

# 1 Energy-Efficient Path Planning in Winter Conditions: A Comparative 2 Study of Traditional Baseline, Q-learning, and Proximal Policy 3 Optimization in a Grid World Environment

4  
5

## 6 1.Abstract

7 Recent investigation in winter route planning has become increasingly important due to  
8 increased resistance and reduced efficiency on winter roads. These conditions  
9 significantly increase energy consumption and introduce uncertainty, raising the risk of  
10 failing to reach the destination. To address the challenge, reinforcement learning (RL) is  
11 a type of machine learning where an agent learns to make decisions by interacting with  
12 an environment, receiving rewards or penalties for its actions to maximize cumulative  
13 rewards. For simulation, we benchmark a standard shortest-path algorithm (BFS) against  
14 two reinforcement learning methods: Q-learning and proximal policy optimization (PPO). All  
15 approaches are tested on identical grid configurations, with the same starting points, goals,  
16 and reward functions. We assess their performance based on cumulative reward, snowy cell  
17 visits (as a proxy for energy cost), trajectory characteristics, and success rate. It is  
18 hypothesized that proximal policy optimization (PPO) algorithms will result in lower  
19 energy consumption than Q-learning and traditional baseline under winter conditions.  
20 The results show that while BFS consistently finds the shortest paths, it fails to consider  
21 energy costs or environmental uncertainty. RL agents, by comparison, adapt more  
22 efficiently to winter conditions. In contrast, Q-learning has an overall better  
23 performance in both conditions than PPO. These results suggest that Q-learning is  
24 better suited for structured winter navigation in grid worlds, while PPO may perform  
25 better in more complex environments with continuous states or decisions.

26

27 Key words: Energy-Efficient, Traditional Baseline, Q-learning, Proximal Policy  
28 Optimization, grid world, reinforcement learning.

29

30

---

31

## 32 2.Introduction

33 As demand in electric vehicles become increasingly widespread, energy efficiency and  
34 route optimization have emerged as critical challenges, particularly in winter when  
35 resistance increases. Freezing temps, plus snow and ice covering the roads, push energy  
36 consumption way up, which means shorter range and trips that feel less predictable. A  
37 2025 study shows that an estimated 50 % of EV driving range can be reduced in cold  
38 climates, including snow and ice covering terrain, highlighting the significant impact of  
39 environmental conditions on energy consumption [1]. Furthermore, when snow or  
40 water remains on the road surface, vehicle tires must continuously displace through ice  
41 as they roll and move, hence forcing the vehicle to draw more power [2]. On top of that,  
42 water cools tires more effectively than air alone, altering their mechanical properties

43 and further pushing rolling resistance higher. Previous studies have shown that rolling  
44 resistance can rise by approximately 30% to 40% under wet or snowy conditions.  
45 Ultimately, these factors make winter driving a major concern for EV efficiency and  
46 reinforce the need for energy-aware routing strategies.

47

48 To approach these challenges, we have explored a few computational methods to allow  
49 agents to learn effective navigation strategies aimed at minimizing energy consumption.  
50 In this study, a 50x50 grid is used as an abstraction routing map that circles key  
51 structures of Canadian snow distribution, rather than exact geographical terrain. This  
52 abstraction supports controlled comparison of different routing methods while  
53 preserving the essential energy-related difficulties.

54

55 More specifically, the research will include a traditional baseline algorithm, and two  
56 types of reinforcement learning, which are Q-learning and Proximal Policy optimization  
57 algorithm. The traditional baseline routing computes the shortest path based on static  
58 cost metrics and follows the predefined route without adaptation or learning (basic  
59 routing). Q-learning, a value-based reinforcement learning method, learns through  
60 incremental dynamic programming processes with computational requirements and it  
61 is well suited for agents to improve and refine action values and achieve effective  
62 performance in controlled Markovian domains [3]. In contrast, PPO is a policy-gradient  
63 algorithm that primarily optimizes a stochastic policy and achieves greater training  
64 stability through constrained policy updates with a clipped surrogate objective function  
65 [4].

66

67 This comparison evaluates the effectiveness of the proposed routing methods in  
68 identifying energy-efficient paths. Performance is evaluated through energy related  
69 costs, overall routing efficiency, and observed navigation behavior under both  
70 deterministic (non-slip) and stochastic (slip) settings. Because PPO balances stable  
71 policy updates with adaptive learning in uncertain environments, it is hypothesized to  
72 be the most effective method implemented for winter routing.

73

74 The following section goes into the methodology and experimental setup in more detail.

75

76

---

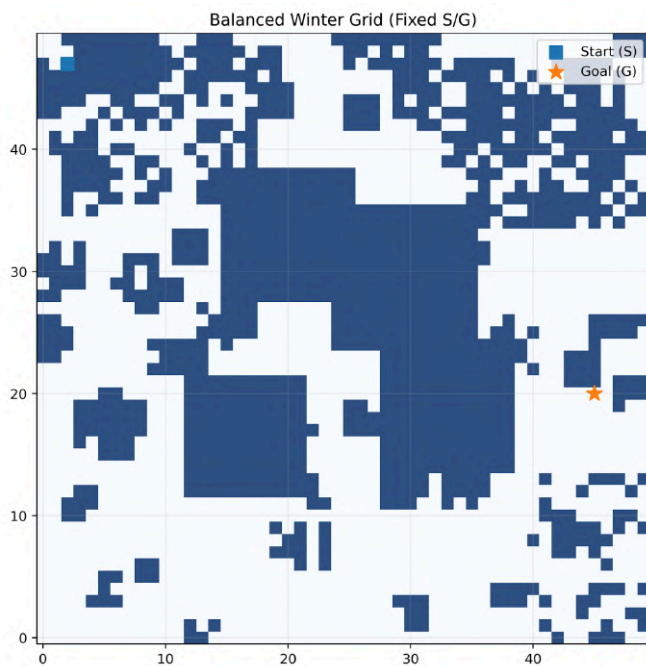
77

## 78 **3.Methods**

### 79 **3.1 Environment Design**

80 A custom winter navigation grid environment was designed to evaluate each  
81 reinforcement (RL) agent under stochastic and cost-sensitive conditions. The  
82 environment consists of a 50 x 50 grid world containing 2500 cells, including both  
83 normal terrain and snow-covered terrain. Specifically, two transition settings were  
84 evaluated: a slip snow condition and non-slip snow condition (same grid map). The

85 simulation contains a fixed start (47, 2) and a fixed goal (20,45) point, and are located in  
86 the top left corner and middle right respectively. At each time step, the agent can take  
87 either one of the four directions, up, down, left, and right; each step on a non-snow grid  
88 has an energy cause of -1.5 per step. Snow is modeled as spatially correlated regions with  
89 graded intensity, capturing the typical uneven accumulation patterns in Canadian winters.  
90 When snow is modeled as a binary state, abrupt reward discontinuities make PPO's  
91 advantage estimates unstable, resulting in increased gradient variance. By contrast, a  
92 continuous representation of snow thickness provides smoother cost transitions, which  
93 help reduce gradient variance and stabilize learning, making it more efficient. These  
94 snow types are classified as "Near snow", "Edge snow", and "Core snow".  
95 To evaluate robustness under different winter conditions, two transition settings were  
96 considered using the same grid map: a non-slip (deterministic) condition and a slip  
97 (stochastic) condition. In the deterministic setting, action always results in the intended  
98 movement, and snow cells only add additional energy loss. In the stochastic setting,  
99 intended actions may be replaced with unintended movements due to unstable control  
100 on icy surfaces, with a higher probability of deviating left or right, simulating loss  
101 control of traction during winter driving on ice and snow.  
102



103

104

105

106 Figure 1: shows a typical winter grid layout used in all experiments.

107

### 108 3.2 Reinforcement Learning Algorithms/ Methodologies:

109 The routing strategies and learning methods used in this study are described in this  
110 section.

111

#### 112 3.2.1 Traditional Baseline (BFS)

113 Traditional Baseline is a non-learning shortest path strategy used as a comparison base  
114 for reinforcement learning methods. It operates on the 50x50 grid map, it considers  
115 four directions to move (up, down, left, right). The system calculates the shortest path  
116 to the destination, without considering energy savings, which means it does not adapt  
117 to environmental feedback or uncertainty.

118

### 119 3.2.2 Q-learning Implementation

120 Tabular Q-learning was used as the value-based reinforcement learning baseline. The agent  
121 maintains a discrete Q (s, a) table, which is updated iteratively based on observed state transitions.  
122 An  $\epsilon$ -greedy policy was used for exploration. The value of  $\epsilon$  begins at 0.6 and decays  
123 exponentially with a factor of 0.9995 per episode until it reaches 0.05. In this 50x50 grid  
124 setting, the decay schedule allows broad initial exploration before shifting emphasis to  
125 exploitation. Actions are restricted to four possibilities—left (0), down (1), right (2), and  
126 up (3)—consistent with standard discrete grid-world formulations and the discrete grid  
127 structure. The Q-value update follows the classic temporal-difference form:

128

$$129 Q(s, a) \leftarrow Q(s, a) + \alpha [R + \gamma \max_{a'} Q(s', a') - Q(s, a)] \quad [5]$$

130

131 We fix the learning rate  $\alpha$  at 0.15 and the discount factor  $\gamma$  at 0.99. These values handle the  
132 stochastic transitions (is\_slippery=True) and support planning over long horizons in the large  
133 grid [3]. State  $s$  denotes the flattened grid position index. The action  $a$  selects one of the four  
134 movement directions.  $Q(s, a)$  represents the estimated discounted cumulative reward  
135 starting from state  $s$  and action  $a$  [3,5].

136

### 137 3.2.3 Reward system

138 In this research, the reward system is designed to represent energy efficiency under  
139 winter conditions, it helps agents find the best paths. At each step, the system gives a  
140 negative value to show how significant a step is to energy saving, with higher cause on  
141 snow grids. A positive 700 is given only when the agents successfully reach the  
142 destination, while within the maximum steps of 1000 steps. It encourages agents to  
143 reach the destination with considerations on energy saving. We initialize the cumulative  
144 reward to 0 at the beginning of each episode. Each will result in negative values,  
145 because by doing this, agents don't need to consider the prior bias, since it assumes  
146 nothing at first, etc. As in the settings, we have Near-snow (light snow), Edge-snow  
147 (medium snow), and Core-snow (deep snow); each of them has different additional  
148 values which are -0.2, -0.5, and -2.0 respectively [6].

149

$$150 r_t = r_{step} - c_{energy}(s_t) + r_{goal} \quad [6]$$

151

152 The basic reward function follows standard reinforcement learning practice by  
153 combining step penalties, energy-related costs, and a terminal goal reward.

154

### 155 3.2.4 Proximal Policy Optimization

156 Proximal Policy Optimization (PPO) served as the policy-gradient method in this study  
157 for energy-efficient routing under winter conditions. PPO learns a stochastic policy by  
158 directly outputting action probabilities for each state, which helps the agent cope with  
159 uncertainty on slippery roads.

160

$$161 L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \quad [4]$$

162

163 where  $r_t(\theta)$  is the probability ratio between the new and previous policies. The clipping  
164 mechanism constrains policy updates to improve stability.

165

166 The PPO agent was trained using fixed values across all experiments. The discount factor was set  
167 to  $\gamma=0.99$ , and Generalized Advantage Estimation (GAE) was used with  $\lambda=0.95$ . The policy  
168 network consisted of two fully connected hidden layers with 512 units each, using ReLU  
169 activation functions. Entropy regularization was applied to encourage exploration during  
170 training. PPO training was conducted for 1,200,000 time steps, using large rollout buffers  
171 and multiple optimization epochs to improve training stability [4].

172

173 Policy-gradient methods work by directly adjusting a parameterized policy to increase  
174 expected reward. They learn a stochastic policy, allowing for more flexibility in  
175 uncertain environments, where the same action can lead to different outcomes. In  
176 contrast, q-learning estimates state-action values and usually converges with a  
177 deterministic policy based on these estimates. Although it can learn in unfamiliar  
178 environments, its action selection may be less reliable in highly unpredictable  
179 situations. As a result, policy-gradient and value-based methods show varying strengths  
180 depending on the level of uncertainty in the winter routing task.

181

### 182 3.2.5 Curriculum Learning for PPO

183 PPO training followed a two-phase curriculum learning approach designed to improve  
184 training stability and sample efficiency under sparse rewards and energy-sensitive  
185 penalties [10].

186

#### 187 Phase 1: Navigation Learning Phase

188 Navigation Learning Phase In phase 1, PPO was trained using a simplified reward  
189 function that focuses only on positive feedback for reaching the goal. Energy costs in  
190 snow grids were not included. This made the feedback denser and more immediate. The  
191 agent quickly learned basic navigation and achieved early success in the 50×50 winter  
192 grid. The policy learned in this phase served as the starting point for the next training  
193 stages.

194

#### 195 Phase 2: Energy-Aware Optimization Phase

196 Energy-Aware Optimization Phase In phase 2, the training continued with the complete  
197 energy-aware reward function, which now included penalties for moving across  
198 snow-covered areas. The agent had to balance successful arrival with lower energy  
199 consumption. All reported results, ablation studies, and comparisons were based solely  
200 on the Phase 2 reward function.

201

202

203

### 204 **3.3 Evaluation factors**

205 To compare the performances of all routing methods under identical conditions, a final  
206 evaluation was conducted after training. Each method BFS, Q-learning, and PPO was  
207 evaluated using the following metrics:

208

- 209 • Total Reward: calculated as the cumulative reward obtained over the episode.
- 210
- 211 • Number of steps: steps taken from the start to the goal (less than 1000 steps).
- 212
- 213 • Average snow cell visited: Snow visits were calculated by counting the number of  
214 snow-covered cells traversed in each episode and averaging this value across all  
215 evaluation episodes.
- 216
- 217 • Success rate: defined as the percentage of episodes in which the agent reached  
218 the goal within the maximum step limit.

219

220 The BFS baseline was evaluated only once on the fixed grid map, as it produces a  
221 deterministic path. Reinforcement learning agents were evaluated over multiple  
222 episodes. All methods were compared using the same-energy aware reward function  
223 and environment configuration to ensure fairness.

224

225 A complete summary of environment settings and hyperparameters is provided in  
226 Appendix A.

227

228

229

---

## 230 **4. Results**

231 The experiments compare BFS, Q-learning, and PPO on the same 50×50 winter grid,  
232 looking at both non-slip (deterministic) and slip (stochastic) cases.

233

### 234 **4.1 Evaluation Setup and Metrics**

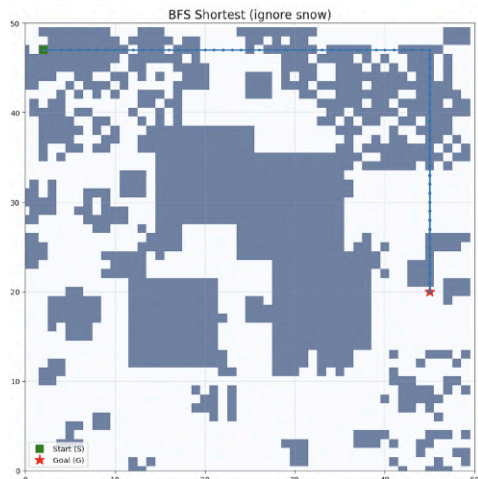
235 The grid layout, start and goal positions, and reward parameters were kept exactly the  
236 same across all methods. For each run, total reward, number of steps, average snow  
237 cells crossed per episode, and success in reaching the goal were recorded.

238

## 239 4.2 Trajectory Visualizations and Quantitative Results

240

### 241 4.2.1 Deterministic (Non-Slip) Winter Condition



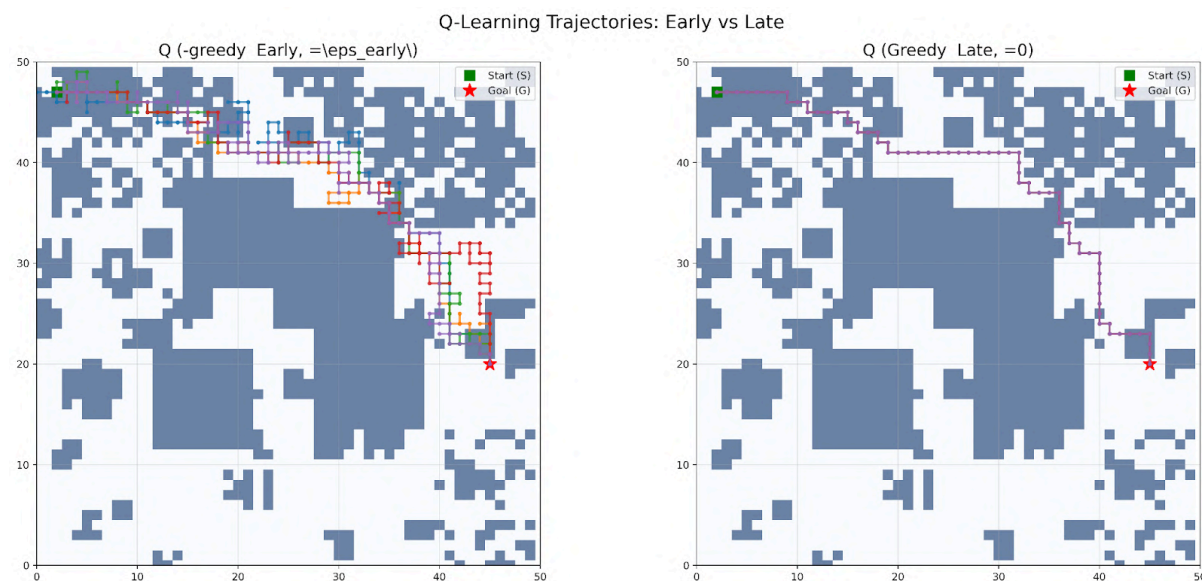
242

243 Figure 2: illustrates the BFS path under non-slip conditions. Since BFS always finds the  
244 shortest route, the trajectory goes straight from the start to the goal with minimal cells  
245 visited without any deviation.

246

247

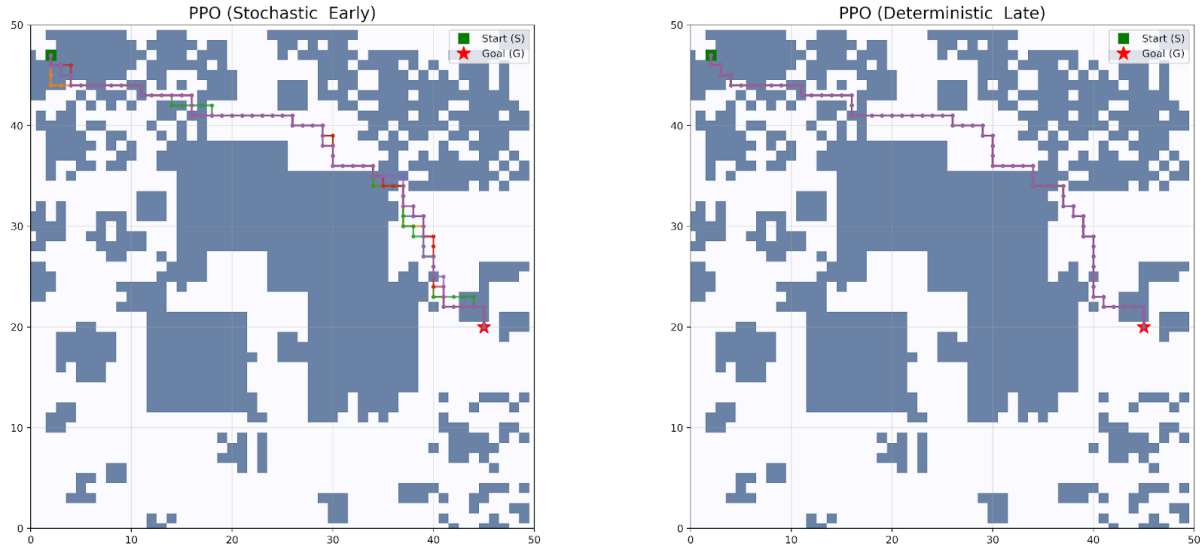
248



249

250 Figure 3: illustrates the navigation trajectories produced by the Q-learning agent under  
251 deterministic (non-slip) winter conditions during early training with an  $\epsilon$ -greedy policy and late  
252 training with a greedy policy. The trajectories demonstrate the transition from exploratory  
253 behavior to a stable path toward the goal.

PPO Trajectories: Pseudo Early vs Late



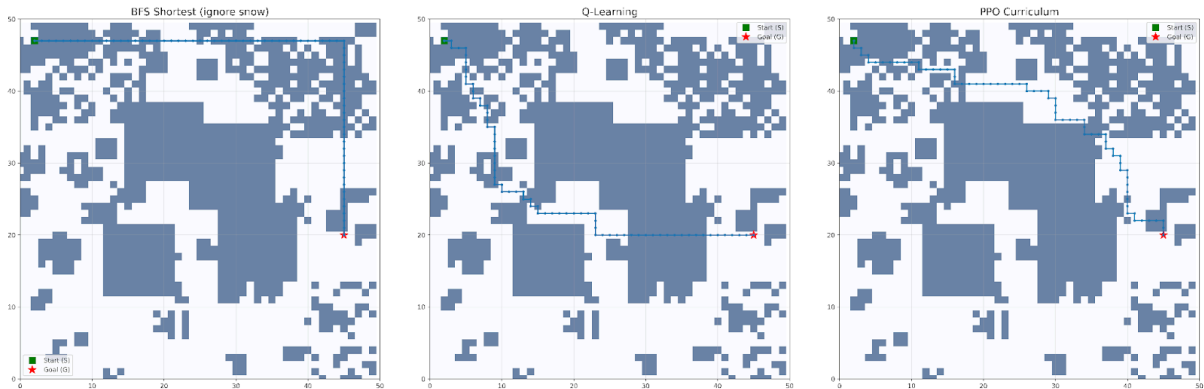
254

255 Figure 4: illustrates the navigation trajectories produced by the PPO agent during early  
256 and late stages of training under deterministic (non-slip) winter conditions.

257

258

3-Way Trajectory Comparison (Balanced Winter Grid)



259

260 Figure 5: compares the trajectories of BFS, Q-learning and PPO in the non-slip winter  
261 setting.

262

263

264 === Final Comparison Table ===

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	<b>575.30</b>	<b>70.00</b>	<b>31.00</b>	<b>100.0%</b>
<b>Q-Learning</b>	<b>589.50</b>	<b>70.00</b>	<b>7.00</b>	<b>100.0%</b>
<b>PPO Curriculum</b>	<b>587.50</b>	<b>70.00</b>	<b>11.00</b>	<b>100.0%</b>

265 Table 1: summarizes the main performance metrics for all three methods under  
266 deterministic conditions, including total reward, steps to goal, average snow cells  
267 crossed, and success rate.

268

269

## 270 4.2.2 Slip Winter Condition

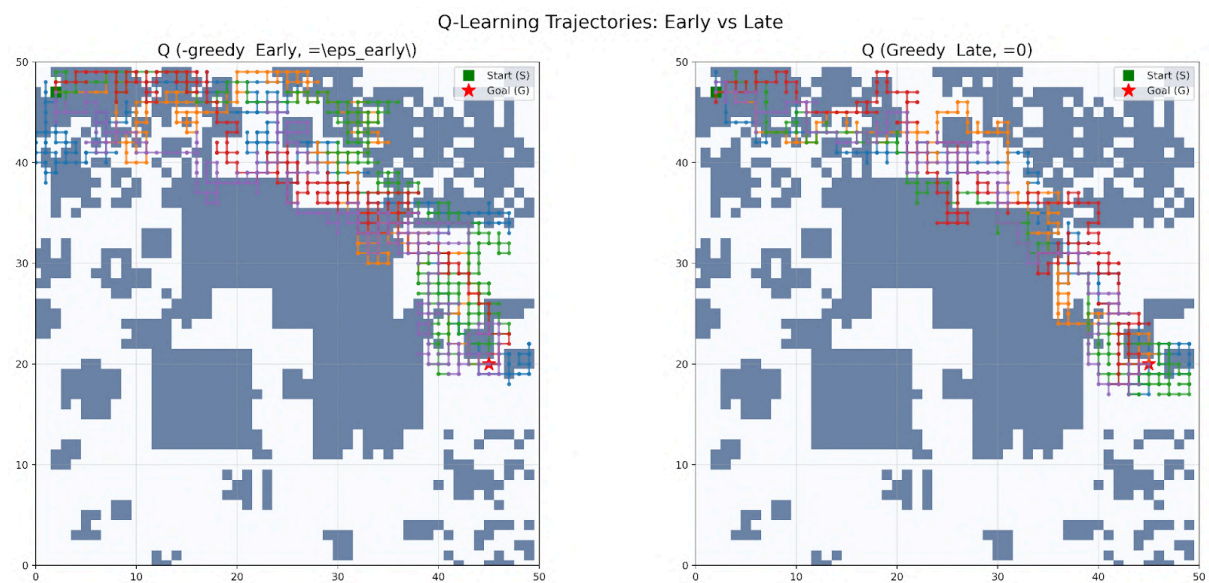


271

272 Figure 6: illustrates the navigation path produced by the BFS baseline under slip  
273 (stochastic) winter conditions.

274

275

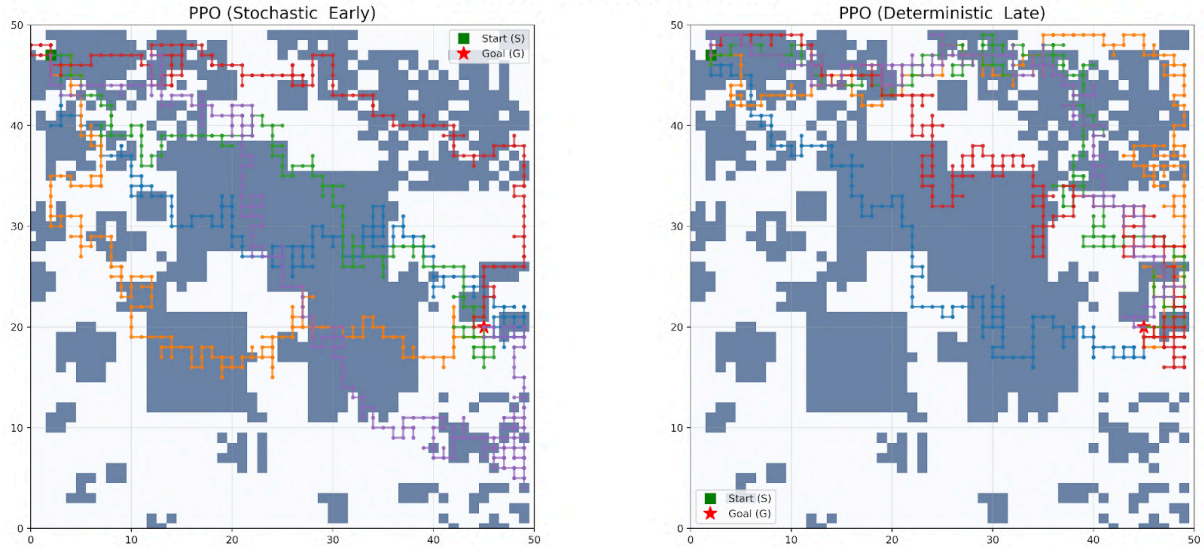


276

277 Figure 7: illustrates the navigation trajectories produced by the Q-learning agent during  
278 early and late stages of training under slip (stochastic) winter conditions.

279

PPO Trajectories: Pseudo Early vs Late



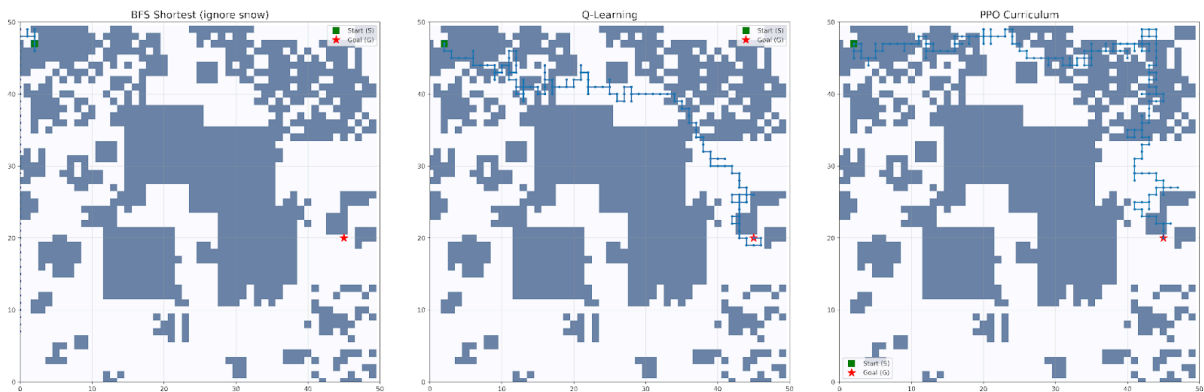
280

281 Figure 8: illustrates the navigation trajectories produced by the PPO agent during early  
282 and late stages of training under slip (stochastic) winter conditions.

283

284

3-Way Trajectory Comparison (Balanced Winter Grid)



285

286 Figure 9: compares the navigation trajectories of BFS, Q-learning, and PPO under slip  
287 winter conditions.

288

289

290 === Final Comparison Table ===

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	<b>-1639.20</b>	<b>1000.00</b>	<b>188.00</b>	<b>0.0%</b>
<b>Q-Learning</b>	<b>309.41</b>	<b>224.2</b>	<b>65.8</b>	<b>100.0%</b>
<b>PPO Curriculum</b>	<b>224.1</b>	<b>255.8</b>	<b>100.9</b>	<b>100.0%</b>

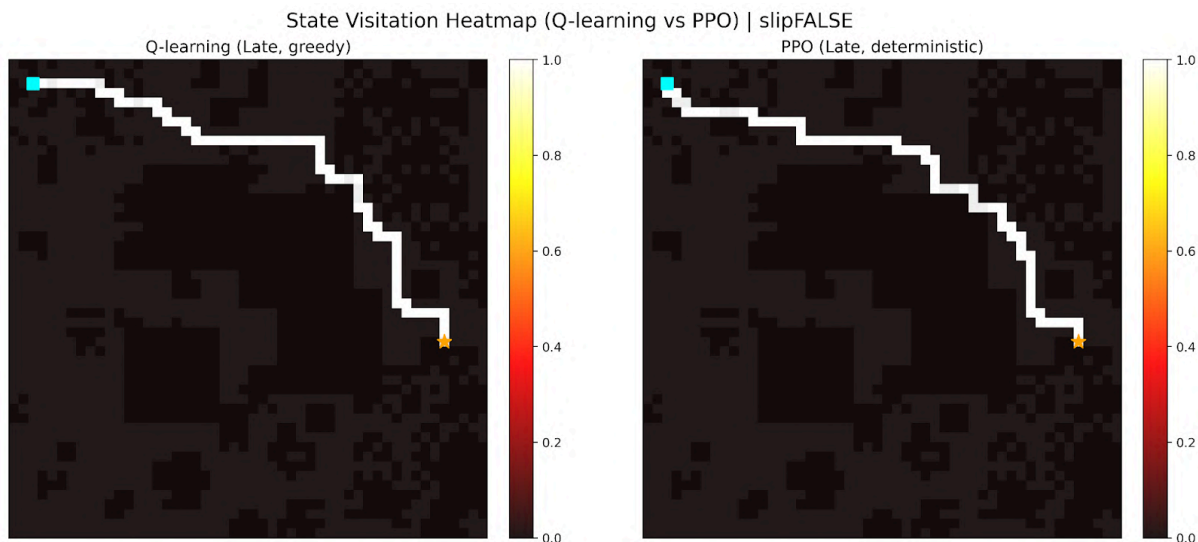
291 Table 2: summarizes the quantitative performance metrics for BFS, Q-learning, and PPO  
292 under slip winter conditions, using the same evaluation metrics as in the deterministic  
293 case.

294

295

### 296 4.3 Spatial State Visitation Heatmaps

297 Spatial state visitation heatmaps offer a grid-based way to see how often an agent  
298 passes through each cell during navigation. In our 50×50 winter grid, each cell  
299 corresponds to a state, and the color intensity reflects visitation frequency across  
300 episodes. Darker areas indicate cells that are visited more frequently, while lighter or  
301 white cells indicate rarely or never visited locations. Such visualizations give a clear  
302 picture of the agent's overall path distribution and behavior patterns. We generated  
303 these heatmaps using aggregated visitation counts normalized to the [0,1] range.

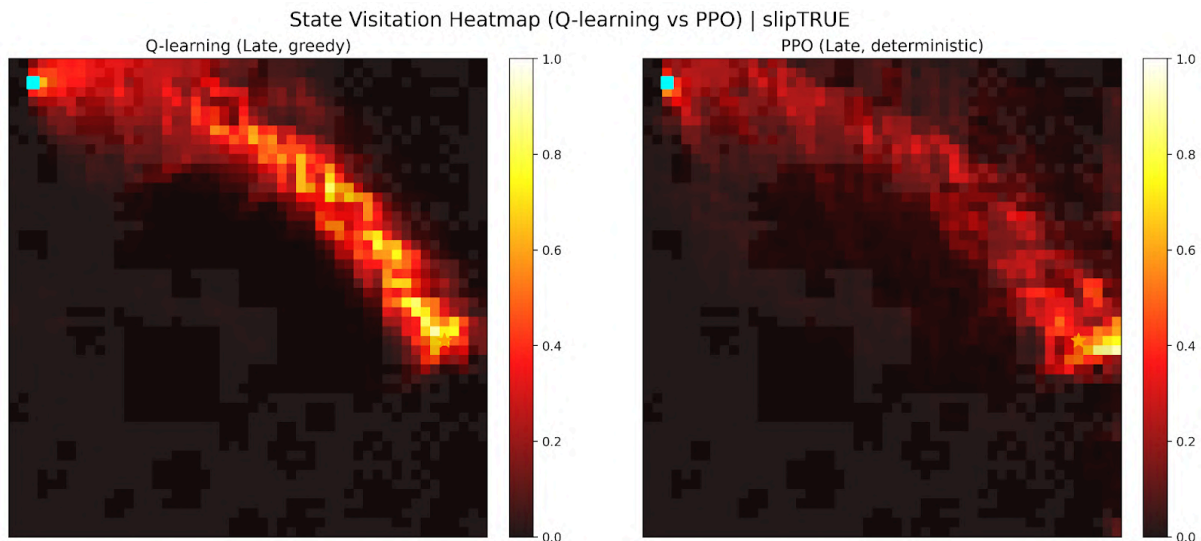


304

305 Figure 10: displays the visitation heatmaps for Q-learning and PPO under  
306 deterministic (non-slip) winter conditions. The results are aggregated  
307 from 10 evaluation episodes. Both agents concentrate most visits along a  
308 narrow corridor connecting the start to the goal, with near-maximum  
309 intensity ( $\approx 1.0$ ) along the primary route and very low visitation  
310 elsewhere.

311

312



313

314 Figure 11: shows the visitation intensity for each grid cell in the slip case. Compared with  
 315 the non-slip condition, visitation is more widely distributed across the grid rather than  
 316 remaining within a narrow corridor. The heatmap aggregates data from 10 evaluation  
 317 episodes.

318

#### 319 4.4 Limitations & Uncertainty

320 Several sources of uncertainty and limitations appeared in this study. Both Q-learning  
 321 and PPO showed noticeable performance differences across training runs, which was  
 322 expected given the stochastic nature of the environment and the learning algorithms.

323 Under slip conditions, the same policy could produce quite different routes from  
 324 episode to episode because actions did not always lead to the expected outcome. In  
 325 addition, stochastic exploration strategies and random initialization of value functions  
 326 (or policy networks) exposed agents to slightly different states, transitions, and rewards  
 327 across runs. This variability made it harder to get perfectly consistent results.

328 An unexpected issue was that even the BFS baseline sometimes failed to reach the goal  
 329 within the 1,000-step limit under slip conditions, resulting in a number of unsuccessful  
 330 evaluation episodes.

331

332 A full numerical breakdown of evaluation metrics is provided in Appendix B.

333

334

335

## 336 5. Discussion

### 337 5.1 Restatement of Hypothesis and Summary of Findings

338 The study hypothesized that reinforcement learning-based agents, particularly Proximal  
 339 Policy Optimization, would achieve more energy-efficient winter navigation paths than  
 340 a traditional shortest-path baseline and a value-based Q-learning agent under both  
 341 deterministic and slip winter conditions. The experiment result did not support the  
 342 hypothesis. Q-learning achieved higher cumulative reward than PPO under both

343 deterministic and slip conditions; it also did not outperform Q-learning in terms of  
344 cumulative reward or snow-cell avoidance. Both reinforcement learning methods  
345 perform better results than traditional BFS across these environments.

346

## 347 **5.2 Interpretation of Q-Learning and PPO Performance**

348 Q-learning achieved a higher performance than Proximal Policy Optimization agent in  
349 both deterministic and stochastic winter environments. One possible explanation was  
350 that a discrete and clear grid-world environment and relatively small environment  
351 would favor more for Q-learning, therefore allowing Q-learning to efficiently estimate  
352 optimal state [7]. In contrast, PPO relied on function approximation and stochastic  
353 policy updates, which may have required more training data or longer training time to  
354 converge to an equally optimal policy. Additionally, its stochastic policy may have  
355 introduced suboptimal choices that were not effective, therefore reducing its total  
356 rewards relative to Q-learning [3].

357

## 358 **5.3 Effects of Slip (Stochastic) Winter Conditions**

359 Under slip (stochastic) conditions, all agents showed broader trajectory dispersion and  
360 increased visits to previously visited states because of slipping, as reflected in the  
361 spatial state visitation heatmaps. This behavior is consistent with probabilistic transition  
362 dynamics, in which the same action could lead to different movement outcomes across  
363 episodes, such as deviating left in one run and right in another. When the grid is  
364 slippery, agents don't follow one stable path anymore. As a result, the routes spread out,  
365 and their performance becomes less consistent across episodes. These observations are  
366 consistent with prior navigation studies showing that stochastic environments increase  
367 policy variance and reduce convergence stability in reinforcement learning tasks [8].

368

## 369 **5.4 Interpretation of BFS Baseline Behavior**

370 The Traditional baseline trajectories did not take account for winter grid costs or  
371 stochastic transitions, it focused on the fastest way to the destination. In non-slip  
372 conditions this approach naturally produced the optimal route. However, under slip  
373 conditions, BFS exceeded the maximum step limit in multiple evaluation episodes,  
374 reflecting its inability to adapt its policy in response to transition winter grid costs or  
375 stochastic transitions.

376

377 Because BFS always assumed deterministic movement, each slip caused deviations from  
378 the planned path. These deviations often led to inefficient loops and longer paths.  
379 Unlike reinforcement learning methods, BFS did not have reward feedback or policy  
380 updates, which limited its ability to adjust navigation behavior under changing  
381 environmental dynamics. Therefore, making it the least effective method.

382

## 383 **5.5 Limitations and Future Directions**

384 Several limitations should be noted when interpreting these findings. All experiments  
385 were carried out in a simulated 50×50 grid environment with a relatively small and

386 simple state space, which differs significantly from real-world scenarios. This setup  
387 likely favored value-based methods like Q-learning, since it could learn precise  
388 state-action values for every cell. In contrast, PPO depends on neural network function  
389 approximation and stochastic policy updates, which may need more training steps or  
390 data to perform well in such discrete, small-scale settings.

391

392 The number of evaluation episodes was also limited, which reduced the statistical  
393 reliability of the results. Noticeable performance variability was observed across  
394 training runs, mainly due to the stochastic environment, random action outcomes, and  
395 probabilistic transitions under slip conditions.

396 The simulated environment simplified real-world winter driving conditions and did not  
397 capture vehicle dynamics, sensor noise, road-surface variation, or realistic  
398 energy-consumption models, these are factors that might lead to changes. As a result,  
399 these findings to real-world winter navigation scenarios were limited [9].

400 Future research could evaluate reinforcement learning agents in larger scale,  
401 continuous, or more realistic winter navigation environments with higher-dimensional  
402 state spaces and more complex conditions. Then it may be better to reflect real-world  
403 uncertainty and could allow policy-gradient methods such as PPO to fully function their  
404 theoretical advantages in handling stochastic and high-dimensional control problems.

405

406

---

407

## 408 **6. Conclusion**

409 This study compared a traditional BFS baseline with two reinforcement learning  
410 methods—Q-learning and Proximal Policy Optimization (PPO)—for energy-efficient  
411 navigation in a 50×50 winter grid under both deterministic (non-slip) and stochastic  
412 (slip) conditions. The results did not support the original hypothesis that PPO would  
413 produce more energy-efficient paths than Q-learning or BFS. Instead, Q-learning  
414 achieved the highest cumulative rewards in both settings and outperformed PPO in  
415 snow-cell avoidance. Although both reinforcement learning methods clearly  
416 outperformed the BFS baseline, BFS struggled under slip conditions due to its inability  
417 to adapt to stochastic transitions.

418

419 These findings indicate that in small and structured state spaces, value-based methods  
420 can be more effective than policy-gradient approaches when environmental complexity  
421 is limited. More generally, the results emphasize that the choice of reinforcement  
422 learning method should be guided by the structure and constraints of the task, rather  
423 than by theoretical advantages alone.

424

425

---

426

## 427 Reference

428

- 429 [1]Mao, R., Xu, W., Qian, Y., Li, X., Li, Y., Li, G., & Zhang, H. (2025).  
430 Understanding the Determinants of Electric Vehicle Range: A Multi-  
431 Dimensional Survey. *Sustainability*, 17(10), 4259. [https://doi.org/  
432 10.3390/su17104259](https://doi.org/10.3390/su17104259)
- 433 [2]Carlson, A., & Vieira, T. (2021). *The effect of water and snow on the  
434 road surface on rolling resistance* (VTI Report 971A). Swedish National  
435 Road and Transport Research  
436 Institute. [https://www.diva-portal.org/smash/get/diva2:1542142/  
437 FULLTEXT01.pdf](https://www.diva-portal.org/smash/get/diva2:1542142/FULLTEXT01.pdf)
- 438 [3]Watkins, C.J.C.H., Dayan, P. Q-learning. *Mach Learn* 8, 279–292  
439 (1992).  
440 <https://doi.org/10.1007/BF00992698>
- 441 [4]Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O.  
442 (2017). *Proximal Policy Optimization Algorithms* (No.  
443 arXiv:1707.06347). arXiv. <https://doi.org/10.48550/arXiv.1707.06347>
- 444 [5] Sutton, R. S., & Barto, A. G. (n.d.). Reinforcement Learning: An  
445 Introduction.  
446 [https://web.stanford.edu/class/psych209/Readings/  
447 SuttonBartoIPRLBook2ndEd.pdf](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)
- 448 [6]Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and  
449 Reinforcement Learning. *Neuron*, 36(2), 285–298. [https://doi.org/  
450 10.1016/S0896-6273\(02\)00963-7](https://doi.org/10.1016/S0896-6273(02)00963-7)
- 451 [7] Tan, C. (2025). Comparative Study of Reinforcement Learning  
452 Performance Based on PPO and DQN Algorithms. *Applied and  
453 Computational Engineering*, 175(1), 30–36. [https://doi.org/  
454 10.54254/2755-2721/2025.AST24879](https://doi.org/10.54254/2755-2721/2025.AST24879)
- 455 [8]Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino,  
456 A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., &  
457 Hadsell, R. (2017). *Learning to Navigate in Complex Environments* (No.  
458 arXiv:1611.03673). arXiv. <https://doi.org/10.48550/arXiv.1611.03673>
- 459 [9]Chukwurah, N., Adebayo, A. S., Ajayi, O. O., & Anfo Pub. (2024).  
460 *Sim-to-Real Transfer in Robotics: Addressing the Gap between  
461 Simulation and Real- World Performance*. 05(01), 33–39. [https://doi.org/  
462 10.54660/IJFMR.2024.5.1.33-39](https://doi.org/10.54660/IJFMR.2024.5.1.33-39)
- 463 [10]Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009).  
464 Curriculum learning. *Proceedings of the 26th Annual International  
465 Conference on Machine Learning*, 41–48. [https://doi.org/  
466 10.1145/1553374.1553380](https://doi.org/10.1145/1553374.1553380)
- 467  
468

469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484

## 485 **Appendices**

### 486 **Appendix A Experimental Configuration and Algorithmic Details:**

487

#### 488 **A.1 Environment Configuration**

489 All experiments were conducted in a custom 50×50 winter grid environment (2,500  
490 discrete states).

- 491 • Start position: (47, 2)
- 492 • Goal position: (20, 45)
- 493 • Maximum steps per episode: 1000
- 494 • Action space: {up, down, left, right}

495 Two transition settings were evaluated:

496 Deterministic (slip = FALSE)

497 The intended action is executed exactly.

498 Stochastic (slip = TRUE)

499 With probability 0.8, the intended action is executed.

500 With probability 0.2, the agent slips to a random adjacent direction.

501 All algorithms were evaluated on the same fixed grid layout to ensure fairness.

502

### 503 A.2 Reward Function

504 The reward function models energy-aware winter navigation.

505 At each time step:

- 506 • Step penalty:  $-1.5$
- 507 • Snow penalties:
  - 508 ○ Near snow:  $-0.2$
  - 509 ○ Edge snow:  $-0.5$
  - 510 ○ Core snow:  $-2.0$
- 511 • Goal reward:  $+700$

512 The cumulative episode reward is:

$$513 R = \sum_{t=1}^T (-1.5 - C_{snow}(s_t)) + 700 \cdot 1_{goal\ reached}$$

514

515 where  $C_{snow}(s_t)$  denotes the terrain penalty and

516  $1_{goal\ reached}$  indicates successful termination.

### 517 A.3 Q-Learning Configuration

518 Tabular Q-learning was implemented with the following hyperparameters:

- 519 • Learning rate  $\alpha=0.15$
- 520 • Discount factor  $\gamma=0.99$
- 521 • Exploration strategy:  $\epsilon$ -greedy
- 522 • Initial  $\epsilon = 0.60$
- 523 • Minimum  $\epsilon = 0.05$
- 524 • Exponential decay per episode =  $0.9995$
- 525 • Training episodes =  $15,000$
- 526 • Max steps per episode =  $1000$

527 The Q-update rule is:

$$528 Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

529 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) was used.

530

## 531 A.4 Proximal Policy Optimization (PPO) Curriculum Configuration

532 PPO was implemented as a stochastic policy-gradient method.

533 Total training timesteps: 1,200,000

- 534 • Phase 1 (deterministic environment): 300,000 timesteps
- 535 • Phase 2 (stochastic environment): 900,000 timesteps

536 Evaluation was conducted using deterministic action selection.

537

538 PPO Objective Function

539 PPO optimizes the clipped surrogate objective:

$$540 L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)]$$

541 where

$$542 r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

543 and  $\hat{A}_t$  is the generalized advantage estimate.

544 The clipping mechanism constrains policy updates to maintain stability.

545

546 PPO Hyperparameters

- 547 • Discount factor  $\gamma=0.99$
- 548 • GAE parameter  $\lambda=0.95$
- 549 • Entropy regularization enabled
- 550 • Neural network architecture: two hidden layers (512 units each, ReLU activation)

551

## 552 A.5 Evaluation Protocol

553 For each condition (slip FALSE / TRUE):

- 554 • One trained model per algorithm
- 555 • Maximum evaluation length: 1000 steps
- 556 • Metrics recorded:

- 557 ○ Total cumulative reward
- 558 ○ Number of steps
- 559 ○ Snow cell visits
- 560 ○ Success rate (%)

561 State visitation heatmaps were generated by aggregating visitation frequencies across  
 562 evaluation runs.

563

---

## 564 **Appendix B Detailed Quantitative Results:**

### 565 **B.1 Deterministic (slip = FALSE)**

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	<b>575.30</b>	<b>70.00</b>	<b>31.00</b>	<b>100.0%</b>
<b>Q-Learning</b>	<b>589.50</b>	<b>70.00</b>	<b>7.00</b>	<b>100.0%</b>
<b>PPO Curriculum</b>	<b>587.50</b>	<b>70.00</b>	<b>11.00</b>	<b>100.0%</b>

566 All methods reached the goal within 70 steps.

567 Q-learning achieved the highest cumulative reward due to reduced snow traversal.

568

### 569 **B.2 Stochastic (slip = TRUE)**

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	<b>-1639.2</b>	<b>1000</b>	<b>188.0</b>	<b>0%</b>
<b>Q-Learning</b>	<b>309.4</b>	<b>224.2</b>	<b>65.8</b>	<b>100.0%</b>
<b>PPO Curriculum</b>	<b>224.1</b>	<b>255.8</b>	<b>100.9</b>	<b>100.0%</b>

570 Under stochastic dynamics:

- 571 ● BFS fails due to inability to adapt.
- 572 ● Q-learning demonstrates greater robustness.
- 573 ● PPO maintains success but exhibits higher trajectory dispersion.

