; if it was worse than expected, the Q-value is decreased.

An important feature of Q-learning is that it is **off-policy**. The update uses the value of the best next action through the maximization term $\max_{a'} Q(s', a')$, regardless of which action the agent actually follows during behavior. This means that the policy used to generate experience may differ from the greedy policy learned. In practice, this is particularly useful because the agent must continue to explore the environment while also improving its value estimates. A common strategy is the ϵ-greedy policy, in which the agent selects a random action with probability ϵ and otherwise selects the action with the highest current Q-value:

$$\pi(a \mid s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|\mathcal{A}|}, & \text{if } a = \arg\max_{a'} Q(s, a') \\ \frac{\epsilon}{|\mathcal{A}|}, & \text{otherwise.} \end{cases} \tag{22}$$

This exploration mechanism is essential because the agent can only learn accurate action values for state-action pairs that it visits sufficiently often. If the policy becomes greedy too early, it may converge to suboptimal behavior simply because it has not explored enough of the environment. Conversely, excessive exploration may slow convergence. The balance between exploration and exploitation is therefore one of the key practical issues in applying Q-learning.

For this thesis, tabular Q-learning is particularly appropriate as an initial reinforcement learning framework. First, it provides a transparent and interpretable baseline that fits well with the comparative nature of the study. Second, its update mechanism is mathematically simple and directly grounded in the MDP formalism introduced earlier. Third, it allows the behavior of predator and prey agents to be analyzed through explicit action-value estimates rather than through a less interpretable black-box model. In the EcoSim environment, Q-learning is used to allow agents to learn ecological behavior from repeated interaction with a grid-based world, where state representations encode local environmental conditions, and rewards reflect ecological objectives such as survival, resource acquisition, and system stability. In this sense, Q-learning serves as the first step in evaluating whether adaptive agent-based learning can provide a meaningful alternative or complement to classical ecological equation-based modeling.

2.3.3 Multi-Agent reinforcement learning

While standard reinforcement learning is typically formulated for a single agent interacting with an environment, many real systems involve multiple decision-making entities whose behavior affects one another. In such settings, the environment is no longer passive from the perspective of each agent; part of the environment consists of other agents, each of which may also be learning and adapting over time. This gives rise to the framework of multi-agent reinforcement learning (MARL),

in which several agents simultaneously interact within a shared environment and learn policies based on individual or collective reward signals [16, 14].

Formally, MARL extends the reinforcement learning setting by introducing multiple agents $i \in 1, \ldots, n$, each with its own action and potentially its own observation, reward, and policy. At time step t, the system is in some global state $s \in S$, and each agent selects action $a_i \in A_i$. Together, these form a **joint action**

$$\mathbf{a} = (a_1, a_2, \ldots, a_n) \tag{23}$$

which determines the next state and the rewards received by the agents. The transition dynamics can therefore be written as

$$P(s' \mid s, a_1, a_2, \ldots, a_n) \tag{24}$$

Similarly, each agent may receive its own reward function $R_i(s, \mathbf{a}, s')$, or in cooperative settings, agents may share a common reward. The main difference from the single-agent case is that the outcome of an action now depends not only on the current environment state, but also on the simultaneously chosen actions of the other agents.

This creates one of the central challenges of MARL: non-stationarity. In single-agent reinforcement learning, the environment is typically assumed to have fixed transition and reward dynamics. In multi-agent settings, however, the effective environment faced by one agent changes as the policies of other agents change during learning. From the perspective of any individual learner, the transition dynamics therefore appear non-stationary, even if the global system itself follows fixed rules [16, 14]. This makes convergence substantially more difficult, since an agent must learn in an environment whose behavior is itself evolving.

A common formalism for MARL is the stochastic game, also known as a Markov game, which generalizes the MDP to multiple agents. A Markov game can be represented as a tuple:

$$\mathcal{G} = (\mathcal{S}, \mathcal{A}_1, \ldots, \mathcal{A}_n, P, R_1, \ldots, R_n, \gamma), \tag{25}$$

where $\mathcal{S}$ is the state space, each $\mathcal{A}_i$ is the action space of agent i, P is the transition function over joint actions, R_i is the reward function for agent i, and γ is the discount factor. In this setting, each agent seeks to optimize its own expected discounted return, which may depend strongly on the evolving behavior of the others. The policy of agent i can be written as $\pi_i(a_i \mid o_i)$, where o_i denotes the local observation available to the agent. This is especially important in partially observable environments, where agents do not directly observe the full global state.

Although many modern MARL methods rely on deep neural networks, the conceptual foundations also apply to tabular settings. In a tabular multi-agent setup, each agent may maintain its own value estimates and update them through local experience, even though the environment dynamics depend on the joint behavior of all agents. This kind of setup is particularly appropriate for a proof-of-concept ecological simulator such as the one developed in this thesis, where predator and prey agents interact repeatedly in a shared grid-based environment. Each species learns from local outcomes, but the ecological dynamics emerge from the collective interaction of all agents. In this sense, MARL provides the formal bridge between standard single-agent Q-learning and the adaptive predator–prey ecosystem simulation used in this study [24, 16].

2.4 Agent-based ecosystem simulation

Agent-based simulation provides an alternative modeling paradigm to aggregate equation-based ecological approaches by representing ecological systems from the perspective of individual entities and their local interactions. Instead of describing population change only through global state variables and differential equations, agent-based models (ABMs) represent systems as collections of autonomous agents that perceive, act, and interact within an environment. System-level ecological behavior then emerges from the repeated local decisions of these agents. This bottom-up perspective is especially useful in ecological settings where spatial structure, local interaction, heterogeneity, and adaptive behavior play an important role in shaping overall dynamics [21, 37].

A central strength of agent-based ecosystem simulation is that it allows ecological complexity to be represented more explicitly at the micro level. Individual organisms can differ in location, internal state, behavior, or access to resources, and these differences may influence how the larger system evolves. In contrast to classical population models, which typically describe populations as homogeneous aggregates, agent-based simulation can preserve variation between entities and thereby represent ecological processes that depend on local conditions or decentralized interaction. This is particularly important in ecosystems where movement, neighborhood effects, spatial exclusion, and resource competition contribute significantly to population-level outcomes [37, 20].

Agent-based models have become increasingly important in ecology because they are well-suited for representing systems in which global dynamics arise from local ecological rules. Rather than imposing a single equation for the whole population, the model designer specifies behavioral rules for individual agents and environmental processes, and then studies the emergent consequences of those repeated interactions. This makes ABMs especially relevant for ecosystems with adaptive organisms, spatially distributed resources, and complex feedback loops, where aggregate equations may conceal the mechanisms through which ecological patterns actually arise [21, 20]. In this sense, agent-based simulation is not simply a computational alternative to classical ecology, but a different modeling perspective that prioritizes process, interaction, and emergence.

The concept of emergence is particularly important in this context. In ecological systems, many large-scale patterns are not directly imposed, but arise from repeated local interactions among organisms and between them and their environment. Population cycles, clustering, competition effects, and spatial self-organization may all emerge from simple rules operating at the level of individuals. Levin emphasizes that ecological complexity is often the result of self-organization across scales, meaning that system-level regularities cannot always be understood solely by examining aggregate variables in isolation [20]. Agent-based simulation is especially valuable precisely because it provides a way to study such emergent behavior computationally.

For ecological applications, agent-based simulation also offers a natural way to incorporate spatial and behavioral realism. Agents may move across explicit environments, consume resources locally, compete for territory, reproduce under local constraints, or respond to nearby threats. These features make ABMs particularly suitable for predator-prey systems, where ecological outcomes depend strongly on encounters, movement, local density, etc. As a result, agent-based ecosystem simulation has become a useful framework for studying ecological dynamics that are difficult to represent cleanly through fully aggregated equations alone [21, 37].

2.5 Reinforcement learning in ecology and environmental management

Reinforcement learning has become increasingly relevant in ecological and environmental research because many problems in these domains can naturally be framed as sequential decision-making under uncertainty. In environmental systems, decisions often have long-term consequences, must be taken with incomplete knowledge of future conditions, and involve trade-offs between short-term gains and long-term system stability. This makes RL especially attractive since its central objective is to learn policies that maximize cumulative reward through repeated interaction with an environment rather than through static one-step prediction alone [19, 23].

A major reason for this relevance is that ecological and environmental management problems are often dynamic and feedback-driven. Conservation planning, natural-resource management, ecosystem monitoring, and intervention design all involve actions whose effects unfold over time and may alter the future state of the system. Unlike classical optimization methods that assume a fixed decision context, RL explicitly accounts for the fact that current decisions shape future opportunities and constraints. This makes it well-suited to settings where management must adapt to changing system states, uncertain outcomes, and nonlinear ecological responses [28, 19].

Within ecology, more specifically, RL has been proposed not only as a tool for management and policy support but also as a framework for studying adaptive behavior in ecological systems themselves. The idea of ecological reinforcement learning emphasizes that learning agents and ecological environments are deeply connected: the structure of the environment, the distribution of resources, and the costs and rewards of behavior all affect what strategies are learned and how adaptation unfolds over time [6]. From this perspective, RL is not merely a computational technique applied to ecology from the outside but a framework capable of representing ecological interaction, adaptation, and feedback in a formally consistent way.

This becomes especially relevant in predator-prey systems, where the behavior of one species directly shapes the strategic environment of the other. Recent studies have shown that reinforcement learning can be used to simulate prey-predator interaction in a way that produces adaptive and ecologically meaningful dynamics. Park et al., for example, model predator and prey populations as learning agents and show that co-learning can generate oscillatory population behavior resembling classical ecological cycles [24]. Such work is important because it suggests that RL-based ecosystem simulation may capture aspects of ecological interaction that are difficult to specify fully in advance through fixed equations alone, especially when behavior depends on repeated feedback from the environment.

At the same time, reinforcement learning has also attracted interest in environmental management because it can support decisions in settings characterized by uncertainty, delayed outcomes, and limited resources. In conservation, for example, RL has been used to study action selection under budget constraints and uncertain ecological dynamics, showing that learned policies may provide flexible alternatives to static intervention rules [19]. More broadly, AI-based methods have been proposed as tools to improve biodiversity protection by supporting prioritization, prediction, and decision-making in complex systems where the scale of the problem exceeds what can easily be handled through just manual analysis [28]. In this sense, RL stands at the intersection of two important developments, the use of AI and the implementation of learning-based simulations.

Nevertheless, RL in ecology and environmental management remains an emerging area rather than a generally used field. Many RL approaches still face challenges related to interpretability, scalability, reward specification, and ecological realism. In addition, the success of RL models

depends strongly on how the environment is formulated, how objectives are encoded, and how closely the simulation learning problem reflects the ecological system of interest [6, 24]. These limitations are important, but they do not diminish the value of RL as a research direction. Instead, they highlight the need for comparative studies that evaluate the feasibility of its use.

2.6 Research gap and positioning of this thesis

The literature reviewed in the previous sections shows that two strong but only partially connected traditions currently exist. On the one hand, classical ecological modeling provides a long-established mathematical framework for describing population dynamics, predator-prey interaction, and ecosystem regulation through interpretable equation-based models [18, 29, 13]. On the other hand, recent work in RL has demonstrated that adaptive agents can be used both to study ecological interaction and to support environmental decision making under uncertainty [6, 24, 19, 28]. However, despite the promise of this second line of work, there is still a limited amount of research that directly compares RL-based ecosystems and the opportunities that the new technology brings to the classical methods in a grounded and relevant way.

More specifically, much of the classical ecological literature prioritizes interpretability, analytical tractability, and population-level description, but often struggles with adaptive behavior, partial observability, non-stationarity, and context-dependent interaction. Conversely, RL-based approaches offer flexibility, feedback-driven adaptation, and the ability to model decentralized interacting agents, but they are often studied either in abstract learning environments or in highly application-specific management problems rather than in explicit comparison with traditional ecological baselines [13, 4, 24]. As a result, an important gap remains between the theoretical strengths of RL and a sequential decision-making framework and its evaluation as a scientifically meaningful alternative or complement to well-established ecological modeling practices.

A second gap concerns the intersection between ecosystem simulation and environmental policy support. Existing work on ecological decision support systems and AI for biodiversity protection suggests that environmental decision-making increasingly requires tools that can deal with uncertainty, dynamic feedback, and long-term consequences [10, 35, 28]. At the same time, RL research in ecology has shown that learning agents can generate adaptive system behavior, particularly in predator-prey settings [6, 24]. Yet there remains relatively little work that explicitly asks whether RL-based simulation can help bridge these two domains: that is, whether it can function not only as a computational learning method, but also as a plausible modeling framework for ecological systems whose behavior matters for policy reasoning.

This thesis is positioned precisely within that gap. It does not attempt to replace classical ecological theory, nor does it claim to build a fully realistic policy-ready ecological DSS. Instead, it investigates whether reinforcement learning can provide a useful complementary approach for ecosystem simulation by comparing an RL-based predator-prey model against the classical ecological ODE baselines in a controlled setting. The focus on predator-prey dynamics provides a tractable, but non-trivial, ecological context, while the comparison with equation-based models makes it possible.

3 Methodology

3.1 Research design overview

This thesis adopts a comparative modeling design to study predator-prey ecosystem dynamics. It compares a classical ecological baseline, based on ordinary differential equations, with a reinforcement-learning-based ecosystem simulator in which predator and prey agents interact within a spatial environment and adapt their behavior. The purpose of this design is to assess whether a learning-based simulation framework can provide a useful complementary perspective to classical ecological modeling [13, 19].

The study is both comparative and exploratory. It is comparative because both modeling approaches are examined within the same general predator-prey setting. It is exploratory because the thesis does not assume in advance that one modeling paradigm is universally superior, but instead investigates the kinds of ecological behavior each can represent and the assumptions under which those behaviors emerge [6].

Overall, the methodology should be understood as a concept validation framework for examining the role of reinforcement learning in ecosystem simulation. It does not attempt to replace classical ecological models or to construct a fully realistic environmental decision-support system. Rather, it provides a structured way to study whether an RL-based predator-prey simulator can capture relevant ecological dynamics and offer useful insight alongside a classical aggregate baseline [17, 29, 24].

3.2 General comparison framework

The comparison is carried out in a predator-prey context and is organized around two distinct modeling setups. The first is a classical ODE baseline that represents predator and prey populations using continuous state variables and fixed interaction terms. The second is an RL-based simulator, which represents the same ecological setting through individual agents, local interactions, and learned decision-making.

The framework compares these two setups under matched ecological conditions as far as possible. The goal is not to force exact equivalence between two practically different techniques, but to evaluate how they differ in the dynamics they generate and in the assumptions they take.

The comparison includes both quantitative and qualitative elements. Quantitatively, it examines outputs such as population trajectories, persistence or extinction behavior, and other measures related to system stability. Qualitatively, it considers the overall character of the resulting dynamics, including oscillatory behavior, coexistence patterns, and the extent to which the RL-based model generates adaptive ecological interaction that is not directly imposed by the baseline equations [26, 3].

3.3 Classical ecological baseline

3.3.1 Selected ODE models

The classical ecological baseline in this thesis consists of two deterministic predator-prey ODE models implemented separately from the RL simulator. The first is the standard Lotka-Volterra model, used as the simplest aggregate benchmark for predator-prey interaction. In this formulation,

prey growth is modeled as exponential in the absence of predators, while predator decline is modeled as exponential in the absence of prey. The implemented system is given by:

$$\frac{dN}{dt} = \alpha N - \beta NP \tag{26}$$

$$\frac{dP}{dt} = \delta\beta NP - \gamma P \tag{27}$$

where N denotes prey population, P predator population, α prey birth rate, β the predation rate coefficient, and γ the predator mortality rate. This model serves as the most basic classical reference because it captures cyclical prey-predator interaction while remaining mathematically compact and interpretable.

The second implemented baseline is the logistic-prey model, corresponding to the Rosenzweig-MacArthur-style extension of Lotka-Volterra. This formulation introduces prey carrying capacity and is therefore more directly aligned with the ecological constraints built into the RL simulator. The implemented system is

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) - aNP \tag{28}$$

$$\frac{dP}{dt} = eaNP - dP \tag{29}$$

where r is the prey intrinsic growth rate, K the prey carrying capacity, a the attack rate, e the predator conversion efficiency, and d the mortality rate. Relative to the standard Lotka-Volterra system, this model adds explicit density dependence on the prey side and therefore provides a more ecologically constrained baseline.

3.3.2 Assumptions and parameters

The ODE baseline is built on a set of assumptions that distinguish it clearly from the RL-based simulation. First, both models are formulated in continuous time and solved numerically deterministically. Second, predator and prey are represented as homogeneous populations, rather than individuals. Third, the interaction coefficients are fixed throughout each simulation, so the models do not represent adaptation, local decision-making, or time-varying ecological behavior. Finally, spatial structure, environmental stochasticity, and individual heterogeneity are absent from the ODE formulation. These assumptions are appropriate for a classical benchmark, but they also delimit the ecological mechanisms that the baseline can represent [13, 8, 7].

The parameters of the two models follow their standard biological interpretations. In the Lotka-Volterra system, the fitted parameters are prey birth rate α, predation rate β, predator conversion efficiency δ, and predator mortality γ. The logistic-prey system shares the fitted parameters, but with the introduction of the prey growth rate r and the prey carrying capacity K. In addition to these ecological coefficients, the ODE analysis also specifies numerical-simulation parameters such as solver type, tolerance settings, number of output points, initial conditions, and simulation horizon [29, 5].

Parameter selection in the implemented ODE analysis follows two different strategies. For the historical hare–lynx and Isle Royale case studies, parameters are obtained by numerical fitting

using differential evolution, with model quality assessed through trajectory fit and information criteria. For the RL-calibrated ODE scenario, parameters are instead derived analytically from the RL simulator configuration, using approximate translations from grid-world quantities such as carrying capacity, action radius, step horizon, and energy decay. This produces an illustrative ODE parameterization grounded in the simulator design rather than in historical ecological data.

Methodologically, several limitations of these parameter choices are important. The logistic-prey model does not always behave as a substantively distinct model from standard Lotka–Volterra, since fitted carrying-capacity values can collapse toward the optimization boundary, effectively removing the logistic constraint. Likewise, some ecological regimes, most notably the Isle Royale wolf crash, cannot be reproduced well by either deterministic ODE model because they reflect structural breaks outside the scope of fixed-parameter aggregate equations. Finally, the RL-calibrated ODE parameters should be interpreted as approximate scale translations rather than as validated ecological estimates. For this reason, the ODE baseline is used in the thesis primarily as a classical comparative benchmark rather than as a fully realistic ecological forecasting model [25, 30, 31].

3.4 RL-based simulation approach

3.4.1 Environment design

The reinforcement-learning-based ecosystem simulator is implemented as a discrete two-dimensional grid-world environment in which predator and prey agents interact over time. The spatial substrate consists of a bounded grid of size 150×150, where each location corresponds to a discrete cell identified by integer (x, y) coordinates. The environment is not wrapped toroidally; instead, all movement and lookup operations are constrained by hard spatial boundaries, so agents cannot move beyond the limits of the grid. This design makes the simulator explicitly spatial while retaining a relatively simple and computationally tractable structure.

At the substrate level, the environment currently implements a simplified tile-based resource system. The two operative tile types are grass and empty. Grass tiles provide consumable energy to prey agents and regenerate after depletion according to a fixed regrowth delay, whereas empty tiles provide no resources. Resource regeneration is deterministic and local: once a grass tile is consumed, it becomes unavailable for a fixed number of steps before becoming edible again. This produces a renewable but spatially distributed food source that constrains prey movement and survival.

The simulator advances in discrete timesteps, and ecological change occurs through repeated update cycles. At a high level, each step consists of agent action, environmental update, and post-action ecological processing. In the single-agent training wrapper, the main learning agent acts first, followed by the background agents; after all actions are applied, resources regenerate, dead agents are removed, reproduction is processed, and episode termination conditions are checked. In the multi-agent training wrapper, all currently alive agents are first collected and randomly shuffled, after which each acts once during the step. Resource regeneration, epsilon decay, dead-agent cleanup, and reproduction are then processed in sequence. This means that the environment is not dully synchronous in the strict game-theoretic sense, but rather sequential and order-sensitive, with randomization used in the multi-agent setting to reduce systematic update bias.

