Xin, X, Liu, K, Yu, Y and Yang, Z

Developing robust traffic navigation scenarios for autonomous ship testing: an integrated approach to scenario extraction, characterization, and sampling in complex waters

Transportation Research Part C 178 (2025) 105246

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Developing robust traffic navigation scenarios for autonomous ship testing: an integrated approach to scenario extraction, characterization, and sampling in complex waters

Xuri Xin [a], Kezhong Liu [b], Yuerong Yu [b], Zaili Yang [a,*]

[a] Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, UK
[b] School of Navigation, Wuhan University of Technology, Wuhan, Hubei, China

ARTICLE INFO

Keywords:
Maritime transportation
MASS performance validation
Scenario-based testing
Scenario extraction and characterization
Optimization sampling
AIS data mining

ABSTRACT

The fast development of Maritime Autonomous Surface Ships (MASSs) marks a significant advancement in the shipping industry, offering enhanced performance, cost-effectiveness, and environmental sustainability. This study aims to develop a holistic methodology for extracting, characterizing, and sampling maritime traffic navigation scenarios to support the testing and validation of MASSs, ensuring their reliability before widespread deployment. It begins with a precise scenario extraction technique that can effectively capture dynamic interactions among ships over time, enhanced by an in-depth analysis of ship motion dynamics, geographical complexities, and spatial-temporal nested interdependencies. Subsequently, the extracted scenarios are characterized using newly created metrics and advanced models in a structured and integrated manner. This allows for the classification and parameterization of ship motion patterns, conflict complexities, and encounter types, thereby enhancing the interpretability of traffic co-behaviors. Finally, a hierarchical greedy sampling strategy is developed to adaptively select representative scenarios from a sufficiently realistic set, striking a balance between comprehensive scenario coverage and efficiency in MASS testing. Extensive experimental analyses validate the efficacy of the proposed methodology. It precisely identifies real evolutionary multi-ship encounter situations, finely characterizes scenarios to support the encoding, explaining, and understanding of dynamic traffic behaviors, and systematically selects representative scenarios by incorporating multiple selection principles. Consequently, this methodology makes new contributions to the pioneering development of an accurate, interpretable, and representative set of real-world traffic navigation scenarios for autonomous testing. This is crucial for assessing and validating the advancements in MASSs and their emerging functionalities, thereby promoting highly and fully automated navigation.

1. Introduction

Maritime Autonomous Surface Ships (MASSs) are increasingly viewed as the future of the shipping industry, offering superior performance, cost-effectiveness, and environmental sustainability over traditional manned ships. These ships boast substantial benefits like optimized space utilization, reduced design and operational costs, decreased fuel consumption, and minimized human error from

* Corresponding author.
E-mail address: Z.Yang@ljmu.ac.uk (Z. Yang).

https://doi.org/10.1016/j.trc.2025.105246
Received 3 February 2025; Received in revised form 6 May 2025; Accepted 18 June 2025
Available online 23 June 2025

fatigue or harsh working conditions (Bakdi et al., 2021; Gu and Wallace, 2021). Alongside these benefits, advancements in technologies such as Artificial Intelligence (AI), cloud computation, the Internet of Things (IoT), and Blockchain, are profoundly enhancing the integration and sophistication of maritime transportation systems (Cao et al., 2025; Xin et al., 2025; Yu et al., 2021). These technological enhancements in perception, communication, and automation are making the shift toward more advanced electromechanical and digitally controlled systems not only viable but also increasingly feasible, steering the industry away from traditional mechanical operations toward autonomous or remotely controlled functions. However, the implementation of these autonomous systems introduces new challenges, particularly in areas such as technology readiness and reliability, human interaction, regulations and standards, environmental adaptability, and safety and security (Cassará et al., 2023; Chang et al., 2024; Gil et al., 2025; Lu et al., 2025; Thombre et al., 2020). Increasingly complex traffic situations and intricate cyber-physical-social interactions also lead to new vulnerabilities. Consequently, there is a widespread agreement among manufacturers, experts, and maritime authorities that MASSs must undergo extensive testing and be proven as safe as manned ships before widespread adoption (Bakdi et al., 2021; Zhu et al., 2022). Rigorous and comprehensive testing is essential to establish trust and confirm the safety, security, and reliability of MASSs, ensuring that these future autonomous ships are fully certified and ready for broad deployment.

The autonomous navigation system is a critical enabling technology for MASSs, responsible for perception, alerting, decision-making, and control to ensure safe navigation (Akdağ et al., 2022; Wang et al., 2024c; Wróbel et al., 2022). Its effectiveness must be validated and verified not just through basic situational awareness and geographical risk assessments but also through dynamic decision-making in various encounter scenarios (Gil et al., 2024b, 2024a; Niu et al., 2023; Zhang et al., 2024). This capability is crucial for avoiding collisions with dynamic entities, such as other MASSs equipped with different levels of computer-assisted functionalities or manned ships. In this context, notable industrial initiatives have been undertaken to launch commercial MASS projects worldwide, including the "YARA BIRKELAND" project in Norway, "Samsung Heavy Industries T-8 Tugboat" in South Korea, the "MAXCMAS" project in the United Kingdom, and the "DFFAS" project in Japan, which involve rigorous physical tests on autonomous navigation functions. It is crucial to acknowledge that the reliability of test outcomes depends on the scope and variety of the scenarios tested. Current physical tests, however, typically involve a limited range of scenarios due to the high costs and risks associated with full-scale real ship testing. In the realm of road autonomy, an effective strategy begins with preliminary simulation tests across a broad range of scenarios, which allows for numerous, quicker, more affordable, and safer trials. This is followed by focused tests for critical scenarios in real-world trials (Feng et al., 2020; Ma et al., 2024). Regardless of the testing strategies employed, to maximize the efficacy of these tests, the scenarios must thoroughly encompass potential real-world traffic navigation conditions (Bolbot et al., 2022; Wang et al., 2024b). Essentially, the strategic formulation of a testing scenario library that closely resembles real-world navigation scenarios and adequately covers potential situations is key to building trust in the test results, thereby helping to ascertain whether a MASS can operate safely and efficiently.

Over the past two decades, the advancement of the Automatic Identification System (AIS) has significantly enhanced the availability of detailed ship trajectory data, facilitating the creation of diverse maritime traffic navigation scenarios (Bakdi et al., 2021; Liu et al., 2024a, 2022a; Liu et al., 2024b; Rong et al., 2024; Xin et al., 2024; Yu et al., 2025; Zhang et al., 2023). Compared to other scenario generation methods, such as rule-based and random-based approaches, AIS-based scenario generation offer several distinct advantages: 1) The scenarios closely replicate real-world navigation conditions by accurately depicting ships' maneuverability, motion patterns, and the spatial-temporal relationships among ships; 2) extensive AIS data ensures thorough coverage of various potential maritime traffic situations associated with realistic risks and complexities; and 3) the scenarios offer in-depth insights into the entire ship encounter process, not just the initial moment (Wang et al., 2024b). Moreover, previous research has shown that AIS data-based scenarios are effective for exploring collision risk assessments (Lotovskyi et al., 2024; Zhang et al., 2021), identifying collision avoidance behaviors (Liu et al., 2021), supporting intent prediction (Jia et al., 2024), and examining the distribution patterns of scenario parameters (Zhu et al., 2022). This robust body of evidence underscores the utility of AIS-generated scenarios in contributing realistic test scenarios that are crucial for comprehensive testing and verification of MASSs.

Nevertheless, research on AIS data-based scenario formulation is still in its infancy. The growing volume of maritime traffic, diversified ship movements, complex geographical landscapes, and intricate spatial–temporal dependencies among ships pose significant challenges to fully realizing the above advantages and achieving the effectiveness and efficiency of current AIS-based scenario production approaches for autonomous testing (Wang et al., 2024b; Yu et al., 2021). These challenges are particularly pressing given the mix of autonomous and manned ships expected in future traffic. Specifically, the state-of-the-art research exhibits several critical shortcomings that remain largely unaddressed: 1) They fail to assure the accuracy of the traffic navigation scenarios they produced because they do not comprehensively account for the impacts of dynamic ship motion and maritime geographical features on the real multiple interdependent interactions among ships over time; 2) they are inadequate at fully classifying and parameterizing scenario characteristics such as ship motion patterns, conflict risks and complexities, and encounter types in a structured and integrated manner, hindering the process of encoding, explaining, and understanding dynamic traffic co-behaviors; and 3) they are incapable of sufficiently exploring the diversity of scenarios from various perspectives, nor do they possess effective techniques to select representative scenarios from a sufficiently realistic set that allows both high coverage and efficient testing. Undoubtedly, addressing these limitations could lead to the creation of a more reliable and comprehensive set of traffic navigation scenarios, critical for conducting trustworthy and extensive autonomy tests.

Motivated by the challenges outlined above, this study aims to propose a holistic methodology for extracting, characterizing, and sampling traffic navigation scenarios from extensive historical ship trajectory data. Firstly, it precisely extracts traffic scenarios that satisfy various constraints in complex traffic environments. These scenarios are then thoroughly characterized to encode traffic behaviors, enhancing interpretability. Finally, the methodology adaptively selects a subset of scenarios, guided by multiple selection principles, that effectively cover the characteristics of the original scenario set to facilitate efficient autonomous testing. Consequently,

this approach recommends an accurate, interpretable, and representative set of traffic navigation scenarios for simulation-based testing or sea trials. Such trials are vital for assessing and validating the advancements in MASSs and their emerging functionalities, thereby promoting highly and fully automated navigation.

The new contributions of this paper are summarized as follows.

1) A novel scenario extraction approach is developed to identify scenarios that capture the continuous spatial-temporal interaction process among ships. This approach precisely reflects the real interactions among multiple ships over time, taking into account the effects of potential ship motion dynamics and constrained waterway geography. Additionally, it allows for the observation of dynamic co-behaviors among ships over time, including variations in {the number of ships, conflict frequency and severity, types of encounters}, and other elements involved in each scenario.
2) Three key scenario characteristics, including ship motion patterns, conflict complexities, and encounter types, are revealed in a structured and holistic manner. The characterization process designs new metrics and employs a range of advanced models to facilitate the classification and parameterization of these scenarios, thereby enhancing the encoding and understanding of their behaviors.
3) A hierarchical greedy sampling technique is proposed to adaptively select representative scenarios from a sufficiently realistic set by examining the diversity of scenarios across three characteristic types. This procedure ensures adequate coverage of potential scenarios while minimizing the number of scenarios required for autonomous testing, thereby achieving a balance between test coverage and efficiency.

The remainder of this paper is organized as follows. Section 2 provides an overview of the current research related to scenario generation methods. Section 3 presents the developed methodology for extracting, characterizing, and sampling traffic scenarios in detail. Section 4 examines the results of the application analysis and explores their implications. Lastly, Section 5 concludes the paper and proposes future research directions.

2. Literature review

Accurate and effective autonomous testing represents a cornerstone in the development of MASSs and their emerging functionalities. The importance of such testing is reinforced by a wealth of literature dedicated to assessing and validating autonomous systems, which provides crucial support for ensuring their reliability and trustworthiness. Research in this area concentrates on several critical testing aspects: safety, authenticity, cost-effectiveness, and coverage (Wang et al., 2023). Commonly employed scenario generation methods for testing include artificial-based, random-based, and AIS data-based approaches. Subsequent sections will provide a critical analysis of the benefits and drawbacks of these methods, along with an exploration of the existing research gaps.

2.1. Artificial-based scenario generation methods

Artificial-based methods generate scenario sets that are either rule-based or driven by expert input, taking into account various factors such as the number of encountered ships, types of encounters, International Regulations for Preventing Collisions at Sea (COLREGs), and environmental parameters. There is an extensive body of work related to autonomous system verification based on these artificially generated scenarios. For instance, Du et al. (2020) developed specific scenarios to assess the performance of their anti-collision algorithms for stand-on ships when the give-way ship does not promptly comply with COLREGs. Ahmed et al. (2021) employed the Imuzu-proposed 22 cases to test their autonomous collision avoidance algorithms. Liu et al. (2022) created mixed multi-ship encounter scenarios involving both autonomous and manned ships to analyze the effectiveness of the collision avoidance module in terms of rule compliance, collision risk, and drift distance. Sawada et al. (2024) developed diverse one-on-one and one-on-two encounter scenarios that take into account risks, COLREGs, and the characteristics of the encounter angle of ships.

Despite these efforts, artificial-based methods exhibit two significant limitations that affect their real-world application effectiveness. First, due to their reliance on manual efforts, the predefined representative scenarios are restrictive, typically including no more than 100 scenarios, which limits their ability to cover potential traffic situations extensively (Zhu et al., 2022). Second, the manually designed scenarios are too simplistic to adequately describe the entire traffic navigation process, making it difficult to mimic the true ship behaviors throughout the encounter process (Wang et al., 2024b). As a result, while autonomous functions may perform well in these artificial-based navigation scenarios, they may falter under broader or more unpredictable conditions.

2.2. Random-based scenario generation methods

Compared to artificial-based methods, random-based scenario generation addresses the limitation of covering all possible situations, enabling the thorough verification of autonomous functions across a variety of complex environments and extreme conditions. This approach can generate a vast array of scenarios spanning diverse geographical areas (e.g., open waters, coastal regions, and archipelago areas), meteorological conditions (e.g., wind speeds, current speeds, and wave conditions), and navigational settings (e.g., ship positions, speeds, headings, and types). For instance, Johansen et al. (2016) designed a comprehensive set of test scenarios from single to multi-obstacle avoidance, and from predictable to complex random obstacle distributions, to evaluate the efficacy of their anti-collision strategy. Pedersen et al. (2020) designed an automatic test scenario generation module using digital twin technology to simulate environmental conditions and maritime traffic interactions. Bolbot et al. (2022) introduced a systematic procedure for

generating traffic scenarios for collision avoidance system testing. This procedure involves generating potential traffic parameters, identifying hazardous scenarios, and extracting representative scenarios. Torben et al. (2023) proposed a scenario set generation methodology that utilizes Gaussian Process (GP) and Signal Temporal Logic (STL) models to enhance the test coverage for the formal verification of autonomous ship control systems. Zhou and Zhang (2023) developed a testing platform to test the collision avoidance capabilities of autonomous ships in diverse settings such as open waters, coastal regions, and archipelagos with varying traffic densities, equipped with a synthetic map generator and a random encounter scenario generator.

While the above studies demonstrate that random-based methods can efficiently generate a wide range of scenarios for comprehensive testing, the virtually infinite scenario space includes a substantial number of invalid cases that fail to adequately capture critical risk and complexity aspects (Bakdi et al., 2021). Several recent studies have addressed this limitation. For instance, Shi et al. (2023) developed probabilistic ship behavior models from real traffic data to construct a naturalistic sailing environment, into which they sparsely injected adversarial maneuvers to create a Naturalistic-and-Adversarial Sailing Environment (NASE), thereby maintaining realism while significantly increasing the yield of critical encounters. Similarly, Zhu et al. (2025) formulated scenario generation as a heterogeneous multi-agent Markov Decision Process (MDP), integrating a six-factor Fuzzy Analytic Hierarchy Process (FAHP) collision risk index with an enhanced Obstacle Zone by Target (OZT)/Kalman-filtered Independence Dueling Double Deep Q-Network (Independence D3QN) and a "targeted-pressure" reward mechanism to adaptively steer background ships toward high-risk interactions. Nevertheless, these approaches face challenges in selecting representative traffic scenarios that accurately and strongly correlate with real-world conditions, primarily because their generated scenarios are not fully derived from actual traffic data, potentially leading to test outcomes that do not reliably reflect the performance of autonomous systems under real-world conditions.

