Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep Reinforcement Learning-Based AUV Trajectory Planning in 3D Time-Varying Current Environments

Jun Xiao
Xiamen University
Ranshuai Zheng
Xiamen University
Xiaolin Bai
xbai@xmu.edu.cn
Xiamen University https://orcid.org/0000-0001-5044-509X
Zhiduo Tan
State Key Laboratory of Robotics Shenyang Institute of Automation Chinese Academy of Sciences
Jiancheng Yu
State Key Laboratory of Robotics Shenyang Institute of Automation Chinese Academy of Sciences
Zhiyu Liu
Xiamen University

Research Article

Keywords: Autonomous Underwater Vehicle, Trajectory Planning, Deep Reinforcement Learning, Time-varying Current, Spatio-temporal Encoding

Posted Date: June 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9924700/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Deep Reinforcement Learning-Based AUV Trajectory Planning in 3D Time-Varying Current Environments

Jun Xiao[a], Ranshuai Zheng[a], Xiaolin Bai[a,b,*], Zhiduo Tan[c,d], Jiancheng Yu[c,d] and Zhiyu Liu[a,b]

[a]State Key Laboratory of Marine Environmental Science & College of Ocean and Earth Sciences, Xiamen University, Xiamen, 361102, China
[b]Fujian Ocean Innovation Center & Laboratory Fujian Institute, Xiamen, 361102, China
[c]State Key Laboratory of Robotics Shenyang Institute of Automation Chinese Academy of Sciences Shenyang 110016 China
[d]Institutes for Robotics and Intelligent Manufacturing Chinese Academy of Sciences Shenyang 110016 China

ARTICLE INFO

Keywords:
Autonomous Underwater Vehicle
Trajectory Planning
Deep Reinforcement Learning
Time-varying Current
Spatio-temporal Encoding

ABSTRACT

Autonomous Underwater Vehicles (AUVs) are highly dependent on the effective utilization of complex ocean currents to optimize mission duration and energy consumption during long-range deployments. Compared to traditional graph-search or heuristic algorithms, Reinforcement Learning (RL) is inherently advantageous for dynamic current compensation under multiple constraints. However, a critical limitation persists: existing RL methods are restricted to reactive navigation, which relies on instantaneous in-situ observations and ignores the long-range temporal correlations of time-varying ocean currents. To address this limitation, this paper proposes a novel Global Trajectory Planning framework based on Deep Reinforcement Learning (DRL). We use high-resolution CMEMS GLORYS12V1 reanalysis data to construct a high-fidelity 3D dynamic environment that closely mirrors real-world oceanographic conditions. To enable the agent to proactively predict and adapt to hydrodynamic evolution, we employ a spatio-temporal composite encoding mechanism (Auxiliary module, AUX), which effectively extends the agent's perception beyond purely spatial considerations to encompass the temporal evolution of the flow field. The experimental results show a 4.62% reduction in travel time compared to the fastest traditional method, A*. Further validation in four globally distinct regions proves that the algorithm can autonomously discover vertical shear in the 3D flow field and execute depth switching, fully demonstrating the universality and robustness of the algorithm in complex environments.

1. Introduction

Autonomous Underwater Vehicles (AUVs) have transitioned from auxiliary tools to essential platforms for deep-marine resource exploration and defense missions. Operating in communication-constrained and uncertain underwater environments, AUVs rely on robust autonomous decision-making (Yu et al., 2024). As a high-level decision module, trajectory planning aims to generate optimal, collision-free trajectories in complex 3D environments while satisfying physical constraints, such as propulsion velocity limits and safe-depth bounds. However, dynamic marine flows serve as both substantial disturbances and critical opportunities for energy optimization: neglecting flow dynamics induces trajectory deviation and premature energy depletion, whereas proactive exploitation of environmental hydrodynamics can substantially enhance mission endurance (Liu et al., 2011; Fan et al., 2025).

Traditional geometric search and stochastic sampling algorithms constitute the foundation of trajectory planning. Specifically, the A^* algorithm (Hart et al., 1968) and the Fast Marching Method (FMM) (Sethian, 1996) derive globally cost-optimal solutions through deterministic grid discretization. Conversely, the Artificial Potential Field (APF) (Khatib, 1986) and Rapidly-exploring Random Trees (RRT) (LaValle, 1998) improve computational efficiency and obstacle avoidance by employing virtual force fields and stochastic sampling, respectively. To address the curse of dimensionality—the exponential growth of computational overhead in high-dimensional state spaces—research has introduced dimensionality reduction and Voronoi-based hybrid frameworks (Nie et al., 2020; Zhao et al., 2022). Advanced variants, including Informed RRT* and Dynamic Artificial Potential Fields (DAPF), further optimize trajectory costs while mitigating local optima stagnation (Zhao et al., 2025a; Zhang et al., 2025a). In multi-agent coordination, FMM and rotational potential fields are utilized to coordinate rendezvous points and evade trapping regions (Kim et al., 2024; Pan et al., 2021). However, these algorithms predominantly rely on static environmental assumptions, constraining their efficacy in time-varying, nonlinear hydrodynamic flows.

*Corresponding author
xbai@xmu.edu.cn (X. Bai)
ORCID(s): 0000-0001-5044-509X (X. Bai)

Bio-inspired swarm intelligence algorithms are utilized to address nonlinearities and multi-objective optimization challenges (Wu et al., 2025). Recent developments in Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and the Snake Optimizer improve search efficiency and trajectory quality in dynamic environments, as well as address heterogeneous multi-agent path planning in complex water missions (Xin et al., 2019; Lim et al., 2018; Wen et al., 2024; Zhao et al., 2025b). Despite their optimization capabilities, these algorithms entail significant computational overhead due to iterative convergence requirements and parameter sensitivity, limiting their applicability in time-critical and instantaneous engineering scenarios.

Deep Reinforcement Learning (DRL) serves as a robust paradigm for sequential decision-making in uncertain and dynamic environments. Within continuous control domains, algorithms such as DDPG and curriculum learning-enhanced Proximal Policy Optimization (PPO) facilitate end-to-end trajectory planning and robust 3D tracking amidst environmental disturbances (Carlucho et al., 2018; Havenstrøm et al., 2021). To improve energy efficiency and navigational safety, value-based methodologies integrated with Prioritized Experience Replay (PER) and Long Short-Term Memory (LSTM)-based obstacle prediction architectures are utilized (Huang et al., 2025; Zhu et al., 2022). Recent advancements refine these frameworks by mitigating training instabilities arising from flow-induced errors and minimizing sensory energy expenditure (Li et al., 2023; Liu et al., 2025a). Furthermore, state-of-the-art approaches integrate Large Language Models (LLMs) and human-in-the-loop mechanisms to enhance generalization capabilities and sample efficiency (Wen et al., 2025; Jiang et al., 2022). Consequently, the application of DRL encompasses diverse operational scenarios, including Electronic Chart Display and Information System (ECDIS)-assisted planning, multi-AUV cooperative observation, and edge computing paradigms (Marchel et al., 2025; Zhang et al., 2025b; Li et al., 2025).

Despite progress, existing AUV trajectory planning frameworks encounter bottlenecks within the intricate dynamics of marine environments. First, a discrepancy persists between idealized flow modeling and empirical marine conditions; most studies rely on simplified mathematical models or Gauss-Markov processes to simulate flows, failing to capture the intrinsic historical patterns of real-world marine data (Li and Yu, 2023; Hadi et al., 2022). Second, although recent investigations have integrated 3D spatial constraints, obstacle avoidance, and time-varying flows using empirical and assimilated data for trajectory planning within specific regions such as the Luzon Strait (Wen et al., 2025), a gap remains regarding algorithmic generalization across a broader spectrum of globally diverse and hydrodynamically complex marine regions. Furthermore, existing research seldom incorporates algorithmic enhancements explicitly targeting the highly time-varying nature of marine flows. Although certain frameworks attempt to anticipate hydrodynamic evolution, they rely on computationally exhaustive iterative integration that hinders real-time onboard planning (Smith et al., 2010). Conversely, most conventional RL models remain restricted to the spatial perception of instantaneous, in-situ flows, limiting the capacity to anticipate future hydrodynamic evolution. Consequently, agents are susceptible to myopic decision-making—prioritizing immediate kinematic advantages at the expense of global energy optimality and trajectory stability throughout extended deployments.

To address the limitations in contemporary AUV trajectory planning research, the primary contributions of this study are as follows:

(1) Establishment of a high-fidelity training environment driven by global marine reanalysis data: Departing from traditional simplified mathematical models, this framework integrates high-resolution empirical marine data—incorporating dynamic flows and relevant operational depth strata—to narrow the sim-to-real gap. This environment serves as a robust benchmark for training and evaluating AUV trajectory planning algorithms.

(2) Enhancement of autonomous navigation and generalization via DRL: By addressing the coupled effects of 3D spatial constraints, intricate topographical obstacles, and dynamic environmental perturbations, this study transcends the limitations of navigation planning restricted to single geographic domains. Beyond validating baseline efficacy, this research rigorously evaluates algorithmic generalizability across physically diverse and challenging environments, addressing a significant void in existing literature.

(3) Development of an auxiliary temporal prediction module (which we term AUX) to mitigate dynamic uncertainties:To overcome the reactive limitations of traditional RL in dynamic settings, this module extends the agent ' s perception into the temporal dimension of the flow field, enabling the anticipation of hydrodynamic evolution. Leveraging a spatio-temporal encoding mechanism, the agent counteracts disturbances induced by time-varying flows, enhancing both long-term trajectory optimality and navigational robustness.

2. Problem Modeling

This section establishes the mathematical framework for the AUV trajectory planning and tracking

problem. Specifically, it formulates optimization objectives, the high-fidelity simulation environment derived from marine reanalysis data, the AUV kinematic model, energy expenditure dynamics, and core evaluation metrics.