Episodes begin with an environment reset, during which both the ecological substrate and the

agent population are initialized. The tile map may either be generated procedurally according to a predefined tile distribution or loaded from a test map. Agents are placed at random, discrete positions on the grid. The environment supports both a single-agent training configuration, in which one focal learning agent is embedded in a larger ecological population, and a multi-agent configuration, in which all active agents are learning agents. Episode length is bounded by a fixed step horizon, and episodes terminate either when the maximum number of steps is reached or when the relevant death conditions are satisfied, depending on the wrapper used.

From an implementation perspective, the environment stores ecological state through several coordinated data structures. These include the tile grid itself, an occupancy grid, a flat list of agents, and a position-indexed mapping that associates grid cells with the agents currently occupying them. Importantly, the environment allows multiple agents to occupy the same cell, meaning that movement does not enforce exclusive spatial occupancy. Local ecological interactions are resolved through local spatial lookup rules on the discrete grid. As a result, the environment should be understood as a discretized spatial ecosystem with local lookup rules rather than as a physically realistic continuous-space model.

At the ecological level, the environment incorporates several key mechanisms relevant to predator-prey dynamics. These include renewable food sources, random initial population placement, local offspring placement during reproduction, immediate death and removal when agents expire, and direct predation events when predators successfully consume prey. Reproduction depends on threshold and neighborhood conditions, while predation and death are processed directly at the environment level to maintain consistency of the population state. Taken together, these mechanics define the ecological substrate within which learning takes place: the RL agents do not act in an abstract benchmark environment, but in a spatially explicit world with resource renewal, mortality, reproduction, and species interaction.

One important methodological implication of this design is that the environment deliberately prioritizes controlled ecological abstraction over full biological realism. The world is bounded and discrete, resources are modeled in a simplified tile-based way, and interactions are local but not physically continuous. These simplifications are appropriate for a proof-of-concept ecosystem simulator because they make agent learning and population dynamics computationally manageable while preserving the core ecological elements required for predator-prey simulation. At the same time, they also define part of the methodological scope of the study, since any results obtained from the simulator must be interpreted relative to this simplified but explicit environmental structure.

3.4.2 Agent types

The RL-based ecosystem simulator implements three agent classes: a shared base class, Agent, and two species-specific subclasses, Prey and Predator. The base class defines the common structural elements of all agents, including position, energy state, movement behavior, and the general observation interface, while the two subclasses implement specific feeding, reproduction, and interaction logic. There are no distinct non-learning or background subclasses; instead, both prey and predator agents may optionally be assigned a reinforcement-learning object through a reference field, allowing the same ecological agent class to function either as a learning agent or as a non-learning agent depending on the training setup.

At the shared level, each agent is defined by a unique identifier, a discrete spatial position, a species label, and an internal energy value. Agents also store a vision radius, a speed parameter, a

reward bookkeeping field, and an optional pointer to a Q-learning object used externally by the training system. In addition, both prey and predator agents maintain a reproduction cooldown, which is decremented over time and constrains the frequency of reproduction events. The effective internal survival dynamics of agents are therefore driven by energy and reproduction-related variables rather than by a full dual-resource metabolism.

The two species differ primarily in movement speed, feeding mechanism, and reproductive constraints. Prey agents consume grass resources from nearby tiles and use the corresponding energy gain to maintain survival and enable reproduction. Predator agents instead consume prey agents directly through local predation events, gaining a fraction of the prey's energy after a successful kill. Predators also move faster than prey, which creates an asymmetric interaction structure consistent with their ecological role. Reproduction is implemented for both species through a sexual pairing process requiring a nearby mate, sufficient energy, a reproduction cooldown expiry, and a free nearby location for offspring placement. In addition to these shared constraints, the prey population is further limited by a carrying capacity boundary that is computed as a function of the size of the grid, approximating the constraints of a real ecosystem.

At the lifecycle level, all agents lose energy over time and through movement, and they die when their energy falls to zero or below. Death is handled explicitly by removing the agent from the environment's population list, spatial position map, and occupancy grid, thereby preventing dead agents from influencing the simulation state. When reproduction succeeds, offspring are created as new instances of the appropriate subclass and initialized with species-specific offspring energy values. This means that the simulation models ecological population turnover directly at the agent level rather than through aggregate birth-death rates alone.

From a methodological perspective, the agent design reflects the hybrid character of the simulator. Ecologically, the agents represent individual predator and prey organisms with local interaction rules and explicit survival constraints. Computationally, they serve as the carriers of the RL behavior when linked to a learning module. This separation between ecological embodiment and learning control is important because it allows the same underlying ecosystem to be used under different training configurations without redefining the species logic itself. At the same time, it also implies that the biological realism of the model is bounded by the simplicity of the implemented agent state, particularly in the absence of richer physiological or behavioral variables beyond energy, spatial position, and reproduction status.

3.4.3 State representation

The reinforcement-learning agents in the simulator operate under partial observability. Rather than accessing the full ecological state of the environment, each agent receives a compact local observation vector derived from its own internal state and from nearby ecological features. This design means that the RL state does not encode the complete simulator configuration, such as global population counts, distant agents, or the full tile map. Instead, it provides a species-specific local summary intended to support tractable decision-making.

In the current implementation, the observation vector has dimensionality six and follows a shared structural format for both species:

$$o = [e,\, d,\, \Delta x,\, \Delta y,\, I_{\text{target}},\, c]\,. \tag{30}$$

Here, e is the normalized energy level of the agent, d is the normalized distance to the nearest

ecologically relevant target, Δx and Δy are normalized direction components, I_{target} is a binary target-detection indicator, and c is a binary context variable whose interpretation depends on the species. Although the vector structure is shared, the ecological meaning of the target and of the final dimension differs between prey and predator agents.

For prey agents, the observation uses conditional encoding. The prey vector can be written as:

$$o^{\text{prey}} = [e,\, d_{\text{primary}},\, \Delta x_{\text{primary}},\, \Delta y_{\text{primary}},\, I_{\text{food}},\, I_{\text{pred}}]\,, \tag{31}$$

where (d_{primary}, $\Delta x_{\text{primary}}$, and $\Delta y_{\text{primary}}$ encode either the nearest food source or the nearest predator, depending on context), I_{food} indicates whether food is sufficiently close to be considered locally detected, and I_{pred} indicates whether a predator is locally detected. When $I_{\text{pred}} = 0$, the primary distance and direction variables encode the nearest grass tile, and the prey is in foraging mode. When $I_{\text{pred}} = 1$, the same variables encode the nearest predator, and the prey is in evasion mode. This conditional encoding allows the same state dimensions to support both food-seeking and directional escape behavior without increasing the size of the tabular state space.

For predator agents, the target is the nearest prey agent. The observation is therefore oriented toward local hunting behavior. The predator vector can be written as:

$$o^{\text{pred}} = [e,\, d_{\text{prey}},\, \Delta x_{\text{prey}},\, \Delta y_{\text{prey}},\, I_{\text{prey}},\, I_{\text{density}}]\,, \tag{32}$$

where d_{prey} is the normalized Manhattan distance to the nearest prey, Δx_{prey} and Δy_{prey} encode prey direction, I_{prey} is the corresponding detection flag, and I_{density} is a binary prey-density indicator that distinguishes between sparse and prey-rich local neighborhoods. Thus, both species use all dimensions of the observation vector, but the final context variable captures different ecological information for prey and predator agents.

Observation values are computed from local spatial queries and then normalized to the interval $[0, 1]$. Energy is scaled by the maximum agent energy. Distances are computed using Manhattan distance ($|dx| + |dy|$) and then normalized by twice the vision radius. Direction components are linearly mapped from signed offsets to the unit interval. Detection indicators are binary, but the thresholds used to determine detection are not fully symmetric across species. Prey food detection depends on whether the nearest grass resource lies within the action radius, whereas predator target detection depends on whether the normalized prey distance is below a threshold of 0.5. Prey-side predator detection is also binary and determines whether the observation is interpreted in foraging mode or evasion mode.

Before learning, the observation vector is converted into a discrete state by the tabular Q-learning component. Discretization is applied uniformly across both species and maps the six continuous or binary observation components into a tuple of discrete indices. The implemented discretization scheme is:

- energy: 5 bins
- target distance: 3 bins
- direction x: 3 bins
- direction y: 3 bins
- target detected: 2 bins

- context variable: 2 bins

This yields a theoretical discrete state space of size:

$$|\mathcal{S}| = 5 \times 3 \times 3 \times 3 \times 2 \times 2 = 540. \tag{33}$$

The same discretization mapping is used for prey and predator agents; species differences arise only in the way the observation vector is populated before discretization. As a result, the RL system uses a shared tabular state architecture with species-specific ecological content.

Methodologically, this state representation provides a compact and computationally manageable abstraction of the local ecological situation. It captures survival-relevant variables such as energy, target proximity, coarse direction, and—for prey—an explicit predator-danger signal. At the same time, the representation remains intentionally coarse. It excludes most global ecological information and compresses directional information into only three bins per axis. The current design is therefore best interpreted as a simplified local ecological state abstraction suitable for tabular learning rather than as a complete state description of the full ecosystem.

3.4.4 Action space

The RL-based ecosystem simulator uses a discrete action space of size ten, shared by both predator and prey agents. The action set consists of eight movement actions, one feeding action, and one idle action. Formally, the action space can be written as:

$$\mathcal{A} = \{a_0, a_1, \ldots, a_9\}, \tag{34}$$

where a_0 to a_7 correspond to directional movement, a_8 corresponds to eating, and a_9 corresponds to idling. The movement component is eight-directional, including the four cardinal directions and the four diagonals. This gives agents a richer spatial control scheme rather than a purely four-neighbor grid model and allows both predator pursuit and prey avoidance to emerge through directional movement in a spatially explicit environment.

More specifically, the movement actions correspond to up, down, left, right, up-right, down-right, up-left, and down-left. These actions are implemented uniformly across species, but their ecological consequences are not fully identical because predators and prey differ in movement speed. Prey agents use the default movement speed of 1, whereas predators move at speed 2. Since movement energy cost scales with speed, predator movement is more expensive, reflecting the greater energetic cost of faster locomotion. Movement actions are not pre-masked at the policy-selection stage when they would lead outside the grid boundary; instead, such actions are allowed to be selected and simply fail harmlessly at execution time if they would take the agent out of bounds.

The two non-movement actions are eat and idle. The idle action represents deliberate inaction for a single timestep. The eat action is shared as an action index across both species, but its logic is species-dependent. This is an important design choice because it preserves a common RL control structure while still allowing species roles to differ through their interaction with the environment.

Notably, the actions do not include reproduction explicitly. Rather than being action-directed, reproduction is handled automatically by the environment when the conditions are met. This means that the action policy directly controls survival-oriented behavior, while longer-term population turnover is only indirectly shaped by what the learned policy enables.

The action-selection mechanism also includes limited action masking. In particular, the eat action is removed from the valid action set when the discrete state indicates that no target is currently detected. This prevents agents from repeatedly selecting the eat action in states where no food or prey is locally available according to the observation model. No analogous masking is used for movement actions, which means that spatially invalid movement choices remain possible and are filtered only during execution.

Methodologically, the action space is intentionally compact. It contains enough structure to support ecologically meaningful local behavior, movement, feeding, and inactivity, without becoming too large for tabular Q-learning. At the same time, its simplicity also imposes limitations. Since reproduction is not an explicit action, the agent cannot learn reproductive timing as a direct policy choice. The action space should therefore be understood as a simplified ecological control interface suitable for proof-of-concept spatial predator–prey learning rather than as a complete behavioral model of real organisms.

3.4.5 Reward design

The reward system of the simulator is formulated as an event-based per-step reward function designed to combine individual survival incentives, species-specific ecological interaction signals, and reproductive success. At each timestep, the reward returned to the learning agent is not a single monolithic quantity, but the sum of several possible terms activated by the agent's action outcome and by the ecological state observed after the action. In general form, the step reward can be written as:

$$r_t = r_t^{\text{step}} + r_t^{\text{eat}} + r_t^{\text{detect}} + r_t^{\text{repr}} + r_t^{\text{death}} \tag{35}$$

where r_t^{step} is a generic step penalty, r_t^{eat} is the feeding-related reward, r_t^{detect} is a species-specific detection term, r_t^{repr} is the reproduction reward, and r_t^{death} is the terminal death penalty when applicable. Not every term is active at every step, but this decomposition describes the logic of the implemented reward structure.

Let $\lambda_s < 0$ denote the generic step penalty. The base event contribution can be written as

$$r_t^{\text{step}} = \lambda_s \tag{36}$$

This term acts as a weak pressure against inefficient behavior and encourages agents to avoid unproductive action sequences. In addition to this base cost, the reward system includes species-specific feeding terms. For prey agents, eating grass yields a reward proportional to the energy gained from the consumed resource. If E_t^{gain} denotes energy gained from food consumption and λ_e the energy-reward scale, the prey feeding reward is

$$r_{t,\text{eat}}^{\text{prey}} = \lambda_e E_t^{\text{gain}} \tag{37}$$

For predators, successful predation yields a reward based not on the prey's full energy but on a fixed transferred fraction of that energy. Let $I_t^{\text{kill}} \in \{0, 1\}$ denote whether the eat action resulted in a successful kill. If E_t^{prey} denotes the prey's energy before consumption, then the predator's feeding reward is

$$r_{t,\text{eat}}^{\text{pred}} = \left(\lambda_e \left(0.25\, E_t^{\text{prey}}\right) + \lambda_h\right) I_t^{\text{kill}} \tag{38}$$

where λ_h is the hunting success bonus, applied only when a kill occurs.

The reward system also includes species-specific detection shaping. For prey, nearby predator detection introduces a penalty, whereas for predators, nearby prey detection introduces a bonus. Let $I_t^{det} \in \{0, 1\}$ denote the relevant binary detection signal. Then the detection term can be written as

$$r_{t,\text{detect}}^{\text{prey}} = \lambda_p I_t^{\text{pred\_det}}, \qquad \lambda_p < 0 \tag{39}$$

$$r_{t,\text{detect}}^{\text{pred}} = \lambda_d I_t^{\text{prey\_det}}, \qquad \lambda_d > 0 \tag{40}$$

For prey, this means that being locally exposed to predators is penalized. For predators, it means that successfully locating prey is rewarded even before a kill count. This makes the reward sensitive not only to consumption events, but also to ecologically meaningful awareness states.

Death is treated as the strongest negative event in the system. If the learning agent dies, a fixed negative terminal penalty is applied. In mathematical notation, this reward can be written as

$$r_t^{\text{death}} = \begin{cases} \lambda_{\text{death}}, & \text{if the agent dies at step } t, \\ 0, & \text{otherwise} \end{cases} \qquad \lambda_{\text{death}} < 0 \tag{41}$$

A reproduction reward is also defined using I_t^{repr}, which denotes a successful reproduction event, and its contribution can be written as

$$r_t^{\text{repr}} = \lambda_r I_t^{\text{repr}}, \qquad \lambda_r > 0 \tag{42}$$

Combining the above components, the implemented prey and predator rewards can be summarized as

$$r_t^{\text{prey}} = \lambda_s + \lambda_e E_t^{\text{gain}} + \lambda_p I_t^{\text{pred\_det}} + \lambda_r I_t^{\text{repr}} + r_t^{\text{death}} \tag{43}$$

$$r_t^{\text{pred}} = \lambda_s + \left(\lambda_e \left(0.25\, E_t^{\text{prey}}\right) + \lambda_h\right) I_t^{\text{kill}} + \lambda_d I_t^{\text{prey\_det}} + \lambda_r I_t^{\text{repr}} + r_t^{\text{death}} \tag{44}$$

These expressions represent the intended effective reward structure at the learning level, with the note that not all terms are activated at every step.

3.4.6 Learning Algorithm

The learning component of the simulator is based on tabular Q-learning. As mentioned earlier, both prey and predator agents use the same Q-learning class, and each learning agent is associated with its own Q-table. States are represented as discrete tuples derived from the observation discretization already described, and each state maps to a vector of action values over the ten-action space. Unseen states are initialized lazily through a zero-valued default structure, which makes the learning system sparse and state-driven [36].

The implemented update rule follows the standard off-policy Q-learning form. For a transition (s_t, a_t, r_t, s_{t+1}), the value update is

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right) \tag{45}$$

with the bootstrap term removed when the next state is terminal. Here, α denotes the learning rate and γ the discount factor. Action selection during training is performed using an ϵ-greedy policy, while evaluation uses greedy action selection with exploration disabled. The implementation also includes action masking for the eat action, which is removed from the valid action set when the discretized state indicates that no target is currently detected.

Although the core method is standard Q-learning, the implementation contains a few additional mechanisms that affect learning dynamics. First, when a state has no exact Q-table entry, the policy can fall back to the nearest known state as a heuristic for action selection. Second, the broader training setup supports policy persistence through save/load functions based on serialized Q-tables, allowing learned policies to be reused across training phases or experimental runs. These elements make the implementation more than a purely textbook tabular baseline, even though its core learning rule remains standard Q-learning [16, 14].

From a methodological perspective, this choice of algorithm is appropriate for the thesis because it provides a transparent and interpretable reinforcement learning baseline that can be directly linked to the discretized ecological state representation. At the same time, it should be noted that some of the theoretical convergence assumptions of single-agent tabular Q-learning are weakened in the multi-agent setting used here, since the effective environment may become non-stationary when multiple agents learn simultaneously. The algorithm should therefore be understood as a tractable proof-of-concept learning method for ecological simulation rather than as a guaranteed-convergent solution method in the full multi-agent case [36, 14, 16].

3.5 Multi-agent training setup

The training setup used in this thesis consists of two implemented protocols. The first is a 1v1 dual-agent setting in which one prey agent and one predator agent learn simultaneously in a minimal environment. The second, and primary, protocol is an alternating training setup in which only one focal agent learns at a time while the remaining agents act with frozen policies. This second protocol is used to produce the final policies analyzed in the study.

In the 1v1 setting, both predator and prey maintain separate Q-tables and update them independently after each step. The environment contains exactly one prey and one predator, with no background population. Turn order alternates across episodes so that neither species systematically benefits from acting first. This protocol is used as a controlled warm-up stage in which basic pursuit and survival behavior can emerge in a simplified setting.

In the alternating training setup, learning is organized into cycles with two phases. In the first phase, the prey agent is the only learner, and the predator population is frozen. In the second phase, the predator agent becomes the only learner, and the prey population is frozen. The learning environment contains one focal agent together with background prey and predator agents. Background agents do not update their Q-values during the phase. Opponent agents use the frozen policy inherited from the previous phase, while same-species background agents use a frozen copy of the local learner's policy taken at the start of the phase. This design reduces non-stationarity during training by ensuring that only one policy changes at a time [2, 27].

Each episode follows the same basic structure. The environment is reset with a deterministic seed, actions are executed, ecological updates are applied, reproduction is resolved, and the focal learner performs a Q-update using the full reward returned for that step. This reward includes action-related terms and, when reproduction occurs, the reproduction bonus as well. Episodes

terminate when the focal agent dies or when the predefined step limit is reached. Exploration follows an ϵ-greedy schedule, with ϵ decaying once per episode rather than once per step. During evaluation, exploration is disabled, and the learned policy is run greedily without Q-updates.

Methodologically, the alternating protocol is the main mechanism used to make tabular Q-learning tractable in a predator-prey setting. By freezing the opponent and same-species background agents within each phase, the focal learner experiences a more stable environment than it would under fully simultaneous multi-agent learning. This does not remove non-stationarity completely across the full training process, but it reduces it substantially and makes the learned policies more stable and interpretable for comparative analysis [2, 14].

3.6 Case studies/datasets

This thesis uses two historical ecological datasets and one synthetic calibration scenario. The historical datasets are used exclusively for the ODE baseline, while the RL simulator is trained and evaluated in procedurally generated environments rather than on empirical ecological records. As a result, the case studies serve primarily as a classical reference system for parameter fitting, qualitative comparison, and analysis of the limits of aggregate ecological models.

The first case study is the Hudson Bay hare-linx dataset, consisting of annual predator and prey population counts from 1900 to 1920 [30, 31, 34]. This is the main empirical benchmark used for fitting the Lotka-Volterra and logistic-prey models, and it provides the strongest classical baseline in the project because both models can reproduce its oscillatory behavior reasonably well. The second case study is the Isle Royale wolf-moose dataset, consisting of annual counts from 1959 to 2011 [25, 22]. This dataset is also used for ODE fitting and analysis, but mainly as a stress case that highlights the limitations of deterministic predator-prey models under real ecological perturbations and structural breaks.