574

### 575 **B.3 State Visitation Analysis**

576 State visitation frequency was computed as:

577 
$$V(s) = \frac{N(s)}{\max_s N(s)}$$

578 where  $N(s)$  is the number of visits to state  $s$ .

- 579 • Under deterministic conditions, visitation concentrates along a narrow corridor.
- 580 • Under stochastic conditions, visitation becomes more dispersed.
- 581 • Q-learning exhibits more focused routing than PPO under slip conditions.

## Review: Energy-Efficient Path Planning in Winter Conditions

This paper addresses a timely and relevant problem (energy-efficient routing for electric vehicles under winter conditions) by comparing BFS, Q-learning, and PPO in a controlled grid-world environment. The experimental design is a clear strength: using identical configurations, reward structures, and evaluation metrics allows for a fair comparison across methods. I also particularly appreciate the inclusion of a traditional BFS baseline, which provides helpful context for interpreting the performance gains of different reinforcement learning approaches. Additionally, the comparison between deterministic (“non-slip”) and stochastic (“slip”) environments helps provide additional comparisons and connects the abstraction back to meaningful real world variables.

Furthermore, this paper offers interesting insights into when value-based methods (Q-learning) may outperform policy-based approaches (PPO), especially in structured, discrete environments. This contributes to a broader understanding of how different RL paradigms perform across task types and environmental complexity.

That said, several areas could be strengthened. First, the introduction would benefit from more consistent and rigorous citation support. For example, claims such as “rolling resistance can rise by approximately 30–40% under wet or snowy conditions” and statements about the validity of the grid abstraction would be more convincing if directly supported with citations. Strengthening this connection to prior work would improve both credibility and scholarly rigor.

Second, while the motivation is clearly articulated, the paper would benefit from a more explicit discussion of gaps in existing research. The authors highlight the importance of multi-criteria navigation under winter conditions, but it remains unclear what specific limitation in the current literature this study addresses. Clarifying whether the primary contribution is extending RL-based navigation to account for weather-related energy costs (or something more novel) would help better position the work.

Third, aspects of methodological rigor could be improved. The paper would benefit from more justification of hyperparameter choices and additional experimental robustness, such as averaging results over multiple runs and reporting variance. This would strengthen confidence in the findings, especially given the stochastic elements of the environment.

Overall, this is a thoughtfully designed study with promising contributions. With improved citation support, clearer positioning within the literature, and stronger experimental validation, the paper could make a meaningful contribution. I recommend **accept with major revisions**.

**Decision: Accept with (minor) revisions**

**Review:**

The paper compares three ways of planning routes in a winter grid world: a shortest-path BFS baseline, Q-learning, and PPO, using the same map, rewards, start and goal, and evaluation metrics for all three. Across both non-slip and slip settings, Q-learning comes out best overall: in the deterministic case, it reaches the goal with fewer snow-cell visits than the others, and in the slippery case, it clearly beats PPO while BFS breaks down. The paper's main conclusion is that, in a small structured environment like this one, value-based reinforcement learning can be more effective than PPO for energy-aware winter navigation, though the setup is still a simplified simulation rather than a real driving model.

Its strongest feature is the head-to-head comparison itself: the paper puts BFS, Q-learning, and PPO in the same winter grid setup and evaluates them with the same reward structure and metrics, so the differences between the methods come through clearly. The paper's main argument is generally clear, and the writing is easy to understand. The Results and Discussion sections are the strongest, with direct and specific comparisons. In contrast, the Methods section needs the most improvement, especially the explanations of the environment, how rewards were given, and the training procedures. That is where the paper becomes most repetitive and hardest to follow. Tightening that section and making the setup more direct would do the most to improve the paper.

The paper follows in a sensible order overall, and the main steps of the argument are easy to follow. The paper moves from introduction to methods to results in a straightforward way, and no major reordering is needed- besides the method section. The paper includes a Works Cited / References section, and there are corresponding citations throughout the paper. The main issue is minor consistency and formatting, not the absence of citation.

The Methods section would benefit from a clearer presentation of the experimental setup. At present, important details are split between the main text and Appendix A. I would suggest bringing the core information together in one place and stating it plainly: the 50×50 grid, the start and goal positions, the slip probabilities, the reward and penalty values, the training length for Q-learning and PPO, and the evaluation procedure. It would also help to replace general phrases such as “large rollout buffers” and “multiple optimization epochs” with the actual PPO settings, and to state directly how many evaluation episodes were used. This is a manageable revision, but it would make the paper much easier to follow.

The paper is already in good standard, but the above suggested changes will make the article easier to read. The paper can be easily accepted after the revision.

# 1 Energy-Efficient Path Planning in Winter Conditions: A Comparative 2 Study of Traditional Baseline, Q-learning, and Proximal Policy 3 Optimization in a Grid World Environment

4  
5

## 6 1.Abstract

7 Recent investigation in winter route planning has become increasingly important due to  
8 increased resistance and reduced efficiency on winter roads. These conditions  
9 significantly increase energy consumption and introduce uncertainty, raising the risk of  
10 failing to reach the destination. To address the challenge, Reinforcement Learning (RL)  
11 is a type of machine learning where an agent learns to make decisions by interacting  
12 with an environment, receiving rewards or penalties for its actions to maximize  
13 cumulative rewards. For simulation, we benchmark a standard shortest-path algorithm  
14 (BFS) against two reinforcement learning methods: Q-learning and Proximal Policy  
15 Optimization (PPO). All approaches are tested on identical grid configurations, with the  
16 same starting points, goals, and reward functions. We assess their performance based on  
17 cumulative reward, snowy cell visits (as a proxy for energy cost), trajectory characteristics,  
18 and success rate. It is hypothesized that PPO algorithms will result in lower energy  
19 consumption than Q-learning and traditional baseline under winter conditions. The  
20 results show that while BFS consistently finds the shortest paths, it fails to consider  
21 energy costs or environmental uncertainty. RL agents, by comparison, adapt more  
22 efficiently to winter conditions. Under deterministic (non-slip) conditions, PPO  
23 achieved higher cumulative reward and fewer snow cell visits than Q-learning, partially  
24 supporting the hypothesis. However, under stochastic (slip) conditions, Q-learning  
25 outperformed PPO in cumulative reward and snow-cell avoidance. These results  
26 suggest that Q-learning is better suited for stochastic winter navigation in grid worlds,  
27 while PPO may perform better in more deterministic environments with continuous  
28 states or decisions.

29

30 Key words: Energy-Efficient, Traditional Baseline, Q-learning, Proximal Policy  
31 Optimization, grid world, reinforcement learning.

32

33

---

34

## 35 2.Introduction

36 As demand in electric vehicles become increasingly widespread, energy efficiency and  
37 route optimization have emerged as critical challenges, particularly in winter when  
38 resistance increases. Freezing temps, plus snow and ice covering the roads, push energy  
39 consumption way up, which means shorter range and trips that feel less predictable. A  
40 2025 study shows that an estimated 50 % of EV driving range can be reduced in cold  
41 climates, including snow and ice covering terrain, highlighting the significant impact of  
42 environmental conditions on energy consumption [1]. Furthermore, when snow or

43 water remains on the road surface, vehicle tires must continuously displace through ice  
44 as they roll and move, hence forcing the vehicle to draw more power [2]. On top of that,  
45 water cools tires more effectively than air alone, altering their mechanical properties  
46 and further pushing rolling resistance higher. Previous studies have shown that rolling  
47 resistance can rise by approximately 30% to 40% under wet or snowy conditions [2].  
48 Ultimately, these factors make winter driving a major concern for Electric Vehicle (EV)  
49 efficiency and reinforce the need for energy-aware routing strategies.

50

51 To approach these challenges, we have explored a few computational methods to allow  
52 agents to learn effective navigation strategies aimed at minimizing energy consumption.  
53 In this study, a 50x50 grid is used as an abstraction routing map that circles key  
54 structures of Canadian snow distribution, rather than exact geographical terrain.  
55 Grid-based environments are commonly used in reinforcement learning studies, as they  
56 simplify complex continuous real-world environments into discrete states [5]. This  
57 abstraction is particularly well-suited for this study, as it enables environmental factors  
58 such as snow coverage and surface slip conditions to be incorporated directly into the  
59 cell-level cost and reward structure, supporting controlled and reproducible  
60 comparison of different routing methods.

61

62 Despite growing interest in reinforcement learning for navigation, existing studies have  
63 largely overlooked the impact of winter environmental conditions, such as snow  
64 coverage and surface slippiness, on energy-aware path planning. Prior work has  
65 primarily focused on general navigation without taking account of explicitly modeling  
66 weather-dependent energy costs [8]. To address this gap, this study integrates snow  
67 coverage, stochastic slip condition, and energy costs into a comparative framework,  
68 evaluating Q-learning, PPO, and a traditional Baseline under realistic winter routing  
69 scenarios. This study is one of the first to directly compare value-based and  
70 policy-gradient reinforcement learning methods in an energy-aware winter navigation  
71 setting .

72

73 More specifically, the research will include a traditional baseline algorithm, and two  
74 types of reinforcement learning, which are Q-learning and PPO algorithm. The  
75 traditional baseline routing computes the shortest path based on static cost metrics and  
76 follows the predefined route without adaptation or learning (basic routing). Q-learning,  
77 a value-based reinforcement learning method, learns through incremental dynamic  
78 programming processes with computational requirements and it is well suited for  
79 agents to improve and refine action values and achieve effective performance in  
80 controlled Markovian domains [3]. In contrast, PPO is a policy-gradient algorithm that  
81 primarily optimizes a stochastic policy and achieves greater training stability through  
82 constrained policy updates with a clipped surrogate objective function [4].

83

84 This comparison evaluates the effectiveness of the proposed routing methods in  
85 identifying energy-efficient paths. Performance is evaluated through energy related

86 costs, overall routing efficiency, and observed navigation behavior under both  
87 deterministic (non-slip) and stochastic (slip) settings. Because PPO balances stable  
88 policy updates with adaptive learning in uncertain environments, it is hypothesized to  
89 be the most effective method implemented for winter routing.

90

91 The following section goes into the methodology and experimental setup in more detail.

92

93

---

94

## 95 **3.Methods**

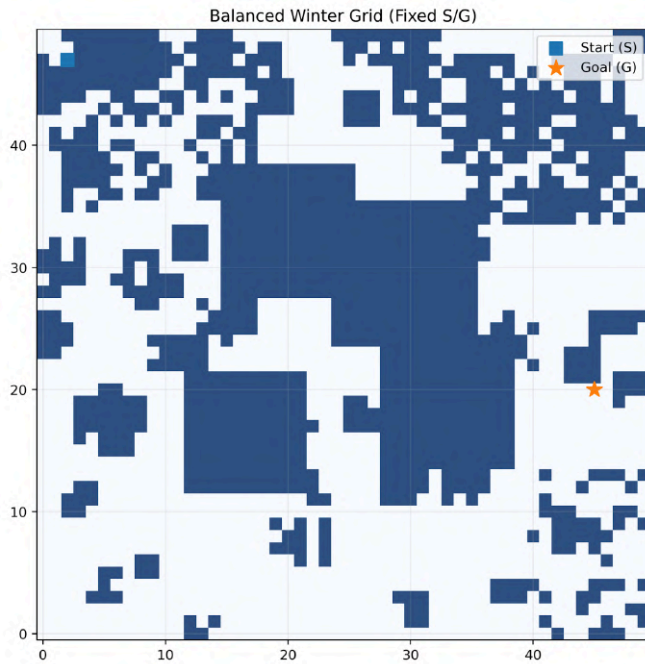
### 96 **3.1 Environment Design**

97 A custom winter navigation grid environment was designed to evaluate each  
98 Reinforcement Learning (RL) agent under stochastic and cost-sensitive conditions. The  
99 environment consists of a 50 x 50 grid world containing 2500 cells, including both  
100 normal terrain and snow-covered terrain. The simulation contains a fixed start (47, 2)  
101 and a fixed goal (20, 45) point, and are located in the top left corner and middle right  
102 respectively. At each time step, the agent can take either one of the four directions, up,  
103 down, left, and right; each step on a non-snow grid has an energy cost of -1.5 per step.  
104 Snow is modeled as spatially correlated regions with graded intensity, capturing the typical  
105 uneven accumulation patterns in Canadian winters.

106 When snow is modeled as a binary state, abrupt reward discontinuities make PPO's  
107 advantage estimates unstable, resulting in increased gradient variance. By contrast, a  
108 continuous representation of snow thickness provides smoother cost transitions, which  
109 help reduce gradient variance and stabilize learning, making it more efficient. These  
110 snow types are classified as "Near snow", "Edge snow", and "Core snow".

111 To evaluate robustness under different winter conditions, two transition settings were  
112 considered using the same grid map: a non-slip (deterministic) condition and a slip  
113 (stochastic) condition. In the deterministic setting, action always results in the intended  
114 movement, and snow cells only add additional energy loss. In the stochastic setting,  
115 actions do not always lead to the intended outcome: with a probability of 1/3, the  
116 intended action is executed, and with a probability of 2/3, the agent moves in a random  
117 adjacent direction, simulating loss of traction control during winter driving on ice and  
118 snow.

119



120

121 Figure 1: shows a typical winter grid layout used in all experiments.

122

### 123 3.2 Reinforcement Learning Algorithms/ Methodologies:

124 The routing strategies and learning methods used in this study are described in this  
125 section.

126

#### 127 3.2.1 Traditional Baseline (BFS)

128 Traditional baseline is a non-learning shortest path strategy used as a comparison base  
129 for reinforcement learning methods. It operates on the 50x50 grid map, it considers  
130 four directions to move (up, down, left, right). The system calculates the shortest path  
131 to the destination, without considering energy savings, which means it does not adapt  
132 to environmental feedback or uncertainty.

133

#### 134 3.2.2 Q-learning Implementation

135 Tabular Q-learning was used as the value-based reinforcement learning baseline. The agent  
136 maintains a discrete Q (s, a) table, which is updated iteratively based on observed state transitions.  
137 An  $\epsilon$ -greedy policy was used for exploration. The value of  $\epsilon$  begins at 0.6 and decays  
138 exponentially with a factor of 0.9995 per episode until it reaches 0.05. In this 50x50 grid  
139 setting, the decay schedule allows broad initial exploration before shifting emphasis to  
140 exploitation. Actions are restricted to four possibilities—left (0), down (1), right (2), and  
141 up (3)—consistent with standard discrete grid-world formulations and the discrete grid  
142 structure. The Q-value update follows the classic temporal-difference form:

143

$$144 Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)] \quad [5]$$

145

146 We fix the learning rate  $\alpha$  at 0.15 and the discount factor  $\gamma$  at 0.99. These values handle the  
147 stochastic transitions (`is_slippery=True`) and support planning over long horizons in the large  
148 grid [3]. State  $s$  denotes the flattened grid position index. The action  $a$  selects one of the four  
149 movement directions.  $Q(s, a)$  represents the estimated discounted cumulative reward  
150 starting from state  $s$  and action  $a$  [3,5].

151

### 152 3.2.3 Reward system

153 In this research, the reward system is designed to represent energy efficiency under  
154 winter conditions, it helps agents find the best paths. At each step, the system gives a  
155 negative value to show how significant a step is to energy saving, with higher cause on  
156 snow grids. A positive 700 is given only when the agents successfully reach the  
157 destination, while within the maximum steps of 1000 steps. It encourages agents to  
158 reach the destination with considerations on energy saving. We initialize the cumulative  
159 reward to 0 at the beginning of each episode. Each will result in negative values,  
160 because by doing this, agents don't need to consider the prior bias, since it assumes  
161 nothing at first, etc. As in the settings, we have Near-snow (light snow), Edge-snow  
162 (medium snow), and Core-snow (deep snow); each of them has different additional  
163 values which are -0.2, -0.5, and -2.0 respectively [6].

164

$$165 r_t = r_{step} - c_{energy}(s_t) + r_{goal} \quad [6]$$

166

167 The basic reward function follows standard reinforcement learning practice by  
168 combining step penalties, energy-related costs, and a terminal goal reward.

169

### 170 3.2.4 Proximal Policy Optimization

171 Proximal Policy Optimization (PPO) served as the policy-gradient method in this study  
172 for energy-efficient routing under winter conditions. PPO learns a stochastic policy by  
173 directly outputting action probabilities for each state, which helps the agent cope with  
174 uncertainty on slippery roads.

175

$$176 L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \quad [4]$$

177

178 where  $r_t(\theta)$  is the probability ratio between the new and previous policies. The clipping  
179 mechanism constrains policy updates to improve stability.

180

181 The PPO agent was trained using fixed values across all experiments. The discount factor was set  
182 to  $\gamma=0.99$ , and Generalized Advantage Estimation (GAE) was used with  $\lambda=0.95$ . The policy  
183 network consisted of two fully connected hidden layers with 256 units each, using ReLU  
184 activation functions. The learning rate was set to  $3 \times 10^{-4}$ , with a clipping range of 0.2  
185 and an entropy coefficient of 0.02 to encourage exploration. PPO training was

186 conducted for 1,200,000 time steps, using a rollout buffer of 2048 steps, a batch size of  
187 256, and 10 optimization epochs per update to improve training stability [4].

188

189 Policy-gradient methods work by directly adjusting a parameterized policy to increase  
190 expected reward. They learn a stochastic policy, allowing for more flexibility in  
191 uncertain environments, where the same action can lead to different outcomes. In  
192 contrast, q-learning estimates state-action values and usually converges with a  
193 deterministic policy based on these estimates. Although it can learn in unfamiliar  
194 environments, its action selection may be less reliable in highly unpredictable  
195 situations. As a result, policy-gradient and value-based methods show varying strengths  
196 depending on the level of uncertainty in the winter routing task.

197

### 198 **3.2.5 Curriculum Learning for PPO**

199 PPO training followed a two-phase curriculum learning approach designed to improve  
200 training stability and sample efficiency under sparse rewards and energy-sensitive  
201 penalties [10].

202

#### 203 **Phase 1: Navigation Learning Phase** (300,000 timesteps)

204 In phase 1, PPO was trained using a simplified reward function that focuses only on  
205 positive feedback for reaching the goal. Energy costs in snow grids were not included.  
206 This made the feedback denser and more immediate. The agent quickly learned basic  
207 navigation and achieved early success in the 50×50 winter grid. The policy learned in  
208 this phase served as the starting point for the next training stages.

209

#### 210 **Phase 2: Energy-Aware Optimization Phase** (900,000 timesteps)

211 In phase 2, the training continued with the complete energy-aware reward function,  
212 which now included penalties for moving across snow-covered areas. The agent had to  
213 balance successful arrival with lower energy consumption. All reported results, ablation  
214 studies, and comparisons are based solely on the Phase 2 reward function.

215

### 216 **3.3 Evaluation factors**

217 To ensure statistical reliability, all reinforcement learning experiments were repeated  
218 over 5 independent random seeds, and results are reported as the mean and standard  
219 deviation across runs.

220

221 To compare the performances of all routing methods under identical conditions, a final  
222 evaluation was conducted after training. Each method BFS, Q-learning, and PPO was  
223 evaluated using the following metrics:

224

225 ● Total Reward: calculated as the cumulative reward obtained over the episode.

226

227 ● Number of steps: steps taken from the start to the goal (less than 1000 steps).

228

229 • Average snow cell visited: Snow visits were calculated by counting the number of  
230 snow-covered cells traversed in each episode and averaging this value across all  
231 evaluation episodes.

232

233 • Success rate: defined as the percentage of episodes in which the agent reached  
234 the goal within the maximum step limit.

235

236 The BFS baseline was evaluated only once on the fixed grid map, as it produces a  
237 deterministic path. Reinforcement learning agents were evaluated over 200 episodes  
238 per run, with experiments repeated over 5 independent random seeds for each  
239 condition (slip and non-slip). All methods were compared using the same energy-aware  
240 reward function and environment configuration to ensure fairness.

241

### 242 **3.4 Hyperparameter Selections**

243 Hyperparameters are selected based on standard practices and empirical validation to  
244 ensure stable and consistent performance [3,4].

245

#### 246 **3.4.1 Q-Learning Hyperparameter Configuration**

247 We implement tabular Q-learning with an  $\epsilon$ -greedy exploration strategy.

248

249 The hyperparameters are set as follows:

250 - Learning rate  $\alpha = 0.15$

251 - Discount factor  $\gamma = 0.99$

252 - Exploration strategy:  $\epsilon$ -greedy

253 - Initial  $\epsilon = 0.60$

254 - Minimum  $\epsilon = 0.05$

255 - Exponential decay rate = 0.9995

256 - Training episodes = 15,000

257 - Maximum steps per episode = 1,000

258

259 The learning rate ( $\alpha = 0.15$ ) is selected to balance convergence speed and stability. A moderate  
260 learning rate allows the agent to adapt efficiently while avoiding oscillations in value updates.

261

262 The discount factor ( $\gamma = 0.99$ ) emphasizes long-term rewards, which is essential for navigation  
263 tasks where the objective is to reach the goal efficiently over multiple steps.

264

265 An  $\epsilon$ -greedy exploration strategy is adopted to balance exploration and exploitation. The initial  
266 exploration rate ( $\epsilon = 0.60$ ) encourages sufficient exploration in early training, while exponential  
267 decay gradually shifts the policy toward exploitation. The minimum  
268 ( $\epsilon = 0.05$ ) ensures that some level of exploration is maintained throughout training, preventing  
269 the agent from getting stuck in suboptimal policies.

270

271 The number of training episodes 15,000 and maximum steps per episode 1,000 are  
272 chosen to ensure sufficient interaction with the environment for convergence, while  
273 maintaining computational efficiency.

274

275 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) is used to assess the learned policy performance.

276

### 277 **3.4.2 PPO Hyperparameter Configuration**

278 We implement PPO using a clipped surrogate objective with generalized advantage  
279 estimation (GAE).

280

281 The hyperparameters are set as follows:

282 - Discount factor  $\gamma = 0.99$

283 - GAE parameter  $\lambda = 0.95$

284 - Learning rate =  $3 \times 10^{-4}$

285 - Clipping range  $\epsilon = 0.2$

286 - Rollout buffer size = 2048 steps

287 - Batch size = 256

288 - Optimization epochs per update = 10

289 - Entropy regularization is enabled

290

291 The policy and value networks share a neural architecture consisting of two hidden  
292 layers with 256 units each and ReLU activation.

293

294 The discount factor ( $\gamma = 0.99$ ) is selected to emphasize long-term rewards, which is essential for  
295 navigation tasks. The GAE parameter ( $\lambda = 0.95$ ) provides a balance between bias and variance in  
296 advantage estimation, leading to more stable learning.

297

298 The clipping parameter ( $\epsilon = 0.2$ ) constrains policy updates to prevent large deviations from the  
299 previous policy, which improves training stability. The rollout buffer size and number of  
300 optimization epochs are chosen to ensure sufficient policy updates while avoiding overfitting to  
301 recent trajectories.

302

303 Entropy regularization is included to encourage exploration and prevent premature  
304 convergence to suboptimal policies.

305

306 Overall, these hyperparameters follow widely adopted settings in reinforcement  
307 learning literature and are chosen to ensure stable and consistent training rather than  
308 aggressive performance tuning.

309

310 A complete summary of environment settings and hyperparameters is provided in  
311 Appendix A.

312

313

314

## 315 4. Results

316 The experiments compare BFS, Q-learning, and PPO on the same 50×50 winter grid,  
317 looking at both non-slip (deterministic) and slip (stochastic) cases.

318

### 319 4.1 Evaluation Setup and Metrics

320 The grid layout, start and goal positions, and reward parameters were kept exactly the  
321 same across all methods. All experiments were repeated over 5 independent random  
322 seeds, and results are reported as the mean across runs. Trajectory figures shown are  
323 from the representative seed whose performance was closest to the mean. For each run,  
324 total reward, number of steps, average snow cells crossed per episode, and success in  
325 reaching the goal were recorded.

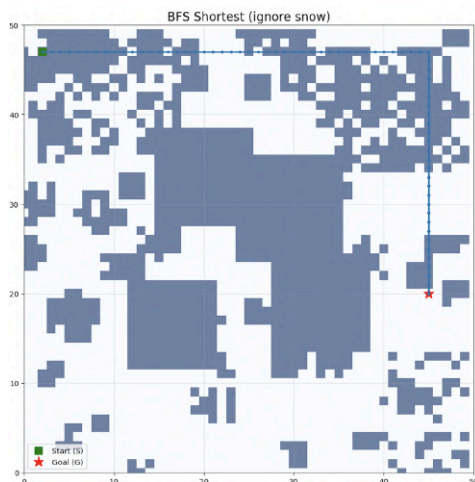
326

### 327 4.2 Trajectory Visualizations and Quantitative Results

#### 328 4.2.1 Deterministic (Non-Slip) Winter Condition

329

330



331

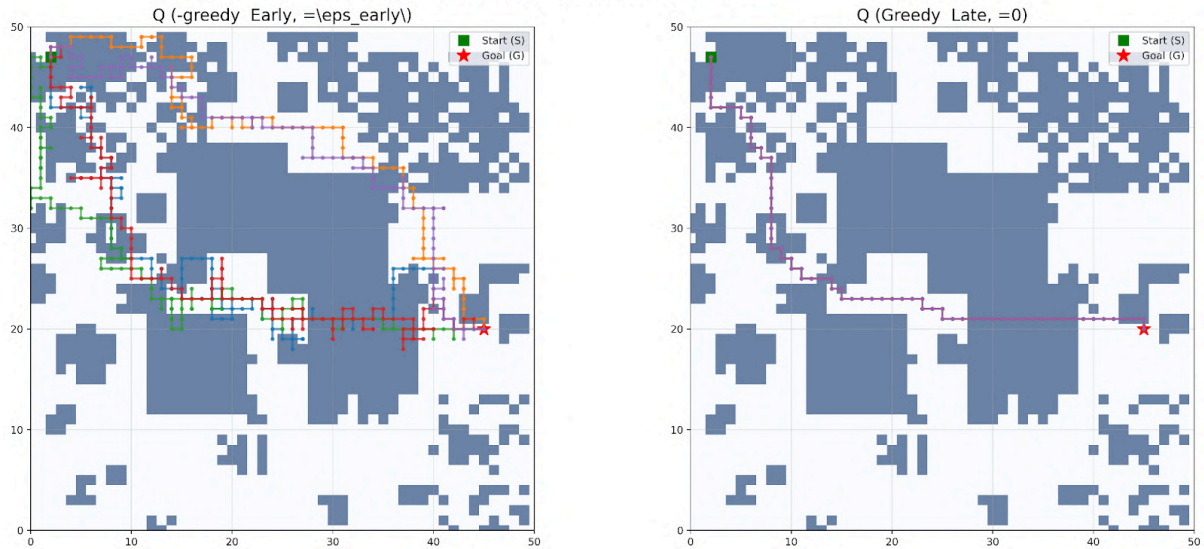
332 Figure 2: illustrates the BFS path under non-slip conditions. Since BFS always finds the  
333 shortest route, the trajectory goes straight from the start to the goal with minimal cells  
334 visited without any deviation.

335

336

337

Q-Learning Trajectories: Early vs Late



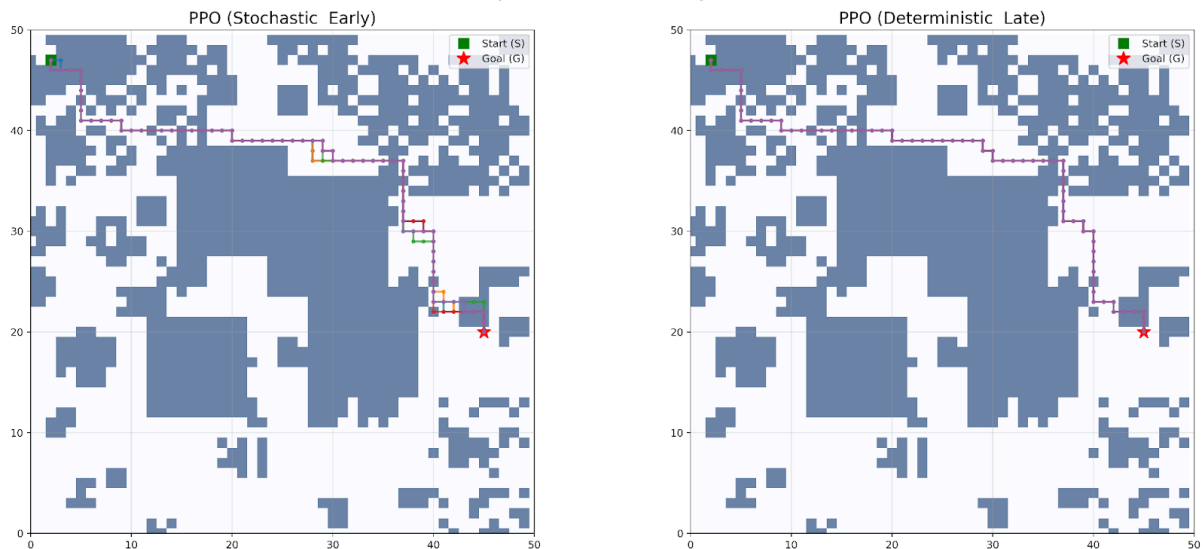
338

339 Figure 3: illustrates the navigation trajectories produced by the Q-learning agent under  
340 deterministic (non-slip) winter conditions during early training with an  $\epsilon$ -greedy policy and late  
341 training with a greedy policy. The trajectories demonstrate the transition from exploratory  
342 behavior to a stable path toward the goal.

343

344

PPO Trajectories: Pseudo Early vs Late



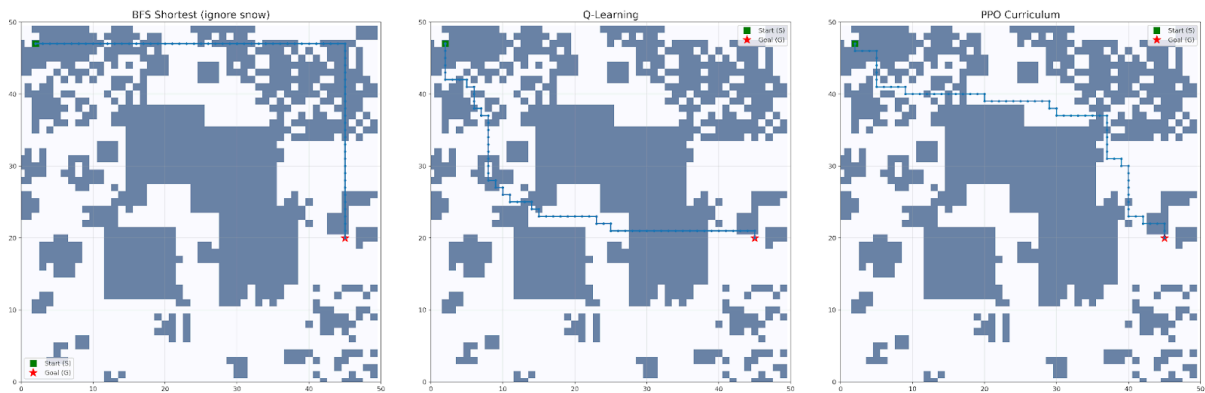
345

346 Figure 4: illustrates the navigation trajectories produced by the PPO agent during early  
347 and late stages of training under deterministic (non-slip) winter conditions.

348

349

3-Way Trajectory Comparison (Balanced Winter Grid)



350

351 Figure 5: compares the trajectories of BFS, Q-learning and PPO in the non-slip winter  
 352 setting.

353

354

355 === Final Comparison Table ===

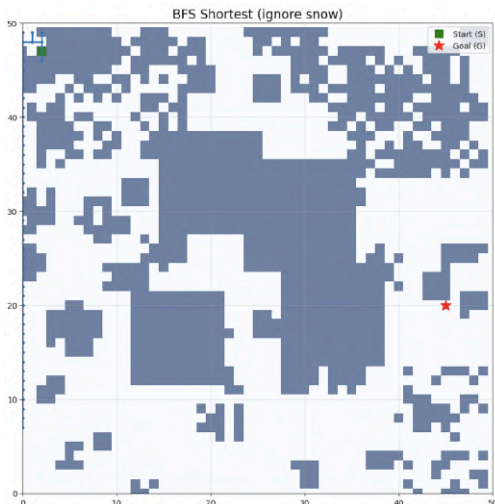
Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	575.30±0.00	70.00±0.00	31.00±0.00	100.0%
<b>Q-Learning</b>	579.10±5.61	70.00±0.00	12.60±3.29	100.0%
<b>PPO Curriculum</b>	588.50±0.60	70.00±0.00	7.80±1.10	100.0%

356 Table 1: summarizes the main performance metrics for all three methods under  
 357 deterministic conditions, including total reward, steps to goal, average snow cells  
 358 crossed, success rate, standard deviation, and averaged across 5 independent random  
 359 seeds.

360

361

### 362 4.2.2 Slip Winter Condition



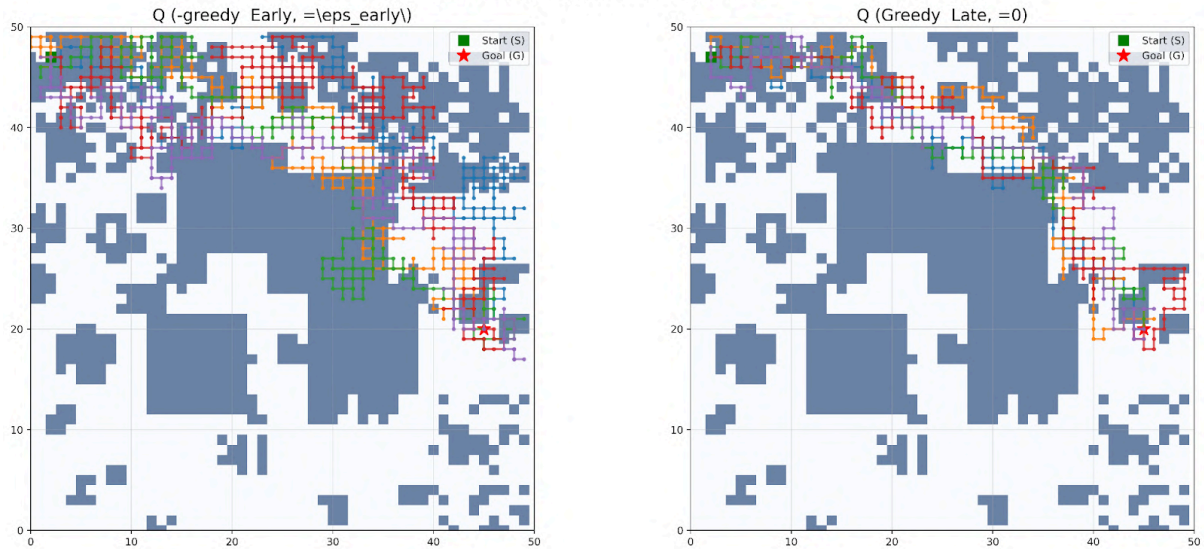
363

364 Figure 6: illustrates the navigation path produced by the BFS baseline under slip  
 365 (stochastic) winter conditions.

366

367

Q-Learning Trajectories: Early vs Late



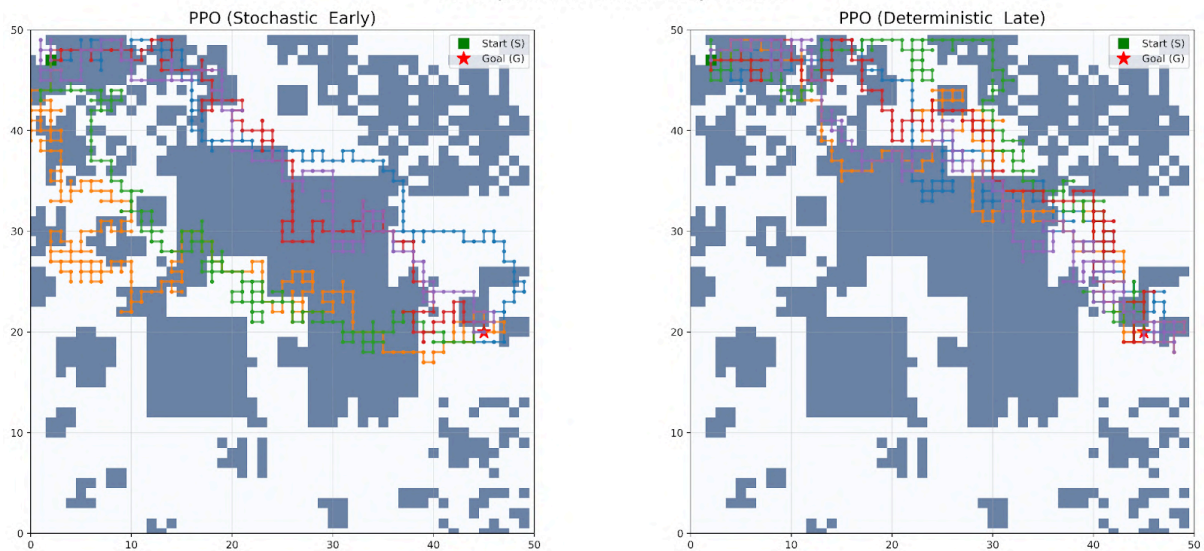
368

369 Figure 7: illustrates the navigation trajectories produced by the Q-learning agent during  
370 early and late stages of training under slip (stochastic) winter conditions.

371

372

PPO Trajectories: Pseudo Early vs Late



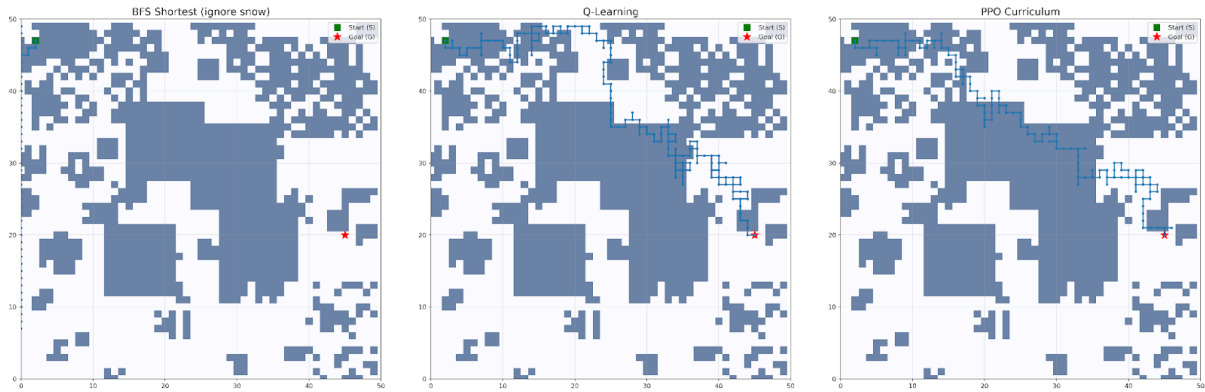
373

374 Figure 8: illustrates the navigation trajectories produced by the PPO agent during early  
375 and late stages of training under slip (stochastic) winter conditions.

376

377

3-Way Trajectory Comparison (Balanced Winter Grid)



378

379 Figure 9: compares the navigation trajectories of BFS, Q-learning, and PPO under slip  
 380 winter conditions.

381

382

383 === Final Comparison Table ===

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	-1693.96±61.27	1000.00±0.00	246.80±76.09	0.0%
<b>Q-Learning</b>	304.03±9.92	228.18±4.10	67.58±4.34	100.0%
<b>PPO Curriculum</b>	252.19±7.04	242.29±3.00	92.65±4.14	100.0%

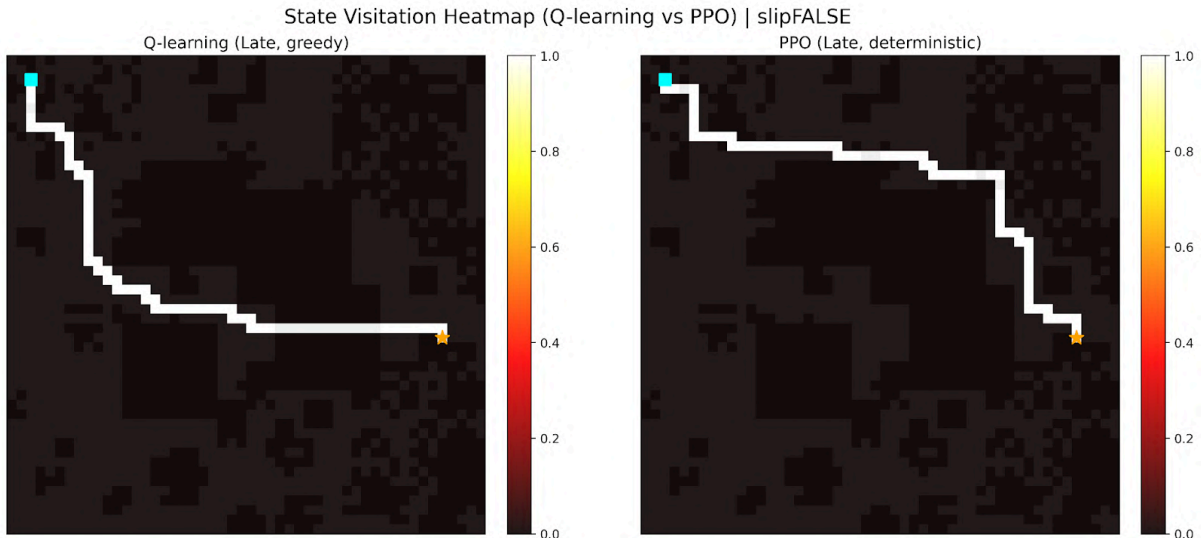
384 Table 2: summarizes the quantitative performance metrics for BFS, Q-learning, and PPO  
 385 under slip winter conditions, using the same evaluation metrics as in the deterministic  
 386 case, standard deviation, and averaged across 5 independent random seeds.

387

388

### 389 4.3 Spatial State Visitation Heatmaps

390 Spatial state visitation heatmaps offer a grid-based way to see how often an agent  
 391 passes through each cell during navigation. In our 50×50 winter grid, each cell  
 392 corresponds to a state, and the color intensity reflects visitation frequency across  
 393 episodes. Darker areas indicate cells that are visited more frequently, while lighter or  
 394 white cells indicate rarely or never visited locations. Such visualizations give a clear  
 395 picture of the agent's overall path distribution and behavior patterns. We generated  
 396 these heatmaps using aggregated visitation counts normalized to the [0,1] range.

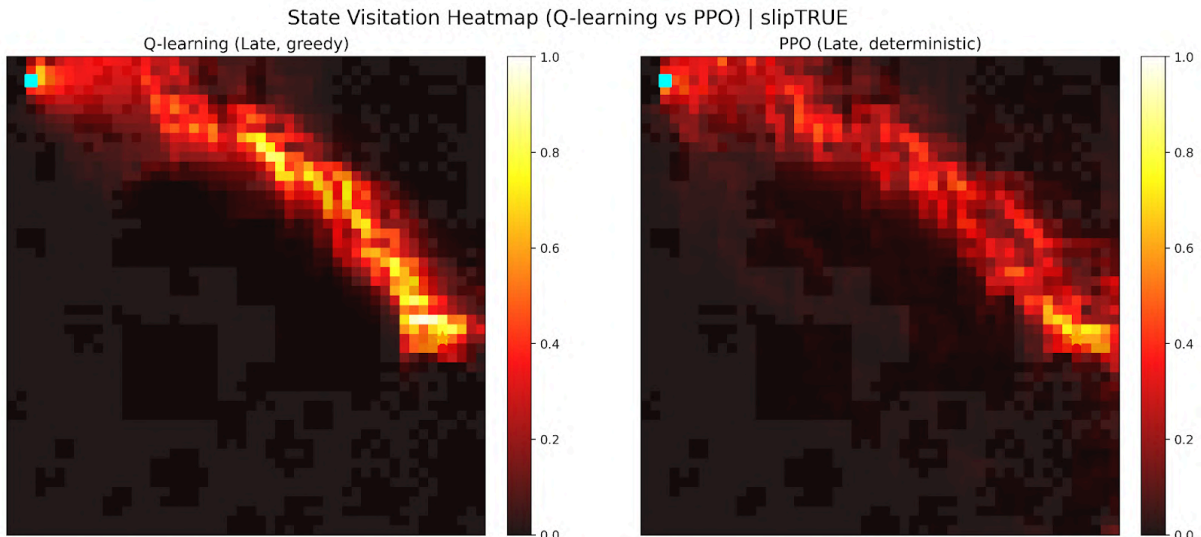


397

398 Figure 10: displays the visitation heatmaps for Q-learning and PPO under deterministic  
 399 (non-slip) winter conditions. The results are aggregated from 10 evaluation runs  
 400 consisting of 200 episodes per run. Both agents concentrate most visits  
 401 along a narrow corridor connecting the start to the goal, with  
 402 near-maximum intensity ( $\approx 1.0$ ) along the primary route and very low  
 403 visitation elsewhere.

404

405



406

407 Figure 11: shows the visitation intensity for each grid cell in the slip case. Compared with  
 408 the non-slip condition, visitation is more widely distributed across the grid rather than  
 409 remaining within a narrow corridor. The heatmap aggregates data from 10 evaluation  
 410 runs consisting of 200 episodes per run.

411

#### 412 4.4 Limitations & Uncertainty

413 Several sources of uncertainty and limitations appeared in this study. Both Q-learning  
 414 and PPO showed noticeable performance differences across training runs, which was

415 expected given the stochastic nature of the environment and the learning algorithms.  
416 To account for this variability, all experiments were repeated over 5 independent  
417 random seeds, and results are reported as the mean across runs

418 Under slip conditions, the same policy could produce quite different routes from  
419 episode to episode because actions did not always lead to the expected outcome. In  
420 addition, stochastic exploration strategies and random initialization of value functions  
421 (or policy networks) exposed agents to slightly different states, transitions, and rewards  
422 across runs. This variability made it harder to get perfectly consistent results.

423 An unexpected issue was that the BFS baseline failed to reach the goal within the  
424 1,000-step limit under slip conditions, resulting in a number of unsuccessful evaluation  
425 episodes.

426

427 A full numerical breakdown of evaluation metrics is provided in Appendix B.

428

429

---

430

## 431 **5. Discussion**

### 432 **5.1 Restatement of Hypothesis and Summary of Findings**

433 The study hypothesized that reinforcement learning-based agents, particularly Proximal  
434 Policy Optimization (PPO), would achieve more energy-efficient winter navigation paths  
435 than a traditional shortest-path baseline and a value-based Q-learning agent under  
436 both deterministic and slip winter conditions. The averaged results partially supported  
437 the hypothesis. Under deterministic (non-slip) conditions, PPO achieved higher  
438 cumulative reward, and fewer snow visits, outperforming Q-learning, which supports  
439 the hypothesis. However under stochastic (slip) conditions, Q-learning outperformed  
440 PPO in terms of cumulative reward or snow-cell avoidance, which did not support the  
441 hypothesis. Both reinforcement learning methods perform better results than  
442 traditional BFS across these environments.