2.3. AIS data-based scenario generation methods

The digitalization revolution has greatly enhanced the accessibility and quality of AIS-based trajectory data, which is now extensively utilized to bolster the development and construction of intelligent transportation systems (Ribeiro et al., 2023; Yu et al., 2023; Zhang et al., 2024). Due to the unique benefits of AIS data, including extensive coverage and verifiable information, and its integration with detailed geographic and environmental data, considerable efforts have been made to construct AIS data-based traffic scenarios for autonomous testing. For example, Bakdi et al. (2021) devised a framework that uses a high-performance computational approach to analyze historical data and create realistic testbed scenarios. This framework integrates detailed traffic data from AIS with high-resolution digital maps, ship registries, and nautical charts to analyze collision and grounding risks, thereby producing robust and authentic testing scenarios. Zhu et al. (2022) developed a scenario-based validation method for autonomous collision avoidance systems by extracting ship encounter scenarios from historical AIS data and generating virtual test scenarios that mirror the parameter probability distributions observed in actual scenarios. Zhou et al. (2024) introduced a technique for parametric modeling and encounter scenario generation for autonomous anti-collision testing in inland waterways, utilizing AIS and radar data. To identify critical encounter scenarios for autonomous testing support, Wang et al. (2024a) proposed a comprehensive framework that assesses both the collision risk and exposure frequency of AIS-based traffic scenarios. Furthermore, Wang et al. (2024b) examined the spatial-temporal dependencies among ships, the complexity of traffic scenarios, and the application strategies for tests, aiming to enhance the rigor of autonomous testing.

Despite the popularity and substantial capabilities of AIS data-based scenarios to deliver sufficiently realistic test scenarios for comprehensive autonomous testing, state-of-the-art research reveals several critical shortcomings that remain largely unaddressed, particularly in terms of accurate scenario extraction, characterization, and sampling, as detailed below:

1. Proper encounter situation recognition is the basis for precisely extracting traffic navigation scenarios. Previous studies have predominantly relied on combining distance parameters and metrics based on the Closest Point of Approach (CPA) to define the spatial-temporal encounter relationships among ships (Cho et al., 2020; Wang et al., 2024a; Zhang et al., 2023). However, they often overlook the impact of factors like potential ship motion dynamics and constrained waterway geography. Meanwhile, few studies have delved into creating dynamic evolution scenarios that feature a continuous spatial-temporal interaction process among multiple ships, which would allow for observing dynamic behaviors such as ships joining and leaving scenarios (Bakdi et al., 2021). Developing advanced models that integrate these elements and account for scenario dynamics offers the potential to accurately and comprehensively capture the evolutionary interactions among multiple ships over time, yet this remains a significant unresolved issue.
2. Analyzing traffic scenario characteristics goes beyond basic assessments of encounter risks and types, which necessitates a comprehensive disclosure of scenario details from a broader perspective. Failing to fully expose these characteristics could fundamentally compromise the accurate interpretation of traffic behaviors within the scenario, potentially leading to unconscious biases in autonomous testing. Although numerous studies focus on either the encounter risks or types (with a few addressing both), they often overlook the trajectory motion patterns of ships involved in the scenario, even though these patterns are crucial for revealing typical navigational behaviors in specific environments and assisting autonomous ships in either replicating or optimizing these behaviors (Li et al., 2022). Additionally, current scenario characterizations concerning encounter risks and types are limited to two major aspects. Firstly, most models inadequately capture the dynamic conflict risks and complexities arising from ship motion dynamics, maneuverability constraints, and intricate topological interactions among multiple ships (Wang et al., 2024a; Zhang et al., 2023). There is also a significant shortfall in the use of advanced classification and parameterization techniques for hierarchically quantifying the complexity levels of conflicts within traffic scenarios (Bakdi et al., 2021). These elements are crucial for conducting accurate and detailed analyses of dangerous situations, especially for traffic scenarios originating from

complex maritime environments characterized by high traffic density, diverse motions, and complex traffic compositions. Secondly, while studies typically identify specific encounter types at the initial moments (Hwang and Youn, 2021; Wang et al., 2024a), they often fail to consider these types throughout the encounter process, hindering the systematic encoding and comprehensive understanding of traffic co-behaviors during encounters. To achieve an accurate and comprehensive understanding of extracted traffic scenarios, it is essential to conduct a systematic analysis of scenario characteristics. This involves jointly revealing the ship motion patterns, conflict complexities, and encounter types throughout the entire scenario.

3. The enormous volume of AIS data makes it possible to generate countless traffic navigation scenarios for autonomous ship testing. Thus, exploring the diversity of these scenarios is pivotal for selecting representative scenarios to optimize testing efficiency. However, most studies have assessed scenario diversity from a single angle, focusing only on elements such as collision risk or encounter type (Bolbot et al., 2022; Chen et al., 2021; Hwang and Youn, 2021). There has been no systematic approach that examines scenario diversity across multiple dimensions. For instance, no existing approach jointly integrates factors like ship motion patterns, conflict complexities, and encounter types during screening processes to ensure the selection of scenarios that are truly representative of real-world conditions. Consequently, the development of advanced sampling techniques, guided by multiple selection principles, is crucial for achieving comprehensive coverage and efficiency in autonomous testing. This involves choosing a smaller, yet fully representative subset of scenarios that comprehensively captures the diversity of the original set.

In summary, the identified limitations significantly hinder the progress of MASS implementation, requiring a novel study addressing all three gaps holistically. Therefore, this study is devoted to developing a new holistic framework for the extraction, characterization, and sampling of traffic scenarios. This framework will integrate advanced and new models to collectively overcome these challenges, thereby supporting comprehensive and reliable autonomous testing and verification.

3. Methodology: Traffic scenario extraction, characterization, and sampling

Fig. 1 presents a systematic framework for extracting, characterizing, and sampling traffic scenarios from historical AIS data, structured into three key modules. The first module develops an advanced dynamic scenario extraction method that integrates enhanced CPA-based parameters, distance, and obstacle spatial distribution to accurately capture the real encounter relationships among ships over time. The second module characterizes the extracted scenarios from three perspectives: 1) The trajectory movement patterns of involved tested ships are analyzed by systematically combining the Dynamic Time Warping (DTW) model and the Ordering Points To Identify the Clustering Structure (OPTICS) algorithm; 2) the conflict complexity levels are meticulously quantified and parameterized by holistically defining and synthesizing 4 conflict-related indicators, employing the advanced Quaternion Ship Domain (QSD) model, a spatial risk model, and the Fuzzy Clustering Iterative (FCI) model; and 3) the encounter types are detailed for the first time through both momentary encounter features and those derived from the entire encounter process. Given the extensive number of traffic navigation scenarios extracted from AIS data, the third module designs a hierarchical greedy sampling technique to selectively identify representative scenarios for autonomous testing, based on the results from scenario characterization along with advanced

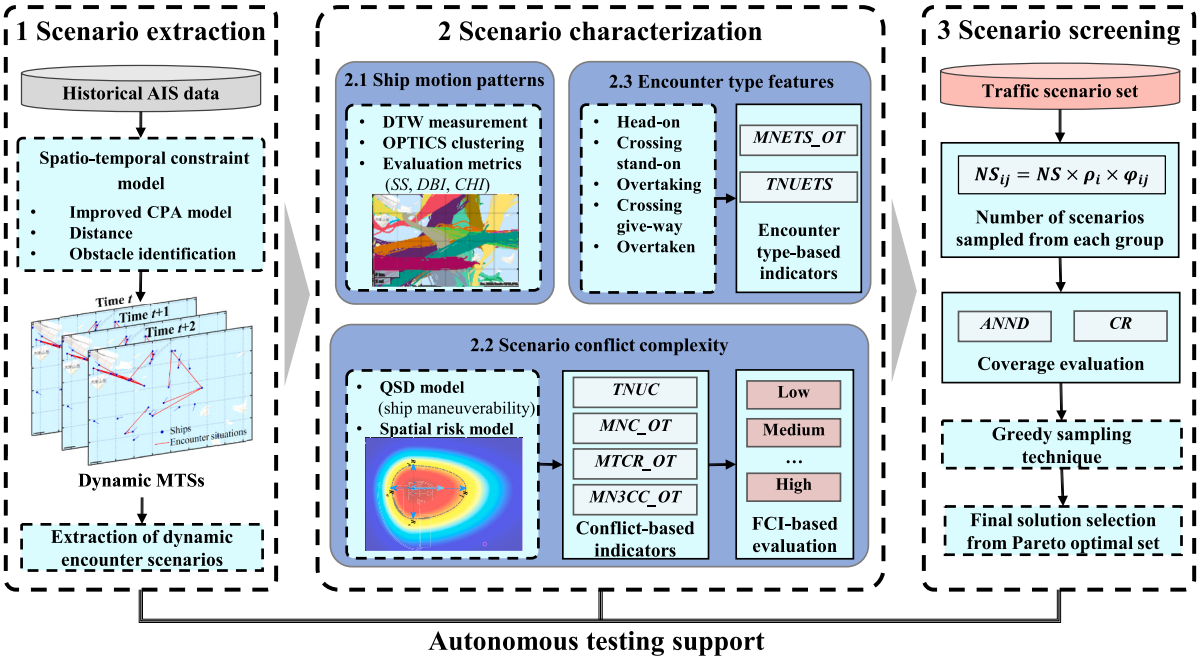

Fig. 1. Research framework.

spatial coverage metrics such as Average Nearest Neighbor Distance (*ANND*) and Coverage Radius (*CR*). Detailed technical descriptions of each module are provided in the subsequent subsections.

3.1. Traffic scenario extraction

Ship traffic navigation scenarios represent a dynamic evolutionary process that captures collective behavior and interactions among ships over time. At any given moment, the encounter relationships between any ship pair can be determined using the following spatial-temporal constraint model (Cho et al., 2020; Zhang et al., 2023; Zhu et al., 2022):

$$ER_{ij}(t) = \begin{cases} 1 & D_{ij}(t) \le \gamma_D \cap DCPA_{ij}(t) \le \gamma_{DCPA} \cap 0 < TCPA_{ij}(t) \le \gamma_{TCPA} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $D_{ij}(t)$, $DCPA_{ij}(t)$, and $TCPA_{ij}(t)$ represent the distance, Distance to the Closest Point of Approach (DCPA), and Time to the Closest Point of Approach (TCPA) between ships i and j at time t, respectively. The thresholds γ_D, γ_{DCPA}, and γ_{TCPA} are set for these measurements. $ER_{ij}(t) = 1$ indicates that ships i and j have an encounter relationship, while $ER_{ij}(t) = 0$ indicates the absence of such a relationship.

However, this model's effectiveness diminishes in complex encounter scenarios where ships exhibit varied motion dynamics. This limitation arises because DCPA and TCPA calculations assume that the encountering ships maintain a constant speed over a finite lookahead horizon. To address this, an enhanced CPA-based model in Zhang et al. (2015) was developed to precisely describe the relative spatial-temporal proximity relationship between ships by considering their potential dynamic motions, which is employed in this study. This model describes the future dynamic trajectories of encountering ships using a series of waypoints and calculates the CPAs between all pairs of consecutive waypoints between ships, subsequently identifying the minimal CPA. Detailed discussions on this model and its effectiveness are available in prior studies (Zhang et al., 2015). Moreover, Eq. (1) fails to account for the impact of complex water topography on the spatial-temporal proximity relationship between ships. In challenging water environments, encountering ships may be separated by obstacles such as landmasses, islands, and shallow waters. In such cases, even if their spatial-temporal proximity relationship satisfies Eq. (1), it does not constitute a real encounter situation.

Therefore, by incorporating the effects of both ship motion dynamics and water topography, Eq. (1) is refined and Eq. (2) is introduced to accurately identify genuine encounter situations, as follows:

$$RER_{ij}(t) = \begin{cases} 1 & D_{ij}(t) \le \gamma_D \cap IDCPA_{ij}(t) \le \gamma_{DCPA} \cap 0 < ITCPA_{ij}(t) \le \gamma_{TCPA} \cap O_{ij}(t) = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $IDCPA_{ij}(t)$ and $ITCPA_{ij}(t)$ denote the improved CPA-based parameters, and $O_{ij}(t) = 0$ indicates that there are no obstacles between ships at their CPA during the encounter. The spatial distribution of obstacles (i.e., unnavigable water areas) is inferred by analyzing the spatial density distribution of ship traffic based on historical trajectory data (Xin et al., 2023). Specifically, Kernel Density Estimation (KDE) is employed to differentiate between navigable and unnavigable areas. The investigated water area is divided into a series of grids, and the spatial probability distribution of AIS data points within each grid is evaluated against a defined threshold to determine its navigability. Additionally, the threshold values for γ_D, γ_{DCPA}, and γ_{TCPA} are set at 6 nautical miles, 1 nautical mile, and 15 min, respectively, based on both reference and research water characteristics (Bakdi et al., 2021; Zhu et al., 2022).

Building on the above formula, the spatial-temporal dependencies between any two ships at a given traffic snapshot can be identified, defined as a Momentary Traffic Situation (MTS). Each MTS, comprising a set of complex interdependent encounter relationships among ships, forms a graph structure. As shown in Fig. 2, these graph structures are sequentially constructed by analyzing

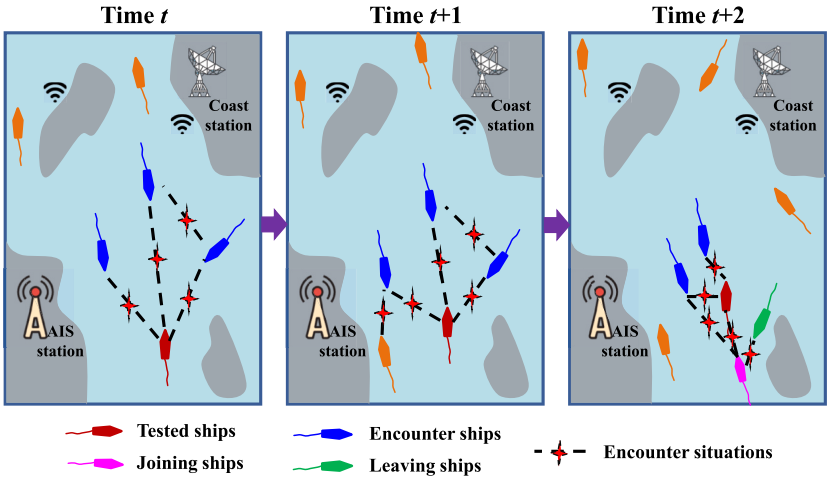

Fig. 2. MTSs involving a dynamic encounter scenario over time.

the encounter relationships between all ship pairs in each MTS. Therefore, this approach facilitates the construction of spatial-temporal dependencies between all ships in a regional area over time.

During the autonomous ship testing process, it is crucial to analyze the interactions between the autonomously functioning ship and other ships as they move. In this setup, one ship is designated as the tested ship (i.e., the own ship) with autonomous functions, while the other ships serve as target ships involved in encounters with the tested ship. Utilizing the graph relationships constructed at each moment, any ship can be designated as the tested ship, allowing for the extraction of target ships with which it has encounter relationships from its initiation to the end. This creates a complete and dynamically evolving encounter scenario, characterized by dynamic interactions between the tested ship and target ships over time. Fig. 2 shows the target ships (marked in blue) that have encounter relationships with ship *A* (marked in red) over time, presenting a sequence of temporally dependent traffic encounter situations. By assigning different ships as the tested ship, various scenarios featuring distinct interactions among ships in temporal sequences can be captured, allowing for the formulation of a comprehensive library of scenarios for autonomous testing.

3.2. Traffic scenario characterization

This subsection systematically characterizes the extracted traffic scenarios across three key aspects: ship motion patterns, conflict complexities, and encounter types. The motion patterns shed light on the typical navigational behaviors of tested ships; conflict complexities highlight the layered differences in scenario conflict criticality; and encounter type features showcase the richness of multi-ship encounter situations. These characterizations enable effective scenario encoding, thereby offering a comprehensive understanding of traffic co-behaviors within each scenario and guiding their application in subsequent testing phases. Table 1 provides a summary of the scenario characterization metrics and their corresponding descriptions.

3.2.1. Motion pattern classification of tested ships

Trajectory clustering is the primary method for uncovering hidden movement patterns. It puts trajectories with similar spatial-temporal behaviors into the same clusters, while trajectories between clusters exhibit relatively low spatial-temporal similarities. Therefore, this study integrates advanced distance measurement methods and clustering techniques to analyze the motion patterns of tested ships in the extracted scenarios, thereby revealing typical navigational behaviors in specific environments and supporting autonomous ships in mimicking or optimizing these behaviors. Ship trajectory clustering involves two crucial components: first, the selection of distance/similarity measurement models, which must fully consider the spatial-temporal characteristics of trajectories, such as dynamic changes, trajectory length, and shape features; second, the adoption of clustering algorithms, focusing on the techniques and approaches used to effectively and efficiently group the trajectories.