In addition, the project includes an RL-calibrated synthetic scenario in which a logistic predator-prey model is parameterized approximately from the simulator configuration. This scenario is not fitted to empirical data and is not used for direct benchmarking. Its role is only illustrative; it provides a rough ODE-scale translation of the RL environment in order to examine how the simulator's ecological assumptions would look in a classical aggregate formulation.

Methodologically, these three cases play different roles. The hare-lynx dataset provides the main empirical ODE benchmark, Isle Royale is used to expose model limitations, and the RL-calibrated scenario serves as an approximate bridge between the classical and simulation-based parts of the thesis. None of them is directly used for RL training, which remains fully simulation-based.

3.7 Evaluation Metrics

The evaluation metrics used in this thesis are grouped into three categories: ODE model diagnostics, RL simulation diagnostics, and cross-model comparative metrics. These metrics are not identical across the two pipelines. Instead, the ODE models and the RL simulator are first evaluated within their own methodological context, after which comparison is performed through post-hoc fitting and diagnostic interpretation [3, 26].

For the ODE baseline, evaluation is based primarily on model fit, model selection, and dynamical analysis. Model fit is quantified separately for predator and prey using root mean squared error (RMSE), while model selection between Lotka–Volterra and Rosenzweig–MacArthur is performed

using the Akaike Information Criterion (AIC). In addition, the ODE analysis includes equilibrium and local stability analysis through Jacobian eigenvalues, phase portraits with nullclines, and parameter sensitivity analysis around the coexistence equilibrium. Together, these metrics are used to assess both empirical fit and qualitative dynamical structure [26, 12].

For the RL simulator, evaluation focuses on learning performance and long-horizon ecological behavior. During training, cumulative episode reward and episode outcome are logged to monitor convergence. After training, frozen policies are evaluated over long-horizon runs using per-step records of prey and predator populations, births, deaths, predation events, extinction status, and coexistence score. The coexistence score is a normalized metric in the interval $[0, 1]$ that combines prey viability, predator viability, and the balance between the two populations; in practice, it is used as a time-averaged indicator of whether both species remain present in a reasonably stable ecological proportion over the evaluation horizon. From these trajectories, additional summary metrics are derived, including population coefficient of variation, oscillation period, survival fraction, extinction step, demographic rates, and mean individual lifespan. Here, survival fraction denotes the mean fraction of the 2000-step evaluation horizon during which both species remained alive, averaged across all evaluation seeds. These metrics are used to assess whether the learned policies produce stable and ecologically meaningful dynamics beyond short-term reward maximization.

Cross-model comparison is performed by fitting discrete Lotka–Volterra parameters to RL-generated population trajectories. This allows RL dynamics to be expressed approximately in the same aggregate parameter space as the classical ODE models. The resulting fit provides a quantitative basis for comparing trajectory shape, oscillation behavior, and equilibrium tendency, even though the ODE and RL pipelines do not share a fully unified evaluation metric set.

3.8 Limitations of the methodology

This methodology has several limitations that should be made explicit when interpreting the results. First, the classical ODE baseline and the RL simulator do not represent the ecological system at the same level of abstraction. The ODE models are deterministic, population-level, and spatially homogeneous, whereas the RL simulator is discrete, spatially explicit, and based on individual agents. As a result, the comparison is necessarily between two different modeling paradigms rather than between two directly equivalent implementations of the same ecological process [13, 8].

The ODE baseline is limited by its aggregate structure. It cannot represent local spatial interactions, within-species heterogeneity, or adaptive behavior. Its parameters are not fitted to a fully empirical ecological field setting, but are instead derived from simulator-based calibration or estimated from synthetic trajectories. This means that the ODE analysis is best interpreted as a benchmark for aggregate ecological dynamics within the project rather than as a biologically validated population model [13, 4].

The RL simulator is also a strong simplification of real ecological systems. Agents act on a coarse local state representation, the action space is intentionally limited, and the environment is procedurally generated rather than based on an empirical landscape. In addition, offspring inherit parental policies directly, which improves population continuity in the simulation but reduces biological realism. The resulting system is therefore suitable for studying learned predator-prey interaction in a controlled setting, but not for making strong claims about real ecosystem behavior [21, 37].

The learning setup introduces further limitations. Although alternating training reduces non-

stationarity by freezing opponents within each phase, it remains only an approximation to the full multi-agent learning problem. Background agents do not adapt during a phase; same-species background policies can lag behind the focal learner, and no formal convergence criterion is imposed beyond the fixed training budget. Moreover, training and long-horizon evaluation are not conducted under exactly the same interaction structure, which introduces a distributional gap between learning and final assessment [16, 14, 2].

Finally, the comparison methodology itself remains partly indirect. The ODE and RL pipelines do not share a single common metric set, and there is no exact temporal alignment between continuous ODE time and discrete simulation steps. Cross-model comparison, therefore, relies on post-hoc fitting, trajectory interpretation, and qualitative dynamical correspondence rather than on a perfectly unified quantitative benchmark. For this reason, the results of the thesis should be interpreted as evidence about the comparative usefulness of two modeling approaches in a controlled experimental setting, rather than as a definitive one-to-one validation of one approach against the other [26, 3, 12].

4 Experimental setup

4.1 Experimental goals

The experiments in this thesis are designed to evaluate the behavior of the RL-based predator-prey simulator and to compare it with the classical ODE baseline under controlled conditions. More specifically, the experimental setup serves three main goals. First, it assesses whether the learned predator and prey policies generate coherent population dynamics over time, rather than only maximizing short-term reward at the individual level. Second, it examines whether these learned dynamics display ecologically meaningful properties such as persistence, oscillation, collapse, or coexistence. Third, it evaluates to what extent the resulting RL population behavior can be related to, approximated by, or contrasted with the dynamics produced by the classical ODE models.

The experiments are therefore not intended merely as a benchmark of predictive accuracy. Instead, they are designed to study the extent to which reinforcement learning can function as a useful ecosystem modeling approach in its own right. In this sense, the goal is twofold: to evaluate the internal quality of the RL simulator as a learning-based ecological system, and to assess its value as a complementary modeling framework relative to the more interpretable but more constrained classical ecological baseline.

At a practical level, this means that the experiments focus on three types of outcomes. The first is learning performance, including whether agents improve their policies over training. The second is ecological performance, including the stability and structure of the resulting predator-prey dynamics during long-horizon evaluation. The third is comparative performance, including the extent to which RL-generated trajectories can be interpreted in relation to fitted Lotka-Volterra-type population dynamics.

4.2 Compared setups

4.2.1 ODE baseline

The ODE baseline used in the experiments consists of two classical predator-prey models: the standard Lotka-Volterra system and the logistic-prey extension corresponding to the Rosenzweig-MacArthur formulation. These models serve as the classical reference setup against which the RL-based simulator is interpreted. Their role in the experimental chapter is not to reproduce the full agent-based dynamics of the simulator, but to provide aggregate population-level benchmarks for trajectory fitting, equilibrium analysis, and dynamical comparison.

Experimentally, the ODE baseline is applied in three contexts. First, both models are fitted to the historical hare-lynx dataset, which serves as the main empirical benchmark for predator-prey oscillatory dynamics. Second, both models are also fitted to the Isle Royale wolf-moose dataset, which is used as a more difficult case, highlighting the limitations of deterministic aggregate models under real ecological perturbations. Third, an RL-calibrated synthetic scenario is constructed by translating selected simulator parameters into approximate ODE coefficients, allowing the aggregate implications of the RL environment design to be examined in classical form.

For each of these cases, the ODE analysis produces fitted trajectories for predator and prey populations, together with phase portraits, nullclines, equilibrium estimates, and sensitivity analyses. In this sense, the ODE baseline provides the experimental reference for classical ecological behavior, against which the population-level dynamics generated by the RL simulator can later be compared.

4.2.2 RL setup

The RL setup used in the experiments consists of a spatial predator-prey simulator in which agents interact in a discrete grid environment and learn behavior through tabular Q-learning. Unlike the ODE baseline, which represents populations at the aggregate level, the RL setup models the system through individual predator and prey agents whose local interactions generate population-level dynamics over time. The experimental role of this setup is to determine whether learned policies can produce coherent ecological behavior, including survival, predation, reproduction, and long-term coexistence.

Experimentally, the RL setup is based on two training configurations. The first is a simplified 1v1 warm-up setting with one predator and one prey, used to support the emergence of basic pursuit and escape behavior under minimal conditions. The second, and primary, configuration is the alternating multi-agent training protocol, in which one focal learner is trained at a time against frozen background populations. This second setup is the main source of the policies later used for long-horizon evaluation and comparison with the ODE baseline.

The RL experiments are conducted entirely in procedurally generated environments rather than on empirical ecological datasets. As a result, the RL setup is evaluated through its internal learning performance and through the ecological structure of the population dynamics it produces during long-horizon rollouts. In this sense, the RL setup functions as the experimental learning-based counterpart to the classical ODE baseline: it is the system through which the thesis studies whether adaptive agent behavior can serve as a useful complementary framework for ecosystem modeling.

4.3 Training and evaluation configuration

4.3.1 Hyperparameters

The RL experiments use a common tabular Q-learning configuration together with a small number of protocol-specific overrides. At the global level, the learning rate is set to $\alpha = 0.1$, the discount factor to $\gamma = 0.95$, the default exploration schedule to $\epsilon_{start} = 1.0, \epsilon_{decay} = 0.995$, and $\epsilon_{min} = 0.01$. The tabular state representation contains 540 discrete states, derived from the six-dimensional observation discretization, and the action space contains 10 actions. The standard episode horizon is 500 steps.

The 1v1 warm-up protocol changes ϵ_{decay} to 0.9985, and a higher ϵ_{min} equal to 0.02. Furthermore, it also adds a protocol-specific predation bonus of 5.0, which is separate from the hunting success bonus stated on the predator's reward function. This protocol also uses a shorter episode horizon of 350 steps. These values are chosen to make early learning in the minimal predator-prey setting less noisy and to encourage faster emergence of basic pursuit and escape behavior.

In the alternating training protocol, the default Q-learning hyperparameters are retained, but two additional exploration settings are used operationally. First, background agents that act with frozen policies use a near-greedy exploration rate of 0.01. Second, when a checkpoint is loaded between cycles, its effective ϵ value is capped at 0.3, since exploration is not serialized with the saved Q-table. This prevents previously learned policies from being reset into full re-exploration at the start of a new cycle.

The experimental population and environment settings are also fixed. The 1v1 protocol is run on a 100×100 grid with one prey and one predator. The main alternating protocol uses a 150×150 grid with 30 prey and 10 predators. Training is typically organized into 350 episodes per phase, with the total number of cycles chosen experimentally depending on the run. These values define the practical scale at which the RL experiments are conducted.

4.3.2 Training settings

Two training configurations are used in the experiments. The first is a simplified 1v1 warm-up protocol, and the second is the main alternating multi-agent protocol. Together, they provide a staged training procedure in which basic predator-prey behavior is first learned in a minimal setting and then transferred to a richer ecological environment.

In the 1v1 warm-up protocol, one prey agent and one predator agent are trained simultaneously on a 100×100 grid with no background population and no reproduction. Each episode ends either when the predator kills the prey or when the maximum horizon of 350 steps is reached. Both agents update their Q-tables after every step, and the turn order alternates across episodes so that neither species systematically benefits from acting first. This protocol is generally run for 1500 episodes and is used to generate an initial checkpoint for the main training stage.

The primary training procedure is the alternating protocol, which uses the full 150×150 environment with 30 prey and 10 predators. Training is organized into cycles, each composed of two phases. In the first phase, the prey agent is the focal learner, and the predator population is frozen. In the second phase, the predator agent becomes the focal learner, and the prey population is frozen. Background agents act during every step but do not update their Q-values. Opponent agents use the frozen policy inherited from the previous phase, while same-species background agents use a frozen copy of the focal learner's policy taken at the start of the phase. This reduces

non-stationarity by ensuring that only one policy changes during a given phase.

Each phase runs for 350 episodes, unless stated otherwise, and the total number of cycles is chosen experimentally depending on the run. Within each episode, the environment is reset with a deterministic seed, actions are executed, ecological updates are applied, reproduction is resolved, and the focal learner performs a Q-update using the full reward returned for that step. This reward includes both immediate action-related terms and the reproduction bonus when reproduction occurs. Training does not use early stopping; each run proceeds until the configured episode budget is exhausted.

This training design combines simplicity and stability. The 1v1 protocol supports the emergence of basic local strategies, while the alternating protocol makes tabular Q-learning more tractable in the larger multi-agent setting by freezing the opponent and background policies within each phase. The result is a training process that does not eliminate non-stationarity, but reduces it enough to produce stable checkpoints for later long-horizon evaluation.

4.3.3 Evaluation procedure

Evaluation is performed at two levels: within-training evaluation and post-training long-horizon evaluation. Together, these two procedures are used to assess both short-term policy quality and the longer-term ecological dynamics generated by the learned predator-prey interactions.

During training, evaluation is carried out in greedy mode with exploration disabled. In the alternating protocol, a short evaluation is performed at the end of each phase by temporarily setting $\epsilon = 0$ and running 10 evaluation episodes without Q-updates. The resulting average reward is logged as a phase-level performance indicator. In the 1v1 warm-up protocol, meaningful greedy evaluation is performed only after training has been completed.

The main experimental evaluation is conducted after training using frozen prey and predator checkpoints. In this stage, both species act greedily with $\epsilon = 0$, and no learning updates are performed. The evaluation is run in the multi-agent environment rather than the single-focal-agent training wrapper, allowing the learned policies to be tested under full population dynamics. Each checkpoint pair is evaluated across three independent random seeds, and each run lasts up to 2000 steps.

During long-horizon evaluation, the simulator records per-step information on predator and prey population sizes, energy distributions, births, deaths, predation events, extinction status, and coexistence score. From these trajectories, a set of summary metrics is derived, including population coefficient of variation, oscillation period, survival fraction, extinction step, demographic rates, and mean individual lifespan. In this way, the evaluation procedure is designed not only to measure immediate policy performance but also to determine whether the learned policies produce stable and ecologically interpretable system-level dynamics over extended horizons.

4.3.4 Checkpoint selection

Checkpoints are saved at the end of each phase of the alternating training protocol, with prey and predator policies stored separately. This means that each cycle produces one prey checkpoint after the prey-training phase and one predator checkpoint after the predator-training phase. The saved files contain the learned Q-table together with basic metadata such as the state and action dimensions, and a small configuration snapshot. The exploration state is not serialized, so

checkpoints preserve learned values but not the current ϵ level.

In experimental practice, checkpoints are not selected through an explicit model-selection criterion such as best evaluation reward or best coexistence score. Instead, the final evaluation uses the most recent completed checkpoint available for each species. In this sense, checkpoint selection is based on the training stage rather than on post-hoc optimization over many saved models. This keeps the evaluation procedure simple and consistent with the staged training design.

Checkpoints also play an operational role during training. In the alternating protocol, the policy learned in one phase is carried forward as the frozen opponent in the next phase. When checkpoints are reloaded between training stages, the effective exploration rate is manually controlled, since the saved files do not store ϵ. As a result, checkpoints function both as final experimental artifacts for evaluation and as intermediate policy carriers within the alternating training process.

4.4 Data collection and logging

The experimental pipeline records data at both training time and evaluation time. During training, episode-level logs are used to monitor learning progress. These include cumulative reward, episode outcome, and protocol-specific evaluation summaries. In the alternating protocol, greedy evaluation scores are also recorded at phase boundaries, providing a compact indication of policy quality without exploration noise.

For post-training evaluation, data collection is more detailed. Long-horizon runs record per-step information on prey and predator population sizes, coexistence score, energy distribution statistics, cumulative births, cumulative deaths, and predation events. These per-step records are written to CSV files and later aggregated into per-seed summary tables. From the logged trajectories, additional derived quantities are computed, including population coefficient of variation, oscillation period, survival fraction, extinction step, demographic rates, and mean individual lifespan.

Model checkpoints are also part of the experimental record. Prey and predator Q-tables are saved separately at the end of each training phase, allowing policies to be reused, evaluated, and compared across cycles. Taken together, the logging system supports three complementary functions: monitoring training behavior, evaluating long-horizon ecological dynamics, and preserving intermediate and final learned policies for later analysis.

4.5 Reproducibility details

The experimental pipeline is designed to be reproducible through fixed configuration values, deterministic seeding, and explicit checkpointing. Training episodes use deterministic seed formulas that vary systematically across runs, cycles, phases, and episodes, so that each episode is distinct while remaining reproducible. Evaluation is also performed under controlled seeds, with long-horizon rollouts repeated across three independent seed values.

Reproducibility is further supported through serialized checkpoints. Prey and predator policies are saved separately as pickle files containing the learned Q-tables, state and action dimensions, and a configuration snapshot. These checkpoints allow training to be resumed, evaluation to be repeated, and trained policy pairs to be reused consistently across experiments.

At the same time, some implementation details place limits on strict reproducibility. In particular, the exploration parameter ϵ is not serialized with saved checkpoints and must therefore be controlled manually when models are reloaded. In addition, training and long-horizon evaluation are run

in different environment wrappers, which means that policies are reproduced across experiments but not always under identical step-order semantics. For this reason, reproducibility in this thesis should be understood primarily as reproducibility of the experimental pipeline and logged outputs, rather than as exact identity of all internal simulator conditions across every stage.

5 Results

5.1 ODE baseline results

5.1.1 Historical case-study fits

The ODE baseline results serve two complementary purposes in this thesis. First, they show how well classical predator-prey models can reproduce known ecological time series under fitting conditions. Second, they provide an analytical reference point for interpreting the behavior of the RL-based simulator. To this end, two classical models were analyzed throughout: the standard Lotka-Volterra (LV) system and the Rosenzweig-MacArthur-style logistic-prey model (RM). These models were fitted to two historical predator-prey datasets, and the RM model was additionally parameterized from the RL environment configuration to obtain an analytical baseline under simulator-inspired assumptions.

The first part of the ODE analysis examined the behavior of the two classical models on historical ecological data. Two case studies were used: the Hudson Bay hare-lynx series and the Isle Royale moose-wolf series. These datasets play different roles in the analysis. The Hare-Lynx system provides a relatively clean benchmark for oscillatory predator-prey behavior, whereas Isle Royale acts as a more difficult test case in which exogenous ecological shocks and structural changes are known to play an important role.

For the Hare-Lynx data, both models captured the overall cyclical structure of the observed dynamics well. However, the Lotka-Volterra model consistently performed slightly better than the logistic-prey model. Its prey RMSE was 5.20, compared with 5.63 for the RM model, and its predator RMSE was 3.32, compared with 3.48 for RM. This ranking was also reflected in the AIC values: LV achieved 84.44, whereas RM achieved 89.37. Since the RM model introduces one additional free parameter through the carrying capacity K, the lower AIC of LV indicates that the extra flexibility of the logistic term did not translate into a better fit in this case. In practice, the fitted RM carrying capacity reached the optimization boundary ($K \approx 600$), indicating that the logistic term contributed little and that the fitted RM dynamics were close to the unconstrained LV formulation. This is an important result in itself: for the Hare-Lynx series, predator-prey interaction appears to dominate density limitation within the fitted time window.