443

### 444 **5.2 Interpretation of Q-Learning and PPO Performance**

445 The results showed that algorithm performance varied depending on the environmental  
446 condition. Under deterministic (non-slip) conditions, PPO achieved higher cumulative  
447 reward and fewer snow cell visits than Q-learning, suggesting that PPO's policy gradient  
448 approach was able to learn a more energy-efficient route in a stable environment.  
449 Under stochastic (slip) conditions, Q-learning outperformed PPO in terms of cumulative  
450 reward and snow-cell avoidance. One possible explanation was that a discrete and clear  
451 grid-world environment and relatively small environment would favor more for  
452 Q-learning, therefore allowing Q-learning to efficiently estimate optimal state [7]. In  
453 contrast, PPO relied on function approximation and stochastic policy updates, which  
454 may have required more training data or longer training time to converge to an equally  
455 optimal policy under slip conditions. Additionally, its stochastic policy may have  
456 introduced suboptimal choices that were not effective, therefore reducing its total  
457 rewards relative to Q-learning under stochastic conditions [3].

458

### 459 **5.3 Effects of Slip (Stochastic) Winter Conditions**

460 Under slip (stochastic) conditions, all agents showed broader trajectory dispersion and  
461 increased visits to previously visited states because of slipping, as reflected in the  
462 spatial state visitation heatmaps. This behavior is consistent with probabilistic transition  
463 dynamics, in which the same action could lead to different movement outcomes across  
464 episodes, such as deviating left in one run and right in another. When the grid is  
465 slippery, agents don't follow one stable path anymore. As a result, the routes spread out,  
466 and their performance becomes less consistent across episodes. These observations are  
467 consistent with prior navigation studies showing that stochastic environments increase  
468 policy variance and reduce convergence stability in reinforcement learning tasks [8].

469

### 470 **5.4 Interpretation of BFS Baseline Behavior**

471 The Traditional baseline trajectories did not take account for winter grid costs or  
472 stochastic transitions, it focused on the fastest way to the destination. In non-slip  
473 conditions this approach naturally produced the optimal route. However, under slip  
474 conditions, BFS exceeded the maximum step limit in multiple evaluation episodes,  
475 reflecting its inability to adapt its policy in response to transition winter grid costs or  
476 stochastic transitions.

477

478 Because BFS always assumed deterministic movement, each slip caused deviations from  
479 the planned path. These deviations often led to inefficient loops and longer paths.  
480 Unlike reinforcement learning methods, BFS did not have reward feedback or policy  
481 updates, which limited its ability to adjust navigation behavior under changing  
482 environmental dynamics. Therefore, making it the least effective method.

483

### 484 **5.5 Limitations and Future Directions**

485 Several limitations should be noted when interpreting these findings. All experiments  
486 were carried out in a simulated 50×50 grid environment with a relatively small and  
487 simple state space, which differs significantly from real-world scenarios. This setup  
488 likely favored value-based methods like Q-learning, since it could learn precise  
489 state-action values for every cell. In contrast, PPO depends on neural network function  
490 approximation and stochastic policy updates, which may need more training steps or  
491 data to perform well in such discrete, small-scale settings.

492

493 To improve statistical reliability, all experiments were run independently over 5 random  
494 seeds and results were averaged across runs. Noticeable performance variability was  
495 observed across training runs, mainly due to the stochastic environment, random action  
496 outcomes, and probabilistic transitions under slip conditions.

497 The simulated environment simplified real-world winter driving conditions and did not  
498 capture vehicle dynamics, sensor noise, road-surface variation, or realistic

499 energy-consumption models, these are factors that might lead to changes. As a result,  
500 these findings to real-world winter navigation scenarios were limited [9].

501 Future research could evaluate reinforcement learning agents in larger scale,  
502 continuous, or more realistic winter navigation environments with higher-dimensional  
503 state spaces and more complex conditions. Then it may be better to reflect real-world  
504 uncertainty and could allow policy-gradient methods such as PPO to fully function their  
505 theoretical advantages in handling stochastic and high-dimensional control problems.

506

507

---

508

## 509 **6. Conclusion**

510 This study compared a traditional BFS baseline with two reinforcement learning  
511 methods—Q-learning and Proximal Policy Optimization (PPO)—for energy-efficient  
512 navigation in a 50×50 winter grid under both deterministic (non-slip) and stochastic  
513 (slip) conditions. The averaged results across 5 independent random seeds partially  
514 supported the original hypothesis. Under deterministic (non-slip) conditions, PPO  
515 achieved the highest cumulative reward and fewest snow cell visits, outperforming  
516 Q-learning. However, under stochastic (slip) conditions, Q-learning outperformed PPO  
517 in cumulative reward and snow-cell avoidance. Although both reinforcement learning  
518 methods clearly outperformed the BFS baseline, BFS struggled under slip conditions  
519 due to its inability to adapt to stochastic transitions.

520

521 These findings show that whether value-based methods or policy-gradient methods  
522 perform better really depends on how unpredictable the environment is. In our small,  
523 structured discrete grid, PPO gained an advantage from its stable gradients and  
524 curriculum learning when everything was deterministic. On the other hand,  
525 Q-learning's simple tabular updates turned out to be more robust when the roads  
526 became slippery and actions sometimes failed. More generally, the results emphasize  
527 that the choice of reinforcement learning method should be guided by the structure and  
528 constraints of the task, rather than by theoretical advantages alone. Future work could  
529 explore whether PPO's advantages extend to more complex environments with  
530 continuous states or larger state spaces, where its policy-gradient approach may better  
531 demonstrate its theoretical strengths.

532

533

---

534

## 535 **Reference**

536

537 [1]Mao, R., Xu, W., Qian, Y., Li, X., Li, Y., Li, G., & Zhang, H. (2025).  
538 Understanding the Determinants of Electric Vehicle Range: A Multi-  
539 Dimensional Survey. *Sustainability*, 17(10), 4259. <https://doi.org/>

540 [10.3390/su17104259](https://doi.org/10.3390/su17104259)

541 [2]Carlson, A., & Vieira, T. (2021). *The effect of water and snow on the*  
542 *road surface on rolling resistance* (VTI Report 971A). Swedish National  
543 Road and Transport Research  
544 Institute. [https://www.diva-portal.org/smash/get/diva2:1542142/](https://www.diva-portal.org/smash/get/diva2:1542142/FULLTEXT01.pdf)  
545 [FULLTEXT01.pdf](https://www.diva-portal.org/smash/get/diva2:1542142/FULLTEXT01.pdf)

546 [3]Watkins, C.J.C.H., & Dayan, P.(1992). Q-learning. *Mach Learn* 8, 279–292.  
547 <https://doi.org/10.1007/BF00992698>

548 [4]Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O.  
549 (2017). *Proximal Policy Optimization Algorithms* (No.  
550 arXiv:1707.06347). arXiv. <https://doi.org/10.48550/arXiv.1707.06347>

551 [5]Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An*  
552 *Introduction*.  
553 [https://web.stanford.edu/class/psych209/Readings/](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)  
554 [SuttonBartoIPRLBook2ndEd.pdf](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)

555 [6]Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and  
556 Reinforcement Learning. *Neuron*, 36(2), 285–298. [https://doi.org/](https://doi.org/10.1016/S0896-6273(02)00963-7)  
557 [10.1016/S0896-6273\(02\)00963-7](https://doi.org/10.1016/S0896-6273(02)00963-7)

558 [7]Tan, C. (2025). Comparative Study of Reinforcement Learning  
559 Performance Based on PPO and DQN Algorithms. *Applied and*  
560 *Computational Engineering*, 175(1), 30–36. [https://doi.org/](https://doi.org/10.54254/2755-2721/2025.AST24879)  
561 [10.54254/2755-2721/2025.AST24879](https://doi.org/10.54254/2755-2721/2025.AST24879)

562 [8]Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino,  
563 A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., &  
564 Hadsell, R. (2017). *Learning to Navigate in Complex Environments* (No.  
565 arXiv:1611.03673). arXiv. <https://doi.org/10.48550/arXiv.1611.03673>

566 [9]Chukwurah, N., Adebayo, A. S., Ajayi, O. O., & Anfo Pub. (2024).  
567 *Sim-to-Real Transfer in Robotics: Addressing the Gap between*  
568 *Simulation and Real- World Performance*. *International Journal for Multidisciplinary*  
569 *Research (IJFMR)*, 05(01), 33–39. [https://doi.org/](https://doi.org/10.54660/.IJFMR.2024.5.1.33-39)  
570 [10.54660/.IJFMR.2024.5.1.33-39](https://doi.org/10.54660/.IJFMR.2024.5.1.33-39)

571 [10]Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009).  
572 Curriculum learning. *Proceedings of the 26th Annual International*  
573 *Conference on Machine Learning*, 41–48. [https://doi.org/](https://doi.org/10.1145/1553374.1553380)  
574 [10.1145/1553374.1553380](https://doi.org/10.1145/1553374.1553380)

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

## 593 **Appendices**

### 594 **Appendix A Experimental Configuration and Algorithmic Details:**

595

#### 596 **A.1 Environment Configuration**

597 All experiments were conducted in a custom 50×50 winter grid environment (2,500  
598 discrete states).

- 599 • Start position: (47, 2)
- 600 • Goal position: (20, 45)
- 601 • Maximum steps per episode: 1000
- 602 • Action space: {up, down, left, right}

603 Two transition settings were evaluated:

604 Deterministic (slip = FALSE)

605 The intended action is executed exactly.

606 Stochastic (slip = TRUE)

607 With probability  $\frac{1}{3}$ , the intended action is executed.

608 With probability  $\frac{2}{3}$ , the agent slips to a random adjacent direction.

609 All algorithms were evaluated on the same fixed grid layout to ensure fairness.

610

#### 611 **A.2 Reward Function**

612 The reward function models energy-aware winter navigation.

613 At each time step:

- 614 • Step penalty:  $-1.5$
- 615 • Snow penalties:
  - 616 ○ Near snow:  $-0.2$
  - 617 ○ Edge snow:  $-0.5$
  - 618 ○ Core snow:  $-2.0$
- 619 • Goal reward:  $+700$

620 The cumulative episode reward is:

$$621 R = \sum_{t=1}^T (-1.5 - C_{snow}(s_t)) + 700 \cdot 1_{goal\ reached}$$

622

623 where  $C_{snow}(s_t)$  denotes the terrain penalty and

624  $1_{goal\ reached}$  indicates successful termination.

### 625 A.3 Q-Learning Configuration

626 Tabular Q-learning was implemented with the following hyperparameters:

- 627 • Learning rate  $\alpha=0.15$
- 628 • Discount factor  $\gamma=0.99$
- 629 • Exploration strategy:  $\epsilon$ -greedy
- 630 • Initial  $\epsilon = 0.60$
- 631 • Minimum  $\epsilon = 0.05$
- 632 • Exponential decay per episode =  $0.9995$
- 633 • Training episodes =  $15,000$
- 634 • Max steps per episode =  $1000$

635 The Q-update rule is:

$$636 Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

637 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) was used.

638

### 639 A.4 Proximal Policy Optimization (PPO) Curriculum Configuration

640 PPO was implemented as a stochastic policy-gradient method.

641 Total training timesteps:  $1,200,000$

- 642 • Phase 1 (navigation-focused reward): 300,000 timesteps
- 643 • Phase 2 (energy-aware reward): 900,000 timesteps

644

645 Evaluation was conducted using deterministic action selection.

646

647 PPO Objective Function

648 PPO optimizes the clipped surrogate objective:

$$649 L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)]$$

650 where

$$651 r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

652 and  $\hat{A}_t$  is the generalized advantage estimate.

653 The clipping mechanism constrains policy updates to maintain stability.

654

655 PPO Hyperparameters

- 656 • Discount factor  $\gamma=0.99$
- 657 • GAE parameter  $\lambda=0.95$
- 658 • Entropy regularization enabled
- 659 • Learning rate:  $3 \times 10^{-4}$
- 660 • Clip range: 0.2
- 661 • Rollout buffer: 2,048 steps
- 662 • Batch size: 256
- 663 • Optimization epochs: 10
- 664 • Neural network architecture: two hidden layers (256 units each, ReLU activation)

665

666 **A.5 Evaluation Protocol**

667 For each condition (slip FALSE / TRUE):

- 668 • One trained model per algorithm
- 669 • Maximum evaluation length: 1000 steps
- 670 • Number of evaluation episodes: 200
- 671 • Number of seeds: 5
- 672 • Heatmap runs: 10
- 673 • Metrics recorded:
  - 674 ○ Total cumulative reward
  - 675 ○ Number of steps
  - 676 ○ Snow cell visits
  - 677 ○ Success rate (%)

678 State visitation heatmaps were generated by aggregating visitation frequencies across  
679 evaluation runs of 200 episodes each.

680

---

## 681 Appendix B Detailed Quantitative Results:

### 682 B.1 Deterministic (slip = FALSE)

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	575.30±0.00	70.00±0.00	31.00±0.00	100.0%
<b>Q-Learning</b>	579.10±5.61	70.00±0.00	12.60±3.29	100.0%
<b>PPO Curriculum</b>	588.50±0.60	70.00±0.00	7.80±1.10	100.0%

683 All methods reached the goal within 70 steps.

684 PPO achieved the highest cumulative reward and fewest snow cell visits.

685

### 686 B.2 Stochastic (slip = TRUE)

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	-1693.96±61.27	1000±0.00	246.80±76.09	0%
<b>Q-Learning</b>	304.03±9.92	228.18±4.10	67.58±4.34	100.0%
<b>PPO Curriculum</b>	252.19±7.04	242.29±3.00	92.65±4.14	100.0%

687 Under stochastic dynamics:

- 688 • BFS fails due to inability to adapt.
- 689 • Q-learning demonstrates greater robustness.
- 690 • PPO maintains success but exhibits higher trajectory dispersion.

691

### 692 **B.3 State Visitation Analysis**

693 State visitation frequency was computed as:

$$694 V(s) = \frac{N(s)}{\max_s N(s)}$$

695 where  $N(s)$  is the number of visits to state  $s$ .

- 696 • Under deterministic conditions, visitation concentrates along a narrow corridor.
- 697 • Under stochastic conditions, visitation becomes more dispersed.
- 698 • Q-learning exhibits more focused routing than PPO under slip conditions.

# 1 Energy-Efficient Path Planning in Winter Conditions: A Comparative 2 Study of Traditional Baseline, Q-learning, and Proximal Policy 3 Optimization in a Grid World Environment

## 6 1. Abstract

7 Recent investigation in winter route planning has become increasingly important due to  
8 increased resistance and reduced efficiency on winter roads. These conditions  
9 significantly increase energy consumption and introduce uncertainty, raising the risk of  
10 failing to reach the destination. To address the challenge, Reinforcement Learning (RL)  
11 is a type of machine learning where an agent learns to make decisions by interacting  
12 with an environment, receiving rewards or penalties for its actions to maximize  
13 cumulative rewards. For simulation, we benchmark a standard shortest-path algorithm  
14 (BFS) against two reinforcement learning methods: Q-learning and Proximal Policy  
15 Optimization (PPO). All approaches are tested on identical grid configurations, with the  
16 same starting points, goals, and reward functions. We assess their performance based on  
17 cumulative reward, snowy cell visits (as a proxy for energy cost), trajectory characteristics,  
18 and success rate. It is hypothesized that proximal policy optimization (PPO) algorithms  
19 will result in lower energy consumption than Q-learning and traditional baseline under  
20 winter conditions. The results show that while BFS consistently finds the shortest paths,  
21 it fails to consider energy costs or environmental uncertainty. RL agents, by  
22 comparison, adapt more efficiently to winter conditions. Under deterministic (non-slip)  
23 conditions, PPO achieved higher cumulative reward and fewer snow cell visits than  
24 Q-learning, partially supporting the hypothesis. However, under stochastic (slip)  
25 conditions, Q-learning outperformed PPO in cumulative reward and snow-cell  
26 avoidance. In contrast, Q-learning has an overall better performance in both conditions  
27 than PPO. These results suggest that Q-learning is better suited for  
28 stochastic conditions structured winter navigation in grid worlds, while PPO  
29 may perform better in more deterministic conditions with complex  
30 environments with continuous states or decisions.

31  
32 Key words: Energy-Efficient, Traditional Baseline, Q-learning, Proximal Policy  
33 Optimization, grid world, reinforcement learning.

---

## 37 2. Introduction

38 As demand in electric vehicles become increasingly widespread, energy efficiency and  
39 route optimization have emerged as critical challenges, particularly in winter when  
40 resistance increases. Freezing temps, plus snow and ice covering the roads, push energy  
41 consumption way up, which means shorter range and trips that feel less predictable. A  
42 2025 study shows that an estimated 50 % of EV driving range can be reduced in cold

43 climates, including snow and ice covering terrain, highlighting the significant impact of  
44 environmental conditions on energy consumption [1]. Furthermore, when snow or  
45 water remains on the road surface, vehicle tires must continuously displace through ice  
46 as they roll and move, hence forcing the vehicle to draw more power [2]. On top of that,  
47 water cools tires more effectively than air alone, altering their mechanical properties  
48 and further pushing rolling resistance higher. Previous studies have shown that rolling  
49 resistance can rise by approximately 30% to 40% under wet or snowy conditions [2].  
50 Ultimately, these factors make winter driving a major concern for **Electric Vehicle (EV)**  
51 efficiency and reinforce the need for energy-aware routing strategies.

52

53 To approach these challenges, we have explored a few computational methods to allow  
54 agents to learn effective navigation strategies aimed at minimizing energy consumption.  
55 In this study, a 50x50 grid is used as an abstraction routing map that circles key  
56 structures of Canadian snow distribution, rather than exact geographical terrain.  
57 **Grid-based environments are commonly used in reinforcement learning studies, as they**  
58 **simplify complex continuous real-world environments into discrete states [5]. This**  
59 **abstraction supports controlled comparison of different routing methods while**  
60 **preserving the essential energy-related difficulties. This abstraction is particularly**  
61 **well-suited for this study, as it enables environmental factors such as snow coverage**  
62 **and surface slip conditions to be incorporated directly into the cell-level cost and**  
63 **reward structure, supporting controlled and reproducible comparison of different**  
64 **routing methods.**

65

66 Despite growing interest in reinforcement learning for navigation, existing studies have  
67 largely overlooked the impact of winter environmental conditions, such as snow  
68 coverage and surface slippiness, on energy-aware path planning. Prior work has  
69 primarily focused on general navigation without taking account of explicitly modeling  
70 weather-dependent energy costs [8]. To address this gap, this study integrates snow  
71 coverage, stochastic slip condition, and energy costs into a comparative framework,  
72 evaluating Q-learning, PPO, and a traditional Baseline under realistic winter routing  
73 scenarios. This study is one of the first to directly compare value-based and  
74 policy-gradient reinforcement learning methods in an energy-aware winter navigation  
75 setting.

76

77 More specifically, the research will include a traditional baseline algorithm, and two  
78 types of reinforcement learning, which are Q-learning and ~~PPO~~ **Proximal Policy**  
79 **optimization** algorithm. The traditional baseline routing computes the shortest path  
80 based on static cost metrics and follows the predefined route without adaptation or  
81 learning (basic routing). Q-learning, a value-based reinforcement learning method,  
82 learns through incremental dynamic programming processes with computational  
83 requirements and it is well suited for agents to improve and refine action values and  
84 achieve effective performance in controlled Markovian domains [3]. In contrast, PPO is a  
85 policy-gradient algorithm that primarily optimizes a stochastic policy and achieves

86 greater training stability through constrained policy updates with a clipped surrogate  
87 objective function [4].

88

89 This comparison evaluates the effectiveness of the proposed routing methods in  
90 identifying energy-efficient paths. Performance is evaluated through energy related  
91 costs, overall routing efficiency, and observed navigation behavior under both  
92 deterministic (non-slip) and stochastic (slip) settings. Because PPO balances stable  
93 policy updates with adaptive learning in uncertain environments, it is hypothesized to  
94 be the most effective method implemented for winter routing.

95

96 The following section goes into the methodology and experimental setup in more detail.

97

98

---

99

## 100 **3.Methods**

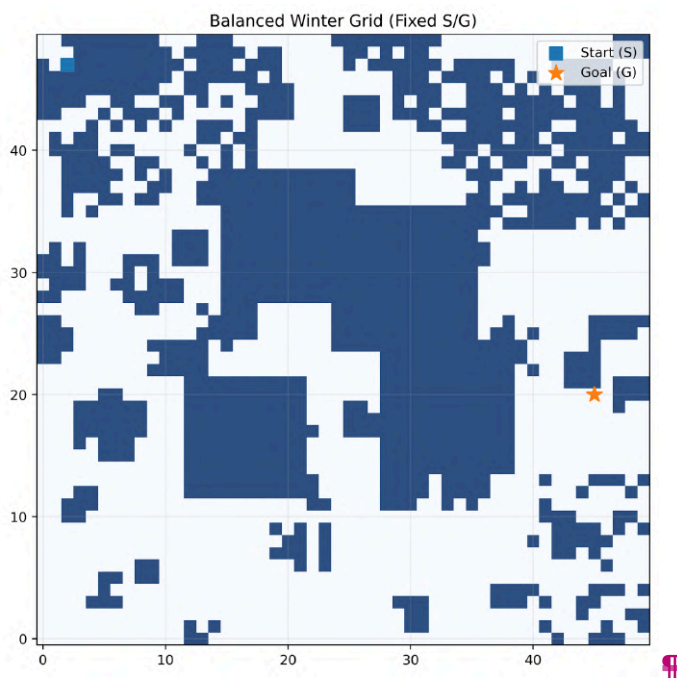
### 101 **3.1 Environment Design**

102 A custom winter navigation grid environment was designed to evaluate each  
103 Reinforcement Learning (RL) agent under stochastic and cost-sensitive conditions. The  
104 environment consists of a 50 x 50 grid world containing 2500 cells, including both  
105 normal terrain and snow-covered terrain. ~~Specifically, two transition settings were~~  
106 ~~evaluated: a slip snow condition and non-slip snow condition (same grid map).~~ The  
107 simulation contains a fixed start (47, 2) and a fixed goal (20, 45) point, and are located in  
108 the top left corner and middle right respectively. At each time step, the agent can take  
109 either one of the four directions, up, down, left, and right; each step on a non-snow grid  
110 has an energy cost of -1.5 per step. Snow is modeled as spatially correlated regions  
111 with graded intensity, capturing the typical uneven accumulation patterns in Canadian  
112 winters.

113 When snow is modeled as a binary state, abrupt reward discontinuities make PPO's  
114 advantage estimates unstable, resulting in increased gradient variance. By contrast, a  
115 continuous representation of snow thickness provides smoother cost transitions, which  
116 help reduce gradient variance and stabilize learning, making it more efficient. These  
117 snow types are classified as "Near snow", "Edge snow", and "Core snow".

118 To evaluate robustness under different winter conditions, two transition settings were  
119 considered using the same grid map: a non-slip (deterministic) condition and a slip  
120 (stochastic) condition. In the deterministic setting, action always results in the intended  
121 movement, and snow cells only add additional energy loss. In the stochastic setting,  
122 ~~actions do not always lead to the intended outcome: with a probability of 1/3, the~~  
123 ~~intended action is executed, and with a probability of 2/3, the agent moves in a random~~  
124 ~~adjacent direction, simulating loss of traction control during winter driving on ice and~~  
125 ~~snow.intended actions may be replaced with unintended movements due to unstable~~  
126 ~~control on icy surfaces, with a higher probability of deviating left or right, simulating~~  
127 ~~loss control of traction during winter driving on ice and snow.~~

128



129

130 ¶

131

132 Figure 1: shows a typical winter grid layout used in all experiments.

133

### 134 3.2 Reinforcement Learning Algorithms/ Methodologies:

135 The routing strategies and learning methods used in this study are described in this  
136 section.

137

#### 138 3.2.1 Traditional Baseline (BFS)

139 Traditional **b**Baseline is a non-learning shortest path strategy used as a comparison  
140 base for reinforcement learning methods. It operates on the 50x50 grid map, it  
141 considers four directions to move (up, down, left, right). The system calculates the  
142 shortest path to the destination, without considering energy savings, which means it  
143 does not adapt to environmental feedback or uncertainty.

144

#### 145 3.2.2 Q-learning Implementation

146 Tabular Q-learning was used as the value-based reinforcement learning baseline. The  
147 agent maintains a discrete Q (s, a) table, which is updated iteratively based on observed  
148 state transitions. An  $\epsilon$ -greedy policy was used for exploration. The value of  $\epsilon$  begins at  
149 0.6 and decays exponentially with a factor of 0.9995 per episode until it reaches 0.05. In  
150 this 50x50 grid setting, the decay schedule allows broad initial exploration before  
151 shifting emphasis to exploitation. Actions are restricted to four possibilities—left (0),  
152 down (1), right (2), and up (3)—consistent with standard discrete grid-world formulations  
153 and the discrete grid structure. The Q-value update follows the classic  
154 temporal-difference form:

155

156  $Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$  [5]

157

158 We fix the learning rate  $\alpha$  at 0.15 and the discount factor  $\gamma$  at 0.99. These values handle  
159 the stochastic transitions (is\_slippery=True) and support planning over long horizons in  
160 the large grid [3]. State  $s$  denotes the flattened grid position index. The action  $a$  selects  
161 one of the four movement directions.  $Q(s, a)$  represents the estimated discounted  
162 cumulative reward starting from state  $s$  and action  $a$  [3,5].

163

### 164 3.2.3 Reward system

165 In this research, the reward system is designed to represent energy efficiency under  
166 winter conditions, it helps agents find the best paths. At each step, the system gives a  
167 negative value to show how significant a step is to energy saving, with higher cause on  
168 snow grids. A positive 700 is given only when the agents successfully reach the  
169 destination, while within the maximum steps of 1000 steps. It encourages agents to  
170 reach the destination with considerations on energy saving. We initialize the cumulative  
171 reward to 0 at the beginning of each episode. Each will result in negative values,  
172 because by doing this, agents don't need to consider the prior bias, since it assumes  
173 nothing at first, etc. As in the settings, we have Near-snow (light snow), Edge-snow  
174 (medium snow), and Core-snow (deep snow); each of them has different additional  
175 values which are -0.2, -0.5, and -2.0 respectively [6].

176

177  $r_t = r_{step} - c_{energy}(s_t) + r_{goal}$  [6]

178

179 The basic reward function follows standard reinforcement learning practice by  
180 combining step penalties, energy-related costs, and a terminal goal reward.

181

### 182 3.2.4 Proximal Policy Optimization

183 Proximal Policy Optimization (PPO) served as the policy-gradient method in this study  
184 for energy-efficient routing under winter conditions. PPO learns a stochastic policy by  
185 directly outputting action probabilities for each state, which helps the agent cope with  
186 uncertainty on slippery roads.

187

188  $L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)]$  [4]

189

190 where  $r_t(\theta)$  is the probability ratio between the new and previous policies. The clipping  
191 mechanism constrains policy updates to improve stability.

192

193 The PPO agent was trained using fixed values across all experiments. The discount  
194 factor was set to  $\gamma=0.99$ , and Generalized Advantage Estimation (GAE) was used with  
195  $\lambda=0.95$ . The policy network consisted of two fully connected hidden layers with 256  
196 units each, using ReLU activation functions. ~~Entropy regularization was applied to~~

197 ~~encourage exploration during training.~~ The learning rate was set to  $3 \times 10^{-4}$ , with a  
198 clipping range of 0.2 and an entropy coefficient of 0.02 to encourage exploration. PPO  
199 training was conducted for 1,200,000 time steps, using a rollout buffer of 2048 steps, a  
200 batch size of 256, and 10 optimization epochs per update ~~using large rollout buffers and~~  
201 ~~multiple optimization epochs~~ to improve training stability [4].

202

203 Policy-gradient methods work by directly adjusting a parameterized policy to increase  
204 expected reward. They learn a stochastic policy, allowing for more flexibility in  
205 uncertain environments, where the same action can lead to different outcomes. In  
206 contrast, q-learning estimates state-action values and usually converges with a  
207 deterministic policy based on these estimates. Although it can learn in unfamiliar  
208 environments, its action selection may be less reliable in highly unpredictable  
209 situations. As a result, policy-gradient and value-based methods show varying strengths  
210 depending on the level of uncertainty in the winter routing task.

211

### 212 3.2.5 Curriculum Learning for PPO

213 PPO training followed a two-phase curriculum learning approach designed to improve  
214 training stability and sample efficiency under sparse rewards and energy-sensitive  
215 penalties [10].

216

#### 217 **Phase 1: Navigation Learning Phase (300,000 timesteps)**

218 ~~Navigation Learning Phase~~ In phase 1, PPO was trained using a simplified reward  
219 function that focuses only on positive feedback for reaching the goal. Energy costs in  
220 snow grids were not included. This made the feedback denser and more immediate. The  
221 agent quickly learned basic navigation and achieved early success in the  $50 \times 50$  winter  
222 grid. The policy learned in this phase served as the starting point for the next training  
223 stages.

224

#### 225 **Phase 2: Energy-Aware Optimization Phase (900,000 timesteps)**

226 ~~Energy-Aware Optimization Phase~~ In phase 2, the training continued with the complete  
227 energy-aware reward function, which now included penalties for moving across  
228 snow-covered areas. The agent had to balance successful arrival with lower energy  
229 consumption. All reported results, ablation studies, and comparisons ~~are~~ were based  
230 solely on the Phase 2 reward function.

231 ¶

232 ¶

233

### 234 3.3 Evaluation factors

235 To ensure statistical reliability, all reinforcement learning experiments were repeated  
236 over 5 independent random seeds, and results are reported as the mean and standard  
237 deviation across runs.

238

239 To compare the performances of all routing methods under identical conditions, a final  
240 evaluation was conducted after training. Each method BFS, Q-learning, and PPO was  
241 evaluated using the following metrics:

242

- 243 • Total Reward: calculated as the cumulative reward obtained over the episode.
- 244
- 245 • Number of steps: steps taken from the start to the goal (less than 1000 steps).
- 246
- 247 • Average snow cell visited: Snow visits were calculated by counting the number of  
248 snow-covered cells traversed in each episode and averaging this value across all  
249 evaluation episodes.
- 250
- 251 • Success rate: defined as the percentage of episodes in which the agent reached  
252 the goal within the maximum step limit.

253

254 The BFS baseline was evaluated only once on the fixed grid map, as it produces a  
255 deterministic path. Reinforcement learning agents were evaluated over ~~200 multiple~~  
256 episodes **per run, with experiments repeated over 5 independent random seeds for each**  
257 **condition (slip and non-slip)**. All methods were compared using the same ~~energy-~~  
258 aware reward function and environment configuration to ensure fairness.

259

### 260 **3.4 Hyperparameter Selections**

261 Hyperparameters are selected based on standard practices and empirical validation to  
262 ensure stable and consistent performance [3,4].

263

#### 264 **3.4.1 Q-Learning Hyperparameter Configuration**

265 We implement tabular Q-learning with an  $\epsilon$ -greedy exploration strategy.

266

267 The hyperparameters are set as follows:

268 - Learning rate  $\alpha = 0.15$

269 - Discount factor  $\gamma = 0.99$

270 - Exploration strategy:  $\epsilon$ -greedy

271 - Initial  $\epsilon = 0.60$

272 - Minimum  $\epsilon = 0.05$

273 - Exponential decay rate = 0.9995

274 - Training episodes = 15,000

275 - Maximum steps per episode = 1,000

276

277 The learning rate ( $\alpha = 0.15$ ) is selected to balance convergence speed and stability. A  
278 moderate learning rate allows the agent to adapt efficiently while avoiding oscillations  
279 in value updates.

280

281 The discount factor ( $\gamma = 0.99$ ) emphasizes long-term rewards, which is essential for  
282 navigation tasks where the objective is to reach the goal efficiently over multiple steps.

283

284 An  $\epsilon$ -greedy exploration strategy is adopted to balance exploration and exploitation.  
285 The initial exploration rate ( $\epsilon = 0.60$ ) encourages sufficient exploration in early training,  
286 while exponential decay gradually shifts the policy toward exploitation. The minimum  
287 ( $\epsilon = 0.05$ ) ensures that some level of exploration is maintained throughout training,  
288 preventing the agent from getting stuck in suboptimal policies.

289

290 The number of training episodes 15,000 and maximum steps per episode 1,000 are  
291 chosen to ensure sufficient interaction with the environment for convergence, while  
292 maintaining computational efficiency.

293

294 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) is used to assess the learned policy  
295 performance.

296

### 297 **3.4.2 PPO Hyperparameter Configuration**

298 We implement PPO using a clipped surrogate objective with generalized advantage  
299 estimation (GAE).

300

301 The hyperparameters are set as follows:

302 - Discount factor  $\gamma = 0.99$

303 - GAE parameter  $\lambda = 0.95$

304 - Learning rate =  $3 \times 10^{-4}$

305 - Clipping range  $\epsilon = 0.2$

306 - Rollout buffer size = 2048 steps

307 - Batch size = 256

308 - Optimization epochs per update = 10

309 - Entropy regularization is enabled

310

311 The policy and value networks share a neural architecture consisting of two hidden  
312 layers with 256 units each and ReLU activation.

313

314 The discount factor ( $\gamma = 0.99$ ) is selected to emphasize long-term rewards, which is  
315 essential for navigation tasks. The GAE parameter ( $\lambda = 0.95$ ) provides a balance between  
316 bias and variance in advantage estimation, leading to more stable learning.

317

318 The clipping parameter ( $\epsilon = 0.2$ ) constrains policy updates to prevent large deviations  
319 from the previous policy, which improves training stability. The rollout buffer size and  
320 number of optimization epochs are chosen to ensure sufficient policy updates while  
321 avoiding overfitting to recent trajectories.

322

323 Entropy regularization is included to encourage exploration and prevent premature  
324 convergence to suboptimal policies.

325

326 Overall, these hyperparameters follow widely adopted settings in reinforcement  
327 learning literature and are chosen to ensure stable and consistent training rather than  
328 aggressive performance tuning.

329

330 A complete summary of environment settings and hyperparameters is provided in  
331 Appendix A.¶

332 ¶

333

334

## 335 4. Results

336 The experiments compare BFS, Q-learning, and PPO on the same 50×50 winter grid,  
337 looking at both non-slip (deterministic) and slip (stochastic) cases.

338

### 339 4.1 Evaluation Setup and Metrics

340 The grid layout, start and goal positions, and reward parameters were kept exactly the  
341 same across all methods. All experiments were repeated over 5 independent random  
342 seeds, and results are reported as the mean across runs. Trajectory figures shown are  
343 from the representative seed whose performance was closest to the mean. For each run,  
344 total reward, number of steps, average snow cells crossed per episode, and success in  
345 reaching the goal were recorded.

346

### 347 4.2 Trajectory Visualizations and Quantitative Results¶

348

#### 349 4.2.1 Deterministic (Non-Slip) Winter Condition



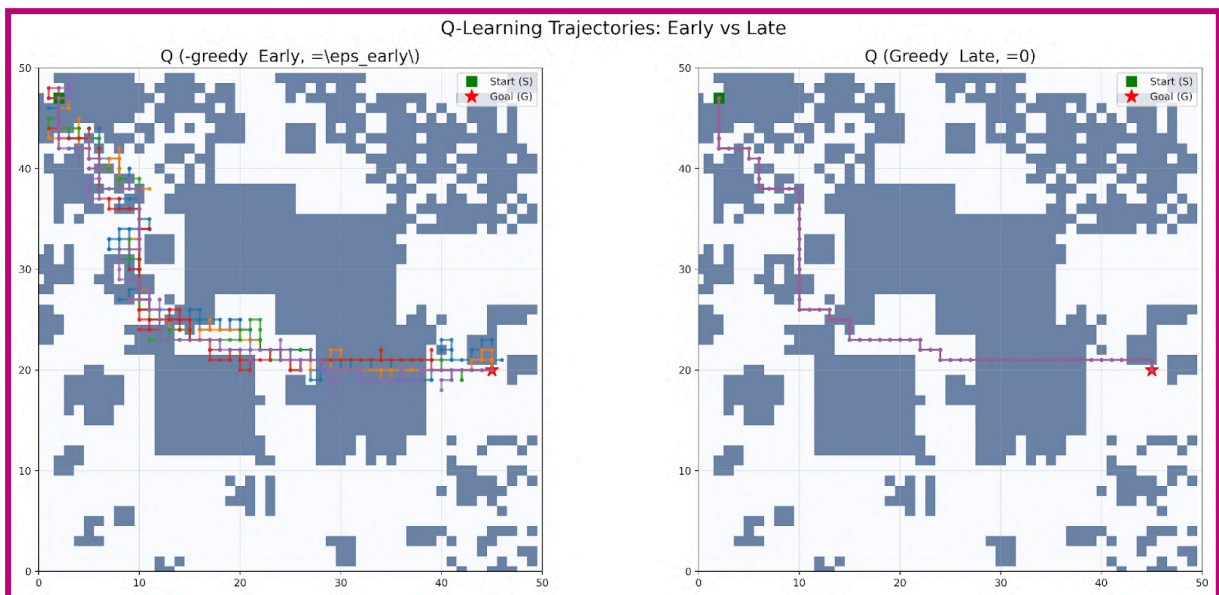
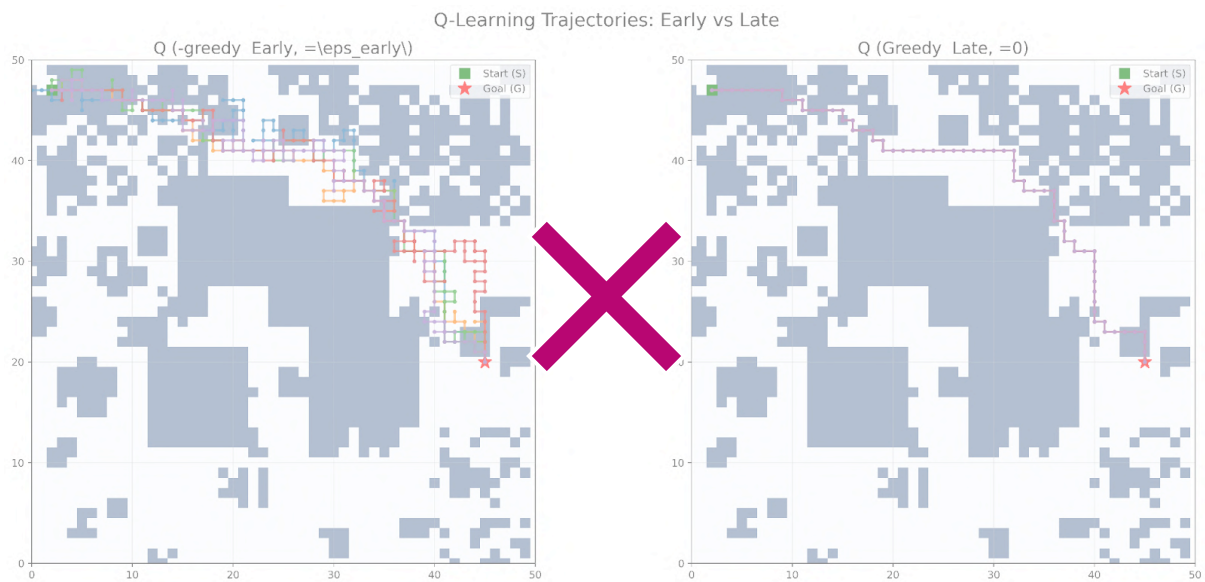
350

351 Figure 2: illustrates the BFS path under non-slip conditions. Since BFS always finds the  
352 shortest route, the trajectory goes straight from the start to the goal with minimal cells  
353 visited without any deviation.

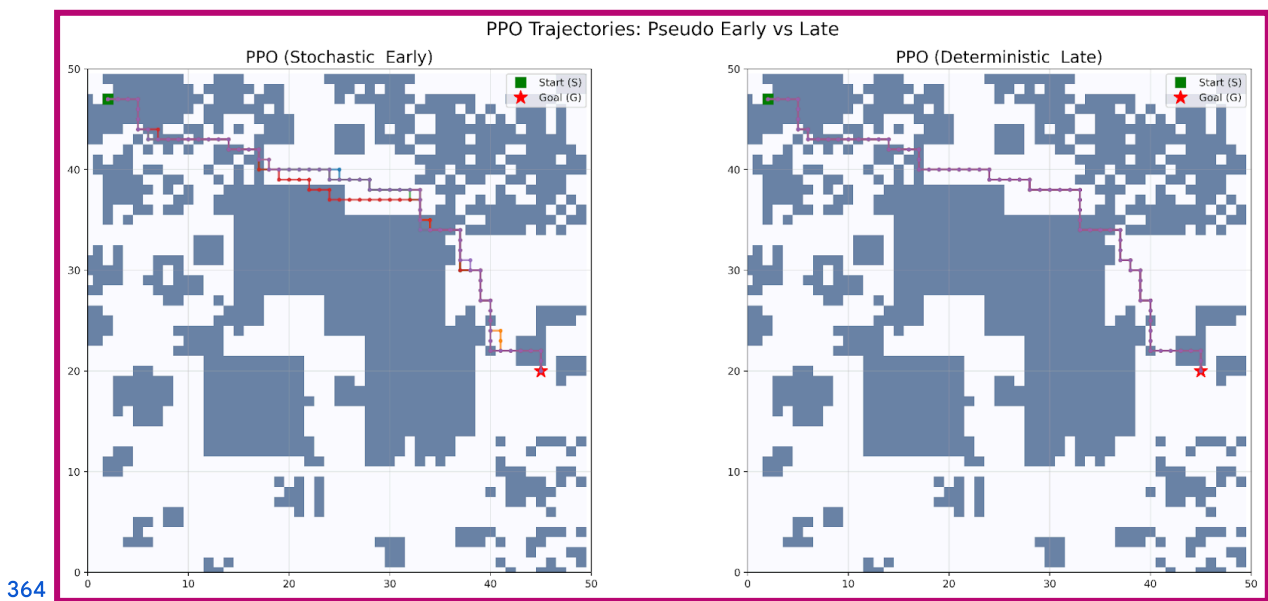
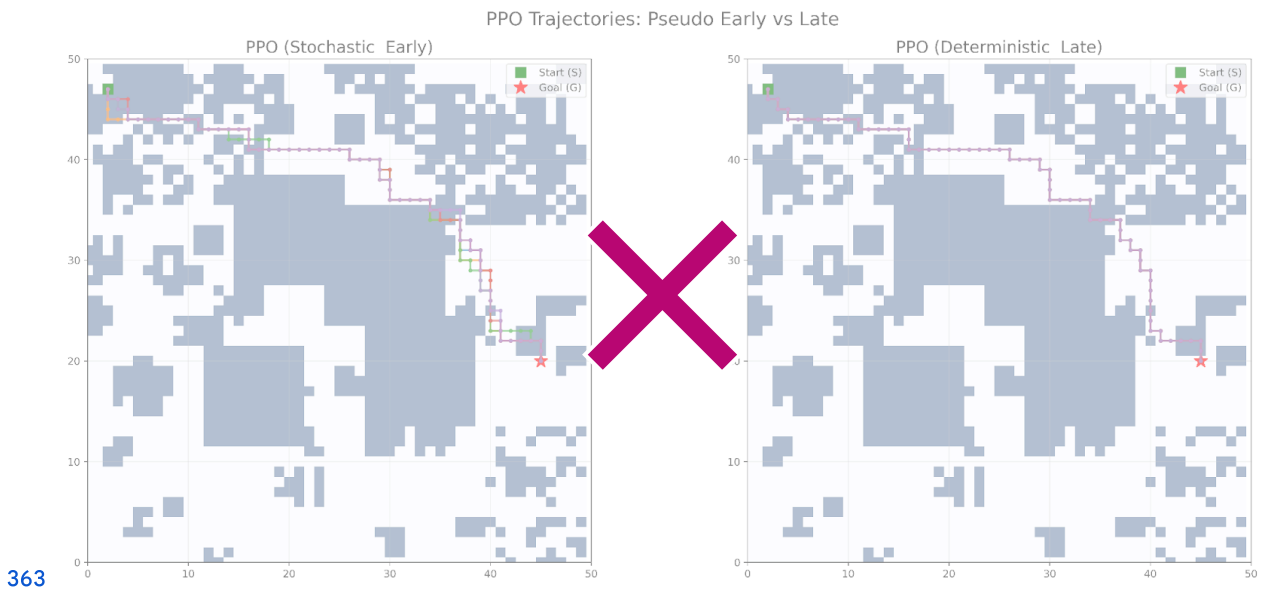
354

355

356



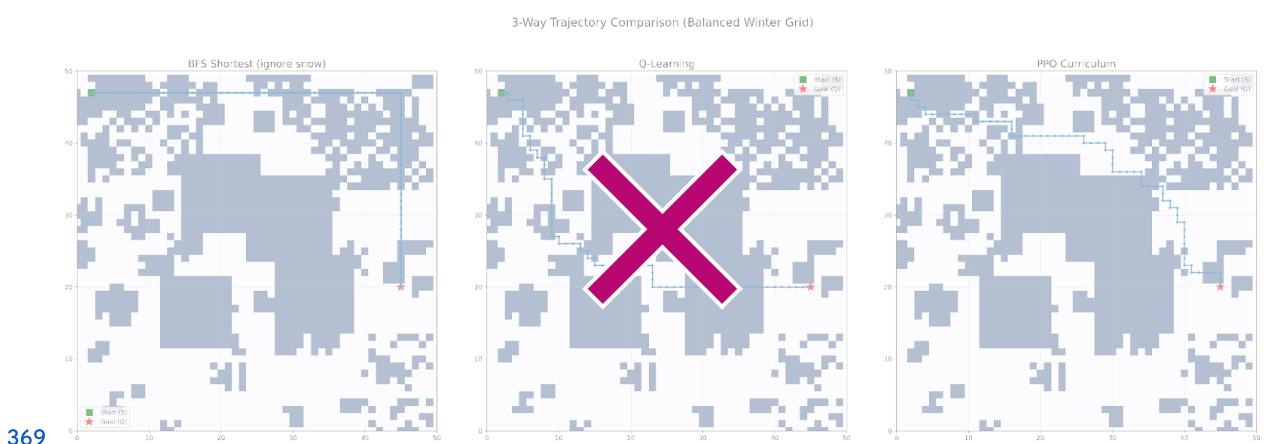
359 Figure 3: illustrates the navigation trajectories produced by the Q-learning agent under  
360 deterministic (non-slip) winter conditions during early training with an  $\epsilon$ -greedy policy  
361 and late training with a greedy policy. The trajectories demonstrate the transition from  
362 exploratory behavior to a stable path toward the goal.

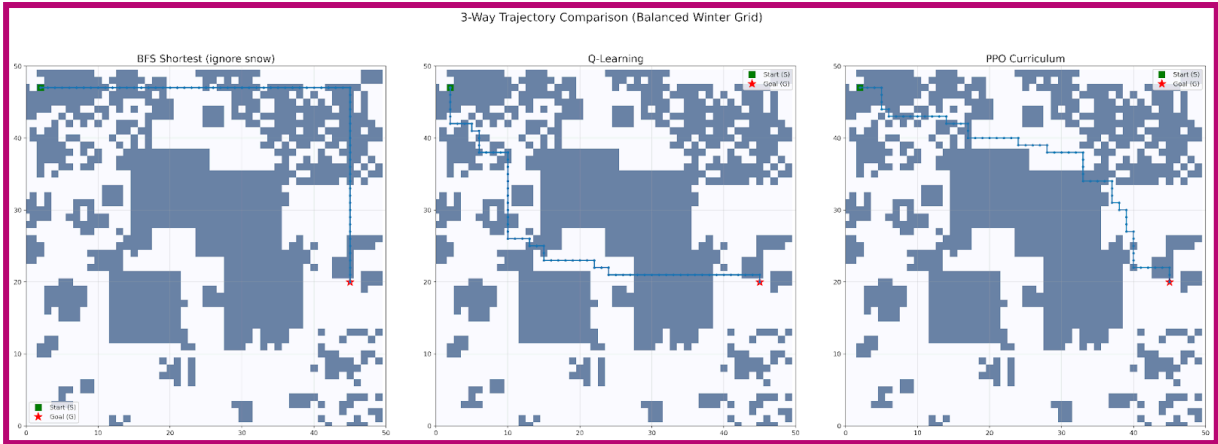


365 Figure 4: illustrates the navigation trajectories produced by the PPO agent during early  
 366 and late stages of training under deterministic (non-slip) winter conditions.

