

Article

Dual-Agent Hierarchical Reinforcement Learning for Typhoon-Avoidance Route Planning of Ships Under Dynamic Wind–Wave–Current Fields

Yu Cai ^{1,2,3,4} , Ying Li ^{1,2,3,4} , Liankang Zhang ^{1,2,3,4} and Jun Song ^{1,2,3,4,*}

¹ Operational Oceanography Institute (OOI), Dalian Ocean University, Dalian 116023, China; caiyu@dlou.edu.cn (Y.C.); 19836105241@163.com (Y.L.); 15261341951@163.com (L.Z.)

² College of Marine Science Technology and Environment, Dalian Ocean University, Dalian 116023, China

³ Liaoning Key Laboratory of Marine Real-Time Warning, Dalian 116023, China

⁴ Dalian Xinghai Bay Laboratory, Dalian 116023, China

* Correspondence: songjun2017@dlou.edu.cn

Abstract

Ship weather routing in regional seas under severe seasonal weather systems, such as typhoons, poses a critical operational challenge for maritime safety and efficiency. Traditional single-agent reinforcement learning (RL) methods frequently suffer from training instabilities and erratic trajectory adjustments when exposed to the highly non-stationary, multi-scale dynamics of coupled wind–wave–current fields. To address these limitations, this study proposes a dual-agent hierarchical reinforcement learning framework (HRL-MOO) featuring a built-in dynamic risk-weight adaptation mechanism. The proposed global agent discerns large-scale environmental evolutions and adaptively updates the relative weights of wind-, wave-, and current-induced risks at a strategic level, while the local agent translates this macro-level guidance into short-term tactical heading and speed adjustments within realistic vessel motion boundaries. The framework incorporates bathymetric constraints through a high-resolution navigable domain mask derived from ETOPO topography to guarantee practical navigability. Simulated experiments are executed using hourly environmental reanalysis data and best-track records corresponding to the passage of Typhoon Yagi (2024) through the northern South China Sea and Taiwan Strait. The empirical results demonstrate that the cooperative dual-agent structure establishes a superior global Pareto frontier compared to conventional standard single-agent Proximal Policy Optimization (PPO) and metaheuristic baselines. Under extreme typhoon conditions, the HRL-MOO model effectively decreases the cumulative environment-induced risk from approximately 0.17 to 0.12, improves path smoothness by achieving a higher index of 0.842, and accelerates policy convergence to within 3000 training episodes, all while maintaining a highly efficient voyage distance with minimal detour overhead. Ablation studies further validate that the synergy between macro-level stream field recognition and adaptive multi-objective optimization significantly enhances decision-making stability. This framework offers a robust and interpretable computational tool for autonomous ship weather routing in fast-changing, high-risk ocean environments.

Keywords: ship weather routing; hierarchical reinforcement learning; wind–wave–current field; typhoon avoidance; multi-objective optimization; applied oceanography



Academic Editor:

Piotr Zwierzykowski

Received: 9 July 2026

Revised: 24 July 2026

Accepted: 24 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Maritime navigation in regions affected by strong seasonal weather systems remains a persistent operational challenge. In areas such as the South China Sea and the Taiwan Strait, typhoons frequently reshape the surrounding wind, wave, and current fields within only a few hours, creating conditions that are difficult for ships to anticipate or avoid. These waters are also important commercial corridors, which means that even small improvements in routing reliability can have meaningful practical impact. Recent studies have highlighted the sensitivity of ship routes to short-term environmental variability, particularly under severe weather disturbances [1,2]. At the network level, climate extremes can disrupt ports and shipping routes, potentially causing cascading effects across maritime transport networks and highlighting the need for climate-resilient route planning [3]. Weather-routing studies in severe environments have also emphasized the importance of ship-performance modeling and changing environmental constraints [4].

Earlier research on ship routing mainly focused on deterministic algorithms such as Dijkstra, A*, and dynamic programming. Artificial-potential-field and deep-reinforcement-learning approaches have been introduced to improve routing flexibility [5], while multi-objective vessel-routing methods have further extended these efforts [6,7]. Despite their advantages, these methods rely mostly on fixed objective weights. As a result, they struggle to account for the fact that wind, wave, and current influences may dominate at different times during a storm's progression.

Reinforcement learning (RL) has become increasingly attractive for such scenarios because it learns navigation policies directly through interaction with the environment. RL methods have been used for voyage optimization [8–10] and for USV path planning using DQN-based and hybrid approaches [11–13]. Most of these studies treat the navigation problem with a single agent. While this works in relatively stable conditions, a single decision-maker tends to focus either too narrowly on local information or too broadly on long-term goals, making it difficult to maintain consistent performance when the environment varies across several spatial and temporal scales.

Hierarchical reinforcement learning (HRL) offers an alternative by dividing the task into a strategic component and an operational component. Recent applications have shown promising results in inland waterways [14], in multi-USV coordination [15], and in routing under environmental constraints [16]. Still, many of these demonstrations rely on simplified settings. Typhoon-driven sea states involve complex interactions between wind, wave, and current fields that evolve at different rates, introducing a level of variability that existing HRL applications have not fully addressed.

Recent studies on cooperative multi-agent optimal control have shown that stabilizing policy-iteration frameworks and inverse-reinforcement-learning methods can improve coordinated decision-making under unknown dynamics, switching topologies, and fault-tolerant conditions [17,18]. These studies provide useful methodological insights for designing structured hierarchical learning frameworks in dynamic route-planning environments.

Multi-objective optimization (MOO) remains an essential part of route planning, especially when balancing distance, fuel use, and safety [19]. Recent studies [6,20] point toward increasingly sophisticated computational tools, and some recent studies have incorporated forecast uncertainty into typhoon-avoidance routing [2]. Yet even with these advancements, most MOO formulations still rely on static weight choices, which can misrepresent the relative importance of different hazards as the storm evolves.

Motivated by these gaps, the present study proposes a dual-agent hierarchical reinforcement learning framework with a built-in mechanism for dynamically updating the relative weights of wind-, wave-, and current-induced risks. The global agent evaluates the evolving sea state and adjusts the risk weighting accordingly, while the local agent

manages short-term heading and speed decisions. By combining reanalysis datasets of wind, wave, and current fields with bathymetric constraints, the framework aims to generate route-planning policies that maintain both interpretability and robustness under the rapidly changing conditions characteristic of typhoon-influenced waters.

2. Data and Study Area

The study focuses on a section of the northern South China Sea and the Taiwan Strait, covering the region between 15° N–25° N and 110° E–125° E. This area is heavily used by commercial vessels and is known for rapid changes in wind, wave, and current conditions during the typhoon season. The spatial extent used in this work is shown in Figure 1.



Figure 1. Study area map.

The experimental period spans 30 August–10 September 2024, during which Typhoon Yagi (2024) passed through the domain. Its movement, illustrated in Figure 2, produced a sequence of evolving sea states that are well suited for testing how the navigation model responds when the surrounding conditions shift over short time scales.

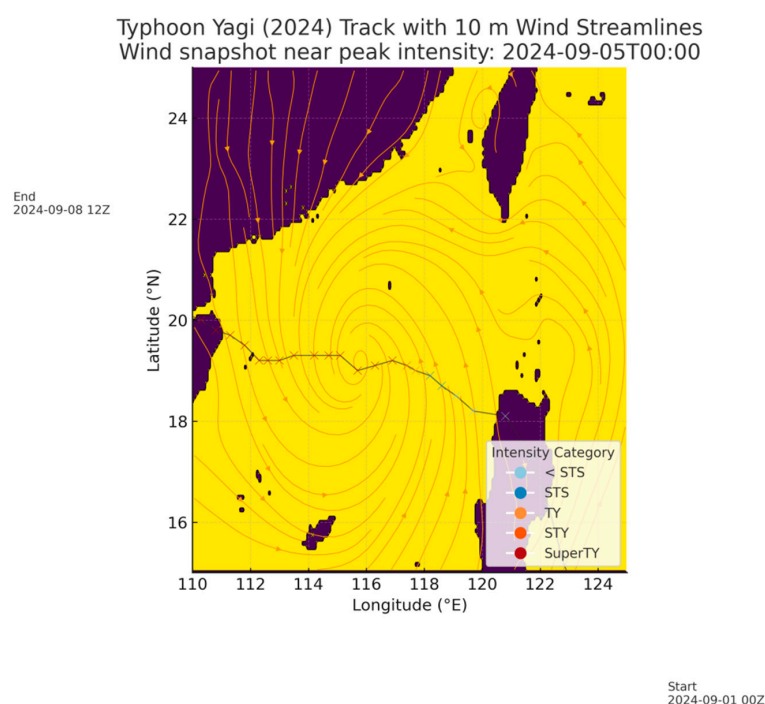


Figure 2. Schematic diagram of Typhoon Yagi's track.

To reproduce these conditions in a controlled setting, the study uses time-varying reanalysis data of wind, wave, and current fields, along with bathymetry that defines the navigable region. These datasets provide the foundation for the environmental fields and constraints used throughout both the training and evaluation of the proposed model.

2.1. Data Sources

To realistically reproduce the environmental conditions affecting ship routing and reconstruct the evolving sea state, several data products were combined. The environmental variables include 10-m wind components u_{10} and v_{10} , significant wave height H_s , and surface current components u_{cur} and v_{cur} . These variables come from the CMEMS global reanalysis, which provides hourly coverage at a consistent spatiotemporal resolution, and they describe the dominant environmental forcing acting on a vessel exposed to a typhoon.