For Isle Royale, the picture was very different. Both models fitted the wolf trajectory only moderately well, but neither model reproduced the moose series satisfactorily. The LV model yielded a prey RMSE of 336.91 and a predator RMSE of 7.63, while the RM model yielded a prey RMSE of 342.95 and a predator RMSE of 7.10. Again, LV was preferred by AIC (624.93 versus 628.81), despite RM having the additional prey carrying-capacity term. As in the Hare-Lynx case, the fitted RM carrying capacity was pushed to the search boundary ($K \approx 5000$), which again implies that the logistic term was effectively inactive over the fitted range. The poor moose fit is not merely numerical noise, but reflects a structural mismatch between the models and the system. In particular, the historical collapse of the wolf population around the early 1980s, followed by long

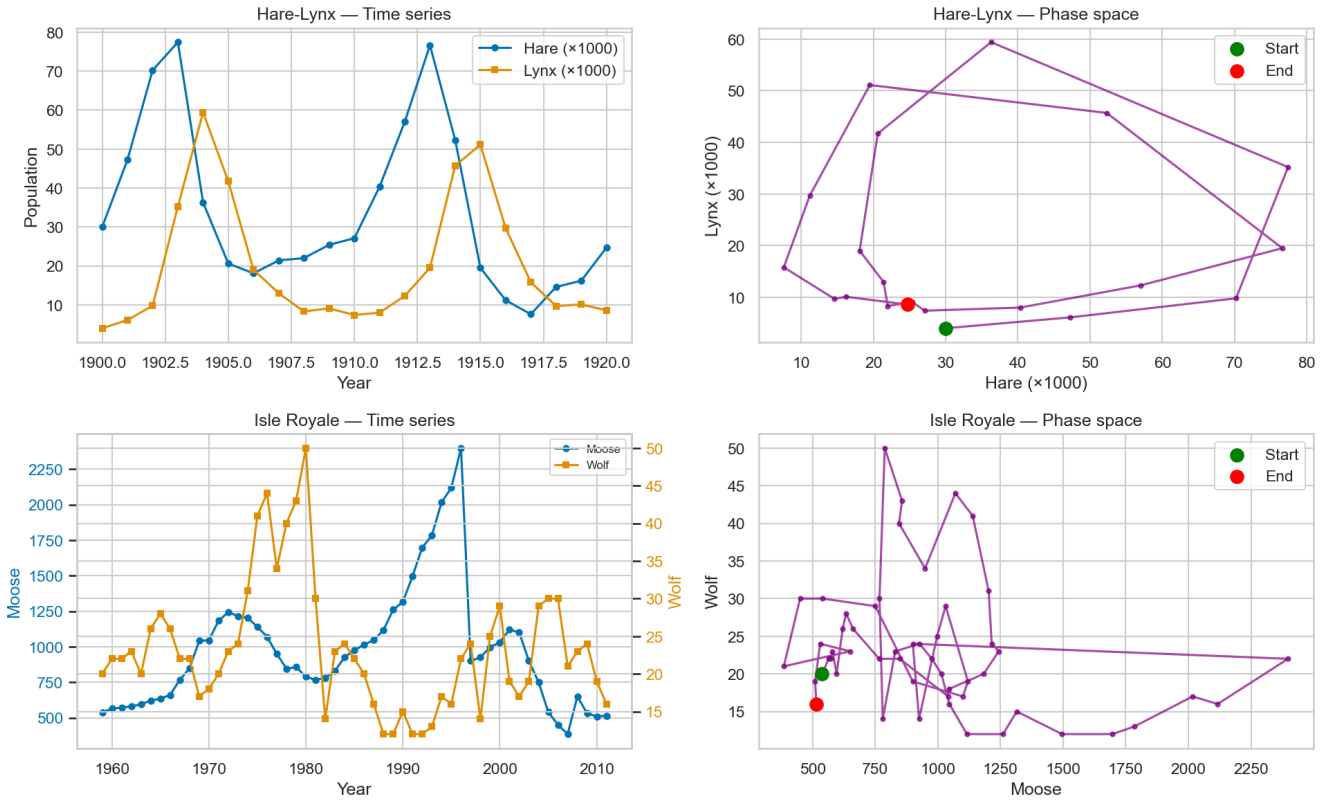

Figure 1: Historical datasets used in the ODE baseline analysis. The figure shows the raw time series and corresponding phase-space trajectories for the Hare-Lynx and Isle Royale systems.

recovery dynamics, cannot be represented adequately by a two-dimensional deterministic ODE with fixed coefficients. The Isle Royale results, therefore, serve less as a successful model fit and more as evidence of the limitations of classical aggregate models when ecological dynamics are shaped by strong exogenous disruptions.

These fit differences are summarized in Table 1. The table shows that, in both datasets, the simpler Lotka-Volterra model slightly outperformed the logistic-prey extension according to both fit quality and information criteria. This does not mean that LV is a better ecological model in general, but it does indicate that, within the fitted windows used here, the additional RM structure did not provide enough explanatory gain to justify its extra parameter.

Table 1: Model comparison for the two historical case studies. Lower RMSE and lower AIC indicate a better fit.

Dataset	Model	k	RMSE prey	RMSE pred	AIC
Hare-Lynx	Lotka-Volterra	4	5.205	3.318	84.443
Hare-Lynx	Rosenzweig-MacArthur	5	5.627	3.484	89.373
Isle Royale	Lotka-Volterra	4	336.913	7.635	624.929
Isle Royale	Rosenzweig-MacArthur	5	342.952	7.096	628.807

Beyond fit quality, the two models also differ structurally in their local stability properties. For both case studies, the Lotka-Volterra equilibrium is a neutral centre, with purely imaginary

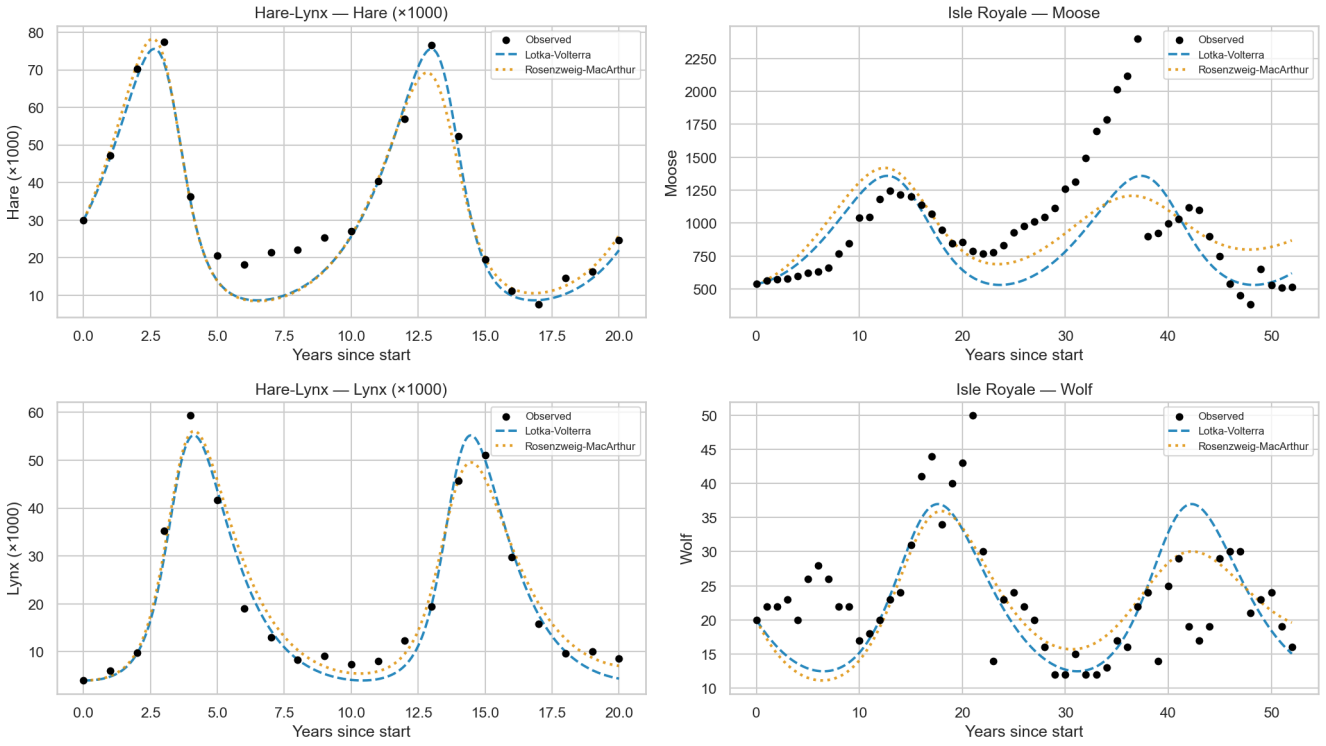

Figure 2: Observed population trajectories and fitted ODE trajectories for the Hare-Lynx and Isle Royale case studies. The LV model slightly outperforms the RM model in both datasets according to RMSE and AIC.

eigenvalues, implying perpetual oscillations without convergence. By contrast, the RM equilibrium is a stable spiral in both datasets, with small but negative real parts. For Hare-Lynx, the fitted RM equilibrium is approximately $(N^*, P^*) = (30.2, 21.0)$, with eigenvalues $-0.016 \pm 0.671i$. For Isle Royale, the fitted RM equilibrium is approximately $(N^*, P^*) = (961.6, 23.2)$, with eigenvalues $-0.024 \pm 0.262i$. The negative real parts are small, which means that convergence is slow and oscillations remain visible over many cycles. This distinction matters for later comparison with the RL simulator, because it implies that even in the classical model, long horizons are needed to distinguish transient oscillations from genuine long-run stabilization.

Table 2: Stability summary at the coexistence equilibrium for the fitted historical models.

Dataset	Model	N^*	P^*	Stability
Hare-Lynx	Lotka-Volterra	30.841	19.501	Neutral centre
Hare-Lynx	Rosenzweig-MacArthur	30.178	21.048	Stable spiral
Isle Royale	Lotka-Volterra	880.968	22.543	Neutral centre
Isle Royale	Rosenzweig-MacArthur	961.601	23.185	Stable spiral

The sensitivity analysis of the RM model provides an additional perspective on which parameters most strongly determine equilibrium structure. Across both datasets, the strongest effects are associated with the attack rate a, the predator conversion efficiency e, and the predator mortality rate d. Increasing either a or e by 10% reduces equilibrium prey density N^* by roughly 9.1% in both

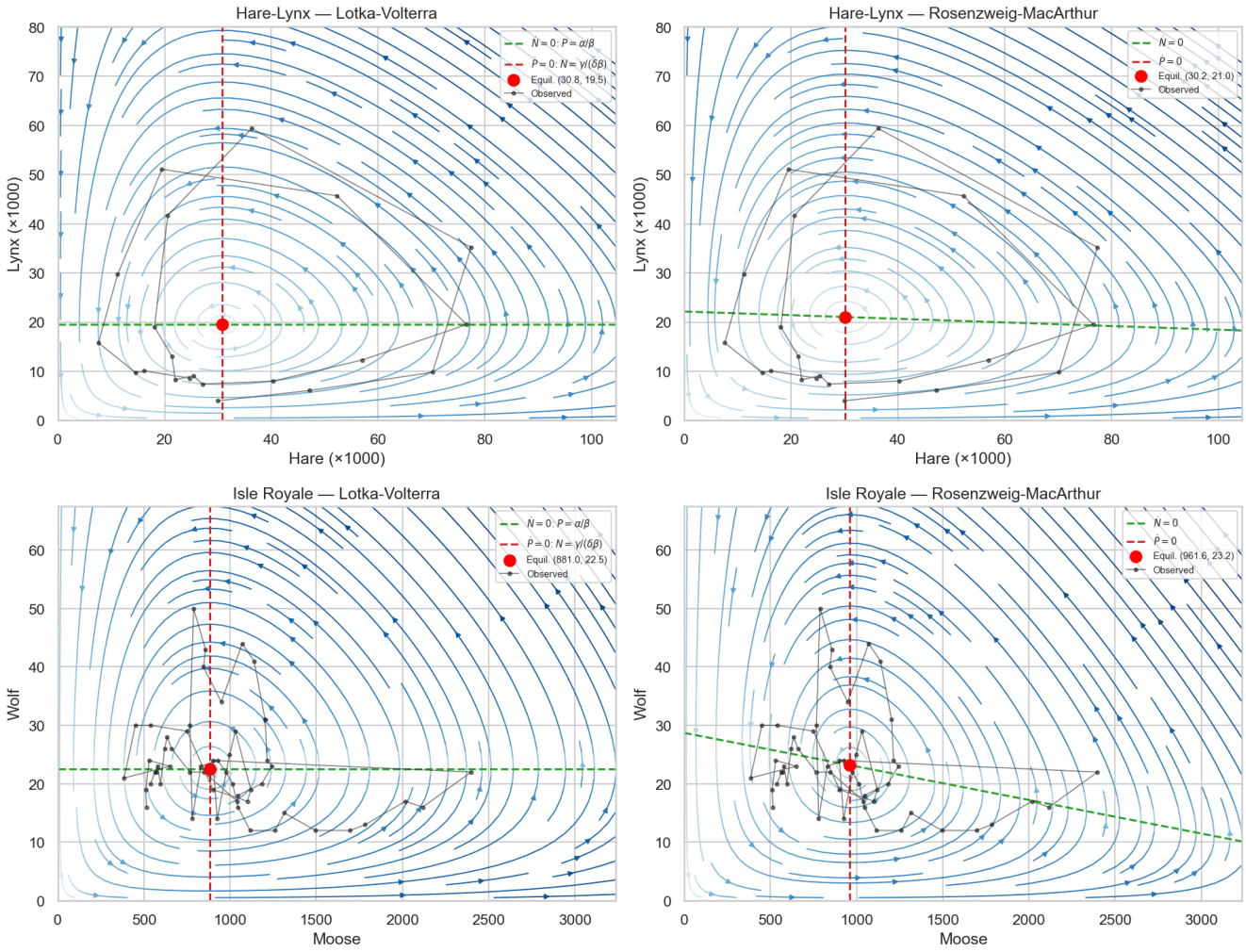

Figure 3: Phase portraits, nullclines, and coexistence equilibria for the fitted LV and RM models. The LV equilibrium is a neutral centre, whereas the RM equilibrium is a stable spiral.

case studies, whereas increasing d by 10% raises N^* by 10%. By contrast, the carrying capacity K has almost no effect on equilibrium prey density and only a small effect on equilibrium predator density. This follows directly from the equilibrium structure of the RM model: prey equilibrium is set primarily by predation pressure and predator mortality rather than by the environmental carrying capacity itself. In the context of this thesis, this result is especially relevant because it indicates that the balance between predation efficiency and predator maintenance cost is the main structural driver of predator viability, both in the fitted ODE models and in the later RL-calibrated analysis.

Taken together, the historical case-study fits show that classical ODEs can reproduce predator-prey oscillations reasonably well in stationary or approximately stationary settings, but struggle when ecological history contains strong external perturbations. The Hare-Lynx series, therefore, validates the use of ODEs as a meaningful classical reference, while the Isle Royale series reveals the limits of that reference. This contrast is methodologically useful for the rest of the thesis: it shows both why the ODE baseline is worth comparing against, and why a more adaptive

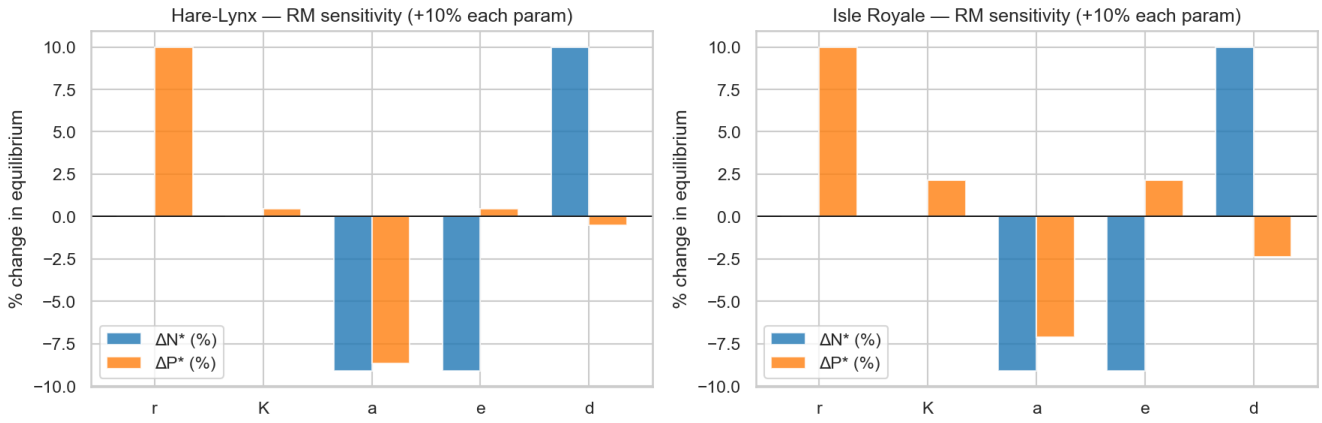

Figure 4: Sensitivity of the RM equilibrium to a +10% increase in each parameter. The dominant effects are associated with attack rate, conversion efficiency, and predator mortality.

Table 3: Sensitivity of the Rosenzweig-MacArthur equilibrium to a +10% increase in each parameter.

Dataset	Parameter	ΔN^* (%)	ΔP^* (%)
Hare-Lynx	r	0.0	+10.0
Hare-Lynx	K	0.0	+0.5
Hare-Lynx	a	-9.1	-8.7
Hare-Lynx	e	-9.1	+0.5
Hare-Lynx	d	+10.0	-0.5
Isle Royale	r	0.0	+10.0
Isle Royale	K	0.0	+2.2
Isle Royale	a	-9.1	-7.1
Isle Royale	e	-9.1	+2.2
Isle Royale	d	+10.0	-2.4

simulation framework may be needed when ecological dynamics move beyond the assumptions of fixed-parameter aggregate models.

5.1.2 RL-calibrated ODE scenario

In addition to the historical fits, a Rosenzweig-MacArthur model was parameterized directly from the simulator configuration in order to obtain an analytical baseline for the RL environment itself. This RL-calibrated ODE scenario is not intended as a precise ecological fit, but as an order-of-magnitude translation of the simulator's structural assumptions into a classical aggregate form. In particular, it provides a way to ask what the current RL parameter regime would imply if interpreted through a deterministic predator-prey model.

Under the current simulator configuration, the calibrated parameters are $r = 0.4$, $K = 337$, $a = 0.2793$, $e = 0.1$, and $d = 0.7059$. These values produce a coexistence equilibrium at approximately $(N^*, P^*) = (25.3, 1.3)$, with eigenvalues $-0.015 \pm 0.511i$, indicating a stable spiral with slow convergence. The forward simulation was initialized at the current RL-inspired population split $(30, 10)$, which places the predator population far above the equilibrium level. As a result, the

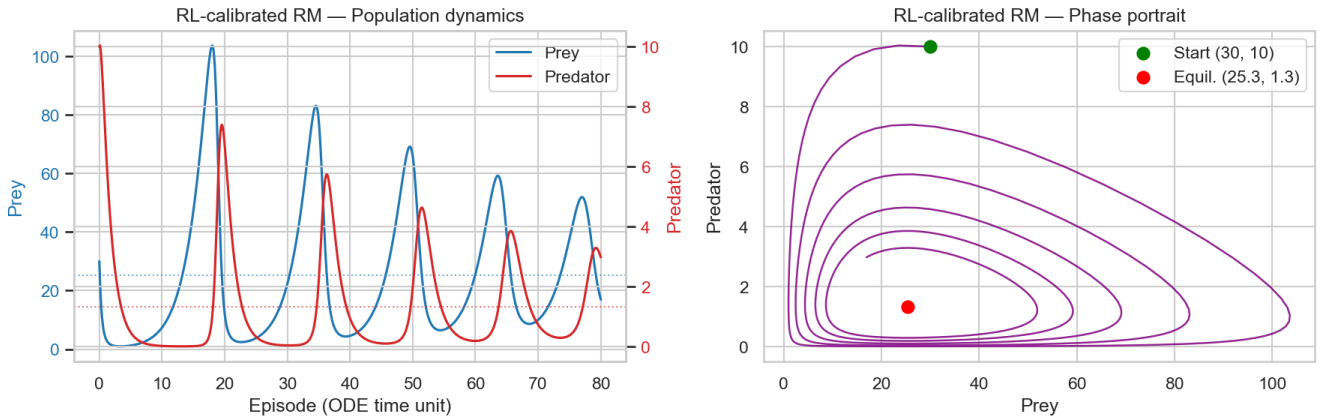

Figure 5: Forward simulation of the RL-calibrated RM model from the initial population split $(30, 10)$. The trajectory spirals toward a low-predator equilibrium, indicating structurally weak predator persistence under the calibrated parameter regime.

ODE trajectory shows a large predator overshoot, a sharp prey decline, and then slowly damped oscillations spiraling toward a state with low predator abundance. This behavior is exactly what would be expected from the RM model when a system starts far above its long-run predator equilibrium.

Table 4: RL-calibrated Rosenzweig-MacArthur parameter set used as an analytical baseline for the simulator.

Parameter	Value	Interpretation
r	0.4000	Prey intrinsic growth rate
K	337.0	Prey carrying capacity
a	0.2793	Attack rate
e	0.1000	Predator conversion efficiency
d	0.7059	Predator mortality rate

At the same time, this translation must be interpreted carefully. The attack-rate parameter a is derived geometrically from simulator settings and therefore represents an upper-bound approximation rather than an empirically validated encounter rate. In effect, it assumes idealized hunting efficiency under uniform mixing assumptions, whereas in the actual simulator, predators do not successfully hunt at every step, prey are not uniformly distributed, and encounter opportunities depend strongly on learned behavior and spatial clustering. The calibrated ODE should therefore not be read as an exact macro-level equivalent of the simulator, but as a coarse analytical baseline under optimistic predation assumptions.