2.1. Optimization Objective

The optimization problem aims to determine an optimal trajectory that balances mission duration and energy economy. The objective function J is formulated as a weighted sum of the total travel time T_f and total energy expenditure, where λ is a weighting coefficient and P represents the power expenditure depending on the control input $u(t)$. This optimization process is subject to a series of constraints comprising physical limits and environmental safety requirements: the AUV must satisfy fixed initial and terminal states $(\mathbf{x}_{\text{start}}, \mathbf{x}_{\text{goal}})$; the kinematics are governed by system dynamics f including current perturbations; the propulsion velocity $\mathbf{v}_r$ is bounded by the maximum propulsion speed $v_{\max} = 1$ m/s; for vertical safety constraints, the AUV must strictly remain within the allowable operating depth bounds $[z_{\min}, z_{\max}]$ (e.g., 0-60 m); the planar trajectory must strictly avoid predefined obstacle regions O; finally, the total mission time is constrained by an upper threshold $T_{\max}$ (set to 222.22 h).

$$
\begin{aligned}
\min_{u(t)} \quad & J = T_f + \lambda \int_0^{T_f} P(u(t))\, dt \\
\text{s.t.} \quad & \mathbf{x}(0) = \mathbf{x}_{\text{start}}, \quad \mathbf{x}(T_f) = \mathbf{x}_{\text{goal}} \\
& \dot{\mathbf{x}}(t) = f(\mathbf{x}(t), u(t), v_{\text{curr}}(\mathbf{x}(t), t)) \\
& \|v_r(t)\| \le v_{\max}, \quad \forall t \\
& z_{\min} \le z(t) \le z_{\max}, \quad \forall t \\
& (x(t), y(t)) \notin O, \quad \forall t \\
& T_f \le T_{\max}
\end{aligned}
\tag{1}
$$

2.2. Environment Modeling

The dynamic environmental flow field is sourced from the GLORYS12V1 global physical reanalysis dataset. This dataset provides high spatio-temporal resolution across relevant operational layers with a horizontal grid spacing of 1/12° (approximately 8 km). Such high fidelity enables the precise resolution of complex flow patterns and gradient structures, which are critical for energy-optimal trajectory planning. Furthermore, to establish high-fidelity physical boundaries for the simulated environment, the framework integrates high-resolution elevation data from the GEBCO dataset. This dataset provides precise topographical profiles that, in conjunction with predefined operational altitude bounds, delineate navigable contours and restricted zones, thereby imposing rigorous geometric constraints upon the agent's state space.

Within the 3D Euclidean space $\mathbb{R}^3$, the marine flow is formulated as a depth-dependent, time-varying horizontal vector field. Given that vertical hydrodynamic velocities are negligible relative to their horizontal counterparts in open-marine domains, this model isolates the eastward u and northward v components, while strictly preserving their spatial dependence on the depth stratum z:

$$
\mathbf{V}_c(x, y, z, t) = [u(x, y, z, t), v(x, y, z, t), 0]^T \tag{2}
$$

Through multi-dimensional interpolation of the discrete GLORYS12V1 gridded data and GEBCO bathymetric datasets, the continuous hydrodynamic velocity profiles and localized navigability attributes at the AUV's instantaneous coordinates are derived. To enhance policy robustness, the training regimen encompasses extensive episodic rollouts, where the hydrodynamic environment for each simulation episode is initialized via stochastic sampling of temporal indices. This stochasticity ensures the model guarantees navigational safety within bathymetric bounds while developing a generalized capacity to proactively exploit highly dynamic flows, thereby preventing overfitting to isolated spatio-temporal flow patterns.

2.3. Kinematic Modeling

For kinematic formulation, the AUV is modeled as a point mass traversing a dynamic 3D marine flow field. Let $\mathbf{p}(t) = [\lambda(t), \phi(t), z(t)]^T$ denote the instantaneous 3D coordinates of the AUV at time t, where λ, ϕ, and z represent longitude, latitude, and depth, respectively. The ground velocity $\mathbf{v}$ of the AUV is governed by the vector summation of its propulsion velocity and the ambient marine flow.

In practical marine deployments, due to hydrodynamic shielding induced by hull geometry and subgrid scale uncertainties within marine flow data, the influence of the flow on the AUV deviates from linear superposition. Consequently, this kinematic formulation incorporates a flow influence factor η and a hydrodynamic drag coefficient C_d. As an empirical parameter characterizing the equivalent kinematic transfer efficiency of ambient flows, η typically approximates unity and is calibrated based on mission-specific constraints and localized hydrodynamic profiles.

Let the horizontal control action be $\mathbf{a}_{xy} \in [-1, 1]^2$, yielding a nominal propulsion velocity $\mathbf{v}_r = \mathbf{a}_{xy} \cdot v_{max}$. To optimize the efficiency ratio between baseline load and hydrodynamic drag for long-endurance AUVs, the maximum propulsion velocity is set to $v_{max} = 1$ m/s (Hobson et al., 2012). The horizontal kinematics follow the modified equation:

$$
\mathbf{v}(t) = \left(\mathbf{v}_r(t) + \eta \cdot \mathbf{v}_c(\mathbf{p}, t)\right) \cdot (1 - C_d) \tag{3}
$$

where $\mathbf{v}_c(\mathbf{p}, t)$ denotes the instantaneous marine flow velocity extracted from the GLORYS12V1 dataset. Vertically, the AUV executes ascending or descending maneuvers via buoyancy regulation driven by depth control directives, updating the vertical state.

As the simulation environment utilizes a global geodetic coordinate system, the ground velocity vector $\mathbf{v}$ is transformed into the time derivatives of longitude and latitude using the instantaneous latitude ϕ and a geodetic conversion factor ($R \approx 111,000$ m/degree):

To maintain simulation fidelity within high-resolution marine flow data exhibiting spatio-temporal gradients and to mitigate integration drift inherent in the first-order Euler scheme, this framework adopts the second-order Runge-Kutta (RK2) method for the discrete kinematic evolution of AUV coordinates over a time step Δt. Finally, bathymetric constraints are imposed on the state space to ensure navigation within permissible depth strata and navigable marine domains.

2.4. Energy Consumption Model

Given the finite battery capacity of AUVs, energy expenditure is a primary objective in trajectory planning. Expenditure is predominantly dictated by the propulsive power required to counteract hydrodynamic drag. Assuming the drag F_d is proportional to the square of the propulsion velocity magnitude ($\|\mathbf{v}_p\|^2$), the propulsive power P_{prop} exhibits a cubic dependency on this velocity (Liu et al., 2025b):

$$P_{prop} \propto \|\mathbf{v}_r\|^3 \tag{4}$$

Moreover, for a mission spanning a fixed distance D, the transit time t is inversely proportional to the propulsion velocity magnitude ($t \propto D/\|\mathbf{v}_p\|$). Consequently, the total energy expenditure E_{total} scales quadratically with the propulsion velocity, as the reduced transit time only partially offsets the cubic escalation in power:

$$E_{total} = P_{prop} \cdot t \propto \|\mathbf{v}_r\|^3 \cdot \frac{D}{\|\mathbf{v}_r\|} \propto \|\mathbf{v}_r\|^2 \tag{5}$$

This quadratic relationship ($E_{total} \propto \|\mathbf{v}_r\|^2$) highlights the nonlinear energy cost associated with high-speed navigation. Additionally, AUVs have a baseline power load that consumes energy continuously over time. Therefore, the formulated reward function implicitly encourages the agent to optimize this trade-off by introducing a time penalty term and effectively utilizing ocean currents. By navigating with the flow, the AUV maximizes its absolute ground velocity $\mathbf{v}$ (thereby reducing mission time t and lowering baseline energy consumption) while maintaining a moderate propulsion velocity magnitude $\|\mathbf{v}_r\|$ (minimizing propulsive power P_{prop}), ultimately achieving minimum total energy consumption.

2.5. Evaluation Metrics

To rigorously evaluate the performance of the proposed algorithm against traditional baseline algorithms such as A^*, FMM, and APF, the following quantitative evaluation metrics are adopted, as summarized in Table 1.

Table 1
Evaluation Metrics for AUV trajectory Planning and Tracking

Metric	Symbol	Unit
Average Reward	R_{avg}	N/A
Convergence Speed	N_{conv}	Step
Success Rate	SR	%
Navigation Time	T_{nav}	s
Energy Consumption	E_{total}	J
Computation Time	t_{comp}	ms

3. Methodology

3.1. Observation Space

To enable the agent to comprehensively perceive 3D ocean environment dynamics and execute effective long-term trajectory planning, the observation space is formulated as 2D feature maps dynamically extracted based on the AUV's instantaneous depth. Although actual ocean currents exhibit complex vertically stratified structures, horizontal hydrodynamic profiles at the AUV's corresponding depth are sufficient to provide essential macro-environmental states for the navigation task while reducing state-space dimensionality.

As depicted in Fig. 1, the marine environment is formulated as multiple layers. Conditioned on the AUV's instantaneous depth, three channels of 2D feature maps are extracted from the 3D hydrodynamic field. Specifically, the primary channel constitutes the Navigation Status Map, which integrates the AUV's instantaneous spatial coordinates, target coordinates, and land-sea topological information. This furnishes explicit geometric and physical constraints, directing the AUV toward the destination while ensuring the planned trajectory strictly remains within the navigable marine domains at the corresponding depth. The remaining two channels encode the u-component (velocity along the x-axis) and v-component (velocity along the y-axis) of the localized hydrodynamics. These maps characterize the hydrodynamic profiles at the corresponding depth stratum, enabling the agent to effectively exploit prevailing flows and circumvent adverse flows.

3.2. Reinforcement Learning Architecture Based on Spatio-Temporal Encoding

This research employs the PPO algorithm as the core decision engine for global trajectory planning. To effectively process the high-dimensional, multi-channel spatio-temporal data derived from the observation space, the network architecture is redesigned to incorporate a spatio-temporal encoding mechanism. As illustrated in Fig. 2, the architecture comprises three core modules: Observation Input, Spatio-Temporal Encoding, and the PPO Agent.

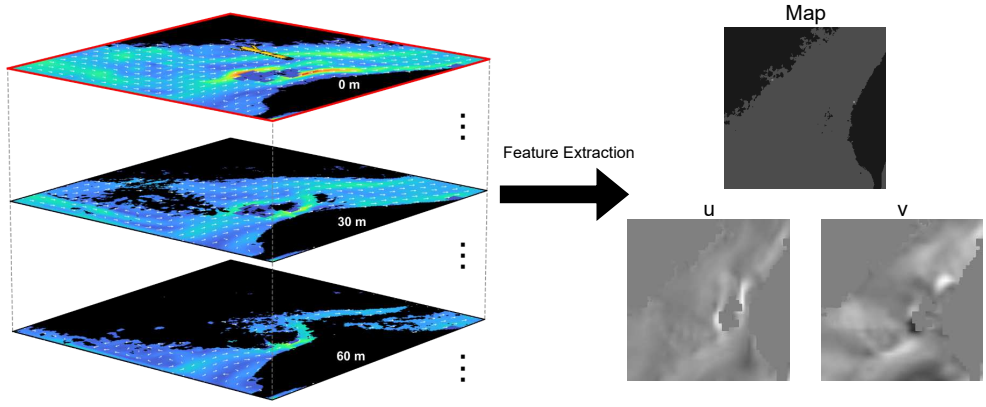


Figure 1: Input Feature Construction: The vertical stratified structure on the left represents the 3D ocean current environment; the right side illustrates three 2D feature map channels extracted based on the AUV's current depth.