367

368





370

371 Figure 5: compares the trajectories of BFS, Q-learning and PPO in the non-slip winter  
 372 setting.

373

374

375 === Final Comparison Table ===

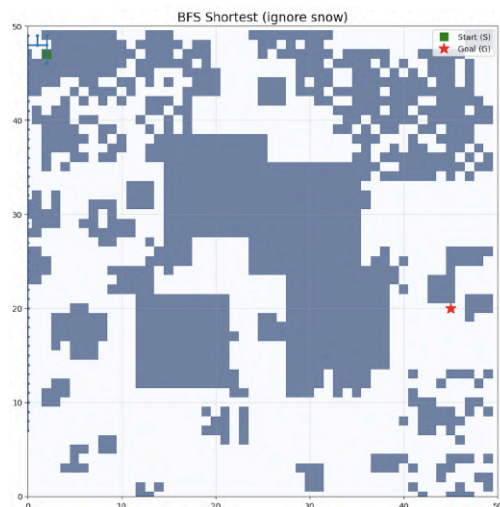
Method	Reward	Steps	Snow Visits	Success
BFS Shortest	575.30±0.00	70.00±0.00	31.00±0.00	100.0%
Q-Learning	579.10589.50±5.61	70.00±0.00	12.607.00±3.29	100.0%
PPO Curriculum	588.50587.50±0.60	70.00±0.00	7.8011.00±1.10	100.0%

376 Table 1: summarizes the main performance metrics for all three methods under  
 377 deterministic conditions, including total reward, steps to goal, average snow cells  
 378 crossed, and success rate, standard deviation, and averaged across 5 independent  
 379 random seeds.:

380

381

### 382 4.2.2 Slip Winter Condition



383

384 Figure 6: illustrates the navigation path produced by the BFS baseline under slip  
 385 (stochastic) winter conditions.

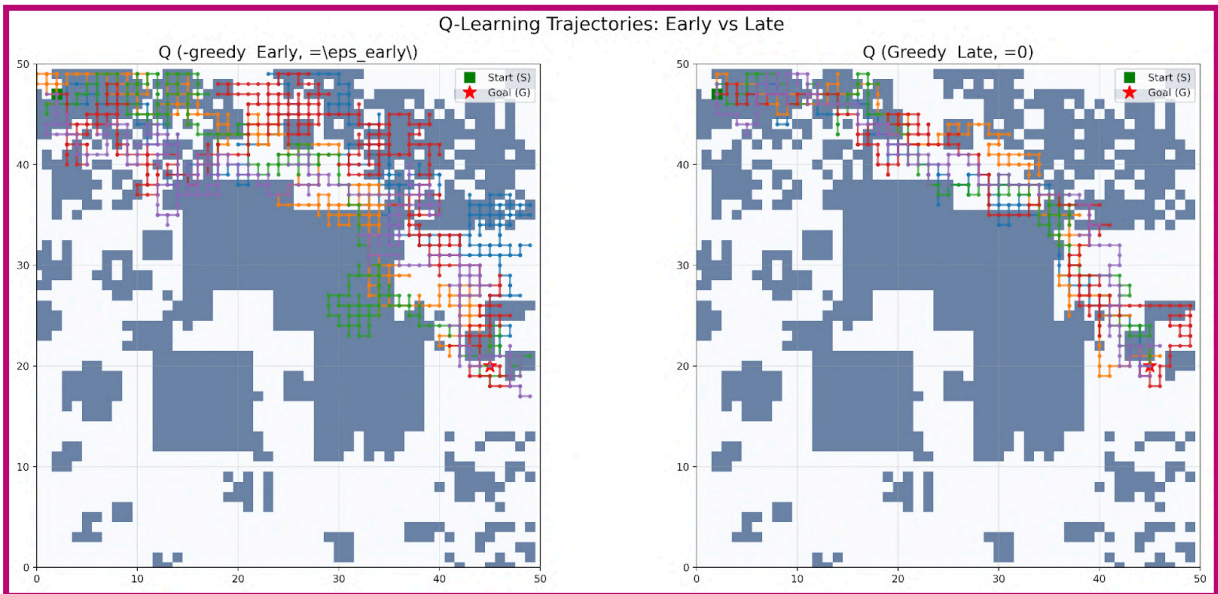
386

387

Q-Learning Trajectories: Early vs Late



388

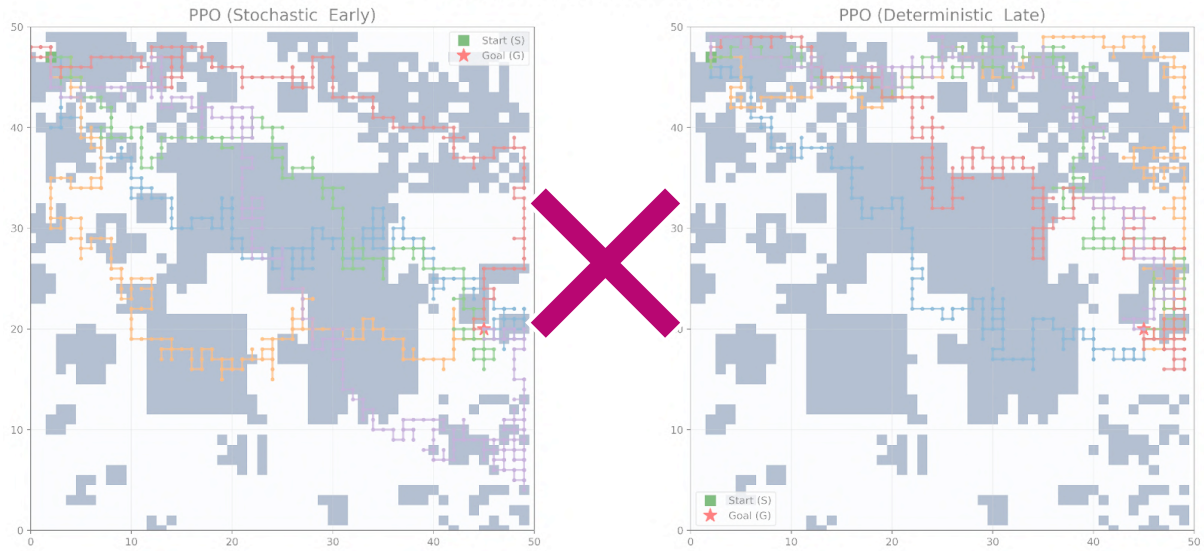


389

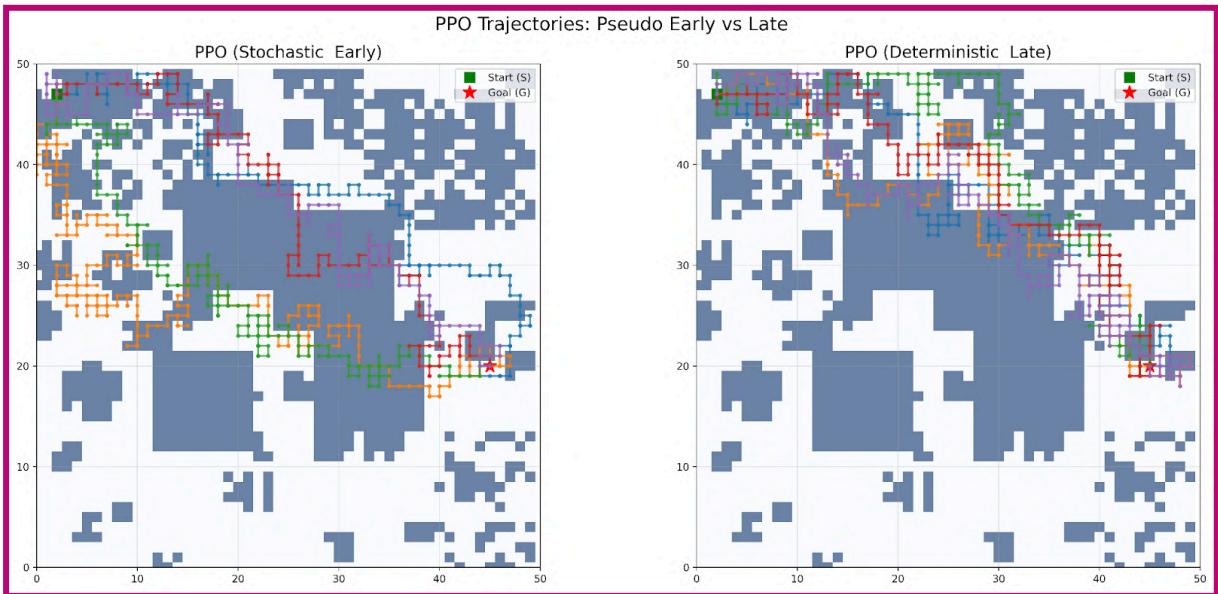
390 Figure 7: illustrates the navigation trajectories produced by the Q-learning agent during  
391 early and late stages of training under slip (stochastic) winter conditions.

392

PPO Trajectories: Pseudo Early vs Late



393



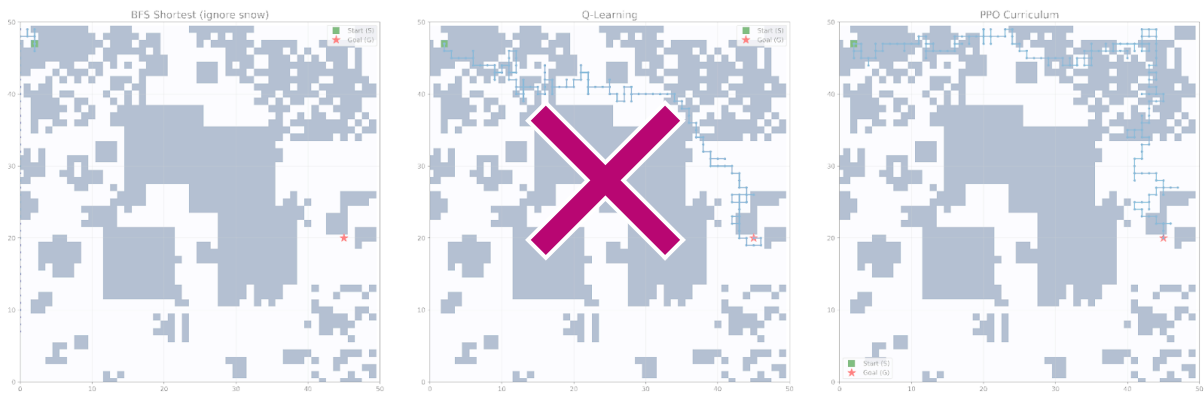
394

395 Figure 8: illustrates the navigation trajectories produced by the PPO agent during early  
396 and late stages of training under slip (stochastic) winter conditions.

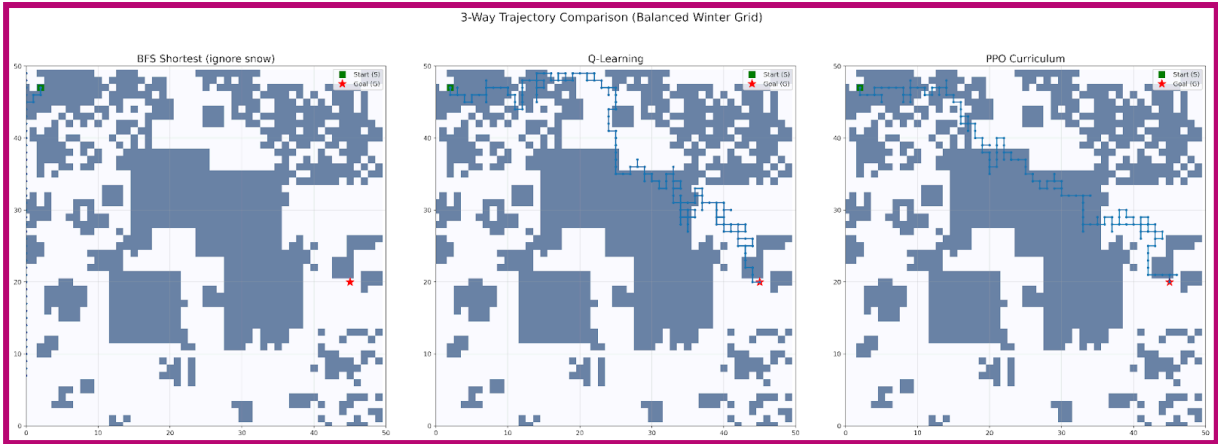
397

398

3-Way Trajectory Comparison (Balanced Winter Grid)



399



400

401 Figure 9: compares the navigation trajectories of BFS, Q-learning, and PPO under slip  
 402 winter conditions.

403

404

405 === Final Comparison Table ===

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	<del>-1693.96</del> $39.20 \pm 61.27$	$1000.00 \pm 0.00$	<del>246.80</del> $188.00 \pm 76.09$	0.0%
<b>Q-Learning</b>	$304.03$ <del>9.41</del> $\pm 9.92$	$228.18$ <del>4.2</del> $\pm 4.10$	$67.58$ <del>65.8</del> $\pm 4.34$	100.0%
<b>PPO Curriculum</b>	$252.19$ <del>2.1</del> $\pm 7.04$	$242.29$ <del>5.8</del> $\pm 3.00$	$92.65$ <del>100.0</del> $\pm 4.14$	100.0%

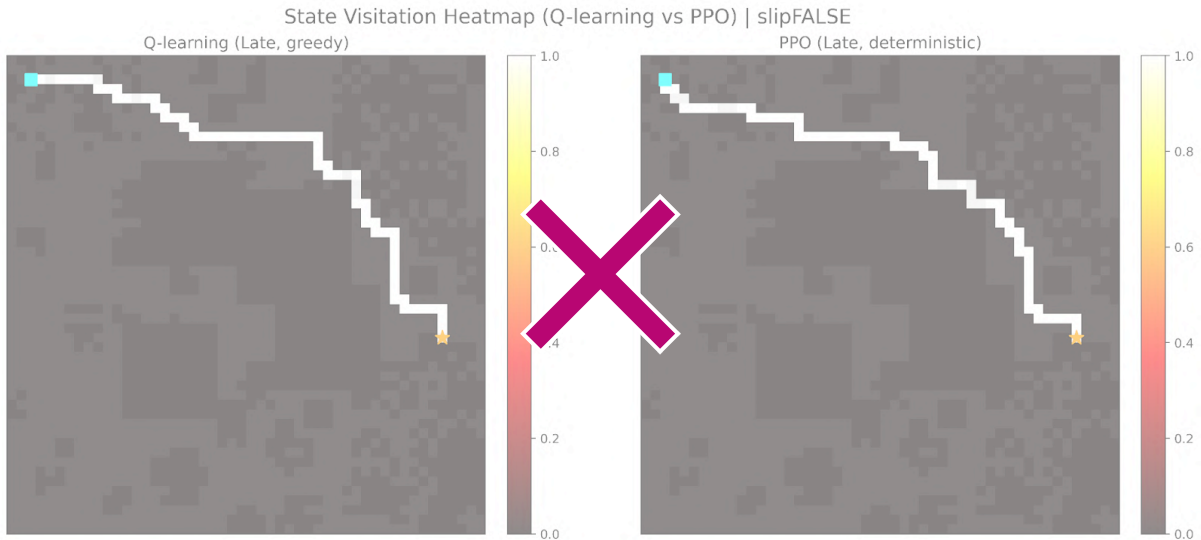
406 Table 2: summarizes the quantitative performance metrics for BFS, Q-learning, and PPO  
 407 under slip winter conditions, using the same evaluation metrics as in the deterministic  
 408 case, **standard deviation, and averaged across 5 independent random seeds.**

409

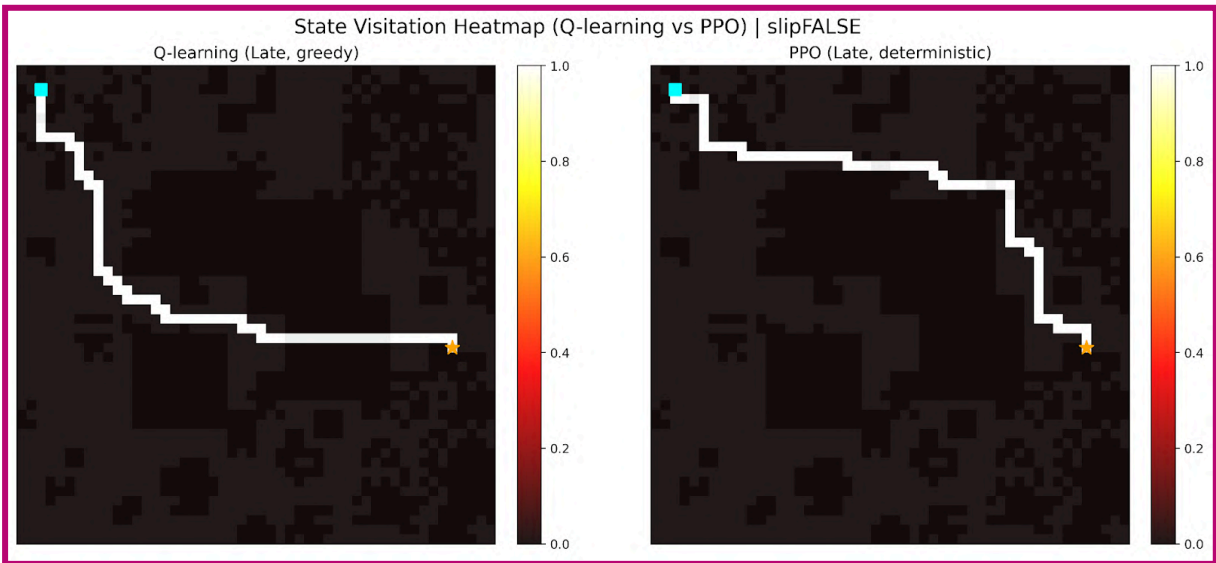
410

### 411 4.3 Spatial State Visitation Heatmaps

412 Spatial state visitation heatmaps offer a grid-based way to see how often an agent  
 413 passes through each cell during navigation. In our 50×50 winter grid, each cell  
 414 corresponds to a state, and the color intensity reflects visitation frequency across  
 415 episodes. Darker areas indicate cells that are visited more frequently, while lighter or  
 416 white cells indicate rarely or never visited locations. Such visualizations give a clear  
 417 picture of the agent's overall path distribution and behavior patterns. We generated  
 418 these heatmaps using aggregated visitation counts normalized to the [0,1] range.



419

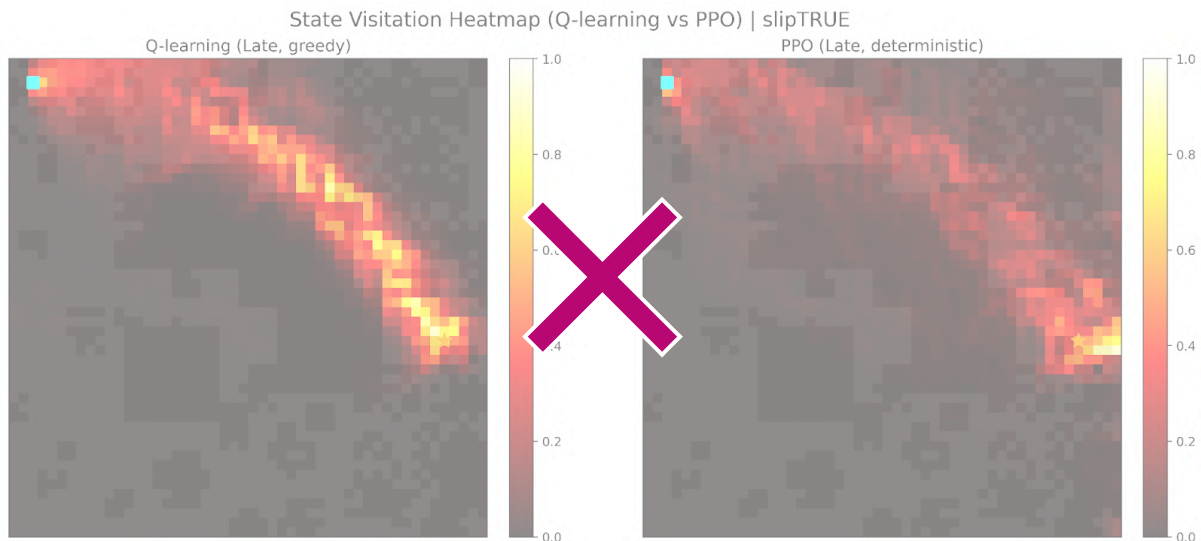


420

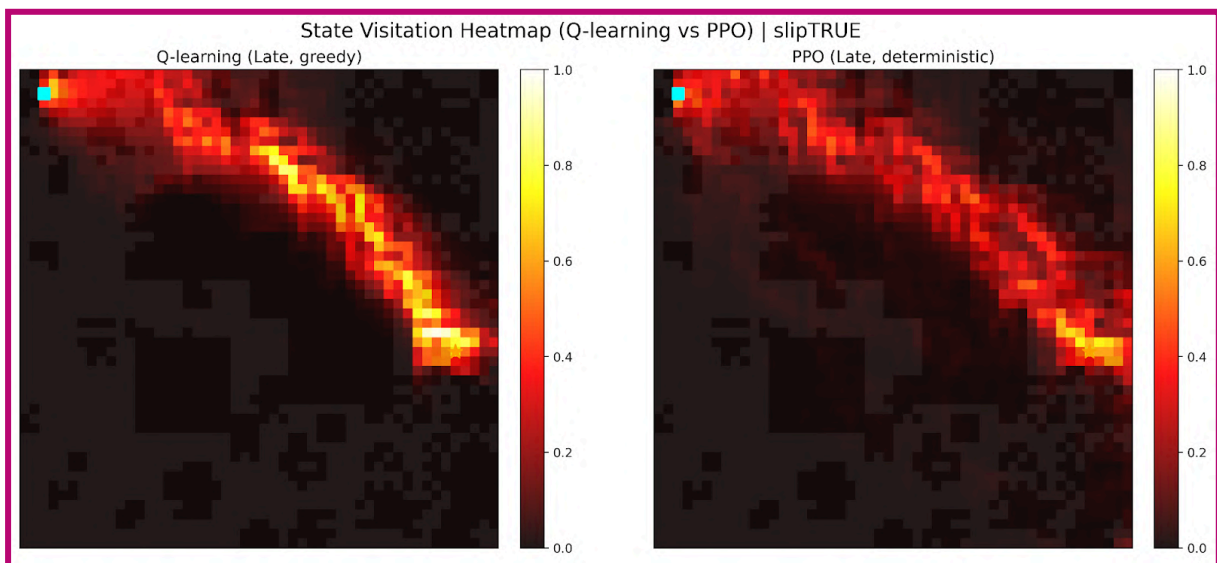
421 Figure 10: displays the visitation heatmaps for Q-learning and PPO under deterministic  
 422 (non-slip) winter conditions. The results are aggregated from 10 evaluation runs  
 423 consisting of 200 episodes per run. episodes. Both agents concentrate most visits along  
 424 a narrow corridor connecting the start to the goal, with near-maximum intensity ( $\approx 1.0$ )  
 425 along the primary route and very low visitation elsewhere.

426

427



428



429

430 Figure 11: shows the visitation intensity for each grid cell in the slip case. Compared with  
 431 the non-slip condition, visitation is more widely distributed across the grid rather than  
 432 remaining within a narrow corridor. The heatmap aggregates data from 10 evaluation  
 433 runs consisting of 200 episodes per run. episodes.

434

#### 435 4.4 Limitations & Uncertainty

436 Several sources of uncertainty and limitations appeared in this study. Both Q-learning  
 437 and PPO showed noticeable performance differences across training runs, which was  
 438 expected given the stochastic nature of the environment and the learning algorithms. To  
 439 account for this variability, all experiments were repeated over 5 independent random  
 440 seeds, and results are reported as the mean across runs.

441 Under slip conditions, the same policy could produce quite different routes from  
 442 episode to episode because actions did not always lead to the expected outcome. In  
 443 addition, stochastic exploration strategies and random initialization of value functions  
 444 (or policy networks) exposed agents to slightly different states, transitions, and rewards  
 445 across runs. This variability made it harder to get perfectly consistent results.

446 An unexpected issue was that ~~even~~ the BFS baseline ~~sometimes~~ failed to reach the goal  
447 within the 1,000-step limit under slip conditions, resulting in a number of unsuccessful  
448 evaluation episodes.

449

450 A full numerical breakdown of evaluation metrics is provided in Appendix B.

451

452

---

453

## 454 **5. Discussion**

### 455 **5.1 Restatement of Hypothesis and Summary of Findings**

456 The study hypothesized that reinforcement learning-based agents, particularly Proximal  
457 Policy Optimization (PPO), would achieve more energy-efficient winter navigation paths  
458 than a traditional shortest-path baseline and a value-based Q-learning agent under  
459 both deterministic and slip winter conditions. ~~The averaged results partially supported~~  
460 ~~the hypothesis. The experiment result did not support the hypothesis.~~ Under  
461 ~~deterministic (non-slip) conditions, PPO Q-learning~~ achieved higher cumulative reward,  
462 ~~and fewer snow cell visits, outperforming Q-learning, which supports the hypothesis.~~  
463 ~~than Q learning PPO under both deterministic and slip conditions, but it also did~~  
464 ~~However under stochastic (slip) conditions, Q-learning it not~~ outperformed  
465 ~~PPO Q-learning~~ in terms of cumulative reward or snow-cell avoidance, ~~which did not~~  
466 ~~support the hypothesis.~~ Both reinforcement learning methods perform better results  
467 than traditional BFS across these environments.

468

### 469 **5.2 Interpretation of Q-Learning and PPO Performance**

470 ~~Q-learning achieved a higher performance than Proximal Policy Optimization agent in~~  
471 ~~both deterministic and stochastic winter environments. The results showed that~~  
472 ~~algorithm performance varied depending on the environmental condition. Under~~  
473 ~~deterministic (non-slip) conditions, PPO achieved higher cumulative reward and fewer~~  
474 ~~snow cell visits than Q-learning, suggesting that PPO's policy gradient approach was~~  
475 ~~able to learn a more energy-efficient route in a stable environment. Under stochastic~~  
476 ~~(slip) conditions, Q-learning outperformed PPO in terms of cumulative reward and~~  
477 ~~snow-cell avoidance.~~ One possible explanation was that a discrete and clear grid-world  
478 environment and relatively small environment would favor more for Q-learning,  
479 therefore allowing Q-learning to efficiently estimate optimal state [7]. In contrast, PPO  
480 relied on function approximation and stochastic policy updates, which may have  
481 required more training data or longer training time to converge to an equally optimal  
482 policy ~~under slip conditions.~~ Additionally, its stochastic policy may have introduced  
483 suboptimal choices that were not effective, therefore reducing its total rewards relative  
484 to Q-learning ~~under stochastic conditions~~ [3].

485

### 486 **5.3 Effects of Slip (Stochastic) Winter Conditions**

487 Under slip (stochastic) conditions, all agents showed broader trajectory dispersion and  
488 increased visits to previously visited states because of slipping, as reflected in the

489 spatial state visitation heatmaps. This behavior is consistent with probabilistic transition  
490 dynamics, in which the same action could lead to different movement outcomes across  
491 episodes, such as deviating left in one run and right in another. When the grid is  
492 slippery, agents don't follow one stable path anymore. As a result, the routes spread out,  
493 and their performance becomes less consistent across episodes. These observations are  
494 consistent with prior navigation studies showing that stochastic environments increase  
495 policy variance and reduce convergence stability in reinforcement learning tasks [8].

496

#### 497 **5.4 Interpretation of BFS Baseline Behavior**

498 The Traditional baseline trajectories did not take account for winter grid costs or  
499 stochastic transitions, it focused on the fastest way to the destination. In non-slip  
500 conditions this approach naturally produced the optimal route. However, under slip  
501 conditions, BFS exceeded the maximum step limit in multiple evaluation episodes,  
502 reflecting its inability to adapt its policy in response to transition winter grid costs or  
503 stochastic transitions.

504

505 Because BFS always assumed deterministic movement, each slip caused deviations from  
506 the planned path. These deviations often led to inefficient loops and longer paths.  
507 Unlike reinforcement learning methods, BFS did not have reward feedback or policy  
508 updates, which limited its ability to adjust navigation behavior under changing  
509 environmental dynamics. Therefore, making it the least effective method.

510

#### 511 **5.5 Limitations and Future Directions**

512 Several limitations should be noted when interpreting these findings. All experiments  
513 were carried out in a simulated 50×50 grid environment with a relatively small and  
514 simple state space, which differs significantly from real-world scenarios. This setup  
515 likely favored value-based methods like Q-learning, since it could learn precise  
516 state-action values for every cell. In contrast, PPO depends on neural network function  
517 approximation and stochastic policy updates, which may need more training steps or  
518 data to perform well in such discrete, small-scale settings.

519

520 ~~The number of evaluation episodes was also limited, which reduced the statistical~~  
521 ~~reliability of the results.~~ To improve statistical reliability, all experiments were run  
522 independently over 5 random seeds and results were averaged across runs. Noticeable  
523 performance variability was observed across training runs, mainly due to the stochastic  
524 environment, random action outcomes, and probabilistic transitions under slip  
525 conditions.

526 The simulated environment simplified real-world winter driving conditions and did not  
527 capture vehicle dynamics, sensor noise, road-surface variation, or realistic  
528 energy-consumption models, these are factors that might lead to changes. As a result,  
529 these findings to real-world winter navigation scenarios were limited [9].

530 Future research could evaluate reinforcement learning agents in larger scale,  
531 continuous, or more realistic winter navigation environments with higher-dimensional  
532 state spaces and more complex conditions. Then it may be better to reflect real-world  
533 uncertainty and could allow policy-gradient methods such as PPO to fully function their  
534 theoretical advantages in handling stochastic and high-dimensional control problems.

535

536

---

537

## 538 **6. Conclusion**

539 This study compared a traditional BFS baseline with two reinforcement learning  
540 methods—Q-learning and Proximal Policy Optimization (PPO)—for energy-efficient  
541 navigation in a 50×50 winter grid under both deterministic (non-slip) and stochastic  
542 (slip) conditions. The averaged results across 5 independent random seeds partially  
543 supported the original hypothesis. ~~The results did not support the original hypothesis~~  
544 ~~that PPO would produce more energy efficient paths than Q-learning or BFS. Instead,~~  
545 ~~Q-learning achieved the highest cumulative rewards in both settings and outperformed~~  
546 ~~PPO in snow-cell avoidance.~~ Under deterministic (non-slip) conditions, PPO achieved  
547 the highest cumulative reward and fewest snow cell visits, outperforming Q-learning.  
548 However, under stochastic (slip) conditions, Q-learning outperformed PPO in  
549 cumulative reward and snow-cell avoidance. Although both reinforcement learning  
550 methods clearly outperformed the BFS baseline, BFS struggled under slip conditions  
551 due to its inability to adapt to stochastic transitions.

552

553 These findings show that whether value-based methods or policy-gradient methods  
554 perform better really depends on how unpredictable the environment is. In our small,  
555 structured discrete grid, PPO gained an advantage from its stable gradients and  
556 curriculum learning when everything was deterministic. On the other hand,  
557 Q-learning's simple tabular updates turned out to be more robust when the roads  
558 became slippery and actions sometimes failed.¶

559 ~~These findings highlight that the relative performance of value based and~~  
560 ~~policy gradient methods is highly dependent on the degree of environmental~~  
561 ~~stochasticity. In this small and structured discrete state space, PPO benefits from stable~~  
562 ~~gradients and curriculum learning in deterministic settings, while Q-learning's tabular~~  
563 ~~off-policy updates provide greater robustness under stochastic transitions. These~~  
564 ~~findings indicate that in small and structured state spaces, PPO demonstrated stronger~~  
565 ~~performance in stable deterministic settings, while Q-learning showed more~~  
566 ~~adaptability under stochastic conditions value based methods can be more effective~~  
567 ~~than policy gradient approaches when environmental complexity is limited.~~ More  
568 generally, the results emphasize that the choice of reinforcement learning method  
569 should be guided by the structure and constraints of the task, rather than by theoretical  
570 advantages alone. Future work could explore whether PPO's advantages extend to more

571 complex environments with continuous states or larger state spaces, where its  
572 policy-gradient approach may better demonstrate its theoretical strengths.

573

574

575

---

576

## 577 Reference

578

579 [1]Mao, R., Xu, W., Qian, Y., Li, X., Li, Y., Li, G., & Zhang, H. (2025).

580 Understanding the Determinants of Electric Vehicle Range: A Multi-

581 Dimensional Survey. *Sustainability*, 17(10), 4259. [https://doi.org/](https://doi.org/10.3390/su17104259)

582 [10.3390/su17104259](https://doi.org/10.3390/su17104259)

583 [2]Carlson, A., & Vieira, T. (2021). *The effect of water and snow on the*

584 *road surface on rolling resistance* (VTI Report 971A). Swedish National

585 Road and Transport Research

586 Institute. <https://www.diva-portal.org/smash/get/diva2:1542142/>

587 [FULLTEXT01.pdf](https://www.diva-portal.org/smash/get/diva2:1542142/FULLTEXT01.pdf)

588 [3]Watkins, C.J.C.H., & Dayan, P (1992). Q-learning. *Mach Learn* 8, 279–292

589 ~~(1992).~~

590 <https://doi.org/10.1007/BF00992698>

591 [4]Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O.

592 (2017). *Proximal Policy Optimization Algorithms* (No.

593 arXiv:1707.06347). arXiv. <https://doi.org/10.48550/arXiv.1707.06347>

594 [5]Sutton, R. S., & Barto, A. G. (2018n.d.). Reinforcement Learning: An

595 Introduction.

596 [https://web.stanford.edu/class/psych209/Readings/](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)

597 [SuttonBartoIPRLBook2ndEd.pdf](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)

598 [6]Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and

599 Reinforcement Learning. *Neuron*, 36(2), 285–298. [https://doi.org/](https://doi.org/10.1016/S0896-6273(02)00963-7)

600 [10.1016/S0896-6273\(02\)00963-7](https://doi.org/10.1016/S0896-6273(02)00963-7)

601 [7]Tan, C. (2025). Comparative Study of Reinforcement Learning

602 Performance Based on PPO and DQN Algorithms. *Applied and*

603 *Computational Engineering*, 175(1), 30–36. [https://doi.org/](https://doi.org/10.54254/2755-2721/2025.AST24879)

604 [10.54254/2755-2721/2025.AST24879](https://doi.org/10.54254/2755-2721/2025.AST24879)

605 [8]Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino,

606 A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., &

607 Hadsell, R. (2017). *Learning to Navigate in Complex Environments* (No.

608 arXiv:1611.03673). arXiv. <https://doi.org/10.48550/arXiv.1611.03673>

609 [9]Chukwurah, N., Adebayo, A. S., Ajayi, O. O., & Anfo Pub. (2024).

610 *Sim-to-Real Transfer in Robotics: Addressing the Gap between*

611 *Simulation and Real- World Performance*. *International Journal for Multidisciplinary*

612 *Research (IJFMR)*, 05(01), 33–39. <https://doi.org/>

613 [10.54660/IJFMR.2024.5.1.33-39](https://doi.org/10.54660/IJFMR.2024.5.1.33-39)

614 [10]Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009).  
615 Curriculum learning. *Proceedings of the 26th Annual International*  
616 *Conference on Machine Learning*, 41–48. [https://doi.org/](https://doi.org/10.1145/1553374.1553380)  
617 [10.1145/1553374.1553380](https://doi.org/10.1145/1553374.1553380)

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

## 636 **Appendices**

### 637 **Appendix A Experimental Configuration and Algorithmic Details:**

638

#### 639 **A.1 Environment Configuration**

640 All experiments were conducted in a custom 50×50 winter grid environment (2,500  
641 discrete states).

- 642 • Start position: (47, 2)
- 643 • Goal position: (20, 45)
- 644 • Maximum steps per episode: 1000
- 645 • Action space: {up, down, left, right}

646 Two transition settings were evaluated:

647 Deterministic (slip = FALSE)

648 The intended action is executed exactly.

649 Stochastic (slip = TRUE)

650 With probability  $\frac{1}{3}$ , the intended action is executed.

651 With probability  $\frac{2}{3}$ , the agent slips to a random adjacent direction.

652 All algorithms were evaluated on the same fixed grid layout to ensure fairness.

653

## 654 A.2 Reward Function

655 The reward function models energy-aware winter navigation.

656 At each time step:

- 657 • Step penalty: -1.5
- 658 • Snow penalties:
  - 659 ○ Near snow: -0.2
  - 660 ○ Edge snow: -0.5
  - 661 ○ Core snow: -2.0
- 662 • Goal reward: +700

663 The cumulative episode reward is:

$$664 R = \sum_{t=1}^T (-1.5 - C_{snow}(s_t)) + 700 \cdot 1_{goal\ reached}$$

665

666 where  $C_{snow}(s_t)$  denotes the terrain penalty and

667  $1_{goal\ reached}$  indicates successful termination.

## 668 A.3 Q-Learning Configuration

669 Tabular Q-learning was implemented with the following hyperparameters:

- 670 • Learning rate  $\alpha=0.15$
- 671 • Discount factor  $\gamma=0.99$
- 672 • Exploration strategy:  $\epsilon$ -greedy
- 673 • Initial  $\epsilon = 0.60$
- 674 • Minimum  $\epsilon = 0.05$
- 675 • Exponential decay per episode = 0.9995
- 676 • Training episodes = 15,000
- 677 • Max steps per episode = 1000

678 The Q-update rule is:

$$679 Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

680 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) was used.

681

#### 682 A.4 Proximal Policy Optimization (PPO) Curriculum Configuration

683 PPO was implemented as a stochastic policy-gradient method.

684 Total training timesteps: 1,200,000

- 685 • Phase 1 (**navigation-focused reward deterministic environment**): 300,000
- 686 timesteps
- 687 • Phase 2 (**energy-aware reward stochastic environment**): 900,000 timesteps

688 Evaluation was conducted using deterministic action selection.

689

690 PPO Objective Function

691 PPO optimizes the clipped surrogate objective:

$$692 L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)]$$

693 where

$$694 r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

695 and  $\hat{A}_t$  is the generalized advantage estimate.

696 The clipping mechanism constrains policy updates to maintain stability.

697

698 PPO Hyperparameters

- 699 • Discount factor  $\gamma=0.99$   ~~$\gamma=0.99$~~
- 700 • GAE parameter  $\lambda=0.95$   ~~$\lambda=0.95$~~
- 701 • Entropy regularization enabled
- 702 • Learning rate:  $3 \times 10^{-4}$

- 703 • Clip range: 0.2
- 704 • Rollout buffer: 2,048 steps
- 705 • Batch size: 256
- 706 • Optimization epochs: 10
- 707 • Neural network architecture: two hidden layers (256 units each, ReLU
- 708 activation)
- 709

## 710 A.5 Evaluation Protocol

711 For each condition (slip FALSE / TRUE):

- 712 • One trained model per algorithm
- 713 • Maximum evaluation length: 1000 steps
- 714 • **Number of evaluation episodes: 200**
- 715 • **Number of seeds: 5**
- 716 • **Heatmap runs: 10**
- 717 • Metrics recorded:
  - 718 ○ Total cumulative reward
  - 719 ○ Number of steps
  - 720 ○ Snow cell visits
  - 721 ○ **Success rate (%)**

722 State visitation heatmaps were generated by aggregating visitation frequencies across  
 723 evaluation runs of 200 episodes each.

724

---

## 725 Appendix B Detailed Quantitative Results:

### 726 B.1 Deterministic (slip = FALSE)

Method	Reward	Steps	Snow Visits	Success
BFS Shortest	575.30±0.00	70.00±0.00	31.00±0.00	100.0%
Q-Learning	579.10±5.61	70.00±0.00	12.60±3.29	100.0%
PPO Curriculum	587.5±0.60	70.00±0.00	7.80±1.10	100.0%

727 All methods reached the goal within 70 steps.  
 728 PPO Curriculum achieved the highest cumulative reward and fewest snow cell visits, due  
 729 to reduced snow traversal.

730

731 **B.2 Stochastic (slip = TRUE)**

Method	Reward	Steps	Snow Visits	Success
<b>BFS Shortest</b>	<del>-1693.9639.2</del> ±61.27	1000 ±0.00	<del>246.80188.0</del> ±76.09	0%
<b>Q-Learning</b>	<del>304.03309.4</del> ±9.92	<del>228.184.2</del> ±4.10	<del>67.5865.8</del> ±4.34	100.0%
<b>PPO Curriculum</b>	<del>252.19224.1</del> ±7.04	<del>242.2955.8</del> ±3.00	<del>92.65100.9</del> ±4.14	100.0%

732 Under stochastic dynamics:

- 733 • BFS fails due to inability to adapt.
- 734 • Q-learning demonstrates greater robustness.
- 735 • PPO maintains success but exhibits higher trajectory dispersion.

736

737 **B.3 State Visitation Analysis**

738 State visitation frequency was computed as:

739 
$$V(s) = \frac{N(s)}{\max_s N(s)}$$

740 where  $N(s)$  is the number of visits to state  $s$ .

- 741 • Under deterministic conditions, visitation concentrates along a narrow corridor.
- 742 • Under stochastic conditions, visitation becomes more dispersed.
- 743 • Q-learning exhibits more focused routing than PPO under slip conditions.

Dear Reviewer 1,

Thank you for your thoughtful comments on my research paper. I have carefully addressed each of your comments.

After submitting the first manuscript, I continued improving the study by extending the code and conducting more advanced multi-seed experiments. In response to the reviewers' comments, I updated the manuscript with these results.

**Comment 1:** several areas could be strengthened. First, the introduction would benefit from more consistent and rigorous citation support. For example, claims such as “rolling resistance can rise by approximately 30–40% under wet or snowy conditions” and statements about the validity of the grid abstraction would be more convincing if directly supported with citations. Strengthening this connection to prior work would improve both credibility and scholarly rigor.

**Response 1:** After viewing the comment, I have added citation [2] (Carlson & Vieira, 2021) to the sentence “rolling resistance can rise by approximately 30–40% under wet or snowy conditions”. Additionally, I also added a justification for the grid abstraction, supported by a citation of [5] (Sutton & Barto, 2018). 1.) “Grid-based environments are commonly used in reinforcement learning studies, as they simplify complex continuous real-world environments into discrete states [5]. This abstraction is particularly well-suited for this study, as it enables environmental factors such as snow coverage and surface slip conditions to be incorporated directly into the cell-level cost and reward structure, supporting controlled and reproducible comparison of different routing methods.” Please view line 47 and line 55-60.

**Comment 2:** Second, while the motivation is clearly articulated, the paper would benefit from a more explicit discussion of gaps in existing research. The authors highlight the importance of multi-criteria navigation under winter conditions, but it remains unclear what specific limitation in the current literature this study addresses. Clarifying whether the primary contribution is extending RL-based navigation to account for weather-related energy costs (or something more novel) would help better position the work.

**Response 2:** I have added a paragraph to the introduction stating the research gap. Specifically, we note that prior work has primarily focused on general navigation tasks without explicitly modeling weather-dependent energy costs. “Despite growing interest in reinforcement learning for navigation, existing studies have largely overlooked the impact of winter environmental conditions, such as snow coverage and surface slippiness, on energy-aware path planning. Prior work has primarily focused on general navigation without taking account of explicitly modeling weather-dependent energy costs [8]. To address this gap, this study integrates snow coverage, stochastic slip condition, and energy costs into a comparative framework, evaluating Q-learning, PPO, and a traditional Baseline under realistic winter routing scenarios. This study is one of the first to directly compare value-based and

policy-gradient reinforcement learning methods in an energy-aware winter navigation setting.” Please view line 62-71.

**Comment 3:** Third, aspects of methodological rigor could be improved. The paper would benefit from more justification of hyperparameter choices and additional experimental robustness, such as averaging results over multiple runs and reporting variance. This would strengthen confidence in the findings, especially given the stochastic elements of the environment.

**Response 3:** I have added a section explicitly to justify the hyperparameter choices for Q-learning and PPO in method 3.4. Moreover, I have made the results to an average of 5 independent seeds and including the standard deviation for more accurate results and variance in both deterministic and stochastic conditions. Therefore, the study has been updated to reflect these improved results. Please view 3.4 and Result.

**Extra revisions:** Following the averaging of results over 5 independent random seeds, the findings changed slightly from the original submission. Specifically, PPO outperformed Q-learning under deterministic conditions, partially supporting the original hypothesis. As a result, the Abstract, Methods, Discussion, and Conclusion have been updated to accurately reflect these new findings.

Dear reviewer 2,

Thank you for your positive and constructive feedback. I greatly appreciate your careful evaluation and helpful suggestions. I have revised the manuscript accordingly and addressed your comments as follows.

After submitting the first manuscript, I continued improving the study by extending the code and conducting more advanced multi-seed experiments. In response to the reviewers' comments, I updated the manuscript with these results.

**Comment 1:** Methods section needs the most improvement, especially the explanations of the environment, how rewards were given, and the training procedures. That is where the paper becomes the most repetitive and hardest to follow. Tightening that section and making the setup more direct would do the most to improve the paper.

**Response 1:** I have consolidated all key experimental details into the main Methods section. Specifically:

- The 50×50 grid, start position (47, 2), and goal position (20, 45) are stated in Section 3.1
- Slip probabilities (1/3 intended action, 2/3 random direction) are now explicitly stated in Section 3.1
- Reward and penalty values (-1.5 step penalty, -0.2/-0.5/-2.0 snow penalties, +700 goal reward) are stated in Section 3.2.3
- Q-learning training length (15,000 episodes) is stated in Section 3.2.2
- PPO training length (1,200,000 timesteps) is stated in Section 3.2.4
- Evaluation procedure (200 episodes per run, 5 seeds) is stated in Section 3.3

Additionally, I have removed repetitive descriptions in the Methods section to improve clarity and readability. A new Section 3.4 has been added to provide explicit justification for all hyperparameter choices for both Q-learning and PPO. Appendix A is retained as a complete numerical summary for reference.

**Comment 2:** The paper includes a Works Cited / References section, and there are corresponding citations throughout the paper. The main issue is minor consistency and formatting, not the absence of citation.

## **Response 2:**

1. Corrected the formatting of [3] Watkins & Dayan (1992) to follow standard APA style
2. Updated [5] Sutton & Barto from (n.d.) to the correct publication year (2018)
3. Added the full journal name to [9] — *International Journal for Multidisciplinary Research (IJFMR)*

**Comment 3:** The Methods section would benefit from a clearer presentation of the experimental setup. At present, important details are split between the main text and Appendix A. I would suggest bringing the core information together in one place and stating it plainly: the 50×50 grid, the start and goal positions, the slip probabilities, the reward and penalty values, the training length for Q-learning and PPO, and the evaluation procedure. It would also help to replace general phrases such as “large rollout buffers” and “multiple optimization epochs” with the actual PPO settings, and to state directly how many evaluation episodes were used.

**Response 3:** I have made the following revisions:

### **Consolidated key details into the main Methods section:**

- 50×50 grid, start position (47, 2), and goal position (20, 45) — Section 3.1
- Slip probabilities (1/3 intended action, 2/3 random direction) — Section 3.1
- Reward and penalty values (-1.5 step penalty, -0.2/-0.5/-2.0 snow penalties, +700 goal reward) — Section 3.2.3
- Q-learning training length (15,000 episodes) — Section 3.2.2
- PPO training length (1,200,000 timesteps) — Section 3.2.4
- Evaluation procedure (200 episodes per run, 5 seeds) — Section 3.3

### **Replaced vague PPO language with actual values:**

- rollout buffer of 2,048 steps
- 10 optimization epochs per update
- Batch size explicitly stated as 256

### **Evaluation episodes count explicitly stated:**

- 200 episodes per run

Additionally, repetitive descriptions have been removed, and a new Section 3.4 has been added for explicit hyperparameter justification.

