IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 September 2025, accepted 26 September 2025,
date of publication 14 October 2025, date of current version 23 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3621188

RESEARCH ARTICLE

Reinforcement Learning-Assisted Secure Reliable Underwater Wireless Acoustic Communications

ABDALLAH S. GHAZY[1], GEORGES KADDOUM[1], (Senior Member, IEEE), CHAMESEDDINE TALHI[1], NAVEED IQBAL[2,3], (Senior Member, IEEE), AND ALI HUSSEIN MUQAIBEL[2,3], (Senior Member, IEEE)

[1]École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada
[2]Electrical Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia
[3]Interdisciplinary Research Center for Communication Systems and Sensing (IRC-CSS), King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia

Corresponding author: Abdallah S. Ghazy (Abdallah.Ghazy@ejust.edu.eg)

This research was supported by the Mitacs grant IT32533 and the Innovation for Defense Excellence and Security (IDEaS) program of the Department of National Defence (DND). The work done by Naveed Iqbal and Ali Hussein Muqaibel is supported by King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, through the Interdisciplinary Research Center for Communication Systems and Sensing under Grant No. INCS2503.

ABSTRACT In recent days, there has been an increasing demand for the deployment of autonomous underwater vehicles (AUVs) for tactical wireless acoustic communications. This requires secure and reliable AUV communications to protect sensitive data. However, existing methods such as cryptography and channel coding introduce extra overheads and computational complexity. This is primarily due to the inherent challenges posed by acoustic communication systems, such as limited bandwidth and low energy efficiency. To overcome these challenges, we propose using intelligent reflecting surfaces (IRSs) in conjunction with reinforcement learning (RL) techniques, resulting in what is termed as RL-assisted Buoyed-IRS-AUV (RL-BIA) links. The RL-BIA links facilitate simultaneous secure and reliable communications by dynamically adjusting its beam width and IRS's depth in response to seawater turbulence induced by wind and tide. We introduce a comprehensive link model that accounts for pointing errors, path loss, interference, and noise. Additionally, we developed an RL model adaptable to BIA links. To integrate channel secrecy and outage probability, a non-convex Max-Min optimization problem is formulated and solved iteratively using Q-learning and State-Action-Reward-State-Action (SARSA) algorithms. Numerical results demonstrate that at a wind speed of 8.5 meters per second, the proposed approach significantly enhances channel secrecy, with the RL-BIA link achieving a remarkable 400% improvement compared to the RL-assisted buoyed-AUV (RL-BA) link.

INDEX TERMS Underwater wireless acoustic communications, autonomous underwater vehicles, intelligent reflecting surfaces, reinforcement learning, channel secrecy.

I. INTRODUCTION

In recent years, the demand for smart autonomous underwater vehicles (AUVs) has been steadily increasing across various warfare domains, including self-defense, border surveillance, and reconnaissance [1], [2]. These smart AUVs often rely on underwater wireless acoustic communications (UWACs) to network, a vulnerability that could be exploited by adversarial AUVs [1], [2]. The need for secure and reliable communications has become more critical in response to heightened security threats and concerns [3], [4], [5]. However, achieving simultaneous security and reliability in underwater wireless communication is challenging due to dynamic conditions induced by wind and sea current turbulence. In addition, the inherent broadcast nature of underwater channels, significant signal attenuation, limited bandwidth, interference, and noise further complicate the task. These factors create an unpredictable environment

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Masini.

where maintaining robust communication while ensuring security requires sophisticated strategies and optimization techniques [2], [6], [7], [8], [9], [10].

Researchers have proposed various methods to achieve secure and reliable underwater wireless communications such as cryptography [11], watermarking [12], authentication [13], and key management [14]. Conversely, reliable communication methods encompass error control coding techniques [15], [16], adaptive modulation [17], [18], multiple-input multiple-output (MIMO) techniques [19], diversity techniques [20], interference mitigation techniques, like adaptive filtering and cancellation, and minimizing interference effects [21]. Nonetheless, these methods often introduce significant overhead and demand high computational complexity, leading to increased power consumption. This poses critical challenges for UWAC systems, which are constrained by limited bandwidth and power resources [7].

Rather than relying solely on traditional security methods, underwater physical layer security (UPLS) offers an alternative approach by leveraging the unique characteristics of the underwater environment to enhance security [3], [4], [5]. Researchers have proposed various approaches to demonstrate the effectiveness of UPLS techniques. Stojanovic et al. employed spread spectrum methods, such as direct sequence and frequency hopping, to reduce the probability of interception in underwater environments [22]. Zhou et al. introduced a spectral scrambling scheme that renders interception and deciphering of information challenging for potential eavesdroppers [23]. Wang and Wang proposed a technique involving signal overlapping at eavesdroppers while ensuring collision-free reception at the authorized user [24].

Directional communication plays a crucial role in enhancing of the UPLS. For example, Zhao demonstrated the effectiveness of directional communications using an acoustic parametric matrix, which significantly reduces the likelihood of adversarial signal detection [25]. In contrast, omni-directional routing protocols, such as Depth-Based Routing [26], inherently broadcast signals in all directions, making transmitted information more vulnerable to eavesdropping and traffic analysis attacks. Conversely, directional routing protocols, such as Vector-Based Forwarding [27], have enhanced communication confidentiality. They achieve this by confining the signal within a narrow beam directed toward the intended destination, thereby minimizing the chances of interception by unintended or malicious nodes. As well, Romdhane and Kaddoum [28] proposed a directional routing scheme, thereby improving link security in underwater scenarios. The foundational understanding of underwater acoustic multipath channel behavior, as presented in [6], justifies the adoption of adaptive and direction-aware routing strategies.

Directional communications can be enhanced using Intelligent Reflecting Surfaces (IRS) technology. Recently, researchers have leveraged this technology to improve physical layer security (PLS) [29], [30], [31], [32], [33]. IRSs are artificial constructs designed to improve wireless communication by reshaping and redirecting beams with programmable flexibility, including adjustments in width, orientation, and intensity. By leveraging their passive or active surfaces and precisely controlling reflection angles and beam patterns, IRSs optimize signal propagation by focusing energy toward legitimate users while suppressing or nullifying signal strength at eavesdroppers. For instance, Liu et al. [29] proposed an IRS-aided scheme to enhance physical layer security (PLS) in NOMA networks by jamming eavesdroppers with artificial noise, optimizing power allocation and IRS configurations for improved secrecy rates. Jin et al. [30] explored improper Gaussian signaling to mitigate artificial noise's impact on legitimate users, achieving enhanced secrecy by optimizing RIS reflection coefficients and transmit covariance matrices. Additionally, Sarawar et al. [31] analyzed the secrecy of RISs integrated to underwater optical wireless communication, deriving closed-form secrecy metrics and offering deployment guidelines for secure communications in diverse environments.

When IRSs are deployed in dynamic and unpredictable underwater environments, reinforcement learning (RL) technology could provide a means to navigate uncertainty and non-stationarity. RL enables IRSs to make real-time decisions by continuously interacting with the underwater environment through trial and error. Numerous studies in the literature have proposed RL-assisted strategies to address underwater challenges [28], [34], [35], [36], [37]. Romdhane and Kaddoum [28] proposed RL-assisted techniques to optimize the beamwidth and orientation, thereby improving link quality and communication success rates across various underwater scenarios. Huang and Wang [36] presented an approach utilizing RL to determine optimal emergency response modes for isolated underwater sensor nodes. However, to the best of our knowledge, the integration of RL technology with IRS architectures has not yet been proposed to enhance the security of underwater wireless communications.

Table 1 compares recent underwater acoustic communication studies involving IRS, RL, or both. Several works employed RL-based optimization without IRS, such as [36], which enhanced network connectivity using RL with multi-objective optimization and adaptive clustering. These methods lacked IRS-based physical-layer control to counteract multipath fading or signal degradation. Conversely, studies like [38], [39], and [40] explored IRS-assisted schemes to mitigate fading, extend range, and enhance BER but relied on static or heuristic designs without learning-based adaptability. The authors in [41] and [42] proposed IRS-based turbulence and hybrid surface models to improve reliability and efficiency, and [43] incorporated higher-order mode optical beams with IRS to boost link quality. However, these approaches assumed quasi-static environments and lacked real-time adaptability. Ghazy et al. [33] introduced a buoyed-IRS-AUV (BIA) link, optimizing beamforming

TABLE 1. Summary and comparison of relevant works in the literature on underwater communications.

Reference	Year	The Aims of the Paper	The Novelty in the System Model	Is IRS Included ?	Is the Machine Learning Included?
[38]	2021	Increase Bit Rate and Mitigate Fading	Designing a Hardware for IRS	✓	✗
[39]	2022	Increase Bit Rate and Extend Communication Range	IRS Operation Protocol	✓	✗
[40]	2022	Improve BER and Link Reliability	Blockage and Pointing Error Models	✓	✗
[28]	2022	Maintain Link Alignment	RL-Assisted Uncertainty Node Positions	✗	✓
[36]	2023	Network Connectivity	Integrating RL, Multi-Objective Optimization, and Adaptive Clustering	✗	✓
[41]	2023	Enhance Link Reliability	OTOPS-Based Turbulence Model	✓	✗
[42]	2023	Improve Bit Error Rate, Spectral and Energy Efficiencies, and Link Reliability.	Joined IRS and Planar Mirror Surface Model	✓	✗
[43]	2024	Improve Link Reliability	Higher-Order Mode Optical Beam	✓	✗
[33]	2024	Enhance Concurrently Channel Secrecy and Link Reliability	Sea waves and Tide Speed-Based Analytic Optimization-Framework	✓	✗
Our work	2025	Enhance Concurrently Channel Secrecy and Link Reliability	Sea waves and Tide Speed-Based Numerical Optimization-Framework	✓	✓

and IRS depth based on sea dynamics. Their two-phase algorithm achieved superior performance over benchmarks, though with increased computational cost.

This paper proposes an RL-assisted IRS approach to address the limitations of the two-phase algorithm in [33]. While RL-assisted IRS has been explored in terrestrial systems, such as [44], underwater acoustic environments introduce unique challenges, such as wind-induced waves and tide-induced currents, that are difficult to model. Unlike the offline-online approach in [33], our RL-assisted buoyed-IRS-to-AUV (RL-BIA) link adapts in real time without prior environmental knowledge, efficiently handling high-dimensional, non-convex optimization. The RL-BIA model features a tailored multi-dimensional state-action-reward structure, enabling robust beamforming and IRS depth control under uncertain underwater conditions. This proposed model represents a significant advancement over prior terrestrial and underwater studies, offering a dynamic and environment-aware solution for secure, reliable underwater communications. The key contributions of this paper are outlined below:

- To the best of our knowledge, this is the first instance of integrating RL technology with IRS architecture to establish secure and reliable wireless communications in a dynamic and unpredictable underwater environment without relying on prior information or estimates.
- Development of a comprehensive RL model that accounts for fluctuations in wind and sea current speeds influencing BIA links, featuring multidimensional sets of states, actions, and rewards.
- The channel secrecy rate (CSR) and outage probability are integrated into a non-convex optimization problem. Two iterative algorithms, Q-learning and State-Action-Reward-State-Action (SARSA), are proposed to solve this problem in real time.
- We compare the CSR performance of RL-BIA links using Q-learning, SARSA, and exhaustive search algorithms. Additionally, we analyze the impact of RL hyperparameters on RL-BIA links. Finally, the CSR performance of RL-BIA links is contrasted with that of RL-BA links.

The remainder of this paper is organized as follows: Section II introduces the framework of the paper. Section III presents the channel and parameters of the RL-BIA link. Section V introduces the RL-assisted system model, formulates the channel secrecy in a Max-Min optimization problem, and presents the Q-learning and SARSA algorithms. Section VI conducts a numerical evaluation of RL metrics and channel secrecy rate for the BIA links, comparing them with the BA links results. Finally, Section VII summarizes our findings and suggests potential avenues for future research.

Notation: In this article, vectors are denoted by boldface letters, e.g., $\boldsymbol{X}$ and $\boldsymbol{x}$. Probability and cumulative distribution functions of x are denoted by $f(x)$ and $F(x)$, respectively. The Normal, Cox-Munk, Uniform, and Rayleigh probability distribution functions (PDFs) are denoted as $\boldsymbol{N}$, $\boldsymbol{CM}$, $\boldsymbol{U}$, and $\boldsymbol{R}$, respectively.

II. PAPER FRAMEWORK

In this section, we present the link topologies and the overall framework of the paper. Figure 1(a) illustrates the RL-assisted Buoyed-IRS-AUV (RL-BIA) and RL-assisted Buoyed-AUV (RL-BA) link topologies, shown in green and yellow, respectively, while Figure 1(b) depicts the flow of the proposed framework.