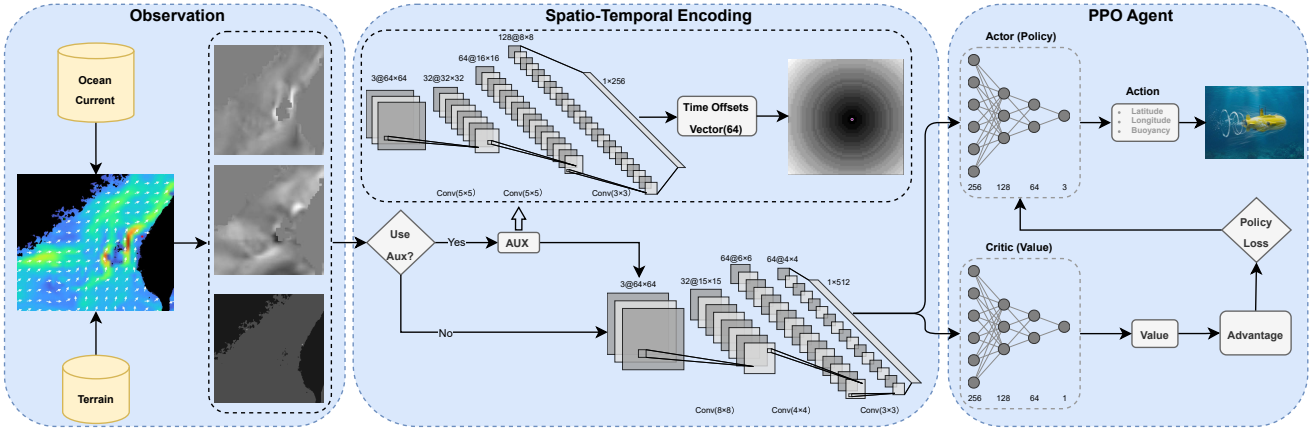


Figure 2: trajectory Planning Network Architecture. The architecture includes observation inputs, a spatio-temporal encoder (covering spatial feature extraction and AUX future flow prediction), and a PPO decision agent consisting of an Actor and a Critic.

Mechanism and Motivation of the Future Flow Prediction Module (AUX). In an ideal steady hydrodynamic environment, distal flow information is static and remains valid over time. However, in actual time-varying marine environments, distal flow fields evolve significantly during long-endurance transits, making trajectories based on instantaneous static observations prone to failure. To address this, the AUX module is introduced. Its core mechanism involves estimating the time delay required for the AUV to reach different spatial locations (derived from the propulsion velocity and distance) and updating the distal hydrodynamic information to its corresponding future state. After AUX processing, the feature map is transformed from a single-frame spatial slice into a "spatio-temporal composite map"—where different pixels encode hydrodynamic profiles at different coordinates and future times. This allows the agent to anticipate and exploit dynamic environmental patterns via temporal shifting. For areas near the AUV, the impact of temporal shifting is negligible, ensuring the reliability of short-range trajectory planning and making the AUX module a "lossless" enhancement to baseline efficacy. In summary, the core logic of the AUX module aligns closely with actual physical processes, demonstrating strong physical interpretability.

Self-Supervised Training Methodology of the AUX Module. To cultivate this predictive capability, a self-supervised learning paradigm is adopted, leveraging the AUV's empirical trajectory rollouts. The training labels are constructed retrospectively at the conclusion of each navigational episode. For a specific observation sampled at time t and spatial coordinate p_t, a sequence of N concentric radial boundaries $\{r_j\}_{j=1}^{N}$ is defined, scaling outwards to a maximum traversal distance r_{max}. By scanning the recorded subsequent trajectory, the actual temporal delay Δt_j required for the AUV to first intersect each radius r_j is extracted. These temporal delays are normalized against a predefined maximum forecasting horizon T_{max}, generating a ground-truth delay vector $\mathbf{y} \in [0,1]^N$. The AUX network outputs a corresponding predicted vector $\hat{\mathbf{y}}$ via its terminal Sigmoid activation. The network parameters are optimized using the Mean Squared Error (MSE)

loss function:

$$\mathcal{L}_{AUX} = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} (\hat{y}_{i,j} - y_{i,j})^2 \quad (6)$$

where M denotes the batch size sampled from a dedicated experience replay buffer. This buffer archives historical trajectory states and labels, enabling off-policy periodic batch optimization (via the Adam optimizer) every set number of episodes. This decoupled training mechanism prevents catastrophic forgetting of diverse environmental interactions and stabilizes the feature representation alongside the high-frequency, on-policy temporal-difference updates of the PPO agent.

Structural Implementation of the Spatio-Temporal Encoding Network. The agent's baseline input comprises instantaneous hydrodynamic and topographical states concatenated into a $3 \times 64 \times 64$ three-channel tensor. The encoding network utilizes a conditional routing mechanism to regulate data flow: when the AUX module is enabled, the input passes through a dedicated flow prediction pipeline consisting of sequential convolutional layers (two 5×5 and one 3×3), transforming feature maps into 32, 64, and 128 channels. The resulting 256-dimensional flattened vector is projected onto the aforementioned N-element Temporal Offset Vector (where $N = 64$). This 1D vector is then expanded into a 2D concentric mask via a rotational expansion mechanism, establishing a grayscale gradient from black to white centered on the AUV's position. This mapping assigns temporal offsets to the spatial grid: whiter regions (distal to the AUV) represent flow data projected further into the prediction. Finally, regardless of the active branch, the features are processed by the backbone architecture through three progressive convolutional layers (kernels: 8×8, 4×4, 3×3) for multi-scale spatial feature extraction, yielding a 512-dimensional latent feature vector for the PPO agent.

Decision Agent (Actor-Critic). The latent feature vector is processed by the Actor (policy) and Critic (value) branches. Both networks employ independent Multi-Layer Perceptrons (MLPs) with hidden layer sizes of 256, 128, and 64 neurons. The Actor outputs three-dimensional continuous control actions: latitudinal, longitudinal (for horizontal navigation), and buoyancy (for depth regulation) commands. The Critic provides a scalar state-value estimate used to compute the advantage function and facilitate policy gradient updates.

3.3. Reward Function

The reward function constitutes the fundamental driver of the learning process. To direct the AUV toward the destination while maintaining safety and efficiency, a composite reward function r_{total} is formulated as:

Table 2
Hyperparameters and Experimental Settings

Parameter	Value
PPO Algorithm	
Learning Rate (α)	3×10^{-4}
Horizon / Steps per Update (T)	2048
Mini-batch Size	64
Update Epochs (K)	10
Discount Factor (γ)	0.99
GAE Parameter (λ)	0.95
Clipping Range (ϵ)	0.1
Target KL Divergence	0.015
Entropy Coefficient (c_2)	0.005
Value Function Coefficient (c_1)	0.5
Max Gradient Norm	0.5
Network Architecture	
Policy Type	CnnPolicy
Extracted Feature Dimension	256
Hidden Layers (Actor & Critic)	[128, 64]
Activation Function	ReLU
Training Environment	
Total Timesteps	10×10^6
Number of Parallel Environments	4
Input Resolution ($H \times W$)	64×64
Max Steps per Episode	400
Simulation Step Size (Δt)	2000 s

$$r_{total} = r_{progress} + r_{ocean} - r_{collision} - r_{time} - r_{energy} \quad (7)$$

The components are defined as follows:

- $r_{progress}$: Navigation progress reward, comprising a dense component proportional to the decrement in Euclidean distance to the destination and a sparse bonus R_{goal} upon goal achievement.
- r_{ocean}: Flow exploitation reward, incentivizing the strategic utilization of currents by rewarding alignment with the target direction and penalizing opposing flows.
- $r_{collision}$: Collision penalty, imposing a substantial negative reward for obstacle contact and triggering immediate episode termination.
- r_{time}: Temporal penalty, consisting of a constant step-wise cost to incentivize minimized mission duration.
- r_{energy}: Energy expenditure reward, proportional to the calculated consumption (modeled as $\propto \|\mathbf{v}_r\|^3$).

Table 3
Performance Comparison of trajectory Planning Algorithms in Different Flow Environments

Environment	Method	Success (%)	Time (h)	Energy (kWh)	Power (W)	Comp (ms)
Steady Flow	RL	100	54.90	1.98	40.00	627
	APF	100	108.68	3.41	34.82	3920
	FMM	100	61.04	2.06	37.41	1882
	A*	100	56.78	2.04	40.00	382
Time-varying Flow	RL	99	55.97	2.01	39.95	782
	RL (AUX)	100	55.17	1.98	39.98	1598
	RL (Energy Penalty)	100	81.40	1.10	13.00	896
	APF	100	111.32	3.12	31.16	4860
	FMM	100	62.82	2.12	37.47	1984
	A*	100	57.84	2.08	40.00	476

3.4. Hyperparameter Settings

The hyperparameters utilized for training the neural network are tabulated inTable 2. These values were determined through empirical tuning to balance exploration and convergence.

In the PPO configuration, the CnnPolicy is utilized to process visual inputs with a learning rate of 3×10^{-4}. Specifically, the clipping range (ϵ) constrains the ratio between new and old policies to prevent excessive updates; concurrently, the target KL divergence quantifies the distributional shift, serving as an early stopping mechanism to regulate the policy evolution step size. Initial experiments revealed that default PPO parameters resulted in policy collapse during late-stage training, characterized by a sharp decline in the reward curve. To mitigate this, through empirical tuning, ϵ is set to a conservative 0.1 and the target KL divergence is limited to 0.015. This configuration stabilizes policy updates, facilitating consistent convergence. While strict step-size limits slightly reduce exploration to enhance stability, these parameters represent a critical trade-off between environmental exploration and stable policy convergence.