(1) Trajectory distance measurement model

Ship trajectories are time series characterized by continuous timestamps, along with geographical positions (i.e. longitude and latitude), velocity, Course Over Ground (COG), and other parameters. This study employs DTW to calculate distances between different trajectories, as it effectively handles deformations along the time axis of these series (Li et al., 2022). Moreover, compared to methods like Euclidean Distance, Longest Common Subsequence, Hausdorff Distance, and Fréchet Distance, DTW offers significant advantages in analyzing data of varying lengths or series with time-shifting patterns (Berndt and Clifford, 1994). Detailed computational specifics for DTW can be found in Appendix A.

(2) Trajectory clustering algorithm

Table 1
Summary of scenario characterization metrics and their description.

Metrics	Category	Description
Trajectory motion pattern	Motion behavior	Clustered patterns of ship trajectories derived from spatial-temporal features using DTW and OPTICS.
Conflict complexity level	Conflict risk	Hierarchical classification of scenario danger levels using conflict-related indicators and FCI model.
Encounter type characteristics	Encounter type	A dynamic matrix representing the evolution of encounter types (e.g., head-on, overtaking, crossing) throughout the scenario.
Total Number of Unique Conflicts (*TNUC*)	Conflict-based indicator	Number of distinct target ships involved in conflict with the tested ship during the scenario.
Maximum Number of Conflicts at One Timestamp (*MNC_OT*)	Conflict-based indicator	Maximum number of simultaneous conflicts between the tested ship and target ships at a single time point.
Maximum Total Conflict Risk at One Timestamp (*MTCR_OT*)	Conflict-based indicator	Maximum cumulative conflict risk value between tested and all target ships at a single time point.
Maximum Number of 3-Cycle Conflicts at One Timestamp (*MN3CC_OT*)	Conflict-based indicator	Maximum number of triangular conflict loops involving the tested ship at a single time point.
Maximum Number of Encountered Target Ships across different encounter types at One Timestamp (*MNETS_OT*)	Encounter type-based indicator	Number of target ships involved in various encounter types (e.g., head-on, crossing) with the tested ship at the peak encounter moment.
Total Number of Unique Encountered Target Ships across different encounter types (*TNUETS*)	Encounter type-based indicator	Total number of distinct target ships that engage in different encounter types with the tested ship throughout the scenario.

OPTICS is a renowned clustering algorithm that builds upon the classic DBSCAN algorithm to effectively handle large-scale datasets with varied density distributions. Compared to other traditional clustering methods, including prototypical, hierarchical, and graph-based algorithms, OPTICS provides numerous advantages:1) It automatically discerns cluster structures from data without the need to pre-specify the number of clusters; 2) OPTICS does not rely on assumptions about the shape of clusters, enabling the identification of clusters with any shape; 3) it shows high robustness in figuring out noise and outliers; and 4) through a reachability plot, OPTICS unveils the hierarchical structure and density distribution within the data, facilitating detailed exploration and interpretation of clusters at various levels (Rong et al., 2022). Due to these advantages, this study employs OPTICS to process ship trajectory data with varying densities and complex structures, effectively revealing the underlying clustering patterns within the data. A detailed description of OPTICS is provided in Appendix B.

3.2.2. *Conflict complexity evaluation of traffic scenarios*

Evaluating the conflict complexity levels of traffic scenarios uncovers varying levels of danger within testbed environments. To precisely evaluate the dynamic conflict risk between ships in diverse situations, a refined conflict risk evaluation approach incorporating the QSD model is utilized. This model accounts for factors such as ships' maneuverability limits, potential motion dynamics, and constrained water geography. Subsequently, four conflict-based indicators are holistically proposed to assess the conflict complexity of a complete traffic scenario from multiple perspectives. Additionally, an advanced FCI model is applied to synthesize these indicators to provide detailed insights into the hierarchical levels of conflict complexity. By innovatively integrating the aforementioned steps, it can achieve a precise and systematic quantification of the conflict complexity of traffic scenarios (See Part 2.2 in Fig. 1).

(1) Conflict risk qualification

A conflict typically occurs when ships approach each other within an unsafe distance over a predetermined future timeframe. To evaluate this risk, various ship domain models with distinct geometries have been developed (Szlapczynski and Szlapczynska, 2017; Zarzycki et al., 2025). Among these, the QSD model stands out as a widely used and practical approach, incorporating key characteristics of ships during encounters (Silveira et al., 2022; Wang, 2010). Shaped by the navigational rules known as COLREGs, this model adopted an asymmetric design with four distinct radii defining the fore, aft, starboard, and port boundaries (see Fig. 3). These radii, which reflect the ship's maneuverability and solely depend on its size and speed, are specified as follows:

$$\begin{cases} R_f = \left(1 + 1.34\sqrt{{k_{AD}}^2 + \left(\frac{k_{DT}}{2}\right)^2}\right)L \\ R_a = \left(1 + 0.67\sqrt{{k_{AD}}^2 + \left(\frac{k_{DT}}{2}\right)^2}\right)L \\ R_s = (0.2 + k_{DT})L \\ R_p = (0.2 + 0.75k_{DT})L \end{cases} \tag{3}$$

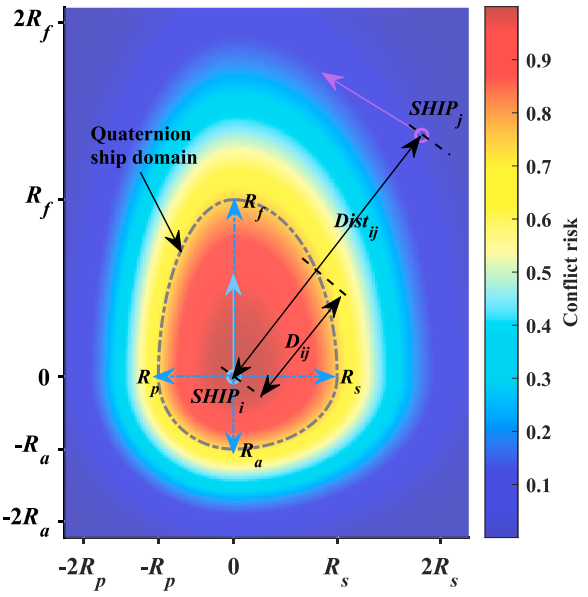

Fig. 3. Illustration of instantaneous conflict risk quantification using the QSD model and spatial risk model.

where R_f, R_a, R_s, and R_p represent the domain radii at the fore, aft, starboard, and port sides, respectively. L denotes the ship length, and k_{AD} and k_{DT} are coefficients linked to the advance (A_D) and tactical diameter (D_T), respectively. These coefficients, indicative of the ship's fundamental maneuvering capabilities, are calculated using the equations below:

$$\begin{cases} k_{AD} = \dfrac{A_D}{L} = 10^{0.3591 log_{10}(V) + 0.0952} \\ k_{DT} = \dfrac{D_T}{L} = 10^{0.5441 log_{10}(V) - 0.0795} \end{cases} \tag{4}$$

where V denotes the ship's speed in knots.

Traditional ship domain-based models typically assess conflict risk through a binary system, labeling situations as either safe or dangerous based on whether ships' domains overlap or invade each other. This study refines these evaluations by combining the QSD model with a spatial risk model that employs an exponential decay function. This approach allows for the calculation of more granular conflict risk scores that continuously range from 0 to 1, where 0 signifies a safe encounter and 1 indicates a collision (Bakdi et al., 2021). Additionally, the criterion that neither ship's domain area should be invaded is adopted as the safety standard, chosen for its straightforward implementation and consideration of both encountering ships' characteristics (Szlapczynski and Szlapczynska, 2017). For example, with two ships, $SHIP_i$ and $SHIP_j$ as illustrated in Fig. 3, an instantaneous conflict risk score for $SHIP_i$, acting as the own ship, can be calculated using the exponential decay function outlined below (Wang, 2010).

$$IRS_{i\to j}^{t} = e^{-\left(\frac{Dist_{ij}^{t}}{D_{ij}^{t}}\left(\ln\left(\frac{1}{r_0}\right)\right)^{\frac{1}{3}}\right)^{3}} \tag{5}$$

where $Dist_{ij}^{t}$ represents the distance between $SHIP_i$ and $SHIP_j$ at time t, and D_{ij}^{t} indicates the distance from the center of $SHIP_i$ to its domain boundaries along the line connecting the two ships (see Fig. 3). The decay parameter r_o, set at 0.5, defines the relationship between domain size and conflict risk score. Similarly, the instantaneous risk score $IRS_{j\to i}^{t}$, with $SHIP_j$ as the own ship, can be calculated. Given that $IRS_{i\to j}^{t}$ and $IRS_{j\to i}^{t}$ might differ due to varying degrees of domain invasion, depending on which ship is considered as the reference, the higher of the two scores is used as the instantaneous conflict risk between the two ships. This approach adheres to the worst-case scenario principle commonly applied in safety sciences, as follows:

$$ICR_{ij}^{t} = \max\left(IRS_{i\to j}^{t}, IRS_{j\to i}^{t}\right) \tag{6}$$

where ICR_{ij}^{t} represents the instantaneous conflict risk between $SHIP_i$ and $SHIP_j$ at time t.

Note that Eq. (6) calculates the instantaneous conflict risk between the encountering ships at a specific moment, whereas conflict risk generally assesses potential risk over an upcoming predefined period (Hernandez-Romero et al., 2019). As such, the dynamic conflict risk between $SHIP_i$ and $SHIP_j$ at time t (i.e., DCR_{ij}^{t}) is defined by the highest value of $ICR_{ij}^{t'}$ within this future timeframe $[t, t+T]$, as outlined below:

$$DCR_{ij}^{t} = \max_{t' \in [t, t+T]} ICR_{ij}^{t'} \tag{7}$$

where T denotes the conflict detection horizon set at 15 min, aligning with the parameter thresholds used in the encounter situation identification model described in Section 3.1. Computing the instantaneous conflict risk at every future moment within $[t, t+T]$ (as per Eq. (7)) would require significant computational resources. To address this, the minimum passing distance between ship pairs within this timeframe is first determined using the improved CPA model (Zhang et al., 2015). The instantaneous conflict risk at this minimum passing distance is then assessed based on Eqs. (5)-(6), which is used to represent the dynamic conflict risk in Eq. (7). Additionally, following the approach of the encounter situation identification model, if any part of the line connecting the ships at their closest passing points intersects with an obstacle-marked area, the situation is directly deemed a safe ship-pair encounter, i.e., the conflict risk between ships is set to 0.

(2) Scenario conflict complexity measurement indicators

The conflict dynamics between a tested ship and target ships in a complete traffic navigation scenario are exceedingly complex, involving varying degrees of conflict severity with different target ships at distinct moments. Hence, the conflict complexity can be analyzed from multiple perspectives. This study develops four representative indicators to systematically describe different aspects of conflict complexity throughout a complete traffic scenario. The first three indicators assess the number and severity of conflicts, while the fourth reveals the topological characteristics that highlight the challenges in resolving these conflicts. These indicators are being used holistically for the first time for this purpose and are defined as follows:

- **Total Number of Unique Conflicts (*TNUC*):** the total number of target ships that engage in conflicts with tested ship throughout the scenario.

- **Maximum Number of Conflicts at One Timestamp (*MNC_OT*):** the highest number of conflicts occurring between target ships and tested ship at a single timestamp.
- **Maximum Total Conflict Risk at One Timestamp (*MTCR_OT*):** the highest cumulative conflict risk value between all target ships and tested ship at a single timestamp.
- **Maximum Number of 3-Cycle Conflicts at One Timestamp (*MN3CC_OT*):** the greatest number of 3-cycle conflicts involving tested ship occurring at a single timestamp. A 3-cycle conflict refers to a case where three ships form a closed loop of conflict relationships (see Fig. 4), illustrating complex structural conflicts among multiple ships.

Evidently, the larger these indicators are, the higher the corresponding conflict complexity.

(3) Scenario conflict complexity evaluation

Evaluating scenario conflict complexity inherently necessitates a multi-index assessment, as multiple indicators are utilized to uncover different aspects of the conflicts. This study adopts the FCI model due to its strong hierarchical structuring and precise classification capabilities. Traditional methods like Gray Relational Analysis (GRA), the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), and Analytic Hierarchy Process (AHP), depend heavily on subjective qualitative standards and experience-based criteria. In contrast, the FCI model combines fuzzy theory and clustering algorithms, excelling in 1) handling uncertainty, randomness, and fuzziness within datasets and 2) effectively clustering multivariate data without the need for prior information. The efficacy of its application across various benchmark experiments and real-world scenarios has been thoroughly demonstrated in previous studies (He et al., 2011; Lu et al., 2017).

The essence of FCI lies in its data-driven techniques for determining the optimal memberships of each scenario across different complexity levels. Once these memberships are established, the conflict complexity level of each traffic scenario can be identified based on its maximum membership, detailed as:

$$SCC_i = \text{argmax}_{\{k=1,2,\cdots,c\}} u_{ki} \tag{8}$$

where SCC_i denotes the conflict complexity level of the ith traffic scenario, u_{ki} represents the membership of scenario i to the kth complexity level, and c represents the number of complexity levels. Further details related to FCI are provided in Appendix C.

3.2.3. *Encounter type analysis of traffic scenarios*

Rules 13-15 in Part B of the COLREGs establish the framework for characterizing encounter types between ships. These rules classify encounters into three main categories: head-on, overtaking, and crossing, based on the bearing of the target ship relative to the own ship (Zhu et al., 2022). Given the distribution of responsibility between target ship and own ship, crossing situations are further subdivided into crossing stand-on and crossing give-way, whereas overtaking situations are categorized as either overtaking or being overtaken (Liu et al., 2022b; Wang et al., 2024a). Fig. 5 depicts these five distinct types of encounters.

Currently, most research on ship encounter scenario analysis concentrates primarily on the type of encounters between ships at the initial moment of encounter (Bolbot et al., 2022; Hwang and Youn, 2021; Wang et al., 2024b). However, a ship traffic navigation scenario is inherently dynamic and continuously evolving. Therefore, it is crucial to account for the encounter type features throughout the entire navigational process. At any specific moment, the encounter types between the tested ship (i.e., the own ship) and target ships in a scenario can be represented as a five-element vector, with each element representing the number of a type of encounter as shown in Fig. 5. For example, the vector [2, 1, 0, 0, 0] indicates two head-on encounters and one crossing stand-on encounter between the tested ship and target ships. Consequently, the evolving encounter types between the tested ship and target ships in a complete traffic scenario can be systematically represented as a newly created matrix, as detailed below:

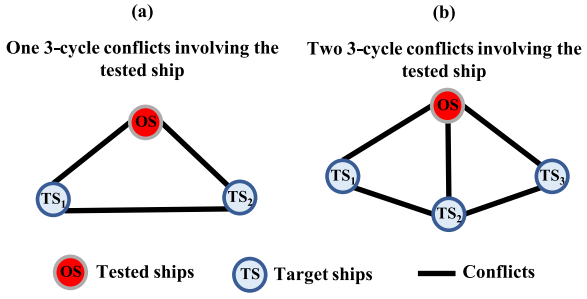

Fig. 4. Illustration of 3-cycle conflicts.

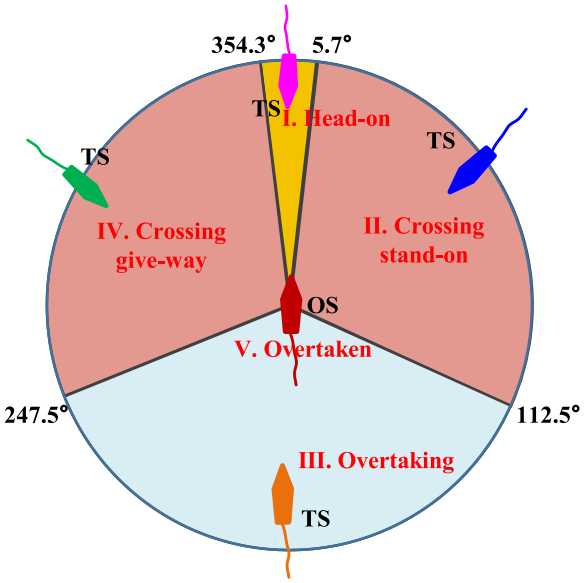

Fig. 5. Illustration of different ship encounter types.

$$ET_i = \begin{bmatrix} ET_{i,1,1} & ET_{i,1,2} & \cdots & ET_{i,1,5} \\ ET_{i,2,1} & ET_{i,2,2} & \cdots & ET_{i,2,5} \\ \vdots & \vdots & \ddots & \vdots \\ ET_{i,T,1} & ET_{i,T,2} & \cdots & ET_{i,T,5} \end{bmatrix} \tag{9}$$

where each row in ET_i represents the number of {head-on, crossing stand-on, overtaking, crossing give-way, and overtaken} encounters at one given moment in the ith scenario. For example, $ET_{i,2,1}$ represents the number of head-on encounters at time 2.