A. TOPOLOGY MODEL

Figure 1 (a) showcases the RL-BA and RL-BIA links depicted in yellow and green, respectively. The illustration portrays a downlink communication scenario involving a transmitter (Tx) buoyed on the seawater surface and a legitimate receiver (Rx) positioned in the seawater. An IRS, mounted on a hovering drone, is strategically positioned between the Tx and Rx to facilitate communications. Nearby,

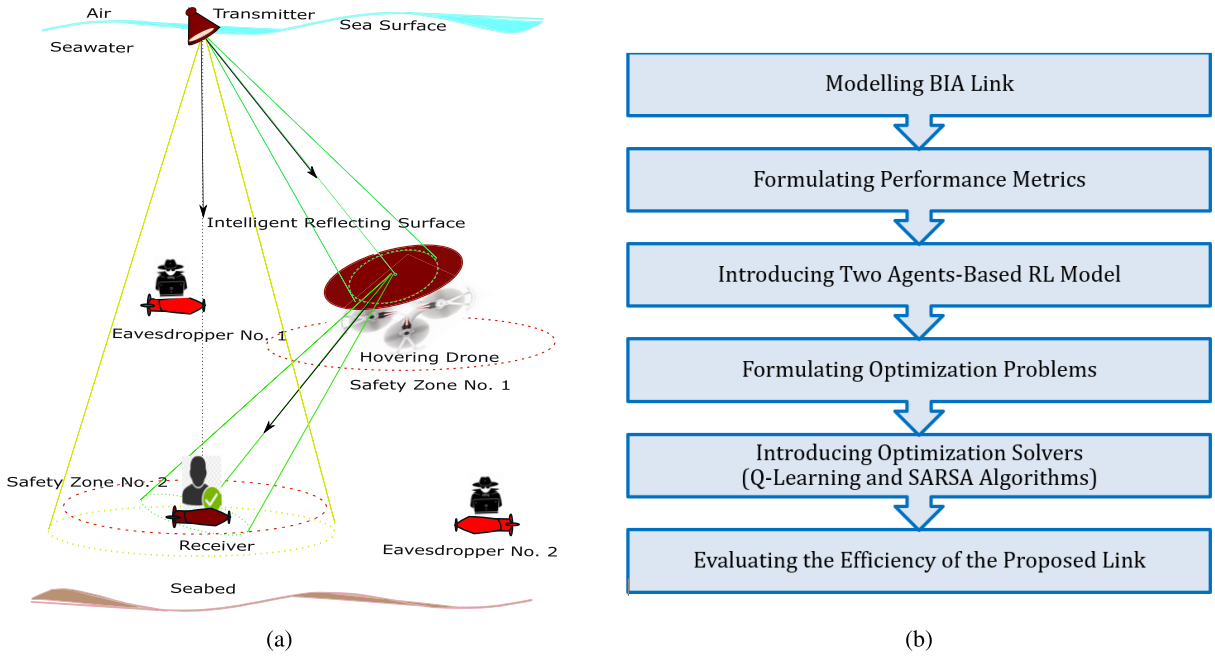


FIGURE 1. Subfigure (a) highlights the RL-BIA and RL-BA link topologies in green and yellow, respectively, while subfigure (b) illustrates the flow of the proposed framework.

eavesdroppers 1 and 2 (Ex1 and Ex2) hover close to the Rx and IRS, respectively, attempting to intercept communications. In this dynamic and challenging environment, the Tx experiences displacement and disorientation due to wind-driven sea waves, while the IRS, Ex1, Rx, and Ex2 are affected by displacement and disorientation caused by tide-induced sea currents [45], [46]. Additionally, vibrations from the propulsion systems further impact system stability.

The reference RL-BA link comprises three essential components: the Tx, Rx, and Ex2. The simple link faces a critical trade-off between security and reliability. Broadening the acoustic beamwidth enhances communication reliability while compromising security. Conversely, selecting a narrower beamwidth enhances security but diminishes communication reliability.

The proposed RL-BIA link encompasses five key components: the Tx, IRS, Ex1, Rx, and Ex2, enabling concurrent secure and reliable communications through the strategic deployment of a sizable IRS between the Tx and Rx. This placement leverages the IRS substantial aperture and short inter-component distances to mitigate geometric losses effectively. Moreover, shorter inter-component distances necessitate narrower beamwidths for the Tx-IRS and IRS-Rx links. Additionally, the link adapts its configuration to varying channel conditions to further enhance secrecy and reliability. It adjusts beamwidths to precisely focus coverage on the IRS and Rx, preventing potential information leakage to Ex1 and Ex2. The link also dynamically adjusts the IRS depth to optimize its proximity to the most dynamically varying component, i.e., the Tx or Rx. This combination of features empowers RL-BIA links to deliver secure and reliable communications, even amidst swinging movements in seawaters. However, it's crucial to note that, unlike the RL-BA link, the IRS component in the RL-BIA link presents a potential hotspot susceptible to eavesdropping. To ensure comprehensive security for RL-BIA links, eavesdropping threats at both the IRS and Rx locations are addressed simultaneously through a Max-Min optimization problem.

B. RL-MODEL

In this paper, we propose using an RL model to minimize the information leakages to the eavesdroppers, i.e., Ex1 and Ex2, concurrently, and enhance the link reliability between the Tx and legitimate users, i.e, IRS and Rx. The proposed RL model makes the Tx and IRS learn optimal adjustments for the beamwidth and depth in real time, through trial and error by interacting with the link fluctuations caused by wind-induced waves and tide-induced currents.

The RL model operates in two modes: learning and tracking. In learning mode, the Tx and IRS (i.e., agents) explore different settings and evaluate their rewards based on received feedback. If the feedback is positive, the agents recognize it as a good setting and consider it for future use; otherwise, they discard it. Since the eavesdroppers (Ex1 and Ex2) do not send feedback, a worst-case scenario estimation is applied. For instance, assuming Ex1 and Ex2 are positioned at horizontal distances equal to the radii of safety zones 1 and 2 from the IRS and Rx, respectively, the Tx and IRS compute the channel secrecy accordingly. Safety zones are physical areas around the transmitter where eavesdroppers are excluded. The agents balance exploration and exploitation over hundreds of iterations. Eventually, they

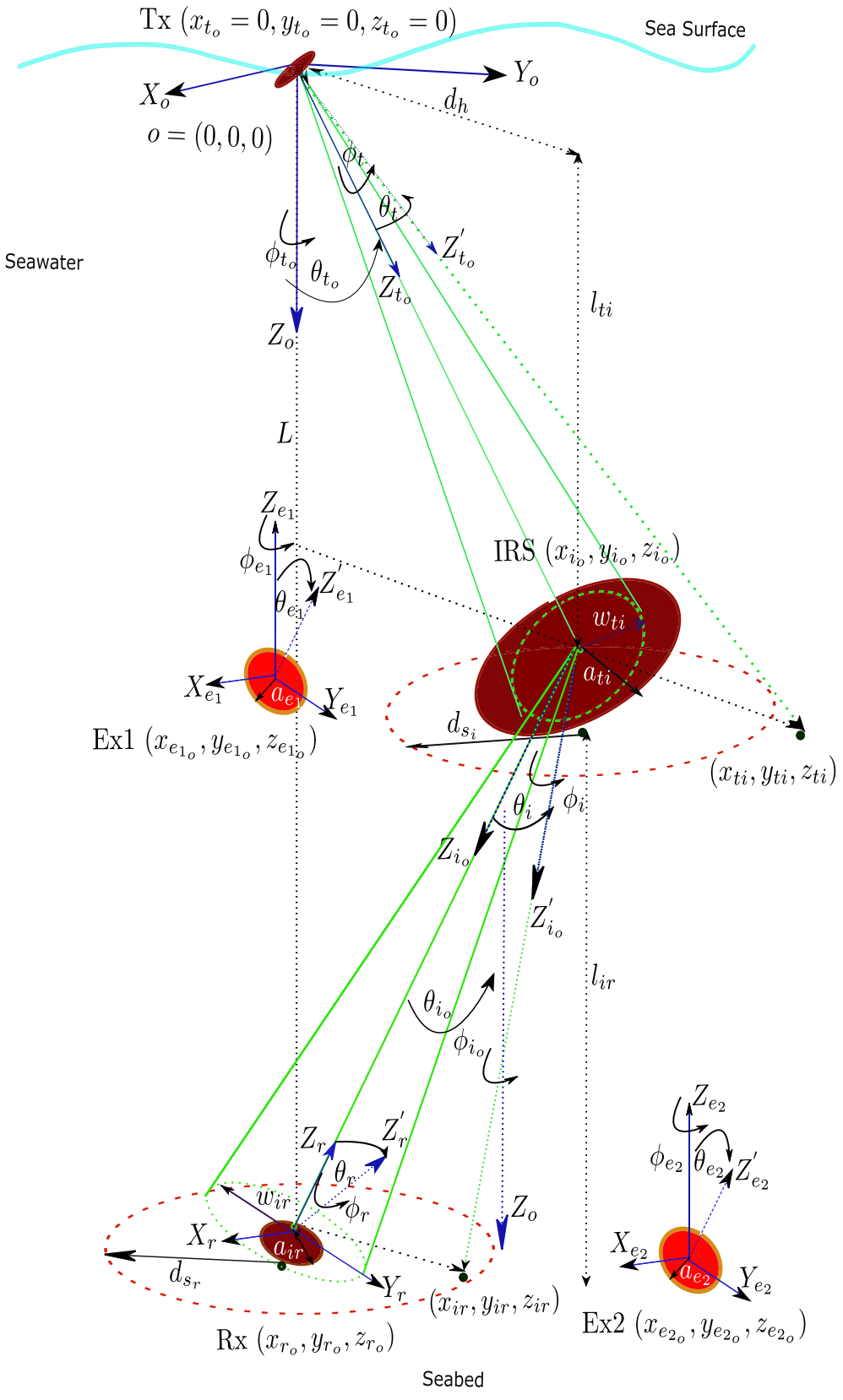


FIGURE 2. Model of RL-BIA links: The model highlights important parameters of the system configuration, link coordination, and orientation.

transition from random to optimized/sub-optimized settings and link performance.

In tracking mode, the agents apply their learned policies to autonomously adapt to environmental variations, maintaining optimal channel secrecy and reliability. The RL model retrains only when significant environmental changes occur, as indicated by discrepancies between the current and optimal rewards. Major environmental changes require extensive training to reconverge, whereas minor changes may not necessitate retraining.

C. FRAMEWORK FLOW

Figure 1(b) illustrates the flow of the framework, outlining a step-by-step development of the proposed RL-BIA system. The paper begins by modeling the BIA underwater communications link, accounting for the dynamic nature of sea-surface IRS nodes and mobile AUV receivers. This is followed by the formulation of key performance metrics, i.e., secrecy rate and outage probability, which serve as the basis for evaluating system efficiency. The next stage introduces a two-agent RL model, where both the Tx and the IRS operate as autonomous agents that learn optimal strategies over time. Subsequently, these learning tasks are formulated as non-convex optimization problems that reflect both performance objectives and operational constraints. To solve these problems efficiently, we integrate model-free RL solvers. Specifically, Q-learning and SARSA algorithms are well-suited for real-time, resource-constrained underwater environments. The paper concludes with a thorough evaluation of the proposed link's performance using simulation results, demonstrating its superiority over baseline models under realistic marine conditions. This structure ensures a comprehensive and logically coherent development of the RL-BIA framework from concept to validation.

III. LINK MODEL

This section outlines the link model through three subsections. System Architecture describes the overall link model and its components. Steady-State Link Coordination and Orientation explains how the link maintains alignment under normal conditions, i.e., quiet sea, while Disruptive-State Link Coordination and Orientation examines its response to disturbances caused by wind-induced sea waves and sea currents.

A. LINK ARCHITECTURE

Figure 2 depicts a model of the RL-BIA link, taking into account the system configuration and the displacement and disorientation between its components. The Tx emits an acoustic Gaussian beam toward the IRS with a variable width w_{ti} and a transmitted acoustic power P_{ti}. The IRS, consisting of a large array of microphones and hydrophones with a radius a_{ti}, utilizes an acoustic-to-electric coefficient η_{i_i} and an electric-to-acoustic coefficient η_{i_o} [47], [48]. Upon receiving the incident acoustic beam, the IRS converts it into an electrical signal with a ratio of η_{i_i} and then emits a new acoustic beam towards the Rx with a variable width w_{ir} and a transmitted acoustic power P_{ir}, where $P_{ir} = P_{ti}\,\eta_{i_i}\,\eta_{i_o}$. The Rx, Ex1, and Ex2 utilize hydrophones with radii a_{ir}, a_{e_1}, and a_{e_2}, respectively, which are relatively small compared to the IRS, i.e., $a_{ti} \gg a_{ir}$. Both the Tx and the IRS are equipped with RL chips to dynamically adjust the beam widths, orientations, and depth based on the wind and sea current speeds.

B. LINK COORDINATION AND ORIENTATION IN STEADY STATE

Figure 2 illustrates the link axes, denoted as (X_o, Y_o, Z_o), with the origin positioned at $O = (0, 0, 0)$. The beam axes of the Tx and IRS are labeled as Z_{to} and Z_{io}, respectively. The aperture axes of the Rx, Ex1, and Ex2 are represented as (X_r, Y_r, Z_r), $(X_{e_1}, Y_{e_1}, Z_{e_1})$, and $(X_{e_2}, Y_{e_2}, Z_{e_2})$, respectively. The total link range is given by $L = l_{ti} + l_{ir}$, where l_{ti} and l_{ir} denote the vertical inter-distances between the Tx and IRS, and between the IRS and Rx, respectively. The IRS adjusts its depth, l_{ti}, based on the wind and sea current speeds. The horizontal inter-distances between the Tx and IRS, and between the IRS and Rx, are equal and are denoted as d_h.