Regarding the training configuration, the total timesteps are set to 10×10^6 to ensure sufficient experience acquisition. To enhance sampling efficiency, 4 parallel environments are executed concurrently. The environmental input resolution is standardized to 64×64. Additionally, the maximum steps per episode is capped at 400, with a simulation step size (Δt) of 2000 s. Upon convergence, the AUV's mean mission duration stabilizes at approximately 100 steps (a physical duration of roughly 55.56 h). Consequently, the agent undergoes approximately 10^5 training episodes, facilitating robust extraction of complex hydrodynamic features and trajectory planning strategies.

4. Experimental Results

To systematically evaluate the proposed RL trajectory planning framework, this section conducts comprehensive simulation experiments using a high-resolution ocean reanalysis dataset. The experimental validation encompasses four aspects: Section 4.1 demonstrates the necessity of the vertical dimension through a comparison of 2D and 3D trajectory planning; Section 4.2 validates the core role of the auxiliary prediction (AUX) module in expanding the temporal dimension for dynamic flow fields, as well as benchmark testing; Section 4.3 analyzes seasonal adaptability and depth-switching strategies in monsoon-driven waters; finally, Section 4.4 demonstrates the algorithm' s generalization and robustness across four global marine regions with distinct characteristics.

4.1. Comparison of 2D and 3D AUV Trajectory Planning Scenarios

A salient characteristic of the marine environment is the existence of vertical current shear—whereby flow velocity and orientation vary significantly with depth. To verify the necessity of integrating the vertical dimension into trajectory planning, the performance of the proposed 3D model is compared against a constrained 2D baseline, where the AUV is restricted to a constant shallow depth stratum.

A visual comparison of the planned trajectories is depicted in Fig. 3. Specifically, the background color gradient represents the flow intensity, while the AUV trajectory color indicates its instantaneous cruising depth. In 2D mode (left column), the AUV is continuously exposed to strong surface flows. Lacking vertical maneuvering capabilities, the agent compensates by exhibiting diverse horizontal detour strategies to circumvent adverse surface flows. In contrast, the 3D model (right column) demonstrates a distinct "depth-switching" capability. As observed in the 3D view, the agent maintains a relatively consistent horizontal trajectory; vertically, it autonomously dives to deeper strata with lower hydrodynamic resistance, effectively bypassing strong surface counter-flows.

The quantitative training curves depicted in Fig. 4 further validate these qualitative observations. The most notable difference is reflected in the *flow reward* (middle subplot). Compared to the 2D model (blue

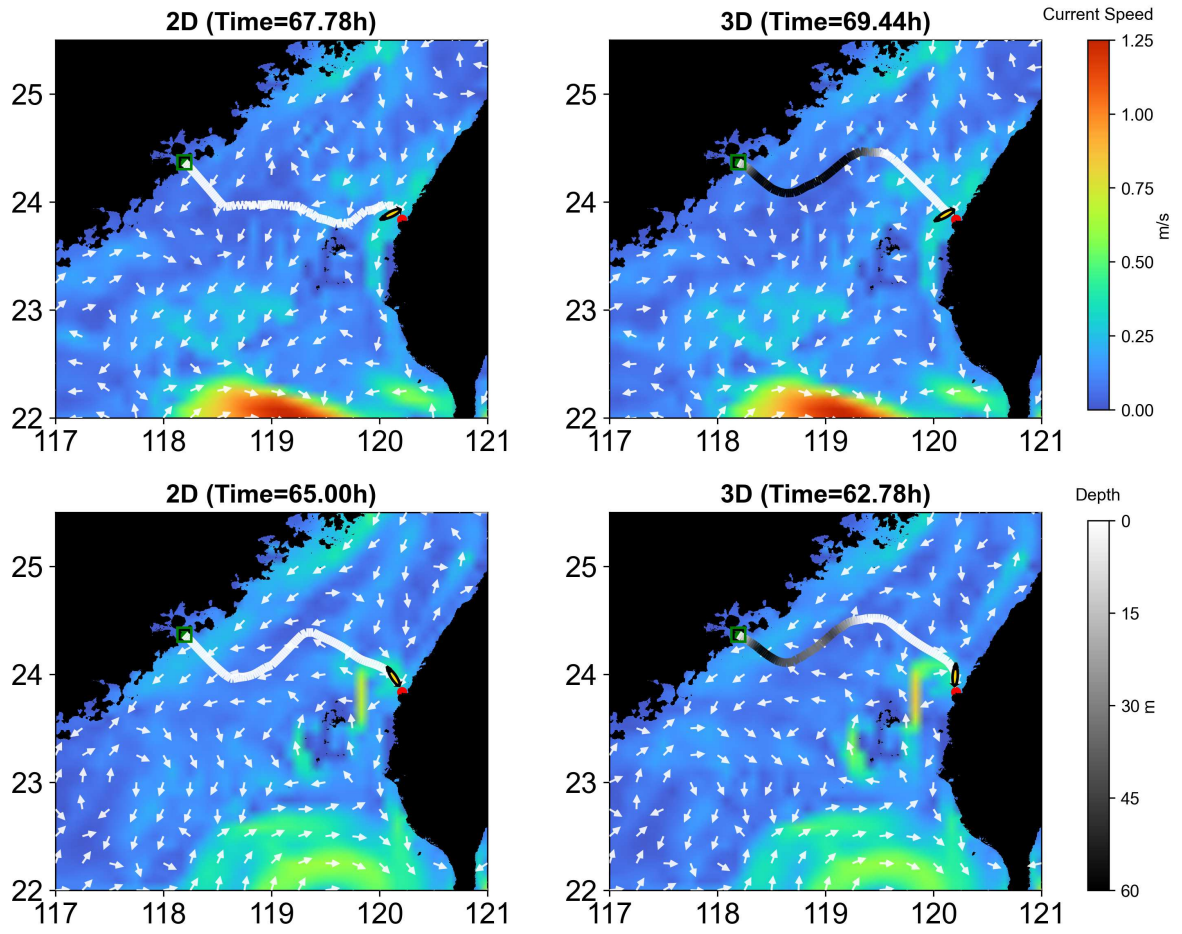


Figure 3: Performance Comparison of Trajectory Planning Algorithms across Diverse Marine Flow Environments

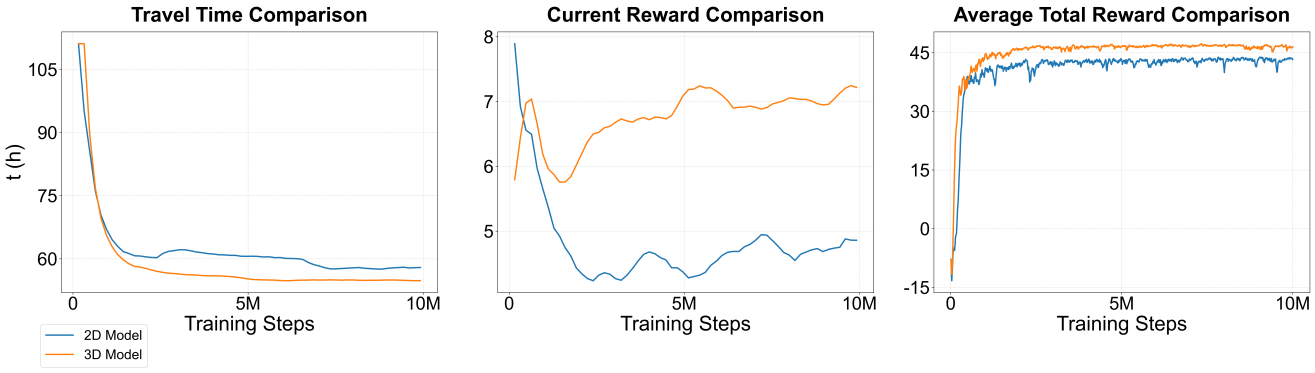


Figure 4: Comparative training curves for the 2D model (blue) and 3D model (orange). From left to right: mission duration, flow exploitation reward, and mean total reward. The 3D model achieves higher flow rewards, indicating efficient utilization of vertical shear.

curve), the 3D model (orange curve) attained higher flow-related rewards, indicating that the 3D agent successfully learned to identify and exploit favorable flow strata rather than passively counteracting surface flows. Consequently, the 3D model converged to a higher *mean total reward* (right subplot) and exhibited superior stability in *mission duration* (left subplot), proving that expanding the action space to 3D is essential for realizing energy-efficient navigation in complex marine environments.

The pronounced performance gain of the 3D model depicted in Fig. 4 is localized to the Taiwan-to-Xiamen route (i.e., the reciprocal of the original mission). Conversely, on the Xiamen-to-Taiwan route, the discrepancy between the two models is marginal. This asymmetry underscores that the interaction between

trajectory orientation and prevailing flow structures dictates the variance in planning outcomes.

4.2. Benchmark Evaluation and Dynamic Flow Field Analysis

To provide a comprehensive evaluation of the proposed RL-based trajectory planning framework, comparative experiments are conducted against three traditional baseline methods—APF, FMM, and the A* algorithm—under both steady and dynamic flow conditions. The quantitative performance metrics for all algorithms are summarized in Table 3, where all values represent the mean statistics derived from 1,000 independent test missions.

Time-Varying Flow Fields and the Role of the AUXiliary Prediction (AUX) Module. Marine currents are inherently time-varying; thus, the "steady-state environment" serves as an ideal baseline for evaluating spatial trajectory planning capabilities. As tabulated in Table 3, the transition from steady to dynamic flow conditions increases the mean mission duration across all algorithms due to persistent environmental disturbances. Absent the AUX module, the RL agent exhibits an increase in mean mission duration from 54.90 h to 55.97 h, with the success rate declining to 99% due to an inability to anticipate hydrodynamic evolution. Upon incorporating the AUX module, the agent (RL-AUX) ensures a 100% success rate through explicit prediction of flow field evolution, with the mission duration reduced to 55.17 h. This enhancement narrows the performance discrepancy between time-varying and steady flows, approaching the ideal steady-state spatial planning baseline. Consequently, the AUX module effectively perceives and mitigates drift in dynamic environments, ensuring mission success while optimizing navigation efficiency.

Asymmetric Benchmark Setup and Robustness. In comparing against traditional baseline algorithms, the structural asymmetry in the evaluation setup is noteworthy: to account for the intrinsic constraints of traditional planners, baseline methods were evaluated within a lenient "soft collision" environment (incorporating smoothed coastlines and artificial safety buffers); conversely, the RL-based methods were tested under a rigorous "hard collision" regime devoid of artificial safety margins. As tabulated in Table 3, notwithstanding the more stringent environmental constraints, the RL-AUX agent attained a 100% success rate, comparable to the baseline algorithms operating under permissive conditions. This demonstrates exceptional environmental adaptability and robustness.