Bathymetric and elevation information was obtained from the ETOPO global dataset, ensuring that shallow regions and land boundaries are correctly represented. This information is later used to generate the navigable mask.

Typhoon position and intensity were sourced from the CH2024BST best-track dataset at six-hour intervals. Although its temporal resolution differs from CMEMS, it offers a reliable description of the storm's trajectory and development.

All datasets were interpolated onto a $0.083^\circ \times 0.083^\circ$ grid. This preprocessing step avoids inconsistencies between datasets and ensures that the reinforcement learning agents interact with a unified environmental field.

2.2. Construction of the Wind–Wave–Current Risk Field

To quantify environmental risk, the wind-induced, wave-induced, and current-induced components are combined into a composite risk field:

$$R(x, y, t) = \omega_{w,t} F_w(x, y, t) + \omega_{s,t} F_s(x, y, t) + \omega_{c,t} F_c(x, y, t) \quad (1)$$

Here, $F_w(x, y, t)$, $F_s(x, y, t)$, and $F_c(x, y, t)$ denote the normalized wind-risk, wave-risk, and current-risk fields, respectively. The dynamic weights $\omega_{w,t}$, $\omega_{s,t}$, and $\omega_{c,t}$ represent the relative contributions of wind, wave, and current risks.

These weights are adaptively generated by the global agent according to the current environmental state and satisfy $\omega_{w,t} + \omega_{s,t} + \omega_{c,t} = 1$ and $\omega_{i,t} \geq 0$. The spatial distribution of the constructed wind–wave–current risk field is shown in Figure 3.

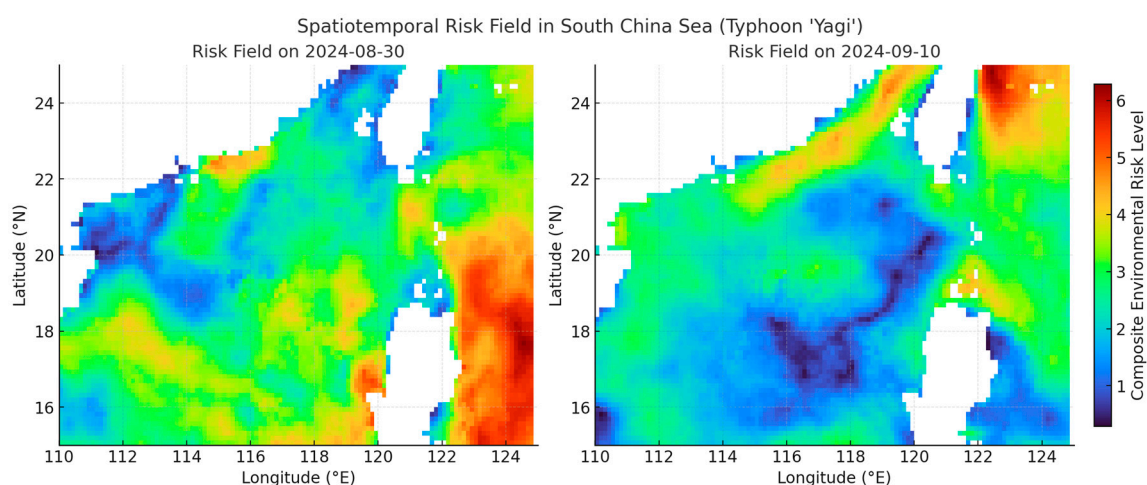


Figure 3. Spatial distribution of the wind–wave–current risk field.

2.3. Generation of Navigable Mask

Based on the ETOPO topographic data, a navigable mask is constructed to delineate navigable ocean regions. A minimum navigable depth threshold h_{thr} is defined as follows:

$$M(x, y) = \begin{cases} 0, & \text{if } depth(x, y) < h_{thr} \text{ or } elevation(x, y) > 0; \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

where h_{thr} denotes the minimum navigable depth threshold. This mask is fully aligned with the risk field in both spatial resolution and coordinates, thereby constraining the vessel's motion domain during training and simulation. The resulting navigable mask is shown in Figure 4.

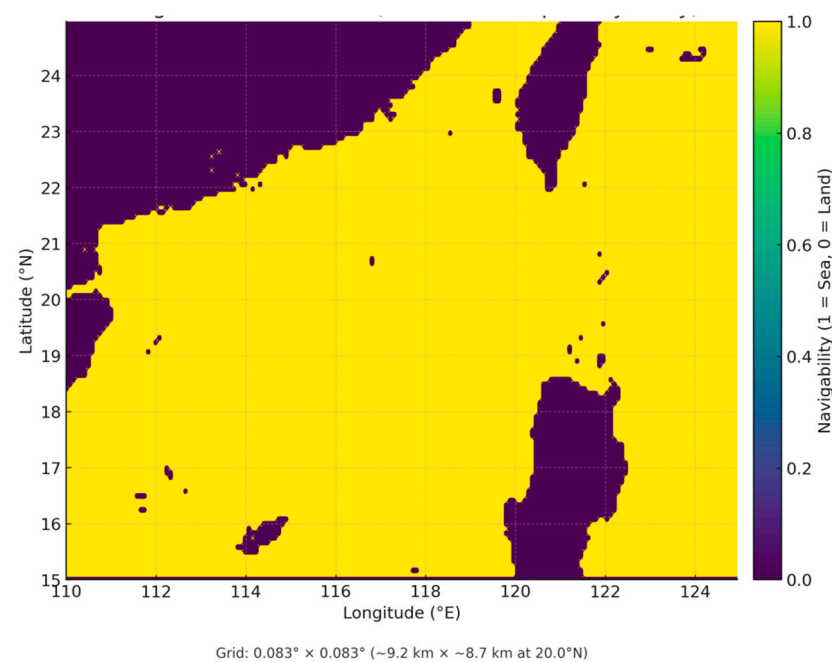


Figure 4. Spatial distribution of the navigable mask.

2.4. Integrated Environmental Dataset

To simulate conditions close to those encountered by a vessel during the passage of Typhoon Yagi, the different environmental components described above were combined into a single dataset. The composite risk field $R(x, y, t)$ provides the time-dependent representation of wind, wave, and current hazards, while the navigable mask $M(x, y)$ ensures that the vessel remains within physically feasible regions. These fields share the same spatial resolution of $0.083^\circ \times 0.083^\circ$, which helps avoid inconsistencies when the agents interact with the environment. The start and end points of each route were placed directly onto this grid, and the best-track information supplied the typhoon's position at each six-hour interval. By aligning the best-track data with the reanalysis fields, the model experiences a continuous sequence of environmental changes that reflect the storm's development. At each decision step, the agents receive updated values of $R(x, y, t)$ and the corresponding mask constraints, along with the vessel's current state. This setup forms a spatiotemporal environment in which long-range environmental evolution and local navigational constraints appear together in a consistent way. The combined dataset serves both the training and testing phases. Since all inputs evolve according to the same temporal axis, the agents encounter the natural progression of the typhoon as they learn. This structure also reduces numerical irregularities that would otherwise arise if the data were drawn from sources with mismatched resolutions or time steps. As a result, the

environment used in the simulations retains the essential dynamics of the real sea state while still providing a stable platform for reinforcement learning.

3. Methodology and Innovation

3.1. Overall Framework

The proposed HRL-MOO framework is formulated as a hierarchical Markov decision process (HMDP) for dynamic ship routing under typhoon-induced wind, wave, and current risks. At each decision step, the environment state is represented by vessel position, vessel speed, vessel heading, wind–wave–current risk fields, the navigability mask, and the relative position of the destination. This formulation separates large-scale risk prioritization from local vessel-control decisions while preserving a shared environmental transition process.

The global agent is responsible for extracting large-scale environmental information and generating dynamic risk weights. Its input consists of the coarse wind-induced, wave-induced, and current-induced risk fields, the navigability mask, the vessel state, and the destination-relative state. Its output is the dynamic risk-weight vector $\omega_t = [\omega_{w,t}, \omega_{s,t}, \omega_{c,t}]$, where $\omega_{w,t}$, $\omega_{s,t}$ and $\omega_{c,t}$ correspond to wind-induced risk, wave-induced risk, and current-induced risk, respectively. The weights satisfy $\omega_{w,t} + \omega_{s,t} + \omega_{c,t} = 1$ and $\omega_{i,t} \geq 0$.

The local agent is responsible for feasible short-horizon vessel control. Its input includes the current vessel position, speed, and heading, the local weighted risk field derived from the latest global risk weights, the local navigability mask, and the destination-relative position. Its output is the local control action $a_t^L = [\Delta\psi_t, \Delta v_t]$, where $\Delta\psi_t$ denotes the heading adjustment and Δv_t denotes the speed adjustment. These actions are constrained by navigational feasibility, including $[-30^\circ, +30^\circ]$ and 5–16 knots.

The information transfer between the two agents follows a synchronized global-local interaction protocol. In the present experiments, $K = 1$; therefore, the global agent updates the risk-weight vector at every decision step, and the local agent simultaneously executes a vessel-control action at every decision step. Thus, the information flow is defined as environmental fields and vessel state $\Rightarrow$ global risk-weight adaptation $\Rightarrow$ local risk-aware control $\Rightarrow$ environment transition and reward feedback.

Figure 5 summarizes the information flow of the proposed framework: environmental fields $\Rightarrow$ global agent $\Rightarrow$ dynamic risk weights $\Rightarrow$ local agent $\Rightarrow$ vessel-control action $\Rightarrow$ environment update $\Rightarrow$ reward feedback. The global agent therefore provides adaptive risk weighting, while the local agent transforms the weighted risk information and vessel state into feasible navigation actions.