Relative to the origin O, the nominal location of the Tx is situated at $(x_{t_o} = 0, y_{t_o} = 0, z_{t_o} = 0)$, while the IRS is positioned at $(x_{i_o} \neq 0, y_{i_o} \neq 0, z_{i_o} = l_{ti})$. The Rx is hovering at $(x_{r_o} = 0, y_{r_o} = 0, z_{r_o} = L)$, and the nominal locations of Ex1 and Ex2 are specified as $(x_{e_{1_o}}, y_{e_{1_o}}, z_{e_{1_o}} = l_{ti})$ and $(x_{e_{2_o}}, y_{e_{2_o}}, z_{e_{2_o}} = L)$, respectively.

To mitigate eavesdropping threats, dashed-red circles in the figure depict safety zones centered at the nominal locations of the IRS and Rx, with radii denoted as d_{s_i} and d_{s_r}, respectively. Consequently, the positions of Ex1 and Ex2 are restricted by these safety zones, and their distances from the nominal locations of the IRS and Rx must satisfy the specified conditions: $\sqrt{(x_{e_{1_o}} - x_{i_o})^2 + (y_{e_{1_o}} - y_{i_o})^2} > d_{s_i}$, $\sqrt{(x_{e_{2_o}} - x_{r_o})^2 + (y_{e_{2_o}} - y_{r_o})^2} > d_{s_r}$.

Figure 2 illustrates the nominal orientations of the link components, aligned to minimize geometric losses. The beam axes of the Tx and IRS are perpendicular to the centers of the IRS and Rx apertures, respectively. Relative to the (X_o, Y_o, Z_o) axes, the nominal polar and azimuthal orientations of the beam axes of the Tx and IRS are $(\theta_{t_o} = \arctan(d_{ti}/l_{ti}),\ \phi_{t_o} = \pi/4)$ and $\left(\theta_{i_o} = \arctan(d_{ir}/l_{ir}),\ \phi_{i_o} = \pi/4\right)$, respectively. The Tx and IRS adjust these nominal angles according to the wind and sea current speeds. The nominal polar and azimuthal orientations of the Rx are $(\theta_{r_o} = \arctan(d_{ir}/l_{ir}),\ \phi_{r_o} = \pi/4)$, while those of Ex1 and Ex2 are aligned with the original axes, i.e., $(\theta_{e_{1_o}} = 0,\ \phi_{e_{1_o}} = 0)$ and $(\theta_{e_{2_o}} = 0,\ \phi_{e_{2_o}} = 0)$, respectively.

C. LINK COORDINATION AND ORIENTATION UNDER DISRUPTIONS

The proposed model for link orientation disruptions is supported by a diverse set of authoritative references that span theory, empirical studies, experiments, and real-world data. Theoretical analyses [7], [28], [49], establish the theoretical

TABLE 2. Definitions for parameters illustrated in Figure 2 and mentioned in Sections III and IV.

Parameter and Symbol	Definition
X_o, Y_o, Z_o	Nominal axes of the link.
$X_i, Y_i, Z_i, X_r, Y_r, Z_r$	Nominal axes of the IRS and Rx apertures.
$X_{e_1}, Y_{e_1}, Z_{e_1}, X_{e_2}, Y_{e_2}, Z_{e_2}$	Nominal axes of the Ex1 and EX2 apertures.
$Z_{t_o}^{'}, Z_{i_o}^{'}, Z_{e_1}^{'}, Z_r^{'}, Z_{e_2}^{'}$	Perturbed beam and aperture axes of Tx, IRS, Ex1, Rx, and Ex2.
l_{ti}, l_{ir}, L	Vertical inter-distances for Tx-IRS, IRS-Rx, and Tx-IRS-Rx links.
z_d, z_w	Seawater depth and depth at which the wind-driven current speed becomes zero.
d_h	Horizontal inter-distance between the Tx and IRS, and between IRS and Rx.
d_{s_i}, d_{s_r}	Radii of safety zones around the IRS and Rx.
$(x_{t_o}, y_{t_o}, z_{t_o}), (x_{i_o}, y_{i_o}, z_{i_o})$	Nominal position of the Tx and IRS.
$(x_{r_o}, y_{r_o}, z_{r_o})$	Nominal position of the Rx.
$(x_{e_{1_o}}, y_{e_{1_o}}, z_{e_{1_o}}), (x_{e_{2_o}}, y_{e_{2_o}}, z_{e_{2_o}})$	Nominal position of Ex1 and Ex2.
$(x_t, y_t, z_t), (x_i, y_i, z_i)$	Perturbed position of the Tx and IRS in X, Y, and Z axes.
(x_r, y_r, z_r)	Perturbed position of the Rx in X, Y, and Z axes.
$(x_{e_1}, y_{e_1}, z_{e_1}), (x_{e_2}, y_{e_2}, z_{e_2})$	Perturbed position of E1 and Ex2 in X, Y and Z axes.
$(\sigma_{x_t}, \sigma_{y_t}, \sigma_{z_t}), (\sigma_{x_i}, \sigma_{y_i}, \sigma_{z_i})$	Standard deviations of displacement for Tx and IRS along X_o, Y_o, and Z_o axes.
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$	Vectors list (x, y, z) coordinates of Tx, IRS, Ex1, Rx, and Ex2.
$\boldsymbol{N}(\boldsymbol{x_o}, \boldsymbol{\sigma_x}), \boldsymbol{N}(\boldsymbol{y_o}, \boldsymbol{\sigma_y})$	Normal PDFs of Tx and IRS position relative to its nominal location.
$\boldsymbol{N}(\boldsymbol{z_o}, \boldsymbol{\sigma_z})$	Normal distribution of Rx position relative to its nominal location.
$\boldsymbol{x_o}, \boldsymbol{y_o}, \boldsymbol{z_o}, \boldsymbol{\sigma_x}, \boldsymbol{\sigma_y}, \boldsymbol{\sigma_z}$	Vectors listing parameters of Normal PDFs for Tx, IRS, Ex1, Rx, and Ex2.
$(\theta_{t_o}, \phi_{t_o}), (\theta_{i_o}, \phi_{i_o}), (\theta_{r_o}, \phi_{r_o})$	Nominal orientations of optical and aperture axes of Tx, IRS and Rx.
$(\theta_{e_{1_o}}, \phi_{e_{1_o}}), (\theta_{e_{2_o}}, \phi_{e_{2_o}})$	Nominal Polar and azimuthal orientations of Ex1 and Ex2.
$(\theta_t, \phi_t), (\theta_i, \phi_i), (\theta_r, \phi_r)$	Perturbed orientations of optical and aperture axes of Tx, IRS and Rx.
$(\theta_{e_1}, \phi_{e_1}), (\theta_{e_2}, \phi_{e_2})$	Perturbed Polar and azimuthal orientations of Ex1 and Ex2.
$\sigma_{\theta_t}, \sigma_{\theta_i}$	Standard deviations of polar disorientations of the Tx and IRS.
$\boldsymbol{\theta}, \boldsymbol{\phi}$	Vectors list angular disorientations (θ, ϕ) of Tx, IRS, Ex1, Rx, and Ex2.
$\boldsymbol{CM}(\boldsymbol{\sigma_\theta}), \boldsymbol{U}(\boldsymbol{a_\phi}, \boldsymbol{b_\phi})$	Cox-Munk and Uniform PDFs for polar and azimuthal disorientations.
$\boldsymbol{\sigma_\theta}, \boldsymbol{a_\phi}, \boldsymbol{b_\phi}$	Vectors listing parameters of Cox-Munk and Uniform PDFs for link components.
$a_{ti}, a_{ir}, a_{e_1}, a_{e_2}$	Radii of the IRS, Rx, Ex1, and Ex2 hydrophones.
$\eta_{i_i}, \eta_{i_o}, \eta_{r_i}$	Acoustic-to-electric and electric-to-acoustic coefficients for IRS and Rx.
μ	Spreading beam factor.
w_{ti}, w_{ir}	Beam widths received by the IRS and Rx.
V_w, V_c	Wind speed at the sea surface and Sea current speed at IRS depth.
V_v, V_t	Speeds of the wind-driven and tide-driven currents.
β_i	Orientation imprecision of the hovering system (OIHS) attached to the IRS.
χ	Subscript parameter, where $\chi := ti$ and $\chi := ir$ for the Tx-IRS and IRS-Rx links.
r_χ	Radial distance between the beam's arrival position and the aperture center.
$A_{l_\chi}(l_\chi, f)$	Specific path loss for distance l_χ tih frequency f for the χ^{th} link.
l_χ, v_χ, A_χ	length, Beam-to-aperture ratio, and Pointing error gain for the χ^{th} link.
h_{ti}, h_{tir}	Channel DC gains for Tx-IRS and Tx-IRS-Rx links.
h_{l_χ}, h_{p_χ}	Path loss and Pointing error for the χ^{th} link.
P_{ti}, P_{ir}	Transmitted acoustic power by the Tx and IRS.
$P_\chi, P_\chi(f)$	Transmitted power and frequency spectrum of χ^{th} link.
f, f_l, f_u	Frequency carrier, lower, and upper frequency spectrum limits in kilohertz.
IN	Total power of the wind interference and thermal noise.

basis, while Operational Oceanographic Products and Services data [43] ensure realistic environmental parameters. Empirical validations from textbooks [45], [50], laboratory experiments replicating our set-up [51], and the modeling of disturbance of buoyed nodes [52] further strengthen the model. Additional theoretical and observational support is drawn from [45], [50], [53], and [54]. Collectively, these sources provide a robust foundation for modeling in realistic marine conditions.

Due to sea waves and currents, the nominal positions of the Tx, IRS, Ex1, Rx, and Ex2 are perturbed, resulting in new positions denoted as (x_t, y_t, z_t), (x_i, y_i, z_i), $(x_{e_1}, y_{e_1}, z_{e_1})$, (x_r, y_r, z_r), and $(x_{e_2}, y_{e_2}, z_{e_2})$, respectively. These positions can be listed in $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ vectors as follows: $\boldsymbol{X} = [x_t, x_i, x_{e_1}, x_r, x_{e_2}]$, $\boldsymbol{Y} = [y_t, y_i, y_{e_1}, y_r, y_{e_2}]$, $\boldsymbol{Z} = [z_t, z_i, z_{e_1}, z_r, z_{e_2}]$, where the $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ vectors represent the displacements relative to the elements of the X_o, Y_o, and Z_o axes, respectively. The elements listed in these vectors are independent random variables, described statistically using Normal probability density functions (PDFs) with means equal to the nominal locations, and standard deviations related to the strengths of the waves and sea currents [10], [49], [51], [55]. These vectors are described using Normal PDFs as: $\boldsymbol{X} \sim \boldsymbol{N}(\boldsymbol{x_o}, \boldsymbol{\sigma_x})$, $\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{y_o}, \boldsymbol{\sigma_y})$, $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{z_o}, \boldsymbol{\sigma_z})$, where the elements listed in $\boldsymbol{x_o}, \boldsymbol{y_o}$, and $\boldsymbol{z_o}$ represent the means, while the elements listed in $\boldsymbol{\sigma_x}$, $\boldsymbol{\sigma_y}$, and $\boldsymbol{\sigma_z}$ denote the standard deviations of the $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ vectors, respectively.

Practically, buoyed and hovering components displace with the same standard deviations in the X_o and Y_o axes, i.e., $\sigma_{x_t} = \sigma_{y_t}$ and $\sigma_{x_i} = \sigma_{y_i}$. Additionally, the standard deviations in the X_o and Y_o axes are significantly larger than the displacements in the Z_o axis, such that $\sigma_{x_t} \gg \sigma_{z_t}$ and $\sigma_{x_i} \gg \sigma_{z_i}$ [7], [49]. Quantifying these deviations as functions of the wind and sea current speeds has not been explored in the existing literature.

Due to sea waves and currents, the nominal orientations of the beam axes of the Tx and IRS, as well as the aperture axis of the Rx, are altered to new orientations represented by the axes Z'_{t_o}, Z'_{i_o}, and Z'_r, respectively. These axes are defined by angles (θ_t, ϕ_t), (θ_i, ϕ_i), and (θ_r, ϕ_r), measured with respect to the Z_{t_o}, Z_{i_o}, and Z_r axes, respectively. Meanwhile, the disorientations of the beam axes of the Ex1 and Ex2 are defined by $(\theta_{e_1}, \phi_{e_1})$, and $(\theta_{e_2}, \phi_{e_2})$, respectively, and they are measured with respect to the Z_o axis.