Navigation Efficiency Analysis. Regarding navigation efficiency, the RL framework demonstrates superior performance relative to traditional baseline methods. The power assessment focuses exclusively on propulsion power, with the peak power constrained to 40 W. In dynamic flow environments, the mean mission duration of RL-AUX (55.17 h) is shorter than the best-performing baseline, A* (57.84 h), and significantly lower than FMM (62.82 h) and APF (111.32 h). Compared to A*, RL-AUX reduces mission duration by approximately 4.6%, while achieving an energy expenditure of 1.98 kWh versus A*'s 2.08 kWh. Furthermore, to explore trade-offs under multi-objective optimization, an RL model with an elevated energy penalty weight (RL-Energy) was evaluated. As tabulated in Table 3, although this strategy extends the mean mission duration to 81.40 h, the agent autonomously converges to a conservative transit strategy maintaining a low mean propulsion power of 13.00 W, reducing the cumulative energy expenditure to 1.10 kWh.

Real-Time Computational Performance. Regarding computational latency (Comp), the A* algorithm achieved the minimum cumulative execution time (476 ms) per episode due to its straightforward heuristic search logic. The proposed RL-AUX model maintained a mean inference latency of 1598 ms throughout the mission duration. Given that the total mission duration for a single AUV deployment typically exceeds 50 h, this millisecond-scale latency is effectively negligible in practical engineering contexts and fully satisfies the requirements for real-time online trajectory replanning. Conversely, FMM (1984 ms) and APF (4860 ms) exhibited higher cumulative computational costs per episode due to intensive full-field numerical solving or high-frequency gradient computations, highlighting the superior execution efficiency of the RL-based framework at the deployment end.

4.3. Seasonal Adaptability in Monsoon-Dominated Regions

The Taiwan Strait constitutes a critical narrow waterway connecting the South China Sea and the East China Sea. It serves as a representative microcosm of marine dynamic processes in the marginal seas of the Northwest Pacific, characterized by complex hydrographic conditions that exhibit significant seasonal variations and reversals. This complexity is primarily manifested in three aspects (Hu and Wang, 2016; Hu et al., 2010): First, pronounced seasonal hydrodynamic variability driven by monsoons. Governed by the East Asian monsoon circulation, the region experiences prevailing southwest monsoons in summer and northeast monsoons in winter, inducing seasonal flow reversals. The hydrodynamic fields in winter and summer are diametrically opposed, while spring and autumn serve as transitional periods with substantially increased flow complexity. Second, significant coupling effects between three-dimensional (3D) flow fields and topography. The Strait features a narrow, elongated topography with distinct water depth stratification. Bottom friction and water column stratification jointly

Table 4
Evaluation of Navigation Performance and Robustness under Seasonal Flow Disturbances

Method	Season	Success (%)	Time (h)	Energy (kWh)	Power (W)	Comp (ms)
RL	Winter	100	57.69	2.08	40.00	978
	Spring	100	54.94	1.98	40.00	685
	Summer	98	53.64	1.92	39.83	515
	Autumn	96	57.31	2.06	40.00	897
RL (AUX)	Winter	100	57.32	2.06	40.00	1918
	Spring	100	54.46	1.96	40.00	1247
	Summer	100	52.36	1.88	39.94	838
	Autumn	100	56.53	2.03	39.95	1832

regulate velocity gradients across different strata, resulting in marked contrasts between surface and deep-water flows. Third, extensive seasonal upwelling. Upwelling zones are distributed along the southwestern and northwestern coasts, the Taiwan Banks, and the Penghu Islands. Influenced jointly by monsoons and topography, their intensity and spatial scope evolve dynamically with the seasons, further enhancing the regional flow field's complexity.

Consequently, these characteristics render the region an ideal testbed for evaluating the environmental adaptability and navigation efficacy of intelligent marine vehicles, motivating its selection for this research. Traditional algorithms frequently fail to accurately capture the dynamic variability and 3D stratification of hydrodynamic fields, exhibiting limited robustness under extreme seasonal marine conditions. Conversely, the proposed RL-based trajectory planning framework dynamically adapts to the seasonally shifting stratified flows in the Taiwan Strait, autonomously detecting environmental fluctuations to facilitate optimal decision-making. To rigorously evaluate the agent's responsiveness to these highly seasonal and complex conditions, environmental sampling is strictly confined to the specific month exhibiting the most pronounced monsoon characteristics for each of the four seasons. This ensures a comprehensive reproduction of the most representative hydrodynamic regimes within the region.

The experimental results (tabulated in Table 4 and visualized in Fig. 5) demonstrate that the vertical maneuvering strategy learned by the AUV adapts to the seasonal hydrodynamic conditions of the Taiwan Strait. The lower panels illustrate the vertical depth profile as well as the flow intensity distribution projected along the target route. In summer, the Taiwan Strait is dominated by the intense southwest monsoon, inducing favorable surface flows aligned with the prescribed route. Under these conditions, the agent adopts an optimal policy to maintain a shallow-water depth, maximizing the exploitation of favorable flows. The RL-AUX agent leverages surface flow alignment to optimize navigation progress, achieving a mean mission duration of 52.36 h and an energy expenditure of 1.88 kWh across all evaluated seasons. Conversely, the baseline RL agent, while capable of following flows, exhibits performance degradation, requiring an additional 1.28 h and 0.04 kWh of energy.

In winter, the Taiwan Strait is governed by the robust northeast monsoon, generating powerful adverse surface flows that impede trajectory progress. Faced with these conditions, the AUV consistently executes an autonomous depth-switching strategy, submerged into deeper strata. Due to benthic friction and water column stratification, the velocity of the adverse flow in deeper layers is substantially attenuated, effectively mitigating the impact of surface counter-flows. Although this maneuvering prevents trajectory deviation, navigating against seasonally dominant flows requires overcoming increased hydrodynamic resistance. Consequently, performance metrics for the winter season represent the extremes: the RL-AUX mission duration increases to 57.32 h, with energy expenditure rising to 2.06 kWh. Spring and autumn, as monsoon transitional periods, exhibit moderate flow conditions without extreme following and counter flows. Accordingly, mission duration and energy expenditure demonstrate transitional trends between the winter and summer extremes (e.g., RL-AUX achieved 54.46 h in spring and 56.53 h in autumn), which corresponds with the seasonal hydrodynamic variation patterns of the region (Hu and Wang, 2016).

The complex, volatile, and seasonally shifting marine conditions of the Taiwan Strait challenge intelligent navigation algorithms, underscoring the utility and efficacy of the AUX module. As tabulated in Table 4, the baseline RL agent exhibits a decline in mission success rates during summer and autumn —characterized by complex flow structures—to 98% and 96%, respectively. Conversely, upon incorporating the AUX module, the RL-AUX strategy maintains a 100% mission success rate across all seasonal scenarios, demonstrating superior adaptability compared to traditional methods. Comparative analysis reveals that while the AUX module slightly increases forward propagation latency (e.g., winter single-step execution time

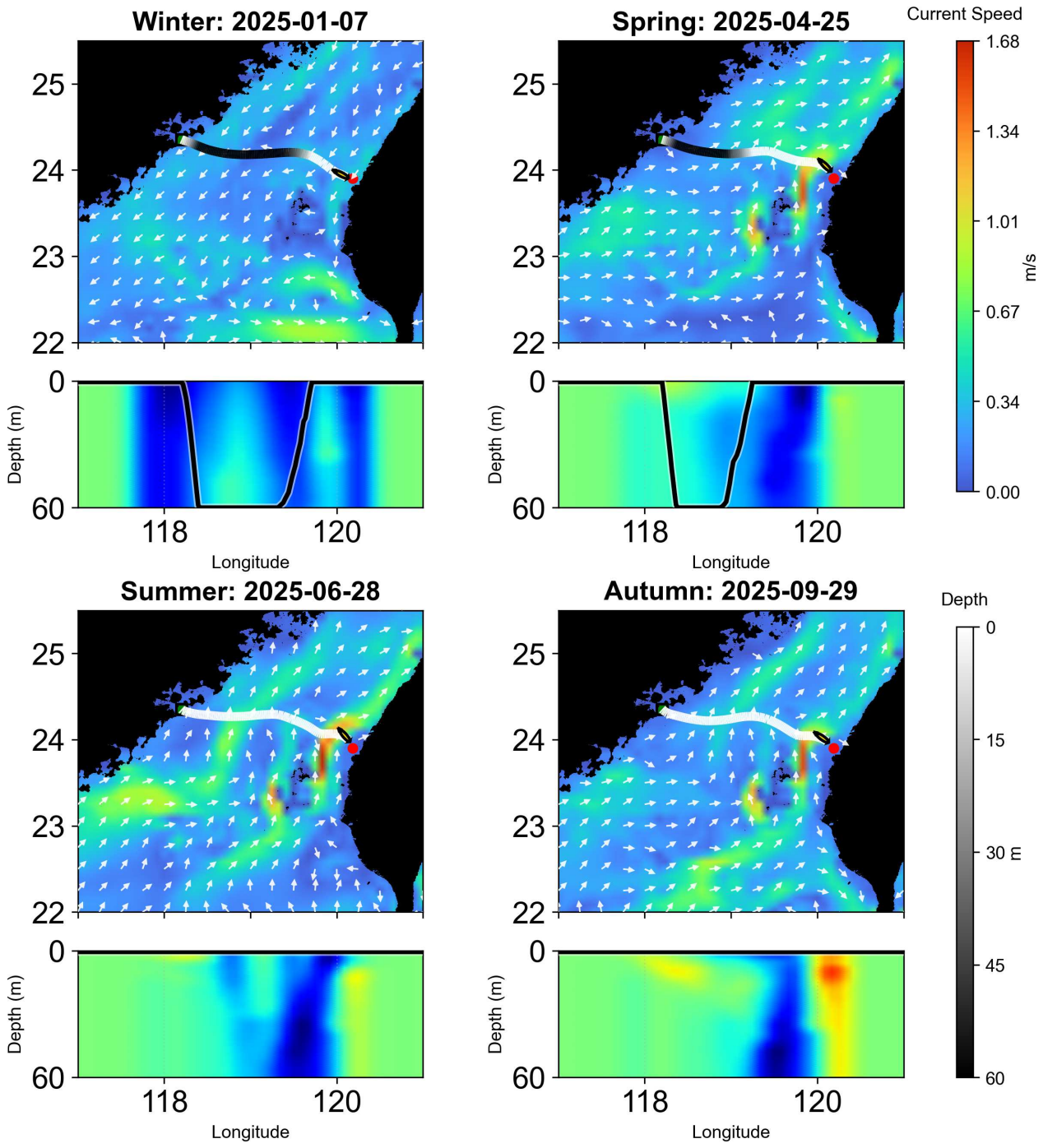


Figure 5: Evaluation of AUV trajectory planning and depth adaptability under varying temporal flow conditions. The figure illustrates horizontal transit trajectories (top) and vertical depth profiles (bottom) within the Taiwan Strait across four distinct temporal periods. The AUV maintains a surface-level trajectory during favorable flow patterns, whereas it adopts a deep-submergence strategy to circumvent intense surface counter-flows. Each subplot depicts the horizontal projection of the AUV trajectory, the vertical profile of hydrodynamic velocities along the trajectory (where the color map represents the flow velocity projected onto the target direction), and the instantaneous vertical position of the AUV (denoted by the black line).

increasing from 978 ms to 1918 ms), it facilitates more precise decision-making tailored to regional marine conditions at the trajectory planning level. Ultimately, this enables the AUV to consistently outperform the baseline RL method in both mean mission duration and energy expenditure across all seasons.