To comprehensively understand the features of scenario encounter types, two new sets of indicators extracted from ET_i are developed. The first set of indicators highlights the instantaneous features of encounter types, while the second set reveals the features over the duration of the encounter process, thereby offering a comprehensive understanding of the scenario encounter types. These include:

- **Maximum Number of Encountered Target Ships across different encounter types at One Timestamp (*MNETS_OT*):** the number of target ships that experience {head-on, crossing stand-on, overtaking, crossing give-way, or overtaken} encounters with the tested ship at the moment when the maximum number of encounters occurs.
- **Total Number of Unique Encountered Target Ships across different encounter types (*TNUETS*):** the total number of target ships that experience {head-on, crossing stand-on, overtaking, crossing give-way, or overtaken} encounters with the tested ship throughout the scenario.

Within these two sets of indicators, a 2×5 matrix can be constructed to describe the encounter type characteristics of a scenario, as follows:

$$ETC_i = \begin{bmatrix} MNETS\_OT_{i,1} & MNETS\_OT_{i,2} & \cdots & MNETS\_OT_{i,5} \\ TNUETS_{i,1} & TNUETS_{i,2} & \cdots & TNUETS_{i,5} \end{bmatrix} \tag{10}$$

where ETC_i represents the encounter type characteristics of the ith scenario.

3.3. Traffic scenario sampling approach

In practice, both sea trials and simulation-based tests for MASSs, particularly the former, are restricted to a certain number of scenarios due to efficiency considerations. However, the abundant AIS data provides a wealth of traffic navigation scenarios suitable for autonomous testing. It is therefore critical to select representative scenarios from a sufficiently realistic set to minimize the required number for testing while ensuring comprehensive coverage of potential real-world situations.

To achieve this, a hierarchical greedy sampling technique is newly developed to strike an optimal balance between test coverage and efficiency. This approach analyzes the diversity of scenarios by utilizing the characterization results of traffic scenarios in Section 3.2. It involves a three-step sampling process, depicted in Fig. 6, which selects scenarios based on scenario characteristics such as

motion patterns, conflict complexities, and encounter types sequentially. The first two steps are straightforward to implement, as the classification of motion patterns and conflict complexity levels for scenarios is well-defined in Sections 3.2.1 and 3.2.2. Assuming the total number of scenarios sampled is *NS*, these steps determine the number of scenarios to be sampled from each motion pattern and conflict complexity level using the following equations:

$$NS_{ij} = NS \times \rho_i \times \varphi_{ij} \tag{11}$$

where ρ_i denotes the percentage of the *i*th motion pattern within the entire extracted scenario set, and φ_{ij} represents the percentage of the *j*th conflict complexity level within the *i*th motion pattern. These two parameters indicate the likelihood of different types of scenarios occurring in real traffic settings, corresponding to their exposure frequency. Consequently, a higher exposure frequency should naturally lead to a greater sample selection ratio.

After determining the sampled number for each motion pattern and conflict complexity level, the third step initiates the sampling process by considering the encounter type characteristics of scenarios. This step addresses two key issues: evaluating the coverage of the sampled scenarios relative to the original scenario set, and designing the sampling strategies.

Regarding the first issue, two tailored metrics are developed and used within this context: Average Nearest Neighbor Distance (*ANND*) and Coverage Radius (*CR*) (Ripley, 1988). The former calculates the average distance from each point in the original set to its closest counterpart in the sampled set, reflecting the overall proximity of the sampled points to the original dataset. The latter metric assesses the maximum of the minimum distances between points in the original set and points in the sampled set, thus highlighting the worst-case coverage situation. *ANND* assesses overall representativeness but may overlook locally uncovered areas. In contrast, *CR* focuses on the maximum distance, enabling the identification of coverage blind spots and providing a different perspective on sample coverage. These two metrics collectively ensure that the coverage of the sampled points is comprehensively assessed from various perspectives, with both aiming for smaller values to indicate better performance. The formulas for these metrics are presented below:

$$ANND = \frac{1}{n}\sum_{i=1}^{n} \min_{j \in SS}\left(ED\_ETC_{ij}\right) \tag{12}$$

where *n* is the number of points in the original sample set and *SS* represents the sampled scenario set.

$$CR = \max_{i \in OS}\left(\min_{j \in SS}\left(ED\_ETC_{ij}\right)\right) \tag{13}$$

where *OS* represents the original sample set. Both metrics utilize ED_ETC_{ij}, the Euclidean distance between the *i*th and *j*th scenarios' encounter type-related matrices, calculated as follows:

$$ED\_ETC_{ij} = \sqrt{\sum_{v=1}^{2}\sum_{y=1}^{5}\left(ETC_i(v,y) - ETC_j(v,y)\right)^2} \tag{14}$$

Regarding the second issue, a greedy sampling technique that optimizes both *ANND* and *CR* is newly proposed. This objective-driven strategy operates as follows: In each iteration, one scenario is selected from the original set and added to the sampled scenario set *SS*. This selection is strategically made to ensure that the addition minimizes either *ANND* or *CR* for the updated *SS*, depending on which metric is currently prioritized. The choice of metric to minimize is made randomly during each iteration, which helps maintain an effective balance between *ANND* and *CR*. This process continues until the number of scenarios in *SS* reaches the predefined target.

Due to the high sensitivity of *SS* generated by the random selection of the optimization metric in each iteration, this strategy is executed numerous times to produce a variety of *SS* solutions, labeled as $\{SS_1, SS_2, \ldots, SS_N\}$. Subsequently, a set of Pareto optimal solutions that are non-dominated by any other solutions is identified. For example, if there are two optimal solutions, SS_i and SS_j, and

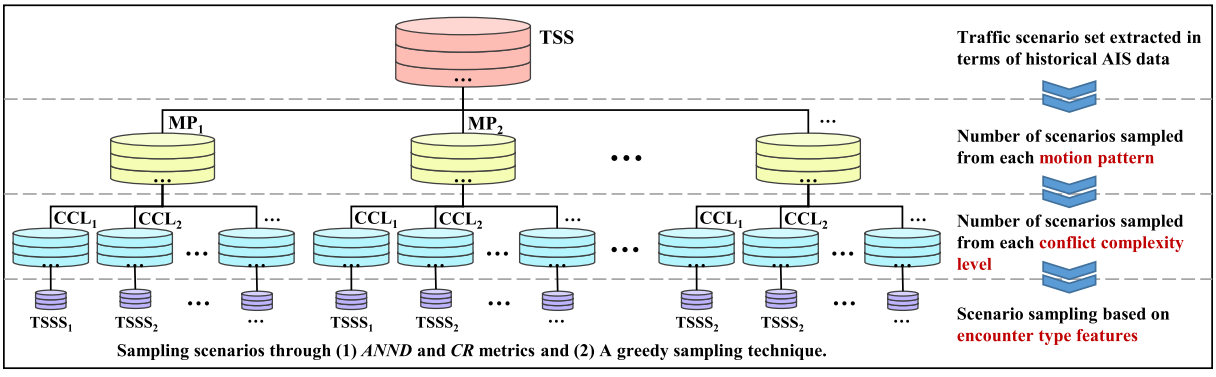

Fig. 6. Traffic scenario sampling procedure.

SS_i has a better *ANND* than SS_j, while SS_j outperforms SS_i in *CR*, then they are considered non-dominated. Finally, the ultimate sampled scenario set is chosen from this Pareto optimal set based on the median values to ensure a balanced consideration of both *ANND* and *CR*.

Based on the aforementioned procedure, the final sampled scenario set effectively covers a diverse range of motion patterns, conflict complexity levels, and encounter types, while also accounting for their exposure frequencies. This holistic approach is expected in theory to support high autonomous test coverage and efficiency, wanting a feasibility demonstration and validation through real cases in Section 4.

4. Application case analyses

4.1. Research area and data collection

To demonstrate the efficacy of the developed methodology, this study selects a major hub within Yangshan Port as the experimental research site. This area ranks among the busiest maritime zones in the world in terms of cargo throughput and serves as the sole deep-water harbor in Shanghai Port capable of accommodating the entry and exit of ultra-large ships. It is renowned for its complex traffic flows and diverse geographical conditions, marked by high traffic densities, varied and unpredictable ship movements, and a wide mix of traffic types, among others. Additionally, as shown in Fig. 7, this zone functions as a critical junction with numerous intersecting routes, significantly increasing the likelihood of ship collision incidents. In response, maritime authorities have designated it as an official precautionary zone, equipped with extensive ship routing systems and navigational regulations. Therefore, utilizing this water area for extracting, characterizing, and sampling traffic navigation scenarios is highly relevant for supporting autonomous testing in environments with complex traffic and geographical challenges. Demonstrating the effectiveness of the methodology in such a challenging environment supports its applicability to comparably complex or less complex waters worldwide.

The designated experimental area spans from 30°28′N to 30°38′N latitude and from 122°3′E to 122°20′E longitude. The historical AIS dataset analyzed comprises over 40 million records, each containing details such as time, Maritime Mobile Service Identity (MMSI) number, latitude, longitude, speed, COG, ship length, and ship type, over a period from April 1, 2020, to June 30, 2020. Recognizing the vulnerability of AIS data to inaccuracies due to technical malfunctions and calibration discrepancies, a comprehensive data cleaning and preprocessing protocol is implemented. Key procedures in this protocol include the elimination of outliers and the verification of trajectory consistency. Additionally, given the variability in AIS message transmission frequencies, a linear interpolation method is utilized to ensure consistent synchronization of AIS data across all ships. These steps significantly improve the reliability and completeness of the data used in subsequent analyses.

4.2. Traffic scenario extraction results

By employing the scenario extraction approach in Section 3.1, this study identifies various traffic navigation scenarios, with each scenario featuring a specific ship designated as the tested subject. Ships with a navigation duration exceeding 60 min are specifically selected as the tested ships for scenario extraction. Predominantly, testing is conducted in hazardous scenarios, as these risk-laden situations are crucial for evaluating autonomous navigation techniques, such as collision avoidance decision-making algorithms. Consequently, the focus is on scenarios that pose conflict risks (i.e., involving encountered ships with $DCR_{ij}^{t} > 0$), ultimately resulting in the extraction of 21,841 risky scenarios from pre-processed AIS data.

Each scenario retains detailed records including the start and end times, information about the tested and target ships at each moment, as well as the conflict risks and encounter types between ships. Fig. 8 displays some examples of the extracted traffic scenarios, where the light blue line indicates the trajectory of the tested ships, and other lines represent the trajectories of target ships. The

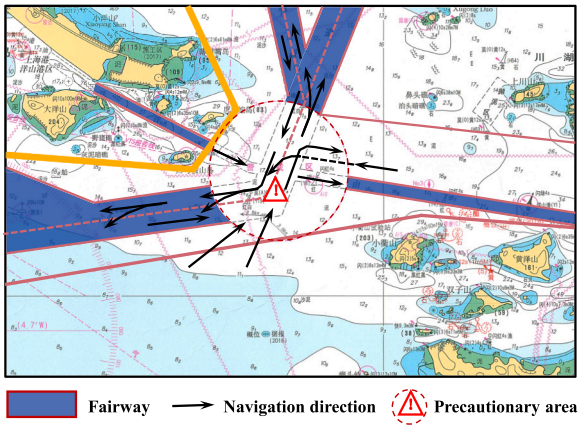

Fig. 7. Research area.

symbols ‘o’ and ‘x’ mark the start and end points of each trajectory, respectively. Dotted red lines connect the co-start points between the tested ship and target ships, while dotted black lines connect the co-end points. These extracted scenarios underscore the complex spatial-temporal dependencies among ships, as well as the dynamic evolution of encounter situations. They illustrate fluctuations in the number of encountered target ships and variations in the relative bearings and distances of these encounters over time. Compared to studies that focus on autonomous testing in straightforward and stable scenarios involving two-ship or multi-ship encounters, typically based on traditional hand-crafted scenarios, these extracted complex and highly dynamic scenarios provide: (i) complete temporal evolution of multi-ship interactions, (ii) continuous speed and heading adjustment behaviors, and (iii) realistic constraints imposed by waterways such as restricted channels, bends, and traffic lanes. They therefore offer a much richer testbed for validating interactions and cooperative behaviors between autonomous and manned ships. A detailed characterization analysis of all extracted scenarios will be presented in the subsequent subsection.

4.3. Traffic scenario characterization results

This subsection presents the characterization results of the extracted traffic scenarios, detailing the movement patterns of tested ships, conflict complexities, and encounter types in sequence. Following this, a systematic scenario analysis is conducted to demonstrate the methodology’s effectiveness in delineating a well-defined traffic scenario across these three characteristic types.

4.3.1. Movement pattern classification results

As outlined in Section 3.2.1, this study systematically combines the DTW and the OPTICS algorithm to identify the motion patterns of the tested ship in each scenario. Therefore, it is crucial to first determine two hyperparameters, *MinPts* and ϵ, for the OPTICS algorithm (see Appendix B). As illustrated in Fig. 9(a), the reachability plots from OPTICS across different *MinPts* values, especially those between 35 and 75, exhibit similar clustering structures. This consistency underscores the robustness of the OPTICS algorithm against variations in parameters, leading to the choice of $MinPts = 50$ for this study. Additionally, the hierarchical clustering structure is identified by capturing valleys in the reachability plot, which makes the initial setting of parameter ϵ, strictly speaking, unnecessary (Rong et al., 2022). In this study, valleys, each representing a cluster, are manually identified by combining the reachability plot with the visualization of trajectory clustering distribution.

Fig. 10 showcases the clustering results for ship movement patterns. As illustrated in Fig. 10(a), distinct clusters reveal different routes and directions of ship movements. Fig. 10(b) displays the number of trajectories within each pattern, highlighting the relative

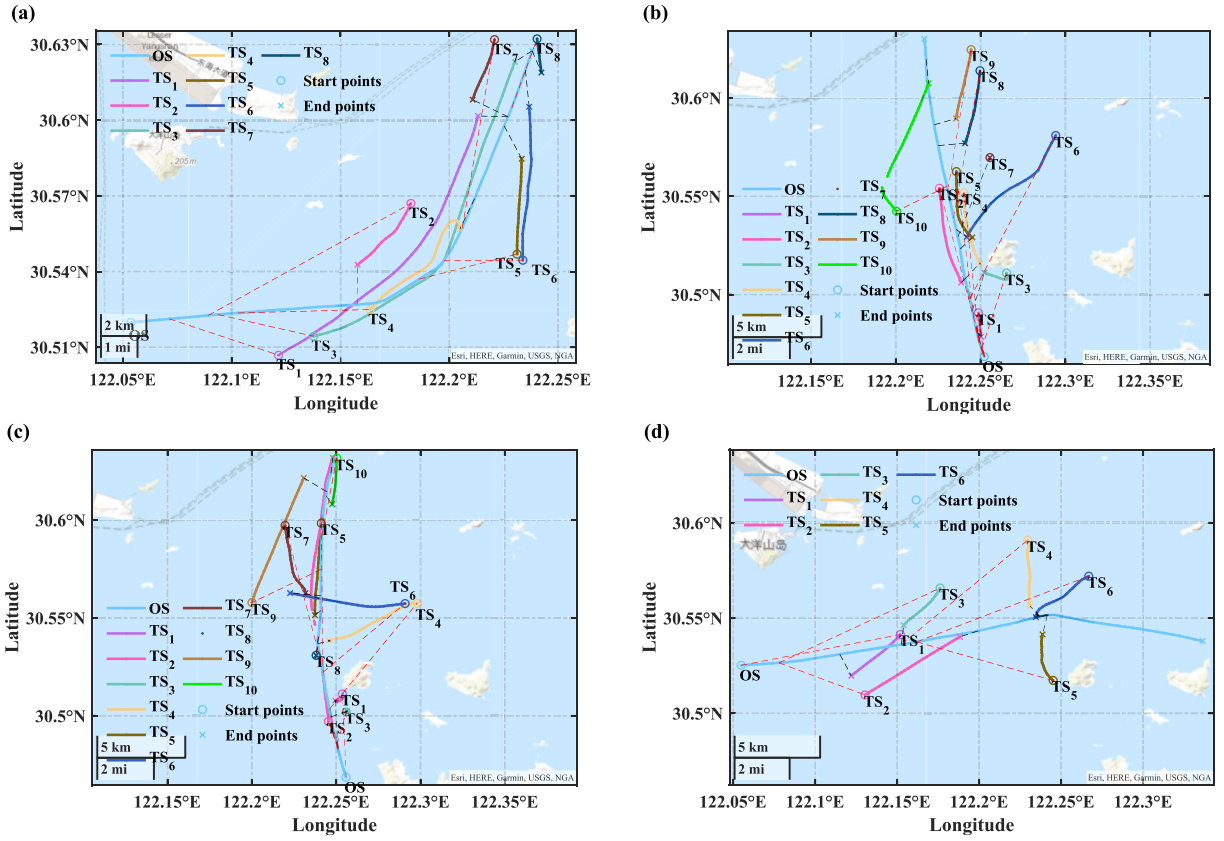

Fig. 8. Illustration of extracted traffic navigation scenarios.