The polar and azimuthal disorientations are listed in $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ vectors, respectively, and given as: $\boldsymbol{\theta} = [\theta_t, \theta_i, \theta_{e_1}, \theta_r, \theta_{e_2}]$, $\boldsymbol{\phi} = [\phi_t, \phi_i, \phi_{e_1}, \phi_r, \phi_{e_2}]$. The angles listed in these vectors are independent random variables, and are described statistically using the Cox-Munk and Uniform probability density functions (PDFs), with means and standard deviations related to the speed of the sea waves and currents [51], [52]. The polar and azimuthal vectors are described using Cox-Munk and Uniform PDFs, respectively, as: $\boldsymbol{\theta} \sim \boldsymbol{CM}(\boldsymbol{\sigma_\theta})$, $\boldsymbol{\phi} \sim \boldsymbol{U}(\boldsymbol{a_\phi}, \boldsymbol{b_\phi})$, where the $\boldsymbol{\sigma_\theta}$ vector lists the parameters of the Cox-Munk PDFs for the Tx, IRS, Ex1, Rx, and Ex2 in order. Additionally, the $\boldsymbol{a_\phi}$ and $\boldsymbol{b_\phi}$ vectors list the parameters of the Uniform PDFs for the Tx, IRS, Ex1, Rx, and Ex2 in order.

In the context of long-range acoustic communication systems, the disorientation of transmitting components, namely the Tx and IRS, exerts a more substantial influence on link dynamics than the receiving components, such as the Rx. Consequently, our analysis focuses on the disorientation of the Tx and IRS, denoted as σ_{θ_t} and σ_{θ_i}, respectively.

The parameter σ_{θ_t} exhibits a linear relationship with the wind speed, V_w, expressed as: $\sigma_{\theta_t} = 0.00512\, V_w + 0.003 \pm 0.004$, where V_w denotes the wind speed measured at the sea surface level, confined within the range $V_w = [0, 16]$ m/s [46], [52], [56], [57]. While σ_{θ_i} lacks quantification in the literature, a similar analogy to σ_{θ_t} can be applied. Specifically, σ_{θ_i} may be linearly associated with the sea current speed and imperfections in the hovering systems, as: $\sigma_{\theta_i}(l_{ti}) = \beta_i \times (0.00512\, V_c(l_{ti}) + 0.003 \pm 0.004)$, $0 \leq \beta_i \leq 1$, where V_c represents the sea current speed and β_i denotes the orientation imprecision of the hovering system (OIHS). The OIHS factor can be obtained from the data associated to the AUV responsible for maintaining the specified depth of the IRS component. The sea current speed at the IRS depth, $V_c(l_{ti})$, depends on factors such as tide speed, wind speed, and IRS depth. Although there is no universally applicable equation for quantifying sea current speeds, empirical models offer a viable approach to addressing this challenge [50]. In this study, an empirical model that formulates sea current speeds is adopted as [33]

$$V_c(l_{ti}) = V_t \left(\frac{z_d - l_{ti}}{z_d}\right)^{(1/7)} + MAX\left\{V_v \left(\frac{z_w - l_{ti}}{z_w}\right), 0\right\}, \tag{1}$$

where V_t denotes the speed of the tide-driven current, confined within [0, 10] m/s range [45], [50], [53], [54]. Additionally, $V_v = (0.0235 \times V_w)$ represents the wind-driven current speed at the sea surface, z_d signifies the depth of seawater, and z_w denotes the depth at which the speed of wind-driven current becomes zero.

IV. SYSTEM PERFORMANCE

In this section, we introduce the channel model and system performance of the RL-BIA links.

A. CHANNEL MODEL

Credible channel modeling is ensured using authoritative references [6], [7], [58], [59], and [60] that span theory, empirical validation, experimentation, and real-world observations.

The channel DC gain of RL-BIA links is influenced by pointing errors and path loss. The channel DC gain of the Tx to IRS (Tx-IRS) link is denoted as h_{ti}, and determined as [58] and [59]

$$h_{ti} = \eta_{i_i}\, h_{l_{ti}}\, h_{p_{ti}} \tag{2}$$

where $h_{l_{ti}}$ and $h_{p_{ti}}$ represent the path loss and pointing errors of the Tx-IRS link, respectively. Similarly, the channel DC gain of the Tx to IRS to Rx (Tx-IRS-Rx) link, denoted as h_{tir}, is calculated as [58] and [59]

$$h_{tir} = \eta_{i_o}\, \eta_{r_i}\, h_{ti}\, h_{l_{ir}}\, h_{p_{ir}}, \tag{3}$$

where, η_{r_i}, $h_{l_{ir}}$, and $h_{p_{ir}}$ denote the electric-acoustic coefficient of the Rx, path loss, and pointing errors associated with the IRS-Rx link, respectively. Optimal passive IRS and Rx components are assumed, hence $\eta_{i_i} = \eta_{i_o} = \eta_{r_i} = 1$. In the subsequent analysis, the subscript χ is used to denote link parameters, where $\chi := ti$ and $\chi := ir$ for the Tx-IRS and IRS-Rx links, respectively.

The presence of seawater leads to signal absorption and beam spreading, resulting in path loss. The path loss of the χ^{th} link is dependent on the frequency, and is calculated as [58] and [59]

$$h_{l_\chi}(l_\chi, f_l, f_u) = \sqrt{\frac{\int_{f_l}^{f_u} A_{l_\chi}^{-1}(l_\chi, f)\, P_\chi(f)\, df}{P_\chi}}, \tag{4}$$

where $A_{l_\chi}(l_\chi, f)$ denotes the specific path loss, l_χ represents the range of the χ^{th} link in kilometers, f is the frequency in kilohertz, P_χ is the transmitted power, $P_\chi(f)$ is the frequency spectrum, and f_l and f_u are the lower and upper-frequency components in that spectrum, respectively. The specific path loss experienced over a distance l_χ and frequency f can be derived using Eqs. (1) and (2) in [7] as

$$A_{l_\chi}(l_\chi, f) = l_\chi^\mu \times 10^{\left[\frac{l_\chi}{10}\left(\frac{0.11 f^2}{1+f^2} + \frac{44 f^2}{4100+f^2} + 2.75\times 10^{-4} f^2\right) + \frac{l_\chi}{10} 0.003 + 3\mu\right]}, \tag{5}$$

where μ is the spreading beam factor. Utilizing the link model given in Fig. 2, the nominal lengths of the Tx-IRS and IRS-Rx

links are computed as

$$l_{ti} = \sqrt{(z_{t_o} - z_{i_o})^2 + (x_{t_o} - x_{i_o})^2 + (y_{t_o} - y_{i_o})^2} \times \sec(\theta_{t_o}),$$
$$l_{ir} = \sqrt{(z_{i_o} - z_{r_o})^2 + (x_{i_o} - x_{r_o})^2 + (y_{i_o} - y_{r_o})^2} \times \sec(\theta_{i_o}). \quad (6)$$

Assuming that the IRS and Rx receive acoustic Gaussian beams, the pointing error of the χ^{th} link is modeled as [60]

$$h_{p_\chi}(r_\chi, a_\chi, , w_\chi) \approx A_\chi \exp\left(-\frac{4\, v_\chi \exp(-v_\chi^2)\, r_\chi^2}{\sqrt{\pi}\ \mathrm{erf}(v_\chi)\, w_\chi^2}\right), \quad (7)$$

where r_χ is the radial distance between the arrival position of the beam axis and the center of the receiving aperture, $w_\chi \in \{w_{ti}, w_{ir}\}$, $a_\chi \in \{a_{ti}, a_{ir}\}$, $A_\chi = \left[\mathrm{erf}\left(v_\chi\right)\right]^2$, and $v_\chi = \frac{\sqrt{\pi}\, a_\chi}{\sqrt{2}\, w_\chi}$. Equation (7) provides a reliable approximation when the beam width is sufficiently larger than the receiving aperture, i.e., $w_\chi > 6\, a_\chi$ [60].

Assuming a high carrier frequency, the impacts of wind interference and thermal noise dominate over other sources of interference and noise [58], [59]. Using equations given in [61, Chapter 6] and performing unit transformations, the wind interference plus thermal noise, IN, can be computed in Watts as [58] and [59]

$$IN(V_w) = \frac{6.3492}{10^{17}} \int_{f_l}^{f_u} \left(10^{(-1.5+2\,\log(f))} + 10^{\left(0.75\, V_w^{1/2}+2\,\log(f)-4\,\log(f+0.4)+5\right)}\right) df, \quad (8)$$

where $\{f, f_l, f_u\}$ are measured in kilohertz, and V_w is measured in meters per second.

B. CHANNEL SECRECY AND OUTAGE PROBABILITY

The channel secrecy rate (CSR) quantifies the difference between the channel capacity of legitimate AUVs (i.e., IRS and Rx) and eavesdropping AUVs (i.e., Ex1 and Ex2). The CSR_χ is computed for the χ^{th} link as

$$CSR_\chi = \mathrm{MAX}(0, \hat{C}_{L_\chi} - \hat{C}_{E_\chi})\ \mathrm{BpCU}, \quad (9)$$

where $\hat{C}_{L_\chi}$ and $\hat{C}_{E_\chi}$ are the average channel capacities of the legitimate AUV and eavesdropping AUV in χ^{th} link, respectively. Moreover, BpCU is the abbreviation of bits per channel use unit. The $\hat{C}_{L_\chi}$ and $\hat{C}_{E_\chi}$ are computed as

$$\hat{C}_{L_\chi} = \int_0^{\chi_u} C_{L_\chi}(h_{L_\chi})\, f_{h_{L_\chi}}(h_{L_\chi})\, dh_{L_\chi}\ \mathrm{BpCU}.$$
$$\hat{C}_{E_\chi} = \int_0^{\chi_u} C_{E_\chi}(h_{E_\chi})\, f_{h_{E_\chi}}(h_{E_\chi})\, dh_{E_\chi}\ \mathrm{BpCU}. \quad (10)$$

where $\chi_u = A_{ti}\, h_{l_{ti}}$ and $\chi_u = A_{ti}\, h_{l_{ti}}\, A_{ir}\, h_{l_{ir}}$ for the Tx-IRS and Tx-IRS-Rx links, respectively. Moreover, $f_{h_\chi}(h_\chi)$ is the PDF of channel DC gain of the χ^{th} link, calculated as in [7]. Assuming a Gaussian channel, the capacity of χ^{th} channel is computed as

$$C_\chi(h_\chi) = \log_2(1 + SINR_\chi(h_\chi))\ \mathrm{BpCU}, \quad (11)$$

where $SINR_\chi(h_\chi)$ is the signal-to-interference plus noise ratio, and it is obtained as

$$SINR_\chi(h_\chi) = \frac{P_\chi\, h_\chi^2}{IN}. \quad (12)$$

On the other hand, the outage probability of the χ^{th} channel capacity, $F_{c_\chi}(c_{th})$, is computed as

$$F_{c_\chi}(c_{th}) = \int_0^{c_{th}} f_{c_\chi}(c_\chi)\, dc_\chi, \quad (13)$$

where c_χ is the channel capacity and c_{th} is its threshold. Moreover, $f_{c_\chi}(c_\chi)$ is the PDF of the χ^{th} channel's capacity, calculated as

$$f_{c_\chi}(c_\chi) = f_{h_\chi}(c_\chi) \left|\frac{dh_\chi}{dc_\chi}\right|, \quad (14)$$

where $f_{h_\chi}(c_\chi)$ represents the PDF of the channel DC gain for the χ^{th} link.

V. TWO AGENTS-BASED RL MODEL

The RL-BIA link aims to jointly optimize channel secrecy and outage probability by tuning the Tx beamwidth, IRS beamwidth, and IRS depth. Given the complexity of this non-convex, multi-parameter optimization in dynamic underwater environments, the RL is employed. This section outlines the RL model, formulates the optimization problem, and presents Q-learning and SARSA solvers.

A. RL MODEL

The interaction between the Tx, IRS, and environment is modeled as a Markov Decision Process (MDP), defined as:

$$\mathbb{M} = \{\mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{R}, \mathbb{F}\} \quad (15)$$

where $\mathbb{S}$, $\mathbb{A}$, $\mathbb{P}$, $\mathbb{R}$, and $\mathbb{F}$ represent the state space, action space, transition probabilities, reward function, and hyperparameters, respectively. The agents—Tx and IRS—iteratively optimize their beamwidth and depth based on feedback from the IRS and receiver (Rx).

1) STATE SPACE $\mathbb{S}$

The Tx state is defined by its beamwidth, while the IRS state includes beamwidth and depth. Their discrete state spaces are:

$$\mathbb{S}_{Tx} = \{w_{min}, w_{min} + \Delta_{w_{ti}}, \ldots, w_{max}\},$$
$$\mathbb{S}_{IRS} = \{(w_{ir}, l_{ti}) \mid w_{ir} \in [w_{min}, w_{max}], \; l_{ti} \in [0, L] \text{ in steps of } \Delta_{l_{ti}}\} \quad (16)$$

with w_{min}, w_{max} denoting beamwidth bounds, $\Delta_{w_{ti}}, \Delta_{w_{ir}}$ the beamwidth step sizes for Tx and IRS, and $\Delta_{l_{ti}}$ the IRS depth increment. At time t, the states are:

$$s_t(t) \in \mathbb{S}_{Tx}, \quad w_{min} \le w_{ti} \le w_{max},$$
$$s_i(t) \in \mathbb{S}_{IRS}, \quad 0 \le l_{ti} \le L \quad (17)$$

2) ACTION SPACE $\mathbb{A}$

The Tx updates beamwidth, while the IRS adjusts both beamwidth and depth. Action sets are:

$$\begin{aligned}\mathbb{A}_{Tx} &= \{-\Delta_{w_{ti}}, 0, \Delta_{w_{ti}}\},\\ \mathbb{A}_{IRS} &= \{(\delta_w, \delta_l) \,|\, \delta_w \in \{-\Delta_{w_{ir}}, 0, \Delta_{w_{ir}}\},\\ &\quad \delta_l \in \{-\Delta_{l_{ti}}, 0, \Delta_{l_{ti}}\}\}\end{aligned} \tag{18}$$

with actions at time t given by:

$$a_t(t) \in \mathbb{A}_{Tx}, \quad a_i(t) \in \mathbb{A}_{IRS} \tag{19}$$

3) TRANSITION PROBABILITIES $\mathbb{P}$

State transitions are deterministic; each action deterministically updates the current state, i.e., $P(s'|s, a) \in \{0, 1\}$.