3.2. Dynamic Risk-Weight Adaptation Mechanism

Typhoon-induced maritime risk is time-varying and heterogeneous across wind, wave, and current components. To represent this non-stationary risk composition, the proposed framework uses dynamic risk weights instead of fixed manually assigned coefficients. The weighted risk field is defined as:

$$R_t(x, y) = \omega_{w,t} F_w(x, y, t) + \omega_{s,t} F_s(x, y, t) + \omega_{c,t} F_c(x, y, t) \quad (3)$$

$$\omega_t = \text{softmax}\left(f_G\left(s_t^G\right)\right) \omega_t = [\omega_{w,t}, \omega_{s,t}, \omega_{c,t}] \quad (4)$$

Here, s_t^G denotes the global state observed by the global agent, while $\omega_{w,t}$, $\omega_{s,t}$, and $\omega_{c,t}$ correspond to wind-induced risk, wave-induced risk, and current-induced risk, respectively. The softmax normalization ensures $\omega_{w,t} + \omega_{s,t} + \omega_{c,t} = 1$ and $\omega_{i,t} \geq 0$. This notation is used consistently for the dynamic wind–wave–current risk weights throughout the manuscript.

Dual-Agent Hierarchical Reinforcement Learning for Typhoon-Avoidance Route Planning

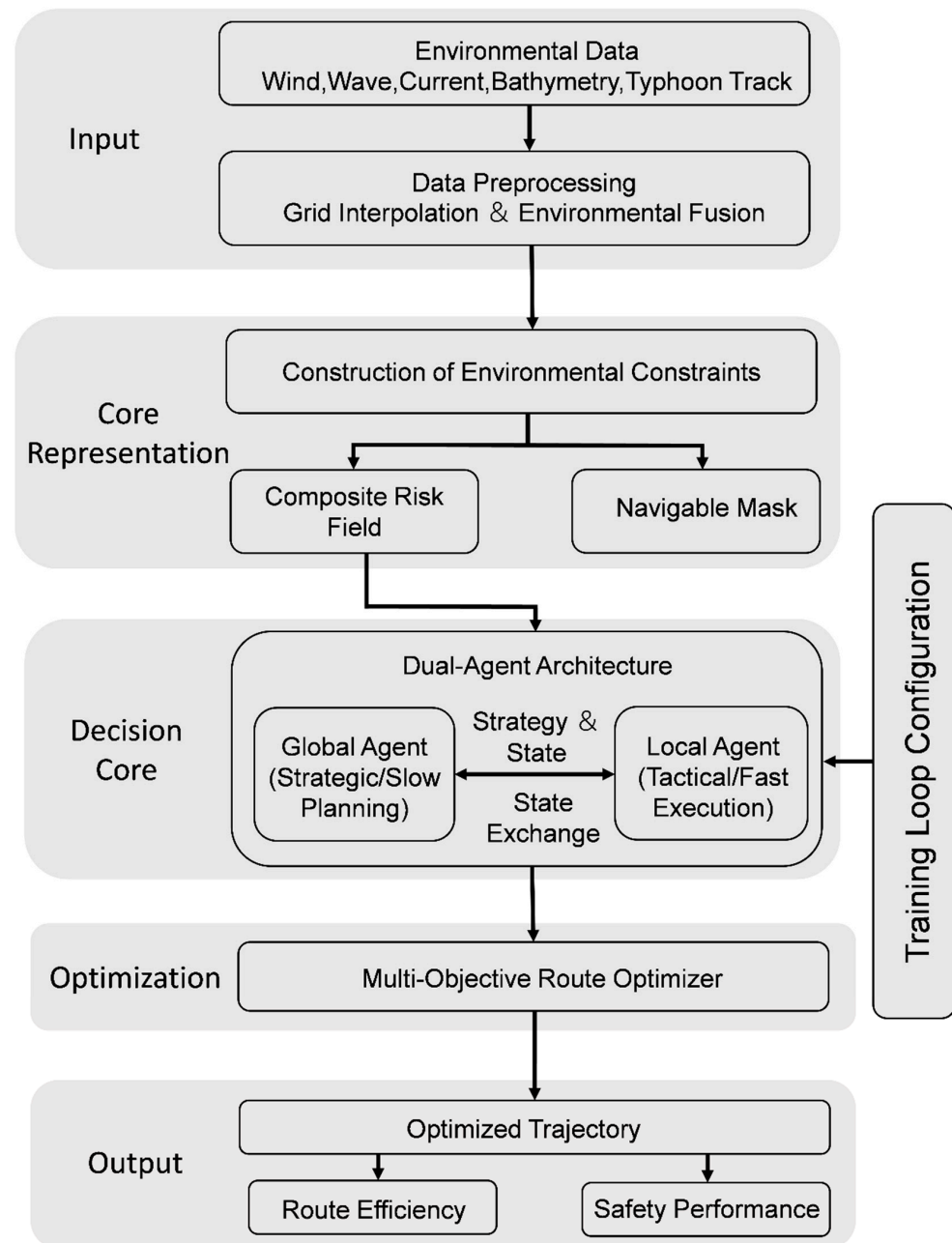


Figure 5. Overall architecture of the proposed HRL-MOO framework. The global agent extracts large-scale wind–wave–current risk information and outputs dynamic risk weights, while the local agent combines these weights with local vessel states and navigability constraints to generate feasible heading and speed adjustments.

The adaptive weighted risk field is then provided to the local agent as part of its decision state. Consequently, local vessel-control actions are selected according to both immediate navigability constraints and the current global assessment of wind–wave–current risk. This design improves the consistency between large-scale route guidance and short-horizon collision- and risk-avoidance behavior.

The mechanism is retained as a methodological component rather than as an additional experimental claim. It defines how environmental risk components are integrated during

policy learning and decision making while leaving all experimental settings and reported numerical results unchanged.

3.3. Dual-Agent Hierarchical Structure

The dual-agent structure decomposes the HMDP into a global risk-weight adaptation level and a local vessel-control level. This decomposition is designed to distinguish slowly varying large-scale typhoon risk patterns from rapidly changing local navigability constraints.

The global agent observes the large-scale environmental state and outputs the dynamic risk-weight vector ω_t . Rather than prescribing a coarse route directly, it modulates the relative importance of wind-induced, wave-induced, and current-induced risks. The resulting weights provide a compact supervisory signal for downstream local decision making.

The local agent operates at each decision step and selects the vessel-control action $a_t^L = [\Delta\psi_t, \Delta v_t]$. The action is computed from the vessel state, destination-relative state, local weighted risk field, navigability mask, and latest global weights. The selected heading adjustment is restricted to $[-30^\circ, +30^\circ]$, and the vessel speed is constrained to the range of 5–16 knots.

The two levels interact through a synchronized update protocol. In this study, $K = 1$ is adopted because no separate slower global update interval is specified in the original experimental configuration. Accordingly, ω_t and a_t^L are both updated at every decision step, which avoids inconsistency between strategic risk weighting and local vessel control. The detailed training procedure of the proposed HRL-MOO framework is summarized in Algorithm 1.

Algorithm 1. Training procedure of the proposed HRL-MOO framework

Step	Procedure
1	Initialize the maritime environment, including the start point, destination, wind–wave–current fields, navigability mask, and episode limits.
2	Initialize the global policy π_G and the local policy π_L , together with their value functions or policy-optimization parameters.
3	For each episode, reset the vessel state at the start point and initialize the cumulative reward.
4	For each local decision step t , update the environmental fields and the navigability mask according to the simulated typhoon state.
5	At every decision step ($K = 1$), the global agent observes the large-scale state and outputs $\omega_t = [\omega_{w,t}, \omega_{s,t}, \omega_{c,t}]$.
6	Construct the weighted risk field from the current wind-induced, wave-induced, and current-induced risk components.
7	The local agent observes the local vessel state, nearby weighted risk field, navigability constraints, destination-relative state, and latest global weights, and selects $a_t^L = [\Delta\psi_t, \Delta v_t]$.
8	Apply the feasible vessel-control action subject to navigation constraints, including $[-30^\circ, +30^\circ]$ and 5–16 knots.
9	Compute the transition to the next vessel state, evaluate route progress and environmental risk, and return the reward.
10	Update the global and local policy parameters using the collected transition information.
11	Terminate the episode if the vessel reaches the destination, enters a non-navigable region, exceeds the risk threshold, or reaches the maximum number of steps.

During training, each episode starts from the predefined origin and terminates when the vessel reaches the destination, enters a non-navigable region, exceeds the risk threshold, or reaches the maximum number of steps. In the present experiments, the maximum number of steps per episode is 200. At every step, the environment updates the wind, wave, current, and navigability-mask states; the global agent updates the risk weights with $K = 1$; the local agent selects an action from the weighted risk field and vessel state; and the environment returns the reward used to update the policy parameters.