The experimental results validate that when encountering volatile and complex marine conditions such as those in the Taiwan Strait, the AUX module exhibits superior environmental adaptability, robustness, and global optimization capabilities. The framework

precisely captures monsoon-driven flow variations, accounts for hydrodynamic discrepancies across vertical strata, and effectively addresses the challenges of dynamic environment perception and multi-process coupling, adapting to the specific hydrodynamic requirements of monsoon-dominated shelf straits. The pronounced seasonal variations demonstrated in the experiments provide robust empirical evidence for the vertical adaptive capability of the proposed trajectory planning framework. The RL agent dispenses with rigid, pre-defined depth profiles; instead, by integrating instantaneous 3D flow features, it autonomously evaluates environmental conditions, dynamically switching between exploiting surface following-flows in summer and circumventing adverse flows through deep-submergence maneuvers in winter. This approach maximizes mission efficiency under extreme seasonal transitions, further underscoring the methodological rigor and representativeness of the Taiwan Strait as a study region.

4.4. Evaluation of Effectiveness and Generalization Capability Across Different Ocean Regions

To qualitatively evaluate the adaptability of the learned policy and its generalization across diverse marine regions, we selected three representative areas with distinct hydrodynamic features for testing. These comprise a stable open-sea region, a dense reef region with complex bathymetry, and an extreme, highly variable polar region. Together with the monsoon-dominated strait represented by the Taiwan Strait, these characterize the primary categories of global marine environments. The planned trajectories are illustrated in Fig. 6, Fig. 7, and Fig. 8. Each visualization consists of four temporal snapshots, depicting the AUV's instantaneous position, 3D transit trajectory, and the hydrodynamic field at different timesteps. These figures demonstrate the temporal dynamics and spatial distribution of the flows as well as the AUV's 3D spatial response, validating the algorithm's efficacy and generalization capability.

4.4.1. Pacific Ocean: Open Ocean With Stable Flows

The Tuamotu Archipelago, located in the Central Pacific, serves as a representative combination of a stable open-sea region and atoll geomorphology, constituting the least complex environment within the evaluation framework. Its primary characteristics include a year-round stable flow field with low velocities, a uniform thermohaline structure, and negligible seasonal variability (Dumas et al., 2012). The bathymetry is characterized by alternating scattered atolls and expansive open-sea regions. Complex yet weak circumflows and wake vortices develop between atolls (Patzert and Bernstein, 1976), in the absence of intense flow interference. Relative to the other evaluated regions, its hydrodynamic environment is the most quiescent, serving as an ideal baseline for assessing the algorithm's foundational adaptability.

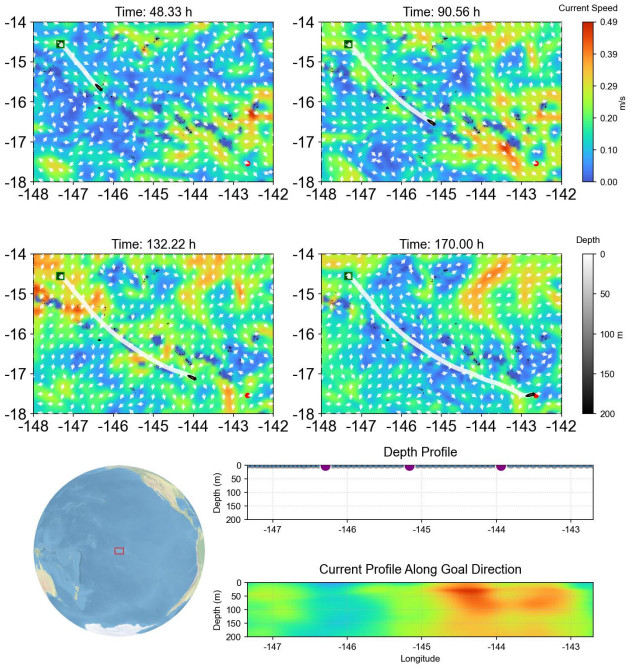


Figure 6: Visualization of AUV trajectory planning in the Tuamotu Archipelago (stable open-sea region and atoll formations). Snapshots illustrate the instantaneous spatial trajectory and hydrodynamic field.

In this evaluation, the AUV's origin was positioned at a northern atoll in the Tuamotu Archipelago (14.55°S, 147.34°W), with the destination at a southern atoll (17.50°S, 142.65°W), primarily traversing open-sea regions and scattered atoll formations. The temporal snapshots in Fig. 6 depict the complete transit from the origin through open-sea areas while circumventing atoll formations, as well as the instantaneous flow field distribution at respective timesteps.

The navigation results demonstrate that in a stable, low-velocity flow regime, the agent requires no intricate vertical maneuvering or extensive detours; efficient transit is achieved through straightforward heading adjustments. When traversing atoll formations, the agent accurately perceives micro-scale circumflows and wake vortices, modulating its heading to circumvent localized adverse flows. It consistently maintains optimal navigation efficiency, ultimately reaching the destination with a streamlined trajectory devoid of redundant maneuvers or depth oscillations. This result indicates that the algorithm possesses excellent foundational adaptability in open-sea regions with stable flow fields, The algorithm is capable of recognizing uncomplicated bathymetry and micro-scale flow variability, thereby

establishing a baseline for evaluation in more complex waters.

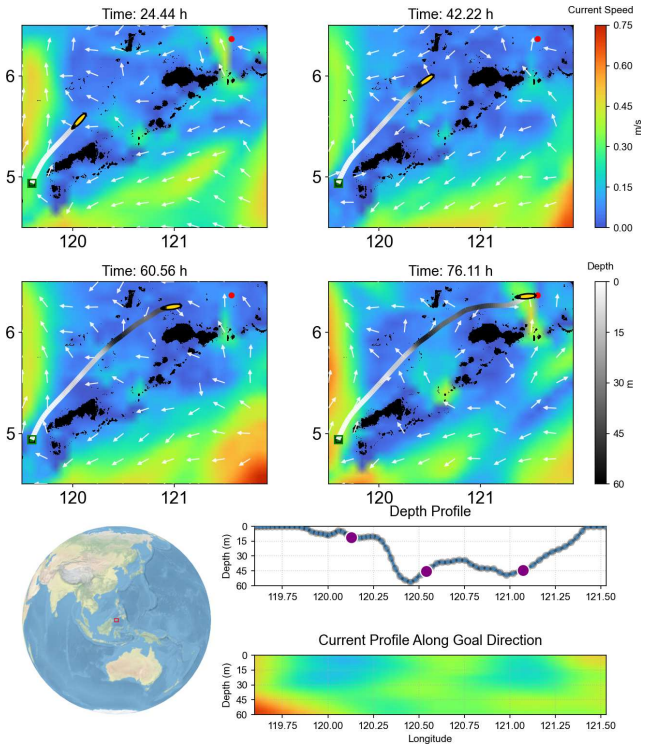


Figure 7: Same as Fig. 6, but for the Sulu Sea (dense island reef topography).

4.4.2. *Sulu Sea: Dense Reef Bathymetry in Low-Latitude*

The Sulu Sea, located in Southeast Asia and bounded by Indonesia, Malaysia, and the Philippines, constitutes a representative region characterized by dense island reefs and complex bathymetry, representing moderate geometric complexity within the evaluation framework. This tropical semi-enclosed basin features intricate terrain with dense distributions of islands and submerged reefs, as well as sharply undulating seabed profiles. The complex geometric boundaries significantly constrain the overall circulation (Metzger and Hurlburt, 1996). Influenced by tidal dynamics, the flow field exhibits a "locally complex, globally quiescent" characteristic—flow velocities are low in open waters, yet localized intense flows and complex circumflows develop within reefs and straits due to bathymetric constraints (Gordon et al., 2011). The thermohaline structure is influenced by tidal forces, exhibiting vertical stratification. Particularly in straits and shoals, tide-driven surges intensify the vertical mixing of localized water masses (Han et al., 2009). Relative to the Tuamotu Archipelago, the primary challenge involves obstacle avoidance among dense reefs and adaptation to localized complex flows; compared to the Arctic Ocean, this region exhibits more frequent high-intensity flow variations and significantly more intricate geometric constraints.

In alignment with the distribution of active tidal zones in the Sulu Sea, the AUV origin was positioned near the Sibutu Strait (4.94°N, 119.60°E), with the destination near Pangutaran Island (6.40°N, 121.56°E). The trajectory traverses dense island reef clusters, requiring robust obstacle avoidance and adaptation to localized flow variability induced by tides. The navigation results (Fig. 7) demonstrate that the agent exhibits robust obstacle avoidance and localized flow adaptation. Facing densely distributed reefs (with inter-reef spacing in specific zones below 1 km), the agent autonomously generates optimal detours even when direct trajectories are obstructed, ensuring zero collision risk against islands and submerged reefs. Regarding localized flow fluctuations and tide-induced circumflows, the agent perceives instantaneous flow variability, modulating its heading and depth to exploit favorable localized flows. This proactive adjustment enhances efficiency and prevents trajectory deviation caused by adverse circumflows. As the evaluation incorporates tidal dynamics of the Sulu Sea (specifically during spring tides included in the evaluation), these results validate the algorithm's adaptability within environments characterized by coupled complex bathymetry and dynamic flows. This aligns with existing research regarding navigational constraints in the Sulu Sea, substantiating the framework's capability to address core regional hydrodynamic challenges.