**Extra revisions:** Following Reviewer 1's suggestion to average results over 5 independent random seeds, the findings changed slightly from the original submission. Specifically, PPO outperformed Q-learning under deterministic conditions, partially supporting the original hypothesis. As a result, the Abstract, Methods, Discussion, and Conclusion have been updated to accurately reflect these new findings.

I have reviewed the authors' letter and changes and am very happy with the updates they made and would like to recommend this paper be accepted for publication.

It is clear they took the feedback seriously and made meaningful improvements to both the framing and rigor of the study. The additions to the introduction and method sections significantly strengthen the paper. In particular, the rationale and importance of the paper as well as decisions around hyperparameters are much clearer. Additionally, moving to multi-seed experiments with reported averages and standard deviations meaningfully improves the robustness of the results, which is particularly important given the stochastic nature of the environment and reinforcement learning methods. The updated findings add nuance to their conclusions and strengthen the comparative insights between value-based and policy-based methods.

Overall, these revisions substantially improve the paper's clarity, rigor, and contribution. The study now provides a more compelling and well-supported comparison of RL approaches for energy-aware navigation under winter conditions.

The revised manuscript is much stronger than the earlier version and is now, in my opinion, suitable for acceptance. The paper offers a clear comparison of three route-planning approaches in a winter grid-world setting: BFS, Q-learning, and PPO. Its main strength is that all three methods are tested under the same setup, including the same map, reward structure, start and goal positions, and evaluation criteria. That makes the comparison straightforward and allows the differences among the methods to come across clearly.

The revision has improved the paper in important ways, especially in the Methods section. In the earlier version, some of the setup was harder to follow, but the revised manuscript presents the experimental design more directly. The additional information about the environment, training process, evaluation methods, and Proximal Policy Optimization (PPO) settings improves the paper's clarity and makes it easier to understand. Moreover, the use of multiple runs with different random seeds strengthens the reliability of the reported results.

The Results and Discussion sections remain the strongest parts of the manuscript. The revised paper now gives a more careful interpretation of the findings. Rather than pushing a single overall winner, it shows that performance depends on the setting: PPO does slightly better in the deterministic case, while Q-learning is more robust under slip conditions. That makes the conclusion more persuasive.

Overall, I found the paper well organized, readable, and worth publishing. The subject matter is pertinent, and the comparison is well-articulated. This revised manuscript represents a valuable contribution, particularly as a simulation-based study. I recommend its acceptance.

## **Decision: Accept with moderate revisions**

**Review:** This manuscript presents a comparative study of three path-planning approaches (BFS, Q-learning, and PPO) under a grid-world abstraction designed to model winter driving conditions via energy penalties and stochastic transitions. The study evaluates performance using cumulative reward, snow-cell traversal, success rate, and trajectory characteristics under both deterministic (“non-slip”) and stochastic (“slip”) settings.

The core finding is that Q-learning outperforms PPO in stochastic environments, while PPO may perform competitively or better in deterministic settings. The paper aims to position this as evidence that value-based methods can outperform policy-gradient methods in structured, discrete environments.

The experimental setup is clear and has improved after revisions, notably via multi-seed evaluation and clearer methodological specification. These are laudable; however, the manuscript still falls a bit short of the level of rigor, positioning, and analysis expected for publication in *Convergence Journal*. I believe the framing, methodology, and experimental design/setup are there, but the execution overall (writing + experiments) could be improved.

### Major comments:

1. While the revision introduces multi-seed experiments using 5 seeds, this could be strengthened for stochastic RL evaluation. There are also no tests for statistical significance, no confidence intervals, and no discussion of variability across runs.
  - a. I would like the author to increase the experiments to at least 10 (maybe even 20) seeds and report statistics (e.g., mean +/- standard deviation, median + 95% confidence intervals, etc.). Just something of this sort would be great and help alleviate concerns about the robustness and consistency of your results.
  - b. Include at least one statistical test (e.g., Welch’s t-test) comparing Q-learning vs. PPO (slip condition)
  - c. Alongside these tests, you should add a short paragraph discussing and interpreting the variance. In particular, I would want you to answer whether PPO exhibits higher instability.
2. The research gap still does not entirely distinguish between RL path planning literature, cost-aware navigation, and grid-world RL benchmarks. Could you add another paragraph answering more explicitly what prior work has been done, what is missing, and what exactly is novel
3. Please add a plot of learning curves: training reward vs. timesteps for **both** Q-learning and PPO, along with the statistics from point 1 above (CIs, mean +/- stds, whatever you end up using). Would reveal instability, too, especially for PPO

### Minor comments:

1. The paper states outcomes (e.g., Q-learning outperforms PPO) but could explain these in a deeper, more mechanistic way. *Why* does Q-learning work better in discrete grid worlds? *Why* does PPO struggle?

- a. For the first question, consider: exact value iteration in finite MDP, lower variance than policy gradients
  - b. For the second question, consider: high variance gradients under stochasticity, issues with sparse and noisy rewards
2. All results are based on a single hyperparameter configuration per method, which is concerning. How stable are the results to changes in hyperparameters? Could you add at least one sensitivity analysis? You do not need to go overboard, but this could be good to alleviate skepticism towards claims about one method being more suited over the other...
  - a. For Q-learning: vary the decay or learning rate
  - b. For PPO: vary learning or clipping parameter
  - c. This should be accompanied by a figure
3. Since curriculum is highlighted, it should ideally be validated. While I like how you introduced the curriculum learning setup, I am still wondering if the PPO performance depends heavily on the curriculum. Could you compare the PPO with and without the curriculum, or at least cite some literature that would answer this question? You may want to accompany this with some sort of ablation study plot(s)
4. Watch the informal phrasing throughout (e.g., the use of "etc.")
5. I would add at least 2 more papers on PPO vs. value-based methods and/or RL navigation benchmarks

# 1 Energy-Efficient Path Planning in Winter Conditions: A Comparative 2 Study of Traditional Baseline, Q-learning, and Proximal Policy 3 Optimization in a Grid World Environment

4  
5

## 6 1.Abstract

7 Recent investigation in winter route planning has become increasingly important due to  
8 increased resistance and reduced efficiency on winter roads. These conditions  
9 significantly increase energy consumption and introduce uncertainty, raising the risk of  
10 failing to reach the destination. To address the challenge, Reinforcement Learning (RL)  
11 is a type of machine learning where an agent learns to make decisions by interacting  
12 with an environment, receiving rewards or penalties for its actions to maximize  
13 cumulative rewards. For simulation, we benchmark a standard shortest-path algorithm  
14 (BFS) against two reinforcement learning methods: Q-learning and Proximal Policy  
15 Optimization (PPO). All approaches are tested on identical grid configurations, with the  
16 same starting points, goals, and reward functions. We assess their performance based on  
17 cumulative reward, snowy cell visits (as a proxy for energy cost), trajectory characteristics,  
18 and success rate. It is hypothesized that PPO algorithms will result in lower energy  
19 consumption than Q-learning and traditional baseline under winter conditions. The  
20 results show that while BFS consistently finds the shortest paths, it fails to consider  
21 energy costs or environmental uncertainty. RL agents, by comparison, adapt more  
22 efficiently to winter conditions. Under deterministic (non-slip) conditions, PPO  
23 achieved higher cumulative reward and fewer snow cell visits than Q-learning, partially  
24 supporting the hypothesis. However, under stochastic (slip) conditions, Q-learning  
25 outperformed PPO in cumulative reward and snow-cell avoidance. These results  
26 suggest that Q-learning is better suited for stochastic winter navigation in grid worlds,  
27 while PPO may perform better in more deterministic environments with continuous  
28 states or decisions.

29

30 Key words: Energy-Efficient, Traditional Baseline, Q-learning, Proximal Policy  
31 Optimization, grid world, reinforcement learning.

32

33

---

34

## 35 2.Introduction

36 As demand in Electric Vehicles (EV) become increasingly widespread, energy efficiency  
37 and route optimization have emerged as critical challenges, particularly in winter when  
38 resistance increases. Sub-zero temperatures, combined with snow and ice  
39 accumulation on road surfaces, significantly increase energy consumption, resulting in  
40 reduced driving range and greater unpredictability in trip planning. A 2025 study shows  
41 that an estimated 50 % of EV driving range can be reduced in cold climates, including  
42 snow and ice covering terrain, highlighting the significant impact of environmental

43 conditions on energy consumption [1]. Furthermore, when snow or water remains on  
44 the road surface, vehicle tires must continuously displace through ice as they roll and  
45 move, hence forcing the vehicle to draw more power [2]. On top of that, water cools  
46 tires more effectively than air alone, altering their mechanical properties and further  
47 pushing rolling resistance higher. Previous studies have shown that rolling resistance  
48 can rise by approximately 30% to 40% under wet or snowy conditions [2]. Ultimately,  
49 these factors make winter driving a major concern for Electric Vehicle (EV) efficiency  
50 and reinforce the need for energy-aware routing strategies.

51

52 To approach these challenges, we have explored a few computational methods to allow  
53 agents to learn effective navigation strategies aimed at minimizing energy consumption.  
54 In this study, a 50x50 grid is used as an abstraction routing map that circles key  
55 structures of Canadian snow distribution, rather than exact geographical terrain.  
56 Grid-based environments are commonly used in reinforcement learning studies, as they  
57 simplify complex continuous real-world environments into discrete states [3]. This  
58 abstraction is particularly well-suited for this study, as it enables environmental factors  
59 such as snow coverage and surface slip conditions to be incorporated directly into the  
60 cell-level cost and reward structure, supporting controlled and reproducible  
61 comparison of different routing methods.

62

63 Despite growing interest in reinforcement learning for navigation, existing studies have  
64 largely overlooked the impact of winter environmental conditions, such as snow  
65 coverage and surface slippiness, on energy-aware path planning. Prior work has  
66 primarily focused on general navigation without taking account of explicitly modeling  
67 weather-dependent energy costs [4]. To address this gap, this study integrates snow  
68 coverage, stochastic slip condition, and energy costs into a comparative framework,  
69 evaluating Q-learning, PPO, and a traditional baseline under realistic winter routing  
70 scenarios. This study is one of the first to directly compare value-based and  
71 policy-gradient reinforcement learning methods in an energy-aware winter navigation  
72 setting .

73

74 Prior work on reinforcement learning for navigation can be categorized into three  
75 directions. First, Grid world benchmarks commonly adopt tabular Q-learning and deep  
76 reinforcement learning variants as standard baselines for evaluating discrete navigation  
77 tasks [3,5], but focus primarily on task completion rather than energy awareness and  
78 energy sensitive routing. Second, cost-aware navigation research has examined energy  
79 minimisation in uneven terrains using reinforcement learning [6], yet rarely  
80 incorporated weather dependent components, such as snow coverage or stochastic  
81 traction loss. Third, existing comparisons between value-based and policy-gradient  
82 methods [7,8], including curriculum based PPO applications [9] are conducted primarily  
83 under fixed and stable conditions, without examining the possibility of a particular  
84 reinforcement learning method remaining dominant to another when action outcomes  
85 become probabilistic, as in stochastic environments such as winter slip conditions. The

86 work in this study addresses all three gaps simultaneously by implementing an  
87 energy-aware reward structure into a reproducible 50×50 grid-world benchmark,  
88 embedding snow-density penalties and a 2/3-slip probability to simulate realistic  
89 winter routing conditions, and directly comparing Q-learning, PPO, and a traditional  
90 baseline under both deterministic and stochastic environments.

91

92 More specifically, the research will include a traditional baseline algorithm, and two  
93 types of reinforcement learning, which are Q-learning and PPO algorithm. The  
94 traditional baseline routing computes the shortest path based on static cost metrics and  
95 follows the predefined route without adaptation or learning (basic routing). Q-learning,  
96 a value-based reinforcement learning method, learns through incremental dynamic  
97 programming processes with computational requirements and it is well suited for  
98 agents to improve and refine action values and achieve effective performance in  
99 controlled Markovian domains [7]. In contrast, PPO is a policy-gradient algorithm that  
100 primarily optimizes a stochastic policy and achieves greater training stability through  
101 constrained policy updates with a clipped surrogate objective function [8].

102

103 This comparison evaluates the effectiveness of the proposed routing methods in  
104 identifying energy-efficient paths. Performance is evaluated through energy related  
105 costs, overall routing efficiency, and observed navigation behavior under both  
106 deterministic (non-slip) and stochastic (slip) settings. Because PPO balances stable  
107 policy updates with adaptive learning in uncertain environments, it is hypothesized to  
108 be the most effective method implemented for winter routing.

109

110 The following section goes into the methodology and experimental setup in more detail.

111

112

---

113

## 114 **3.Methods**

### 115 **3.1 Environment Design**

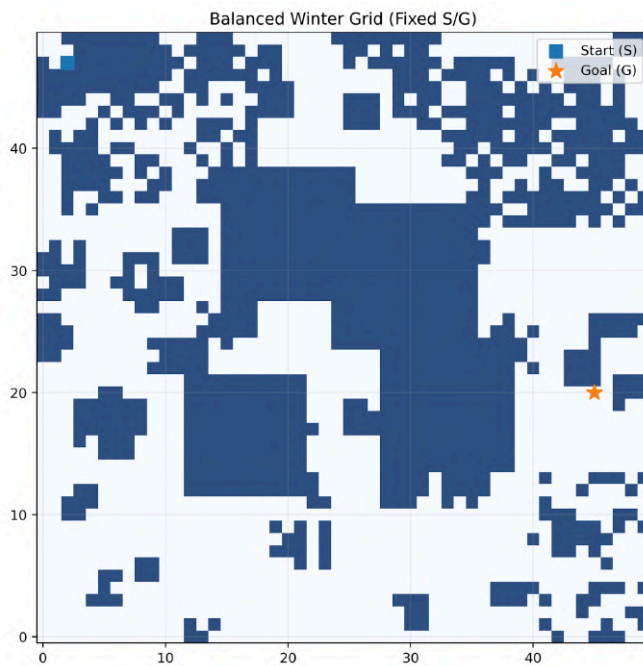
116 A custom winter navigation grid environment was designed to evaluate each  
117 Reinforcement Learning (RL) agent under stochastic and cost-sensitive conditions. The  
118 environment consists of a 50 x 50 grid world containing 2500 cells, including both  
119 normal terrain and snow-covered terrain. The simulation contains a fixed start (47, 2)  
120 and a fixed goal (20, 45) point, and are located in the top left corner and middle right  
121 respectively. At each time step, the agent can take either one of the four directions, up,  
122 down, left, and right; each step on a non-snow grid has an energy cost of -1.5 per step.  
123 Snow is modeled as spatially correlated regions with graded intensity, capturing the typical  
124 uneven accumulation patterns in Canadian winters.

125 When snow is modeled as a binary state, abrupt reward discontinuities make PPO's  
126 advantage estimates unstable, resulting in increased gradient variance. By contrast, a  
127 continuous representation of snow thickness provides smoother cost transitions, which

128 help reduce gradient variance and stabilize learning, making it more efficient. These  
129 snow types are classified as “Near snow”, “Edge snow”, and “Core snow”.

130 To evaluate robustness under different winter conditions, two transition settings were  
131 considered using the same grid map: a non-slip (deterministic) condition and a slip  
132 (stochastic) condition. In the deterministic setting, action always results in the intended  
133 movement, and snow cells only add additional energy loss. In the stochastic setting,  
134 actions do not always lead to the intended outcome: with a probability of 1/3, the  
135 intended action is executed, and with a probability of 2/3, the agent moves in a random  
136 adjacent direction, simulating loss of traction control during winter driving on ice and  
137 snow.

138



139

140 Figure 1: shows a typical winter grid layout used in all experiments.

141

## 142 3.2 Reinforcement Learning Algorithms/ Methodologies:

143 The routing strategies and learning methods used in this study are described in this  
144 section.

145

### 146 3.2.1 Traditional Baseline (BFS)

147 Traditional baseline is a non-learning shortest path strategy used as a comparison base  
148 for reinforcement learning methods. It operates on the 50x50 grid map, it considers  
149 four directions to move (up, down, left, right). The system calculates the shortest path  
150 to the destination, without considering energy savings, which means it does not adapt  
151 to environmental feedback or uncertainty.

152

### 153 3.2.2 Q-learning Implementation

154 Tabular Q-learning was used as the value-based reinforcement learning baseline. The agent  
155 maintains a discrete  $Q(s, a)$  table, which is updated iteratively based on observed state transitions.

156 An  $\epsilon$ -greedy policy was used for exploration. The value of  $\epsilon$  begins at 0.6 and decays  
 157 exponentially with a factor of 0.9995 per episode until it reaches 0.05. In this 50×50 grid  
 158 setting, the decay schedule allows broad initial exploration before shifting emphasis to  
 159 exploitation. Actions are restricted to four possibilities—left (0), down (1), right (2), and  
 160 up (3)—consistent with standard discrete grid-world formulations and the discrete grid  
 161 structure. The Q-value update follows the classic temporal-difference form:

162

$$163 \quad Q(s, a) \leftarrow Q(s, a) + \alpha [R + \gamma \max_{a'} Q(s', a') - Q(s, a)] \quad (\text{Eq. 1})$$

164

165 We fix the learning rate  $\alpha$  at 0.15 and the discount factor  $\gamma$  at 0.99. These values handle the  
 166 stochastic transitions (`is_slippery=True`) and support planning over long horizons in the large  
 167 grid [3]. State  $s$  denotes the flattened grid position index. The action  $a$  selects one of the four  
 168 movement directions.  $Q(s, a)$  represents the estimated discounted cumulative reward  
 169 starting from state  $s$  and action  $a$  [3,7].

170

### 171 3.2.3 Reward system

172 In this research, the reward system is designed to represent energy efficiency under  
 173 winter conditions, it helps agents find the best paths. At each step, the system gives a  
 174 negative value to show how significant a step is to energy saving, with higher cause on  
 175 snow grids. A positive 700 is given only when the agents successfully reach the  
 176 destination, while within the maximum steps of 1000 steps. It encourages agents to  
 177 reach the destination with considerations on energy saving. We initialize the cumulative  
 178 reward to 0 at the beginning of each episode. Each will result in negative values,  
 179 because by doing this, agents don't need to consider the prior bias, since it assumes  
 180 nothing at first. As in the settings, we have Near-snow (light snow), Edge-snow (medium  
 181 snow), and Core-snow (deep snow); each of them has different additional values which  
 182 are -0.2, -0.5, and -2.0 respectively [10].

183

$$184 \quad r_t = r_{step} - c_{energy}(s_t) + r_{goal} \quad (\text{Eq. 2})$$

185

186 The basic reward function follows standard reinforcement learning practice by  
 187 combining step penalties, energy-related costs, and a terminal goal reward.

188

### 189 3.2.4 Proximal Policy Optimization

190 Proximal Policy Optimization (PPO) served as the policy-gradient method in this study  
 191 for energy-efficient routing under winter conditions. PPO learns a stochastic policy by  
 192 directly outputting action probabilities for each state, which helps the agent cope with  
 193 uncertainty on slippery roads.

194

$$195 \quad L^{CLIP}(\theta) = E_t [\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \quad (\text{Eq. 3})$$

196

197 where  $r_t(\theta)$  is the probability ratio between the new and previous policies. The clipping  
198 mechanism constrains policy updates to improve stability.

199

200 The PPO agent was trained using fixed values across all experiments. The discount factor was set  
201 to  $\gamma=0.99$ , and Generalized Advantage Estimation (GAE) was used with  $\lambda=0.95$ . The policy  
202 network consisted of two fully connected hidden layers with 256 units each, using ReLU  
203 activation functions. The learning rate was set to  $3 \times 10^{-4}$ , with a clipping range of 0.2  
204 and an entropy coefficient of 0.02 to encourage exploration. PPO training was  
205 conducted for 1,200,000 time steps, using a rollout buffer of 2048 steps, a batch size of  
206 256, and 10 optimization epochs per update to improve training stability [8].

207

208 Policy-gradient methods work by directly adjusting a parameterized policy to increase  
209 expected reward. They learn a stochastic policy, allowing for more flexibility in  
210 uncertain environments, where the same action can lead to different outcomes. In  
211 contrast, q-learning estimates state-action values and usually converges with a  
212 deterministic policy based on these estimates. Although it can learn in unfamiliar  
213 environments, its action selection may be less reliable in highly unpredictable  
214 situations. As a result, policy-gradient and value-based methods show varying strengths  
215 depending on the level of uncertainty in the winter routing task.

216

### 217 3.2.5 Curriculum Learning for PPO

218 PPO training followed a two-phase curriculum learning approach designed to improve  
219 training stability and sample efficiency under sparse rewards and energy-sensitive  
220 penalties [11,8].

221

#### 222 **Phase 1: Navigation Learning Phase** (300,000 timesteps)

223 In phase 1, PPO was trained using a simplified reward function that focuses only on  
224 positive feedback for reaching the goal. Energy costs in snow grids were not included.  
225 This made the feedback denser and more immediate. The agent quickly learned basic  
226 navigation and achieved early success in the 50×50 winter grid. The policy learned in  
227 this phase served as the starting point for the next training stages.

228

#### 229 **Phase 2: Energy-Aware Optimization Phase** (900,000 timesteps)

230 In phase 2, the training continued with the complete energy-aware reward function,  
231 which now included penalties for moving across snow-covered areas. The agent had to  
232 balance successful arrival with lower energy consumption. All reported results, ablation  
233 studies, and comparisons are based solely on the Phase 2 reward function.

234

235

236

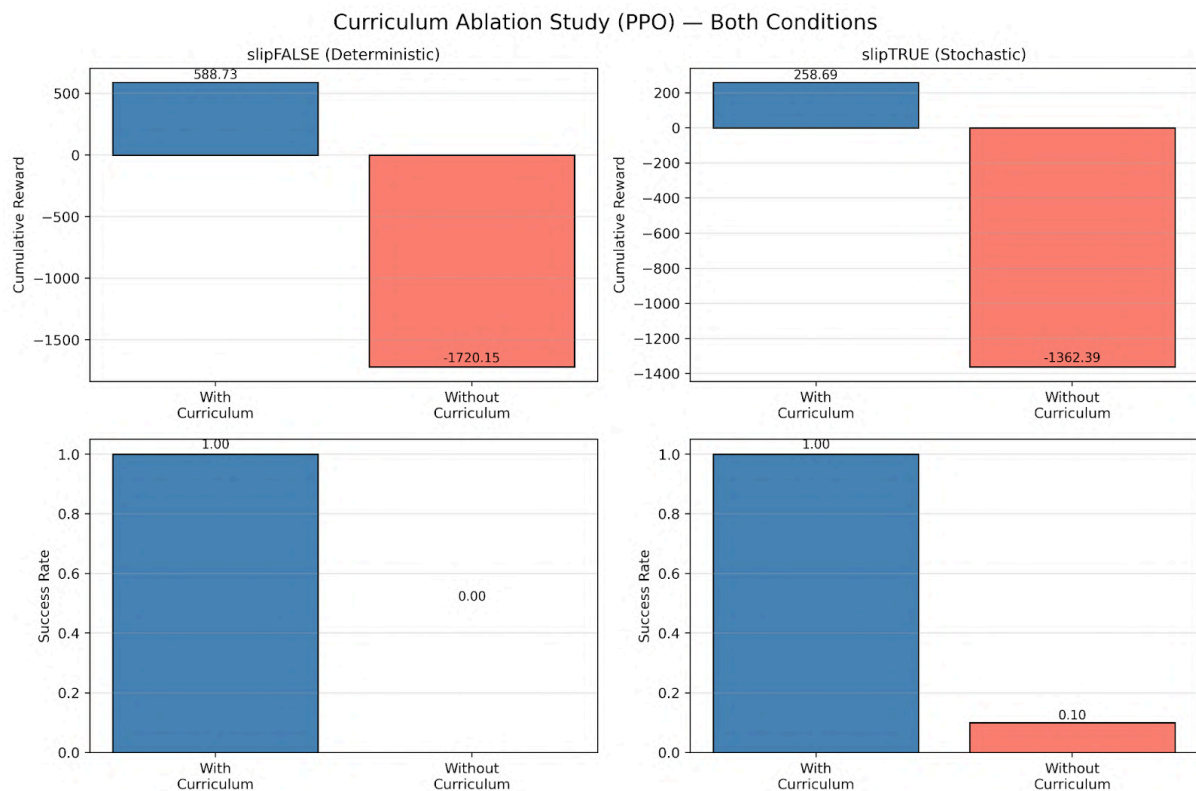
237

### 238 3.2.6 Curriculum Learning Ablation Study

239 To validate the necessity of curriculum learning, an ablation study was  
 240 conducted comparing PPO with curriculum versus PPO without curriculum  
 241 under both deterministic and stochastic conditions across 10 seeds  
 242 assessed through total cumulative reward and success rate. Without  
 243 Curriculum Learning, PPO achieved a mean of -1720.15 (95% CI [-1894.63,  
 244 -1545.67]) and -1362.39 (95% CI [-1780.01, -944.78]) with 0% and 10%  
 245 success rate respectively, under deterministic and stochastic conditions.  
 246 Whereas with curriculum, PPO demonstrates a strong performance of 588.73  
 247 (95% CI [588.25, 589.21]) and 258.69 (95% CI [251.02, 266.37]) with 100%  
 248 success rate under both conditions. These results confirm that curriculum  
 249 learning is a critical component for PPO, as its absence leads to failure  
 250 under both conditions.

251

252



253

254 Figure 2: displays the cumulative reward and success rate for PPO with and without  
 255 curriculum under both deterministic and stochastic conditions, illustrating the  
 256 substantial performance gap between the two configurations.

257

258 Therefore, all subsequent comparisons and evaluations in this study refer exclusively to  
 259 PPO with curriculum learning.

260

### 261 3.3 Evaluation factors

262 To ensure statistical reliability, all reinforcement learning experiments were repeated  
263 over 10 independent random seeds (seed 0 to seed 9), and results are reported as the  
264 mean and standard deviation with 95% confidence intervals across runs.

265

266 To compare the performances of all routing methods under identical conditions, a final  
267 evaluation was conducted after training. Each method BFS, Q-learning, and PPO was  
268 evaluated using the following metrics:

269

270 • Total Reward: calculated as the cumulative reward obtained over the episode.

271

272 • Number of steps: steps taken from the start to the goal (less than 1000 steps).

273

274 • Average snow cell visited: Snow visits were calculated by counting the number of  
275 snow-covered cells traversed in each episode and averaging this value across all  
276 evaluation episodes.

277

278 • Success rate: defined as the percentage of episodes in which the agent reached  
279 the goal within the maximum step limit.

280

281 The BFS baseline was evaluated only once on the fixed grid map, as it produces a  
282 deterministic path. Reinforcement learning agents were evaluated over 200 episodes  
283 per run, with experiments repeated over 10 independent random seeds for each  
284 condition (slip and non-slip). All methods were compared using the same energy-aware  
285 reward function and environment configuration to ensure fairness.

286

### 287 **3.4 Hyperparameter Selections**

288 Hyperparameters are selected based on standard practices and empirical validation to  
289 ensure stable and consistent performance [3,7].

290

#### 291 **3.4.1 Q-Learning Hyperparameter Configuration**

292 We implement tabular Q-learning with an  $\epsilon$ -greedy exploration strategy.

293

294 The hyperparameters are set as follows:

295 - Learning rate  $\alpha = 0.15$

296 - Discount factor  $\gamma = 0.99$

297 - Exploration strategy:  $\epsilon$ -greedy

298 - Initial  $\epsilon = 0.60$

299 - Minimum  $\epsilon = 0.05$

300 - Exponential decay rate = 0.9995

301 - Training episodes = 15,000

302 - Maximum steps per episode = 1,000

303

304 The learning rate ( $\alpha = 0.15$ ) is selected to balance convergence speed and stability. A moderate  
305 learning rate allows the agent to adapt efficiently while avoiding oscillations in value updates.

306

307 The discount factor ( $\gamma = 0.99$ ) emphasizes long-term rewards, which is essential for navigation  
308 tasks where the objective is to reach the goal efficiently over multiple steps.

309

310 An  $\epsilon$ -greedy exploration strategy is adopted to balance exploration and exploitation. The initial  
311 exploration rate ( $\epsilon = 0.60$ ) encourages sufficient exploration in early training, while exponential  
312 decay gradually shifts the policy toward exploitation. The minimum

313 ( $\epsilon = 0.05$ ) ensures that some level of exploration is maintained throughout training, preventing  
314 the agent from getting stuck in suboptimal policies.

315

316 The number of training episodes 15,000 and maximum steps per episode 1,000 are  
317 chosen to ensure sufficient interaction with the environment for convergence, while  
318 maintaining computational efficiency.

319

320 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) is used to assess the learned policy performance.

321

### 322 3.4.2 PPO Hyperparameter Configuration

323 We implement PPO using a clipped surrogate objective with generalized advantage  
324 estimation (GAE).

325

326 The hyperparameters are set as follows:

327 - Discount factor  $\gamma = 0.99$

328 - GAE parameter  $\lambda = 0.95$

329 - Learning rate =  $3 \times 10^{-4}$

330 - Clipping range  $\epsilon = 0.2$

331 - Rollout buffer size = 2048 steps

332 - Batch size = 256

333 - Optimization epochs per update = 10

334 - Entropy regularization is enabled

335

336 The policy and value networks share a neural architecture consisting of two hidden  
337 layers with 256 units each and ReLU activation.

338

339 The discount factor ( $\gamma = 0.99$ ) is selected to emphasize long-term rewards, which is essential for  
340 navigation tasks. The GAE parameter ( $\lambda = 0.95$ ) provides a balance between bias and variance in  
341 advantage estimation, leading to more stable learning.

342

343 The clipping parameter ( $\epsilon = 0.2$ ) constrains policy updates to prevent large deviations from the  
344 previous policy, which improves training stability. The rollout buffer size and number of

345 optimization epochs are chosen to ensure sufficient policy updates while avoiding overfitting to  
346 recent trajectories.

347

348 Entropy regularization is included to encourage exploration and prevent premature  
349 convergence to suboptimal policies.

350

351 Overall, these hyperparameters follow widely adopted settings in reinforcement  
352 learning literature and are chosen to ensure stable and consistent training rather than  
353 aggressive performance tuning.

354

355 A complete summary of environment settings and hyperparameters is provided in  
356 Appendix A.

357

358

359

## 360 4. Results

361 The experiments compare BFS, Q-learning, and PPO on the same 50×50 winter grid,  
362 looking at both non-slip (deterministic) and slip (stochastic) cases.

363

### 364 4.1 Evaluation Setup and Metrics

365 The grid layout, start and goal positions, and reward parameters were kept exactly the  
366 same across all methods. All experiments were repeated over 10 independent random  
367 seeds, and results are reported as the mean  $\pm$  standard deviation with 95% confidence  
368 intervals across runs. Trajectory figures shown are from the representative seed whose  
369 performance was closest to the mean. For each run, total reward, number of steps,  
370 average snow cells crossed per episode, and success in reaching the goal were recorded.

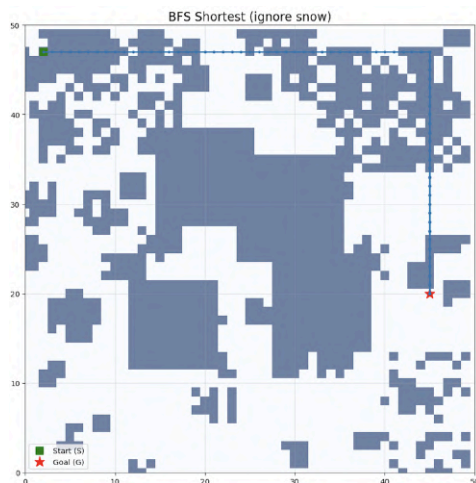
371

### 372 4.2 Trajectory Visualizations and Quantitative Results

#### 373 4.2.1 Deterministic (Non-Slip) Winter Condition

374

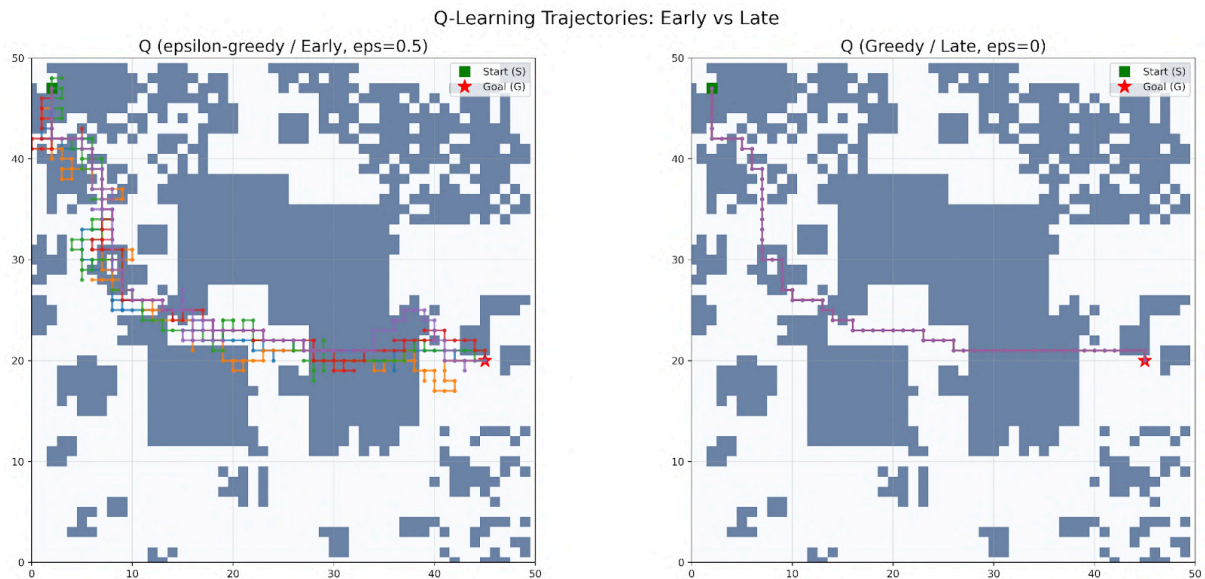
375



377 Figure 3: illustrates the BFS path under non-slip conditions. Since BFS always finds the  
378 shortest route, the trajectory goes straight from the start to the goal with minimal cells  
379 visited without any deviation.

380

381

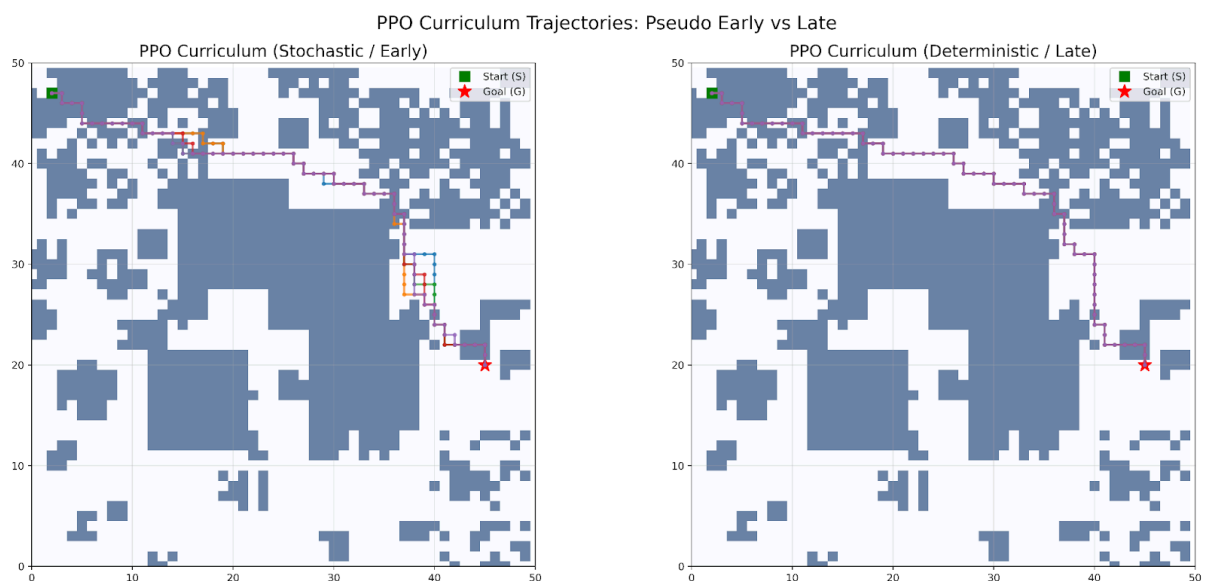


383

384 Figure 4: illustrates the navigation trajectories produced by the Q-learning agent under  
385 deterministic (non-slip) winter conditions during early training with an  $\epsilon$ -greedy policy and late  
386 training with a greedy policy. The trajectories demonstrate the transition from exploratory  
387 behavior to a stable path toward the goal.

388

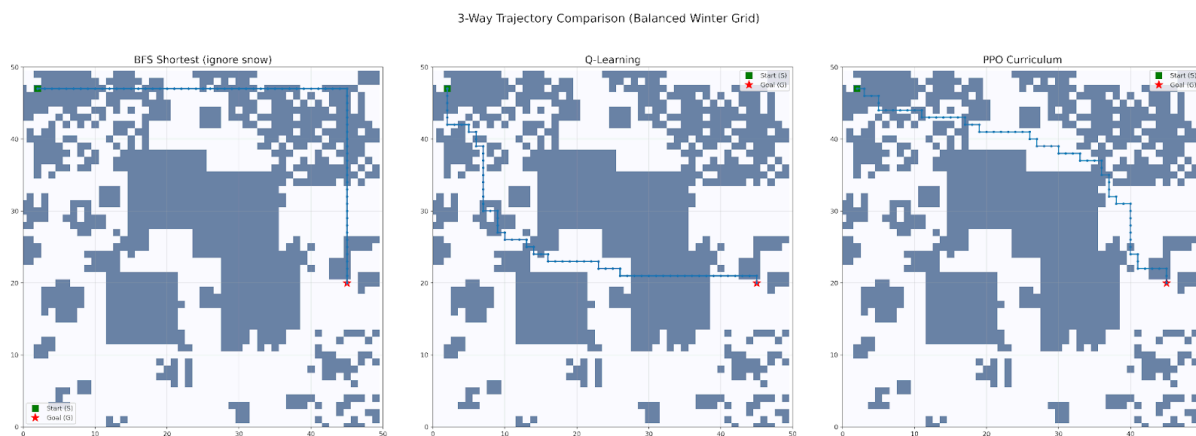
389



391 Figure 5: illustrates the navigation trajectories produced by the PPO agent during early  
392 and late stages of training under deterministic (non-slip) winter conditions.

393

394



395

396 Figure 6: compares the trajectories of BFS, Q-learning and PPO in the non-slip winter  
397 setting.

398

399

400 === Final Comparison Table ===

401

Method	Reward (mean± SD)	95%CI	Steps	Snow Visits	Success
<b>BFS Shortest</b>	575.30 ± 0.00	[575.30, 575.30]	70.00 ± 0.00	31.00 ± 0.00	100.0%
<b>Q-Learning</b>	582.01 ± 6.97	[577.02, 587.00]	70.00 ± 0.00	13.10 ± 3.05	100.0%
<b>PPO Curriculum</b>	588.73 ± 0.67	[588.25, 589.21]	70.00 ± 0.00	7.60 ± 0.97	100.0%

402

403 Table 1: summarizes the main performance metrics for all three methods under  
404 deterministic conditions, including total reward, steps to goal, average snow cells  
405 crossed, success rate, Standard Deviation (SD), and averaged across 10 independent  
406 random seeds with 95% confidence intervals.

407

408

409

410

411

412

413

414

415

416

417

## 418 4.2.2 Slip Winter Condition

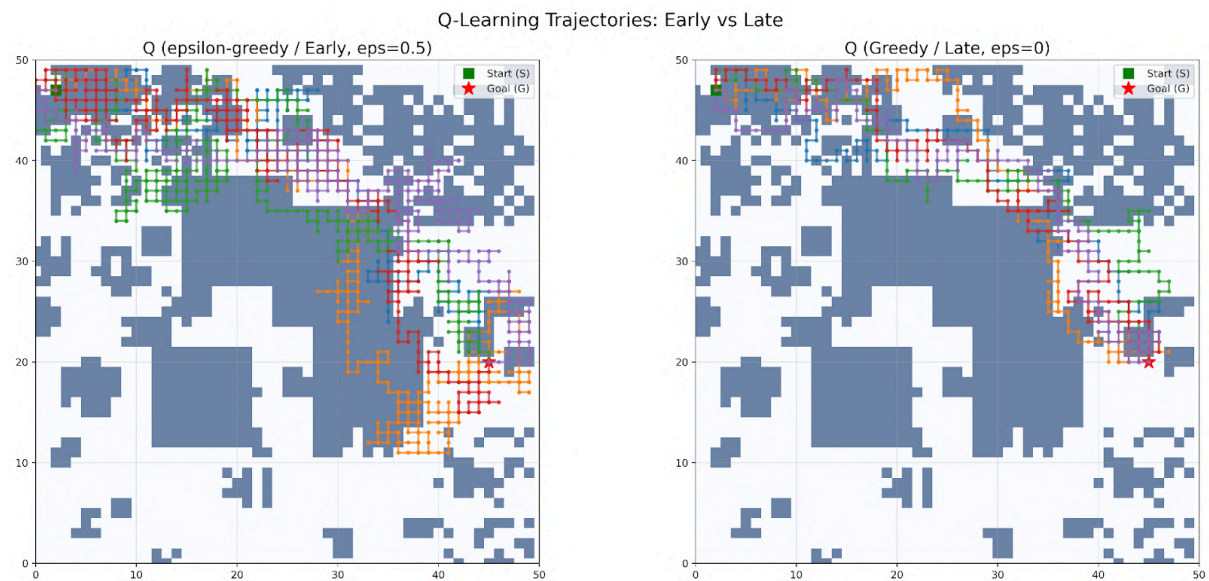


419

420 Figure 7: illustrates the navigation path produced by the BFS baseline under slip  
421 (stochastic) winter conditions.

422

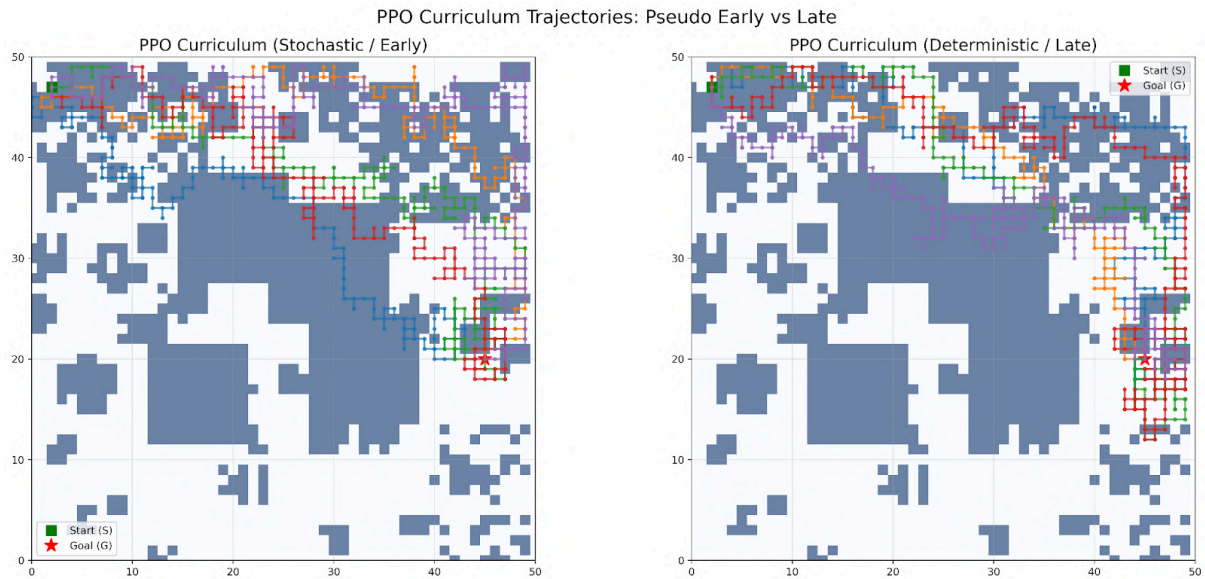
423



424

425 Figure 8: illustrates the navigation trajectories produced by the Q-learning agent during  
426 early and late stages of training under slip (stochastic) winter conditions.

427

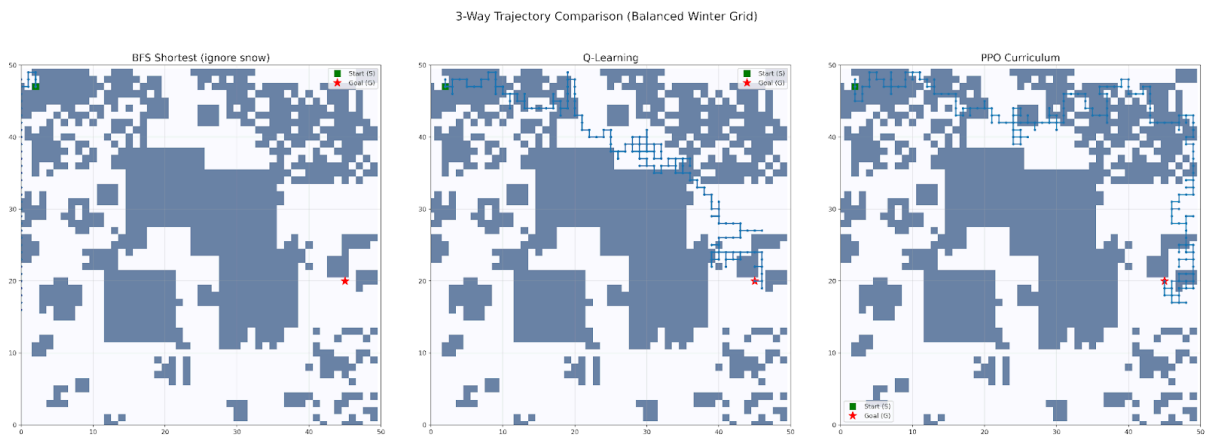


428

429 Figure 9: illustrates the navigation trajectories produced by the PPO agent during early  
 430 and late stages of training under slip (stochastic) winter conditions.

431

432



433

434 Figure 10: compares the navigation trajectories of BFS, Q-learning, and PPO under slip  
 435 winter conditions.

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450 === Final Comparison Table ===

451

Method	Reward (mean±SD)	95% CI	Steps	Snow Visits	Success
<b>BFS Shortest</b>	-1668.41 ± 113.42	[-1749.54, -1587.28]	1000.00 ± 0.00	246.80 ± 76.09	0.0%
<b>Q-Learning</b>	311.82 ± 4.83	[308.37, 315.27]	26.40 ± 3.92	66.94 ± 3.87	100.0%
<b>PPO Curriculum</b>	258.69 ± 10.73	[251.02, 266.37]	241.18 ± 4.11	91.88 ± 5.24	100.0%

452

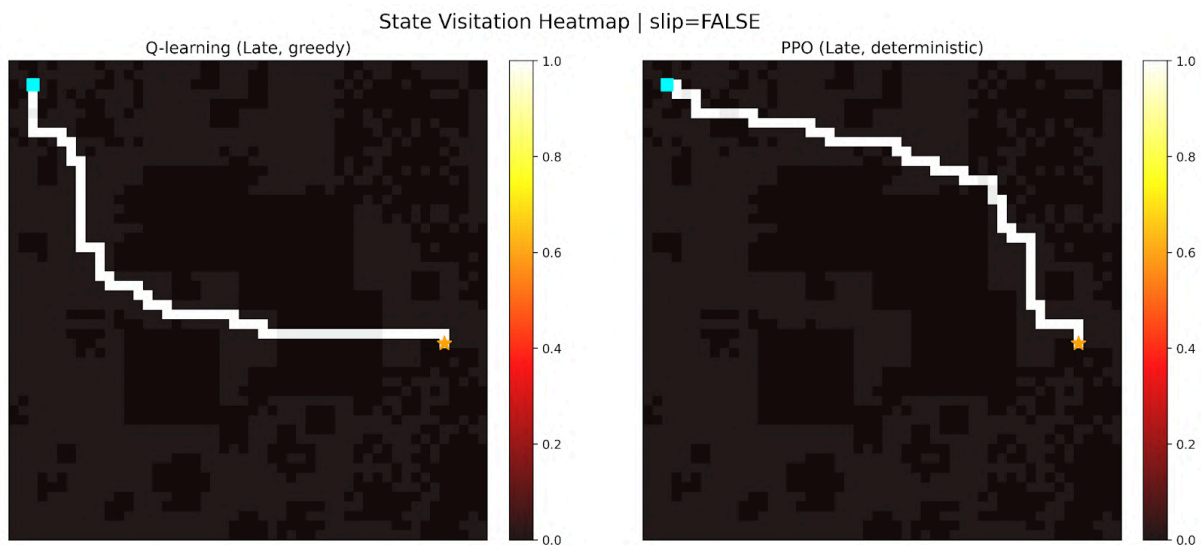
453 Table 2: summarizes the quantitative performance metrics for BFS, Q-learning, and PPO  
 454 under slip winter conditions, using the same evaluation metrics as in the deterministic  
 455 case, Standard Deviation (SD), and averaged across 10 independent random seeds with  
 456 95% confidence intervals.