4) REWARD FUNCTION $\mathbb{R}$

The reward reflects link quality and secrecy performance, dependent on the channel secrecy rate (CSR) and outage probability $F_{c_{tir}}(t)$ compared to a threshold F_{th}. Let CSR_{ti} and CSR_{tir} be the CSR at the IRS and Rx, respectively:

$$\mathbb{R}_{Tx} = \{CSR_{ti},\ 0\}, \quad \mathbb{R}_{IRS} = \{CSR_{tir},\ 0\} \tag{20}$$

At time t, rewards for the Tx and IRS agents are given by:

$$\begin{aligned} r_t(t) &= \begin{cases} CSR_{ti}(t), & \text{if } F_{c_{tir}}(t) \le F_{th} \\ 0, & \text{otherwise,} \end{cases}\\ r_i(t) &= \begin{cases} CSR_{tir}(t), & \text{if } F_{c_{tir}}(t) \le F_{th} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{21}$$

5) HYPERPARAMETERS $\mathbb{F}$

These control learning dynamics. Let:

$$\mathbb{F}_{Tx} = \{\gamma_t, \alpha_t, \epsilon_t\}, \quad \mathbb{F}_{IRS} = \{\gamma_i, \alpha_i, \epsilon_i\} \tag{22}$$

where $\gamma \in [0.5, 0.99]$ is the discount factor balancing future and immediate rewards, $\alpha \in (0, 1]$ is the learning rate, and ϵ is the exploration rate. Smaller ϵ promotes greedy (exploitation) behavior [34], [62].

B. FORMULATION OF OPTIMIZATION PROBLEM

The Tx and IRS agents learn policies by navigating the state-action space and interacting with received rewards over the long term. The Tx's policy, π_{Tx}, maps each state-action pair to the probability $\pi_{Tx}(a_t \mid s_t)$ of taking action a_t in state s_t. Similarly, the IRS's policy, π_{IRS}, maps each state-action pair to the probability $\pi_{IRS}(a_i \mid s_i)$ of taking action a_i in state s_i. Starting from a specific state and following their respective policies, the agents compute the expected total reward. For the Tx, the value of state s_t, denoted as $V^{\pi}_{Tx}(s_t)$, is expressed as:

$$V^{\pi}_{Tx}(s_t) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} \gamma_t^t r_t(t) \,\middle|\, s_t(0) = s_t(t)\right], \tag{23}$$

where $\mathbb{E}$ denotes the expectation under policy π_{Tx}, and T is the total number of iterations. Similarly, for the IRS, the value of state s_i, denoted as $V^{\pi}_{IRS}(s_i)$, is computed as:

$$V^{\pi}_{IRS}(s_i) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T} \gamma_i^t r_i(t) \,\middle|\, s_i(0) = s_i(t)\right], \tag{24}$$

while the Tx and IRS agents can try several policies, optimal policies exist that maximize the average reward for the Tx-IRS and IRS-Rx links, i.e., maximizing channel secrecy while minimizing the outage probability. The agents pursue these optimal policies through exploration (trying new actions) and exploitation (selecting actions that have previously yielded high rewards). Striking a good balance between exploration and exploitation enables the agents to address the non-convexity, multi-dimensionality, and duality nature of the optimization problem, ultimately converging to optimal or suboptimal policies.

1) DEFINING THE OPTIMIZATION OBJECTIVE

The agents aim to learn optimal policies π^*_{Tx} and π^*_{IRS} that maximize their respective value functions while satisfying operational constraints, such as beamwidth and depth limits. The optimal policy for the Tx agent, denoted as $\pi^*_{Tx}(s_t, t)$, can be formulated as optimization problems as

$$\begin{aligned}\pi^*_{Tx}(s_t, t) = \underset{\mathbb{A}_{Tx}}{\text{Arg-Max}} : V^{\pi}_{Tx}(s_t, t)\,,\\ \textit{Subject to}: \quad w_{min} \le w_{ti} \le w_{max}\,,\end{aligned} \tag{25}$$

as well, the optimal policy of the IRS, donated as $\pi^*_{IRS}(s_i, t)$, could be formulated as

$$\begin{aligned}\pi^*_{IRS}(s_i, t) = \underset{\mathbb{A}_{IRS}}{\text{Arg-Max}} : V^{\pi}_{IRS}(s_i, t)\,,\\ \textit{Subject to}: \ 0 \le l_{ti} \le L\,, \quad w_{min} \le w_{ir} \le w_{max}\,,\end{aligned} \tag{26}$$

The above formulations represent a *dual-objective* optimization, which can be computationally challenging to solve in real-time.

2) REFORMULATION FOR TRACTABILITY

to relax the complexity of the optimization problems, Equations (25) and (26), the dual-objective reward function can be transformed into a single-objective reward function by maximizing the channel secrecy rate while constraining the outage probability to remain below a predefined threshold. Equation (25) can be reformulated as:

$$\begin{aligned}\pi^*_{Tx}(s_t, t) = \underset{\mathbb{A}_{Tx}}{\text{Arg-Max}} : V^{\pi}_{Tx}(s_t, t)\,,\\ \textit{Subject to}: \quad F_{c_{ti}} \le F_{th}\,, \quad w_{min} \le w_{ti} \le w_{max}\,,\end{aligned} \tag{27}$$

where $F_{c_{ti}}$ and F_{th} are the outage probability and outage threshold, respectively, of the Tx-IRS link. As well, Eq. (26) could be reformulated as

$$\begin{aligned}\pi^*_{IRS}(s_i, t) = \underset{\mathbb{A}_{IRS}}{\text{Arg-Max}} : V^{\pi}_{IRS}(s_i, t)\,,\\ \textit{Subject to}: \ F_{c_{ir}} \le F_{th}\,, \quad 0 \le l_{ti} \le L\,,\end{aligned}$$

$$w_{min} \leq w_{ir} \leq w_{max}\,, \tag{28}$$

where $F_{c_{ir}}$ is the outage probability of the IRS-Rx link. Note that Eqs. (25)-(26) represent a double-objective reward function with fewer constraints, while Eqs. (27)-(28) represents a single-objective reward function with additional constraints.

3) JOINT OPTIMIZATION VIA MAX-MIN FAIRNESS

Equations (27) and (28) present optimal policies that independently maximize the channel secrecy rates at the IRS and Rx locations for the Tx-IRS and IRS-Rx links, respectively. However, these policies cannot concurrently maximize the CSR at both locations due to the mutual influence between the links. To address this, a policy that maximizes the minimum CSR at the IRS and Rx locations is required. To achieve this, the Max-Min technique is employed. The Max-Min technique could be compromised while maintaining the fairness between the secrecy of the Tx-IRS and IRS-Rx channels. The optimal policy of the Tx-IRS-Rx link, donated as $\overset{*}{\pi}_{BIA}(s_t, s_i, t)$, is formulated as

$$\begin{aligned}\overset{*}{\pi}_{BIA}(s_t, s_i, t) = \underset{\{\mathbb{A}_{Tx},\, \mathbb{A}_{IRS}\}}{\text{Arg-Max-Min}} : \{V^{\pi}{}_{Tx}(s_t, t), V^{\pi}{}_{IRS}(s_i, t)\},\\ \textit{Subject to}: \; F_{c_{tir}} \leq F_{th}\,, \quad 0 \leq l_{ti} \leq L\,,\\ w_{min} \leq w_{ti} \leq w_{max}\,,\\ w_{min} \leq w_{ir} \leq w_{max}\,,\end{aligned} \tag{29}$$

the value functions in this equation, $V^{\pi}{}_{Tx}(s_t, t)$ and $V^{\pi}{}_{IRS}(s_i, t)$, depend on the reward functions, which are computed using Eqs. (21). Clearly, Eq. (29) ensures optimal fairness between the CSR values at the IRS and Rx locations while constraining the outage probability of the Tx-IRS-Rx link to remain below a predefined threshold, F_{th}.

4) COMPUTATIONAL CONSIDERATIONS AND LEARNING STRATEGY

Computing the optimal policy in Eq. (29) is resource-intensive, often exceeding the capabilities of IRSs and AUVs. Iterative algorithms, such as model-assisted RL, provide suboptimal policies efficiently by conducting virtual experiments to update state-values without actual actions. However, the dynamic, unpredictable underwater environment makes establishing a model-assisted RL impractical, necessitating model-free algorithms like Q-learning and SARSA, which, despite slower convergence, are essential for policy optimization. In the next subsection, we introduce Q-learning and SARSA solvers.

C. Q-LEARNING AND SARSA SOLVERS

Q-learning and SARSA are based on the value of state-action pairs, denoted as $Q_{\chi}(s_{\chi}, a_{\chi}, t)$, where $\chi := t$ and $\chi := i$ for the Tx-IRS and IRS-Rx links, respectively. The value of taking action a_{χ} in state s_{χ} under a policy π_{χ}, denoted as $Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)$, is defined as the expected reward from taking that action at the given state and subsequently following policy π_{χ}. The optimal values of the value functions in Eq. (29), denoted as $V^{\pi^*}_{\chi}(s_{\chi}, t)$, can be computed from $Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)$ as:

$$V^{\pi^*}_{\chi}(s_{\chi}, t) = \underset{\mathbb{A}_{\chi}}{\text{Arg Max}} \; : \; Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)\,, \tag{30}$$

This equation indicates that the maximum value of a state is obtained by identifying the maximum value of the actions that can be taken in that state and subsequently following the policy π_{χ}.

The Q-learning algorithm approximates the solution of Eq. (30), i.e., maximizing $Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)$, by the following iterative process [62]:

$$\begin{aligned}&Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)\\ &\leftarrow (1-\alpha_{\chi})\, Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t) + \alpha_{\chi}\\ &\quad \times \left[r_{\chi}(t) + \gamma_{\chi} \max_{a_{\chi}} Q^{\pi}_{\chi}(s_{\chi}(t), a_{\chi}(t+1)) \right],\end{aligned} \tag{31}$$

this equation highlights that the Q-learning algorithm is an off-policy method that updates the action-value functions, $Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)$, by evaluating the maximum value among all possible actions at a given state s_{χ}. The SARSA algorithm approximates the solution of Eq. (30), i.e., maximizing $Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)$, through the following iterative process [62]:

$$\begin{aligned}&Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t)\\ &\leftarrow (1-\alpha_{\chi})\, Q^{\pi}_{\chi}(s_{\chi}, a_{\chi}, t) + \alpha_{\chi}\\ &\quad \times \left[r_{\chi}(t) + \gamma_{\chi}\, Q^{\pi}_{\chi}(s_{\chi}(t+1), a_{\chi}(t+1)) \right],\end{aligned} \tag{32}$$

this equation highlights that, unlike Q-learning, the SARSA algorithm is an on-policy approach. It does not necessarily use the maximum reward for the next state to update the Q-values. Instead, it selects a new action and its corresponding reward based on the same policy that determined the initial action. The colored terms in this equation and Eq. (32) highlight the mathematical difference between the Q-learning and SARSA algorithms.

Equations (15)-(24) and (29)-(32) are implemented using the Pseudocode shown in Algorithms 1 and 2 shown and discussed in Appendix. Algorithms 1 and 2 delineate the core procedures of the Q-learning and SARSA algorithms, respectively. Over numerous episodes and iterations, agents learn which actions maximize channel secrecy and reliability, gradually converging towards suboptimal policies. In the tracking mode, agents apply their learned policies to maximize channel secrecy and reliability performance, autonomously adapting to the environment. The algorithms re-initiate training only when significant environmental changes occur, indicated by discrepancies between current and optimal policies. Significant environmental changes necessitate significant training time to converge the policy again to a suboptimal value, while minor changes may not require retraining. To ensure efficiency, the learning mode interval should be shorter than the tracking mode to minimize the energy consumption and communication interruptions.

TABLE 3. Simulation parameters for the numerical results [7], [28], [33], [68].