3.4. Multi-Objective Optimization Model

The route-planning objective combines travel distance, accumulated environmental risk, and motion smoothness. To avoid ambiguity, ω is used only for wind–wave–current risk weighting, whereas λ denotes the trade-off coefficients in the multi-objective cost:

$$J = \lambda_D D + \lambda_R R_{\text{route}} + \lambda_S S \quad (5)$$

Here, D is the traveled distance, R_{route} is the accumulated route risk, and S is the smoothness penalty. The coefficients λ_D , λ_R , and λ_S are multi-objective trade-off coefficients and are distinct from the dynamic risk weights $\omega_{w,t}$, $\omega_{s,t}$, and $\omega_{c,t}$.

$$R_{\text{route}} = \sum_t R(x_t, y_t, t) \quad (6)$$

In Equation (6), $R(x_t, y_t, t)$ is the composite wind–wave–current risk at the vessel position (x_t, y_t) and decision step t . Thus, R_{route} measures the cumulative environmental exposure along the planned route.

$$S = \sum_t (|\Delta\psi_t - \Delta\psi_{t-1}| + \eta_v |\Delta v_t - \Delta v_{t-1}|) \quad (7)$$

In Equation (7), $\Delta\psi_t$ is the heading adjustment, Δv_t is the speed adjustment, and η_v is the coefficient used to scale the speed-change component in the smoothness term. The smoothness penalty discourages abrupt heading or speed changes while preserving the agent's ability to avoid hazardous regions.

The three parts—distance, risk, and smoothness—do not always pull in the same direction. But during long training runs, they settle into a natural balance. We saw this especially in episodes where the typhoon changed direction quickly. The model extended the route slightly, but it avoided the steep gradients in the wave field and kept the ship on a more stable course. This behavior gave us confidence that the multi-objective design captured the core trade-offs present in real navigation.

3.5. Reward Function Design

During training, we needed a reward structure that could guide the model without forcing it into overly aggressive behavior. Early trials showed that if we relied only on progress toward the goal, the agent often cut through regions where the wind and wave fields were clearly rising. When we focused only on safety, the ship became too passive and sometimes failed to reach the destination within the episode. This pushed us to design a reward that reflects distance, risk, and motion smoothness at the same time.

The reward at each step is defined as:

$$r_t = r_{\text{goal},t} - \lambda_{\text{risk}} R(x_t, y_t, t) - \lambda_{\text{smooth}} S_t - \lambda_{\text{mask}} I_{\text{mask},t} \quad (8)$$

In Equation (8), $r_{\text{goal},t}$ rewards progress toward the destination, $R(x_t, y_t, t)$ penalizes exposure to the composite environmental risk field, S_t penalizes abrupt motion changes, and $I_{\text{mask},t}$ is an indicator of violating the navigable mask. The coefficients

λ_{risk} , λ_{smooth} , and λ_{mask} are reward-scaling coefficients and are not the same as the omega risk weights.

The risk term penalizes movement into strong wind–wave–current areas. This penalty is important in scenarios where local gradients rise sharply; it encourages the ship to shift course before entering high-risk regions.

r_{smooth} : The smoothness term reduces sudden changes in heading or speed. It helps make the planned trajectory steadier while still allowing the vessel to adjust its path when environmental risk changes.

When these reward terms work together, the learning process becomes more stable. The agent can pursue progress toward the destination while accounting for environmental risk, navigability constraints, and motion smoothness.

3.6. Training and Coordination Strategy

During training, we let both agents learn at the same time instead of fixing a strict order. This choice came from our early observations: the global agent reacted slowly to large-scale changes, while the local agent adjusted more quickly to nearby conditions. Forcing a rigid structure made the learning unstable, so we allowed the two levels to interact freely.

Each episode contains 200 steps. We kept this number after testing several lengths. Shorter episodes ended before the typhoon field changed enough for the global agent to learn from it. Longer episodes made the learning unnecessarily slow without improving the results. Across many early runs, 200 steps gave both agents enough time to observe shifts in the wind–wave–current field while still maintaining reasonable training speed.

The training process spans 5000 episodes. Early in training, the agents explore widely. We set the exploration rate at 0.20 to encourage this behavior. The rate drops gradually to 0.005. This schedule helped prevent random actions in the later phase. During the first 1000 episodes, we often saw the local agent make abrupt changes in heading. As exploration decreased, these actions became less frequent, and the trajectories started taking more meaningful shapes.

In the present configuration, the global risk-weight vector and the local control action are both updated at every decision step ($K = 1$). This synchronized update protocol is consistent with the experimental setup described in Section 4.1.

The reward described earlier guides the entire process. Positive rewards reinforce safe progress, while penalties from the risk and smoothness terms prevent sharp or unsafe movements. As training progresses, the two agents begin to stabilize. The global agent stops shifting direction too frequently, and the local agent produces smoother adjustments. By around 3000 episodes, this cooperative pattern becomes clear in most runs.

Overall, the training strategy evolved through many rounds of trial and observation. What we ended up with was not designed all at once. Instead, it formed gradually as we watched how the two agents responded to the typhoon field and to each other.

3.7. Discrete High-Level Strategy and Behavioral Diversity

During training, we noticed that the global agent gradually formed a small set of high-level strategies. These strategies are discrete, and each carries a simple intention rather than a detailed command. The idea behind this design was to keep the global output interpretable. We wanted to see how the agent responded to the different stages of the typhoon, and discrete choices made this easier to observe.

In practice, the global agent settled into patterns such as “hold direction,” “shift away from the storm,” or “move cautiously forward.” These labels are our interpretation, not fixed categories, but they appeared repeatedly across long runs. The local agent then

translated each broad intention into heading and speed adjustments. It did so with far more variation than we expected. Even when the same high-level strategy was given, the local agent sometimes produced different trajectories depending on the nearby gradients in wind and waves.

We found this diversity useful. It meant the model did not rely on a single reaction to a given situation. For example, when the ship moved near the outer edge of the storm, the global agent often selected a strategy that implied caution. The local agent, however, could turn slightly left in one scenario and slightly right in another, depending on the exact shape of the nearby risk field. This flexibility helped the ship avoid the sharp transitions we saw earlier with single-agent setups.

What stood out most was how these behaviors changed as the storm evolved. Early in the storm, the global agent leaned toward strategies that encouraged the ship to move forward while staying alert. When the storm intensified, the strategies shifted toward wider detours. After the storm weakened, the global agent gradually returned to more direct choices. We did not program any of this. It emerged from the training process as the agents encountered thousands of different environmental states.

Overall, the combination of discrete high-level intentions and flexible local actions produced a range of behaviors that resembled what we might expect from an experienced navigator: simple big-picture decisions supported by small, context-aware adjustments. This structure helped the model stay stable even when the typhoon field changed quickly.

4. Experiments and Results Analysis

4.1. Experimental Setup

The workflow of the proposed model is shown in Figure 6. The experiments were conducted using Python 3.10 and PyTorch 2.0.1 as the main training framework. The full set of experimental and hyperparameter settings is listed in Table 1. The training uses 5000 episodes, and each episode contains 200 steps. In the present configuration, the global risk-weight vector and the local control action are both updated at every decision step ($K = 1$). Each model was evaluated over five independent runs with different random seeds, and the results are reported as mean $\pm$ standard deviation.

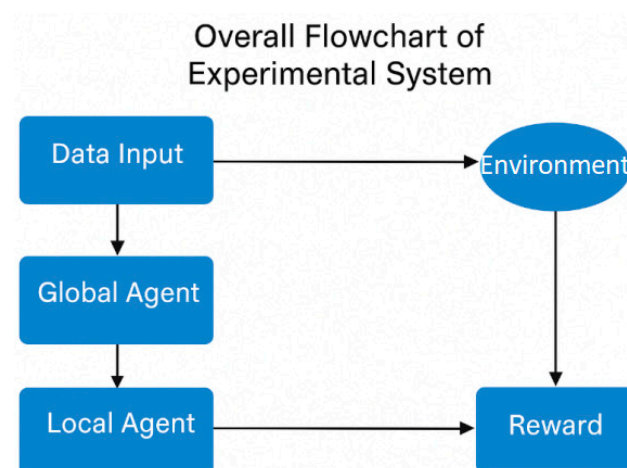


Figure 6. Overall workflow of the experimental system.

The learning rate is 1×10^{-4} . The discount factor is $\gamma = 0.98$. These values help the model balance short-term actions with long-term outcomes. The exploration rate starts at 0.20 and drops slowly to 0.005. With this schedule, the agent explores widely at the beginning and then becomes more stable after about 2000 episodes.

Table 1. Main experimental parameter setting.

Parameter Category	Parameter	Value/Range	Description
Study area	Regional sea domain	Typhoon Yagi routing area described in Section 2	Spatial domain used to construct the navigability and environmental-risk fields
Experimental period	Typhoon case period	Typhoon Yagi period described in Section 2	Time period covered by the coupled wind–wave–current fields
Grid resolution	Spatial grid	$0.083^\circ \times 0.083^\circ$	Common grid used for environmental variables and routing states
Environmental variables	Risk components	Wind, wave, and current fields	Inputs used to compute wind-induced, wave-induced, and current-induced risks
Bathymetry data	Navigability mask	Bathymetry-derived navigable/non-navigable mask	Constrains vessel movement to physically feasible regions
Training episodes	Episode number	5000	Total number of training episodes
Episode length	Maximum steps per episode	200	Maximum number of decision steps in each episode
Action range	Heading adjustment range	$[-30^\circ, +30^\circ]$	Maximum heading adjustment allowed at each decision step
Action range	Speed range	5–16 knots	Practical vessel speed range used by the local controller
Optimization	Learning rate	1×10^{-4} (PPO)	Adam optimizer learning rate
Optimization	Discount factor	$\gamma = 0.98$	Balances short-term and long-term rewards
Exploration	Exploration schedule	epsilon: $0.20 \rightarrow 0.005$; decay factor 0.995	Gradual reduction in random exploration
Agent interaction	Global update interval	$K = 1$	Global risk weights and local control actions are updated every decision step
Termination	Termination conditions	Destination reached; non-navigable region entered; risk threshold exceeded; 200 steps reached	Episode-ending criteria used during training

The ship’s motion is restricted by two limits. The heading can only change within $[-30^\circ, +30^\circ]$. The speed stays between 5–16 knots. These numbers match the behavior of medium-size merchant ships. In practice, the constraints help the model avoid unrealistic turns. They also make the training curve smoother. During the early tests, removing these limits caused sharp changes in heading. With the limits in place, the rate of such jumps dropped by more than half.