4.4.3. *Arctic Ocean: Complex Bathymetry Region in High-Latitude*

The Arctic Ocean represents the most complex and challenging environment within the evaluation framework. Amidst intensifying global warming, continuous sea-ice melting, and the gradual opening of polar shipping routes, marine navigation, observation, and resource exploration in this region are becoming increasingly frequent. Consequently, the demand for AUV deployment in high-latitude regions is rising, which is of substantial practical significance. This extreme high-latitude environment is characterized by a highly variable hydrodynamic field constrained by complex bathymetry, featuring intricate circulation systems such as the Beaufort Gyre, where localized intense flows interact with narrow topographic constraints (Li et al., 2024). The complex thermohaline structure—driven by sub-zero temperatures (year-round below 0°C) and freshwater convergence from melting ice—results in pronounced seawater density stratification. Furthermore, the coupling of ice-ocean stress and sea-ice melting processes enhancing the spatio-temporal fluctuations of the flow field (Leng et al., 2022). Driven by climate warming, eddy kinetic energy and flow variability are projected to increase further (Watanabe et al., 2014), exceeding those in other evaluated

regions and imposing stringent requirements on AUV trajectory planning and environmental adaptability.

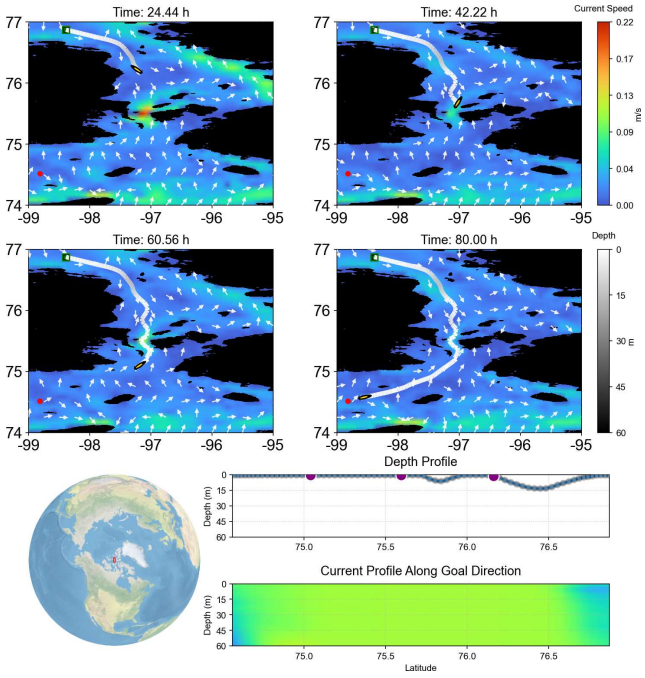


Figure 8: Same as Fig. 6, but for the Arctic Ocean (extremely variable polar flows and complex constraints).

In the context of opening Arctic routes, the dense inter-ice lead region of the Canadian Basin is selected as the study area. The AUV origin is positioned in the waters north of Ellesmere Island (76.87°N, 98.39°W), with the destination located off the northwest coast of the island (74.55°N, 98.82°W). The trajectory primarily traverses inter-island leads and nearshore waters of the Canadian Basin along an approximate meridional orientation. The background hydrodynamic field utilizes conditions from the peak period of summer sea-ice melt. During this period, modulated by sea-ice melting and polar circulation, the flow field exhibits pronounced spatio-temporal fluctuations. Flow velocities in localized intense regions can exceed 0.5 m/s, accompanied by significant circumflows and vortex structures. Enhanced heat flux from the ocean to the sea ice reduces flow field stability, imposing substantial disturbances on autonomous navigation and trajectory tracking.

The experimental results (Fig. 8) demonstrate that the agent exhibits robust environmental adaptability and trajectory planning capabilities under extreme Arctic conditions of the Arctic Ocean. As illustrated in Fig. 8, the algorithm generates a tortuous yet continuous and stable trajectory. The AUV maneuvers through localized intense flows and narrow bathymetric constraints, autonomously identifying collision-free trajectories with minimal hydrodynamic resistance through modulated adjustments of heading and depth. Horizontally, the AUV avoids obstructions from narrow bathymetry while exploiting favorable localized flows from the polar circulation to enhance efficiency; vertically, it adjusts its depth based on the vertical stratification of the hydrodynamic field, maintaining stable transit without oscillation. These results correspond with the dynamic flow characteristics of the Beaufort Gyre region documented in existing literature, validating the algorithm's efficacy in variable polar waters and providing empirical support for future AUV deployments in the Arctic Ocean.

4.4.4. *Comprehensive Comparison Across Different Ocean Regions*

Synthesizing the experimental results across the four representative marine regions—the Tuamotu Archipelago, the Sulu Sea, the Taiwan Strait, and the Arctic Ocean—validates the robust adaptability and generalization capability of the proposed RL-based trajectory planning framework. The evaluation in the Tuamotu Archipelago confirms foundational adaptability to stable flow regimes and uncomplicated bathymetry, while the Sulu Sea results validate robust obstacle avoidance and localized flow adaptation within dense reef clusters and tide-induced circumflows. Furthermore, the experiments in the Taiwan Strait demonstrate proficiency in navigating monsoon-driven flows and vertical shear by optimizing energy expenditure through autonomous vertical maneuvering. Finally, the Arctic Ocean results substantiate stable navigation capabilities amidst highly variable polar flows and severe environmental disturbances. Furthermore, the framework effectively addresses the multi-faceted challenges inherent to the polar environment. This cross-regional performance comprehensively demonstrates the algorithm's capacity to handle diverse hydrodynamic regimes and bathymetric constraints globally.

From a regional oceanographic perspective, the four evaluated regions encompass four primary oceanic categories: open-sea, island-reef zones, straits, and polar regions. These regions span three core flow regimes (stable flow, monsoon-driven flow, and highly variable flow) and three distinct thermohaline profiles (uniform structure, seasonal stratification, and polar low-temperature stratification), effectively representing the hydrodynamic and bathymetric characteristics of major global marine environments. The framework demonstrates high adaptability and feasibility across all regions, validating its capacity to couple 3D spatial perception with temporal prediction for precise capture of flow-field dynamics. Furthermore, the algorithm exhibits significant generalization capability, adapting to regional hydrodynamic and bathymetric variations without requiring site-specific parameter tuning.

From the perspective of practical utility, the evaluation scenarios correspond to actual AUV deployment requirements: the Tuamotu Archipelago represents large-scale marine observation, the Sulu Sea pertains to coastal reef exploration, the Taiwan Strait aligns with nearshore monitoring, and the Arctic Ocean serves polar scientific expeditions. The robust performance across these regions provides a technical foundation for the widespread deployment of AUVs in diverse global waters. Especially given the expansion of Arctic shipping routes and the increasing demand for global marine observation, the framework's generalization capability and adaptability to extreme environments reduce deployment costs while enhancing navigation efficiency and safety across heterogeneous marine regions.

5. Conclusion and Discussion

This paper presents a DRL framework to address energy-efficient AUV trajectory planning within complex dynamic marine environments. The proposed approach integrates high-resolution marine reanalysis data (CMEMS GLORYS12V1) into the training environment to narrow the discrepancy between simulation and reality. In contrast to traditional methods that struggle to process complex environmental variables, the DRL algorithm leverages this massive realistic dataset to empirically learn navigation strategies, ultimately deriving policies that are demonstrably superior and more adaptable to dynamic flows.

The proposed framework focuses on global trajectory planning and effectively addresses the trade-off between long-range energy optimization and adaptability to complex marine environments. Extensive simulations conducted within the Taiwan Strait region yield the following primary conclusions:

(1) **Energy Efficiency via Autonomous Vertical Maneuvering:** The framework captures and exploits the three-dimensional structures of marine flows. By leveraging vertical shear and executing autonomous vertical maneuvering to access favorable flows, the proposed framework significantly reduces mission duration and total energy expenditure compared to the geometrically optimal A* baseline, maintaining a high mission success rate and demonstrating superior energy efficiency.

2) **Robustness in Spatio-temporal Flow Fields:** By incorporating an auxiliary temporal prediction module (AUX), the agent circumvents passive drift artifacts inherent in dynamic environments. The results demonstrate that the method adapts to the temporal evolution of the flow field, maintaining trajectory stability and efficiency even when static flow assumptions are invalid.

Building upon these spatio-temporal adaptive capabilities, this study facilitates a fundamental paradigm shift in environmental interaction. While traditional geometric planners and many existing RL approaches predominantly treat dynamic currents as external disturbances to be resisted or reactively compensated for, the proposed framework redefines the complex flow field as an exploitable navigational asset. This transition from reactive pathfinding to environment-synergistic navigation provides a conceptual blueprint for next-generation autonomous systems to achieve highly efficient, symbiotic operations in deep-sea environments.

Despite these contributions, the current framework exhibits certain limitations that warrant further investigation. Regarding kinematic modeling, the AUV is simplified as a mass point, and the interaction between the vehicle and the marine environment is modeled via the linear vector superposition of propulsion and flow velocities. This idealized approach neglects the inherent 6-DOF nonlinear dynamics and complex fluid-structure interactions, which are critical for precise maneuvering in highly dynamic ocean conditions.

Furthermore, the existing energy expenditure and hydrodynamic resistance models rely primarily on theoretical derivations involving square and cubic velocity terms. Such simplifications overlook several essential engineering constraints: the nonlinear hydrodynamic interference, propulsion system dead-zone and saturation characteristics, and inherent actuator response latencies. Additionally, the model does not fully account for variations in propulsion efficiency or the nonlinear fluctuations of resistance encountered in real-world deployments. Consequently, these omissions may result in a non-negligible discrepancy between the simulated trajectory optimality and the empirical performance observed in actual sea trials.

Future research will prioritize the following objectives. To improve alignment with engineering practice, high-fidelity dynamic models will be incorporated into the simulation framework. Furthermore, instantaneous obstacle avoidance evaluations utilizing data from physical sensors, such as forward-looking sonar will be conducted to assess the avoidance and safety assurance capabilities of the framework when encountering unknown localized obstructions.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (42376007), the MEL-RLAB Joint Fund for Marine Science & Technology Innovation (M&R202301), and Fujian Ocean Innovation Center.