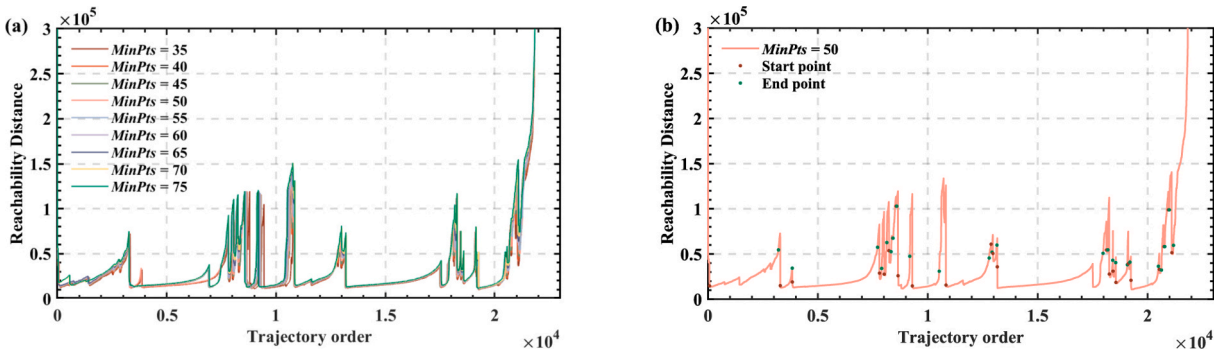

Fig. 9. (a) OPTICS reachability plots across different *MinPts* values; (b) OPTICS reachability plot when *MinPts* = 50.

busyness of different motion patterns. Specifically, Clusters 1, 3, and 14 are the most prominent patterns. Moreover, Fig. 10(c) shows the spatial distributions of each motion pattern, clearly demonstrating their distinct spatial characteristics. Trajectories within the same patterns exhibit consistent motion characteristics, whereas those from different patterns demonstrate contrasting motion characteristics. To validate the accuracy and effectiveness of the clustering results further, three typical indicators including the Silhouette Score (*SS*), Davies-Bouldin Index *(DBI*), and Calinski-Harabasz Index (*CHI*) are adopted. The details of these indicators' calculations are given in Appendix D. The values of Average *SS*, *DBI*, and *CHI* are listed in Table 2, all falling within ideal ranges. These observations affirm the satisfactory and well-structured clustering of ship motion patterns.

Overall, the classification identifies 24 distinct behavior patterns, which can aid autonomous ships in emulating various behaviors

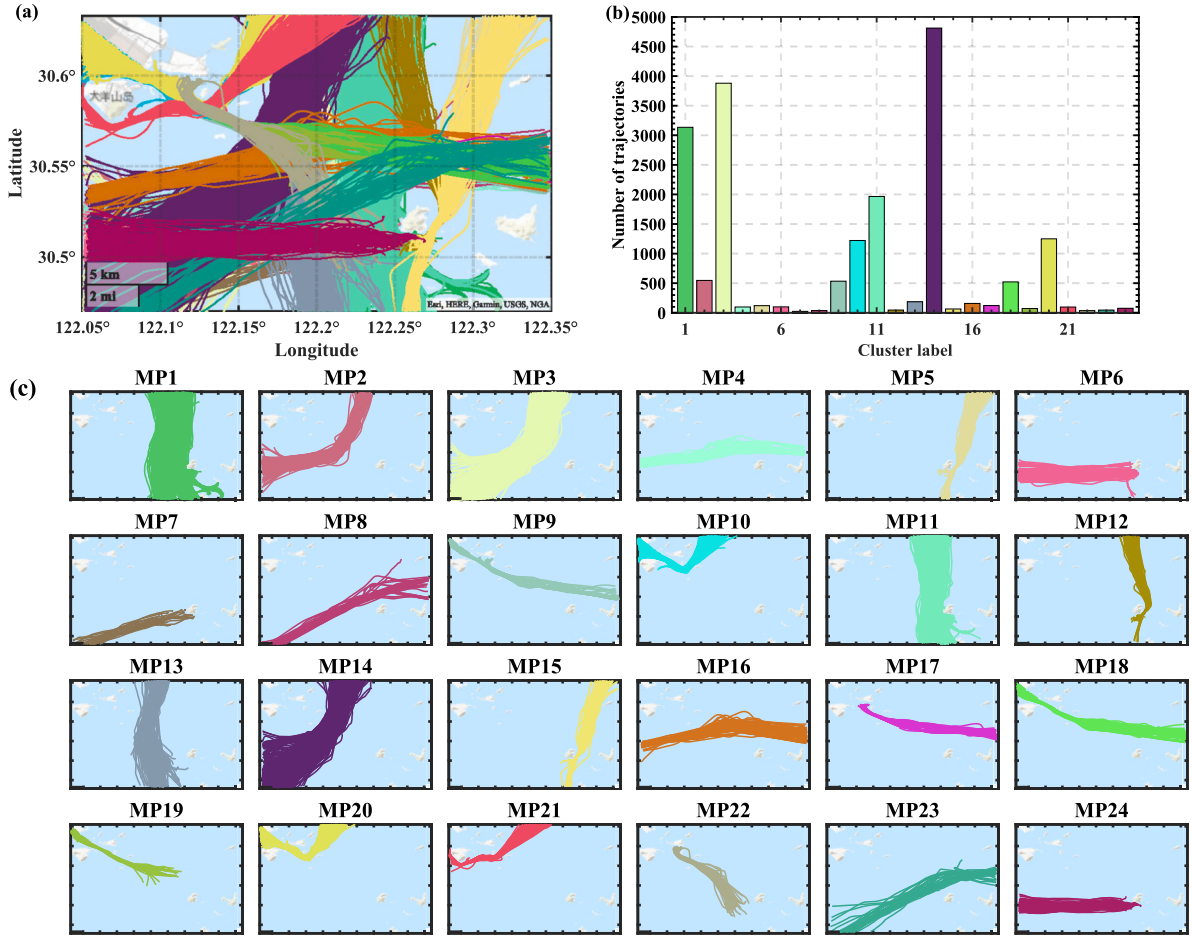

Fig. 10. Clustering results of ship movement patterns based on DTW and OPTICS. (a) Visualization of trajectory clustering distribution; (b) number of trajectories in each trajectory cluster; (c) spatial distribution visualization of each trajectory cluster.

during performance tests at the research site. In particular, the extracted motion patterns, such as Clusters 9, 10, 18, 19, 20, and 21, are clearly governed by the narrow channel, adjacent shoals, and designated traffic lanes, each occurring overwhelmingly within the restricted corridor highlighted in Fig. 7. These patterns therefore effectively represent narrow passage navigation.

4.3.2. Conflict complexity training results

Based on the extracted scenarios, the four conflict-related indicators for each scenario are calculated according to the methodology described in the first and second parts of Section 3.2.2. Fig. 11 illustrates the distribution characteristics of these indicators across all extracted traffic scenarios. To reveal the conflict complexity of each scenario, these indicators are aggregated using the FCI model, which hierarchically determines the conflict complexity levels. Consequently, Fig. 12 showcases the training results of the FCI model.

As depicted in Fig. 12(a), the relationships between the number of input complexity levels and the FCI's objective function (i.e., $F(u_{ki}, s_{jk})$ in Eq. (C.2)) are clarified. It is noted that an elbow point occurs at 4 complexity levels. Hence, in terms of the hand-elbow method, the optimal traffic complexity levels are categorized into 4 groups, labeled from Level 1 to Level 4, ranging from low to high. Fig. 12(b) details the optimization iterations within the FCI process, employing 20 randomly initialized cluster centers (i.e., u_{ki} in Eq. (C.3)) as inputs. The objective function $F(u_{ki}, s_{jk})$ achieves rapid convergence and exhibits remarkable stability, independent of the initial cluster center choices. This highlights the FCI model's effectiveness and resilience in optimization processes. Additionally, Fig. 12(c) reveals the number of scenarios across different complexity levels, highlighting the distribution characteristics of scenario conflict complexities. Furthermore, Fig. 12(d) showcases the final optimal cluster centers for each conflict-related indicator at each complexity level, demonstrating a strong hierarchical structure.

To further validate the efficacy of the FCI model in classifying conflict complexity levels, Fig. 13 illustrates the distributions of traffic complexity classification results based on any three conflict-related indicators. This clearly displays distinct spatial value distributions of conflict-based indicators corresponding to different complexity levels. Specifically, samples within the same complexity levels exhibit close proximity, whereas those from different complexity levels show relatively greater distances. This observation emphasizes the FCI model's effectiveness in hierarchically and precisely classifying conflict complexity levels.

4.3.3. Encounter type analysis results

Fig. 14 illustrates the distribution characteristics of encounter type-related indicators, as detailed in Section 3.2.3, across all extracted traffic scenarios. According to the figure, the maximum number of encountered target ships in head-on, crossing stand-on, overtaking, crossing give-way, and overtaken situations at a specific moment varies from 0-13, 0-18, 0-11, 0-23, and 0-7, respectively. Meanwhile, the total number of unique encountered target ships with respect to these encounter types ranges from 0 to 20, 0-24, 0-14, 0-43, and 0-10. The substantial variation in these indicators among different scenarios underscores the diversity of the extracted scenarios. Unlike traditional designs that typically involve simpler two-ship or multi-ship encounters for autonomous testing, the scenarios depicted here are significantly more complex. They feature a dynamic array of encounter types over time, which more accurately mirrors real-world conditions, particularly in highly complex coastal or port waters. These scenarios incorporate the continuous joining of new ships and the departure of existing ones, constantly updating the encounter dynamics. Using these complex and evolving scenarios for autonomous testing can more successfully build confidence and verify the safety, security, and reliability of MASSs.

4.3.4. Illustration of traffic scenario characterization

Based on the analysis results described above, a comprehensive characterization analysis for any traffic navigation scenario can be conducted. Fig. 15 presents a detailed characterization of a complete traffic navigation scenario. In Fig. 15(a), the movement trajectories of the tested ship and target ships are displayed, showing how they navigate relative to each other. Fig. 15(b) displays the evolving distances between the target ships and the tested ship over time, typically following a pattern where they first move closer together and then apart. Solid lines in the figure indicate scenarios where the target ships and the tested ship either constitute encounter situations or experience real-time ship domain intrusions. It is noted that in most cases, as the distance between the tested ship and target ships shifts from decreasing to increasing, the line transitions from solid to dashed, indicating that they no longer form encounter situations. This transition demonstrates the effectiveness of the encounter situation identification model used in the analysis.

Furthermore, Fig. 15(c)-(e) illustrates momentary encounter situations at specific times, where 'o' symbols mark the current positions of ships, while dotted lines indicate their projected future trajectories as identified by the improved CPA method. The complex QSD visualization reveals the spatial risk distribution at the closest points of approach between ships. With this information, the conflict risks and encounter types between any target ship and the tested ship at each moment can be determined, as depicted in Fig. 15 (f)-(h). There are complex conflicts with spatial dependencies that form graph representations and involve intricate encounter types. Fig. 15(i)-(j) demonstrates the evolutions of conflict-based and encounter type-based characteristics between the tested ship and target ships over time, facilitating the observation of dynamic developments such as the reduction or increase in the number of conflicts, total

Table 2
Trajectory clustering performance under different evaluation metrics.

Evaluation metrics	Average *SS*	*DBI*	*CHI*
Test results	0.6863	0.7199	740,550
Ideal results	>0.5	<1	>(100–9999)

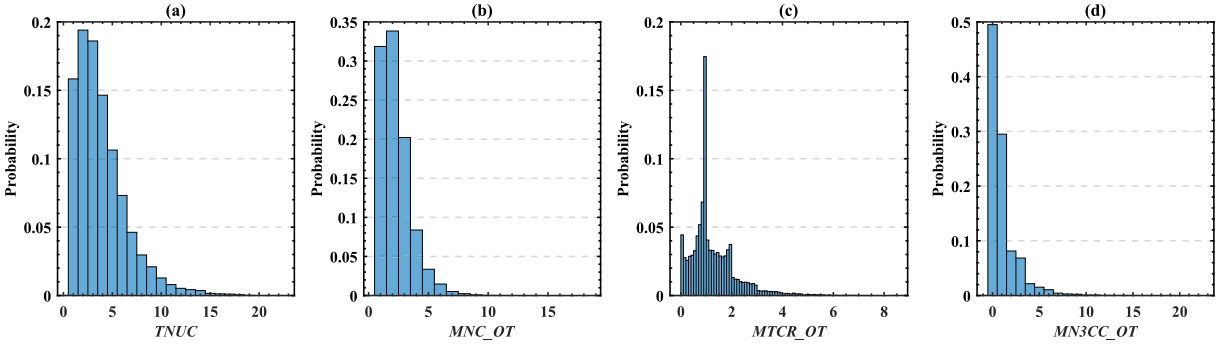

Fig. 11. Distribution of each conflict-related indicator for all extracted traffic scenarios.

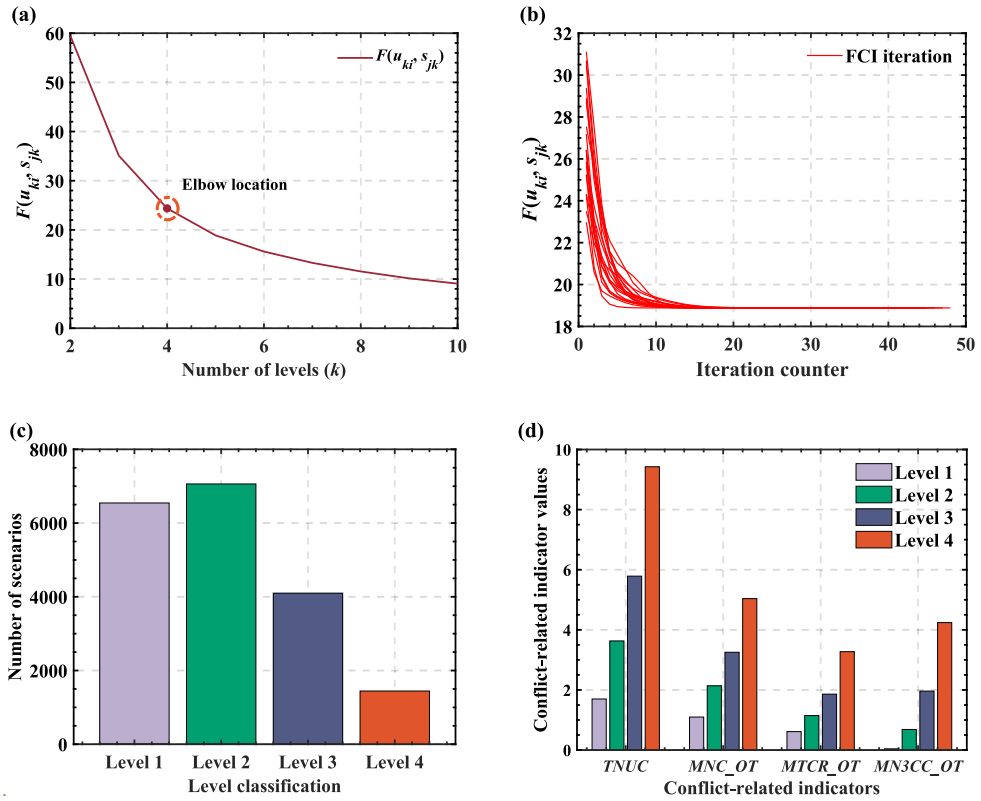

Fig. 12. FCI training results for conflict complexity levels of traffic scenarios (a) Relationships between the number of complexity levels and $F(u_{ri}, s_{jr})$; (b) optimization iterations in the FCI process using 20 randomly initialized cluster centers; (c) number of scenarios categorized by different conflict complexity levels; (d) final optimal cluster centers for each conflict-related indicator at each complexity level.