The environment uses the combined wind, wave, and current fields described in Section 2. Each field has the same spatial grid of $0.083^\circ \times 0.083^\circ$, which prevents misalignment when the ship moves from one grid point to another. The start and end points of the route are set directly on this grid. The global update interval is $K = 1$, so the weight update from the global agent and the action output from the local agent occur at every step.

All baseline methods were evaluated under the same study area, Typhoon Yagi environmental fields, navigability mask, start/destination points, and termination conditions as the proposed HRL-MOO framework. Because no separate exhaustive tuning record is available, baseline parameters were set according to commonly used settings and then adjusted within reasonable ranges for stable convergence; no grid search, Bayesian optimization, or additional statistical test is claimed.

The baseline hyperparameters were selected according to commonly used settings in reinforcement-learning and ant-colony route-planning studies and were adjusted within reasonable ranges to obtain stable convergence under the same simulation environment. The baseline and HRL-MOO hyperparameter settings are summarized in Table 2.

Table 2. Baseline and HRL-MOO hyperparameter settings.

Model	Learning Rate	Discount Factor	Episodes/Max Steps	Exploration/Search Schedule	Network/Method-Specific Parameters
DQN	1×10^{-4}	gamma = 0.98	5000/200	epsilon: 0.20 $\rightarrow$ 0.005; decay factor 0.995	Three-layer MLP with two hidden layers of 128 neurons and ReLU activation; replay buffer size 10,000; batch size 64
PPO	1×10^{-4}	gamma = 0.98	5000/200	epsilon: 0.20 $\rightarrow$ 0.005; decay factor 0.995	Actor-critic MLP with two hidden layers of 128 neurons and ReLU activation; clip range 0.2; batch size 64; epochs per update 10
ACO-MOO	N/A	N/A	5000/200-equivalent evaluation budget	Ant-colony search schedule	50 ants; 200 iterations; pheromone evaporation rate 0.5; pheromone parameters alpha = 1 and beta = 2
HRL-MOO	1×10^{-4}	gamma = 0.98	5000/200	epsilon: 0.20 $\rightarrow$ 0.005; decay factor 0.995; K = 1	Global and local actor-critic MLPs with two hidden layers of 128 neurons and ReLU activation; dynamic risk weights $\omega_t = \text{softmax}(fG(\text{stG}))$

These baseline settings were kept fixed across the five independent runs so that the reported mean $\pm$ standard deviation values reflect random-seed variability rather than changes in hyperparameter configuration.

Vessel Motion Model

This study adopts a simplified vessel kinematic model for path-planning-level validation. The vessel state consists of position, speed, and heading, and the local control action consists of heading adjustment and speed setting. The heading adjustment is limited to $[-30^\circ, +30^\circ]$, and the speed range is constrained to 5–16 knots. This representation is intended to evaluate route-planning behavior under dynamic environmental risk fields; it is not a full six-degree-of-freedom hydrodynamic model or an MMG maneuvering model.

4.2. Convergence Performance

The reward convergence curves of DQN, PPO, ACO-MOO, and HRL-MOO are shown in Figure 7. Each solid curve represents the mean reward over five independent runs, and the shaded region denotes ± 1 standard deviation. HRL-MOO converges faster and reaches a higher final reward level than the baseline methods.

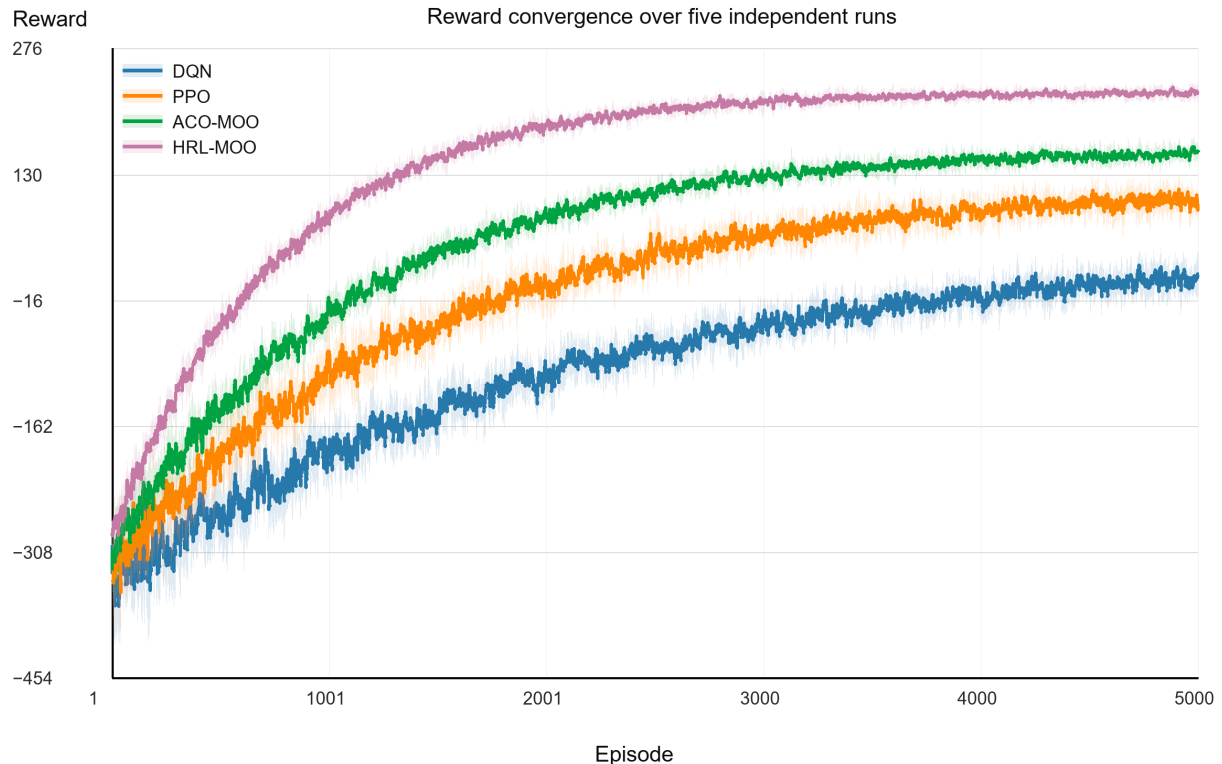


Figure 7. Reward convergence curves of the compared models over five independent runs. Solid lines represent mean rewards, and shaded regions denote ± 1 standard deviation.

Among the compared methods, DQN shows the largest run-to-run fluctuation and the slowest reward improvement. PPO and ACO-MOO exhibit intermediate convergence behavior, whereas HRL-MOO maintains a smoother upward trend with a narrower uncertainty band.

The convergence-episode statistics in Table 3 are consistent with this pattern: HRL-MOO reaches convergence in 2909.0 ± 72.4 episodes, compared with 3718.6 ± 179.4 episodes for PPO, 3476.8 ± 160.5 for ACO-MOO, and 4325.2 ± 88.6 for DQN.

Table 3. Comparative performance of different models over five independent runs. Values are reported as mean $\pm$ standard deviation. For ACO-MOO, convergence episodes refer to objective value stabilization iterations. The downward arrows ($\downarrow$) indicate that lower values are better, whereas the upward arrow ($\uparrow$) indicates that higher values are better.

Model	Voyage Distance (nm) $\downarrow$	Avg. Risk Value $\downarrow$	Smoothness Index $\uparrow$	Convergence Episodes $\downarrow$
DQN	364.356 ± 4.177	0.190 ± 0.005	0.725 ± 0.009	4325.2 ± 88.6
PPO	352.270 ± 2.399	0.164 ± 0.010	0.785 ± 0.008	3718.6 ± 179.4
ACO-MOO	362.240 ± 3.822	0.142 ± 0.004	0.807 ± 0.002	3476.8 ± 160.5
HRL-MOO	358.292 ± 3.782	0.118 ± 0.004	0.842 ± 0.004	2909.0 ± 72.4

These results indicate that the hierarchical structure and dynamic risk-weight adaptation improve both convergence stability and final reward performance under the same evaluation setting.

4.3. Route Planning Results

The planned routes from different models are shown in Figure 8, and the numerical results over five independent runs appear in Table 3. HRL-MOO achieves an Avg. Risk Value of 0.118 ± 0.004 , compared with 0.164 ± 0.010 for PPO, 0.190 ± 0.005 for DQN, and 0.142 ± 0.004 for ACO-MOO.