The Simulation Parameter	Its Value
Discount Factor of the RL algorithms	$\gamma_o = \{0.8, 0.9, 1\}$
Learning rate of the RL algorithms	$\alpha_o = \{0.05, 0.1, 0.2\}$
Exploration Probability of the RL algorithms	$\epsilon_o = \{0.25, 0.50, 0.75\}$
Link length (L)	500 m
Horizontal inter-distances (d_h)	$0.1 \times$ L
Seawater depth and depth at which wind speed is zero	$z_d = L$, $z_w = 100$ m
IRS' aperture radius (a_{ti})	0.5 m
OIHS of drone attached to the IRS (β_i)	$\{1, 0.9, 0.8, 0.5\}$
Rx, Ex1 and Ex2 aperture radius (a_{ir}, a_{e_1}, a_{e_2})	0.1 m
Bandwidth of the Tx and IRS (w_{ti}, w_{ir})	[5:0.25:10] meter
Depth of IRS (l_{ti})	[30:15:270] meter
Transmitted power of the Tx (P_{tir})	2 Watt
Localization's root mean square error	5 m
Gaussian standard deviation (e.g., $\sigma_{x_t}, \sigma_{y_t}, \sigma_{x_j}, \sigma_{y_j}$)	5 m
Outage threshold (F_{th})	0.05
Radius of the safety zones (d_s, d_s)	5 m
Spreading factor of the acoustic beam (μ)	1.5
Acoustic carrier ($f_c = (f_u + f_l)/2$)	12 kHz
Signal bandwidth ($BW = f_u - f_l$)	4 kHz

D. PRACTICAL IMPLEMENTATION CONSIDERATIONS

While the proposed RL-BIA link model offers strong theoretical advantages, its practical deployment in underwater environments poses several challenges. A primary concern is real-time decision-making under limited computational and memory resources typical of AUVs and buoyed platforms. To address this, lightweight reinforcement learning variants can be used, with training conducted offline on surface stations or centralized servers when feasible [63]. Although training time depends on model complexity, it remains manageable with current computing resources since it occurs offboard. The resulting trained policies, which have modest memory requirements, can then be deployed to AUVs for efficient real-time execution with minimal overhead. Computational limitations, memory constraints, and offline training are key considerations to enabling scalable RL-BIA implementation in real-world underwater acoustic communications.

Synchronization between the mobile AUVs and the buoyed IRS is another practical challenge, which can be addressed through underwater acoustic signaling protocols utilizing time-stamping and error correction techniques [64]. In addition, hardware non-idealities -such as actuator delays or limited IRS tuning granularity- may impact performance. However, the adaptive nature of reinforcement learning provides robustness to such uncertainties through continuous online adjustments [65]. Moreover, RL methods typically do not require precise channel state information, which reduces reliance on accurate channel estimation and enhances their robustness to estimation errors [66].

Finally, to support sustainable deployment in energy-constrained underwater platforms, the system design incorporates energy-aware reward functions and efficient training routines that prioritize low-power operation [67].

VI. NUMERICAL RESULTS

In this section, we present numerical results for the proposed RL-BIA system. We deliberately compare the RL-BIA link with basic benchmark methods, i.e., RL-BA, exhaustive search-BA, and exhaustive search-BIA, to establish a clear and interpretable baseline for evaluating the proposed system. These methods provide a controlled and well-understood comparison framework, which is essential during the early stages of developing and validating a new RL-based system. To ensure a fair comparison, we maintain consistency in the simulation parameters across all link configurations, as summarized in Table 3.

A. SIMULATION SETTINGS AND LINK SETUPS

In this subsection, we provide a detailed overview of the simulation settings and link setups used in our experiments.

1) SIMULATION SETTINGS

Table 3 summarizes the key parameters used in the simulations. To ensure practical relevance and robustness, the simulation settings were selected based on well-established studies and realistic constraints of underwater deployments [6], [7], [28], [33], [68].

Physical layer parameters, e.g., acoustic carrier frequency, 12 kHz, bandwidth, 4 kHz, and transmit power, 2 Watt, reflect typical underwater transceiver configurations [7], [68]. IRS related parameters, e.g., beamwidth, $[5, 10]$ meters, and IRS depth, $[30, 270]$ meters, were chosen to match common AUV-to-buoy setups and the sea current profiles observed in shallow and mid-depth environments [33]. Environmental parameters, e.g., wind speed, and sea tides speed, [2, 14] meter/sec, were selected to reflect real-world conditions as reported in prior work [33], [49]. The underwater acoustic channel model, e.g., frequency-dependent absorption, interference, and background noise models, are computed as described in [6], [7], and [68].

The RL behavior of the Tx and IRS agents were adopted from validated studies under similar underwater conditions [28]: the learning rate (α) is set to values, $\{0.05, 0.1, 0.2\}$, to control the Q-table update magnitude, the discount factor (γ) is chosen, $\{0.8, 0.9, 1\}$, to balance the importance of long-term rewards, and the exploration probability (ϵ) is initialized, $\{0.8, 0.9, 1\}$, to support efficient exploration. Training is conducted over 200 episodes, each containing 50-100 iterations. Policy convergence is assumed when the change remains non-significant over consecutive episodes. The agents operate over discrete beamwidths and IRS depth states: the beamwidth ranges from 5 meters to 10 meters in steps of 0.25 meters, while the IRS depth varies from 200 meters to 300 meters in increments of 10 meters. The IRS depth spans a total of 300 meters, varying from 10% to 90% of this distance in increments of 5%. The channel secrecy rate is computed using Eq. (9) at both the IRS and Rx locations, and the outage constraint is enforced with a threshold $F_{th} = 0.2$.

Simulations are executed using Python 3.10 and MATLAB *R2024a* on a system running Windows 11 equipped with an Intel Core *i*8 processor and 64 GB RAM. This setup ensures full reproducibility of the proposed RL-based secure communication system in dynamic underwater environments.

2) LINK SETUPS

We simulate the link setup illustrated in Fig.2. This configuration is supported by diverse sources, including theoretical analyses [7], [28], [49], environmental data from the Operational Oceanographic Products and Services [43], empirical studies [45], [50], laboratory experiments [51], and buoy disturbance models [52]. These collectively provide a solid foundation for setting up realistic marine scenarios.

Initially, the components are deployed with nominal locations and orientations to minimize geometric losses. In this challenging and dynamic environment, the Tx sways on the seawater surface due to sea waves, while the IRS, Ex1, Rx, and Ex2 float in the seawater under the influence of sea currents. To enhance its stability, the IRS is mounted on a hovering drone with $\beta_i = 0.9$. The total link length is $L = 500$ m, while the IRS depth, l_{ti}, is adaptable. The nominal positions (in meters) for the Tx, IRS, and Rx are $(0, 0, 0)$, $(0.07L, 0.07L, l_{ti})$, and $(0, 0, L)$, respectively. To mitigate the eavesdropping risk, safety zones with a radius of 5 meters surround the IRS and Rx. Ex1 and Ex2 are nominally located (in meters) at $(3.6+0.07L, 3.6+0.07L, l_{ti})$ and $(3.6, 3.6, L)$, respectively. The Tx and IRS utilize acoustic Gaussian beams with adaptive beam widths, while the IRS can hover at an adjustable depth beneath the seawater surface. The Rx, Ex1, and Ex2 share a common aperture radius of 0.1 meters, suitable for compact commercial AUVs. In contrast, the IRS has a relatively larger aperture radius of 0.5 meters, designed for larger commercial drones. This larger size allows us to explore the potential of the RL-BIA link under severe wind and tide turbulence, including wind speeds of up to 16 m/s, current speeds of up to 14 m/s, and displacements with a variance of 5 meters in the X_o and Y_o axes. A smaller IRS could be employed for scenarios with less severe dynamic conditions.

B. PERFORMANCE EVALUATION OF Q-LEARNING AND SARSA ALGORITHMS

In this subsection, the performance of Q-learning and SARSA algorithms are evaluated by considering three metrics: the average reward, success rate, and progressing rate of the Q-table. The average reward signifies the average cumulative reward obtained by the RL agents across multiple interactions with the environments. It reflects the effectiveness of the agent's actions in achieving its goals over time. The success rate denotes the percentage of trials in which the RL agent successfully achieve its goal. It provides insights into the agent's ability to accomplish tasks. The progressing rate of the Q-table indicates how quickly the Q-values in the table converge to the suboptimal values as the agent interacts with the environment and receives feedback. This metric measures the rate at which the agent experiences the environment.

Figure 3 displays six sub-figures, labeled (a) to (f), arranged in two rows and three columns. The first and second rows present the metrics of the RL-BIA and RL-BA links, respectively. The first, second, and third columns correspond to the average reward, success rate, and Progressing of the Q-table, respectively. Each sub-figure compares two RL algorithms: Q-Learning and SARSA, highlighted in different colors. The results are plotted against the episode index, representing the number of episodes the agents have completed.

Figures 3(a) and 3(d), the average reward curves exhibit a general upward trend with an increasing episode index, indicating that the agents learn and improve their performance over time. The Q-learning and SARSA algorithms converge to similar average rewards by the end of 200 episodes. However, SARSA tends to oscillate more than Q-Learning due to its partial-greedy nature, introducing additional randomness. The figures show that the RL-BA link achieves higher rewards. Numerically, the RL-BIA and RL-BA links achieve average rewards of 2.5 BpCU and 3.5 BpCU, respectively. This unexpected result is attributed to the weak turbulence conditions considered, such as $V_w = 1$ m/s and $V_t = 3$ m/s, where the addition of IRS components in the RL-BIA link causes performance degradation.

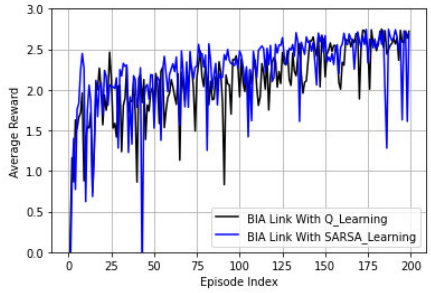
Similarly, Figs. 3(b) and 3(e) illustrate an increasing success rate trend as the episode index rises. Both RL-BIA and RL-BA links eventually achieve a success rate of 1 after around 150 episodes, indicating mastery of the task. Notably, SARSA outperforms Q-Learning in the RL-BA link, possibly due to the latter's greedy exploration strategy affecting performance.

Finally, Figs. 3(c) and 3(f) illustrate the progressing rate curves with increasing episode index. The RL-BA link converges more rapidly compared to the RL-BIA link because its environment is smaller relative to that of the RL-BIA link. By the end of 100 episodes, the RL-BA link reaches optimal or near-optimal values. Additionally, SARSA exhibits more oscillations than Q-Learning, which can be attributed to its partial-greedy approach.

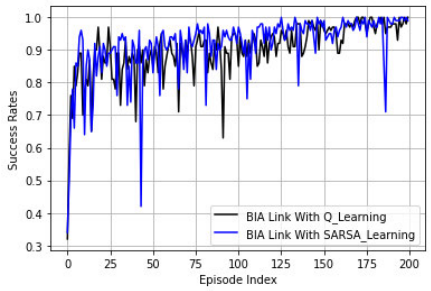
C. CSR PERFORMANCE EVALUATION VERSUS WIND SPEED

In this subsection, Figure 4 illustrates the CSR results, in BpCU unit, for RL-BIA links as a function of wind speed in meters per second. The simulation was repeated three times under identical conditions, as depicted in sub-figures (a), (b), and (c). The results for the Q-learning and exhaustive search algorithms are highlighted in blue and orange, respectively.

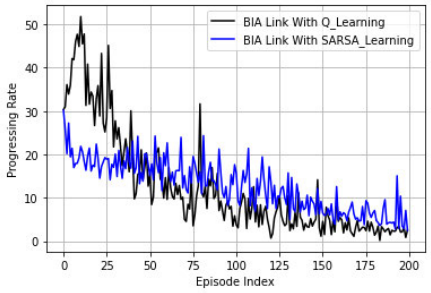
The performance of the Q-learning algorithm exhibits variability and lower effectiveness compared to the exhaustive search method across all three experiments. This variability arises from Q-learning's reliance on exploration and learning processes, which can result in inconsistent outcomes in certain scenarios. Conversely, the exhaustive search scheme consistently outperforms Q-learning, delivering stable and superior results across the experiments. Its performance remains fixed because it systematically evaluates all possible solutions to determine the optimal one, avoiding the



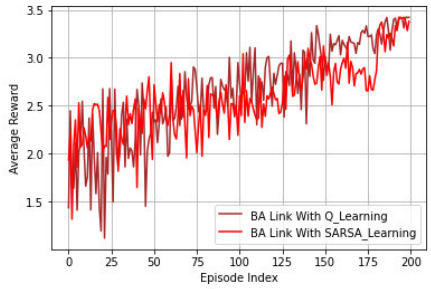
(a) Average reward (in units of BpCU) of RL-BIA links



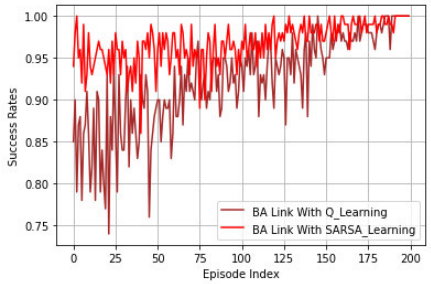
(b) Succeses rate (unit less) of RL-BIA links



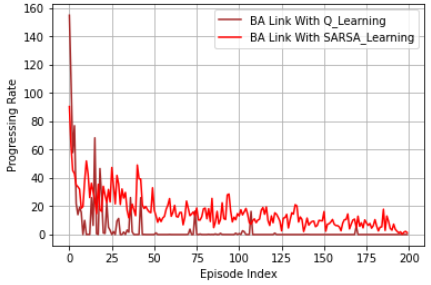
(c) Progressing rate (in units of BpCU) of RL-BIA links



(d) Average reward (in units of BpCU) of RL-BA links

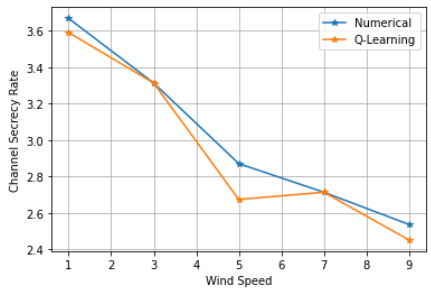


(e) Succeses rate (unit less) of RL-BA links

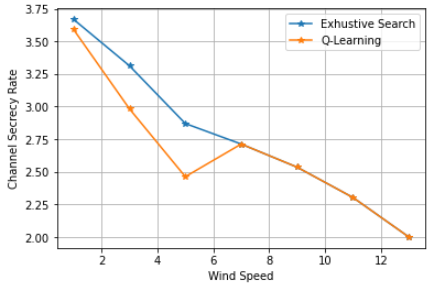


(f) Progressing rate (in units of BpCU) of RL-BA links

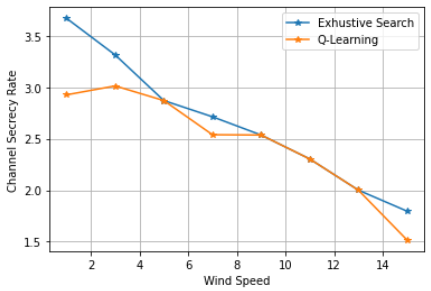
FIGURE 3. Evaluations of the Q-learning and SARSA algorithms implemented with RL-BIA and RL-BA links, focusing on average reward, success rate, and progression rate.