Even with this caveat, the key structural result is robust: the calibrated ODE predicts that the current parameter regime is unfavorable to persistent predator abundance. A predator equilibrium of only $P^* \approx 1.3$ is very small relative to the initialized predator population of 10. In biological terms, the model indicates that predators must operate under very tight energetic constraints. Predator mortality is high relative to energy conversion from prey, so predators can only remain viable if effective predation pressure is sustained at a sufficiently high level. If it is not, the system

is drawn toward a regime in which prey remain present, but predator numbers collapse toward near-extinction levels. This is the most important analytical result of the RL-calibrated ODE scenario, because it generates a clear structural prediction before any learned RL dynamics are considered.

This makes the RL-calibrated baseline especially useful for interpreting the simulator results later in the thesis. The ODE does not explain how predator and prey agents behave locally, but it does indicate what broad population outcome should be expected if the simulator's energetic and interaction coefficients dominate the dynamics in the way assumed by the aggregate model. In that sense, it functions as a structural benchmark: if the later RL experiments also show low predator persistence or frequent predator collapse, that outcome is not merely a learning failure, but a behavior that is already consistent with the underlying system balance implied by the analytical baseline.

5.2 RL training results

5.2.1 1v1 warm-up training

The first stage of the RL training process consisted of a simplified 1v1 warm-up protocol in which one predator and one prey agent were trained against each other in isolation. The purpose of this stage was not to obtain a final ecological model, but to bootstrap basic predator-prey competencies before moving to the more complex alternating multi-agent setting. In this sense, the protocol functioned as a controlled pretraining phase in which the agents could learn simple pursuit and evasion behavior without the added complexity of reproduction, background populations, or large-scale ecological feedback.

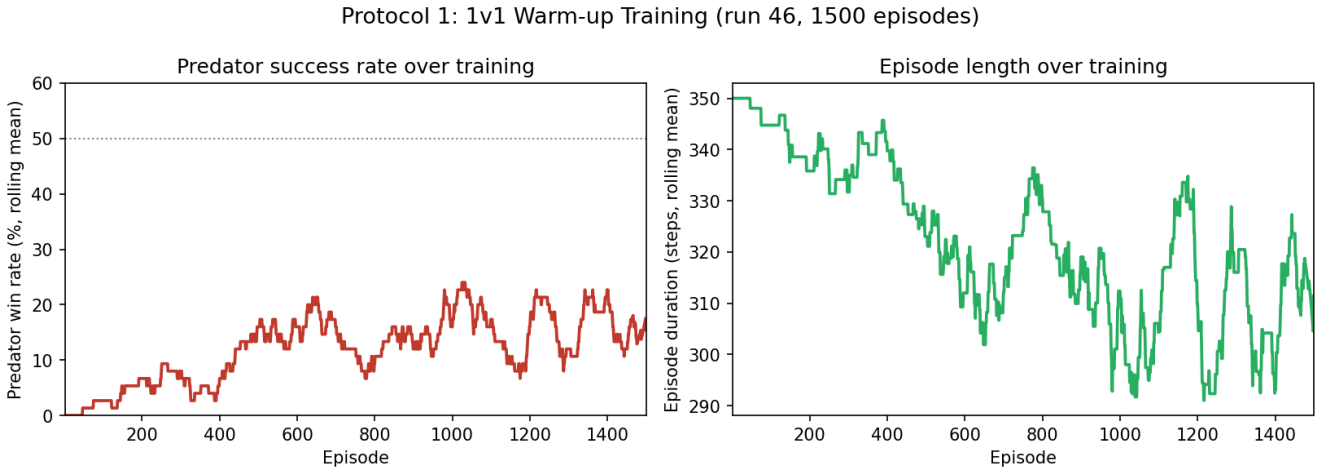

Figure 6: Protocol 1 warm-up training results for run 46 over 1500 episodes. Left: rolling predator win rate (window 75 episodes), rising from near zero to approximately 15–20% by the final episodes. The dotted horizontal line marks the 50% baseline for reference. Right: mean episode duration over training, showing a gradual decrease as the predator improves its ability to terminate encounters.

The training curves show that this warm-up stage produced limited but meaningful learning progress. As shown in Figure 6, the predator's win rate increased from almost 0% at the beginning

of training to roughly 15 - 20% by the later episodes. This indicated that the predator learned to pursue and occasionally catch the prey, but never became dominant. At the same time, the prey continued to survive in the large majority of encounters, which is consistent with the intended role of the protocol: to create agents with basic local competence rather than to solve the predator-prey problem at this stage.

A similar pattern appears in the episode-length curve. Figure 6 shows a gradual decrease in average episode duration over training, from values close to the 350-step horizon toward lower but still relatively high levels. This suggests that both agents became more decisive over time. The predator improved its ability to convert pursuit into successful captures, while the prey learned evasion strategies that reduced fully random wandering but still allowed survival in most episodes.

Taken together, these results show that Protocol 1 succeeded as a warm-up stage. It did not produce a strong predator policy on its own but, instead, produced agents with basic predator-prey competencies; the predator learned to hunt occasionally, and the prey learned to avoid capture often enough to remain viable. This made the resulting checkpoints suitable starting points for the subsequent alternating training protocol, where these local behaviors could be refined in a richer ecological setting.

5.2.2 Alternating training protocol

The second training stage used the alternating multi-agent protocol and constitutes the main learning process of the thesis. Unlike the 1v1 stage, this protocol was designed to expose the agents to a richer ecological setting with background populations, reproduction, and larger-scale predator-prey interactions. Its purpose was therefore not only to refine local pursuit and evasion behavior, but also to test whether such behavior could be stabilized into meaningful population dynamics.

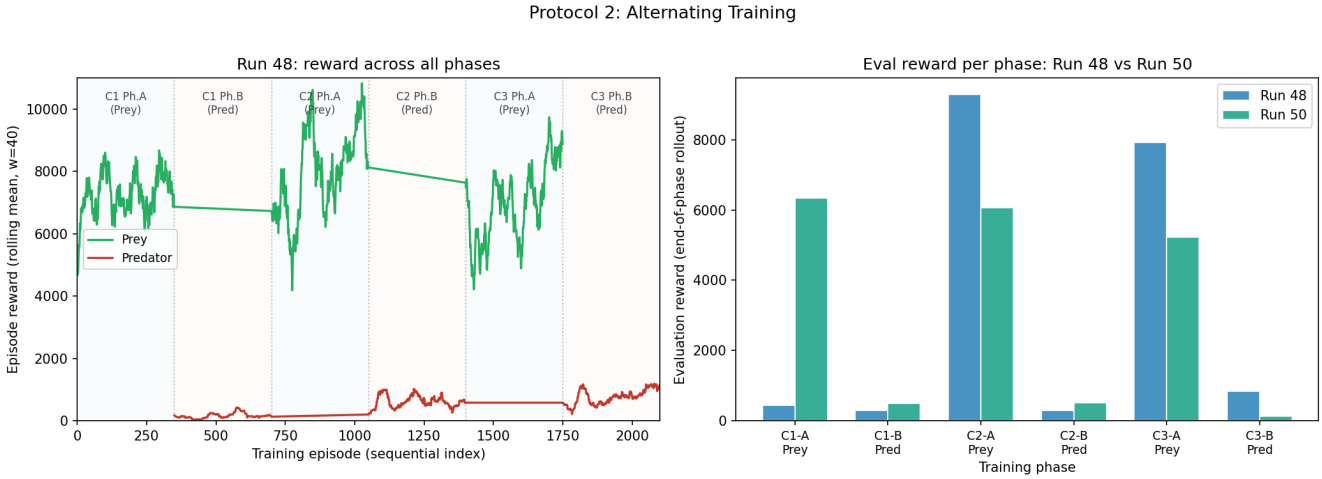

Figure 7: Protocol 2 alternating training results. Left: rolling episode reward (window 40 episodes) for prey (green) and predator (red) across all six phases of run 48, with dotted vertical separators at phase boundaries and alternating shaded bands labeling each phase. Prey reward is consistently higher than predator reward throughout training, with a pronounced increase in cycle 2. Right: end-of-phase evaluation reward for runs 48 and 50, showing strong prey-side performance in both runs but a sharp predator reward drop in the final predator phase of run 50.

The training curves reveal a clear and persistent asymmetry between prey and predator learning. As shown in Figure 7, prey reward remained consistently much higher than predator reward throughout the alternating phases of all runs, often by roughly one order of magnitude more. This indicates that prey policies were generally easier to optimize under the current reward and state design, whereas predator learning remained more difficult and less stable. In particular, the prey phase of cycle 2 shows a marked increase in reward.

The same figure also shows that predator improvement was delayed and much weaker. Predator reward remains comparatively low throughout most of the alternating process and only rises meaningfully in the later predator phase of cycle 3. This delayed increase is significant because it indicates that predator competence did improve, but only after the prey policy had already become relatively strong and stable. In practical terms, this means that the alternating protocol did not produce balanced co-adaptation at the same pace for both species. Instead, prey learning advanced earlier, while predator learning lagged behind and only caught up partially in later phases.

A similar pattern appears in the phase-level evaluation rewards shown for runs 48 and 50. Figure 7 shows that run 50 maintained relatively strong prey-side evaluation performance across phases, while predator evaluation dropped sharply in the final predator phase. This result is consistent with the later ecological analysis. Although the final run produced strong coexistence metrics at the population level, the underlying predator policy was not uniformly robust in a general ecological sense. Rather, it converged toward a strategy that functioned well under the specific spatial structure of the environment.

Taken together, these results show that Protocol 2 was essential for generating the later population-level regimes, but they also reveal an important imbalance in the learning process. The alternating design succeeded in producing continued adaptation after the warm-up stage. At the same time, the training curves already suggest that the predator-prey system was not converging towards a single balanced ecological situation. Instead, the protocol produced increasingly specialized strategies whose strengths and weaknesses became visible only during long-horizon evaluation.

5.3 Long-horizon RL population dynamics

5.3.1 Run 48 cycle 2: fragile coexistence

Run 48 cycle 2 was the first checkpoint in which the alternating training protocol produced sustained coexistence in a substantial fraction of seeds, but this coexistence remained highly unstable. Across the 50-seed evaluation, the checkpoint achieved a mean coexistence score of 0.381 ± 0.391, a survival fraction of 0.836, and 31 out of 50 seeds with both species still alive at step 2000. At the same time, the large standard deviation, which exceeds the mean itself, indicates a strongly bimodal outcome distribution: in some seeds, the model produced healthy coexistence, whereas in many others, the predator collapsed almost completely. This is why the regime is best described as fragile rather than robust.

The population dynamics of a representative successful seed illustrate why this checkpoint initially appears promising. As shown in Figure 8, predator and prey populations can settle into relatively stable oscillatory dynamics, with both species remaining well above extinction levels for the full 2000-step horizon. This is consistent with the best seeds of the run, in which coexistence scores reached values above 0.9. However, these successful trajectories are not representative of the checkpoint as a whole, since 19 out of 50 seeds ended with predator extinction. The main result of

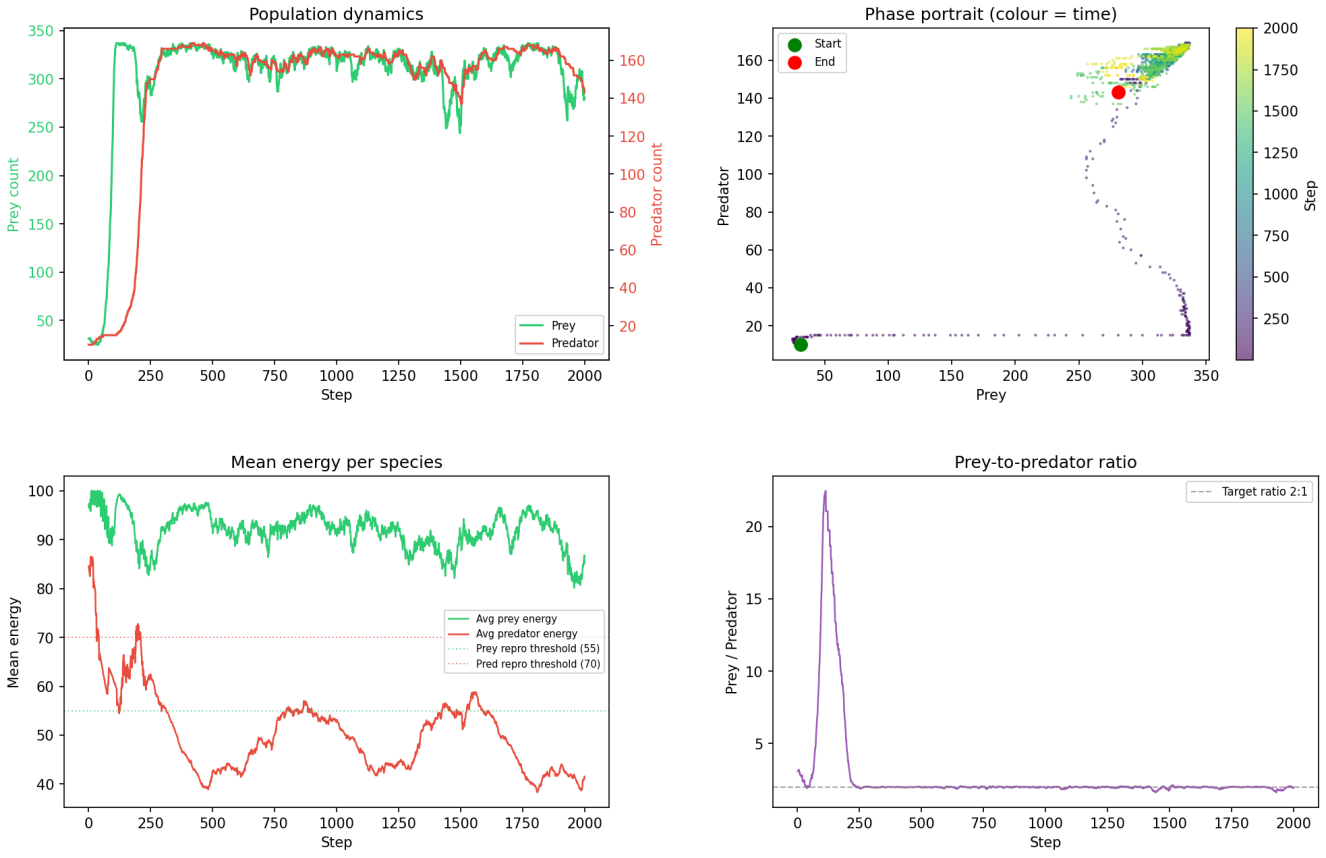

Figure 8: Population dynamics for a representative successful seed of run 48 cycle 2 over 2000 steps. Both prey (blue) and predator (red) remain well above extinction levels throughout the rollout, producing broadly stable oscillatory dynamics. This seed belongs to the 31 out of 50 that retained both species at step 2000; the remaining 19 seeds ended in predator extinction.

run 48 cycle 2 is therefore not stable coexistence itself, but the fact that coexistence had become possible without yet becoming reliable.

The density maps in Figure 9 further show that this coexistence did not yet rely on the strong corner-trapping pattern that later became dominant. Instead, prey and predator activity remained more spatially distributed across the grid, although still unevenly concentrated. Predation events occurred across a broad region of the habitat rather than at a single zone. This makes run 48 cycle 2 important in the overall training story: it represents an intermediate regime in which ecologically plausible interaction patterns begin to emerge, but the learned policies are still too brittle to sustain coexistence consistently across initial conditions.

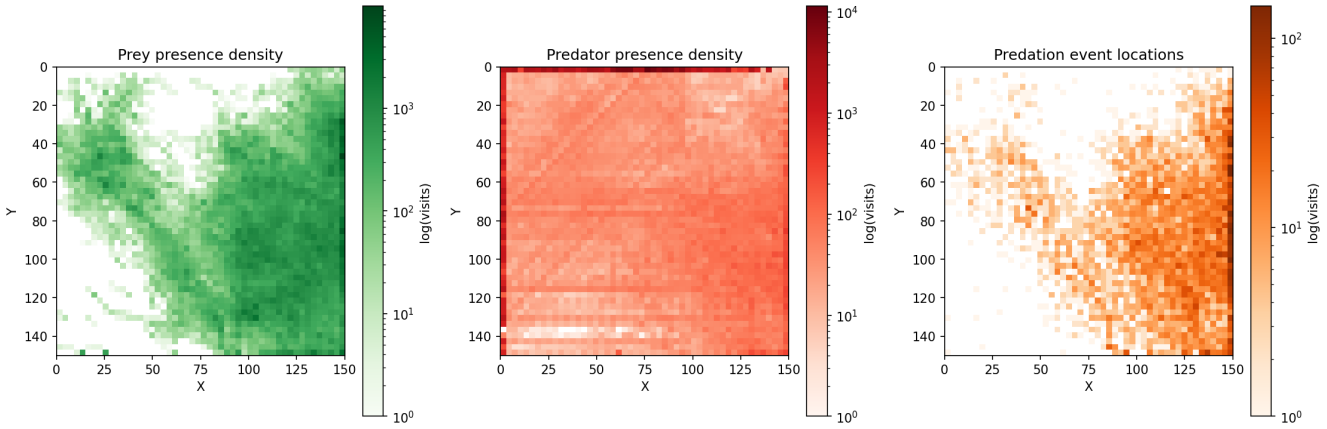

Figure 9: Spatial density maps for a successful seed of run 48 cycle 2, showing accumulated prey presence (left), predator presence (center), and predation events (right) across the 150×150 grid. Activity is broadly distributed across the habitat with no strong boundary concentration, distinguishing this checkpoint from the corner-trapping pattern that emerges in run 50 cycle 3.

To sum up, run 48 cycle 2 shows that the alternating training protocol was capable of producing coexistence-like dynamics, but only in an unstable and seed-dependent way. The checkpoint, therefore, marks an important transition in learning, yet it does not provide a satisfactory ecological solution. Instead, it reveals the first side of the central trade-off observed in the thesis: local predator-prey behavior can become meaningful before ecosystem-level stability becomes robust.

5.3.2 Run 48 cycle 3: ecologically meaningful overhunting

Run 48 cycle 3 produced the most ecologically meaningful local predator behavior observed in the experiments, but failed to maintain viable coexistence at the ecosystem level. Across the 50-seed evaluation, this checkpoint achieved a mean coexistence score of 0.370 ± 0.241 and a survival fraction of 0.716. However, only 4 out of 50 seeds retained both species at step 2000, while the remaining runs were dominated by prey extinction, predator extinction, or mutual collapse. This makes run 48 cycle 3 important, not because it solved the predator-prey problem, but because it reveals a different failure mode from run 48 cycle 2; here, the predator became too effective rather than too fragile.

The representative population trajectory in Figure 10 shows this pattern clearly. After the initial expansion of both species, predator abundance rises to a level that the prey population cannot sustain. Although the prey briefly recovers at intermediate points, it eventually collapses to extinction near the end of the run, after which the predator population also begins to decline. This is consistent with the broader seed-level statistics: 8 out of 50 seeds ended with predators surviving while prey had already gone extinct, and 25 out of 50 ended with both species extinct before the horizon. The main ecological problem in this regime is therefore overhunting rather than under-learning.