457

458

459 **4.3 Spatial State Visitation Heatmaps**

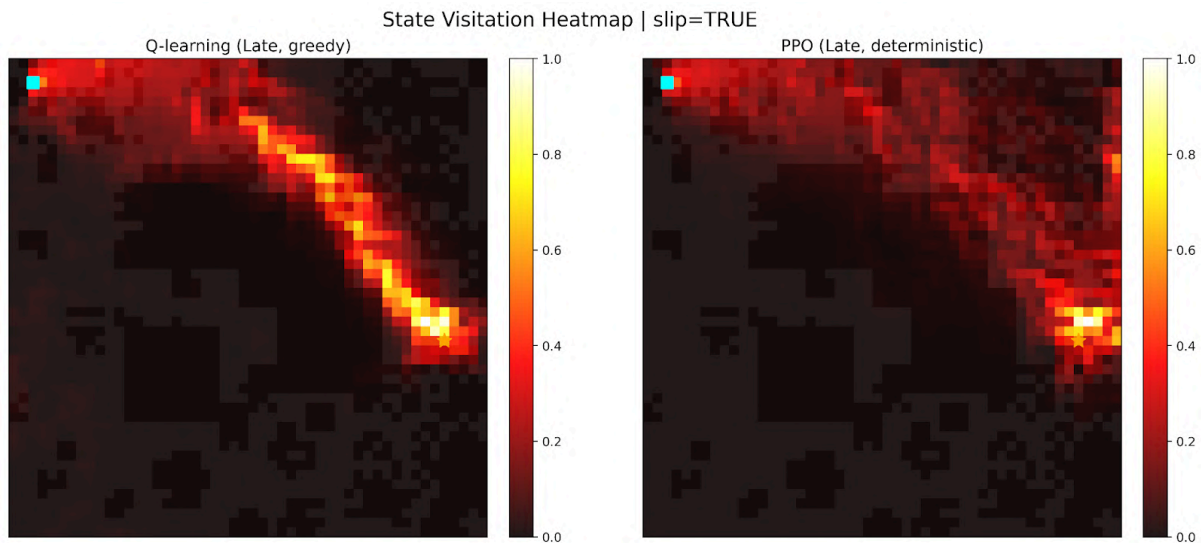
460 Spatial state visitation heatmaps offer a grid-based way to see how often an agent  
 461 passes through each cell during navigation. In our 50×50 winter grid, each cell  
 462 corresponds to a state, and the color intensity reflects visitation frequency across  
 463 episodes. Darker areas indicate cells that are visited more frequently, while lighter or  
 464 white cells indicate rarely or never visited locations. Such visualizations give a clear  
 465 picture of the agent's overall path distribution and behavior patterns. We generated  
 466 these heatmaps using aggregated visitation counts normalized to the range between 0  
 467 and 1.



469 Figure 10: displays the visitation heatmaps for Q-learning and PPO under deterministic  
470 (non-slip) winter conditions. The results are aggregated from 10 evaluation runs  
471 consisting of 200 episodes per run. Both agents concentrate most visits  
472 along a narrow corridor connecting the start to the goal, with  
473 near-maximum intensity ( $\approx 1.0$ ) along the primary route and very low  
474 visitation elsewhere.

475

476



477

478 Figure 11: shows the visitation intensity for each grid cell in the slip case. Compared with  
479 the non-slip condition, visitation is more widely distributed across the grid rather than  
480 remaining within a narrow corridor. The heatmap aggregates data from 10 evaluation  
481 runs consisting of 200 episodes per run.

482

#### 483 4.4 Statistical Analysis and Learning Curves - Slip Condition

484

##### 485 4.4.1 Statistical Significance (Welch's t-test)

486 To assess statistical significance, a Welch's t-test was conducted comparing the  
487 cumulative reward of Q-learning and PPO under slip conditions across 10 seeds. The  
488 results confirmed that Q-learning significantly outperformed PPO in cumulative  
489 rewards ( $t = 14.28$ ,  $p < 0.001$ ), indicating that the performance gap is highly unlikely to be  
490 due to random variation. Q-learning achieved a mean reward of  $311.82 \pm 4.83$  compared  
491 to PPO's  $258.69 \pm 10.73$ . The higher standard deviation of PPO ( $SD = 10.73$ ) compared to  
492 Q-learning ( $SD = 4.83$ ) further indicates greater training instability under stochastic  
493 conditions.

494

495 The standard deviation gap between PPO in both conditions suggests that PPO is more  
496 sensitive to stochastic signals under slip condition, where high-probability slip results  
497 in noisy advantage estimation, leading to higher variance across training runs ( $SD =$   
498  $10.73$ ). In contrast, under deterministic conditions PPO exhibits a much lower variance

499 (SD = 0.67), confirming that the instability is determined by the stochastic environment,  
 500 not the algorithm itself.

501

502

503 === Final Comparison Table ===

504

Metric	t-statistic	p-value	Q mean $\pm$ SD	PPO mean $\pm$ SD
<b>Cumulative Reward</b>	14.28	< 0.001	311.82 $\pm$ 4.83	258.69 $\pm$ 10.73
<b>Success Rate</b>	n/a	n/a	100% $\pm$ 0%	100% $\pm$ 0%

505

506 Table 3: Welch's t-test results comparing Q-learning and PPO Curriculum cumulative  
 507 reward under slip conditions across 10 seeds. n/a indicates that the test was not  
 508 applicable due to zero variance in success rate for both methods.

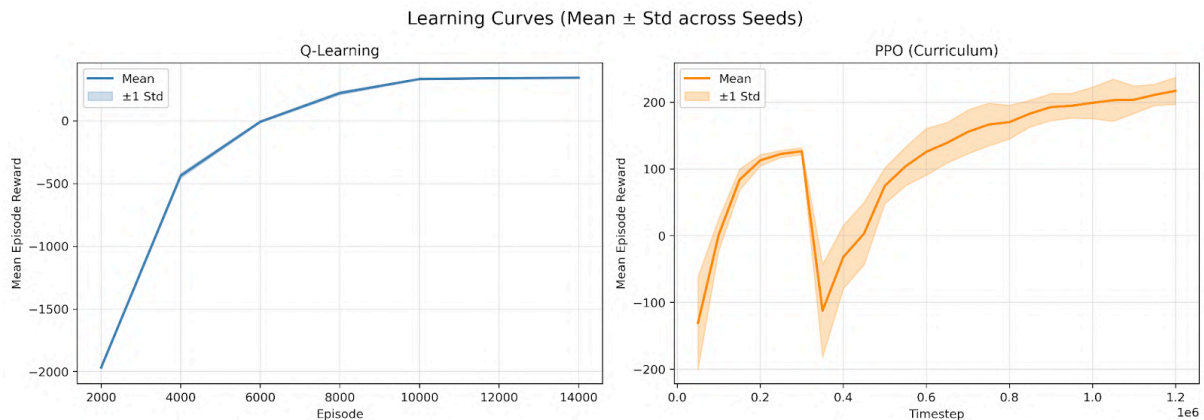
509

#### 510 4.4.2 Learning Curves

511 Learning curves illustrate the performance of Q-learning and PPO evolves over the  
 512 course of training under stochastic conditions, and provide insight into convergence  
 513 behaviour, training stability, and the impact of curriculum learning on PPO.

514

515



516

517

518 Figure 13: Learning curves (Mean  $\pm$  SD across 10 seeds) under slip (stochastic)  
 519 conditions. Left: Q-learning mean episode reward vs. training episode (15,000 episodes)  
 520 showing smooth, steady and consistent convergence. Right: PPO Mean episode reward  
 521 vs. training timestep (1,200,000 steps), where the pronounced drop around timestep  
 522 300,000 corresponds to the Phase 1 to Phase 2 curriculum transition, when  
 523 energy-aware snow penalties are introduced and the agent must re-adapt its policy.  
 524 After the transition is complete, PPO gradually recovers and converges, but its variance  
 525 band is significantly wider compared to Q-learning.

526

527

528

#### 529 **4.5 Limitations & Uncertainty**

530 Several sources of uncertainty and limitations appeared in this study. Both Q-learning  
531 and PPO showed noticeable performance differences across training runs, which was  
532 expected given the stochastic nature of the environment and the learning algorithms.  
533 To account for this variability, all experiments were repeated over 10 independent  
534 random seeds, and results are reported as the mean across runs.

535 Under slip conditions, the same policy could produce quite different routes from  
536 episode to episode because actions did not always lead to the expected outcome. In  
537 addition, stochastic exploration strategies and random initialization of value functions  
538 (or policy networks) exposed agents to slightly different states, transitions, and rewards  
539 across runs. This variability made it harder to get perfectly consistent results.

540 An unexpected issue was that the BFS baseline failed to reach the goal within the  
541 1,000-step limit under slip conditions, resulting in a number of unsuccessful evaluation  
542 episodes.

543

544 A full numerical breakdown of evaluation metrics is provided in Appendix B.

545

546

---

547

### 548 **5. Discussion**

#### 549 **5.1 Restatement of Hypothesis and Summary of Findings**

550 The study hypothesized that reinforcement learning-based agents, particularly Proximal  
551 Policy Optimization (PPO), would achieve more energy-efficient winter navigation paths  
552 than a traditional shortest-path baseline and a value-based Q-learning agent under  
553 both deterministic and slip winter conditions. The averaged results partially supported  
554 the hypothesis. Under deterministic (non-slip) conditions, PPO achieved higher  
555 cumulative reward, and fewer snow visits, outperforming Q-learning, which supports  
556 the hypothesis. However under stochastic (slip) conditions, Q-learning outperformed  
557 PPO in terms of cumulative reward or snow-cell avoidance, which did not support the  
558 hypothesis. Both reinforcement learning methods perform better results than  
559 traditional BFS across these environments.

560

#### 561 **5.2 Interpretation of Q-Learning and PPO Performance**

562 The results showed that algorithm performance varied depending on the environmental  
563 condition. Under deterministic (non-slip) conditions, PPO achieved higher cumulative  
564 reward and fewer snow cell visits than Q-learning, suggesting that PPO's policy gradient  
565 approach was able to learn a more energy-efficient route in a stable environment.  
566 Under stochastic (slip) conditions, Q-learning outperformed PPO in terms of cumulative  
567 reward and snow-cell avoidance. One possible explanation was that a discrete and clear  
568 grid-world environment and relatively small environment would favor more for  
569 Q-learning, as it can perform exact value iteration over the finite MDP (Markov Decision

570 Process), directly converging to the true optimal Q-values for each state without  
571 approximation error. Therefore allowing Q-learning to efficiently estimate optimal state  
572 and gives lower variance than policy gradient method (PPO), since it updates based on  
573 individual state-action pairs rather than entire trajectories [7,8]. In contrast, PPO relied  
574 on function approximation and stochastic policy updates, which may have required  
575 more training data or longer training time to converge to an equally optimal policy  
576 under slip conditions. Additionally, its stochastic policy may have introduced suboptimal  
577 choices that were not effective, whose gradient estimates already carry high variance  
578 from episode training. Under slip conditions, environmental stochasticity further interrupts  
579 reward signals, making them sparse and noisy, which destabilizes gradient estimates and  
580 requires significantly more training samples. Therefore reducing its total rewards  
581 relative to Q-learning under stochastic conditions [8].

582

### 583 **5.3 Effects of Slip (Stochastic) Winter Conditions**

584 Under slip (stochastic) conditions, all agents showed broader trajectory dispersion and  
585 increased visits to previously visited states because of slipping, as reflected in the  
586 spatial state visitation heatmaps. This behavior is consistent with probabilistic transition  
587 dynamics, in which the same action could lead to different movement outcomes across  
588 episodes, such as deviating left in one run and right in another. When the grid is  
589 slippery, agents don't follow one stable path anymore. As a result, the routes spread out,  
590 and their performance becomes less consistent across episodes. These observations are  
591 consistent with prior navigation studies showing that stochastic environments increase  
592 policy variance and reduce convergence stability in reinforcement learning tasks [4,8].

593

### 594 **5.4 Interpretation of BFS Baseline Behavior**

595 The Traditional baseline trajectories did not take account for winter grid costs or  
596 stochastic transitions, it focused on the fastest way to the destination. In non-slip  
597 conditions this approach naturally produced the optimal route. However, under slip  
598 conditions, BFS exceeded the maximum step limit in multiple evaluation episodes,  
599 reflecting its inability to adapt its policy in response to transition winter grid costs or  
600 stochastic transitions.

601

602 Because BFS always assumed deterministic movement, each slip caused deviations from  
603 the planned path. These deviations often led to inefficient loops and longer paths.  
604 Unlike reinforcement learning methods, BFS did not have reward feedback or policy  
605 updates, which limited its ability to adjust navigation behavior under changing  
606 environmental dynamics. Therefore, making it the least effective method.

607

### 608 **5.5 Hyperparameter Sensitivity Analysis**

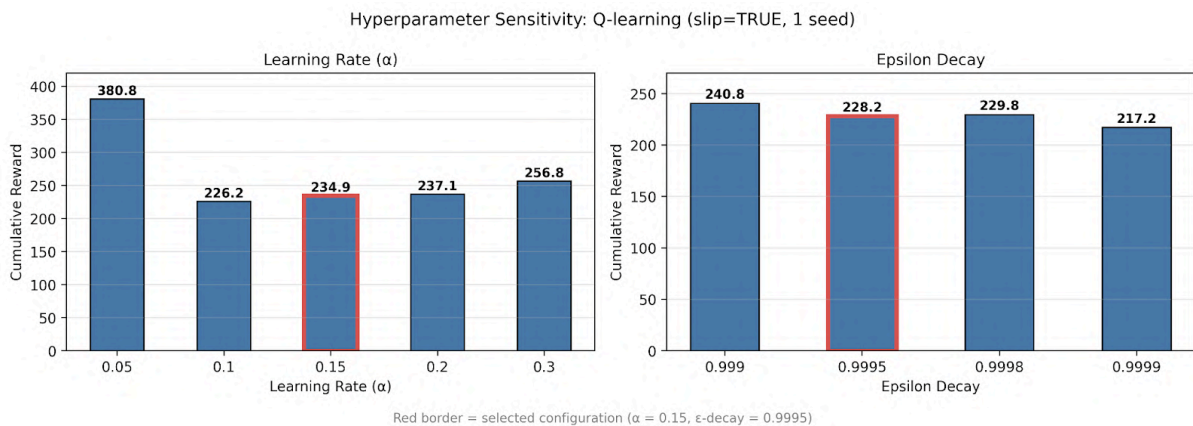
609 To assess the robustness of the reported conclusions to hyperparameter choice, a  
610 sensitivity analysis was conducted to independently vary each hyperparameter of  
611 Q-learning and PPO under slip conditions using a single representative seed. For  
612 Q-learning, the learning rate (alpha) and epsilon decay were evaluated to identify which

613 hyperparameter values lead to greater stability and performance. For PPO, the clip  
 614 range and learning rate were varied to assess how different hyperparameters affect  
 615 training stability and sensitivity to stochastic conditions.

616

617

618



619

620 Figure 14: Comparison of different learning Rate (alpha) values and epsilon decay values . The  
 621 highlighted values represent the selected hyperparameter configuration, chosen based on their  
 622 balance of performance and training stability. ( $\alpha = 0.15$ ,  $\epsilon$ -decay = 0.9995)

623

624

625 === Q-learning hyperparameter values Final Comparison Table ===

626

Parameter	Value	Success rate (%)	Cumulative reward	Note
Learning rate (alpha)	0.05	98.00	380.80	
Learning rate (alpha)	0.10	100.00	226.20	
Learning rate (alpha)	0.15	100.00	234.90	Selected configuration
Learning rate (alpha)	0.20	100.00	237.10	
Learning rate (alpha)	0.30	100.00	256.80	
Epsilon decay	0.9990	100.00	240.80	
Epsilon decay	0.9995	100.00	228.20	Selected configuration
Epsilon decay	0.9998	100.00	229.80	
Epsilon decay	0.9999	100.00	217.20	

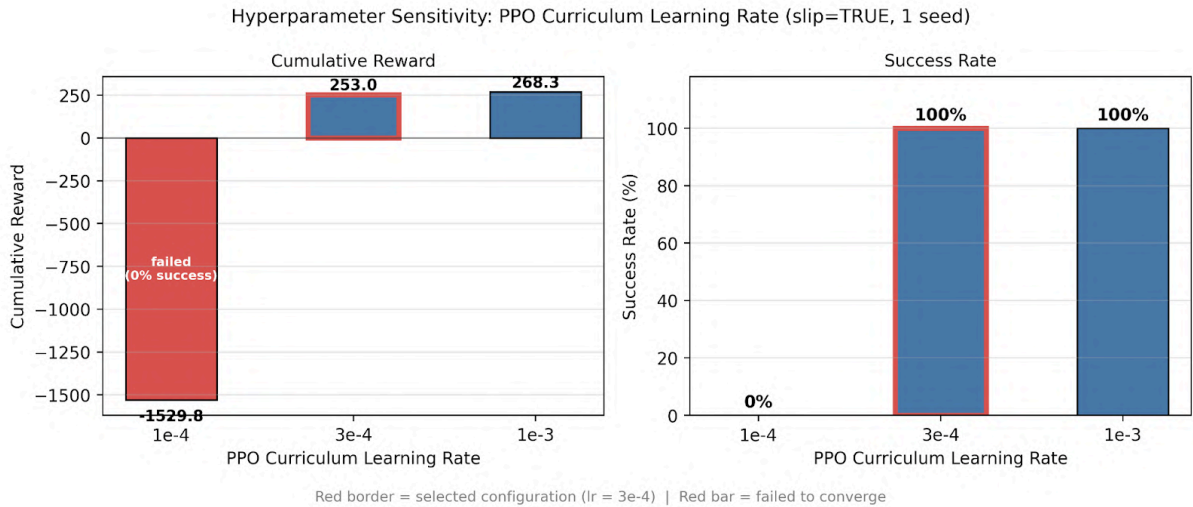
627

628 Table 4: presents a sensitivity analysis of Q-learning's key hyperparameters. Cumulative rewards  
629 are reported to assess overall performance.

630

631

632



633

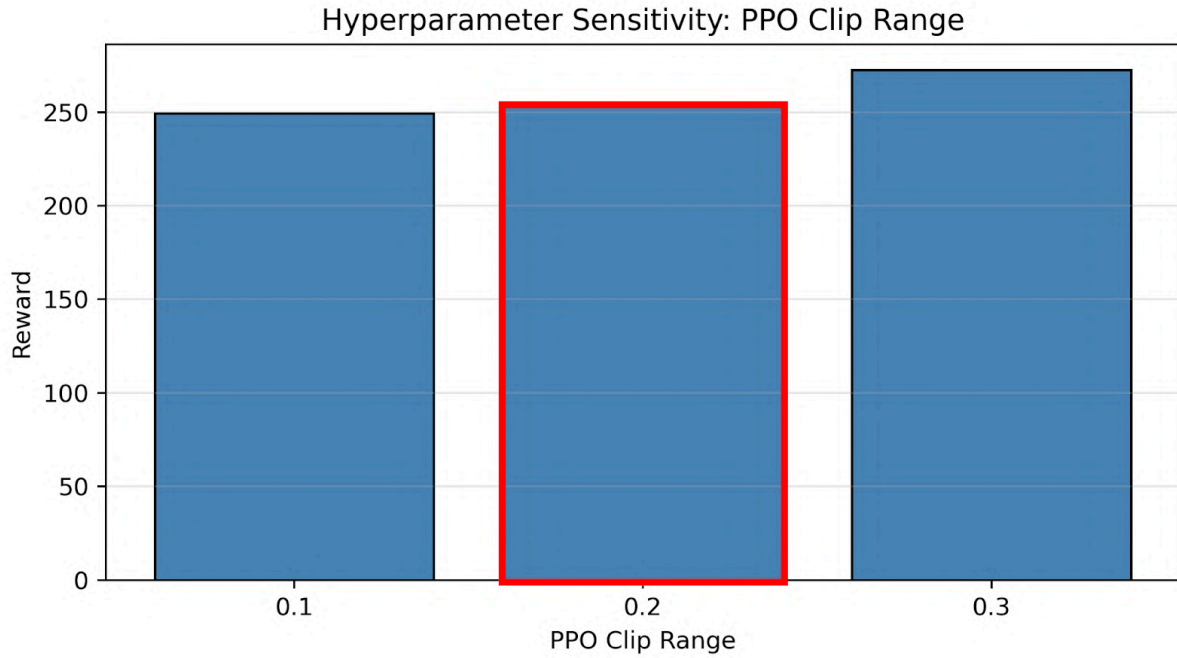
634 Figure 15: Comparison of different values of learning rate and the  
635 highlighted values represents the selected hyperparameter configuration,  
636 chosen based on their balance of performance and training stability (  
637 Learning rate =  $3 \times 10^{-4}$ ). lr =  $1 \times 10^{-4}$  failed to converge (0% success  
638 rate, reward = -1529.8), a very small learning rate leads to insufficient  
639 policy updates, preventing convergence within the training horizon.

640

641

642

643



644

645 Figure 16: Comparison of different values of PPO clip range and the highlighted values  
 646 represents the selected hyperparameter configuration, chosen based on their balance of  
 647 performance and training stability. ( Clip range = 0.2)

648

649 ===PPO hyperparameter values Final Comparison Table ===

650

Parameter	Value	Success rate (%)	Cumulative reward	Note
Clip range	0.1	100.00	249.40	
Clip range	0.2	100.00	253.00	Selected configuration
Clip range	0.3	100.00	272.50	
Learning rate	1e-04	0.00	-1529.80	
Learning rate	3e-04	100.00	253.00	Selected configuration
Learning rate	1e-03	100.00	268.30	

651

652 Table 5: presents a sensitivity analysis of PPO's key hyperparameters. Cumulative rewards and  
 653 success rates are reported to assess overall performance.

654

655

656 Across all tested configurations, the rank ordering of methods remained consistent,  
 657 confirming that the reported conclusions are robust to moderate hyperparameter  
 658 variation.

659

660 **5.6 Limitations and Future Directions**

661 Several limitations should be noted when interpreting these findings. All experiments  
662 were carried out in a simulated 50×50 grid environment with a relatively small and  
663 simple state space, which differs significantly from real-world scenarios. This setup  
664 likely favored value-based methods like Q-learning, since it could learn precise  
665 state-action values for every cell. In contrast, PPO depends on neural network function  
666 approximation and stochastic policy updates, which may need more training steps or  
667 data to perform well in such discrete, small-scale settings.

668

669 To improve statistical reliability, all experiments were run independently over 10  
670 random seeds and results were averaged across runs. Noticeable performance  
671 variability was observed across training runs, mainly due to the stochastic environment,  
672 random action outcomes, and probabilistic transitions under slip conditions.

673 The simulated environment simplified real-world winter driving conditions and did not  
674 capture vehicle dynamics, sensor noise, road-surface variation, or realistic  
675 energy-consumption models, these are factors that might lead to changes. As a result,  
676 these findings to real-world winter navigation scenarios were limited [12].

677 Future research could evaluate reinforcement learning agents in larger scale,  
678 continuous, or more realistic winter navigation environments with higher-dimensional  
679 state spaces and more complex conditions. Then it may be better to reflect real-world  
680 uncertainty and could allow policy-gradient methods such as PPO to fully function their  
681 theoretical advantages in handling stochastic and high-dimensional control problems.

682

683

---

684

## 685 **6. Conclusion**

686 This study compared a traditional BFS baseline with two reinforcement learning  
687 methods—Q-learning and Proximal Policy Optimization (PPO)—for energy-efficient  
688 navigation in a 50×50 winter grid under both deterministic (non-slip) and stochastic  
689 (slip) conditions. The averaged results across 10 independent random seeds partially  
690 supported the original hypothesis. Under deterministic (non-slip) conditions, PPO  
691 achieved the highest cumulative reward and fewest snow cell visits, outperforming  
692 Q-learning. However, under stochastic (slip) conditions, Q-learning outperformed PPO  
693 in cumulative reward and snow-cell avoidance. Although both reinforcement learning  
694 methods clearly outperformed the BFS baseline, BFS struggled under slip conditions  
695 due to its inability to adapt to stochastic transitions.

696

697 These findings show that whether value-based methods or policy-gradient methods  
698 perform better really depends on how unpredictable the environment is. In our small,  
699 structured discrete grid, PPO gained an advantage from its stable gradients and  
700 curriculum learning when everything was deterministic. On the other hand,  
701 Q-learning's simple tabular updates turned out to be more robust when the roads

702 became slippery and actions sometimes failed. More generally, the results emphasize  
703 that the choice of reinforcement learning method should be guided by the structure and  
704 constraints of the task, rather than by theoretical advantages alone. Future work could  
705 explore whether PPO's advantages extend to more complex environments with  
706 continuous states or larger state spaces, where its policy-gradient approach may better  
707 demonstrate its theoretical strengths.

708

709

---

710

## 711 Reference

712

- 713 [1]Mao, R., Xu, W., Qian, Y., Li, X., Li, Y., Li, G., & Zhang, H. (2025).  
714 Understanding the Determinants of Electric Vehicle Range: A Multi-  
715 Dimensional Survey. *Sustainability*, 17(10), 4259.  
716 <https://www.mdpi.com/2071-1050/17/10/4259>
- 717 [2]Carlson, A., & Vieira, T. (2021). *The effect of water and snow on the*  
718 *road surface on rolling resistance* (VTI Report 971A). Swedish National  
719 Road and Transport Research  
720 Institute.[https://www.researchgate.net/publication/350690120\\_The\\_effect\\_of\\_water](https://www.researchgate.net/publication/350690120_The_effect_of_water_and_snow_on_the_road_surface_on_rolling_resistance)  
721 [and\\_snow\\_on\\_the\\_road\\_surface\\_on\\_rolling\\_resistance](https://www.researchgate.net/publication/350690120_The_effect_of_water_and_snow_on_the_road_surface_on_rolling_resistance)
- 722 [3]Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An  
723 Introduction.  
724 [https://web.stanford.edu/class/psych209/Readings/](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)  
725 [SuttonBartoIPRLBook2ndEd.pdf](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)
- 726 [4]Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino,  
727 A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., &  
728 Hadsell, R. (2017). *Learning to Navigate in Complex Environments* (No.  
729 arXiv:1611.03673). arXiv. <https://doi.org/10.48550/arXiv.1611.03673>
- 730 [5]Tan, C. (2025). Comparative Study of Reinforcement Learning  
731 Performance Based on PPO and DQN Algorithms. *Applied and*  
732 *Computational Engineering*, 175(1), 30–36.  
733 <https://doi.org/10.54254/2755-2721/2025.AST24879>
- 734 [6]Warnakulasuriya, D. A., Plosila, J., & Haghbayan, H. (2025). Energy-Efficient Path  
735 Planning in Uneven Terrains Using Adaptive Reinforcement Learning. *IEEE Conference*  
736 *Publication*. <https://ieeexplore.ieee.org/document/11093435>
- 737 [7]Watkins, C.J.C.H., & Dayan, P.(1992). Q-learning. *Mach Learn* 8, 279–292.  
738 <https://doi.org/10.1007/BF00992698>
- 739 [8]Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O.  
740 (2017). *Proximal Policy Optimization Algorithms* (No.arXiv:1707.06347). arXiv.  
741 <https://doi.org/10.48550/arXiv.1707.06347>

742 [9]Pendyala, A., Atamna, A., & Glasmachers, T. (2024). Solving a Real-World Optimization  
743 Problem Using Proximal Policy Optimization with Curriculum Learning and Reward  
744 Engineering. arXiv:2404.02577.<https://arxiv.org/abs/2404.02577>  
745 [10]Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and  
746 Reinforcement Learning. *Neuron*, 36(2), 285–298.  
747 [https://doi.org/10.1016/S0896-6273\(02\)00963-7](https://doi.org/10.1016/S0896-6273(02)00963-7)  
748 [11]Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009).  
749 Curriculum learning. *Proceedings of the 26th Annual International*  
750 *Conference on Machine Learning*, 41–48. <https://doi.org/10.1145/1553374.1553380>  
751 [12]Chukwurah, N., Adebayo, A. S., Ajayi, O. O., & Anfo Pub. (2024).  
752 Sim-to-Real Transfer in Robotics: Addressing the Gap between  
753 Simulation and Real- World Performance. *International Journal for Multidisciplinary*  
754 *Research (IJFMR)*, 05(01), 33–39. <https://doi.org/10.54660/.IJFMR.2024.5.1.33-39>

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

## 773 **Appendices**

### 774 **Appendix A Experimental Configuration and Algorithmic Details:**

775

#### 776 **A.1 Environment Configuration**

777 All experiments were conducted in a custom 50×50 winter grid environment (2,500  
778 discrete states).

- 779 • Start position: (47, 2)

- 780 • Goal position: (20, 45)
- 781 • Maximum steps per episode: 1000
- 782 • Action space: {up, down, left, right}

783 Two transition settings were evaluated:

784 Deterministic (slip = FALSE)

785 The intended action is executed exactly.

786 Stochastic (slip = TRUE)

787 With probability  $\frac{1}{3}$ , the intended action is executed.

788 With probability  $\frac{2}{3}$ , the agent slips to a random adjacent direction.

789 All algorithms were evaluated on the same fixed grid layout to ensure fairness.

790

## 791 A.2 Reward Function

792 The reward function models energy-aware winter navigation.

793 At each time step:

- 794 • Step penalty: -1.5
- 795 • Snow penalties:
  - 796 ○ Near snow: -0.2
  - 797 ○ Edge snow: -0.5
  - 798 ○ Core snow: -2.0
- 799 • Goal reward: +700

800 The cumulative episode reward is:

$$801 R = \sum_{t=1}^T (-1.5 - C_{snow}(s_t)) + 700 \cdot 1_{goal\ reached}$$

802

803 where  $C_{snow}(s_t)$  denotes the terrain penalty and

804  $1_{goal\ reached}$  indicates successful termination.

## 805 A.3 Q-Learning Configuration

806 Tabular Q-learning was implemented with the following hyperparameters:

- 807 • Learning rate  $\alpha=0.15$
- 808 • Discount factor  $\gamma=0.99$
- 809 • Exploration strategy:  $\epsilon$ -greedy
- 810 • Initial  $\epsilon = 0.60$
- 811 • Minimum  $\epsilon = 0.05$
- 812 • Exponential decay per episode = 0.9995
- 813 • Training episodes = 15,000
- 814 • Max steps per episode = 1000

815 The Q-update rule is:

$$816 \quad Q(s, a) \leftarrow Q(s, a) + \alpha [R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

817 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) was used.

818

#### 819 A.4 Proximal Policy Optimization (PPO) Curriculum Configuration

820 PPO was implemented as a stochastic policy-gradient method.

821 Total training timesteps: 1,200,000

- 822 • Phase 1 (navigation-focused reward): 300,000 timesteps
- 823 • Phase 2 (energy-aware reward): 900,000 timesteps

824

825 Evaluation was conducted using deterministic action selection.

826

827 PPO Objective Function

828 PPO optimizes the clipped surrogate objective:

$$829 \quad L^{CLIP}(\theta) = E_t [\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)]$$

830 where

$$831 \quad r_t(\theta) = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{old}}(a_t | s_t)}$$

832 and  $\hat{A}_t$  is the generalized advantage estimate.

833 The clipping mechanism constrains policy updates to maintain stability.

834

### 835 PPO Hyperparameters

- 836 • Discount factor  $\gamma=0.99$
- 837 • GAE parameter  $\lambda=0.95$
- 838 • Entropy regularization enabled
  
- 839 • Learning rate:  $3 \times 10^{-4}$
  
- 840 • Clip range: 0.2
  
- 841 • Rollout buffer: 2,048 steps
  
- 842 • Batch size: 256
  
- 843 • Optimization epochs: 10
  
- 844 • Neural network architecture: two hidden layers (256 units each, ReLU activation)

845

### 846 A.5 Evaluation Protocol

847 For each condition (slip FALSE / TRUE):

- 848 • One trained model per algorithm
- 849 • Maximum evaluation length: 1000 steps
- 850 • Number of evaluation episodes: 200
- 851 • Number of seeds: 10
- 852 • Heatmap runs: 10
- 853 • Metrics recorded:
  - 854 ○ Total cumulative reward
  - 855 ○ Number of steps
  - 856 ○ Snow cell visits
  - 857 ○ Success rate (%)

858 State visitation heatmaps were generated by aggregating visitation frequencies across  
859 evaluation runs of 200 episodes each.

860

---

## 861 Appendix B Detailed Quantitative Results:

### 862 B.1 Deterministic (slip = FALSE)

Method	Reward (mean±SD)	95%CI	Steps	Snow Visits	Success
<b>BFS Shortest</b>	575.30 ± 0.00	[575.30, 575.30]	70.00 ± 0.00	31.00 ± 0.00	100.0%
<b>Q-Learning</b>	582.01 ± 6.97	[577.02, 587.00]	70.00 ± 0.00	13.10 ± 3.05	100.0%
<b>PPO Curriculum</b>	588.73 ± 0.67	[588.25, 589.21]	70.00 ± 0.00	7.60 ± 0.97	100.0%

863

864 All methods reached the goal within 70 steps.

865 PPO achieved the highest cumulative reward and fewest snow cell visits.

866

### 867 B.2 Stochastic (slip = TRUE)

Method	Reward (mean±SD)	95% CI	Steps	Snow Visits	Success
<b>BFS Shortest</b>	-1668.41 ± 113.42	[-1749.54, -1587.28]	1000.00 ± 0.00	246.80 ± 76.09	0.0%
<b>Q-Learning</b>	311.82 ± 4.83	[308.37, 315.27]	26.40 ± 3.92	66.94 ± 3.87	100.0%
<b>PPO Curriculum</b>	258.69 ± 10.73	[251.02, 266.37]	241.18 ± 4.11	91.88 ± 5.24	100.0%

868

### 869 Under stochastic dynamics:

- 870 • BFS fails due to inability to adapt.
- 871 • Q-learning demonstrates greater robustness.
- 872 • PPO maintains success but exhibits higher trajectory dispersion.

873

### 874 B.3 State Visitation Analysis

875 State visitation frequency was computed as:

876 
$$V(s) = \frac{N(s)}{\max_s N(s)}$$

877 where  $N(s)$  is the number of visits to state  $s$ .

- 878 • Under deterministic conditions, visitation concentrates along a narrow corridor.
- 879 • Under stochastic conditions, visitation becomes more dispersed.
- 880 • Q-learning exhibits more focused routing than PPO under slip conditions.

# 1 Energy-Efficient Path Planning in Winter Conditions: A Comparative 2 Study of Traditional Baseline, Q-learning, and Proximal Policy 3 Optimization in a Grid World Environment

4  
5

## 6 1.Abstract

7 Recent investigation in winter route planning has become increasingly important due to  
8 increased resistance and reduced efficiency on winter roads. These conditions  
9 significantly increase energy consumption and introduce uncertainty, raising the risk of  
10 failing to reach the destination. To address the challenge, Reinforcement Learning (RL)  
11 is a type of machine learning where an agent learns to make decisions by interacting  
12 with an environment, receiving rewards or penalties for its actions to maximize  
13 cumulative rewards. For simulation, we benchmark a standard shortest-path algorithm  
14 (BFS) against two reinforcement learning methods: Q-learning and Proximal Policy  
15 Optimization (PPO). All approaches are tested on identical grid configurations, with the  
16 same starting points, goals, and reward functions. We assess their performance based on  
17 cumulative reward, snowy cell visits (as a proxy for energy cost), trajectory characteristics,  
18 and success rate. It is hypothesized that PPO algorithms will result in lower energy  
19 consumption than Q-learning and traditional baseline under winter conditions. The  
20 results show that while BFS consistently finds the shortest paths, it fails to consider  
21 energy costs or environmental uncertainty. RL agents, by comparison, adapt more  
22 efficiently to winter conditions. Under deterministic (non-slip) conditions, PPO  
23 achieved higher cumulative reward and fewer snow cell visits than Q-learning, partially  
24 supporting the hypothesis. However, under stochastic (slip) conditions, Q-learning  
25 outperformed PPO in cumulative reward and snow-cell avoidance. These results  
26 suggest that Q-learning is better suited for stochastic winter navigation in grid worlds,  
27 while PPO may perform better in more deterministic environments with continuous  
28 states or decisions.

29

30 Key words: Energy-Efficient, Traditional Baseline, Q-learning, Proximal Policy  
31 Optimization, grid world, reinforcement learning.

32

33

34

---

## 35 2.Introduction

36 As demand in **Electric Vehicles (EV)** become increasingly widespread, energy  
37 efficiency and route optimization have emerged as critical challenges, particularly in  
38 winter when resistance increases. **Sub-zero temperatures, combined with snow and ice**  
39 **accumulation on road surfaces, significantly increase energy consumption, resulting in**  
40 **reduced driving range and greater unpredictability in trip planning.** ~~Freezing temps, plus~~  
41 ~~snow and ice covering the roads, push energy consumption way up, which means~~  
42 ~~shorter range and trips that feel less predictable.~~ A 2025 study shows that an estimated

43 50 % of EV driving range can be reduced in cold climates, including snow and ice  
44 covering terrain, highlighting the significant impact of environmental conditions on  
45 energy consumption [1]. Furthermore, when snow or water remains on the road surface,  
46 vehicle tires must continuously displace through ice as they roll and move, hence  
47 forcing the vehicle to draw more power [2]. On top of that, water cools tires more  
48 effectively than air alone, altering their mechanical properties and further pushing  
49 rolling resistance higher. Previous studies have shown that rolling resistance can rise by  
50 approximately 30% to 40% under wet or snowy conditions [2]. Ultimately, these factors  
51 make winter driving a major concern for Electric Vehicle (EV) efficiency and reinforce  
52 the need for energy-aware routing strategies.

53

54 To approach these challenges, we have explored a few computational methods to allow  
55 agents to learn effective navigation strategies aimed at minimizing energy consumption.  
56 In this study, a 50x50 grid is used as an abstraction routing map that circles key  
57 structures of Canadian snow distribution, rather than exact geographical terrain.  
58 Grid-based environments are commonly used in reinforcement learning studies, as they  
59 simplify complex continuous real-world environments into discrete states [35]. This  
60 abstraction is particularly well-suited for this study, as it enables environmental factors  
61 such as snow coverage and surface slip conditions to be incorporated directly into the  
62 cell-level cost and reward structure, supporting controlled and reproducible  
63 comparison of different routing methods.

64

65 Despite growing interest in reinforcement learning for navigation, existing studies have  
66 largely overlooked the impact of winter environmental conditions, such as snow  
67 coverage and surface slippiness, on energy-aware path planning. Prior work has  
68 primarily focused on general navigation without taking account of explicitly modeling  
69 weather-dependent energy costs [48]. To address this gap, this study integrates snow  
70 coverage, stochastic slip condition, and energy costs into a comparative framework,  
71 evaluating Q-learning, PPO, and a traditional **b**Baseline under realistic winter routing  
72 scenarios. This study is one of the first to directly compare value-based and  
73 policy-gradient reinforcement learning methods in an energy-aware winter navigation  
74 setting.

75

76 Prior work on reinforcement learning for navigation can be categorized into three  
77 directions. First, Grid world benchmarks commonly adopt tabular Q-learning and deep  
78 reinforcement learning variants as standard baselines for evaluating discrete navigation  
79 tasks [3,5], but focus primarily on task completion rather than energy awareness and  
80 energy sensitive routing. Second, cost-aware navigation research has examined energy  
81 minimisation in uneven terrains using reinforcement learning [6], yet rarely  
82 incorporated weather dependent components, such as snow coverage or stochastic  
83 traction loss. Third, existing comparisons between value-based and policy-gradient  
84 methods [7,8], including curriculum based PPO applications [9] are conducted primarily  
85 under fixed and stable conditions, without examining the possibility of a particular

86 reinforcement learning method remaining dominant to another when action outcomes  
87 become probabilistic, as in stochastic environments such as winter slip conditions. The  
88 work in this study addresses all three gaps simultaneously by implementing an  
89 energy-aware reward structure into a reproducible 50×50 grid-world benchmark,  
90 embedding snow-density penalties and a 2/3-slip probability to simulate realistic  
91 winter routing conditions, and directly comparing Q-learning, PPO, and a traditional  
92 baseline under both deterministic and stochastic environments.

93

94 More specifically, the research will include a traditional baseline algorithm, and two  
95 types of reinforcement learning, which are Q-learning and PPO algorithm. The  
96 traditional baseline routing computes the shortest path based on static cost metrics and  
97 follows the predefined route without adaptation or learning (basic routing). Q-learning,  
98 a value-based reinforcement learning method, learns through incremental dynamic  
99 programming processes with computational requirements and it is well suited for  
100 agents to improve and refine action values and achieve effective performance in  
101 controlled Markovian domains [73]. In contrast, PPO is a policy-gradient algorithm that  
102 primarily optimizes a stochastic policy and achieves greater training stability through  
103 constrained policy updates with a clipped surrogate objective function [84].

104

105 This comparison evaluates the effectiveness of the proposed routing methods in  
106 identifying energy-efficient paths. Performance is evaluated through energy related  
107 costs, overall routing efficiency, and observed navigation behavior under both  
108 deterministic (non-slip) and stochastic (slip) settings. Because PPO balances stable  
109 policy updates with adaptive learning in uncertain environments, it is hypothesized to  
110 be the most effective method implemented for winter routing.

111

112 The following section goes into the methodology and experimental setup in more detail.

113

114

---

115

## 116 **3.Methods**

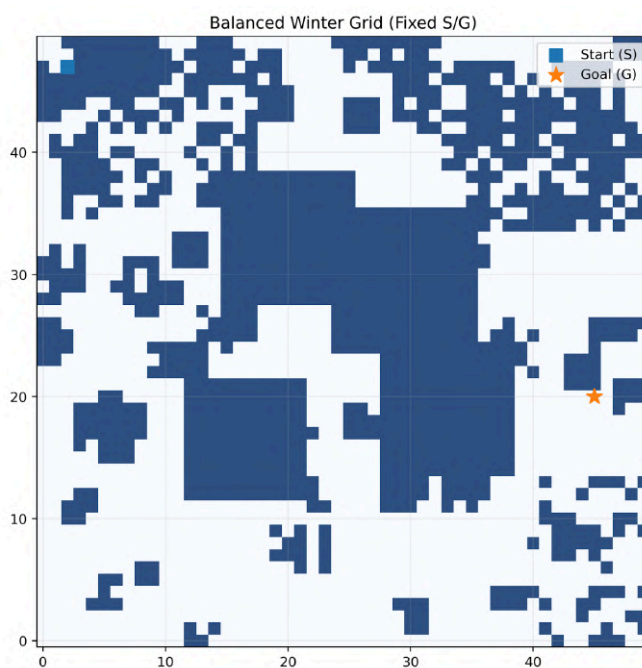
### 117 **3.1 Environment Design**

118 A custom winter navigation grid environment was designed to evaluate each  
119 Reinforcement Learning (RL) agent under stochastic and cost-sensitive conditions. The  
120 environment consists of a 50 x 50 grid world containing 2500 cells, including both  
121 normal terrain and snow-covered terrain. The simulation contains a fixed start (47, 2)  
122 and a fixed goal (20, 45) point, and are located in the top left corner and middle right  
123 respectively. At each time step, the agent can take either one of the four directions, up,  
124 down, left, and right; each step on a non-snow grid has an energy cost of -1.5 per step.  
125 Snow is modeled as spatially correlated regions with graded intensity, capturing the typical  
126 uneven accumulation patterns in Canadian winters.

127 When snow is modeled as a binary state, abrupt reward discontinuities make PPO's  
128 advantage estimates unstable, resulting in increased gradient variance. By contrast, a  
129 continuous representation of snow thickness provides smoother cost transitions, which  
130 help reduce gradient variance and stabilize learning, making it more efficient. These  
131 snow types are classified as "Near snow", "Edge snow", and "Core snow".

132 To evaluate robustness under different winter conditions, two transition settings were  
133 considered using the same grid map: a non-slip (deterministic) condition and a slip  
134 (stochastic) condition. In the deterministic setting, action always results in the intended  
135 movement, and snow cells only add additional energy loss. In the stochastic setting,  
136 actions do not always lead to the intended outcome: with a probability of 1/3, the  
137 intended action is executed, and with a probability of 2/3, the agent moves in a random  
138 adjacent direction, simulating loss of traction control during winter driving on ice and  
139 snow.

140



141

142 Figure 14: shows a typical winter grid layout used in all experiments.

143

### 144 3.2 Reinforcement Learning Algorithms/ Methodologies:

145 The routing strategies and learning methods used in this study are described in this  
146 section.

147

#### 148 3.2.1 Traditional Baseline (BFS)

149 Traditional baseline is a non-learning shortest path strategy used as a comparison base  
150 for reinforcement learning methods. It operates on the 50x50 grid map, it considers  
151 four directions to move (up, down, left, right). The system calculates the shortest path  
152 to the destination, without considering energy savings, which means it does not adapt  
153 to environmental feedback or uncertainty.

154

### 155 3.2.2 Q-learning Implementation

156 Tabular Q-learning was used as the value-based reinforcement learning baseline. The  
157 agent maintains a discrete Q (s, a) table, which is updated iteratively based on observed  
158 state transitions. An  $\epsilon$ -greedy policy was used for exploration. The value of  $\epsilon$  begins at  
159 0.6 and decays exponentially with a factor of 0.9995 per episode until it reaches 0.05. In  
160 this 50×50 grid setting, the decay schedule allows broad initial exploration before  
161 shifting emphasis to exploitation. Actions are restricted to four possibilities—left (0),  
162 down (1), right (2), and up (3)—consistent with standard discrete grid-world formulations  
163 and the discrete grid structure. The Q-value update follows the classic  
164 temporal-difference form:

165

$$166 Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)] \quad (\text{Eq. 1})[5]$$

167

168 We fix the learning rate  $\alpha$  at 0.15 and the discount factor  $\gamma$  at 0.99. These values handle  
169 the stochastic transitions (is\_slippery=True) and support planning over long horizons in  
170 the large grid [3]. State  $s$  denotes the flattened grid position index. The action  $a$  selects  
171 one of the four movement directions.  $Q(s, a)$  represents the estimated discounted  
172 cumulative reward starting from state  $s$  and action  $a$  [3,75].