(a) First experiment



(b) Second experiment



(c) Third experiment

FIGURE 4. Comparison of CSR performance (in BpCU) for RL-BIA links under varying wind speeds using Q-learning (orange) and exhaustive search (blue) across three independent simulation runs shown in sub-figures (a), (b), and (c).

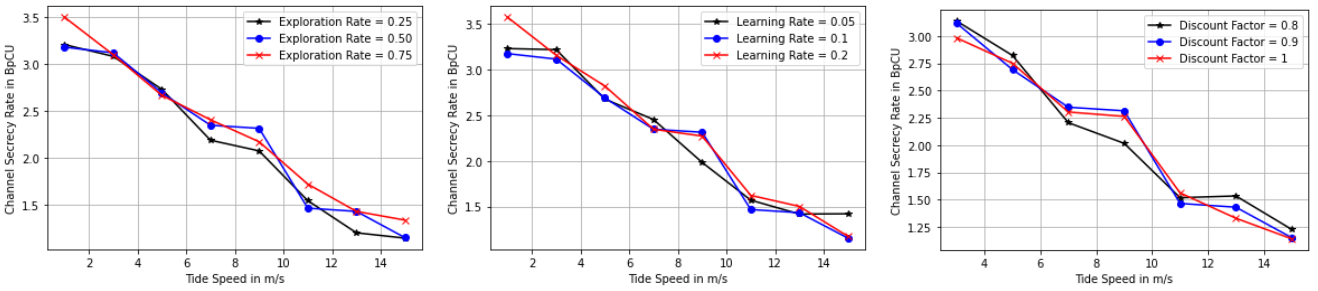
uncertainty associated with learning-based methods. For instance, at $V_w = 5$ m/s, Q-learning records CSR values of 2.7, 2.5, and 2.9 BpCU as shown in sub-figure (a), (b), and (c), respectively. While the exhaustive search consistently achieves a CSR value of 2.9 BpCU in all sub-figures.

Although the exhaustive search scheme outperforms Q-learning, it has significant drawbacks. It requires prior system information (making it a model-based technique) and demands substantial computational resources, with complexity growing exponentially with the size of the state-action space, rendering it impractical for real-time applications. In contrast, the Q-learning technique provides suboptimal solutions without requiring prior system knowledge (as a model-free approach) and operates with significantly greater computational efficiency, as its complexity does not scale exponentially with the state-action space.

D. IMPACTS OF THE HYPERPARAMETERS ON THE SYSTEM PERFORMANCE

In this subsection, we explore the impact of hyperparameters on the CSR performance of the RL-BIA link concerning tide speed in meters per second. We investigate three hyperparameters: exploration rate, learning rate, and discount factor.

Figure 5 comprises three sub-figures, labeled (a) to (c), arranged in a single row with three columns. Each sub-figure depicts the CSR performance obtained using the Q-Learning scheme. The shown curves represent the average of five



(a) Exploration Rate (b) Learning Rate (c) Discount Factor

FIGURE 5. Impact of Q-learning hyperparameters on CSR performance (in BpCU) for RL-BIA links under varying tide speeds. Sub-figures (a)–(c) illustrate the effect of (a) exploration rate, (b) learning rate, and (c) discount factor, respectively.

runs to mitigate the effects of randomness in the Q-Learning scheme results.

Figure 5(a) illustrates the effect of varying the exploration rate across three values: 0.25, 0.50, and 0.75. While the curves generally follow similar trends, the red curve consistently outperforms the others, while the black curve exhibits the poorest performance. Higher exploration rates provide agents with greater flexibility to explore the environment, resulting in the discovery of better rewards. Numerically, at a tide speed of 1 m/s, setting the exploration rate to 0.75 and 0.25 achieves CSR performances of 3.5 and 3.2, respectively. Similarly, at a tide speed of 11 m/s, the same exploration rates yield CSR performances of 1.75 and 1.5, respectively.

Figure 5(b) demonstrates the effect of varying the learning rate across three values: 0.05, 0.1, and 0.2. The red curve consistently outperforms the others, while the black curve exhibits the lowest performance. Higher learning rates enable agents to learn more efficiently from the environment, resulting in better rewards. Numerically, at a tide speed of 1 m/s, setting the learning rate to 0.2 and 0.05 yields CSR performances of 3.6 and 3.2, respectively. Similarly, at a tide speed of 9 m/s, the same learning rates yield CSR performances of 2.8 and 2, respectively.

Figure 5(c) investigates the impact of varying the discount factor across three values: 0.8, 0.9, and 1. The blue curve consistently exhibits superior performance, while the black curve demonstrates the poorest performance. Setting the discount factor to the intermediate value of 0.9 enhances the intelligence of the Q-learning scheme. For instance, at a tide speed of 9 m/s, setting the discount factor to 0.9 and 0.8 yields CSR performances of 2.25 and 2, respectively.

E. CONTRAST CSR PERFORMANCE OF THE RL-BIA AND RL-BA LINKS

In this subsection, the secrecy performance of the RL-BA and RL-BIA links is compared, highlighting the CSR in units of BpCU versus wind speed in meters per second.

Figure 6 contrasts the secrecy performance of the RL-BA and RL-BIA links, where generally, we observe a degradation in channel secrecy with increasing wind speed. Higher wind speeds induce significant link disruptions, which negatively

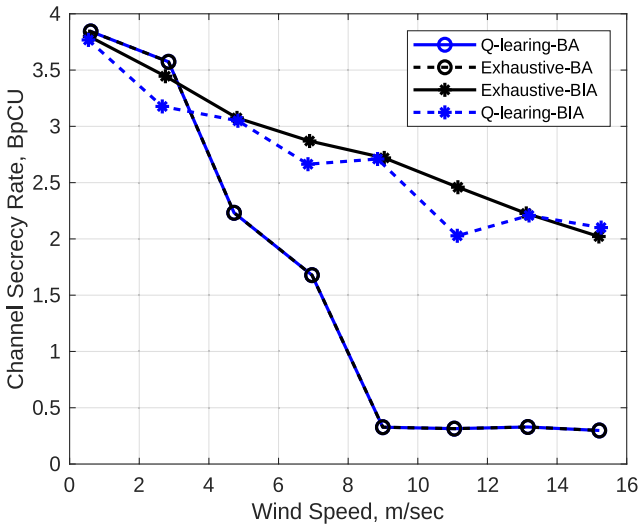


FIGURE 6. Comparison of CSR performance (in BpCU) versus wind speed (in m/sec) for RL-BIA and RL-BA links using Q-learning and exhaustive research.

impact link secrecy. The optimal CSR values for the RL-BIA and RL-BA links are depicted by solid black and dotted black curves, respectively, obtained using the exhaustive search method. Notably, the RL-BIA link significantly outperforms the RL-BA link in medium and strong turbulence conditions, i.e., $V_w \geq 3$ m/s. As an illustration, at a wind speed of 5 meters per second, the optimal CSR values for the RL-BIA and RL-BA links are 2.6 and 2 BpCU, respectively. Similarly, at a wind speed of 8.5 meters per second, the optimal CSR values for the RL-BIA and RL-BA links are 2.5 and 0.5 BpCU, respectively. This means that the RL-BIA scheme enhances the channel secrecy rate by a factor of five, corresponding to a 400% improvement. However, the RL-BA link slightly outperforms the RL-BIA link in weak turbulence conditions, i.e., $V_w \leq 3$ m/s. As an illustration, at a wind speed of 2 meters per second, the realized CSR values for the RL-BIA and RL-BA links are 3.5 and 3 BpCU, respectively. Clearly, RL-BIA links demonstrate superior performance compared to RL-BA links when the IRS

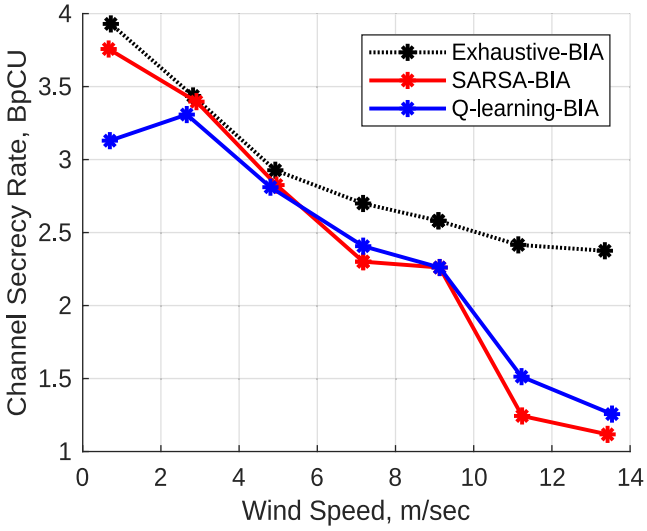


FIGURE 7. CSR performance (in BpCU) versus tide speed (in m/sec) for the RL-BIA link using SARSA (red), Q-learning (blue), and exhaustive search (dotted black).

component is affected by slow tide-induced currents relative to the wind-induced waves impacting the Tx component.

While the Q-Learning scheme achieves optimal CSR results in the RL-BA link (i.e., matching the exhaustive search), its performance in the RL-BIA link falls slightly below the exhaustive search results. This can be attributed to the complexity of the RL-BIA links environment. In simpler environments like RL-BA, the agent can attain optimal results with reasonable training, while in more complex environments like RL-BIA, optimal results may not be achievable. For instance, at a wind speed of 3 meters per second, the exhaustive search and Q-Learning achieve CSR values of 3.4 and 2.75 BpCU, respectively. Similarly, at a wind speed of 11 meters per second, the exhaustive search and Q-Learning yield CSR values of 2.3 and 1.75 BpCU, respectively.

F. CONTRAST THE PERFORMANCE OF THE Q-LEARNING AND SARSA ALGORITHMS

In this subsection, we compare the secrecy performance of the Q-Learning and SARSA algorithms for the RL-BIA link. Figure 7 illustrates the channel secrecy rate metric versus tide speeds in meters per second. The results of the exhaustive search, Q-Learning, and SARSA are depicted using dotted-black, blue, and red colors, respectively. Generally, both the Q-Learning and SARSA algorithms approach optimal results at low tide speeds, but the performance diverges at higher tide speeds. SARSA outperforms Q-Learning at lower tide speeds, while the opposite trend is observed at higher tide speeds. For instance, at a tide speed of 1 meter per second, SARSA and Q-Learning algorithms achieve CSR values of 3.8 and 3.2, respectively. Conversely, at a tide speed of 13 meters per second, SARSA and Q-Learning algorithms achieve CSR values of 1.25 and 1.4, respectively.

VII. CONCLUSION AND FUTURE WORKS

The deployment of traditional Buoyed-AUV (RL-BA) links in underwater environments faced challenges due to the dynamic and unpredictable nature. To overcome these obstacles, the RL-assisted Buoyed-IRS-AUV (RL-BIA) links were introduced as a solution. The RL-BIA link navigated the environment through trial and error. It dynamically adjusted the beamwidth and IRS depth to adapt to seawater turbulence induced by wind and tide speeds. A comprehensive link model was provided, incorporating factors such as pointing errors and path loss, and seamlessly integrating RL technology into the BIA link. Significant improvements were achieved by leveraging Max-Min optimization, which incorporated channel secrecy and outage probability, alongside Q-learning and SARSA algorithms. The learning behavior of the Q-Learning and SARSA algorithms was delved into, showcasing metrics such as average reward, success, and progressing rates. Moreover, the impact of RL hyperparameters on the channel secrecy rate performance is investigated. Comparing RL-BIA and RL-BA links, the former's superiority was demonstrated. The importance of shorter training modes in RL-BIA links for energy conservation and communication sustainability is underscored in this study. We underscored the importance of shorter training modes in RL-BIA links for energy conservation and communication sustainability. Furthermore, optimizing hyperparameters is crucial for maximizing CSR performance. While Q-learning and SARSA algorithms perform similarly overall, the performance may vary under specific wind and tide speeds.