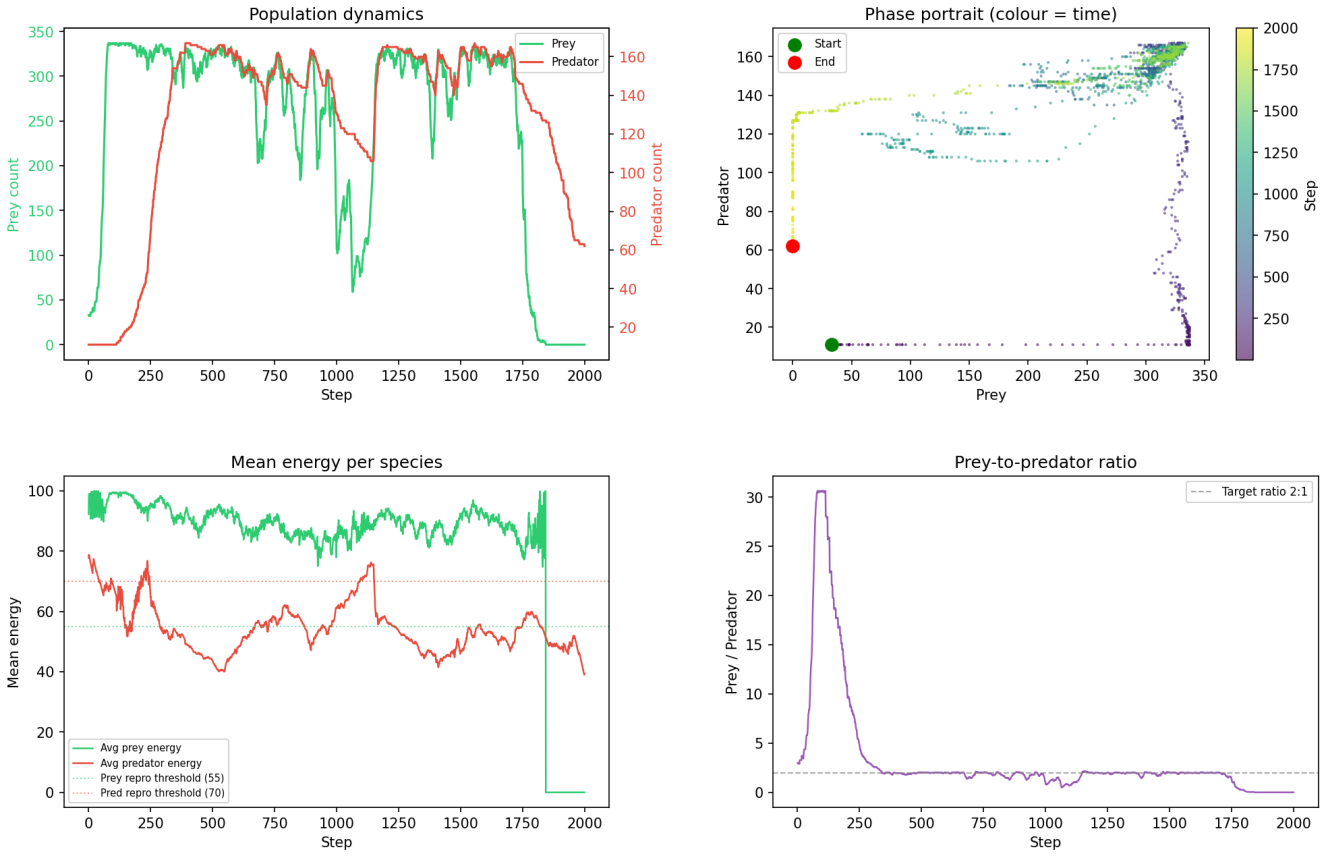

Figure 10: Population dynamics for a representative seed of run 48 cycle 3 over 2000 steps. After an initial expansion phase, predator abundance rises to a level that the prey population cannot sustain. Prey collapses to extinction, after which predator numbers also decline. This overhunting pattern characterizes the majority of the 50 evaluated seeds, with only 4 retaining both species at step 2000.

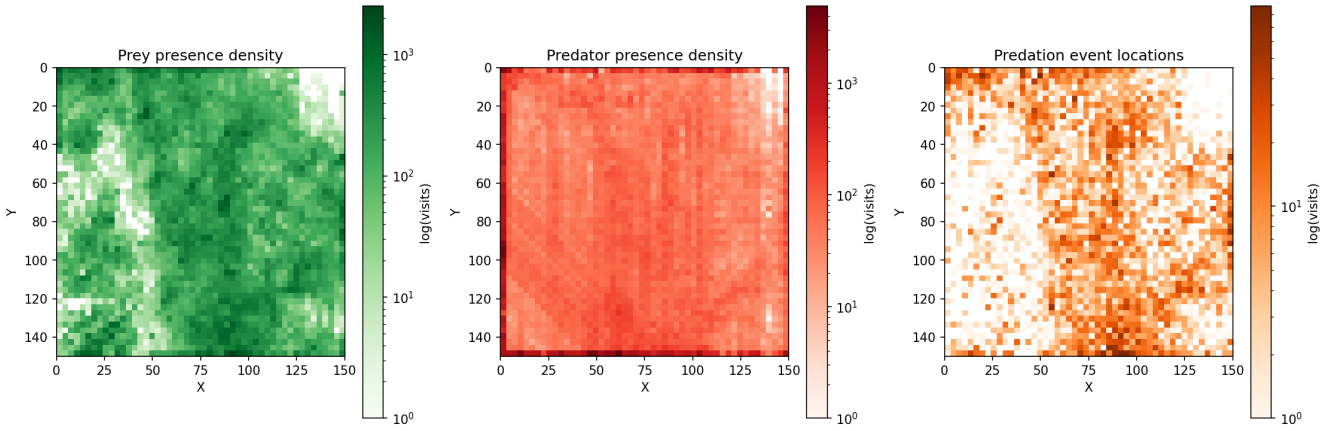

Figure 11: Spatial density maps for a representative seed of run 48 cycle 3, showing accumulated prey presence (left), predator presence (center), and predation events (right) across the 150×150 grid. Unlike run 50 cycle 3, activity is broadly distributed across the habitat with no corner concentration. This distributed hunting pattern is the most ecologically plausible spatial behavior observed across all evaluated checkpoints.

The density maps provide the key reason why this checkpoint remains important despite its poor coexistence metrics. Unlike the later run 50 cycle 3 solution, the activity in Figure 11 is not concentrated in a single corner or narrow boundary zone. Prey presence remains broadly distributed across the habitat, predator movement covers much of the grid, and predation events occur over a wide spatial area. In other words, the agents learned a more distributed and ecologically recognizable interaction pattern: predators actively tracked prey across the environment rather than exploiting a fixed geometric artifact. This makes run 48 cycle 3 the strongest evidence in the thesis that meaningful local predator-prey behavior did emerge under tabular Q-learning.

At the same time, that local behavioral quality did not translate into a stable ecological outcome. The predator policy became sufficiently effective to destabilize the prey population, but the state representation remained too limited to encode broader population consequences of sustained overpredation. Run 48 cycle 3, therefore, illustrates the second side of the central trade-off in the thesis: when the agents learned more ecologically plausible spatial behavior, they lost the capacity to maintain robust coexistence.

5.3.3 Run 50 cycle 3: best aggregate coexistence, model overfitting

Run 50 cycle 3 achieved the strongest aggregate performance of all evaluated checkpoints, but it did so through a spatially degenerate strategy. Across 50 evaluation seeds, it obtained a mean coexistence score of 0.782 ± 0.239, a survival fraction of 0.984, and 48 out of 50 seeds with both species still alive at step 2000. It also showed the lowest variability among the three main regimes, with mean prey and predator coefficients of variation of 0.173 and 0.398, respectively, together with a characteristic oscillation period of approximately 8 - 9 steps. In purely numerical terms, this was the best-performing model produced by the training pipeline.

At the population level, the checkpoint appears highly successful. As shown in Figure 12, the representative best-case seed produces stable long-horizon coexistence, with both species remaining well above extinction levels and the prey-to-predator ratio staying close to the target value of 2:1 for most of the rollout. This is consistent with the best seeds of the run, most notably seed 9, which achieved a coexistence score of 0.949 and never dropped below 0.5 at any point during the full 2000-step horizon. On the basis of population counts alone, run 50 cycle 3 would be the natural candidate for a final model.

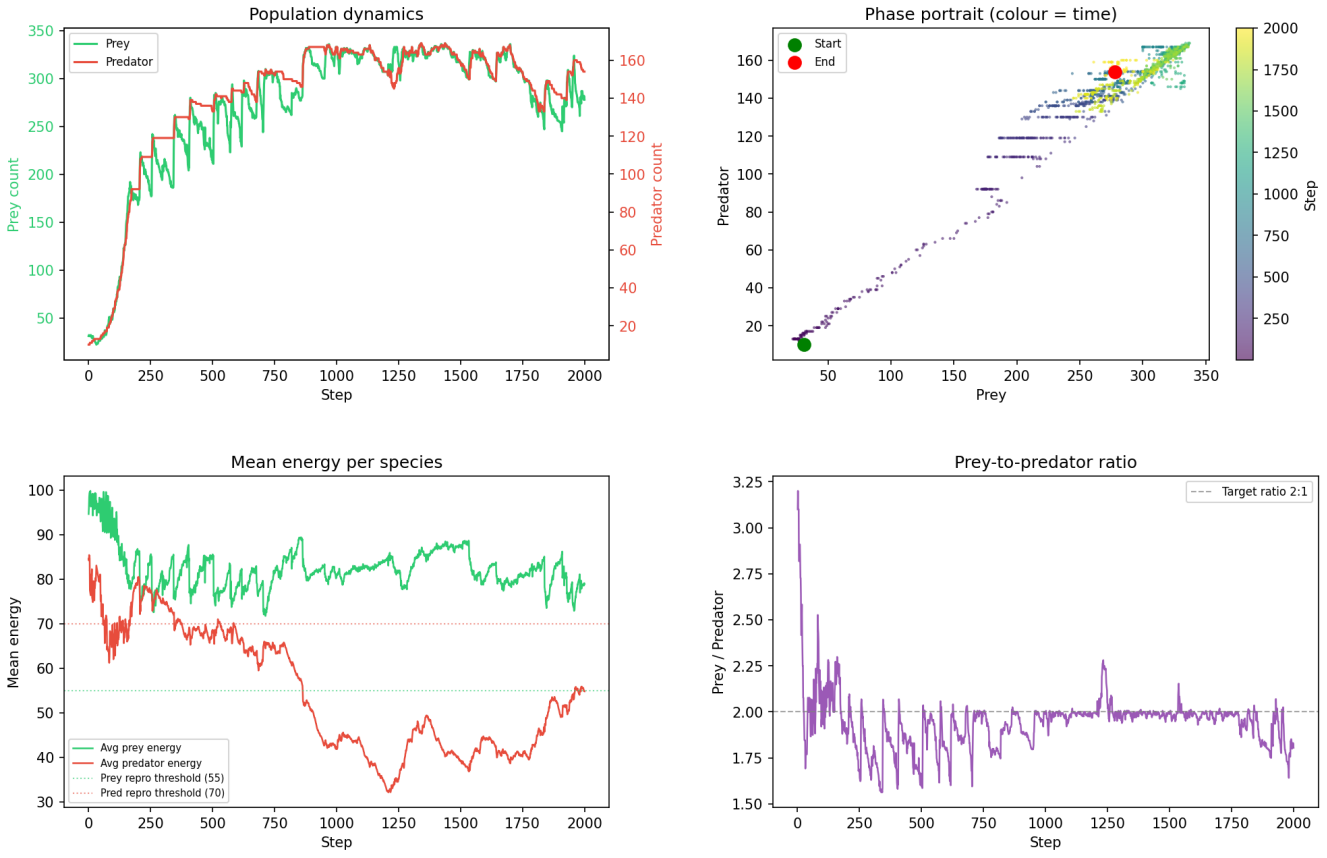

Figure 12: Population dynamics for the best-performing seed of run 50 cycle 3 (random seed 51) over 2000 steps. Both prey (blue) and predator (red) remain well above extinction levels for the full horizon. The coexistence score for this seed is 0.949, the highest observed across all evaluated checkpoints and seeds. The prey-to-predator ratio remains close to the target value of 2:1 for most of the rollout, and cumulative predation events number 3858.

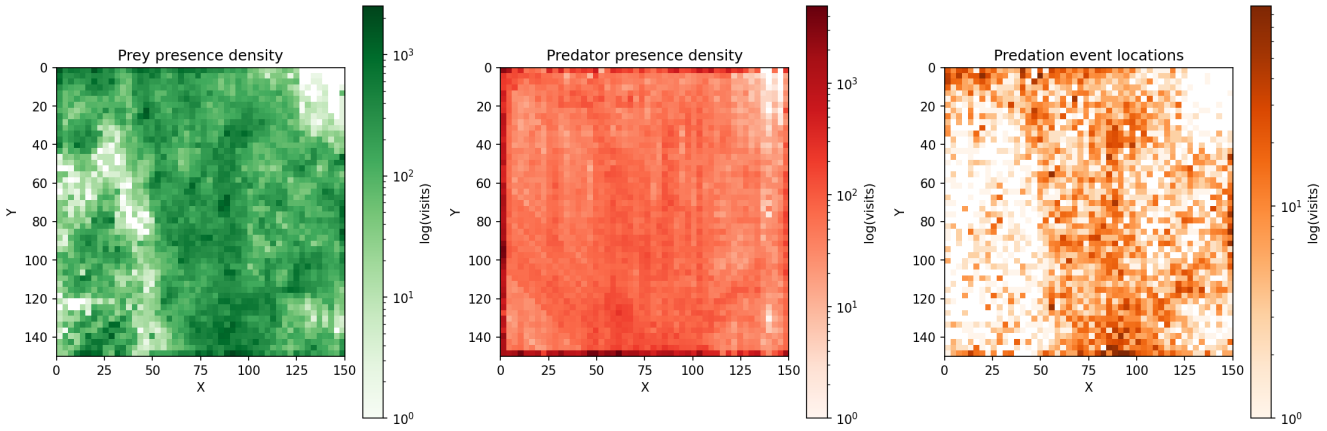

Figure 13: Spatial density maps for run 50 cycle 3 (random seed 51). Prey activity (left) is concentrated almost entirely in the bottom-right corner of the 150×150 grid. Predator density (center) forms broad convergent bands directed toward that corner, and predation events (right) are likewise localized near the boundary. This corner-trapping pattern is the spatial mechanism sustaining the population-level coexistence metrics and constitutes environment-specific overfitting to the hard-wall grid structure.

However, the spatial analysis shows that this success was achieved through model overfitting to the environment geometry rather than through ecologically satisfactory distributed behavior. Figure 13 reveals that prey activity is concentrated almost entirely in the bottom-right corner of the grid, while predator density forms broad convergent bands toward that same region. Predation events are likewise localized near the corner. This means that coexistence is not being maintained through balanced predator-prey interaction across the habitat, but through a corner-trapping equilibrium in which prey reduce exposure by concentrating at a boundary, and predators learn to exploit that concentration.

Run 50 cycle 3 also contains a particularly revealing illustrative case showing why final-step population counts are not sufficient to evaluate ecological success. One representative failure case ended the 2000-step horizon with 321 prey and 162 predators still alive, which would appear successful from a final snapshot alone. However, its time-averaged coexistence score was only 0.021, and it produced only 584 predation events over the entire rollout, compared with 3858 and 4381 predation events in the strongest healthy seeds. This shows that two seeds can end with apparently similar population counts while reflecting fundamentally different ecological regimes. For this reason, time-averaged metrics are essential in the present setting.

This interpretation is reinforced by the fact that the same spatial pattern appears in other high-performing seeds (whose summary can be seen in Figure 14), not only in the best individual case. The result is therefore not a single outlier, but a repeated solution strategy discovered by the policy. In this sense, run 50 cycle 3 represents the clearest example of environment-specific overfitting in the thesis: the learned behavior is highly effective inside the current MDP, but it is not ecologically generalizable.

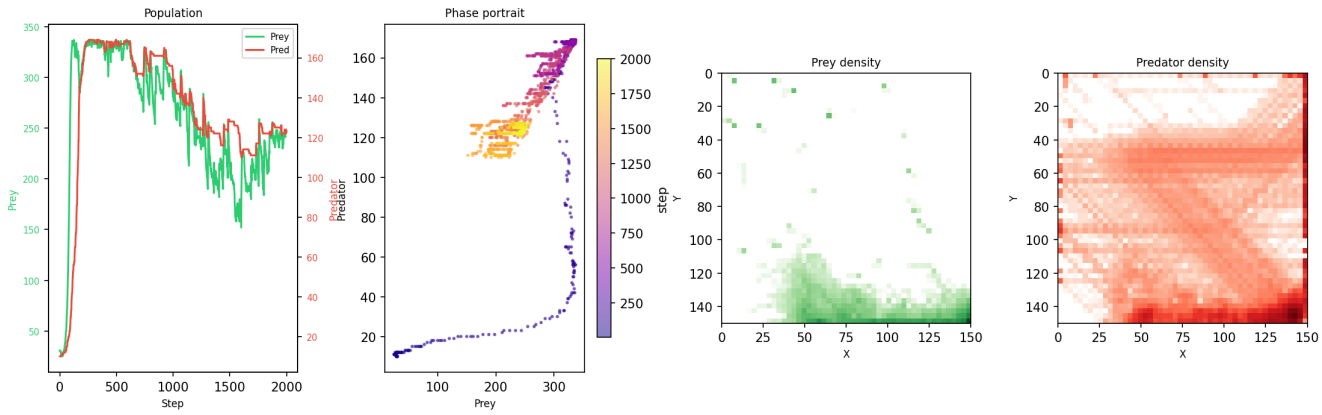

Figure 14: Summary panel for a second high-performing seed of run 50 cycle 3 (random seed 83, coexistence score 0.874, 4381 predation events). The same corner-concentration pattern visible in Figure 13 is reproduced here, confirming that corner-trapping is a consistent policy-level solution across seeds rather than an isolated outcome of a single initial condition.

A likely explanation for this corner-concentration pattern is the asymmetry introduced by the hard grid boundaries. In open space, prey can be approached from many directions, whereas at a corner, the number of effective predator approach directions is greatly reduced. Even though the state representation does not explicitly encode boundary context, repeated reward optimization can still favor corner positions because they increase expected local survival. Predators then adapt

to this prey concentration by converging toward the same region, producing a stable but spatially artificial coexistence pattern.

This interpretation is also supported by the large variation in predation activity across seeds. In the healthy coexistence seeds, cumulative predation events were typically in the range of roughly 3500–4400 over 2000 steps, whereas the clean collapse case recorded only 26 predation events, and the deceptive failure only 584. The spread is therefore ecological as well as numerical: seeds with similar end states may have been produced through very different predation regimes.

The importance of this checkpoint is therefore twofold. On the one hand, it demonstrates that tabular Q-learning can generate strong population-level coexistence metrics and long-horizon oscillatory dynamics resembling classical predator-prey behavior. On the other hand, it shows that these aggregate results can mask a poor underlying ecological mechanism. Run 50 cycle 3 is thus the best checkpoint numerically, but not the best ecological model conceptually. It constitutes the clearest expression of the other side of the central trade-off identified in the thesis: when coexistence becomes robust at the population level, the learned solution may do so by exploiting structural biases in the environment rather than by producing realistic ecosystem dynamics.

5.3.4 Comparative view across the three RL regimes

Taken together, the three evaluated checkpoints define a clear progression in the behavior learned by the tabular Q-learning system. Run 48 cycle 2 was the first regime to produce coexistence in a substantial fraction of seeds, but this coexistence remained sharply bimodal and therefore fragile. Run 48 cycle 3 produced the most ecologically meaningful local predator behavior, with predation distributed across the habitat rather than concentrated near a fixed boundary. However, this came at the cost of ecological instability, since predators became too effective and drove prey to extinction in most runs. Run 50 cycle 3 achieved by far the strongest aggregate coexistence metrics, with high survival and sustained oscillatory population dynamics, but this success was tied to a spatially degenerate corner-trapping strategy.

The comparison, therefore, shows that the three checkpoints should not be interpreted simply as weak, medium, and strong versions of the same ecological model. Instead, they reveal different trade-offs within the same learning framework. Run 48 cycle 2 demonstrates that coexistence can emerge, but not reliably. Run 48 cycle 3 demonstrates that meaningful local hunting behavior can emerge, but not without destabilizing the prey population. Run 50 cycle 3 demonstrates that robust population-level coexistence can be achieved, but in this case, through overfitting to the geometry of the environment.

This comparative pattern is central to the thesis. It shows that tabular Q-learning, under the current state representation and environment design, does not converge toward a single ecologically satisfactory solution. Instead, it resolves the predator-prey tension in different partial ways depending on the checkpoint: through unstable coexistence, through ecologically meaningful but excessive predation, or through numerically successful but spatially artificial coexistence. For this reason, run 50 cycle 3 can be treated as the best-performing checkpoint in quantitative terms, but not as a fully satisfactory final ecological model.

This trade-off strongly suggests that the 540-state tabular representation is not rich enough to encode the spatial and ecological context required for both robust coexistence and ecologically realistic behavior.