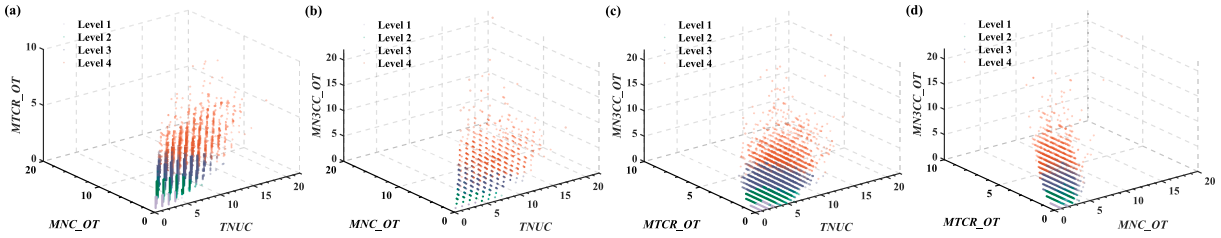

Fig. 13. Visualization of traffic complexity classification results based on any three conflict-related indicators.

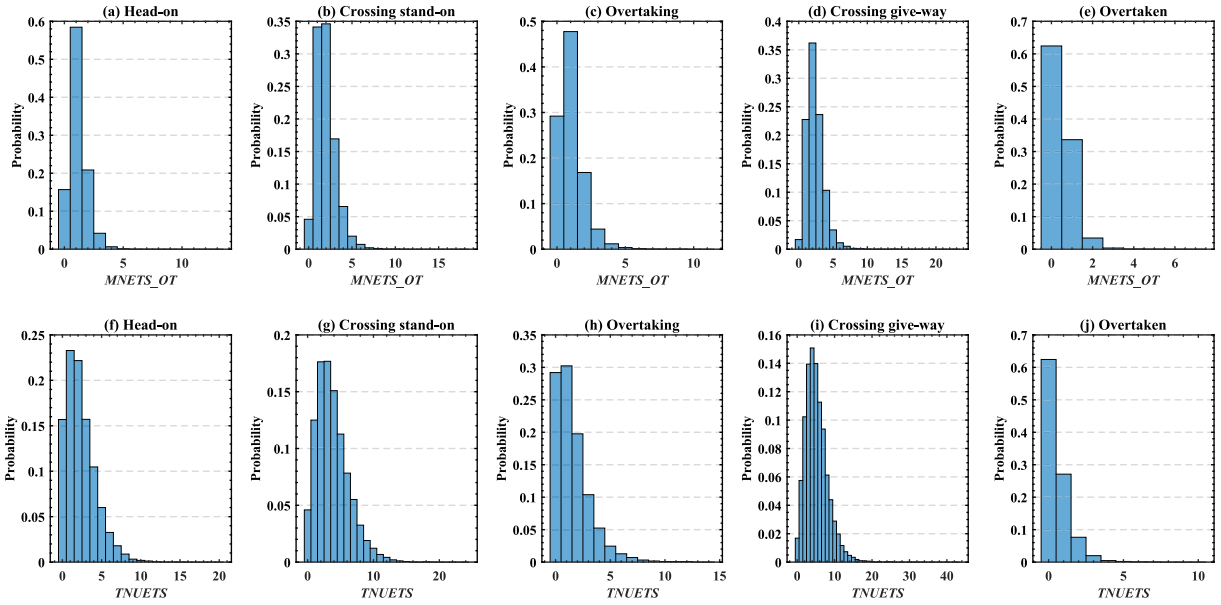

Fig. 14. Distribution of each encounter type-based indicators for all extracted traffic scenarios.

conflict risk, and the number of encountered target ships across different types involved in a scenario. The final descriptive parameters of the depicted scenario are summarized in Table 3. These characterizations aid in encoding and understanding the traffic co-behaviors in each scenario, guiding their application in subsequent autonomous testing phases.

4.4. Traffic scenario sampling results

To demonstrate the efficacy of the proposed scenario sampling framework, 1000 samples are chosen from the overall scenario set for the final testing, serving as a representative example. The proportion of scenarios selected from each motion pattern can be identified in Fig. 10(b), while the proportion from each complexity level within each motion pattern is detailed in Fig. 16. As a result, the number of samples slected from each motion pattern and conflict complexity level can be determined by Eq. (11) in Section 3.3. Based on this, the final test samples are carefully selected from each motion pattern and conflict complexity level, taking into account both the *ANND* and *CR* that are associated with the encounter type features.

For comparative evaluation, three typical sampling methods are utilized alongside the proposed greedy sampling technique. Note that Latin Hypercube Sampling (LHS) and Sobol sampling are excluded as baselines because they first generate virtual design points that must be projected back onto the dataset—a detour that introduces duplicates, enlarges the worst-case coverage radius, adds extra hyperparameters, and offers no coverage guarantee. The competing algorithms, along with key implementation details, are summarized as follows:

- **Random sampling method:** This method involves selecting a fixed number of samples randomly from the original set.
- **Max-Min distance sampling method:** This method aims to maximize the minimum distance between selected samples, thereby ensuring that the samples are as spatially dispersed as possible. It starts with a randomly selected initial sample, followed by selecting the sample with the largest minimum distance to the already selected samples from the remaining pool. This process is repeated until the necessary number of samples is achieved (Pronzato and Müller, 2012).
- **K-means++ based sampling method:** This method begins by randomly selecting the initial sample from the dataset. Subsequent samples are chosen from the remaining samples, with a probability proportional to the square of their distance from the nearest previously selected samples (Arthur and Vassilvitskii, 2006).

Fig. 17 presents a comparison of average performance across various sampling methods applied to two original subsets, each characterized by specific conflict complexity levels and motion patterns. Among these competitors, the proposed greedy sampling technique excels in achieving the best *ANND*, as depicted in Fig. 17(a) and (d). Conversely, the Max-Min distance sampling method achieves the best *CR*, demonstrated in Fig. 17(b) and (e). Notably, the proposed technique significantly outperforms both the random sampling and K-means++ based sampling methods across these evaluation metrics. While the Max-Min distance sampling method surpasses the greedy technique in terms of *CR*, it substantially sacrifices *ANND*, resulting in the weakest performance in that metric. In contrast, the proposed method adeptly integrates both *ANND* and *CR* to steer the sampling process. As illustrated in Fig. 17(c) and (f), the points generated by the greedy technique, in terms of *ANND* and *CR*, are notably closer to the lower left corner of the axis, indicating superior performance. The blue diamond points represent non-dominated solutions, whereas the red diamond points signify

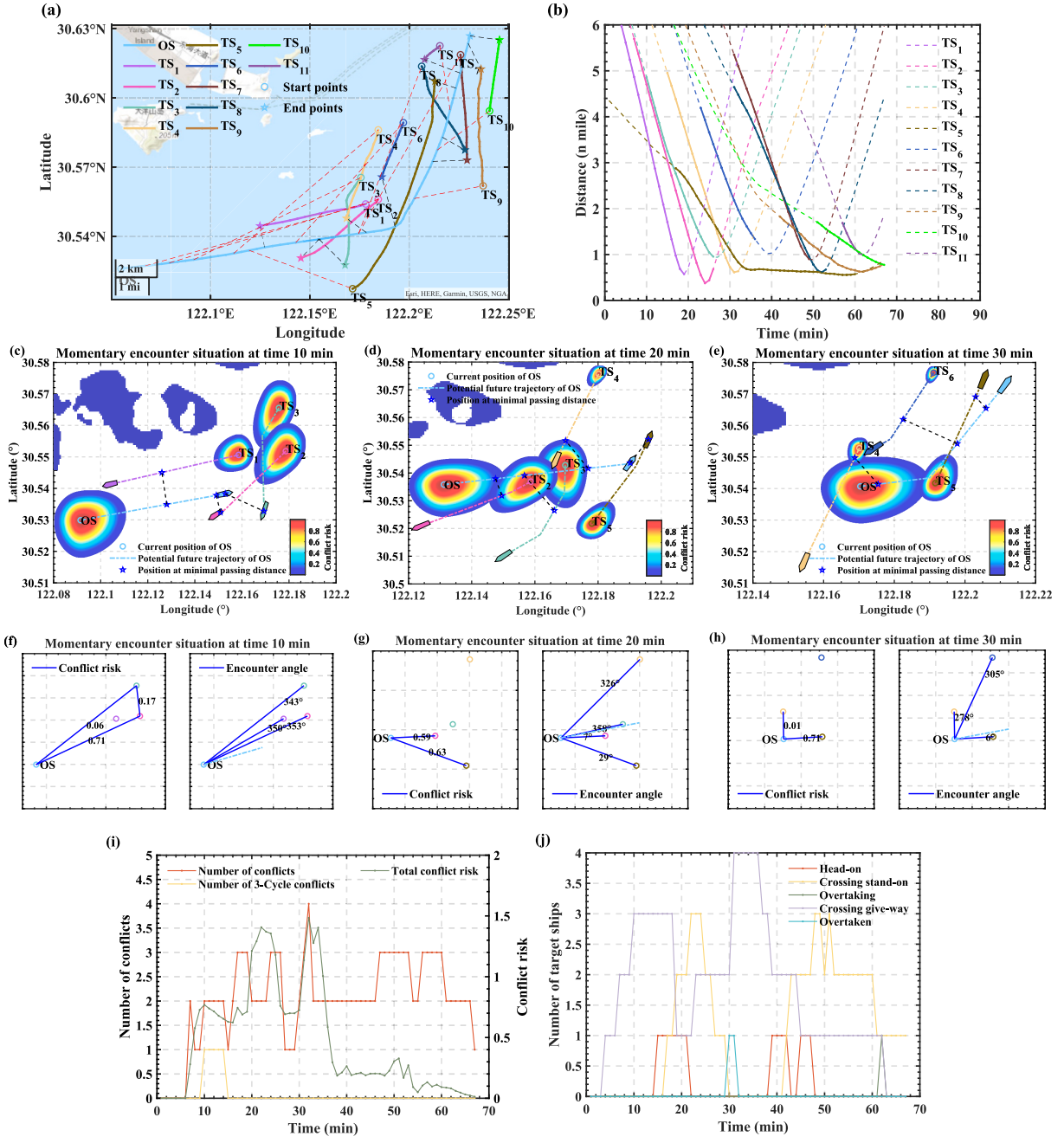

Fig. 15. Depiction of a complete traffic navigation scenario characterization. (a) Trajectories of tested ship and target ships; (b) evolution of distances between tested ship and target ships over time; (c)-(e) illustration of encounter situations at specific moments; (f)-(h) relationships between tested ship and target ships based on conflict risks and encounter types; (i) evolutions of conflict-based indicators between tested ship and target ships; (j) evolutions of encounter type-based indicators between tested ship and target ships.

Table 3
Final characterization results of the depicted scenario.

Motion pattern	Conflict complexity	Encounter type-related indicators
MP2	$\boldsymbol{TNUC_i} = 9$; $\boldsymbol{MNC\_OT_i} = 4$ $\boldsymbol{MTCR\_OT_i} = 1.485$; $\boldsymbol{MN3CC\_OT_i} = 1$ $u_{1:c,i} = [0.087, 0.187, \mathbf{0.546}, 0.180]$ Complexity level = **3**	$\boldsymbol{MNETS\_OT_i} = [1\ 3\ 1\ 4\ 1]$ $\boldsymbol{TNUETS_i} = [4\ 7\ 1\ 9\ 1]$

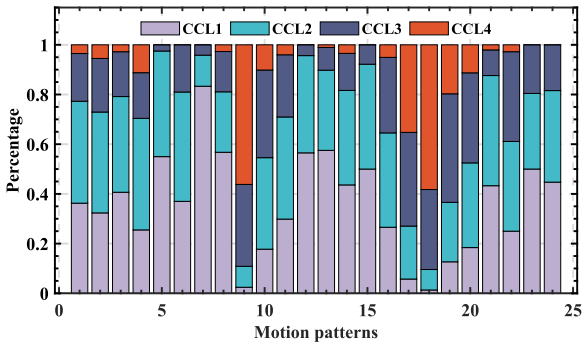

Fig. 16. Distribution of selection percentages of each conflict complexity level across different motion patterns.

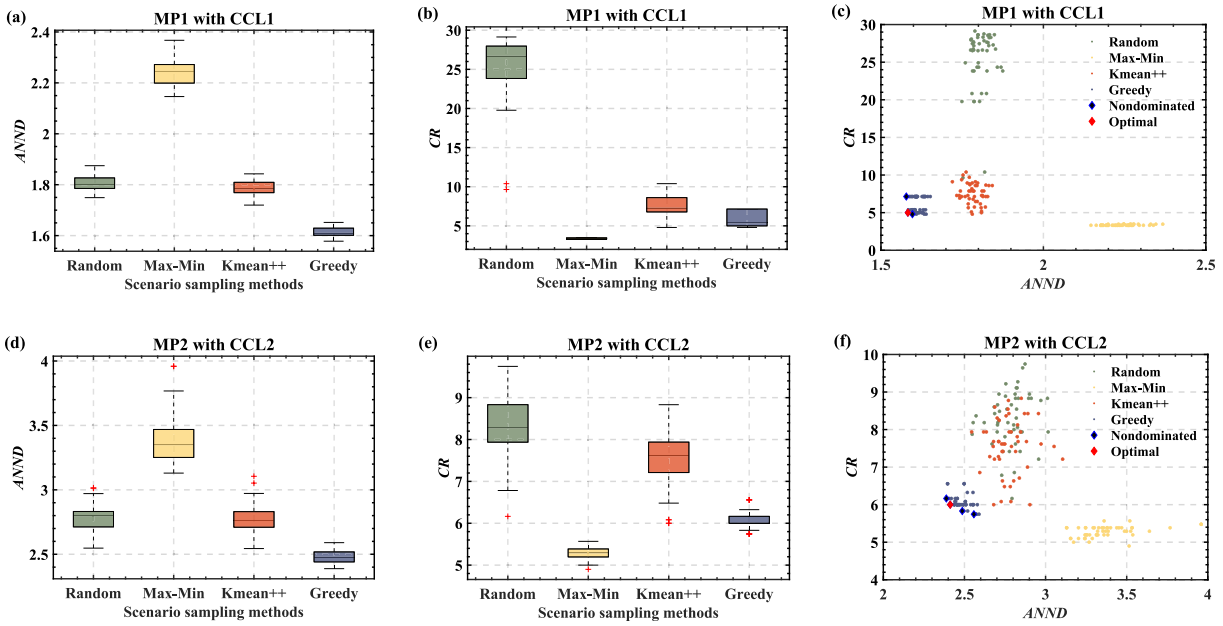

Fig. 17. Comparisons of the proposed greedy strategy with other scenario sampling methods across subsets characterized by specific motion patterns and conflict complexity levels.

the final chosen optimal solutions that achieve a strategic balance between the two metrics.

Furthermore, Fig. 18 showcases overall comparisons between the proposed greedy strategy and other scenario sampling methods across all subsets. These results reinforce those in Fig. 17, further demonstrating the proposed technique's ability to merge two coverage metrics to select representative scenarios. Consequently, the proposed technique exhibits a superior and competitive advantage in comprehensively addressing various coverage performance metrics.