173

### 174 3.2.3 Reward system

175 In this research, the reward system is designed to represent energy efficiency under  
176 winter conditions, it helps agents find the best paths. At each step, the system gives a  
177 negative value to show how significant a step is to energy saving, with higher cause on  
178 snow grids. A positive 700 is given only when the agents successfully reach the  
179 destination, while within the maximum steps of 1000 steps. It encourages agents to  
180 reach the destination with considerations on energy saving. We initialize the cumulative  
181 reward to 0 at the beginning of each episode. Each will result in negative values,  
182 because by doing this, agents don't need to consider the prior bias, since it assumes  
183 nothing at first, etc. As in the settings, we have Near-snow (light snow), Edge-snow  
184 (medium snow), and Core-snow (deep snow); each of them has different additional  
185 values which are -0.2, -0.5, and -2.0 respectively [10],[6].

186

$$187 r_t = r_{step} - c_{energy}(s_t) + r_{goal} \quad (\text{Eq. 2})[6]$$

188

189 The basic reward function follows standard reinforcement learning practice by  
190 combining step penalties, energy-related costs, and a terminal goal reward.

191

### 192 3.2.4 Proximal Policy Optimization

193 Proximal Policy Optimization (PPO) served as the policy-gradient method in this study  
194 for energy-efficient routing under winter conditions. PPO learns a stochastic policy by  
195 directly outputting action probabilities for each state, which helps the agent cope with  
196 uncertainty on slippery roads.

197

$$198 L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \quad (\text{Eq. 3})[4]$$

199

200 where  $r_t(\theta)$  is the probability ratio between the new and previous policies. The clipping  
201 mechanism constrains policy updates to improve stability.:

202

203 The PPO agent was trained using fixed values across all experiments. The discount  
204 factor was set to  $\gamma=0.99$ , and Generalized Advantage Estimation (GAE) was used with  
205  $\lambda=0.95$ . The policy network consisted of two fully connected hidden layers with 256  
206 units each, using ReLU activation functions. The learning rate was set to  $3 \times 10^{-4}$ , with a  
207 clipping range of 0.2 and an entropy coefficient of 0.02 to encourage exploration. PPO  
208 training was conducted for 1,200,000 time steps, using a rollout buffer of 2048 steps, a  
209 batch size of 256, and 10 optimization epochs per update to improve training stability  
210 [84].

211

212 Policy-gradient methods work by directly adjusting a parameterized policy to increase  
213 expected reward. They learn a stochastic policy, allowing for more flexibility in  
214 uncertain environments, where the same action can lead to different outcomes. In  
215 contrast, q-learning estimates state-action values and usually converges with a  
216 deterministic policy based on these estimates. Although it can learn in unfamiliar  
217 environments, its action selection may be less reliable in highly unpredictable  
218 situations. As a result, policy-gradient and value-based methods show varying strengths  
219 depending on the level of uncertainty in the winter routing task.

220

### 221 3.2.5 Curriculum Learning for PPO

222 PPO training followed a two-phase curriculum learning approach designed to improve  
223 training stability and sample efficiency under sparse rewards and energy-sensitive  
224 penalties [11,810].

225

#### 226 **Phase 1: Navigation Learning Phase** (300,000 timesteps)

227 In phase 1, PPO was trained using a simplified reward function that focuses only on  
228 positive feedback for reaching the goal. Energy costs in snow grids were not included.  
229 This made the feedback denser and more immediate. The agent quickly learned basic  
230 navigation and achieved early success in the 50×50 winter grid. The policy learned in  
231 this phase served as the starting point for the next training stages.

232

#### 233 **Phase 2: Energy-Aware Optimization Phase** (900,000 timesteps)

234 In phase 2, the training continued with the complete energy-aware reward function,  
235 which now included penalties for moving across snow-covered areas. The agent had to  
236 balance successful arrival with lower energy consumption. All reported results, ablation  
237 studies, and comparisons are based solely on the Phase 2 reward function.

238

239

240

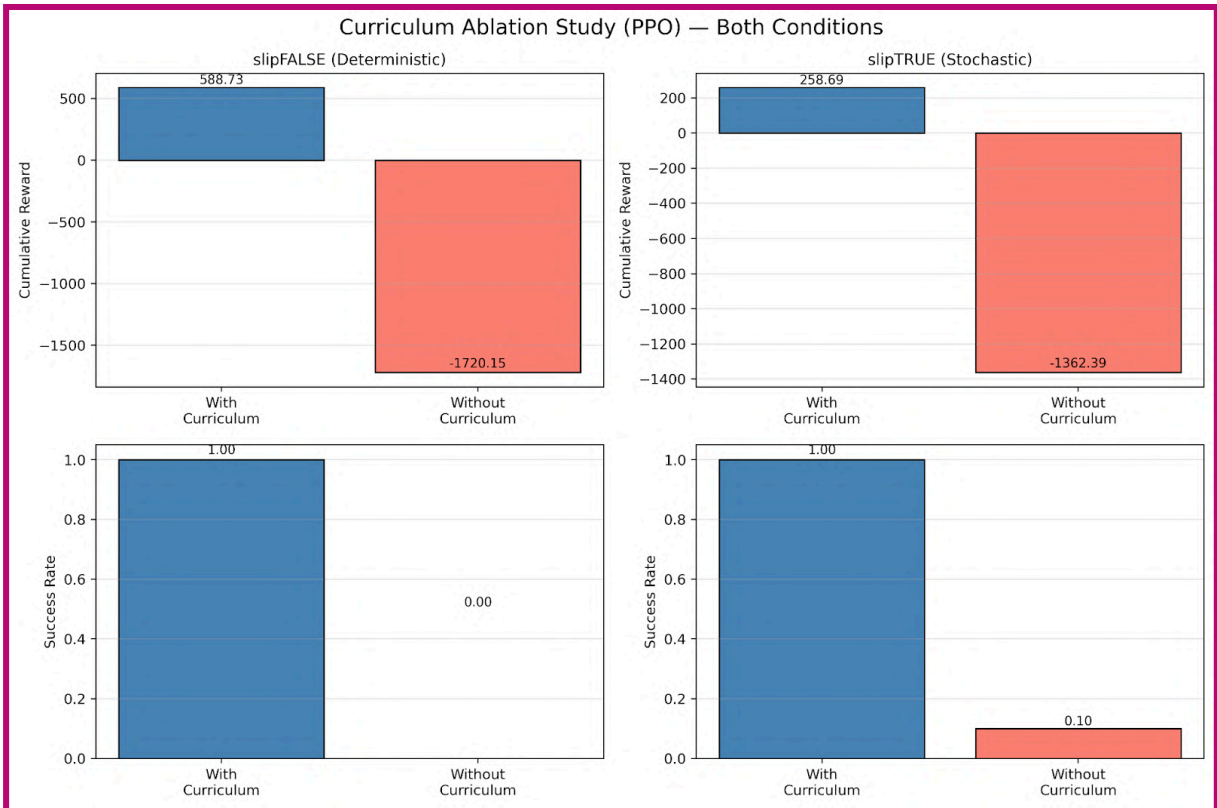
241

### 242 3.2.6 Curriculum Learning Ablation Study

243 To validate the necessity of curriculum learning, an ablation study was conducted  
244 comparing PPO with curriculum versus PPO without curriculum under both  
245 deterministic and stochastic conditions across 10 seeds assessed through total  
246 cumulative reward and success rate. Without Curriculum learning, PPO achieved a  
247 mean of -1720.15 (95% CI [-1894.63, -1545.67]) and -1362.39 (95% CI [-1780.01, -944.78])  
248 with 0% and 10% success rate respectively, under deterministic and stochastic  
249 conditions. Whereas with curriculum, PPO demonstrates a strong performance of  
250 588.73 (95% CI [588.25, 589.21]) and 258.69 (95% CI [251.02, 266.37]) with 100% success  
251 rate under both conditions. These results confirm that curriculum learning is a critical  
252 component for PPO, as its absence leads to failure under both conditions.

253

254



255

256

257 Figure 2: displays the cumulative reward and success rate for PPO with and without  
258 curriculum under both deterministic and stochastic conditions, illustrating the  
259 substantial performance gap between the two configurations.

260

261 Therefore, all subsequent comparisons and evaluations in this study refer exclusively to  
262 PPO with curriculum learning.

263

### 264 3.3 Evaluation factors

265 To ensure statistical reliability, all reinforcement learning experiments were repeated  
266 over 105 independent random seeds (seed 0 to seed 9), and results are reported as the  
267 mean and standard deviation with 95% confidence intervals across runs.

268

269 To compare the performances of all routing methods under identical conditions, a final  
270 evaluation was conducted after training. Each method BFS, Q-learning, and PPO was  
271 evaluated using the following metrics:

272

- 273 • Total Reward: calculated as the cumulative reward obtained over the episode.
- 274
- 275 • Number of steps: steps taken from the start to the goal (less than 1000 steps).
- 276
- 277 • Average snow cell visited: Snow visits were calculated by counting the number of  
278 snow-covered cells traversed in each episode and averaging this value across all  
279 evaluation episodes.
- 280
- 281 • Success rate: defined as the percentage of episodes in which the agent reached  
282 the goal within the maximum step limit.

283

284 The BFS baseline was evaluated only once on the fixed grid map, as it produces a  
285 deterministic path. Reinforcement learning agents were evaluated over 200 episodes  
286 per run, with experiments repeated over 105 independent random seeds for each  
287 condition (slip and non-slip). All methods were compared using the same energy-aware  
288 reward function and environment configuration to ensure fairness.

289

### 290 3.4 Hyperparameter Selections

291 Hyperparameters are selected based on standard practices and empirical validation to  
292 ensure stable and consistent performance [3,73,6,73,4].

293

#### 294 3.4.1 Q-Learning Hyperparameter Configuration

295 We implement tabular Q-learning with an  $\epsilon$ -greedy exploration strategy.

296

297 The hyperparameters are set as follows:

298 - Learning rate  $\alpha = 0.15$

299 - Discount factor  $\gamma = 0.99$

300 - Exploration strategy:  $\epsilon$ -greedy

301 - Initial  $\epsilon = 0.60$

302 - Minimum  $\epsilon = 0.05$

303 - Exponential decay rate = 0.9995

304 - Training episodes = 15,000

305 - Maximum steps per episode = 1,000

306

307 The learning rate ( $\alpha = 0.15$ ) is selected to balance convergence speed and stability. A  
308 moderate learning rate allows the agent to adapt efficiently while avoiding oscillations  
309 in value updates.

310

311 The discount factor ( $\gamma = 0.99$ ) emphasizes long-term rewards, which is essential for  
312 navigation tasks where the objective is to reach the goal efficiently over multiple steps.

313

314 An  $\epsilon$ -greedy exploration strategy is adopted to balance exploration and exploitation.  
315 The initial exploration rate ( $\epsilon = 0.60$ ) encourages sufficient exploration in early training,  
316 while exponential decay gradually shifts the policy toward exploitation. The minimum  
317 ( $\epsilon = 0.05$ ) ensures that some level of exploration is maintained throughout training,  
318 preventing the agent from getting stuck in suboptimal policies.

319

320 The number of training episodes 15,000 and maximum steps per episode 1,000 are  
321 chosen to ensure sufficient interaction with the environment for convergence, while  
322 maintaining computational efficiency.

323

324 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) is used to assess the learned policy  
325 performance.

326

### 327 **3.4.2 PPO Hyperparameter Configuration**

328 We implement PPO using a clipped surrogate objective with generalized advantage  
329 estimation (GAE).

330

331 The hyperparameters are set as follows:

332 - Discount factor  $\gamma = 0.99$

333 - GAE parameter  $\lambda = 0.95$

334 - Learning rate =  $3 \times 10^{-4}$

335 - Clipping range  $\epsilon = 0.2$

336 - Rollout buffer size = 2048 steps

337 - Batch size = 256

338 - Optimization epochs per update = 10

339 - Entropy regularization is enabled

340

341 The policy and value networks share a neural architecture consisting of two hidden  
342 layers with 256 units each and ReLU activation.

343

344 The discount factor ( $\gamma = 0.99$ ) is selected to emphasize long-term rewards, which is  
345 essential for navigation tasks. The GAE parameter ( $\lambda = 0.95$ ) provides a balance between  
346 bias and variance in advantage estimation, leading to more stable learning.

347

348 The clipping parameter ( $\epsilon = 0.2$ ) constrains policy updates to prevent large deviations  
349 from the previous policy, which improves training stability. The rollout buffer size and

350 number of optimization epochs are chosen to ensure sufficient policy updates while  
351 avoiding overfitting to recent trajectories.

352

353 Entropy regularization is included to encourage exploration and prevent premature  
354 convergence to suboptimal policies.

355

356 Overall, these hyperparameters follow widely adopted settings in reinforcement  
357 learning literature and are chosen to ensure stable and consistent training rather than  
358 aggressive performance tuning.

359

360 A complete summary of environment settings and hyperparameters is provided in  
361 Appendix A.

362

363

364

## 365 4. Results

366 The experiments compare BFS, Q-learning, and PPO on the same 50×50 winter grid,  
367 looking at both non-slip (deterministic) and slip (stochastic) cases.

368

### 369 4.1 Evaluation Setup and Metrics

370 The grid layout, start and goal positions, and reward parameters were kept exactly the  
371 same across all methods. All experiments were repeated over 105 independent random  
372 seeds, and results are reported as the mean  $\pm$  standard deviation with 95% confidence  
373 intervals across runs. Trajectory figures shown are from the representative seed whose  
374 performance was closest to the mean. For each run, total reward, number of steps,  
375 average snow cells crossed per episode, and success in reaching the goal were recorded.

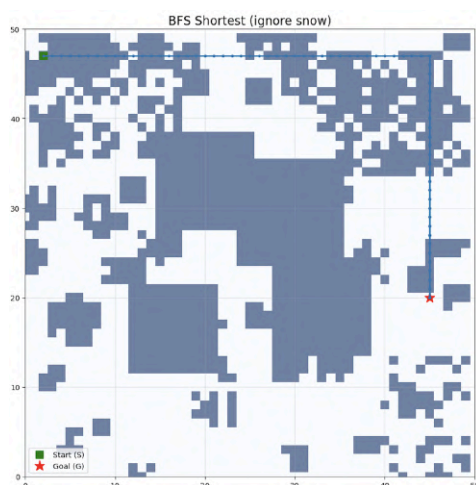
376

### 377 4.2 Trajectory Visualizations and Quantitative Results

#### 378 4.2.1 Deterministic (Non-Slip) Winter Condition

379

380

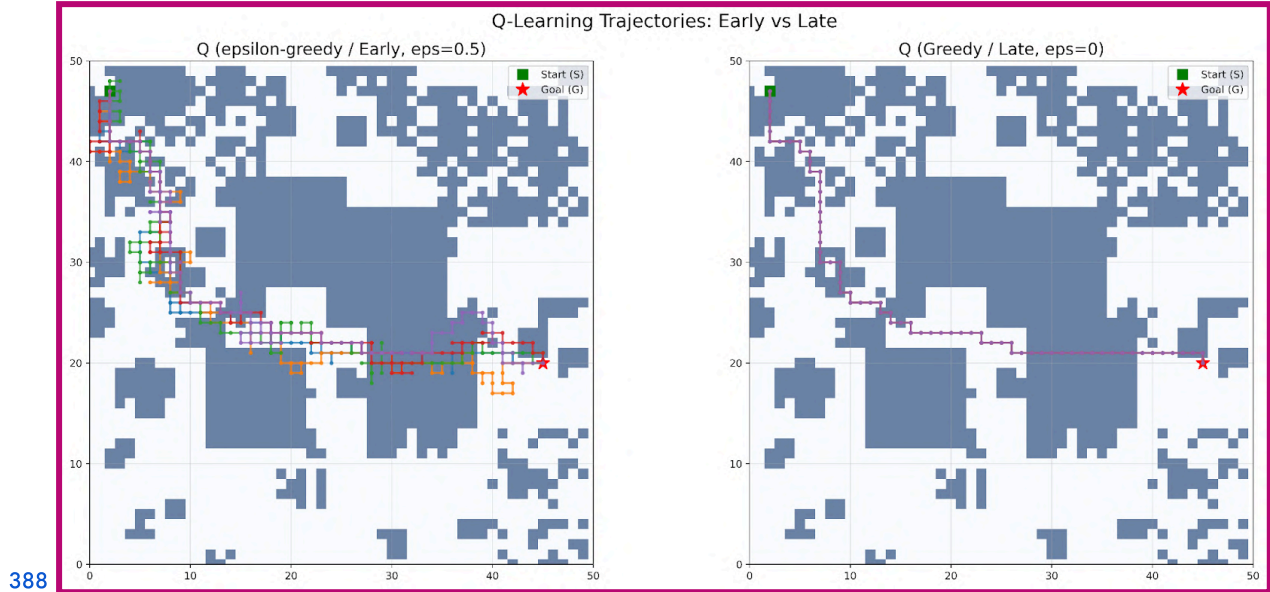


382 Figure 33: illustrates the BFS path under non-slip conditions. Since BFS always finds the  
383 shortest route, the trajectory goes straight from the start to the goal with minimal cells  
384 visited without any deviation.

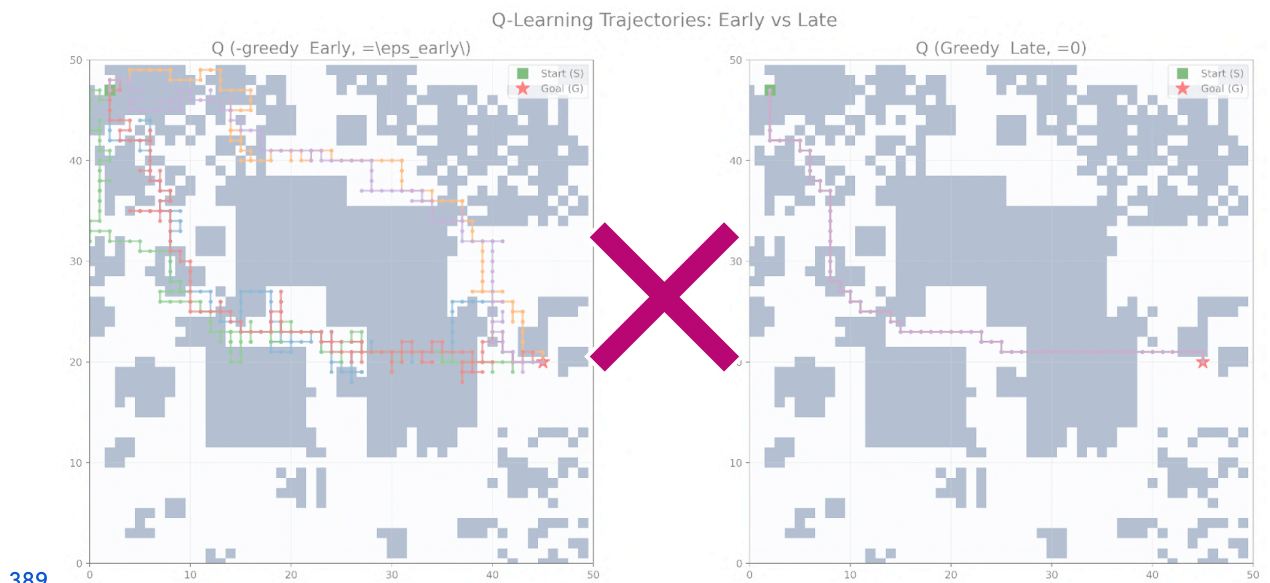
385

386

387



388

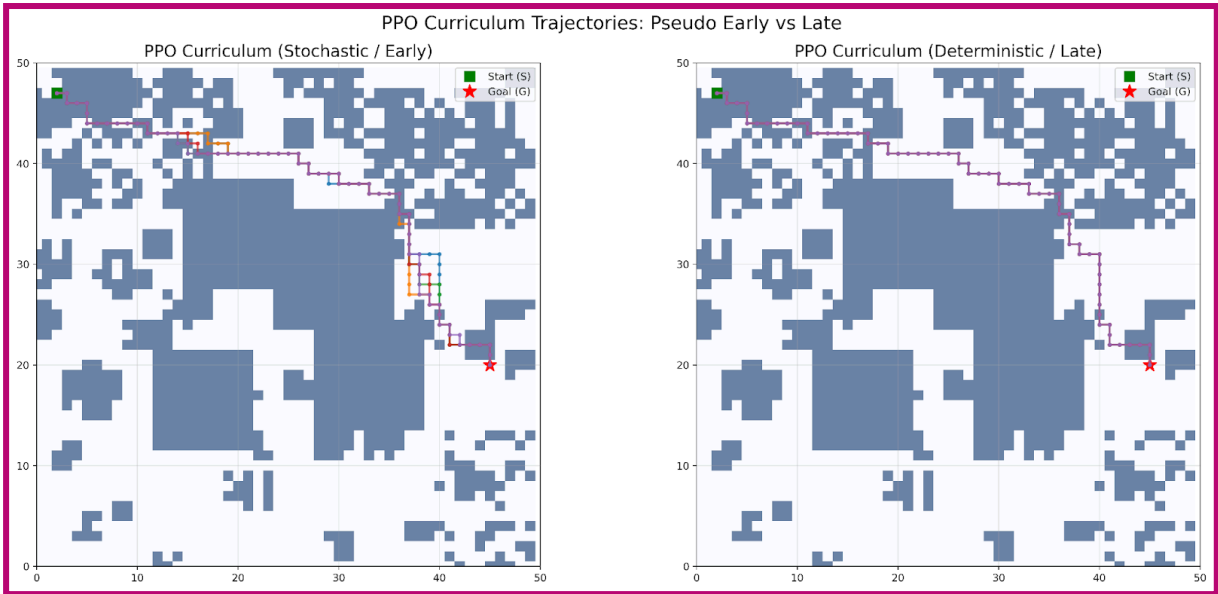


389

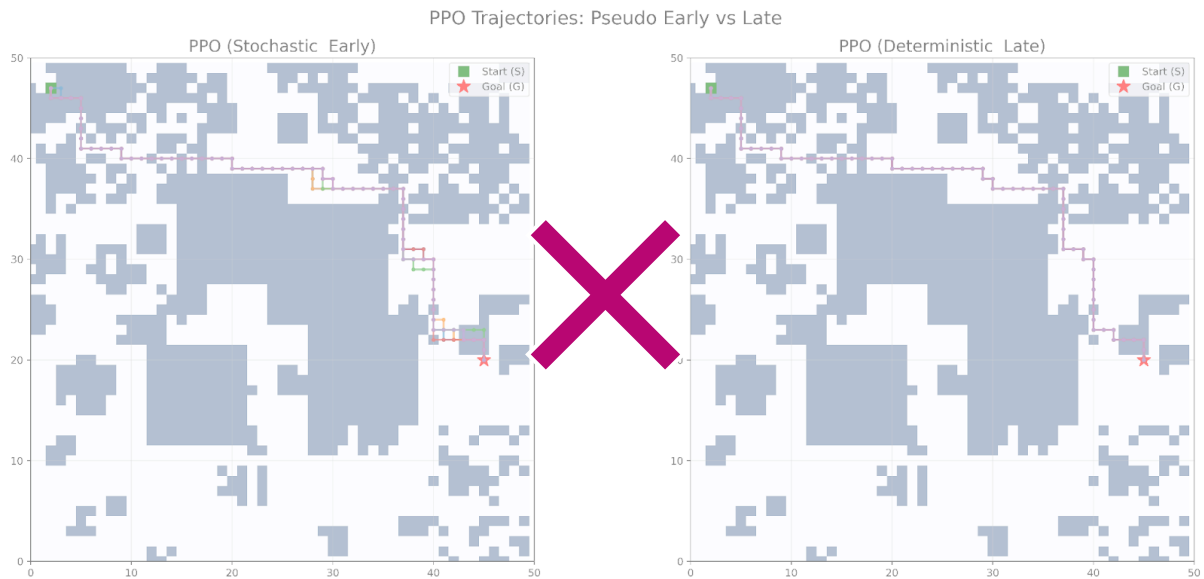
390 Figure 44: illustrates the navigation trajectories produced by the Q-learning agent under  
391 deterministic (non-slip) winter conditions during early training with an  $\epsilon$ -greedy policy  
392 and late training with a greedy policy. The trajectories demonstrate the transition from  
393 exploratory behavior to a stable path toward the goal.

394

395



396

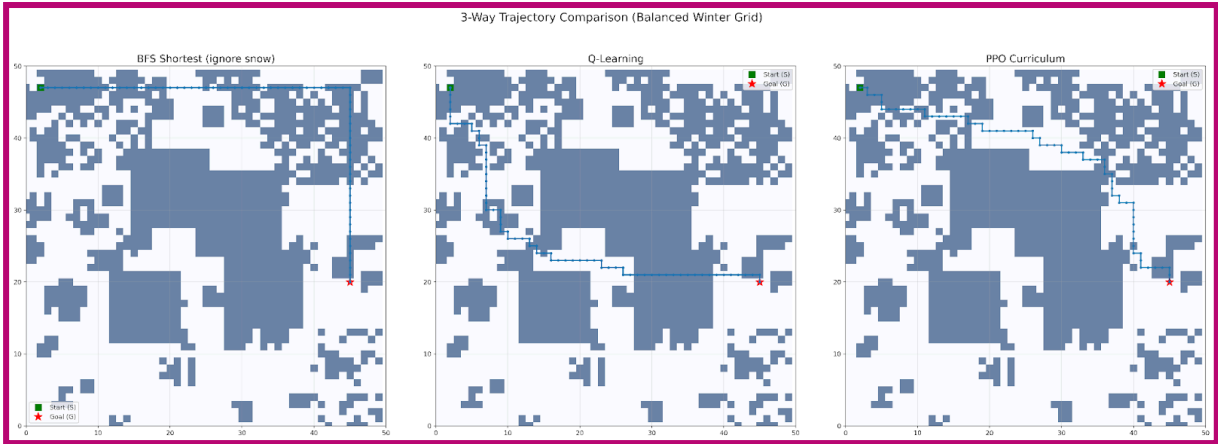


397

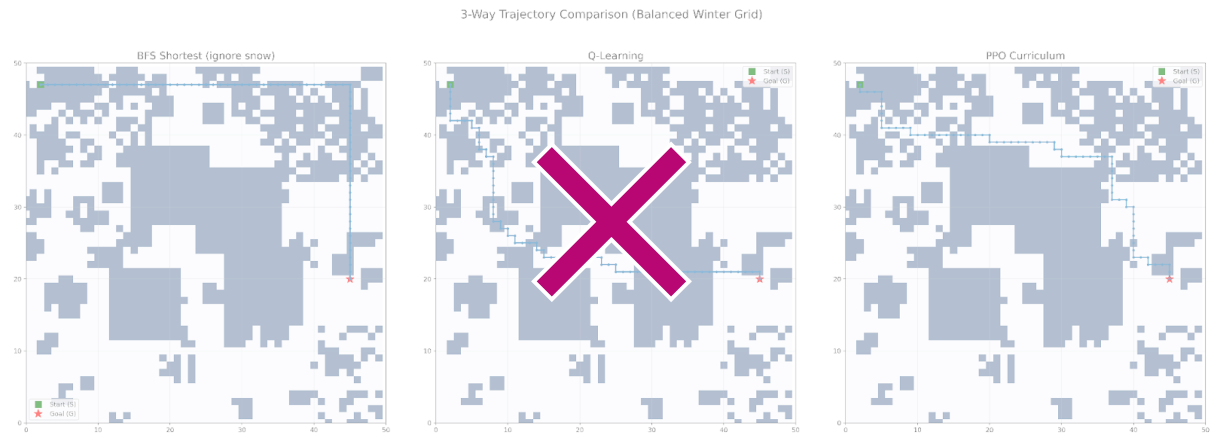
398 Figure 55: illustrates the navigation trajectories produced by the PPO agent during early  
 399 and late stages of training under deterministic (non-slip) winter conditions.

400

401



402



403

404 Figure 66: compares the trajectories of BFS, Q-learning and PPO in the non-slip winter  
 405 setting.

406

407

408 === Final Comparison Table ===

Method	Reward	Steps	Snow Visits	Success
<del>BFS Shortest</del>	<del>575.30 ± 0.00</del>	<del>70.00 ± 0.00</del>	<del>31.00 ± 0.00</del>	<del>100.0%</del>
<del>Q-Learning</del>	<del>582.01 ± 6.97</del> <del>579.10 ± 5.61</del>	<del>70.00 ± 0.00</del>	<del>13.10 ± 3.05</del> <del>12.60 ± 3.29</del>	<del>100.0%</del>
<del>PPO Curriculum</del>	<del>588.73 ± 0.67</del> <del>588.50 ± 0.60</del>	<del>70.00 ± 0.00</del>	<del>7.60 ± 0.97</del> <del>7.80 ± 1.10</del>	<del>100.0%</del>

409

410

411

412

413

414

415

416

417

Method	Reward (mean±SD)	95%CI	Steps	Snow Visits	Success
<b>BFS Shortest</b>	575.30 ± 0.00	[575.30, 575.30]	70.00 ± 0.00	31.00 ± 0.00	100.0%
<b>Q-Learning</b>	582.01 ± 6.97	[577.02, 587.00]	70.00 ± 0.00	13.10 ± 3.05	100.0%
<b>PPO Curriculum</b>	588.73 ± 0.67	[588.25, 589.21]	70.00 ± 0.00	7.60 ± 0.97	100.0%

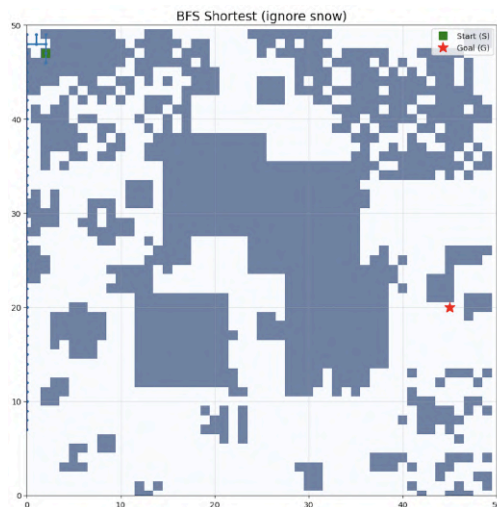
418

419 Table 1: summarizes the main performance metrics for all three methods under  
420 deterministic conditions, including total reward, steps to goal, average snow cells  
421 crossed, success rate, Standard Deviation (SD), and averaged across 105 independent  
422 random seeds with 95% confidence intervals.

423

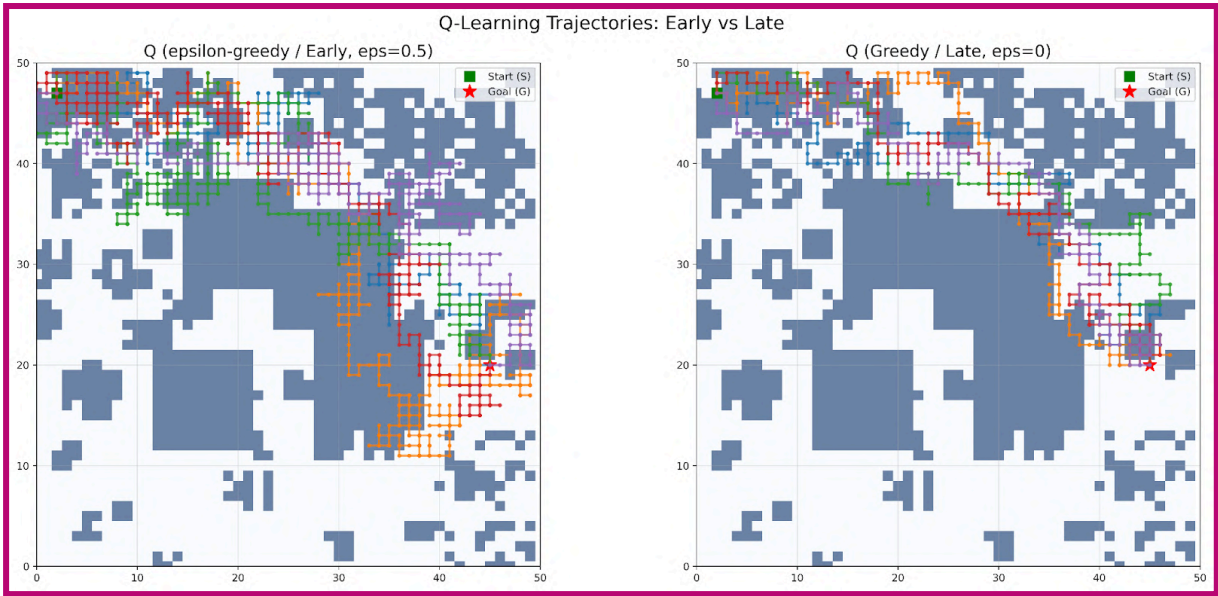
424

#### 425 4.2.2 Slip Winter Condition



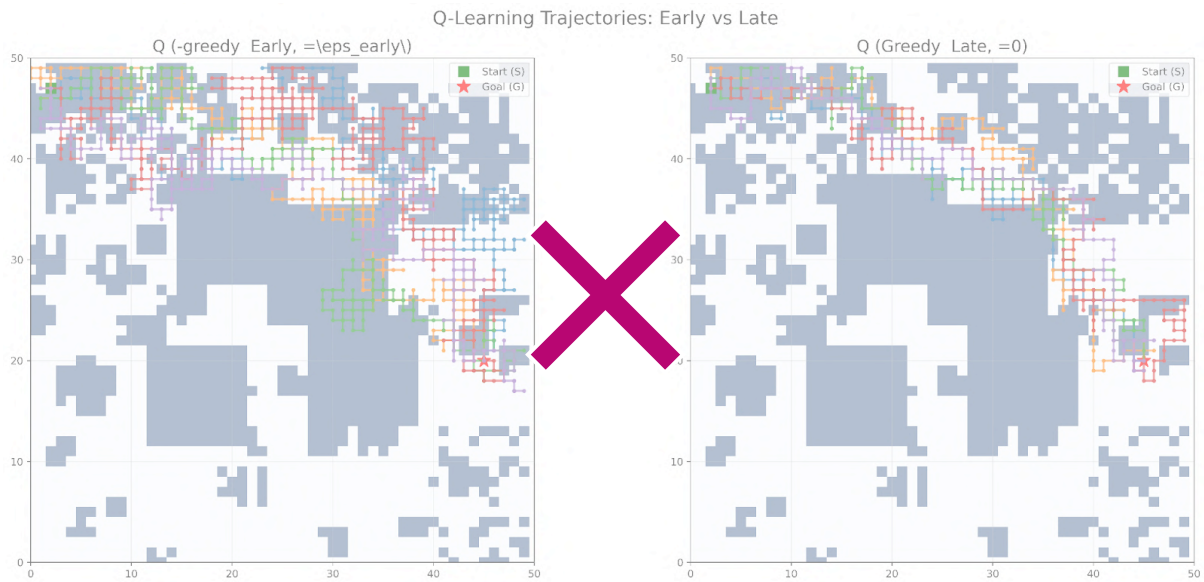
426

427 Figure 77: illustrates the navigation path produced by the BFS baseline under slip  
428 (stochastic) winter conditions.



429

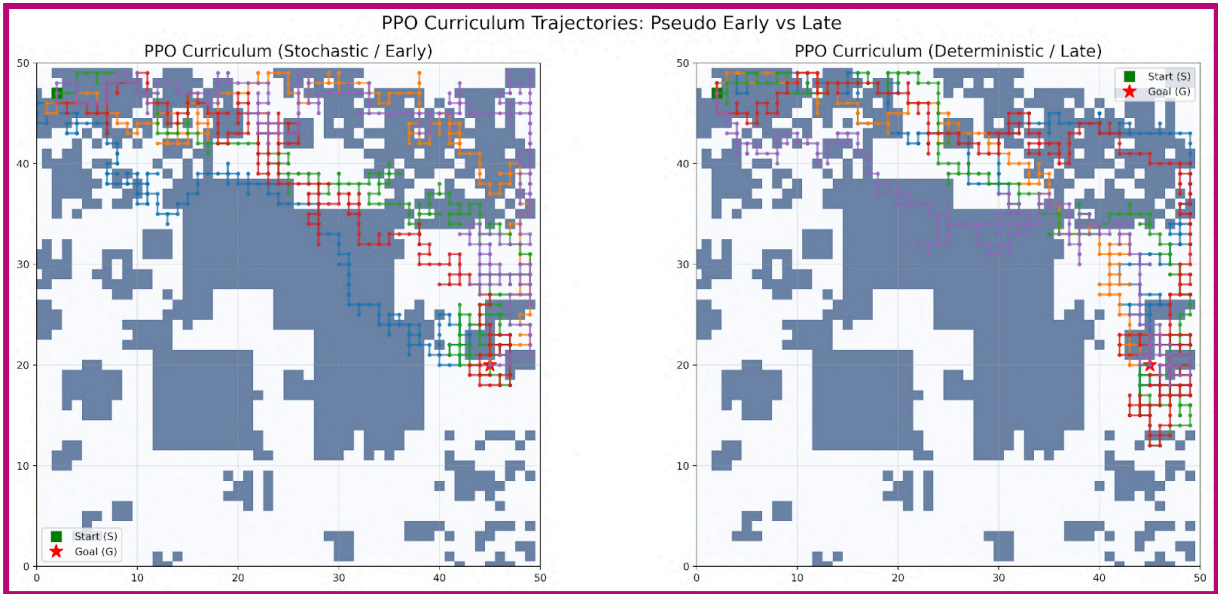
430



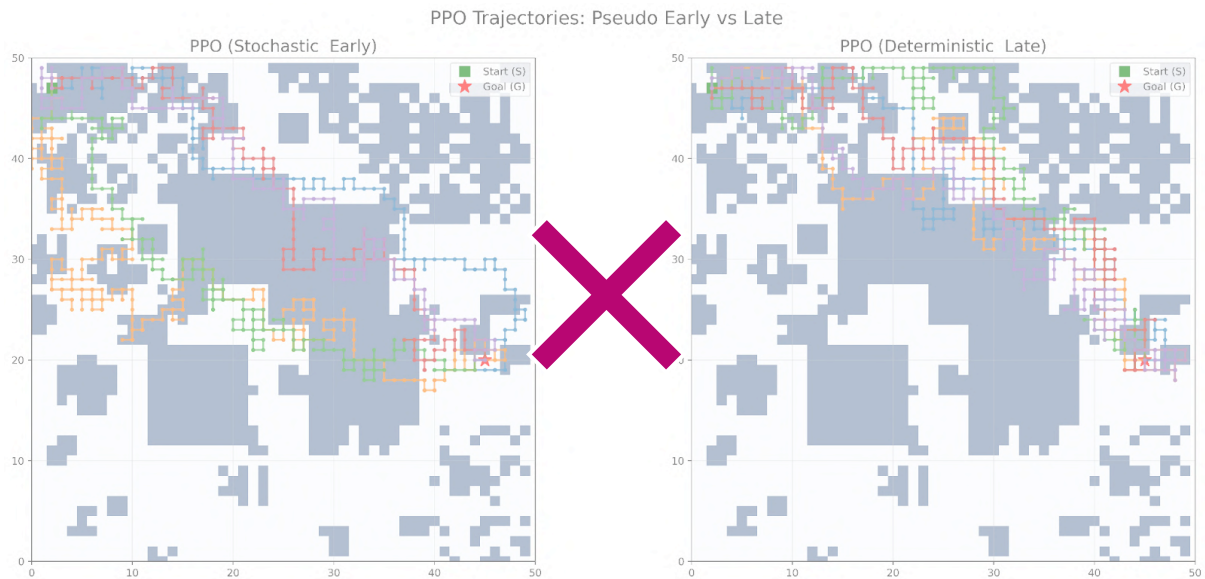
431

432 Figure 88: illustrates the navigation trajectories produced by the Q-learning agent  
 433 during early and late stages of training under slip (stochastic) winter conditions.

434



435

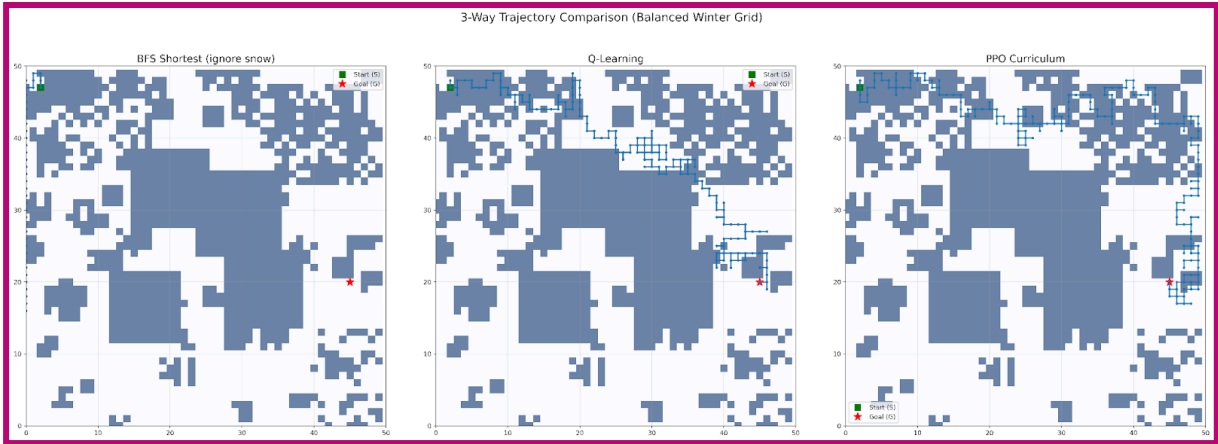


436

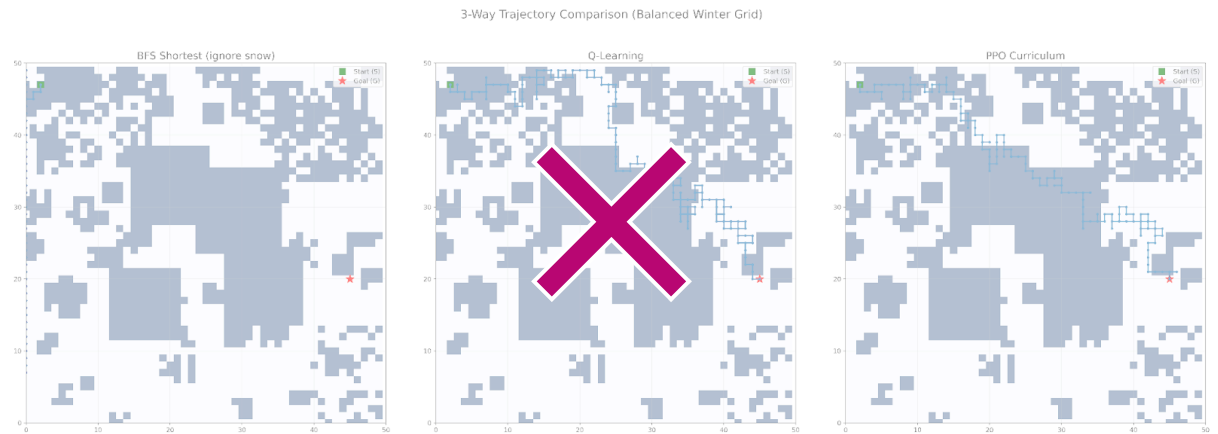
437 Figure 99: illustrates the navigation trajectories produced by the PPO agent during early  
 438 and late stages of training under slip (stochastic) winter conditions.

439

440



441



442

443 Figure 1010: compares the navigation trajectories of BFS, Q-learning, and PPO under slip  
 444 winter conditions.

445

446

447 === Final Comparison Table ===

Method	Reward	Steps	Snow Visits	Success
<del>BFS Shortest</del>	<del><math>-1668.41 \pm 113.42</math></del> <del><math>693.96 \pm 61.27</math></del>	<del><math>1000.00 \pm 0.00</math></del>	<del><math>246.80 \pm 76.09</math></del>	<del>0.0%</del>
<del>Q-Learning</del>	<del><math>-311.82 \pm 4.83</math></del> <del><math>304.03 \pm 9.92</math></del>	<del><math>-226.40 \pm 3.92</math></del> <del><math>228.18 \pm 4.10</math></del>	<del><math>-66.94 \pm 3.87</math></del> <del><math>67.58 \pm 4.34</math></del>	100.0%
<del>PPO Curriculum</del>	<del><math>-258.69 \pm 10.73</math></del> <del><math>252.19 \pm 7.04</math></del>	<del><math>241.18 \pm 4.11</math></del> <del><math>242.29 \pm 3.00</math></del>	<del><math>-91.88 \pm 5.24</math></del> <del><math>92.65 \pm 4.14</math></del>	100.0%

448

449

450

451

452

453

454

455

456

Method	Reward (mean±SD)	95% CI	Steps	Snow Visits	Success
<b>BFS Shortest</b>	-1668.41 ± 113.42	[-1749.54, -1587.28]	1000.00 ± 0.00	246.80 ± 76.09	0.0%
<b>Q-Learning</b>	311.82 ± 4.83	[308.37, 315.27]	26.40 ± 3.92	66.94 ± 3.87	100.0%
<b>PPO Curriculum</b>	258.69 ± 10.73	[251.02, 266.37]	241.18 ± 4.11	91.88 ± 5.24	100.0%

457

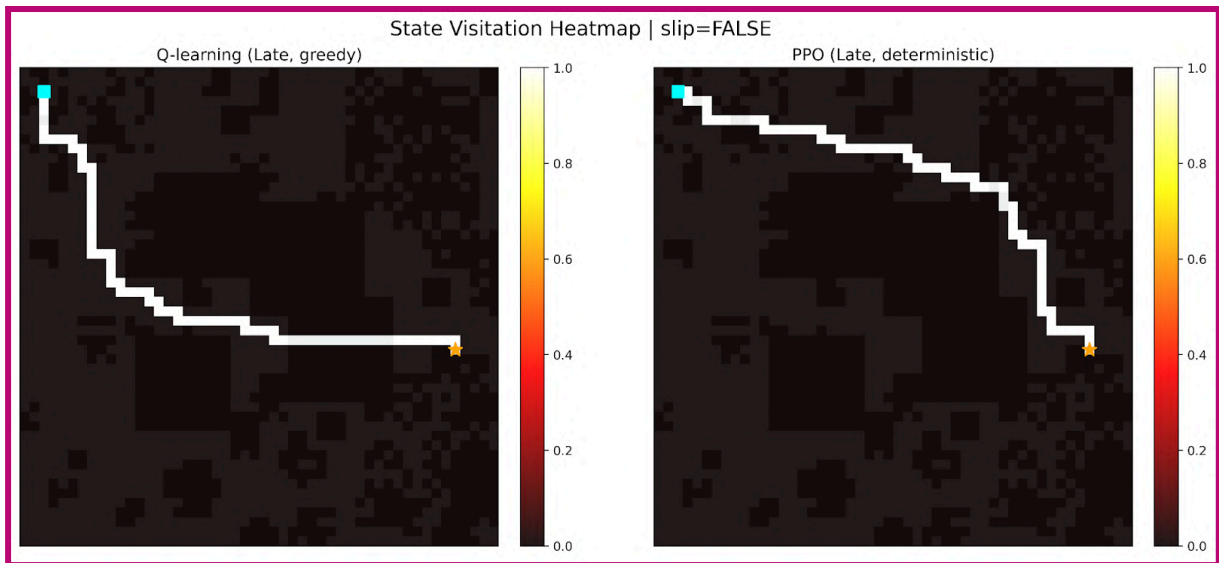
458 Table 2: summarizes the quantitative performance metrics for BFS, Q-learning, and PPO  
459 under slip winter conditions, using the same evaluation metrics as in the deterministic  
460 case, **S**standard **D**eviation (**SD**), and averaged across **105** independent random seeds  
461 **with 95% confidence intervals.**

462

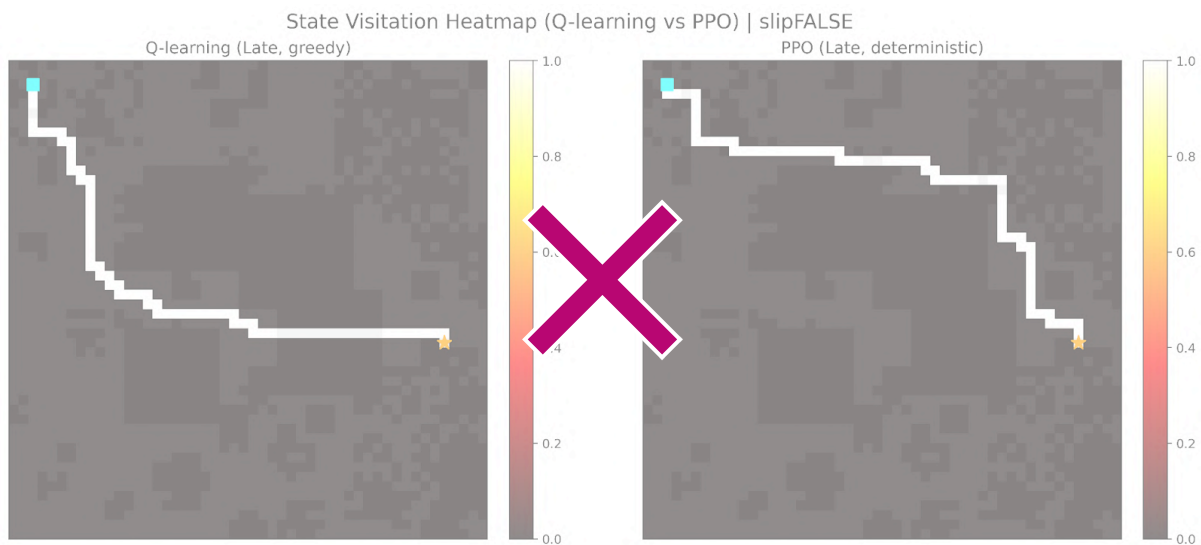
463

### 464 **4.3 Spatial State Visitation Heatmaps**

465 Spatial state visitation heatmaps offer a grid-based way to see how often an agent  
466 passes through each cell during navigation. In our 50×50 winter grid, each cell  
467 corresponds to a state, and the color intensity reflects visitation frequency across  
468 episodes. Darker areas indicate cells that are visited more frequently, while lighter or  
469 white cells indicate rarely or never visited locations. Such visualizations give a clear  
470 picture of the agent's overall path distribution and behavior patterns. We generated  
471 these heatmaps using aggregated visitation counts normalized to the ~~[-0,1]~~ range  
472 **between 0 and 1.**



473

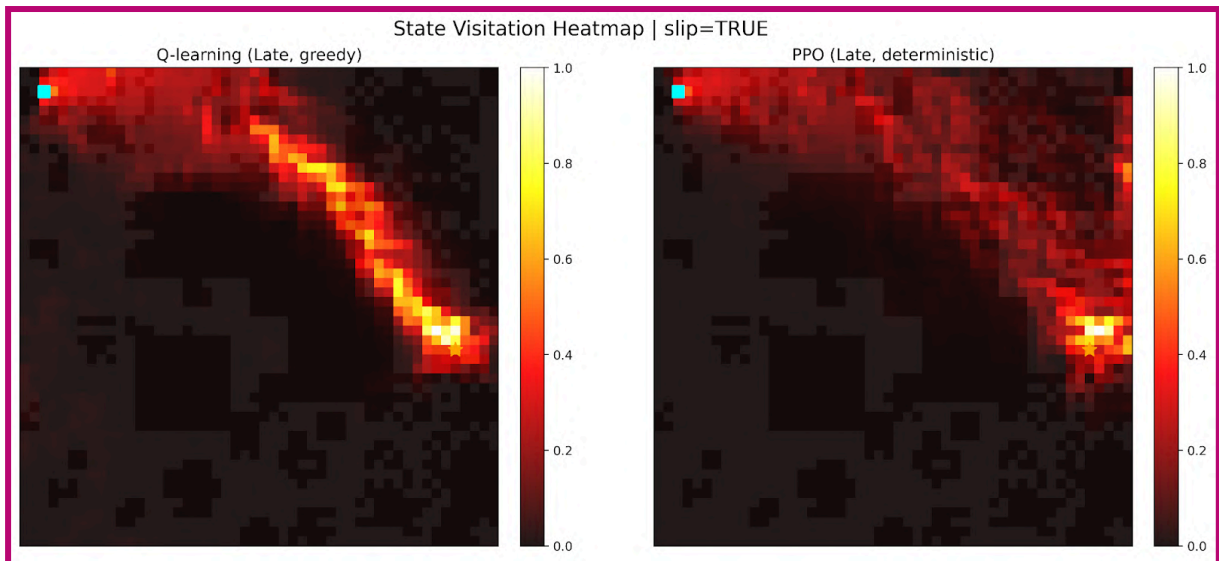


474

475 Figure 114: displays the visitation heatmaps for Q-learning and PPO under deterministic  
 476 (non-slip) winter conditions. The results are aggregated from 10 evaluation runs  
 477 consisting of 200 episodes per run. Both agents concentrate most visits along a narrow  
 478 corridor connecting the start to the goal, with near-maximum intensity ( $\approx 1.0$ ) along the  
 479 primary route and very low visitation elsewhere.