Table 5: Comparison of the three main RL checkpoints over 50 evaluation seeds.

Run	Cycle	Mean coex ± SD	Survival	Both alive	Prey CV	Pred CV	Oscillation period
48	2	0.381 ± 0.391	0.836	31/50 (62%)	0.223	0.604	∼5–6
48	3	0.370 ± 0.241	0.716	4/50 (8%)	0.487	0.616	∼6–7
50	3	0.782 ± 0.239	0.984	48/50 (96%)	0.173	0.398	∼8–9

5.4 Comparison with the ODE baselines

The comparison between the RL checkpoints and the ODE baselines shows that the relationship between the two approaches is only partial. At the population level, some RL trajectories resemble classical predator-prey dynamics, especially when viewed through aggregate counts and oscillatory behavior. However, once the underlying spatial mechanism is taken into account, none of the learned checkpoints can be considered a fully satisfactory analogue of the classical models.

Among the three RL regimes, run 50 cycle 3 is the closest to the classical baselines in purely aggregate terms. Its long-horizon rollouts exhibit sustained coexistence, relatively low population variability, and oscillatory dynamics that qualitatively resemble the cyclical behavior shown by both the fitted historical ODE models and the RL-calibrated Rosenzweig-MacArthur formulation. In particular, the population trajectories in Figure 12 are visually more consistent with the persistent oscillatory regimes of the fitted Hare-Lynx models shown earlier in Figure 2 than are the more unstable trajectories of runs 48 cycle 2 and 48 cycle 3. From this perspective, run 50 cycle 3 appears to provide the strongest population-level correspondence to the classical predator-prey framework.

At the same time, the calibrated ODE baseline highlights an important limitation of this apparent success. The RL-calibrated Rosenzweig-MacArthur model predicted a low predator equilibrium and a trajectory spiraling toward weak predator persistence, as shown in Figure 5. The fact that run 50 cycle 3 instead maintains a large predator population over 2000 steps indicates that the learned simulation is not reproducing the calibrated aggregate dynamics in any direct sense. Rather, it is generating a different effective regime, one that depends strongly on the spatial concentration of prey and on predator convergence toward that concentration. This explains why the density maps in Figure 13 are essential to the comparison: they show that the apparent agreement with classical oscillatory dynamics is only superficial, since the mechanism sustaining coexistence is boundary exploitation rather than distributed ecological interaction.

Run 48 cycle 3 provides a complementary contrast. In that checkpoint, the density maps show a far more spatially distributed pattern of predation than in run 50 cycle 3, which makes its local behavior more ecologically plausible. Yet its population dynamics do not align well with the stable coexistence structure of the fitted ODE baselines. Instead, prey extinction and subsequent predator decline dominate many runs, indicating that the learned policy produces excessive predation pressure relative to long-term coexistence. In this sense, run 48 cycle 3 captures something that the ODE baselines also make clear analytically: predator-prey systems are highly sensitive to the balance between attack pressure and predator maintenance. However, the checkpoint fails to stabilize that balance in practice.

Run 48 cycle 2 lies between these two extremes. Its successful seeds demonstrate that coexistence can emerge, but the large variance across seeds prevents any strong comparison with the classical models. Rather than matching a coherent ODE-like regime, it represents a brittle transitional stage in which coexistence is possible but not yet robust. This is consistent with the broader

pattern observed across the RL experiments: the learned policies do not converge toward a single stable ecological solution, but instead express different trade-offs between local behavior and population-level outcome.

To make this comparison more precise, the Lotka-Volterra and Rosenzweig-MacArthur models were also fitted directly to the RL population trajectories, treating each simulation step as one unit of continuous time. This fitting was performed for the best-performing seed of run 50 cycle 3 (random seed 51) and for a representative pre-extinction window of run 48 cycle 3 (seed 0, steps 1 to 1843). The results are summarized in Table 6.

Table 6: Lotka-Volterra and Rosenzweig-MacArthur parameters fitted to RL population trajectories, compared with the RL-calibrated analytical baseline (Table 4). RMSE values are in agent counts. The dashes in the last row indicate that no fitting was performed: the calibrated parameters were derived analytically from the simulator configuration rather than from trajectory data.

Checkpoint	Model	RMSE prey	RMSE pred	N^*	P^*	Stability
Run 50 c3 (seed9)	LV	68.3	31.5	266.8	143.7	Neutral centre
Run 50 c3 (seed9)	RM	34.8	7.7	294.3	158.1	Stable node
Run 48 c3 (seed0, pre-ext.)	LV	222.3	129.8	93.4	112.1	Neutral centre
Run 48 c3 (seed0, pre-ext.)	RM	80.1	13.6	257.2	150.8	Stable spiral
RL-calibrated RM (Table 4)	RM	—	—	25.3	1.3	Stable spiral

For run 50 cycle 3, the RM model provides a substantially better fit than LV, particularly for predator dynamics, where the RMSE drops from 31.5 to 7.7 (approximately 5% relative to the mean predator count of 152.8). The fitted RM equilibrium ($N^* = 294.3$, $P^* = 158.1$) is also very close to the observed trajectory mean ($\bar{N} = 295.2$, $\bar{P} = 152.8$), and the fitted carrying capacity ($K = 322.7$) is close to the actual RL environment value of 337. In this sense, the RM model captures the mean operating level of the run 50 cycle 3 trajectory well, although it does not reproduce the short-period oscillations, which reflect stochastic agent-level fluctuations rather than deterministic ODE dynamics.

The most important result of this fitting analysis concerns the comparison between the fitted equilibrium and the RL-calibrated analytical baseline from Section 5 (Table 4). That baseline, derived from simulator parameters under uniform-mixing assumptions, predicted $P^* \approx 1.3$. The RM model fitted to the actual run 50 cycle 3 trajectory instead finds $P^* = 158.1$, a difference of approximately 120-fold. This gap is not a fitting artefact. It is a direct consequence of the corner-trapping strategy identified in the spatial analysis: by concentrating prey in the boundary zone, the learned behavior creates a much higher effective local encounter rate than uniform mixing would produce. The calibrated ODE cannot represent this spatial mechanism, and therefore, it severely underestimates the predator abundance that the learned behavior can sustain. The fitting result therefore provides quantitative confirmation that the RL agents operate in a qualitatively different dynamical regime from the one predicted by the calibrated aggregate model.

For run 48 cycle 3, the fitting quality is substantially worse. The RM predator RMSE of 13.6 is higher than for run 50 cycle 3, and the prey RMSE of 80.1 is almost 2.3 times larger. This is expected: the pre-extinction trajectory is not stationary, and no fixed-parameter ODE can simultaneously capture the initial growth phase, the intermediate period of quasi-stability, and the final prey collapse. The fact that RM still provides a much better fit than LV for the predator

series (13.6 versus 129.8) confirms that some structure is present, but the high prey RMSE confirms that the trajectory as a whole is not well-described by either classical model.

Taken together, the comparison with the ODE baselines leads to two main results. First, the RL system can reproduce some classical population-level signatures, especially oscillation and coexistence, under selected checkpoints. Second, these similarities do not imply that the RL agents have learned an ecologically satisfactory version of the classical models. The ODE baselines remain more coherent and interpretable at the system level, whereas the RL results reveal that under the current state representation and environment design, population-level similarity can arise from mechanisms that are spatially artificial or ecologically unstable. For this reason, the comparison supports a qualified conclusion: tabular Q-learning can approximate classical predator-prey dynamics at the level of aggregate counts, but it does not yet provide a convincing ecosystem-level replacement for the classical baselines.

5.5 Summary of findings

Taken together, the results of this chapter show that the comparison between classical ecological models and the RL-based simulator leads to a mixed but informative conclusion. On the ODE side, the historical baseline analysis confirmed that classical predator-prey models remain useful as an aggregate reference system. Both Lotka-Volterra and Rosenzweig-MacArthur reproduced the Hare-Lynx case study reasonably well, while the Isle Royale case demonstrated the limits of fixed-parameter deterministic models under structural ecological disruptions. The RL-calibrated ODE scenario further showed that, under the simulator's current parameter regime, predator persistence should already be expected to be difficult at the aggregate level.

On the RL side, the training results showed that the learning system was able to acquire meaningful local predator-prey competencies. The 1v1 warm-up stage produced basic pursuit and evasion behavior, while the alternating multi-agent protocol generated progressively more specialized predator and prey policies. However, the long-horizon evaluation revealed that these learned strategies did not converge toward a single ecologically satisfactory solution. Instead, the three main checkpoints expressed three different regimes. Run 48 cycle 2 showed that coexistence could emerge, but only in a brittle and strongly seed-dependent way. Run 48 cycle 3 produced the most ecologically meaningful spatial hunting behavior, but this came at the cost of overhunting and widespread prey collapse. Run 50 cycle 3 achieved by far the strongest population-level coexistence metrics, yet did so through a spatially degenerate boundary-concentration strategy.

The main empirical conclusion of the results chapter is therefore not that one checkpoint simply succeeded and the others failed. Rather, the experiments reveal a structural trade-off in the current tabular Q-learning setup. When the learned policies produced more realistic distributed predator-prey interactions, they failed to sustain ecosystem-level coexistence. When they produced the strongest aggregate coexistence, they did so by overfitting to the geometry of the environment. In this sense, the results suggest that tabular Q-learning under the current state representation is capable of learning meaningful local behavior, but not of integrating that behavior into a robust and ecologically convincing ecosystem-level model.

The multi-seed evaluation provides evidence about the sensitivity of the learned policies to stochastic variation in initialization and rollout dynamics, but it does not constitute a full test of robustness under explicit environmental perturbations or noisy conditions.

Finally, the comparison with the ODE baselines reinforces this interpretation. The RL system

can reproduce some classical signatures at the level of aggregate population counts, especially oscillation and coexistence, but these similarities are not sufficient to claim that it has learned an ecologically satisfactory analogue of the classical models. The population-level agreement is real, yet the underlying mechanism often remains spatially artificial or dynamically unstable. The central finding of the chapter is therefore that tabular Q-learning can approximate predator-prey dynamics in limited ways, but it is not enough on its own to deliver a fully credible simulation of ecosystem population dynamics in the present setting.

6 Discussion

6.1 Interpretation of the main findings

The results of this thesis point to a clear but complex conclusion. At a population level, the RL-based simulator was capable of genuinely generating predator-prey dynamics that, in some cases, did resemble classical ecological behavior. In particular, only looking at the numbers, run 50 cycle 3 produced sustained coexistence, relatively low population variability, and oscillatory behavior that appeared broadly consistent with classical predator-prey cycles. At first sight, this suggests that tabular Q-learning can indeed approximate meaningful ecosystem dynamics.

However, the whole set of results shows that this conclusion is only partially valid. The three main checkpoints did not converge toward a single satisfactory ecological solution. Instead, they revealed three different regimes, each solving part of the predator-prey problem, while failing on another. Run 48 cycle 2 showed that coexistence could emerge, but only in a breakable and strongly seed-dependent way. Run 48 cycle 3 produced the most ecologically recognizable local behavior, with predators actively tracking prey across the habitat, but this came at the cost of overhunting and frequent collapse. Run 50 cycle 3 achieved the strongest aggregate coexistence metrics, but did so through spatial overfitting on the MDP.

The central interpretation is, therefore, that the learning system did not fail, but neither did it succeed in solving the ecological problem. The agents clearly learned meaningful local behavior: predators improved their pursuit efficiency, prey improved their evasion behavior, and the alternating protocol produced distinct predator-prey regimes rather than random movement. What failed was the integration of these local behaviors into a robust ecosystem-level solution. So, the system could learn to act meaningfully at a local scale, but it could not consistently translate this into ecologically satisfactory long-term population dynamics.

This interpretation is reinforced by the comparison with the ODE baselines. The RL system was able to reproduce some classical signatures at the level of aggregate population counts, especially oscillation and coexistence, but it did not reproduce them through the same kind of interpretable system-level mechanism. The ODE baselines remained analytically coherent, whereas the RL solutions often depended on unstable or spatially artificial strategies. This means that population-level similarity alone is not sufficient to conclude that the RL system learned a credible ecological model. The quantitative fitting analysis makes this concrete: the Rosenzweig-MacArthur model fitted to the actual run 50 cycle 3 trajectory recovers a predator equilibrium of $P^* = 158.1$, approximately 120 times higher than the $P^* \approx 1.3$ predicted by the RL-calibrated baseline under uniform-mixing assumptions. This gap is direct evidence that the corner-trapping regime is not a minor departure from the calibrated dynamics, but a qualitatively different effective interaction

regime, one that the aggregate model cannot anticipate from the simulator's parameters alone.

Taken together, the main finding of the thesis is not that reinforcement learning is useless for ecosystem simulation, nor that it clearly succeeds either. Rather, it is that tabular Q-learning in the present setting lies in the middle: it is strong enough to generate meaningful local predator-prey behavior and even partial ecosystem-level regularities, but not strong enough to combine ecological realism, robustness, and stable coexistence within a single learned solution. That alone is the key result that shapes the rest of the discussion.

6.2 Why tabular Q-learning was not sufficient

The main reason tabular Q-learning was insufficient in this thesis is that the learned policies operated on a state representation that was too limited to support real ecosystem-level regulation. The agents were able to learn meaningful local behavior, such as pursuit, evasion, and opportunistic feeding, but the 540-state tabular observation space cannot contain enough ecological context to connect those local decisions to long-term population consequences. In practice, this meant that the agents could become locally effective without becoming ecologically sustainable since they lacked information of the bigger picture.

This limitation is clear in the comparison between the three main checkpoints. Run 48 cycle 2 showed that coexistence could emerge, but did so on a fraction of the seeds. Run 48 cycle 3 produced the most ecologically recognizable predator behavior, with distributed hunting across the grid, but the predator became too effective and repeatedly destabilized the prey population. Run 50 cycle 3 achieved the strongest aggregate coexistence metrics, yet only by exploiting a corner-concentration strategy tied to the geometry of the grid. These outcomes suggest that tabular Q-learning was capable of solving parts of the problem, but not of integrating them into a single robust ecological solution.

The main bottleneck was the observation space itself. With only 540 discrete states, the policy could encode coarse local information such as energy, target direction, detection, and limited context, but it could not adequately represent broader ecological conditions such as population balance, habitat-scale spatial structure, or the long-term consequences of predation pressure. As a result, predators could learn that hunting was immediately rewarding without being able to infer that persistent overhunting would eventually collapse the prey population and therefore affect predator's reward as well. In the same way, prey could learn that boundary positions increased short-term survival without representing that this was trapping them against the predator in a future situation.

This helps explain why the learning system repeatedly resolved the task through partial strategies. When the predator adopted a more mobile and ecologically plausible hunting strategy, as in run 48, cycle 3, the prey population was pushed beyond what the system could sustain. When coexistence became numerically robust, run 50 cycle 3, it was achieved through boundary exploitation rather than through distributed ecological balance. The issue was therefore not simply that the agents failed to learn, but that the problem's representational structure encouraged locally rational yet globally insufficient solutions.

A second limitation is that tabular Q-learning scales poorly when a richer ecological context is needed. To correct the failures observed here, the agents would likely need access to more informative observations, such as local density structure, broader population pressure, or boundary awareness. However, incorporating such variables directly into a tabular state space would cause a

rapid combinatorial expansion of the number of states, making convergence and learning increasingly impractical. This means that the weakness of the present approach is not only in the specific policies learned, but also the fact that the tabular framework itself becomes restrictive once the ecological problem demands more expressive representations.

A third structural limitation is multi-agent non-stationarity. In the alternating training protocol, each agent learns against a frozen version of its counterpart, which means the learning target shifts every time the frozen agent is updated during the different phases. The deployed policies were never jointly converged: predator and prey were each trained to respond to a fixed opponent, not to a simultaneously adapting one. This helps explain why the learned strategies are often brittle under the full joint dynamics of evaluation. Unlike the state-space problem, this limitation cannot be resolved simply by enlarging the observation space; it requires either concurrent training methods that explicitly handle non-stationarity or self-play schemes in which both agents adapt together throughout training.

A further consequence of this representational limitation is that the oscillatory dynamics produced by the RL system are not the same kind of dynamics as those in classical ODE models. The Rosenzweig-MacArthur model fitted to the run 50 cycle 3 trajectory yields a stable node, meaning both eigenvalues are real and negative, which predicts overdamped monotonic convergence rather than oscillatory behavior. Yet the actual trajectory clearly shows short-period fluctuations. This indicates that the oscillations visible in the RL population curves are not deterministic ecological cycles of the type described by differential equation models, but stochastic fluctuations arising from individual agent births, deaths, and hunting events. Population-level similarity in terms of oscillatory appearance, therefore, does not imply that the RL system learned the same underlying ecological mechanism; it only means that individual-level demographic noise can produce patterns that superficially resemble ODE-type cycles.

6.3 Ecological realism and simulator limitations

The results show that the simulator captures some ecologically relevant ingredients, but remains a highly simplified ecosystem model. At the most basic level, it includes prey-predator interaction, renewable resources, reproduction, mortality, and spatial movement on a bounded habitat. These features are sufficient to generate non-trivial population dynamics and to make the learning problem ecologically meaningful. In that sense, the simulator goes beyond an abstract reinforcement-learning environment and provides a structured grid in which population-level behavior can emerge.

At the same time, the ecological realism of the simulator is limited by the way these mechanisms are implemented. The environment is a discrete grid with hard boundaries, simplified tile-based resources, and highly reduced agent observations. Predator and prey behavior is driven by a small number of local cues rather than a richer biological or environmental context. As a result, the agents do not perceive the habitat in a way that resembles real organisms, nor do they represent broader ecological pressures such as global scarcity, migration corridors, seasonal variation, or delayed environmental change.

One of the clearest limitations revealed by the results is the role of boundary geometry. The corner-concentration regime in run 50 cycle 3 showed that the simulator's hard walls create an artificial survival advantage for prey at the edges of the grid. Predators then adapt to that prey concentration, producing a stable but spatially unrealistic coexistence pattern. This means that the environment does not merely constrain behavior; it actively shapes which strategies become

optimal. In this case, the strategy that maximized survival inside the simulator was not a generally meaningful ecological strategy, but an exploitation of a structural artifact of the habitat.

A second major limitation is that the simulator operates at an ecological scale that is only partially represented in the agents' observation space. The agents act in a shared population system, but their state representation contains no direct encoding of broader population balance, habitat-wide spatial gradients, or long-term demographic pressure. This creates a mismatch between the scale of the ecological problem and the scale of the information available to the learner. In practical terms, the simulator asks agents to solve a population-regulation problem while only allowing them to perceive a narrow local slice of the system.

The simulator is also limited by its biological abstraction. Reproduction is simplified, individuals have no age structure, no memory, and no explicit physiological diversity beyond energy and species type, and environmental dynamics are restricted to a small set of deterministic or weakly stochastic processes. These simplifications are appropriate for a bachelor's thesis model, but they matter for interpretation: they mean that the results describe the properties of this particular computational setting rather than natural ecological dynamics, and that strategies which appear successful here may not remain viable under more biologically grounded constraints.

For this reason, the simulator should be understood as ecologically suggestive rather than ecologically realistic in a strong sense. It was realistic enough to expose meaningful tensions between local adaptation and ecosystem-level stability, which is a genuine contribution of the thesis. However, it was not realistic enough to support strong claims about how learned policies would transfer to real ecosystems. The value of the simulator lies primarily in showing what tabular Q-learning can and cannot achieve in a controlled predator-prey setting, not in providing a faithful model of natural ecological dynamics.

6.4 Implications for policy-oriented ecosystem simulation

The results of this thesis have mixed implications for policy-oriented ecosystem simulation. On the one hand, they suggest that reinforcement learning can be useful for exploring adaptive ecological behavior in a way that classical aggregate models cannot. The RL simulator was able to generate distinct predator-prey regimes, reveal tensions between predation efficiency and coexistence, and expose how local behavioral rules can scale into very different population-level outcomes. In that sense, RL-based simulation has value as an exploratory tool for studying how ecological strategies and environmental structure interact over time.

On the other hand, the present results also show that this value is not yet sufficient for direct policy use. A policy-oriented ecological model should not only produce plausible aggregate outputs, but should do so through mechanisms that remain interpretable, robust, and ecologically credible. The current tabular Q-learning setup did not meet that standard.