4.5. Scenario real-world applications for autonomous testing

Based on extensive experiments, a comprehensive traffic navigation scenario library is developed. Each scenario in the library contains detailed traffic information from all participating ships over the scenario's duration, as illustrated in Fig. 15. Within each scenario, ships follow a sequence of waypoints from start to finish, with each waypoint recording parameters such as position, speed, and course at specific timestamps. This dataset captures how scenarios are executed by manned ships, serving as a high-fidelity bridge between historical operations and prospective simulations or sea trials for autonomous testing.

For practical testing, each scenario designates a predefined tested ship to act as a MASS, equipped with various autonomous navigation functions, while the remaining ships serve as target ships. The tested ship can operate under four distinct configurations to assess different interaction dynamics: 1) All ships follow recorded waypoints; 2) the tested ship follows waypoints with collision avoidance disabled, while target ships replay their original tracks; 3) the tested ship follows waypoints with collision avoidance enabled, while target ships replay original tracks; and 4) the tested ship autonomously plans, updates, and executes its trajectory using onboard navigation algorithms, while target ships either replay historical tracks or are driven by rule-based, behavior-cloned, or

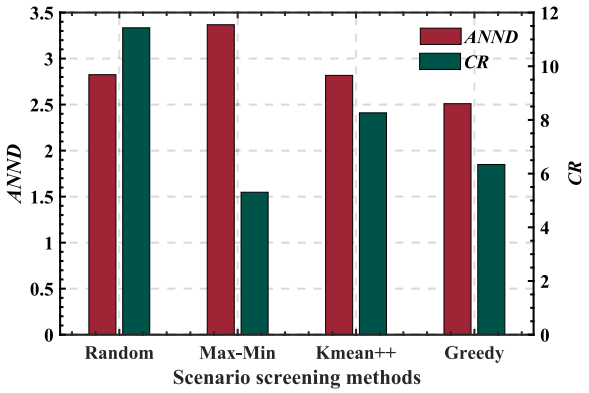

Fig. 18. Overall comparisons between the proposed greedy strategy and other scenario sampling methods across all subsets.

Reinforcement Learning (RL)/game-theoretic policies that respond dynamically to the tested ship in real time. Allowing both the tested and target ships to deviate from historical tracks is critical, as it prevents overfitting to past traffic patterns, evaluates whether autonomy can outperform traditional navigation practices, and enables stress-testing against adversarial behaviors. These setups enable the tested ship to determine and execute its course from origin to destination using different strategies. Upon completion of the tests, performance metrics, including safety, smoothness, regulatory compliance, and environmental efficiency, are applied to evaluate MASS capabilities in accordance with testing requirements.

4.6. Discussions, implications, and limitations

The methodology developed for scenario extraction, characterization, and sampling from extensive historical AIS data establishes a sophisticated traffic navigation scenario library, significantly advancing MASSs, the broader shipping industry, and maritime regulatory authorities. This study delivers multiple benefits:

Firstly, it is pivotal in enhancing the safety and efficiency of MASSs by enabling the simulation and testing of various autonomous navigation algorithms through accurate, reliable, and comprehensive traffic scenarios. Unlike manually constructed fictitious scenarios, this data-driven methodology utilizes extensive historical AIS data to accurately mirror real-world conditions and provide detailed insights into dynamic co-behaviors among ships. The methodology covers all steps from water geography preparation to hierarchical data mining, scenario complexity scoring, and summary statistics, thereby preparing the scenarios for MASS testing. Its scalability also supports the extraction of tailored scenarios across diverse geographical regions and timeframes, accounting for factors such as ship motion dynamics, ship maneuverability, and constrained waterway geography. This adaptability is crucial for comprehensive testing of autonomous ships through varied traffic interactions and dynamic multi-conflict sequences. Additionally, the extensive library of navigation scenarios simulates not only common traffic situations but also extreme and rare encounters, providing a robust testing platform for MASS navigation algorithms and ensuring comprehensive coverage of potential encounters. As such, the scenario library equips MASSs to handle a spectrum of realistic and complex conditions, enhancing system adaptability, refining decision-making algorithms, and ensuring readiness for varied traffic conditions. As autonomous maritime technologies evolve, this scenario library will become increasingly vital for system enhancements and the certification process, ensuring that MASSs can adeptly navigate the complexities of real maritime environments.

Secondly, this methodology offers valuable managerial and operational insights that are transforming maritime operations. The integration of autonomous ships is poised to revolutionize maritime logistics and operations for the shipping industry at large. The extensive traffic scenarios developed through this methodology provide reliable data that supports advanced trajectory prediction, facilitates in-depth analysis of ship maneuver behaviors, and assists in collision avoidance decision-making, thereby enhancing the predictiveness and responsiveness of autonomous systems. Traditional trajectory prediction methods often neglect the impact of multi-ship interaction on potential future dynamic behaviors, which can significantly affect the accuracy of real trajectory predictions, especially in waters with frequent multi-ship encounters. However, the extracted traffic scenarios retain detailed collision risk indicators and multi-ship interaction information over time, which can be used to train advanced AI models for more precise predictions. This supports the training and refinement of prediction models that can accurately anticipate ship dynamics and further predict potential hazards in complex traffic conditions, enhancing risk management and emergency response mechanisms for MASS systems. Moreover, detailed scenario analysis not only reveals the behavioral patterns of ships during different encounters but also evaluates the effectiveness of these maneuvers, crucial for understanding the performance of ships in various situations, particularly assessing the ability of MASSs to comply with and effectively execute maneuvers under international maritime regulations, such as the COLREGs rules. This could lead to optimized route planning, improved fuel efficiency, and reduced operational costs. The shift toward MASSs has the potential to significantly decrease human error, a primary cause of maritime accidents, thereby facilitating safer and more efficient autonomous navigation.

Lastly, the methodology informs regulatory and policy frameworks. Maritime regulatory bodies face significant challenges in framing rules and standards that adequately address the rise of autonomous technologies. The scenario library provides empirical data

that can inform and shape regulatory frameworks to accommodate MASSs. The characterization process proposes new metrics and employs a range of advanced models to facilitate the evaluation, classification, and parameterization of these scenarios, enhancing the encoding and understanding of their behaviors. By understanding the various traffic navigation scenarios that MASSs might encounter, regulators can develop more informed policies that ensure safety, compliance, and cohesiveness in various types of waters. Therefore, this comprehensive methodology not only fosters advancements in MASS technology but also serves as a cornerstone for future regulatory standards, ensuring maritime safety in an era of increasing automation.

Although the proposed methodology extracts a large number of high-risk traffic navigation scenarios based on historical AIS data, it is intrinsically limited to situations that have already occurred. Autonomous ships may face rarer or yet-unseen hazards. A promising way forward is to treat AIS data-based scenarios as a training set and apply advanced data-augmentation techniques (e.g., generative adversarial networks) to synthesize additional, physically plausible high-risk scenarios. These generated cases should then be validated against traffic statistics and expert judgment to ensure adequate coverage without sacrificing realism. Such efforts would broaden hazard coverage, strengthen the robustness and safety assurance of autonomous navigation system testing, and accelerate their regulatory acceptance and real-world deployment.

5. Conclusion

Testing and evaluation are crucial steps in the development and deployment of MASSs. This study presents a systematic methodology to extract, characterize, and sample maritime traffic navigation scenarios from extensive historical AIS data to generate a holistic testing scenario library. This methodology features several innovative aspects: 1) A novel scenario extraction approach that captures continuous spatial-temporal interactions among ships, accounting for potential ship motion dynamics, constrained waterway geography, and evolutionary graph-based interdependencies; 2) a thorough classification and parameterization of scenario characteristics, including ship motion patterns, conflict complexities, and encounter types, which enhances the encoding and understanding of traffic co-behaviors; and 3) a hierarchical greedy sampling technique that adaptively selects representative scenarios from a realistically sufficient set, ensuring comprehensive coverage of potential scenarios while optimizing the efficiency of autonomous testing. Extensive experimental analyses validate the effectiveness of this proposed methodology in producing an accurate, interpretable, and representative set of real-world traffic navigation scenarios for autonomous testing. Consequently, this study assists in establishing trust and confirms the safety, security, and reliability of MASSs by enabling these future autonomous ships to undergo thorough testing with the generated scenario library, thereby advancing highly and fully automated navigation.

Future research could explore the following promising areas. Firstly, the propagation and prediction of traffic movements based on the extracted traffic scenarios warrant further investigation. Accurate prediction depends not only on the static environment and individual information from the predicted ships but also on the interaction information from other dynamic target ships. The extracted scenarios encompass the essential data needed to train AI models aimed at enhancing the accuracy of movement predictions. This would aid in identifying real forthcoming traffic situations, crucial for issuing early collision alerts and reducing response times in risk management. Secondly, it would be highly beneficial to incorporate external environmental factors, such as wind, currents, and wave conditions, into the scenario extraction process. Doing so would enable a more holistic and realistic representation of maritime navigation scenarios, particularly in environmentally sensitive or high-risk regions. Thirdly, generative models could be employed to enrich AIS data-based scenario datasets with rigorously validated, physically plausible high-risk scenarios, thereby extending test coverage to rarer or yet-unseen hazards. Finally, while this study focuses extensively on how to extract and analyze traffic navigation scenarios, future work could valuably explore how to utilize these scenarios for autonomous testing and how to evaluate testing performance. This process would further advance the practical application of MASSs.

CRediT authorship contribution statement

Xuri Xin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kezhong Liu:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Yuerong Yu:** Writing – review & editing, Software, Investigation, Data curation. **Zaili Yang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by a European Research Council project under the European Union's Horizon 2020 research and innovation program (TRUST CoG 2019 864724) and the National Natural Science Foundation of China (Grant No. 52031009).

Appendix A. Procedure of DTW calculation

The core principle of DTW is to identify the optimal match between two sequences and to calculate the cumulative distance of the

optimal alignment path between them. Assuming the trajectories of ships i and j are represented as $Tra_i = \{P_{1,i}, P_{2,i}, \ldots, P_{t,i}, \ldots, P_{n,i}\}$ and $Tra_j = \{P_{1,j}, P_{2,j}, \ldots, P_{t,j}, \ldots, P_{n,j}\}$, with each point $P_{t,i}$ specified as $\{t_i, lat_{t,i}, lon_{t,i}, COG_{t,i}, V_{t,i}\}$. Then, a distance matrix $A = (a_{pq})_{n,i\times n,j}$ can be constructed, where a_{pq} represents the Euclidean distance between the pth point in Tra_i and the qth point in Tra_j. From this distance matrix, a warping path $WP = \{wp_1, wp_2, \ldots, wp_t, \ldots, wp_l\}$ can be identified, comprising a sequence of adjacent elements in A. The length of the warping path, l, must satisfy $\max\{n,i, n,j\} < l \le n,i + n,j - 1$. Additionally, the warping path adheres to the following constraints:

- **Boundary constraint:** The path starts at $wp_1 = a_{1,1}$ and ends at $wp_l = a_{n,i\times n,j}$;
- **Monotonicity:** For $wp_{t\text{-}1} = a_{p'q'}$, and $wp_t = a_{pq}$, it requires $q - q' \ge 0$ and $p - p' \ge 0$;
- **Continuity:** For $wp_{t\text{-}1} = a_{p'q'}$ and $wp_t = a_{pq}$, it requires $q - q' \le 1$ and $p - p' \le 1$.

Various warping paths can be generated based on the aforementioned constraints, and DTW identifies the path with the lowest warping distance to effectively measure the distance between two sequences, as described below:

$$DTW(Tra_i, Tra_j) = \min\left\{\frac{1}{l}\sum_{t=1}^{l} w_t\right\} \tag{A.1}$$

To achieve Eq. (A.1), the minimal DTW distance is iteratively searched using the following equations:

$$D(1,1) = a_{11} \tag{A.2}$$

$$D(p,q) = a_{pq} + \min\{D(p,q-1), D(p-1,q-1), D(p-1,q)\} \tag{A.3}$$

where $D(p, q)$ calculates the cumulative distance of the optimal alignment path between the two sequences Tra_i and Tra_j, starting from the beginning and up to points $P_{p,i}$ and $P_{q,j}$.

Appendix B. Detailed description of OPTICS

The fundamental principle of the OPTICS algorithm revolves around the construction of a reachability plot. This structure depends on two pivotal distance concepts: core distance and reachability distance, both shaped by two key parameters: $\in$, defining the maximum distance for consideration, and *MinPts*, indicating the minimum number of points required to establish a cluster. A point p is considered a core point if there are at least *MinPts* points within $\in$ distance from it. The core distance of point p is the distance from p to its $MinPts^{th}$ nearest neighbor, defined as follows:

$$Core\_Distance(p, \in, MinPts) = \begin{cases} MinPts^{th}\text{smallest distance to Neighbors} & |N_\in(p)| \ge MinPts \\ \infty & \text{otherwise} \end{cases} \tag{B.1}$$

where $|N_\in(p)|$ represents the number of points within the $\in$-neighborhood of point p. The reachability distance of point o relative to point p is determined as the greater of p's core distance or the actual distance from o to p, as follows:

$$Reachability\_Distance(o, p, \in, MinPts) = \begin{cases} \max(Core\_Distance(p, \in, MinPts), Distance(o,p)) & p \text{ is a core point} \\ \infty & \text{otherwise} \end{cases} \tag{B.2}$$

where $Distance(o, p)$ represents the actual distance between points o and p. By employing these definitions, the OPTICS algorithm generates a cluster-ordering including ordered points along with the reachability distance of each point. The reachability plot of these ordered points facilitates visualization of the dataset's clustering structure. More detailed pseudo-code for the OPTICS algorithm procedure is provided in previous work (Rong et al., 2022).

Appendix C. Procedure of FCI implementation

FCI adopts an iterative procedure to continuously refine and determine the optimal clustering center matrix and membership matrix. Consider a dataset with n traffic navigation scenarios, each characterized by m assessment indicators, formatted as $X = (x_{ij})_{n\times m}$, where x_{ij} denotes the jth indicator of the ith scenario. To ensure scale invariance across various indices, a normalization process transforms dataset X into a normalized matrix, as follows:

$$\widehat{x}_{ij} = \left.\frac{x_{ij} - x_{min,j}}{x_{max,j} - x_{min,j}}\right|_{j=1:m} \tag{C.1}$$

where $\widehat{x}_{ij}$ represents the normalized value of x_{ij}, with $x_{min,j}$ and $x_{max,j}$ being the minimal and maximum values of the jth indicators within dataset X , respectively.

Following this, the class center matrix is defined as $S = (s_{jk})_{m\times c}$ and the fuzzy membership matrix as $U = (u_{ki})_{c\times n}$, where c represents the number of classes grouping n scenarios, s_{jk} indicates the class center of the jth indicator in the kth class, and u_{ki} reflects the relative membership of scenario i to the kth class. To derive the optimal S and U, an objective function is formulated to minimize the quadratic sum of the Euclidean distances between the assessment indicators from all scenarios and their respective fuzzy class centers, as follows:

$$\min\left[F\left(u_{ki}, s_{jk}\right)\right] = \min\left\{\sum_{i=1}^{n}\sum_{k=1}^{c}\left(u_{ki}{}^{2}\sum_{j=1}^{m}\left(\widehat{x}_{ij}-s_{jk}\right)^{2}\right)\right\} \tag{C.2}$$

where s_{jk} and u_{ki} adhere to the constraints $0 \le s_{jk} \le 1$, $0 \le u_{ki} \le 1$, and $\sum_{k=1}^{c} u_{ki} = 1$.

To meet the goal set in Eq. (C.2), an iterative method using the Lagrange multiplier method is employed, described by the following equations:

$$u_{ki} = \left[\sum_{h=1}^{c}\frac{\sum_{j=1}^{m}\left(\widehat{x}_{ij}-s_{jk}\right)^{2}}{\sum_{j=1}^{m}\left(\widehat{x}_{ij}-s_{jh}\right)^{2}}\right]^{-1} \tag{C.3}$$

$$s_{jk} = \sum_{i=1}^{n} u_{ki}{}^{2}\widehat{x}_{ij} \Big/ \sum_{i=1}^{n} u_{ki}{}^{2} \tag{C.4}$$

The iterative adjustments for u_{ki} and s_{jk} persist through Eqs. (C.3)-(C.4) until the termination criteria are fulfilled. Detailed guidelines for FCI implementation are documented in prior research (He et al., 2011).

Appendix D. Detailed illustration of clustering performance evaluation metrics

The clustering performance indicators, including Silhouette Score (*SS*), Davies-Bouldin Index *(DBI*), and Calinski-Harabasz Index (*CHI*), are introduced below.