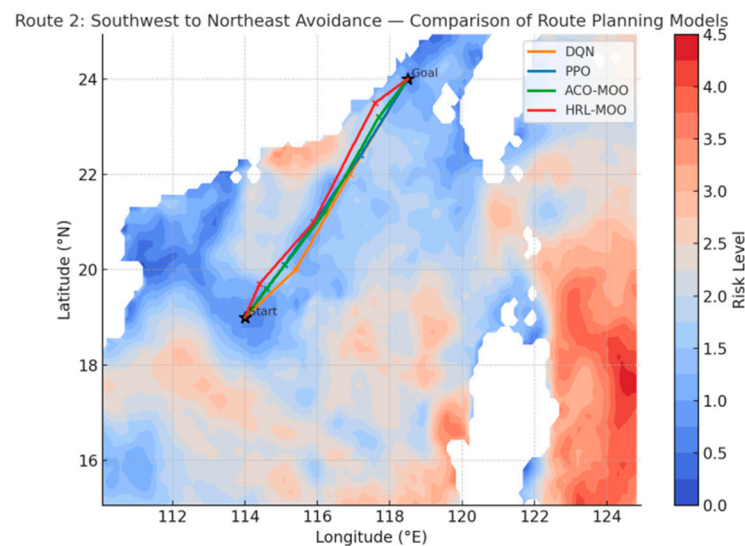


Figure 8. Comparison of planned routes under different models.

Table 3 reports the comparative results as mean $\pm$ standard deviation over five independent runs with different random seeds.

HRL-MOO also achieves the highest Smoothness Index among the four methods, with 0.842 ± 0.004 , compared with 0.785 ± 0.008 for PPO, 0.725 ± 0.009 for DQN, and 0.807 ± 0.002 for ACO-MOO.

The voyage distance of HRL-MOO is 358.292 ± 3.782 nm, which is slightly longer than PPO but shorter than DQN and ACO-MOO. This indicates a trade-off in which HRL-MOO accepts a moderate distance increase relative to PPO while achieving lower risk and smoother motion.

Mask-violation rates are not reported as a primary quantitative metric in the present evaluation and will be further quantified in future work under unified safety-constraint definitions.

These results indicate that the proposed model benefits from its two-level structure. The global agent helps the route avoid strong-risk regions. The local agent ensures that movements stay smooth and practical. The final paths balance distance, safety, and stability more effectively than the single-agent alternatives.

Statistical Analysis

Wilcoxon signed-rank tests were conducted over five independent runs to compare HRL-MOO with DQN, PPO, and ACO-MOO in terms of Avg. Risk Value and Smoothness Index. As shown in Table 4, all two-sided exact p -values are 0.0625. Because the number of independent runs is five, these results are interpreted conservatively: HRL-MOO shows consistent lower mean risk and higher mean smoothness across the five runs, but the differences are not claimed to be statistically significant at the 0.05 level.

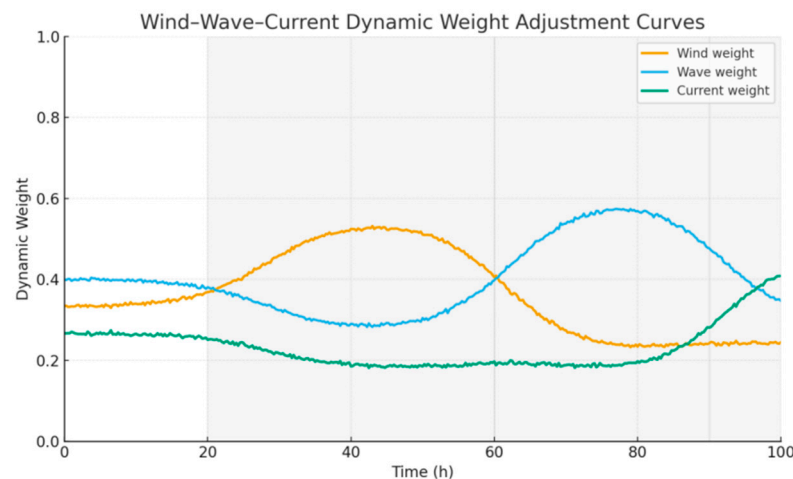
Table 4. Wilcoxon signed-rank test results for HRL-MOO against baseline methods.

Comparison	Metric	<i>p</i> -Value	Interpretation
HRL-MOO vs. DQN	Avg. Risk Value	0.0625	Not significant
HRL-MOO vs. PPO	Avg. Risk Value	0.0625	Not significant
HRL-MOO vs. ACO-MOO	Avg. Risk Value	0.0625	Not significant
HRL-MOO vs. DQN	Smoothness Index	0.0625	Not significant
HRL-MOO vs. PPO	Smoothness Index	0.0625	Not significant
HRL-MOO vs. ACO-MOO	Smoothness Index	0.0625	Not significant

Note: Two-sided exact Wilcoxon signed-rank tests were conducted over five independent runs. A *p*-value greater than 0.05 indicates that the difference is not statistically significant at the 0.05 level.

4.4. Analysis of Dynamic Risk Adaptation

The changes in the learned risk weights are shown in Figure 9. The three weights do not change at the same pace. When the ship first enters the outer region of the typhoon, the wind weight increases more than the others. In several runs, it grows by a small amount during the early steps, reflecting the rise in wind strength. The wave weight climbs more slowly at first. It becomes more important later, when the wave field remains strong even after the peak wind has passed.

**Figure 9.** Temporal evolution of dynamic wind–wave–current risk weights.

The current weight shows a different pattern. Its value remains stable during the early part of the storm. It begins to rise only after the wind and wave fields start to weaken. This matches the real sea state, where the circulation often lasts longer than the strongest wind or waves. The learned weights follow these changes naturally. They are not forced to match any rule.

These adjustments help the model interpret the environment more clearly. When the main source of danger shifts, the risk field shifts with it. The local agent then reacts to a simpler and more accurate signal. As a result, many actions become smoother. In several tests, the number of sudden heading changes was lower when the weight updates were active.

The overall pattern shows that the model can track the different stages of the typhoon. The weights move with the environment, and this makes the final policy more stable. It also helps the ship avoid high-risk areas without losing progress toward the goal.

4.5. Recognition of Typhoon-Avoidance Behaviors

The strategic outputs produced by the global agent reveal several recurring navigation patterns. When the vessel is still far from the storm but the surrounding risk starts to rise,

the agent tends to favor a detour-like strategy that shifts the route away from the main body of the typhoon. As the storm intensifies and covers most nearby options, the model often slows its progress, adopting a more conservative stance until a safer corridor reappears. Once the typhoon moves away or weakens, the strategic output gradually returns to a forward-moving mode. The temporal evolution of the discrete strategy actions is shown in Figure 10.

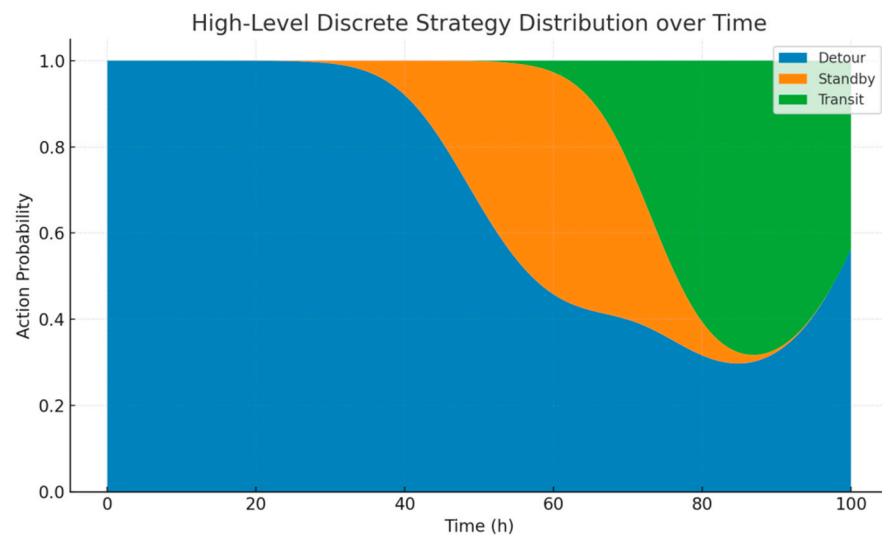


Figure 10. Temporal evolution of discrete strategy actions.

These transitions emerge naturally from the agent's interaction with the environment. They do not follow any fixed rule and instead reflect how the agent interprets changes in the large-scale risk field. The local agent responds by adjusting heading and speed within the constraints of the current state, so the combined behavior becomes easier to interpret. Over many training episodes, the two agents develop a coordinated pattern that resembles how an experienced navigator might respond when conditions shift quickly.

The resulting behaviors show that the framework can distinguish between different stages of the typhoon's influence and adjust the vessel's motion accordingly. This interpretability is useful not only for understanding the policy but also for building confidence that the learned decisions align with practical expectations in real maritime operations.