The comparison with the ODE baselines is especially relevant here. The classical models remain more interpretable and analytically stable, which makes them more suitable as transparent reference tools for high-level reasoning. At the same time, it is worth noting that even the classical baselines are not policy-ready tools in an unconditional sense: the Isle Royale case study in Section 5 showed that fixed-parameter ODE models fail under structural ecological disruptions that fall outside the scope of their assumptions. The RL simulator, by contrast, provides richer adaptive behavior but at the cost of reduced interpretability and weaker ecological control. For policy-oriented ecosystem simulation, this suggests that RL should not currently be viewed as a replacement for classical

ecological modeling, but rather as a complementary approach whose main value lies in exploring adaptive mechanisms, testing behavioral hypotheses, and identifying where aggregate models may miss important local dynamics.

The broader implication is therefore conditional rather than negative. Reinforcement learning may eventually contribute to policy-oriented ecosystem simulation, but only if the learning problem is embedded in a model that better reflects ecological structure at the scale relevant for management. In the present thesis, that condition was not satisfied. As a result, the policy relevance of the findings lies less in the learned policies themselves and more in the methodological lesson they provide: if RL is to become useful for ecological decision support, it must be coupled with richer ecological observations, more realistic spatial structure, and learning methods capable of reasoning over broader and longer-term system consequences.

6.5 Future work

The most immediate direction for future work is to improve the representation of ecological context available to the learning agents. The result of this thesis suggests that the main limitation of the current system is not the absence of learning altogether, but the mismatch between the ecological scale of the problem and the narrow local state representation provided to the agents. A priority would therefore be to enrich the observation space with variables that encode broader population and spatial context, such as local density fields, boundary awareness, neighborhood occupancy structure, or approximate prey-to-predator balance. This would allow agents to respond not only to immediate nearby targets, but also to the wider ecological consequences of their actions.

A second important direction is to improve the learning framework itself. The present study used tabular Q-learning because it provided a transparent and tractable baseline, but the results indicate that this approach becomes restrictive once the ecological problem requires richer state information and more flexible decision rules. Future work should therefore explore function-approximation methods, particularly deep reinforcement learning approaches, that can scale beyond a small discrete state space. Methods such as PPO or other actor-critic variants would be especially relevant if combined with richer observations or continuous movement, since they could support more expressive policies without the combinatorial explosion of tabular representations.

The spatial structure of the simulator should also be revised. The corner-concentration pattern found in run 50 cycle 3 showed that the hard grid boundaries introduced an artificial survival advantage that the agents learned to exploit. A natural next step would be to replace the current bounded geometry with toroidal boundaries, continuous-space movement, or obstacle-based habitats that eliminate the trivial refuge effects and force more ecologically meaningful spatial strategies. Related to this, future versions of the simulator could include more realistic terrain variation, better-designed patchy resources, environmental obstacles, or heterogeneous habitat regions so that movement and predation are shaped by ecological structure rather than by grid artifacts alone.

Another important extension would be to evaluate robustness more directly. In the present thesis, robustness was studied only partially through multi-seed variability. Future work should include dedicated perturbation experiments, such as sudden prey shocks, predator mortality shocks, temporary resource depletion, or spatial habitat change. These would make it possible to assess whether the learned policies remain viable under explicit ecological disturbances rather than only under stochastic variation in initialization and rollout conditions.

At the ecological level, the simulator could also be made more realistic by introducing richer

biological structure. Possible extensions include age classes, species-specific life-history differences, more detailed energy budgets, memory, sex-specific or seasonal reproduction, and variable resource regeneration. These additions would not only make the ecosystem more realistic, but could also help distinguish between strategies that are merely optimal in a simplified artificial setting and strategies that remain viable under more biologically grounded constraints.

Finally, future work should strengthen the connection between simulation and ecological theory. This thesis took a first step by fitting Lotka-Volterra and Rosenzweig-MacArthur models to the RL population trajectories, which revealed the 120-fold discrepancy between the calibrated and fitted predator equilibria. A natural extension would be to apply this fitting procedure across all 50 evaluation seeds and all three checkpoints, rather than only to individual representative seeds, so that the correspondence between RL dynamics and classical models can be assessed as a distribution rather than as isolated examples. Beyond this, future work could study how fitted ODE parameters evolve across training cycles, whether deep RL agents with richer observations produce trajectories that are more consistent with classical ecological regimes, and whether the learned interaction rules can be mapped onto interpretable ecological quantities. The main value of this direction is not only to improve performance, but to determine whether reinforcement learning can eventually serve as a genuinely credible ecosystem-level modeling approach rather than only as a locally adaptive simulation technique.

7 Conclusion

7.1 Answer to the research question

The central research question of this thesis asked how effective reinforcement learning methods are, compared to classical ecological models, in supporting ecosystem policy design. Based on the results obtained, the answer is mixed. Reinforcement learning, in the form implemented here, was effective enough to generate meaningful local behavior and, in some cases, population-level patterns that resembled classical ecological dynamics. However, it was not effective enough to provide a robust, ecologically credible, and interpretable ecosystem-level model comparable to the classical baselines.

More specifically, the study showed that Q-learning could learn partial solutions to the predator-prey problem. When the learned policies display more ecologically recognizable spatial behavior, coexistence tends to collapse through overhunting. When coexistence became strong at the aggregate level, it was sustained through corner-trapping and boundary exploitation rather than through distributed ecological balance. As a result, the RL system did not converge toward a single satisfactory ecological solution.

Compared with the classical models, the RL-based simulator offered a different kind of value. The ODE baselines remained more interpretable, more coherent at the system level, and more suitable as transparent reference models. The RL system contributed adaptive local behavior and exposed mechanisms that aggregate models cannot represent directly, but it did so at the cost of reduced interpretability and weaker ecological control. Therefore, RL functioned as a complementary exploratory framework whose strengths lay in revealing behavioral dynamics and failure modes rather than in delivering a superior policy-support model.

Regarding the first sub-question, on how accurately Q-learning agents can reproduce population

dynamics compared to classical models, the answer is: partially, and with important caveats about the mechanism. The best-performing checkpoint produced sustained coexistence and oscillatory population curves, and the Rosenzweig-MacArthur model fitted to that trajectory achieved a predator RMSE of 7.7, suggesting some classical structure is present. However, the same fitting analysis found a predator equilibrium of $P^* = 158.1$, approximately 120 times higher than the $P^* \approx 1.3$ predicted by the RL-calibrated analytical baseline under uniform-mixing assumptions. This 120 times confirms that the RL agents operate in a qualitatively different dynamical regime driven by spatial concentration, not one that follows from the classical model with the simulator's own parameters. Furthermore, the fitted model predicts a stable node with no oscillations, while the actual trajectory shows clear fluctuations, confirming that those oscillations are stochastic demographic noise rather than deterministic ecological cycles.

Regarding the second sub-question, on robustness and alignment with conservation objectives, the answer is partial. The multi-seed evaluation across 50 initialization seeds showed that the strongest checkpoint maintained coexistence across the majority of seeds and displayed lower population variability than the weaker checkpoints, providing evidence of stochastic robustness under varied rollout conditions. Alignment with survival-based conservation objectives was partially demonstrated through the coexistence score and survival fraction metrics. However, the thesis did not conduct explicit perturbation experiments such as prey shocks or resource depletion, so robustness under active environmental disturbance remains an open question. The alignment that was demonstrated is also limited in ecological relevance because the mechanism sustaining coexistence in the strongest checkpoint was a spatial artifact rather than a generalizable ecological strategy.

The answer to the research question is that reinforcement learning was only partially effective in this setting. It showed clear potential for ecosystem simulation, but tabular Q-learning with the present state representation and environment design was not sufficient to support policy-oriented ecological modeling in a strong sense. The method is capable of approximating some predator-prey dynamics, yet it cannot combine robustness, ecological realism, and stable coexistence within a single learned policy. For this reason, the thesis concludes that reinforcement learning remains a promising direction for ecosystem simulation, but only if future work moves beyond the current tabular setup towards richer ecological observation and more expressive learning methods.

7.2 Main contributions and limitations

The thesis makes two main contributions. First, it develops and evaluates a reinforcement-learning-based predator-prey simulator that is rich enough to generate non-trivial ecological regimes rather than only trivial survival behavior. The study shows that even a relatively simple RL setup can produce meaningful local predator-prey strategies, including pursuit, evasion, and feeding behavior, and can generate population-level dynamics that partially resemble classical ecological oscillations. In that sense, the thesis demonstrates that reinforcement learning can serve as a useful exploratory framework for ecosystem simulation.

Second, the thesis contributes a comparative analysis between classical ecological baselines and learned simulation behavior. By combining fitted ODE models, RL-calibrated analytical baselines, and long-horizon multi-seed evaluation, the study does not merely report whether the RL system performed well or poorly. Instead, it identifies a more informative result: the learned policies produced three distinct ecological regimes, each reflecting a different trade-off between local

behavior and ecosystem-level outcome. This comparative perspective is one of the main strengths of the thesis, because it shows both what reinforcement learning can achieve and where its current limitations become decisive.

Third, the thesis provides a concrete mechanistic diagnosis of where and why RL-based simulation diverges from classical ecological dynamics. The identification of corner-trapping as the mechanism sustaining the strongest aggregate checkpoint, and its quantitative confirmation through the 120-fold discrepancy between the fitted and calibrated predator equilibria, is a specific finding that aggregate models alone could not have produced. The additional result that the fitted Rosenzweig-MacArthur model predicts a stable node while the RL trajectory shows oscillations further establishes that the two modeling paradigms generate dynamics through fundamentally different mechanisms, even when their population-level outputs appear similar.

At the same time, the thesis has several important limitations. The most significant is that the reinforcement-learning component relies on a coarse tabular state representation that cannot adequately encode broader ecological context. As a result, the agents can learn locally effective behavior without being able to regulate their actions in a way that remains sustainable at the ecosystem level. This is the main reason why the learned solutions tended to resolve the predator-prey problem only partially: through brittle coexistence, overhunting, or spatial overfitting.

A second limitation concerns ecological realism. Although the simulator includes predator-prey interaction, spatial structure, resource renewal, reproduction, and mortality, it remains a highly simplified artificial ecosystem. The hard grid boundaries, reduced local observations, and simplified biological assumptions mean that the learned behaviors cannot be interpreted as direct ecological predictions about real systems. In particular, the corner-concentration regime observed in the strongest aggregate checkpoint shows that the policies are sensitive to structural artifacts of the environment.

A third limitation is methodological scope. The thesis provides a robust comparison across seeds and checkpoints, but it does not fully address robustness under explicit environmental perturbations such as shocks, habitat change, or resource collapse. Similarly, while the comparison with ODE baselines was extended through fitted trajectory analysis, this fitting was performed only for representative runs rather than across the full distribution of evaluation seeds. The conclusions are therefore strongest at the level of controlled proof-of-concept comparison rather than full ecological validation.

In conclusion, these contributions and limitations define the thesis clearly. The work shows that reinforcement learning can generate meaningful predator-prey behavior and can reveal important failure modes that classical aggregate models do not capture directly. At the same time, it also shows that tabular Q-learning in the present form is not sufficient to serve as a fully credible ecosystem-level simulation method. That combination of positive evidence and clearly identified limits is, in itself, one of the main contributions of the study.

7.3 Final remarks

This thesis began from the question of whether reinforcement learning could serve as a useful alternative or complement to classical ecological modeling in the context of ecosystem simulation. The answer that emerged is neither a simple success story nor a simple failure. Instead, the study showed that reinforcement learning can generate meaningful local predator-prey behavior and can even reproduce some aggregate signatures of classical ecological dynamics, but that these

achievements are not enough on their own to constitute a robust and ecologically credible ecosystem model.

The most important lesson of the thesis is therefore methodological. When reinforcement learning is applied to ecological systems, the quality of the learned behavior depends not only on the learning algorithm itself, but on the ecological realism of the environment, the information available to the agents, and the degree to which local reward optimization is aligned with long-term population stability. In the present study, tabular Q-learning was able to learn partial solutions, but the limitations of the state representation and simulator structure prevented those solutions from becoming fully satisfactory at the ecosystem level.

At the same time, this result should not be understood as closing the door on reinforcement learning for ecology. On the contrary, the experiments show that RL can be a valuable research tool precisely because it exposes tensions that are less visible in aggregate equation-based models. The fact that the learned system produced fragile coexistence, overhunting, and spatial overfitting is itself informative: it reveals which ecological mechanisms must be represented more carefully if learning-based simulation is to become scientifically and practically useful. And, honestly, getting there will require more than a larger Q-table.

References

[1] Heterogeneous mean-field analysis of the generalized lotka-volterra model on a network. arXiv preprint, 2024. Verify author names and final bibliographic metadata from the PDF before final submission.

[2] Christopher Amato. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2409.03052*, 2024.

[3] Jacqueline Augusiak, Paul J. Van den Brink, and Volker Grimm. Merging validation and evaluation of ecological models to 'evaludation': A review of terminology and a practical approach. *Ecological Modelling*, 280:117–128, 2014.

[4] Thomas Banitz, Maja Schlüter, Emilie Lindkvist, Sonja Radosavljevic, Lars-Göran Johansson, Petri Ylikoski, Rodrigo Martínez-Peña, and Volker Grimm. Model-derived causal explanations are inherently constrained by hidden assumptions and context: The example of baltic cod dynamics. *Environmental Modelling & Software*, 157:105489, 2022.

[5] Corey J. A. Bradshaw and Salvador Herrando-Pérez. Logistic-growth models measuring density feedback are sensitive to population declines, but not fluctuating carrying capacity. *Ecology and Evolution*, 13(4):e10010, 2023.

[6] John D. Co-Reyes, Suvansh Sanjeev, Glen Berseth, Abhishek Gupta, and Sergey Levine. Ecological reinforcement learning. *arXiv preprint arXiv:2006.12478*, 2020.

[7] Huseyin Coskun. Dynamic ecological system analysis: A holistic analysis of compartmental systems. *Heliyon*, 5(9):e02347, 2019.

[8] Maximilien Cosme, Colin Thomas, and Cédric Gaucherel. On the history of ecosystem dynamical modeling: The rise and promises of qualitative models. *Entropy*, 25(11):1526, 2023.

[9] Donald L. DeAngelis and Stephen Yurek. Equation-free modeling unravels the behavior of complex ecological systems. *Proceedings of the National Academy of Sciences*, 112(13):3856–3857, 2015.

[10] Frank D'Erchia, Carl E. Korschgen, Maury Nyquist, Randy Root, Richard S. Sojda, and Peter Stine. A framework for ecological decision support systems: Building the right systems and building the systems right. Information and Technology Report 2001-0002, United States Geological Survey, Reston, VA, 2001.

[11] Ethan R. Deyle, Robert M. May, Stephan B. Munch, and George Sugihara. Tracking and forecasting ecosystem interactions in real time. *Proceedings of the Royal Society B: Biological Sciences*, 283(1822):20152258, 2016.

[12] Graziella V. DiRenzo, Ephraim M. Hanks, and David A. W. Miller. A practical guide to understanding and validating complex models using data simulations. *Methods in Ecology and Evolution*, 14(1):203–217, 2023.

[13] Carlos M. Duarte, Jeffrey S. Amthor, Donald L. DeAngelis, Linda A. Joyce, Roxane J. Maranger, Michael L. Pace, John Pastor, and Steven W. Running. The limits to models in ecology. In Charles D. Canham, Jonathan J. Cole, and William K. Lauenroth, editors, *Models in Ecosystem Science*, pages 437–452. Princeton University Press, Princeton, NJ, 2004.

[14] Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

[15] Alan G. Hildrew and Colin R. Townsend. A predator-prey interaction with role reversal. *Oikos*, 42(1):19–26, 1984. Use this only if it matches the exact paper you intended; your spreadsheet link appears to point to a different later article.

[16] Dom Huh and Prasant Mohapatra. Multi-agent reinforcement learning: A comprehensive survey. *arXiv preprint arXiv:2312.10256*, 2023.

[17] Sharon E. Kingsland. The refractory model: The logistic curve and the history of population ecology. *The Quarterly Review of Biology*, 57(1):29–52, 1982.

[18] Sharon E. Kingsland. Alfred j. lotka and the origins of theoretical population ecology. *Proceedings of the National Academy of Sciences*, 112(31):9493–9495, 2015.

[19] Marcus Lapeyrolerie, Melissa S. Chapman, Kari E. A. Norman, and Carl Boettiger. Deep reinforcement learning for conservation decisions. *Methods in Ecology and Evolution*, 13(11):2649–2662, 2022.

[20] Simon A. Levin. Self-organization and the emergence of complexity in ecological systems. *BioScience*, 55(12):1075–1079, 2005.

[21] Moritz Lippe, Mike Bithell, Nicholas Gotts, Davide Natalini, Pete Barbrook-Johnson, Carlo Giupponi, Maik Hallier, Gert Jan Hofstede, Christophe Le Page, Robin B. Matthews, Maja Schlüter, Pete Smith, Arianna Teglio, and Katrin Thellmann. Using agent-based modelling to simulate social-ecological systems across scales. *GeoInformatica*, 23(2):269–298, 2019.

[22] Brian E. McLaren and Rolf O. Peterson. Wolves, moose, and tree rings on isle royale. *Science*, 266(5190):1555–1558, 1994.

[23] G. Michae. Reinforcement learning for dynamic decision making in ecological management. *International Journal of Knowledge Engineering and Computing Innovations*, 2025. Bibliographic details limited; verify journal metadata before final submission.

[24] Jeongho Park, Juwon Lee, Taehwan Kim, Inkyung Ahn, and Jooyoung Park. Co-evolution of predator-prey ecosystems by reinforcement learning agents. *Entropy*, 23(4):461, 2021.

[25] Rolf O. Peterson. Wolf–moose interaction on isle royale: The end of natural regulation? *Ecological Applications*, 9(1):10–16, 1999.

[26] Edward J. Rykiel. Testing ecological models: the meaning of validation. *Ecological Modelling*, 90(3):229–244, 1996.

[27] Piyush K. Sharma, Rolando Fernandez, Erin G. Zaroukian, Michael R. Dorothy, Anjon Basak, and Derrik E. Asher. Survey of recent multi-agent reinforcement learning algorithms utilizing centralized training. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, volume 11746, page 1174608. SPIE, 2021.

[28] Daniele Silvestro, Stefano Goria, Thomas Sterner, and Alexandre Antonelli. Improving biodiversity protection through artificial intelligence. *Nature Sustainability*, 5(5):415–424, 2022.

[29] Kuldeep Singh, Teekam Singh, Lakshmi Narayan Mishra, Ramu Dubey, and Laxmi Rathour. A brief review of predator-prey models for an ecological system with a different type of behaviors. *Korean Journal of Mathematics*, 32(3):381–406, 2024.

[30] Nils C. Stenseth, Wilhelm Falck, Ottar N. Bjørnstad, and Charles J. Krebs. Population regulation in snowshoe hare and canadian lynx: Asymmetric food web configurations between hare and lynx. *Proceedings of the National Academy of Sciences*, 94(10):5147–5152, 1997.

[31] Nils C. Stenseth, Wilhelm Falck, Kung-Sik Chan, Ottar N. Bjørnstad, Mark O'Donoghue, Howell Tong, Rudy Boonstra, Stan Boutin, Charles J. Krebs, and Nigel G. Yoccoz. From patterns to processes: Phase and density dependencies in the canadian lynx cycle. *Proceedings of the National Academy of Sciences*, 95(26):15430–15435, 1998.

[32] Taro Tahara, Maria Kris Areja, Tetsuya Kawano, Joseph M. Tubay, Jomar F. Rabajante, Hiroshi Ito, Kei Suzuki, Toshio Aonishi, and Satoru Morita. Asymptotic stability of a modified lotka-volterra model with small immigrations. *Scientific Reports*, 8:7029, 2018.

[33] James T. Thorson and Kasper Kristensen. Ecological examples of nonstationarity, nonlinearity, and statistical interactions in dynamic structural equation models. EcoEvoRxiv, 2026. Preprint / in press at Methods in Ecology and Evolution.

[34] Rebecca Tyson, Sheena Haines, and Karen E. Hodges. Modelling the canada lynx and snowshoe hare population cycle: The role of specialist predators. *Theoretical Ecology*, 3(2):97–111, 2010.

[35] Eric Walling and Céline Vaneeckhaute. Developing successful environmental decision support systems: Challenges and best practices. *Journal of Environmental Management*, 264:110513, 2020.

[36] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992.

[37] Bo Zhang and Donald L. DeAngelis. An overview of agent-based models in plant biology and ecology. *Annals of Botany*, 126(4):539–548, 2020.