The *SS* quantifies the quality of clustering by comparing the proximity of points within a cluster to their separation from the nearest other cluster, calculated as follows:

$$SS_i = \frac{a_i - b_i}{\max\{a_i, b_i\}} \tag{D.1}$$

where a_i represents the average distance from point i to all other points within the same cluster, and b_i denotes the average distance from point i to points in the closest different cluster. The score varies from -1 to $+1$, where values nearer to $+1$ represent more robust clustering. Generally, an average *SS* above 0.5 suggests satisfactory clustering.

The *DBI* evaluates clustering by measuring the compactness of clusters and their separation from each other. It is defined as:

$$DBI = \frac{1}{nc}\sum\nolimits_{k=1}^{nc} max_{v\neq y}\left\{\frac{\sigma_v + \sigma_y}{d\left(cc_v, cc_y\right)}\right\} \tag{D.2}$$

where nc is the number of clusters. σ_v is the average distance of all points in cluster v to their cluster center cc_v, and $d(cc_v, cc_y)$ represents the distance between cluster centers cc_v and cc_y. A lower *DBI* value, ideally below 1.0, indicates better clustering by showing that clusters are both well-separated and internally cohesive.

Finally, the *CHI* measures clustering effectiveness based on cluster tightness and separation. It uses the following formula:

$$CHI = \frac{B(nc)/(nc-1)}{W(nc)/(n-nc)} \tag{D.3}$$

where $B(nc)$ represents the total between-cluster dispersion, $W(nc)$ represents the total within-cluster dispersion, and n is the total number of samples. High values of the *CHI* indicate that clusters are both distinct from each other and internally compact. Typically, a high value, in the hundreds or thousands, signals a well-structured clustering.

Data availability

Data will be made available on request.

References

Ahmed, Y.A., Hannan, M.A., Oraby, M.Y., Maimun, A., 2021. COLREGs compliant fuzzy-based collision avoidance system for multiple ship encounters. J. Mar. Sci. Eng. 9, 790.

Akdağ, M., Solnør, P., Johansen, T.A., 2022. Collaborative collision avoidance for maritime autonomous surface ships: a review. Ocean Eng. 250, 110920.

Arthur, D., Vassilvitskii, S., 2006. k-means++: the advantages of careful seeding. Stanford.

Bakdi, A., Glad, I.K., Vanem, E., 2021. Testbed scenario design exploiting traffic big data for autonomous ship trials under multiple conflicts with collision/grounding risks and spatio-temporal dependencies. IEEE Trans. Intell. Transp. Syst. 22, 7914–7930.

Berndt, D.J., Clifford, J., 1994. Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, pp. 359–370.

Bolbot, V., Gkerekos, C., Theotokatos, G., Boulougouris, E., 2022. Automatic traffic scenarios generation for autonomous ships collision avoidance system testing. Ocean Eng. 254, 111309. https://doi.org/10.1016/j.oceaneng.2022.111309.

Cao, Y., Xin, X., Jarumaneeroj, P., Li, H., Feng, Y., Wang, J., Wang, X., Pyne, R., Yang, Z., 2025. Data-driven resilience analysis of the global container shipping network against two cascading failures. Transp. Res. Part E Logist. Transp. Rev. 193, 103857.
Cassará, P., Di Summa, M., Gotta, A., Martelli, M., 2023. E-Navigation: a distributed decision support system with extended reality for bridge and ashore seafarers. IEEE Trans. Intell. Transp. Syst.
Chang, C.-H., Wijeratne, I.B., Kontovas, C., Yang, Z., 2024. COLREG and MASS: analytical review to identify research trends and gaps in the development of autonomous collision avoidance. Ocean Eng. 302, 117652.
Chen, X., Liu, Y., Achuthan, K., Zhang, X., Chen, J., 2021. A semi-supervised deep learning model for ship encounter situation classification. Ocean Eng. 239, 109824.
Cho, Y., Han, J., Kim, J., 2020. Efficient COLREG-compliant collision avoidance in multi-ship encounter situations. IEEE Trans. Intell. Transp. Syst. 23, 1899–1910.
Du, L., Goerlandt, F., Banda, O.A.V., Huang, Y., Wen, Y., Kujala, P., 2020. Improving stand-on ship's situational awareness by estimating the intention of the give-way ship. Ocean Eng. 201, 107110.
Feng, S., Feng, Y., Yu, C., Zhang, Y., Liu, H.X., 2020. Testing scenario library generation for connected and automated vehicles, part I: methodology. IEEE Trans. Intell. Transp. Syst. 22, 1573–1582.
Gil, M., Krata, P., Kozioł, P., Hinz, T., 2024a. A multiparameter simulation-driven analysis of ship turning trajectory concerning a required number of irregular wave realizations. Ocean Eng. 299, 117293.
Gil, M., Montewka, J., Krata, P., 2024b. Predicting a passenger ship's response during evasive maneuvers using Bayesian Learning. Reliab. Eng. Syst. Saf.
Gil, M., Poczęta, K., Wróbel, K., Yang, Z., Chen, P., 2025. Toward using fuzzy grey cognitive maps in manned and autonomous collision avoidance at sea. IEEE J. Ocean Eng.
Gu, Y., Wallace, S.W., 2021. Operational benefits of autonomous vessels in logistics—a case of autonomous water-taxis in Bergen. Transp. Res. Part E Logist. Transp. Rev. 154, 102456.
He, Y., Zhou, J., Kou, P., Lu, N., Zou, Q., 2011. A fuzzy clustering iterative model using chaotic differential evolution algorithm for evaluating flood disaster. Expert Syst. Appl. 38, 10060–10065.
Hernandez-Romero, E., Valenzuela, A., Rivas, D., 2019. A probabilistic approach to measure aircraft conflict severity considering wind forecast uncertainty. Aerosp. Sci. Technol. 86, 401–414.
Hwang, T., Youn, I.-H., 2021. Navigation situation clustering model of human-operated ships for maritime autonomous surface ship collision avoidance tests. J. Mar. Sci. Eng. 9. https://doi.org/10.3390/jmse9121458.
Jia, C., Ma, J., de Vries, B., Kouw, W.M., 2024. Bayesian inference of collision avoidance intent during ship encounters. IEEE Trans. Autom. Sci. Eng.
Johansen, T.A., Perez, T., Cristofaro, A., 2016. Ship collision avoidance and COLREGS compliance using simulation-based control behavior selection with predictive hazard assessment. IEEE Trans. Intell. Transp. Syst. 17, 3407–3422. https://doi.org/10.1109/TITS.2016.2551780.
Li, H., Lam, J.S.L., Yang, Z., Liu, J., Liu, R.W., Liang, M., Li, Y., 2022. Unsupervised hierarchical methodology of maritime traffic pattern extraction for knowledge discovery. Transp. Res. Part C Emerg. Technol. 143, 103856.
Liu, C., Kulkarni, K., Suominen, M., Kujala, P., Musharraf, M., 2024a. On the data-driven investigation of factors affecting the need for icebreaker assistance in ice-covered waters. Cold Reg. Sci. Technol. 221, 104173.
Liu, C., Musharraf, M., Li, F., Kujala, P., 2022a. A data mining method for automatic identification and analysis of icebreaker assistance operation in ice-covered waters. Ocean Eng. 266, 112914.
Liu, J., Zhang, J., Yan, X., Soares, C.G., 2022b. Multi-ship collision avoidance decision-making and coordination mechanism in mixed Navigation Scenarios. Ocean Eng. 257, 111666.
Liu, J., Zhang, J., Yang, Z., Wan, C., Zhang, M., 2024b. A novel data-driven method of ship collision risk evolution evaluation during real encounter situations. Reliab. Eng. Syst. Saf. 249, 110228.
Liu, K., Yuan, Z., Xin, X., Zhang, J., Wang, W., 2021. Conflict detection method based on dynamic ship domain model for visualization of collision risk Hot-spots. Ocean Eng. 242, 110143.
Lotovskyi, E., Rong, H., Teixeira, A.P., 2024. Collision risk assessment in ship encounter scenarios using AIS trajectory data, in: Advances in Maritime Technology and Engineering. CRC Press, pp. 135–142.
Lu, H., Zhang, Y., Zhang, C., Niu, Y., Wang, Z., Zhang, H., 2025. A multi-sensor fusion approach for maritime autonomous surface ships berthing navigation perception. Ocean Eng. 316, 119965.
Lu, S., Shang, Y., Li, Y., 2017. A research on the application of fuzzy iteration clustering in the water conservancy project. J. Clean. Prod. 151, 356–360.
Ma, Y., Jiang, W., Zhang, L., Chen, J., Wang, H., Lv, C., Wang, X., Xiong, L., 2024. Evolving testing scenario generation and intelligence evaluation for automated vehicles. Transp. Res. Part C Emerg. Technol. 163, 104620.
Niu, Y., Zhu, F., Zhai, P., 2023. An autonomous decision-making algorithm for ship collision avoidance based on DDQN with prioritized experience replay. In: 2023 7th International Conference on Transportation Information and Safety (ICTIS). IEEE, pp. 1174–1180.
Pedersen, T.A., Glomsrud, J.A., Ruud, E.L., Simonsen, A., Sandrib, J., Eriksen, B.O.H., 2020. Towards simulation-based verification of autonomous navigation systems. Saf. Sci. 129, 104799. https://doi.org/10.1016/j.ssci.2020.104799.
Pronzato, L., Müller, W.G., 2012. Design of computer experiments: space filling and beyond. Stat. Comput. 22, 681–701.
Ribeiro, C.V., Paes, A., de Oliveira, D., 2023. AIS-based maritime anomaly traffic detection: a review. Expert Syst. Appl. 231, 120561.
Ripley, B.D., 1988. Statistical inference for spatial processes. Cambridge University Press.
Rong, H., Teixeira, A.P., Soares, C.G., 2024. A framework for ship abnormal behaviour detection and classification using AIS data. Reliab. Eng. Syst. Saf. 247, 110105.
Rong, H., Teixeira, A.P., Soares, C.G., 2022. Maritime traffic probabilistic prediction based on ship motion pattern extraction. Reliab. Eng. Syst. Saf. 217, 108061.
Sawada, R., Sato, K., Minami, M., 2024. Framework of safety evaluation and scenarios for automatic collision avoidance algorithm. Ocean Eng. 300, 117506.
Shi, J., Wang, S., Yuan, Q., Liu, X., Hsieh, T.-H., 2023. Generation of naturalistic and adversarial sailing environment (NASE) for intelligent test of autonomous ships. Ocean Eng. 285, 115438.
Silveira, P., Teixeira, A.P., Soares, C.G., 2022. A method to extract the Quaternion Ship Domain parameters from AIS data. Ocean Eng. 257, 111568.
Szlapczynski, R., Szlapczynska, J., 2017. Review of ship safety domains: models and applications. Ocean Eng. 145C, 277–289.
Thombre, S., Zhao, Z., Ramm-Schmidt, H., García, J.M.V., Malkamäki, T., Nikolskiy, S., Hammarberg, T., Nuortie, H., Bhuiyan, M.Z.H., Särkkä, S., 2020. Sensors and ai techniques for situational awareness in autonomous ships: a review. IEEE Trans. Intell. Transp. Syst.
Torben, T.R., Glomsrud, J.A., Pedersen, T.A., Utne, I.B., Sorensen, A.J., 2023. Automatic simulation-based testing of autonomous ships using Gaussian processes and temporal logic. Proc. Inst. Mech. Eng. PART O-JOURNAL RISK Reliab. 237, 293–313. https://doi.org/10.1177/1748006X211069277.
Wang, C., Zhang, X., Gao, H., Bashir, M., Li, H., Yang, Z., 2024a. COLERGs-constrained safe reinforcement learning for realising MASS's risk-informed collision avoidance decision making. Knowledge-Based Syst.
Wang, N., 2010. An intelligent spatial collision risk based on the quaternion ship domain. J. Navig. 63, 733–749.
Wang, S., Zhang, Y., Zhang, X., Gao, Z., 2023. A novel maritime autonomous navigation decision-making system: Modeling, integration, and real ship trial. Expert Syst. Appl. 222, 119825.
Wang, W., Huang, L., Liu, K., Zhou, Y., Yuan, Z., Xin, X., Wu, X., 2024b. Ship encounter scenario generation for collision avoidance algorithm testing based on AIS data. Ocean Eng. 291, 116436.
Wang, W., Liu, K., Huang, L., Xin, X., Wu, X., Yuan, Z., 2024c. Generation and complexity analysis of ship encounter scenarios using AIS data for collision avoidance algorithm testing. Ocean Eng. 312, 119034.
Wróbel, K., Gil, M., Huang, Y., Wawruch, R., 2022. The vagueness of COLREG versus collision avoidance techniques—A discussion on the current state and future challenges concerning the operation of autonomous ships. Sustainability 14, 16516.
Xin, X., Cao, Y., Jarumaneeroj, P., Yang, Z., 2025. Vulnerability assessment of international container shipping networks under national-level restriction policies. Transp. Policy.

Xin, X., Liu, K., Li, H., Yang, Z., 2024. Maritime traffic partitioning: an adaptive semi-supervised spectral regularization approach for leveraging multi-graph evolutionary traffic interactions. Transp. Res. Part C Emerg. Technol. 164, 104670.
Xin, X., Liu, K., Loughney, S., Wang, J., Li, H., Ekere, N., Yang, Z., 2023. Multi-Scale Collision Risk Estimation for Maritime Traffic in complex Port Waters. Reliab. Eng. Syst Saf.
Yu, H., Meng, Q., Fang, Z., Liu, J., Xu, L., 2023. A review of ship collision risk assessment, hotspot detection and path planning for maritime traffic control in restricted waters. J. Navig. 1–27.
Yu, H., Murray, A.T., Fang, Z., Liu, J., Peng, G., Solgi, M., Zhang, W., 2021. Ship path optimization that accounts for geographical traffic characteristics to increase maritime port safety. IEEE Trans. Intell. Transp. Syst. 23, 5765–5776.
Yu, Y., Liu, K., Kong, W., Xin, X., 2025. Time-evolving graph-based approach for multi-ship encounter analysis: Insights into ship behavior across different scenario complexity levels. Transp. Res. Part A Policy Pract. 194, 104427.
Zarzycki, F., Gil, M., Montewka, J., Szłapczyński, R., Szłapczyńska, J., 2025. Declarative ship arenas under favourable conditions. Ocean Eng. 316, 119927.
Zhang, J., Liu, J., Hirdaris, S., Zhang, M., Tian, W., 2023. An interpretable knowledge-based decision support method for ship collision avoidance using AIS data. Reliab. Eng. Syst. Saf. 230, 108919. https://doi.org/10.1016/j.ress.2022.108919.
Zhang, J., Zhang, D., Yan, X., Haugen, S., Soares, C.G., 2015. A distributed anti-collision decision support formulation in multi-ship encounter situations under COLREGs. Ocean Eng. 105, 336–348.
Zhang, M., Montewka, J., Manderbacka, T., Kujala, P., Hirdaris, S., 2021. A big data analytics method for the evaluation of ship-ship collision risk reflecting hydrometeorological conditions. Reliab. Eng. Syst. Saf. 213, 107674.
Zhang, M., Taimuri, G., Zhang, J., Zhang, D., Yan, X., Kujala, P., Hirdaris, S., 2024. Systems driven intelligent decision support methods for ship collision and grounding prevention: present status, possible solutions, and challenges. Reliab. Eng. Syst. Saf.
Zhou, H., Zheng, M., Chu, X., Liu, C., Zhong, C., 2024. Scenario modeling method for collision avoidance testing in inland waterway. Ocean Eng. 298, 117192.
Zhou, Z., Zhang, Y., 2023. A system for the validation of collision avoidance algorithm performance of autonomous ships. Ocean Eng. 280, 114600.
Zhu, F., Niu, Y., Wei, M., Du, Y., Zhai, P., 2025. A high-risk test scenario adaptive generation algorithm for ship autonomous collision avoidance decision-making based on Reinforcement Learning. Ocean Eng. 320, 120344.
Zhu, F., Zhou, Z., Lu, H., 2022. Randomly testing an autonomous collision avoidance system with real-world ship encounter scenario from AIS data. J. Mar. Sci. Eng. 10. https://doi.org/10.3390/jmse10111588.