This study was conducted using the E.U. Copernicus Marine Service Information; the global ocean physics reanalysis data were obtained from the Copernicus Marine Service (CMEMS), (product ID: GLOBAL

_MULTIYEAR_PHY_001_030, DOI: https://doi.org/10.48670/moi-00021). The bathymetric data were derived from the GEBCO 2025 Grid (GEBCO Compilation Group 2025, DOI: https://doi.org/10.5285/37c52e96-24ea-67ce-e063-7086abc05f29), the definitive product of the General Bathymetric Chart of the Oceans (GEBCO) project.

The source codes for this study are publicly available when the paper is accepted.

References

Carlucho, I., De Paula, M., Wang, S., Menna, B.V., Petillot, Y.R., Acosta, G.G., 2018. Auv position tracking control using end-to-end deep reinforcement learning, in: OCEANS 2018 MTS/IEEE Charleston, IEEE. pp. 1–8.

Dumas, F., Le Gendre, R., Thomas, Y., Andréfouët, S., 2012. Tidal flushing and wind driven circulation of ahe atoll lagoon (tuamotu archipelago, french polynesia) from in situ observations and numerical modelling. Marine Pollution Bulletin 65, 425–440. Ahe Atoll and Pearl Oyster Aquaculture in the Tuamotu Archipelago.

Fan, S., Yuan, P., Feng, J., Huang, H., Hu, S., Zhao, Z., Zhu, Y., Wang, K., 2025. Observing typhoon-driven upper ocean dynamics in the south china sea using a virtual mooring underwater glider array: methods and analysis. Journal of Ocean Engineering and Science .

Gordon, A.L., Tessler, Z.D., Villanoy, C., 2011. Dual overflows into the deep sulu sea. Geophysical Research Letters 38.

Hadi, B., Khosravi, A., Sarhadi, P., 2022. Deep reinforcement learning for adaptive path planning and control of an autonomous underwater vehicle. Applied Ocean Research 129, 103326.

Han, W., Moore, A.M., Levin, J., Zhang, B., Arango, H.G., Curchitser, E., Di Lorenzo, E., Gordon, A.L., Lin, J., 2009. Seasonal surface ocean circulation and dynamics in the philippine archipelago region during 2004–2008. Dynamics of Atmospheres and Oceans 47, 114–137.

Hart, P.E., Nilsson, N.J., Raphael, B., 1968. A formal basis for the heuristic determination of minimum cost paths. IEEE transactions on Systems Science and Cybernetics 4, 100–107.

Havenstrøm, S.T., Rasheed, A., San, O., 2021. Deep reinforcement learning controller for 3d path following and collision avoidance by autonomous underwater vehicles. Frontiers in Robotics and AI 7, 566037.

Hobson, B.W., Bellingham, J.G., Kieft, B., McEwen, R., Godin, M., Zhang, Y., 2012. Tethys-class long range auvs-extending the endurance of propeller-driven cruising auvs from days to weeks, in: 2012 IEEE/OES autonomous underwater vehicles (AUV), IEEE. pp. 1–8.

Hu, J., Kawamura, H., Li, C., Hong, H., Jiang, Y., 2010. Review on current and seawater volume transport through the taiwan strait. JOURNAL OF OCEANOGRAPHY 66, 591–610.

Hu, J., Wang, X.H., 2016. Progress on upwelling studies in the china seas. REVIEWS OF GEOPHYSICS 54, 653–673.

Huang, H., Song, K., Chen, Y., Jin, H., Guan, Y., 2025. 3d path planning for auvs under ocean currents by prioritized experience replay mechanism. Neurocomputing 630, 129719.

Jiang, D., Fang, Z., Cheng, C., He, B., Li, G., 2022. Action guidance-based deep interactive reinforcement learning for auv path planning, in: 2022 International Conference on Machine Learning, Control, and Robotics (MLCR), IEEE. pp. 158–165.

Khatib, O., 1986. Real-time obstacle avoidance for manipulators and mobile robots. The international journal of robotics research 5, 90–98.

Kim, J., Park, H.J., Penumarti, A., Shin, J., 2024. Fast marching based rendezvous path planning for a team of heterogeneous vehicles. IEEE Access .

LaValle, S., 1998. Rapidly-exploring random trees: A new tool for path planning. Research Report 9811 .

Leng, H., Spall, M.A., Bai, X., 2022. Temporal evolution of a geostrophic current under sea ice: Analytical and numerical solutions. Journal of Physical Oceanography 52, 1191 – 1204.

Li, X., Wang, Q., Danilov, S., Koldunov, N., Liu, C., Mueller, V., Sidorenko, D., Jung, T., 2024. Eddy activity in the arctic ocean projected to surge in a warming world. NATURE CLIMATE CHANGE 14.

Li, X., Yu, S., 2023. Three-dimensional path planning for auvs in ocean currents environment based on an improved compression factor particle swarm optimization algorithm. Ocean Engineering 280, 114610.

Li, Y., He, X., Lu, Z., Jing, P., Su, Y., 2023. Comprehensive ocean information-enabled auv motion planning based on reinforcement learning. Remote Sensing 15, 3077.

Li, Z., Li, Z., He, J., Xi, M., Wen, J., Yang, J., Wang, J., 2025. Auv path planning in complex ocean environment via deep reinforcement learning assisted by edge computing. IEEE Transactions on Consumer Electronics .

Lim, H.S., Fan, S., Chin, C.K., Chai, S., 2018. Performance evaluation of particle swarm intelligence based optimization techniques in a novel auv path planner, in: 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV), IEEE. pp. 1–7.

Liu, C., Hao, Y.l., Gao, F., 2011. A path planning method for underwater vehicle based on ocean current information, in: 2011 Fourth International Joint Conference on Computational Sciences and Optimization, IEEE. pp. 987–991.

Liu, J., Xu, Y., Song, S., Jiang, L., 2025a. Reducing auv energy consumption through dynamic sensor directions switching via deep reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 18843–18851.

Liu, S., Wang, Z., Wang, T., Zhou, S., Zhang, Y., Jin, P., Yang, G., 2025b. Energy-optimized path planning for fully actuated auvs in complex 3d environments. Journal of Marine Science and Engineering 13, 1269.

Marchel, Ł., Kot, R., Szymak, P., Piskur, P., 2025. Model-based auv path planning using curriculum learning and deep reinforcement learning on a simplified electronic navigation chart. Applied Sciences 15, 6081.

Metzger, E.J., Hurlburt, H.E., 1996. Coupled dynamics of the south china sea, the sulu sea, and the pacific ocean. Journal of Geophysical Research: Oceans 101, 12331–12352.

Nie, Y., Yang, H., Gao, Q., Qu, T., Fan, C., Song, D., 2020. Research on path planning algorithm based on dimensionality reduction method and improved rrt, in: Global Oceans 2020: Singapore–US Gulf Coast, IEEE. pp. 1–4.

Pan, Z., Zhang, C., Xia, Y., Xiong, H., Shao, X., 2021. An improved artificial potential field method for path planning and formation control of the multi-uav systems. IEEE Transactions on Circuits and Systems II: Express Briefs 69, 1129–1133.

Patzert, W.C., Bernstein, R.L., 1976. Eddy structure in the central south pacific ocean. Journal of Physical Oceanography 6, 392 – 394.

Sethian, J.A., 1996. A fast marching level set method for monotonically advancing fronts. proceedings of the National Academy of Sciences 93, 1591–1595.

Smith, R.N., Pereira, A., Chao, Y., Li, P.P., Caron, D.A., Jones, B.H., Sukhatme, G.S., 2010. Autonomous underwater vehicle trajectory design coupled with predictive ocean models: A case study, in: 2010 IEEE International Conference on Robotics and Automation, IEEE. pp. 4770–4777.

Watanabe, E., Onodera, J., Harada, N., Honda, M.C., Kimoto, K., Kikuchi, T., Nishino, S., Matsuno, K., Yamaguchi, A., Ishida, A., Kishi, M.J., 2014. Enhanced role of eddies in the arctic marine biological pump. NATURE COMMUNICATIONS 5.

Wen, J., Li, Z., Xi, M., He, J., 2025. A llm-assisted auv 3d path planning scheme under ocean current interference via reinforcement learning. IEEE Internet of Things Journal .

Wen, Z., Wang, Z., Wen, X., Niu, C., Wang, P., Zhou, D., 2024. A hierarchical planning method for auv search tasks based on the snake optimization algorithm. Sensors 24, 7196.

Wu, J., Liu, C., Filaretov, V., Yukhimets, D., Cai, R., Zheng, A., Zuev, A., 2025. Review of research on cooperative path planning algorithms for auv clusters. Drones 9, 790.

Xin, J., Zhong, J., Yang, F., Cui, Y., Sheng, J., 2019. An improved genetic algorithm for path-planning of unmanned surface vehicle. Sensors 19, 2640.

Yu, D., Wang, H., Cao, X., Wang, Z., Ren, J., Zhang, K., 2024. Enhancing autonomous underwater vehicle decision making through intelligent task planning and behavior tree optimization. Journal of Marine Science and Engineering 12, 791.

Zhang, B., Su, Y., Sun, S., Luo, W., Huang, Q., 2025a. Hybrid path planning algorithm for underactuated auv based on rrt star and apf. Scientific Reports 15, 28493.

Zhang, J., Zhou, W., Deng, X., Yang, S., Yang, C., Yin, H., 2025b. Optimization of adaptive observation strategies for multi-auvs in complex marine environments using deep reinforcement learning. Journal of Marine Science and Engineering 13, 865.

Zhao, E., Cheng, T., Wang, S., Lin, X., 2025a. Underwater 3d path planning for auv in ocean currents based on improved informed rrt* algorithm. Journal of Intelligent & Robotic Systems 111, 116.

Zhao, J., Ma, X., Yang, B., Chen, Y., Zhou, Z., Xiao, P., 2022. Global path planning of unmanned vehicle based on fusion of a$*$ algorithm and voronoi field. Journal of Intelligent and Connected Vehicles 5, 250–259.

Zhao, L., Bai, Y., Paik, J.K., 2025b. Global path planning and waypoint following for heterogeneous unmanned surface vehicles assisting inland water monitoring. Journal of Ocean Engineering and Science 10, 88–108.

Zhu, G., Shen, Z., Liu, L., Zhao, S., Ji, F., Ju, Z., Sun, J., 2022. Auv dynamic obstacle avoidance method based on improved ppo algorithm. IEEE Access 10, 121340–121351.