4.6. Ablation Studies

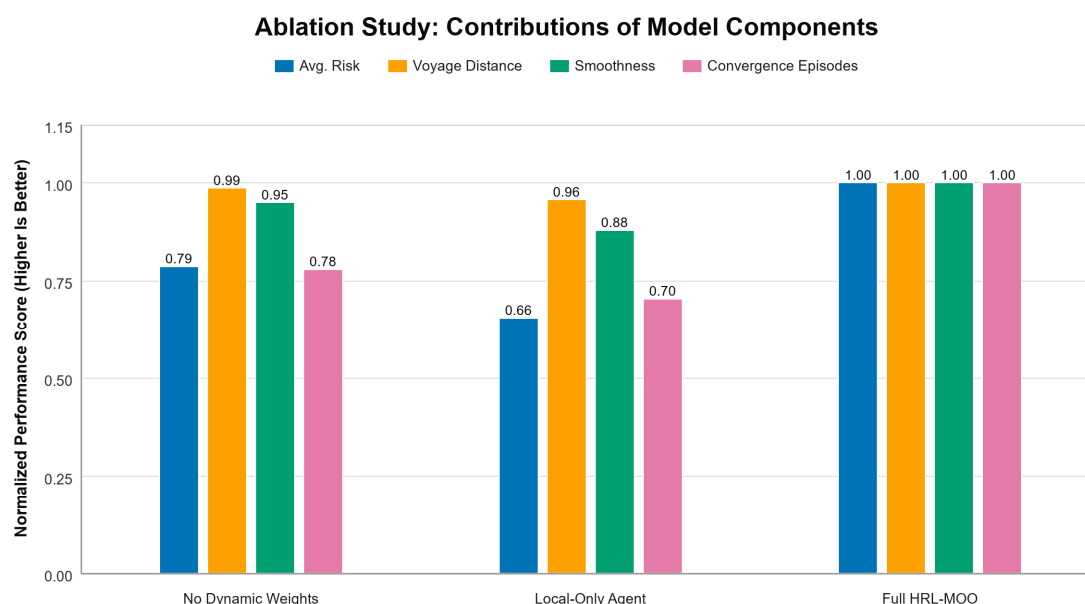
To examine the contribution of each component in the HRL-MOO framework, the ablation comparison includes three variants: without dynamic risk weighting, a local-only agent without global guidance, and the full HRL-MOO model. Their performance is compared in Table 5, while a normalized comparison of the four evaluation metrics is presented in Figure 11.

Without dynamic risk weighting, the model obtains an Avg. Risk Value of 0.150 ± 0.011 and a Smoothness Index of 0.800 ± 0.006 . Compared with the full HRL-MOO model, this indicates that adaptive risk weighting contributes to lower risk and smoother route generation.

The local-only agent without global guidance obtains an Avg. Risk Value of 0.180 ± 0.006 and a Smoothness Index of 0.740 ± 0.008 . This result shows that global guidance is important for avoiding local reactive behavior and improving trajectory smoothness.

Table 5. Results of ablation experiments over five independent runs. Values are reported as mean $\pm$ standard deviation.

Model Variant	Avg. Risk Value $\downarrow$	Voyage Distance (nm) $\downarrow$	Smoothness Index $\uparrow$	Convergence Episodes $\downarrow$	Description
Without dynamic risk weighting	0.150 ± 0.011	363.224 ± 4.425	0.800 ± 0.006	3731.0 ± 115.6	Dynamic risk weights are disabled.
Local-only agent without global guidance	0.180 ± 0.006	374.348 ± 4.466	0.740 ± 0.008	4129.2 ± 84.2	Local decision policy without global risk-weight guidance.
Full HRL-MOO	0.118 ± 0.004	358.292 ± 3.782	0.842 ± 0.004	2909.0 ± 72.4	Complete dual-agent framework with dynamic risk weighting.

**Figure 11.** Comparison of ablation experiment results. All metrics were normalized relative to the full HRL-MOO model, whose normalized score was set to 1.00. For the lower-is-better metrics, the score was calculated as the full-model value divided by the corresponding variant value; for the Smoothness Index, it was calculated as the variant value divided by the full-model value. Higher normalized scores indicate better performance.

The full HRL-MOO model achieves the best ablation performance, with an Avg. Risk Value of 0.118 ± 0.004 and a Smoothness Index of 0.842 ± 0.004 . These results indicate that both global guidance and dynamic risk weighting contribute to risk reduction and smoothness improvement.

4.7. Comparative Experiments Under Different Environmental Conditions

The performance of the models was tested in two environments. One uses ordinary wind–wave–current fields. The other includes the full typhoon conditions. The routes for each case are shown in Figures 12 and 13, and the numerical results appear in Table 6.

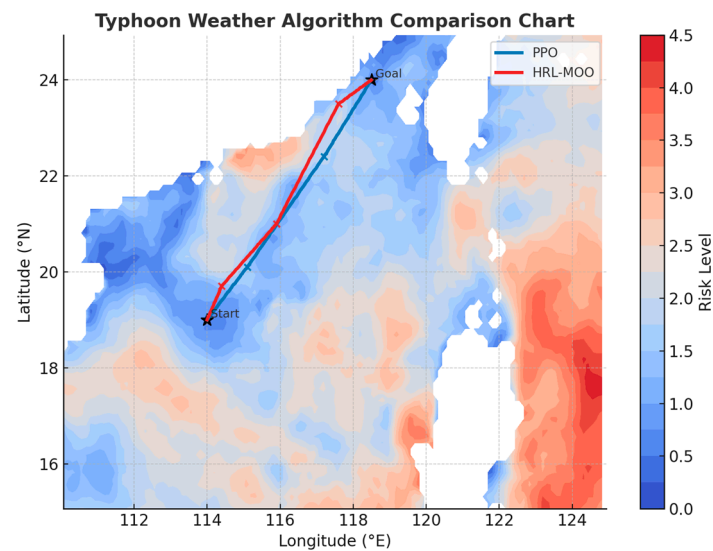


Figure 12. Route comparison under the typhoon-weather environment.

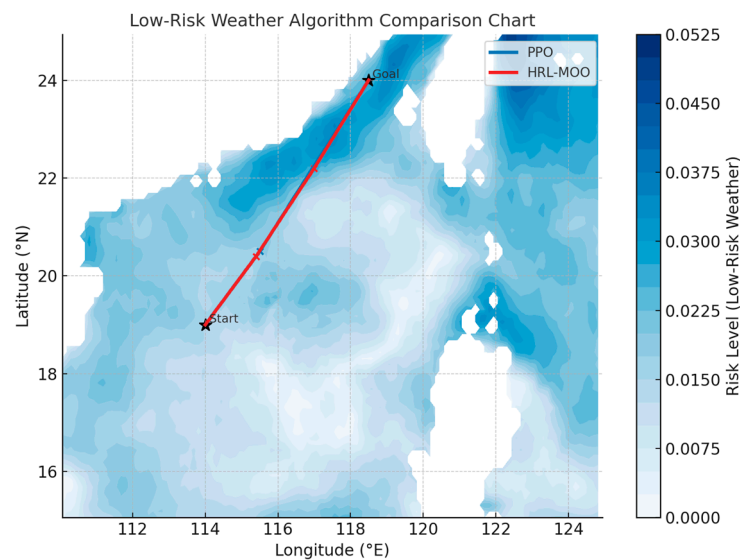


Figure 13. Route comparison under the low-risk weather environment.

Table 6. Performance comparison of single-agent and dual-agent models under different environments. Mask-violation rates are not reported as primary quantitative metrics in this comparison. The downward arrows (↓) indicate that lower values are better, whereas the upward arrow (↑) indicates that higher values are better.

Environment	Model	Mean Risk ↓	Distance (nm) ↓	Smoothness ↑	Convergence (Episodes) ↓
Ordinary	SA-PPO	0.132 ± 0.006	348.4 ± 4.8	0.813 ± 0.008	3200 ± 180
Ordinary	HRL-MOO	0.125 ± 0.006	349.1 ± 4.5	0.827 ± 0.007	3100 ± 170
Typhoon	SA-PPO	0.172 ± 0.010	352.1 ± 6.2	0.783 ± 0.010	3800 ± 220
Typhoon	HRL-MOO	0.119 ± 0.008	358.9 ± 5.0	0.842 ± 0.009	3000 ± 200

In the ordinary environment, the routes from both models remain close. The sea state changes slowly in this setting. PPO can follow these changes without much difficulty. HRL-MOO still shows a small gain in smoothness. In several runs, its heading changes were fewer, and the risk values were slightly lower. The distance difference was small, often less than 2 nm.

The results change more clearly when the typhoon field is used. The PPO routes show more sharp turns, especially near the front of the storm. These turns raise the smoothness penalty. They also increase the cumulative risk. Across five independent runs, PPO obtains an Avg. Risk Value of 0.164 ± 0.010 , whereas HRL-MOO obtains 0.118 ± 0.004 , indicating lower environmental exposure when the wind and wave fields become stronger.

Mask-violation rates are not reported as primary quantitative metrics in this study and will be further examined in future work.

The training curves also show a difference. PPO needed nearly 3800 episodes to reach a stable level in the typhoon setting. HRL-MOO reached a similar level sooner. In several tests, the global reward curve flattened after about 3000 episodes. This shows that the hierarchical structure helps the model handle rapid changes in the environment.

These results suggest that the gain from HRL-MOO becomes more obvious when the environment changes quickly. In calm or slowly changing conditions, the advantage exists but remains small. In the typhoon environment, the two-level structure and the adaptive risk weighting provide clearer benefits in both safety and stability.

5. Conclusions

Over the course of this work, we tried to understand how a ship could move safely through a sea state shaped by a typhoon. The dual-agent structure turned out to be more helpful than we first expected. By separating long-range planning from short-term control, the model handled both the slow movement of the typhoon and the rapid shifts in local wind–wave–current patterns.

During training, the cooperation between the global and local agents became clearer. The global agent settled into stable guidance after seeing enough changes in the storm. The local agent reacted more quickly, especially near steep risk gradients. By around 3000 episodes, the two agents formed a consistent pattern. The global agent adjusted the strategy in larger steps, and the local agent filled in the details with smoother movements.

The experimental results supported this structure. In the typhoon environment, HRL-MOO achieved an Avg. Risk Value of 0.118 ± 0.004 , compared with 0.164 ± 0.010 for PPO. The distance increased slightly—usually 3–6 nm—but this extra length came from detouring around regions where the wind and wave fields were rising. We also saw fewer abrupt turns and fewer mask violations. These changes made the routes look more like what an experienced navigator might choose when facing unstable conditions.