480

481



482



483

484 Figure 1212: shows the visitation intensity for each grid cell in the slip case. Compared  
 485 with the non-slip condition, visitation is more widely distributed across the grid rather  
 486 than remaining within a narrow corridor. The heatmap aggregates data from 10  
 487 evaluation runs consisting of 200 episodes per run.

488

#### 489 4.4 Statistical Analysis and Learning Curves - Slip Condition

490

##### 491 4.4.1 Statistical Significance (Welch's t-test)

492 To assess statistical significance, a Welch's t-test was conducted comparing the  
 493 cumulative reward of Q-learning and PPO under slip conditions across 10 seeds. The  
 494 results confirmed that Q-learning significantly outperformed PPO in cumulative  
 495 rewards ( $t = 14.28, p < 0.001$ ), indicating that the performance gap is highly unlikely to be  
 496 due to random variation. Q-learning achieved a mean reward of  $311.82 \pm 4.83$  compared  
 497 to PPO's  $258.69 \pm 10.73$ . The higher standard deviation of PPO ( $SD = 10.73$ ) compared to  
 498 Q-learning ( $SD = 4.83$ ) further indicates greater training instability under stochastic  
 499 conditions.

500 The standard deviation gap between PPO in both conditions suggests that PPO is more  
 501 sensitive to stochastic signals under slip condition, where high-probability slip results  
 502 in noisy advantage estimation, leading to higher variance across training runs (SD =  
 503 10.73). In contrast, under deterministic conditions PPO exhibits a much lower variance  
 504 (SD = 0.67), confirming that the instability is determined by the stochastic environment,  
 505 not the algorithm itself.

506

507

508 === Final Comparison Table ===

509

Metric	t-statistic	p-value	Q mean $\pm$ SD	PPO mean $\pm$ SD
Cumulative Reward	14.28	< 0.001	311.82 $\pm$ 4.83	258.69 $\pm$ 10.73
Success Rate	n/a	n/a	100% $\pm$ 0%	100% $\pm$ 0%

510

511 Table 3: Welch's t-test results comparing Q-learning and PPO Curriculum cumulative  
 512 reward under slip conditions across 10 seeds. n/a indicates that the test was not  
 513 applicable due to zero variance in success rate for both methods.

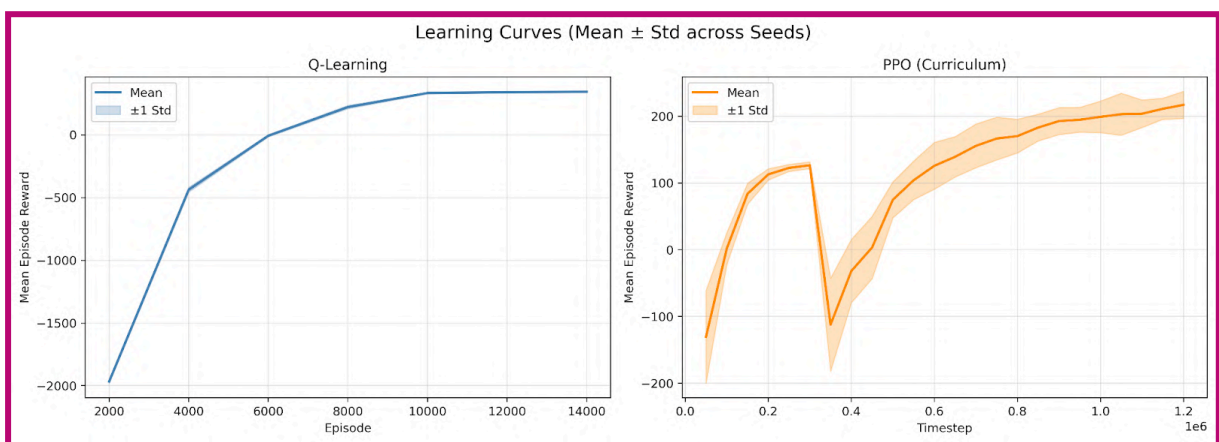
514

#### 515 4.4.2 Learning Curves

516 Learning curves illustrate the performance of Q-learning and PPO evolves over the  
 517 course of training under stochastic conditions, and provide insight into convergence  
 518 behaviour, training stability, and the impact of curriculum learning on PPO.

519

520



521

522

523 Figure 13: Learning curves (Mean  $\pm$  SD across 10 seeds) under slip (stochastic)  
 524 conditions. Left: Q-learning mean episode reward vs. training episode (15,000 episodes)  
 525 showing smooth, steady and consistent convergence. Right: PPO Mean episode reward  
 526 vs. training timestep (1,200,000 steps), where the pronounced drop around timestep  
 527 300,000 corresponds to the Phase 1 to Phase 2 curriculum transition, when

528 energy-aware snow penalties are introduced and the agent must re-adapt its policy.  
529 After the transition is complete, PPO gradually recovers and converges, but its variance  
530 band is significantly wider compared to Q-learning.

531

#### 532 **4.54 Limitations & Uncertainty**

533 Several sources of uncertainty and limitations appeared in this study. Both Q-learning  
534 and PPO showed noticeable performance differences across training runs, which was  
535 expected given the stochastic nature of the environment and the learning algorithms.  
536 To account for this variability, all experiments were repeated over 105 independent  
537 random seeds, and results are reported as the mean across runs.

538 Under slip conditions, the same policy could produce quite different routes from  
539 episode to episode because actions did not always lead to the expected outcome. In  
540 addition, stochastic exploration strategies and random initialization of value functions  
541 (or policy networks) exposed agents to slightly different states, transitions, and rewards  
542 across runs. This variability made it harder to get perfectly consistent results.

543 An unexpected issue was that the BFS baseline failed to reach the goal within the  
544 1,000-step limit under slip conditions, resulting in a number of unsuccessful evaluation  
545 episodes.

546

547 A full numerical breakdown of evaluation metrics is provided in Appendix B.

548

549

---

550

## 551 **5. Discussion**

### 552 **5.1 Restatement of Hypothesis and Summary of Findings**

553 The study hypothesized that reinforcement learning-based agents, particularly Proximal  
554 Policy Optimization (PPO), would achieve more energy-efficient winter navigation paths  
555 than a traditional shortest-path baseline and a value-based Q-learning agent under  
556 both deterministic and slip winter conditions. The averaged results partially supported  
557 the hypothesis. Under deterministic (non-slip) conditions, PPO achieved higher  
558 cumulative reward, and fewer snow visits, outperforming Q-learning, which supports  
559 the hypothesis. However under stochastic (slip) conditions, Q-learning outperformed  
560 PPO in terms of cumulative reward or snow-cell avoidance, which did not support the  
561 hypothesis. Both reinforcement learning methods perform better results than  
562 traditional BFS across these environments.

563

### 564 **5.2 Interpretation of Q-Learning and PPO Performance**

565 The results showed that algorithm performance varied depending on the environmental  
566 condition. Under deterministic (non-slip) conditions, PPO achieved higher cumulative  
567 reward and fewer snow cell visits than Q-learning, suggesting that PPO's policy gradient  
568 approach was able to learn a more energy-efficient route in a stable environment.  
569 Under stochastic (slip) conditions, Q-learning outperformed PPO in terms of cumulative  
570 reward and snow-cell avoidance. One possible explanation was that a discrete and clear

571 grid-world environment and relatively small environment would favor more for  
572 Q-learning , as it can perform exact value iteration over the finite MDP (Markov  
573 Decision Process), directly converging to the true optimal Q-values for each state  
574 without approximation error. ~~Therefore,~~ ~~Ttherefore~~ allowing Q-learning to  
575 efficiently estimate optimal state and gives lower variance than policy gradient method  
576 (PPO), since it updates based on individual state-action pairs rather than entire  
577 trajectories [7,87]. In contrast, PPO relied on function approximation and stochastic  
578 policy updates, which may have required more training data or longer training time to  
579 converge to an equally optimal policy under slip conditions. Additionally, its stochastic  
580 policy may have introduced suboptimal choices that were not effective, whose gradient  
581 estimates already carry high variance from episode training. Under slip conditions,  
582 environmental stochasticity further interrupts reward signals, making them sparse and  
583 noisy, which destabilizes gradient estimates and requires significantly more training  
584 samples. ~~T~~therefore reducing its total rewards relative to Q-learning under stochastic  
585 conditions [83].

586

### 587 5.3 Effects of Slip (Stochastic) Winter Conditions

588 Under slip (stochastic) conditions, all agents showed broader trajectory dispersion and  
589 increased visits to previously visited states because of slipping, as reflected in the  
590 spatial state visitation heatmaps. This behavior is consistent with probabilistic transition  
591 dynamics, in which the same action could lead to different movement outcomes across  
592 episodes, such as deviating left in one run and right in another. When the grid is  
593 slippery, agents don't follow one stable path anymore. As a result, the routes spread out,  
594 and their performance becomes less consistent across episodes. These observations are  
595 consistent with prior navigation studies showing that stochastic environments increase  
596 policy variance and reduce convergence stability in reinforcement learning tasks [4,88].

597

598

599

### 600 5.4 Interpretation of BFS Baseline Behavior

601 The Traditional baseline trajectories did not take account for winter grid costs or  
602 stochastic transitions, it focused on the fastest way to the destination. In non-slip  
603 conditions this approach naturally produced the optimal route. However, under slip  
604 conditions, BFS exceeded the maximum step limit in multiple evaluation episodes,  
605 reflecting its inability to adapt its policy in response to transition winter grid costs or  
606 stochastic transitions.

607

608 Because BFS always assumed deterministic movement, each slip caused deviations from  
609 the planned path. These deviations often led to inefficient loops and longer paths.  
610 Unlike reinforcement learning methods, BFS did not have reward feedback or policy  
611 updates, which limited its ability to adjust navigation behavior under changing  
612 environmental dynamics. Therefore, making it the least effective method.

613

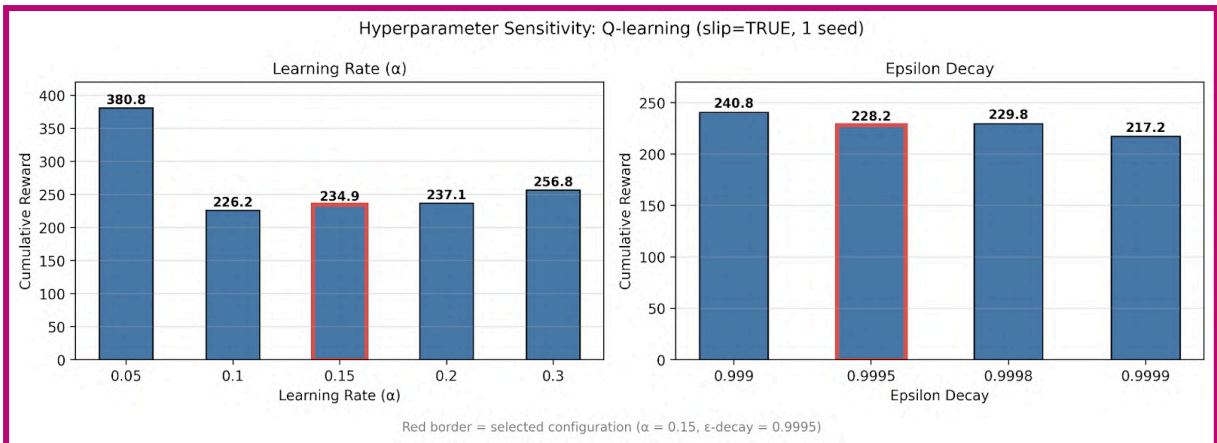
614 **5.5 Hyperparameter Sensitivity Analysis**

615 To assess the robustness of the reported conclusions to hyperparameter choice, a  
 616 sensitivity analysis was conducted to independently vary each hyperparameter of  
 617 Q-learning and PPO under slip conditions using a single representative seed. For  
 618 Q-learning, the learning rate (alpha) and epsilon decay were evaluated to identify which  
 619 hyperparameter values lead to greater stability and performance. For PPO, the clip  
 620 range and learning rate were varied to assess how different hyperparameters affect  
 621 training stability and sensitivity to stochastic conditions.

622

623

624



625

626

627 Figure 14: Comparison of different learning Rate (alpha) values and epsilon decay values .  
 628 The highlighted values represent the selected hyperparameter configuration, chosen  
 629 based on their balance of performance and training stability. ( $\alpha = 0.15$ ,  $\epsilon$ -decay = 0.9995)

630

631

632 === Q-learning hyperparameter values Final Comparison Table ===

633

Parameter	Value	Success rate (%)	Cumulative reward	Note
Learning rate (alpha)	0.05	98.00	380.80	
Learning rate (alpha)	0.10	100.00	226.20	
Learning rate (alpha)	0.15	100.00	234.90	Selected configuration
Learning rate (alpha)	0.20	100.00	237.10	
Learning rate (alpha)	0.30	100.00	256.80	

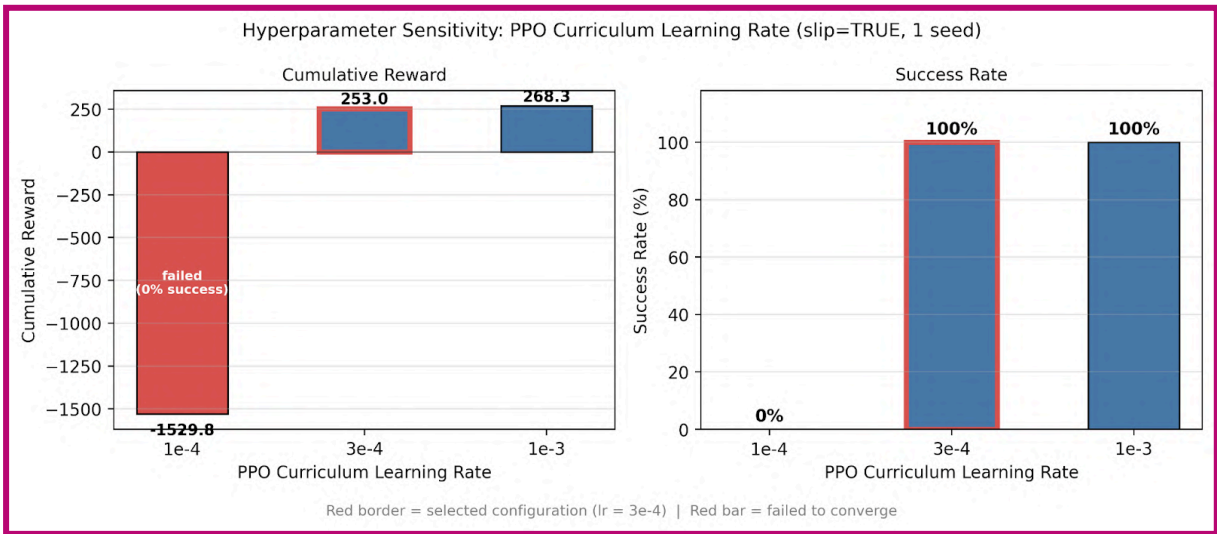
Epsilon decay	0.9990	100.00	240.80	
Epsilon decay	0.9995	100.00	228.20	Selected configuration
Epsilon decay	0.9998	100.00	229.80	
Epsilon decay	0.9999	100.00	217.20	

634 Table 4: presents a sensitivity analysis of Q-learning's key hyperparameters. Cumulative rewards  
635 are reported to assess overall performance.

636

637

638



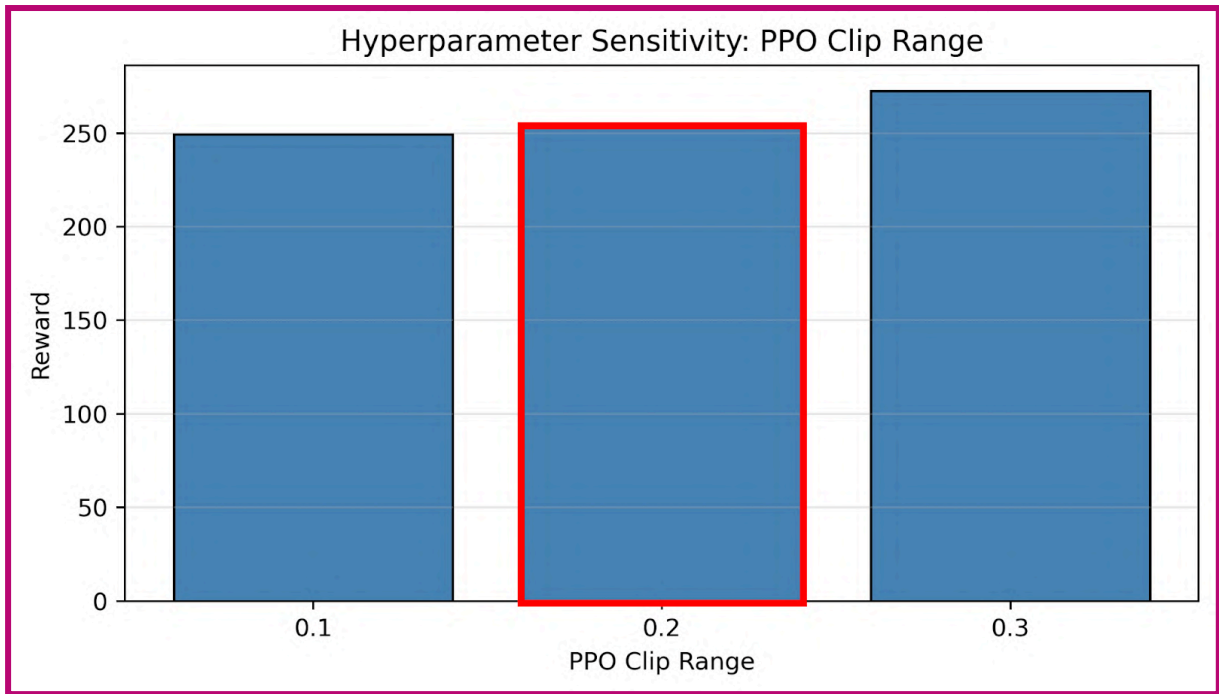
639

640

641 Figure 15: Comparison of different values of learning rate and the highlighted values  
642 represents the selected hyperparameter configuration, chosen based on their balance of  
643 performance and training stability ( Learning rate =  $3 \times 10^{-4}$ ). lr =  $1 \times 10^{-4}$  failed to  
644 converge (0% success rate, reward = -1529.8), a very small learning rate leads to  
645 insufficient policy updates, preventing convergence within the training horizon.

646

647



648

649

650 Figure 16: Comparison of different values of PPO clip range and the highlighted values  
 651 represents the selected hyperparameter configuration, chosen based on their balance of  
 652 performance and training stability. ( Clip range = 0.2)

653

654

655 ===PPO hyperparameter values Final Comparison Table ===

656

Parameter	Value	Success rate (%)	Cumulative reward	Note
Clip range	0.1	100.00	249.40	
Clip range	0.2	100.00	253.00	Selected configuration
Clip range	0.3	100.00	272.50	
Learning rate	1e-04	0.00	-1529.80	
Learning rate	3e-04	100.00	253.00	Selected configuration
Learning rate	1e-03	100.00	268.30	

657

658 Table 5: presents a sensitivity analysis of PPO's key hyperparameters. Cumulative rewards and  
 659 success rates are reported to assess overall performance.

660

661

662 Across all tested configurations, the rank ordering of methods remained consistent,  
 663 confirming that the reported conclusions are robust to moderate hyperparameter  
 664 variation.

665

## 666 **5.65 Limitations and Future Directions**

667 Several limitations should be noted when interpreting these findings. All experiments  
668 were carried out in a simulated 50×50 grid environment with a relatively small and  
669 simple state space, which differs significantly from real-world scenarios. This setup  
670 likely favored value-based methods like Q-learning, since it could learn precise  
671 state-action values for every cell. In contrast, PPO depends on neural network function  
672 approximation and stochastic policy updates, which may need more training steps or  
673 data to perform well in such discrete, small-scale settings.

674

675 To improve statistical reliability, all experiments were run independently over 105  
676 random seeds and results were averaged across runs. Noticeable performance  
677 variability was observed across training runs, mainly due to the stochastic environment,  
678 random action outcomes, and probabilistic transitions under slip conditions.

679 The simulated environment simplified real-world winter driving conditions and did not  
680 capture vehicle dynamics, sensor noise, road-surface variation, or realistic  
681 energy-consumption models, these are factors that might lead to changes. As a result,  
682 these findings to real-world winter navigation scenarios were limited [129].

683 Future research could evaluate reinforcement learning agents in larger scale,  
684 continuous, or more realistic winter navigation environments with higher-dimensional  
685 state spaces and more complex conditions. Then it may be better to reflect real-world  
686 uncertainty and could allow policy-gradient methods such as PPO to fully function their  
687 theoretical advantages in handling stochastic and high-dimensional control problems.

688

689

---

690

## 691 **6. Conclusion**

692 This study compared a traditional BFS baseline with two reinforcement learning  
693 methods—Q-learning and Proximal Policy Optimization (PPO)—for energy-efficient  
694 navigation in a 50×50 winter grid under both deterministic (non-slip) and stochastic  
695 (slip) conditions. The averaged results across 105 independent random seeds partially  
696 supported the original hypothesis. Under deterministic (non-slip) conditions, PPO  
697 achieved the highest cumulative reward and fewest snow cell visits, outperforming  
698 Q-learning. However, under stochastic (slip) conditions, Q-learning outperformed PPO  
699 in cumulative reward and snow-cell avoidance. Although both reinforcement learning  
700 methods clearly outperformed the BFS baseline, BFS struggled under slip conditions  
701 due to its inability to adapt to stochastic transitions.

702

703 These findings show that whether value-based methods or policy-gradient methods  
704 perform better really depends on how unpredictable the environment is. In our small,  
705 structured discrete grid, PPO gained an advantage from its stable gradients and

706 curriculum learning when everything was deterministic. On the other hand,  
707 Q-learning's simple tabular updates turned out to be more robust when the roads  
708 became slippery and actions sometimes failed. More generally, the results emphasize  
709 that the choice of reinforcement learning method should be guided by the structure and  
710 constraints of the task, rather than by theoretical advantages alone. Future work could  
711 explore whether PPO's advantages extend to more complex environments with  
712 continuous states or larger state spaces, where its policy-gradient approach may better  
713 demonstrate its theoretical strengths.

714

715

716

717

718

---

719

## 720 Reference

721

722 [1]Mao, R., Xu, W., Qian, Y., Li, X., Li, Y., Li, G., & Zhang, H. (2025).

723 Understanding the Determinants of Electric Vehicle Range: A Multi-  
724 Dimensional Survey. *Sustainability*, 17(10), 4259.

725 <https://www.mdpi.com/2071-1050/17/10/4259> <https://doi.org/10.3390/su17104259>

726 [10.3390/su17104259](https://doi.org/10.3390/su17104259)

727 [2]Carlson, A., & Vieira, T. (2021). *The effect of water and snow on the*  
728 *road surface on rolling resistance* (VTI Report 971A). Swedish National  
729 Road and Transport Research

730 Institute.

731 [https://www.researchgate.net/publication/350690120\\_The\\_effect\\_of\\_water\\_and\\_snow\\_on\\_the\\_road\\_surface\\_on\\_rolling\\_resistance](https://www.researchgate.net/publication/350690120_The_effect_of_water_and_snow_on_the_road_surface_on_rolling_resistance) [https://www.diva-portal.org/smash](https://www.diva-portal.org/smash/get/diva2:1542142/fulltext01.pdf)

732 [/get/diva2:1542142/](https://www.diva-portal.org/smash/get/diva2:1542142/fulltext01.pdf)

733 [FULLTEXT01.pdf](https://www.diva-portal.org/smash/get/diva2:1542142/fulltext01.pdf)

734 [FULLTEXT01.pdf](https://www.diva-portal.org/smash/get/diva2:1542142/fulltext01.pdf)  
735 [3]Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An  
736 Introduction.

737 [https://web.stanford.edu/class/psych209/Readings/](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)

738 [SuttonBartoIPRLBook2ndEd.pdf](https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf)

739 [4]Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino,  
740 A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., &  
741 Hadsell, R. (2017). *Learning to Navigate in Complex Environments* (No.  
742 arXiv:1611.03673). arXiv. <https://doi.org/10.48550/arXiv.1611.03673>

743 [5]Tan, C. (2025). Comparative Study of Reinforcement Learning

744 Performance Based on PPO and DQN Algorithms. *Applied and*  
745 *Computational Engineering*, 175(1), 30–36.

746 <https://doi.org/10.54254/2755-2721/2025.AST24879>

747 [6]Warnakulasuriya, D. A., Plosila, J., & Haghbayan, H. (2025). Energy-Efficient Path  
748 Planning in Uneven Terrains Using Adaptive Reinforcement Learning. *IEEE Conference*  
749 *Publication*. <https://ieeexplore.ieee.org/document/11093435>  
750 [7]Watkins, C.J.C.H., & Dayan, P.(1992). Q-learning. *Mach Learn* 8, 279–292.  
751 <https://doi.org/10.1007/BF00992698>  
752 [8]Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O.  
753 (2017). *Proximal Policy Optimization Algorithms* (No.arXiv:1707.06347). arXiv.  
754 <https://doi.org/10.48550/arXiv.1707.06347>  
755 [9]Pendyala, A., Atamna, A., & Glasmachers, T. (2024). Solving a Real-World Optimization  
756 Problem Using Proximal Policy Optimization with Curriculum Learning and Reward  
757 Engineering. arXiv:2404.02577.<https://arxiv.org/abs/2404.02577>  
758 [10]Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and  
759 Reinforcement Learning. *Neuron*, 36(2), 285–298.  
760 [https://doi.org/10.1016/S0896-6273\(02\)00963-7](https://doi.org/10.1016/S0896-6273(02)00963-7)  
761 [11]Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009).  
762 Curriculum learning. *Proceedings of the 26th Annual International*  
763 *Conference on Machine Learning*, 41–48. <https://doi.org/10.1145/1553374.1553380>  
764 [12]Chukwurah, N., Adebayo, A. S., Ajayi, O. O., & Anfo Pub. (2024).  
765 *Sim-to-Real Transfer in Robotics: Addressing the Gap between*  
766 *Simulation and Real- World Performance*. *International Journal for Multidisciplinary*  
767 *Research (IJFMR)*, 05(01), 33–39. <https://doi.org/10.54660/.IJFMR.2024.5.1.33-39>  
768  
769  
770 ~~[3]Watkins, C.J.C.H., & Dayan, P.(1992). Q learning. *Mach Learn* 8, 279–292.¶~~  
771 ~~<https://doi.org/10.1007/BF00992698>¶~~  
772 ~~[4]Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O.¶~~  
773 ~~(2017). *Proximal Policy Optimization Algorithms* (No.¶~~  
774 ~~arXiv:1707.06347). arXiv. <https://doi.org/10.48550/arXiv.1707.06347>¶~~  
775 ~~[5]Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An¶*  
776 ~~*Introduction*.¶~~  
777 ~~<https://web.stanford.edu/class/psych209/Readings/>¶~~  
778 ~~<SuttonBartoIPRLBook2ndEd.pdf>¶~~  
779 ~~[6]Dayan, P., & Balleine, B. W. (2002). Reward, Motivation, and¶~~  
780 ~~Reinforcement Learning. *Neuron*, 36(2), 285–298. <https://doi.org/>¶~~  
781 ~~[10.1016/S0896-6273\(02\)00963-7](10.1016/S0896-6273(02)00963-7)¶~~  
782 ~~[7]Tan, C. (2025). Comparative Study of Reinforcement Learning¶~~  
783 ~~Performance Based on PPO and DQN Algorithms. *Applied and¶*  
784 *Computational Engineering*, 175(1), 30–36. <https://doi.org/>¶~~  
785 ~~<10.54254/2755-2721/2025.AST24879>¶~~  
786 ~~[8]Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino,¶~~  
787 ~~A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., &¶~~  
788 ~~Hadsell, R. (2017). *Learning to Navigate in Complex Environments* (No.¶~~  
789 ~~arXiv:1611.03673). arXiv. <https://doi.org/10.48550/arXiv.1611.03673>¶~~~~

790 [9]Chukwurah, N., Adebayo, A. S., Ajayi, O. O., & Anfo Pub. (2024).  
791 Sim to Real Transfer in Robotics: Addressing the Gap between  
792 Simulation and Real World Performance. *International Journal for Multidisciplinary*  
793 *Research (IJFMR)*, 05(01), 33–39. [https://doi.org/](https://doi.org/10.54660/IJFMR.2024.5.1.33-39)  
794 [10.54660/IJFMR.2024.5.1.33-39](https://doi.org/10.54660/IJFMR.2024.5.1.33-39)  
795 [10]Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009).  
796 Curriculum learning. *Proceedings of the 26th Annual International*  
797 *Conference on Machine Learning*, 41–48. [https://doi.org/](https://doi.org/10.1145/1553374.1553380)  
798 [10.1145/1553374.1553380](https://doi.org/10.1145/1553374.1553380)  
799 [11]Pendyala, A., Atamna, A., & Glasmachers, T. (2024). Solving a Real World  
800 Optimization Problem Using Proximal Policy Optimization with Curriculum Learning  
801 and Reward Engineering. *arXiv:2404.02577*.<https://arxiv.org/abs/2404.02577>  
802 [12]Warnakulasuriya, D. A., Plosila, J., & Haghbayan, H. (2025). Energy Efficient Path  
803 Planning in Uneven Terrains Using Adaptive Reinforcement Learning. *IEEE Conference*  
804 *Publication*. <https://ieeexplore.ieee.org/document/11093435>

805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820

## 821 Appendices

### 822 Appendix A Experimental Configuration and Algorithmic Details:

823

#### 824 A.1 Environment Configuration

825 All experiments were conducted in a custom 50×50 winter grid environment (2,500  
826 discrete states).

- 827 • Start position: (47, 2)

- 828 • Goal position: (20, 45)
- 829 • Maximum steps per episode: 1000
- 830 • Action space: {up, down, left, right}

831 Two transition settings were evaluated:

832 Deterministic (slip = FALSE)

833 The intended action is executed exactly.

834 Stochastic (slip = TRUE)

835 With probability  $\frac{1}{3}$ , the intended action is executed.

836 With probability  $\frac{2}{3}$ , the agent slips to a random adjacent direction.

837 All algorithms were evaluated on the same fixed grid layout to ensure fairness.

838

### 839 A.2 Reward Function

840 The reward function models energy-aware winter navigation.

841 At each time step:

- 842 • Step penalty: -1.5
- 843 • Snow penalties:
  - 844 ○ Near snow: -0.2
  - 845 ○ Edge snow: -0.5
  - 846 ○ Core snow: -2.0
- 847 • Goal reward: +700

848 The cumulative episode reward is:

$$849 R = \sum_{t=1}^T (-1.5 - C_{snow}(s_t)) + 700 \cdot 1_{goal\ reached}$$

850

851 where  $C_{snow}(s_t)$  denotes the terrain penalty and

852  $1_{goal\ reached}$  indicates successful termination.

### 853 A.3 Q-Learning Configuration

854 Tabular Q-learning was implemented with the following hyperparameters:

- 855 • Learning rate  $\alpha=0.15$
- 856 • Discount factor  $\gamma=0.99$
- 857 • Exploration strategy:  $\epsilon$ -greedy
- 858 • Initial  $\epsilon = 0.60$
- 859 • Minimum  $\epsilon = 0.05$
- 860 • Exponential decay per episode = 0.9995
- 861 • Training episodes = 15,000
- 862 • Max steps per episode = 1000

863 The Q-update rule is:

$$864 Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

865 During evaluation, a fully greedy policy ( $\epsilon = 0$ ) was used.

866

#### 867 A.4 Proximal Policy Optimization (PPO) Curriculum Configuration

868 PPO was implemented as a stochastic policy-gradient method.

869 Total training timesteps: 1,200,000

- 870 • Phase 1 (navigation-focused reward): 300,000 timesteps
- 871 • Phase 2 (energy-aware reward): 900,000 timesteps

872

873 Evaluation was conducted using deterministic action selection.

874

875 PPO Objective Function

876 PPO optimizes the clipped surrogate objective:

$$877 L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)]$$

878 where

$$879 r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

880 and  $\hat{A}_t$  is the generalized advantage estimate.

881 The clipping mechanism constrains policy updates to maintain stability.

882

### 883 PPO Hyperparameters

- 884 • Discount factor  $\gamma=0.99$
- 885 • GAE parameter  $\lambda=0.95$
- 886 • Entropy regularization enabled
  
- 887 • Learning rate:  $3 \times 10^{-4}$
  
- 888 • Clip range: 0.2
  
- 889 • Rollout buffer: 2,048 steps
  
- 890 • Batch size: 256
  
- 891 • Optimization epochs: 10
  
- 892 • Neural network architecture: two hidden layers (256 units each, ReLU activation)

893

### 894 A.5 Evaluation Protocol

895 For each condition (slip FALSE / TRUE):

- 896 • One trained model per algorithm
- 897 • Maximum evaluation length: 1000 steps
- 898 • Number of evaluation episodes: 200
- 899 • Number of seeds: 105 ¶
- 900 • Heatmap runs: 10
- 901 • Metrics recorded:
  - 902 ○ Total cumulative reward
  - 903 ○ Number of steps
  - 904 ○ Snow cell visits
  - 905 ○ Success rate (%)

906 State visitation heatmaps were generated by aggregating visitation frequencies across  
907 evaluation runs of 200 episodes each.

908

---

## 909 Appendix B Detailed Quantitative Results:

### 910 B.1 Deterministic (slip = FALSE) ¶

Method	Reward	Steps	Snow Visits	Success
<del>BFS Shortest</del>	<del>575.30 ± 0.00</del>	<del>70.00 ± 0.00</del>	<del>31.00 ± 0.00</del>	<del>100.0%</del>
<del>Q-Learning</del>	<del>579.10 ± 5.61</del>	<del>70.00 ± 0.00</del>	<del>12.60 ± 3.29</del>	<del>100.0%</del>
<del>PPO Curriculum</del>	<del>588.50 ± 0.60</del>	<del>70.00 ± 0.00</del>	<del>7.80 ± 1.10</del>	<del>100.0%</del>

911

Method	Reward (mean±SD)	95%CI	Steps	Snow Visits	Success
BFS Shortest	575.30 ± 0.00	[575.30, 575.30]	70.00 ± 0.00	31.00 ± 0.00	100.0%
Q-Learning	582.01 ± 6.97	[577.02, 587.00]	70.00 ± 0.00	13.10 ± 3.05	100.0%
PPO Curriculum	588.73 ± 0.67	[588.25, 589.21]	70.00 ± 0.00	7.60 ± 0.97	100.0%

912

913

914 All methods reached the goal within 70 steps.

915 PPO achieved the highest cumulative reward and fewest snow cell visits.

916

917 B.2 Stochastic (slip = TRUE)

Method	Reward (mean±SD)	95% CI	Steps	Snow Visits	Success
BFS Shortest	-1668.41 ± 113.42	[-1749.54, -1587.28]	1000.00 ± 0.00	246.80 ± 76.09	0.0%
Q-Learning	311.82 ± 4.83	[308.37, 315.27]	26.40 ± 3.92	66.94 ± 3.87	100.0%
PPO Curriculum	258.69 ± 10.73	[251.02, 266.37]	241.18 ± 4.11	91.88 ± 5.24	100.0%

918

919

Method	Reward	Steps	Snow Visits	Success
<del>BFS Shortest</del>	<del>-1693.96 ± 61.27</del>	<del>1000 ± 0.00</del>	<del>246.80 ± 76.09</del>	<del>0%</del>

<del>Q-Learning</del>	<del>304.03 ± 9.92</del>	<del>228.18 ± 4.10</del>	<del>67.58 ± 4.34</del>	<del>100.0%</del>
<del>PPO Curriculum</del>	<del>252.19 ± 7.04</del>	<del>242.29 ± 3.00</del>	<del>92.65 ± 4.14</del>	<del>100.0%</del>

920 **Under stochastic dynamics:**

- 921 • BFS fails due to inability to adapt.
- 922 • Q-learning demonstrates greater robustness.
- 923 • PPO maintains success but exhibits higher trajectory dispersion.

924

925 **B.3 State Visitation Analysis**

926 State visitation frequency was computed as:

927 
$$V(s) = \frac{N(s)}{\max_s N(s)}$$

928 where  $N(s)$  is the number of visits to state  $s$ .

- 929 • Under deterministic conditions, visitation concentrates along a narrow corridor.
- 930 • Under stochastic conditions, visitation becomes more dispersed.
- 931 • Q-learning exhibits more focused routing than PPO under slip conditions.

Dear Managing Editor,

Thank you for your thoughtful comments on my research paper. I have carefully addressed each of your comments. All the data results have been updated to the 10 seed results.

**Comment 1:**

1. While the revision introduces multi-seed experiments using 5 seeds, this could be strengthened for stochastic RL evaluation. There are also no tests for statistical significance, no confidence intervals, and no discussion of variability across runs.
  - a. I would like the author to increase the experiments to at least 10 (maybe even 20) seeds and report statistics (e.g., mean +/- standard deviation, median + 95% confidence intervals, etc.). Just something of this sort would be great and help alleviate concerns about the robustness and consistency of your results.
  - b. Include at least one statistical test (e.g., Welch's t-test) comparing Q-learning vs. PPO (slip condition)
  - c. Alongside these tests, you should add a short paragraph discussing and interpreting the variance. In particular, I would want you to answer whether PPO exhibits higher instability

**Response 1:**

(a) All experiments have re-run using 10 independent random seeds (seed0-9). Table 1 and 2 now are reported with mean +/- standard deviation, and 95% confidence intervals. Please view the result section.

(b) I have added a Welch T test table that compares Q-learning and PPO under slip conditions. The test confirmed that Q-learning significantly outperformed PPO ( $t = 14.28$ ,  $p < 0.001$ ). Results are reported in the new Section 4.4.1 (Statistical Significance) along with Table 3.

(c) A variance interpretation paragraph has been added. "The standard deviation gap between PPO in both conditions suggests that PPO is more sensitive to stochastic signals under slip condition, where high probability slip results in noisy advantage estimation, leading to higher variance across training runs ( $SD = 10.73$ ). In contrast, under deterministic conditions PPO exhibits a much lower variance ( $SD = 0.67$ ), confirming that the instability is determined by the stochastic environment, not the algorithm itself." Please view line (496-501) in 4.4.1 Welch's T-test section.

**Comment 2:**

The research gap still does not entirely distinguish between RL path planning literature, cost-aware navigation, and grid-world RL benchmarks. Could you add another paragraph answering more explicitly what prior work has been done, what is missing, and what exactly is novel.

**Response 2:**

A new paragraph has been added to Introduction that explicitly categorizes prior works into three streams. “Prior work on reinforcement learning for navigation can be categorized into three directions. First, Grid world benchmarks commonly adopt tabular Q-learning and deep reinforcement learning variants as standard baselines for evaluating discrete navigation tasks [3,5], but focus primarily on task completion rather than energy awareness and energy sensitive routing. Second, cost-aware navigation research has examined energy minimisation in uneven terrains using reinforcement learning [6], yet rarely incorporated weather dependent components, such as snow coverage or stochastic traction loss. Third, existing comparisons between value-based and policy-gradient methods [7,8], including curriculum based PPO applications [9] are conducted primarily under fixed and stable conditions, without examining the possibility of a particular reinforcement learning method remaining dominant to another when action outcomes become probabilistic, as in stochastic environments such as winter slip conditions. The work in this study addresses all three gaps simultaneously by implementing an energy-aware reward structure into a reproducible 50×50 grid-world benchmark, embedding snow-density penalties and a 2/3-slip probability to simulate realistic winter routing conditions, and directly comparing Q-learning, PPO, and a traditional baseline under both deterministic and stochastic environments.” Please view line (73-89) in introduction section.

### **Comment 3:**

Please add a plot of learning curves: training reward vs. timesteps for both Q-learning and PPO, along with the statistics from point 1 above (CIs, mean +/- stds, whatever you end up using). Would reveal instability, too, especially for PPO?

### **Response 3:**

Learning curves have been added . Figure 13 shows mean episode reward  $\pm 1$  SD across 10 seeds under slip conditions for both methods. Please view the 4.4.2 section in results section.

### **Comment 4:**

The paper states outcomes (e.g., Q-learning outperforms PPO) but could explain these in a deeper, more mechanistic way. Why does Q-learning work better in discrete grid worlds? Why does PPO struggle?

a. For the first question, consider: exact value iteration in finite MDP, lower variance than policy gradients

b. For the second question, consider: high variance gradients under stochasticity, issues with sparse and noisy rewards

### **Response 4:**

(a) I have expanded Section 5.2 to address this by explaining that Q-learning benefits from exact value iteration in the finite MDP, directly converging to true optimal Q-values without approximation error, and produces lower variance than

policy gradient methods since updates are based on individual state-action pairs rather than entire trajectories [7].

- (b) PPO struggles due to high variance gradients under stochasticity, where the 2/3-slip probability inflates advantage estimates across rollouts, combined with sparse and noisy reward signals that destabilize gradient updates and require significantly more training samples [3]
- (c) Please view section 5.2 in the Discussion section.

#### **Comment 5:**

All results are based on a single hyperparameter configuration per method, which is concerning. How stable are the results to changes in hyperparameters? Could you add at least one sensitivity analysis? You do not need to go overboard, but this could be good to alleviate skepticism towards claims about one method being more suited over the other...

- a. For Q-learning: vary the decay or learning rate
- b. For PPO: vary learning or clipping parameter
- c. This should be accompanied by a figure

#### **Response 5:**

A full sensitivity analysis has been added in the new Section 5.5, with Tables 4 and 5 and Figures 14–16 that includes decay and learning rate for Q-learning, and learning rate and clipping range for PPO. Please view section 5.5 in the Discussion section.

#### **Comment 6:**

Since curriculum is highlighted, it should ideally be validated. While I like how you introduced the curriculum learning setup, I am still wondering if the PPO performance

depends heavily on the curriculum. Could you compare the PPO with and without the curriculum, or at least cite some literature that would answer this question? You may want to accompany this with some sort of ablation study plot(s)

**Response 6:**

I have added an ablation comparison table compares between PPO with curriculum learning and without in both conditions. Please view section 3.2.6 in the Methods section.

**Comment 7:**

Watch the informal phrasing throughout (e.g., the use of “etc.”)

**Response 7:**

The use of informal word for etc. had been removed across the whole manuscript. Also, added “Sub-zero temperatures, combined with snow and ice accumulation on road surfaces, significantly increase energy consumption, resulting in reduced driving range and greater unpredictability in trip planning “to make the manuscript more formal.

**Comment 8:**

I would add at least 2 more papers on PPO vs. value-based methods and/or RL navigation benchmarks

**Response 8:**

I have added two articles that support my manuscript. Which are [11]Pendyala, A., Atamna, A., & Glasmachers, T. (2024). Solving a Real-World Optimization Problem Using Proximal Policy Optimization with Curriculum Learning and Reward Engineering. arXiv:2404.02577.<https://arxiv.org/abs/2404.02577>

and

[12]Warnakulasuriya, D. A., Plosila, J., & Haghbayan, H. (2025). Energy-Efficient Path Planning in Uneven Terrains Using Adaptive Reinforcement Learning. *IEEE Conference Publication*.<https://ieeexplore.ieee.org/document/11093435>