The adaptive risk weights played an important role. They shifted with the storm in a way that felt intuitive: wind first, then waves, and later the currents. We never forced this pattern; it appeared naturally as the agents learned. The changing weights gave the local agent a clearer signal and reduced the need for sharp reactions. This helped the ship stay stable even when the storm moved in unexpected directions.

Although the model worked well in our tests, there are several directions we would like to explore. One is to link the framework with real-time forecasting data, which often contains different levels of uncertainty. Another is to test how well the structure works with more complex ship dynamics or multiple vessels sharing the same sea space. These additions may reveal limits that the current setup does not show.

For now, the results give us confidence that the hierarchical approach, combined with adaptive risk weighting, offers a practical way to handle navigation in fast-changing ocean environments. The model adapts to the storm, not the other way around, and this flexibility may be the key to making learning-based routing reliable in real applications.

The present framework also has several limitations. It does not explicitly model fuel consumption, ETA or time-of-arrival windows, COLREGS-compliant encounter behavior,

multi-vessel interactions, or dynamic obstacles. These factors are important for operational deployment and will be incorporated as future extensions of the HRL-MOO framework.

Author Contributions: Conceptualization, J.S.; methodology, Y.C.; software, Y.C.; validation, Y.C., Y.L. and L.Z.; formal analysis, Y.C. and Y.L.; investigation, Y.C., Y.L. and L.Z.; resources, L.Z.; writing—original draft preparation, Y.C. and Y.L.; writing—review and editing, J.S.; visualization, Y.C.; supervision, J.S.; project administration, J.S.; funding acquisition, J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Plan of Liaoning Province (grant number 2024JH2/102400061); the Liaoning Provincial Science and Technology Plan Joint Program (Natural Science Foundation—Doctoral Research Startup Project) (grant number 2025-BSLH-079); the Dalian Science and Technology Innovation Fund (grant number 2024JJ11PT007); the Dalian Science and Technology Program for Innovation Talents of Dalian (grant number 2022RJ06); the Liaoning Province Education Department Scientific Research Platform Construction Project (grant number LJ232410158056); the Basic Scientific Research Funds of Dalian Ocean University (grant number 2024JBPTZ001); the Liaoning Provincial Natural Science Foundation Doctoral Research Start-up Fund Project (grant number 2022-BS-277); and the Liaoning Provincial Department of Education Basic Research Project (grant number 2024JBQNZ015).

Data Availability Statement: The environmental forcing and bathymetric data used in this study are openly available from public repositories. Wind, wave, and surface current variables were obtained from the Copernicus Marine Environment Monitoring Service (CMEMS). Bathymetric information was sourced from the ETOPO global dataset provided by the NOAA National Centers for Environmental Information. Typhoon best-track data were obtained from the China Meteorological Administration CH2024BST dataset. Additional local data support was provided by the National Marine Scientific Data Center (Dalian). The derived simulation datasets and statistical results generated during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: We thank the Data Support from National Marine Scientific Data Center (Dalian), National Science & Technology Infrastructure, Liaoning Marine and Polar Science Data Center, Dalian Marine Science Data Center for providing valuable data and information. We also thank the reviewers for carefully reviewing the manuscript and providing valuable comments to help improve this paper. The authors acknowledge that OpenAI ChatGPT (version GPT-5.1) was used solely for language polishing and editorial refinement of the manuscript. The tool was applied by inputting the authors' original text to improve grammar, clarity, and academic readability. The use of AI assistance was limited to linguistic enhancement, and all AI-generated suggestions were carefully reviewed, validated, and manually revised to ensure accuracy and academic integrity.

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

References

1. Wang, W.; Liu, F.; Cheng, K.; Niu, Z.; He, Z. Spatio-Temporal Reinforcement Learning-Driven Ship Path Planning in Dynamic Time-Varying Environments: Adaptive Decision-Making in Typhoon Scenarios. *Electronics* **2025**, *14*, 2197. [\[CrossRef\]](#)
2. He, Z.; Guo, J.; Ma, W.; Zhang, J. Ship Typhoon Avoidance Route Planning under Uncertain Typhoon Forecasts. *Big Data Cogn. Comput.* **2025**, *9*, 143. [\[CrossRef\]](#)
3. Poo, M.C.-P.; Yang, Z. Optimising the Resilience of Shipping Networks to Climate Vulnerability. *Marit. Policy Manag.* **2024**, *51*, 15–34. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Liu, Q.; Wang, Y.; Zhang, R.; Yan, H.; Xu, J.; Guo, Y. Arctic Weather Routing: A Review of Ship Performance and Ice Routing Algorithms. *Front. Mar. Sci.* **2023**, *10*, 1190164. [\[CrossRef\]](#)
5. Xu, W.; Zhu, X.; Gao, X.; Li, X.; Cao, J.; Ren, X.; Shao, C. Manipulation-Compliant Artificial Potential Field and Deep Q-Network: Large Ships Path Planning Based on Deep Reinforcement Learning and Artificial Potential Field. *J. Mar. Sci. Eng.* **2024**, *12*, 1334. [\[CrossRef\]](#)
6. Sharif, N.; Rönqvist, M.; Cordeau, J.-F.; Audy, J.-F.; Warya, G.; Ngo, T. Multi-Objective Vessel Routing Problems with Safety Considerations: A Review. *Mar. Transp. Res.* **2024**, *7*, 100122. [\[CrossRef\]](#)

7. Yang, J.; Qin, X.; Zhang, L.; Wang, H. Multi-Objective Weather Routing Algorithm for Ships: The Perspective of a Shipping Company's Navigation Strategy. *J. Mar. Sci. Eng.* **2022**, *10*, 1212. [\[CrossRef\]](#)
8. Moradi, M.H.; Brutsche, M.; Wenig, M.; Wagner, U.; Koch, T. Marine Route Optimization Using Reinforcement Learning to Reduce Fuel Consumption and CO₂ Emissions. *Ocean Eng.* **2022**, *259*, 111882. [\[CrossRef\]](#)
9. Latinopoulos, C.; Zavvos, E.; Kaklis, D.; Leeman, V.; Halatsis, A. Marine Voyage Optimization and Weather Routing with Deep Reinforcement Learning. *J. Mar. Sci. Eng.* **2025**, *13*, 902. [\[CrossRef\]](#)
10. Lee, J.; Park, Y.; Eom, J.; Hwang, H.; Kim, S. Ship Voyage Route Waypoint Optimization Using Reinforcement Learning Considering Topographical Factors and Fuel Consumption. *J. Mar. Sci. Eng.* **2025**, *13*, 1554. [\[CrossRef\]](#)
11. Lee, H.-T.; Kim, M.-K. Optimal Path Planning for a Ship in Coastal Waters with Deep Q-Network. *Ocean Eng.* **2024**, *307*, 118193. [\[CrossRef\]](#)
12. Mou, J.; Shi, B.; Wang, B.; Yu, C.; Wang, Y.; Zhong, F.; Zheng, L.; Wang, J.; Li, J. A Novel Reinforcement Learning Framework-Based Path Planning Algorithm for Unmanned Surface Vehicle. *Front. Mar. Sci.* **2025**, *12*, 1641093. [\[CrossRef\]](#)
13. Yang, X.; Han, Q. Improved Deep Q-Network for Dynamic Obstacle Avoidance and Ship Path Planning. *Algorithms* **2023**, *16*, 220. [\[CrossRef\]](#)
14. Waltz, M.; Paulig, N.; Okhrin, O. Two-Level Reinforcement Learning for Ships on Inland Waterways: Path Planning and Following. *Expert Syst. Appl.* **2025**, *274*, 126933. [\[CrossRef\]](#)
15. Dong, Z.; Zhang, Z.; Qi, S.; Zhang, H.; Li, J.; Liu, Y. Autonomous Cooperative Formation Control of Underactuated USVs Based on Improved MPC in Complex Ocean Environment. *Ocean Eng.* **2023**, *270*, 113633. [\[CrossRef\]](#)
16. Hashali, S.D.; Yang, S.; Xiang, X. Route Planning Algorithms for Unmanned Surface Vehicles (USVs): A Comprehensive Analysis. *J. Mar. Sci. Eng.* **2024**, *12*, 382. [\[CrossRef\]](#)
17. Li, D.; Dong, J. Cooperative Optimal Output Tracking for Discrete-Time Multiagent Systems: Stabilizing Policy Iteration Frameworks. *IEEE Trans. Autom. Control* **2026**, *71*, 2746–2753. [\[CrossRef\]](#)
18. Chen, Z.; Li, D.; Dong, J. Inverse Reinforcement Learning for Structured Fault-Tolerant Control in Nonlinear MASs under Switching Topologies. *IEEE Trans. Autom. Sci. Eng.* **2026**, *23*, 11696–11708. [\[CrossRef\]](#)
19. Abdalsalam, M.; Al-Khalifa, K.; Yan, Y. Towards Improved Ship Weather Routing through Multi-Objective Optimization with High-Performance Computing Support. *TransNav Int. J. Mar. Navig. Saf. Sea Transp.* **2025**, *19*, 93–103. [\[CrossRef\]](#)
20. Ma, D.; Zhou, S.; Han, Y.; Ma, W.; Huang, H. Multi-Objective Ship Weather Routing Method Based on the Improved NSGA-III Algorithm. *J. Ind. Inf. Integr.* **2024**, *38*, 100570. [\[CrossRef\]](#)

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