While this work presents comprehensive simulation-based evaluations of the proposed RL-BIA framework, including comparisons with RL-BA systems, we recognize the importance of broader benchmarking. Future research could expand the comparative analysis to include other state-of-the-art IRS-assisted underwater acoustic communication schemes—both RL-based and non-RL-based, e.g., heuristic methods, offline-optimized IRS approaches, and traditional cryptographic security mechanisms from recent literature, to enhance contextual understanding. Additionally, we acknowledge that the complexity of real-world underwater environments, such as multipath interference, hardware imperfections, and unpredictable dynamics, may not be fully captured in simulation. As a next step, we could implement a hardware prototype and conduct field trials to validate the practicality and robustness of the RL-BIA system under real operating conditions. These efforts will help bridge the gap between simulation and real-world deployment.

APPENDIX
PSEUDOCODE FOR Q-LEARNING AND SARSA ALGORITHMS

The pseudocode in Algorithms 1 and 2 implements the formulations presented in Equations (15)–(24) and (29)–(32). Specifically, Algorithm 1 and Algorithm 2 delineate the core procedures of the Q-learning and SARSA algorithms,

Algorithm 1 Q-Learning Algorithm Pseudocode. The Colored Part in This Pseudocode Highlights the Difference Relative to the SARSA Algorithm Shown in Algo. 2

Inputs: $\mathbb{A}$, $\mathbb{S}$, $\mathbb{R}$, $\mathbb{P}$, and $\mathbb{F}$.
Initialization: Initialize the Q-table with zero values
Begin the Learning Mode
for episode= $\{1, 2, 3, \dots\}$ **do**
  Reset: Reset the environment to initial state, $s_t(t) \in \mathbb{S}_{Tx}$, and $s_i(t) \in \mathbb{S}_{IRS}$.
  if mod(episode index,10)== 0 **then**
    $\epsilon_\chi = \epsilon_\chi + 0.01$
  end if
  for iteration, t=$\{1, 2, 3, \dots\}$ **do**
    Generate random probability p
    if $p \geq \epsilon_\chi$ **then**
      Choose a random action (e.g., Exploration), $a_t(t) \in \mathbb{A}_{Tx}$, and $a_i(t) \in \mathbb{A}_{IRS}$.
    else
      Choose a greedy action (e.g., Exploitation), $a_\chi(t) = \max(Q_\chi(S_\chi(t), a_\chi(t+1)))$.
    end if
    Apply the chosen action, $a_\chi(t)$
    Observe the next state, $s_\chi(t+1)$
    Compute the reward, $r_\chi(t+1)$
    Update Q-value for the current state-action pair:
    $Q_\chi(s_\chi(t), a_\chi(t)) = Q_\chi(s_\chi(t), a_\chi(t)) + \alpha_o \times [$
    $r_\chi(t+1) + \gamma_\chi \times \max(Q_\chi(s_\chi(t+1), a_\chi(t+1))) -$
    $Q_\chi(s_\chi(t), a_\chi(t))]$
    Update the $s_\chi(t)$ to be $s_\chi(t+1)$
  end for
end for
End the Learning Mode
Outputs: Suboptimal policies of the Tx and IRS agents, $\pi_t^*(s_t(t), a_t(t))$ and $\pi_i^*(s_i(t), a_i(t))$, respectively.
Begin the tracking Mode
for Transmitted packet number = $\{1, 2, 3, \dots, \infty\}$ **do**
  if $\pi_t(s_t(t), a_t(t)) \neq \pi_t^*(s_t(t), a_t(t))$ || $\pi_i(s_i(t), a_i(t)) \neq \pi_i^*(s_i(t), a_i(t))$ **then**
    Go to the Learning Mode
  end if
end for
End the tracking Mode

Algorithm 2 SARSA Algorithm Pseudocode. The Colored Part in This Pseudocode Highlights the Difference Relative to the Q-Learning Algorithm Shown in Algo. 1

Inputs: $\mathbb{A}$, $\mathbb{S}$, $\mathbb{R}$, $\mathbb{P}$, and $\mathbb{F}$.
Initialization: Initialize the Q-table with zero values
Begin the Learning Mode
for episode= $\{1, 2, 3, \dots\}$ **do**
  Reset: Reset the environment to initial state, $s_t(t) \in \mathbb{S}_{Tx}$, and $s_i(t) \in \mathbb{S}_{IRS}$.
  if mod(episode index,10)== 0 **then**
    $\epsilon_\chi = \epsilon_\chi + 0.01$
  end if
  Generate random probability p
  if $p \geq \epsilon_\chi$ **then**
    Choose a random action (e.g., Exploration), $a_t(t) \in \mathbb{A}_{Tx}$, and $a_i(t) \in \mathbb{A}_{IRS}$.
  else
    Choose a greedy action (e.g., Exploitation), $a_\chi(t) = \max(Q_\chi(S_\chi(t), a_\chi(t+1)))$.
  end if
  for iteration t=$\{1, 2, 3, \dots\}$ **do**
    Apply the chosen action, $a_\chi(t)$
    Observe the next state, $s_\chi(t+1)$
    Compute the reward, $r_\chi(t+1)$
    Generate random probability p
    if $p \geq \epsilon_\chi$ **then**
      Choose a random action (e.g., Exploration), $a_t(t) \in \mathbb{A}_{Tx}$, and $a_i(t) \in \mathbb{A}_{IRS}$.
    else
      Choose a greedy action (e.g., Exploitation), $a_\chi(t+1) = \max(Q_\chi(S_\chi(t), a_\chi(t+1)))$.
    end if
    Update Q-value for the current state-action pair:
    $Q_\chi(s_\chi(t), a_\chi(t)) = Q_\chi(s_\chi(t), a_\chi(t)) + \alpha_\chi \times [$
    $r_\chi(t+1) + \gamma_\chi \times \max(Q_\chi(s_\chi(t+1), a_\chi(t+1))) -$
    $Q_\chi(s_\chi(t), a_\chi(t))]$
    Update the $s_\chi(t)$ to be $s_\chi(t+1)$
  end for
end for
End the Learning Mode
Outputs: Suboptimal policies of the Tx and IRS agents, $\pi_t^*(s_t(t), a_t(t))$ and $\pi_i^*(s_i(t), a_i(t))$, respectively.
Begin the tracking Mode
for Transmitted packet number = $\{1, 2, 3, \dots, \infty\}$ **do**
  if $\pi_t(s_t(t), a_t(t)) \neq \pi_t^*(s_t(t), a_t(t))$ || $\pi_i(s_i(t), a_i(t)) \neq \pi_i^*(s_i(t), a_i(t))$ **then**
    Go to the Learning Mode
  end if
end for
End the tracking Mode

respectively. The algorithms are color to distinguish different parts, simplifying the comparison process. These algorithms operate in two distinct modes: learning and tracking. During the learning mode (lines 1-26 and 1-34 in Algorithms 1 and 2, respectively), the agents explore solutions through trial and error. The algorithms initiate from random states, execute actions, update states, receive rewards, determine subsequent actions, and repeat this process. Through hundreds of iterations per episode, the agents transition from random to well-developed states. This iterative process continues until the policy converges to a suboptimal value. Rewards are received via feedback links, either equal to the *CSR* or zero. If the outage probability falls below the threshold, the reward is set to the *CSR* value; otherwise, it is set to zero (line

No. 20 in Algorithms 1 and 2). After each iteration, the agents update their policies assisted on the received feedback (lines No. 21 and 29 in Algorithms 1 and 2, respectively). Q-learning updates policies using the maximum expected reward, while SARSA updates policies using actual rewards. Agents take actions to transition to the next state (lines No. 18-20 and 17-21 in Algorithms 1 and 2, respectively), aiming to optimize channel secrecy and reliability. Actions may be random or greedy (i.e., exploitation) based on the exploration probability (lines No. 11-17 and 10-16 in Algorithms 1 and 2, respectively).

Over numerous episodes and iterations, agents learn which actions maximize channel secrecy and reliability, gradually converging towards suboptimal policies. In the tracking mode (lines No. 27-33 and 35-41 in Algorithms 1 and 2, respectively), agents apply their learned policies to maximize channel secrecy and reliability performance, autonomously adapting to the environment. The algorithms re-initiate training only when significant environmental changes occur, indicated by discrepancies between current and optimal policies. Significant environmental changes necessitate significant training time to converge the policy again to a suboptimal value, while minor changes may not require retraining. To ensure efficiency, the learning mode interval should be shorter than the tracking mode to minimize the energy consumption and communication interruptions.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Eng. Verdier Assoume for his valuable technical discussions, constructive feedback, and insightful comments.

REFERENCES

[1] M. Jouhari, K. Ibrahimi, H. Tembine, and J. Ben-Othman, "Underwater wireless sensor networks: A survey on enabling technologies, localization protocols, and Internet of Underwater Things," *IEEE Access*, vol. 7, pp. 96879–96899, 2019.

[2] A. S. Ghazy, "Reliable high-speed short-range underwater wireless optical communication systems," Ph.D. thesis, Electrical and Computer Engineering School, McMaster University, Hamilton, ON, Canada, 2021.

[3] M. T. Anowar, M. H. Khan, M. Alam, M. Kabir, M. Hossen, S. Zahan, M. Hossain, and M. M. Hasan, "A survey of acoustic underwater communications and ways of mitigating security challenges," *Int. J. Res. Eng. Sci. (IJRES)*, vol. 4, no. 6, pp. 43–51, 2016.

[4] S. Jiang, "On securing underwater acoustic networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 729–752, 1st Quart., 2019.

[5] W. Aman, S. Al-Kuwari, M. Muzzammil, M. M. Ur Rahman, and A. Kumar, "Security of underwater and air—Water wireless communication: State-of-the-art, challenges and outlook," *Ad Hoc Netw.*, vol. 142, 2023, Art. no. 103114, doi: 10.1016/j.adhoc.2023.103114. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570870523000343

[6] M. Stojanovic and J. Preisig, "Underwater acoustic communication channels: Propagation models and statistical characterization," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 84–89, Jan. 2009.

[7] A. S. Ghazy, G. Kaddoum, and S. Singh, "Low-latency low-energy adaptive clustering hierarchy protocols for underwater acoustic networks," *IEEE Access*, vol. 11, pp. 50578–50594, 2023.

[8] F. Qu, Z. Wang, L. Yang, and Z. Wu, "A journey toward modeling and resolving Doppler in underwater acoustic communications," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 49–55, Feb. 2016.

[9] A. S. Ghazy, H. S. Khallaf, S. Hranilovic, and M.-A. Khalighi, "Under-sea ice diffusing optical communications," *IEEE Access*, vol. 9, pp. 159652–159671, 2021.

[10] A. S. Ghazy, S. Hranilovic, and M.-A. Khalighi, "Angular MIMO for underwater wireless optical communications: Channel modelling and capacity," in *Proc. 16th Can. Workshop Inf. Theory (CWIT)*, Jun. 2019, pp. 1–6.

[11] G. Dini and A. L. Duca, "A secure communication suite for underwater acoustic sensor networks," *Sensors*, vol. 12, no. 11, pp. 15133–15158, Nov. 2012.

[12] S. Yang, Z. Guo, Q. Ren, and S. Guo, "A covert underwater acoustic communication method based on spread spectrum digital watermarking," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, p. 3230, Oct. 2016.

[13] M. Xu, G. Liu, D. Zhu, and H. Wu, "A cluster-based secure synchronization protocol for underwater wireless sensor networks," *Int. J. Distrib. Sensor Netw.*, vol. 10, no. 4, pp. 1–13, Apr. 2014.

[14] G. Dini and I. Savino, "S2RP: A secure and scalable rekeying protocol for wireless sensor networks," in *Proc. IEEE Int. Conf. Mobile Ad Hoc Sensor Syset.*, Oct. 2006, pp. 457–466.

[15] H. U. Yildiz, "Maximization of underwater sensor networks lifetime via fountain codes," *IEEE Trans. Ind. Informat.*, vol. 15, no. 8, pp. 4602–4613, Aug. 2019.

[16] P. Casari, M. Rossi, and M. Zorzi, "Towards optimal broadcasting policies for HARQ based on fountain codes in underwater networks," in *Proc. 5th Annu. Conf. Wireless Demand Netw. Syst. Services*, Jan. 2008, pp. 11–19.

[17] K. Pelekanakis, L. Cazzanti, G. Zappa, and J. Alves, "Decision tree-based adaptive modulation for underwater acoustic communications," in *Proc. IEEE 3rd Underwater Commun. Netw. Conf.*, Aug. 2016, pp. 1–16.

[18] J. Li, J. Halt, and Y. R. Zheng, "Utilizing JANUS for very high frequency underwater acoustic modem," in *Proc. Global Oceans*, Singapore, Oct. 2020, pp. 1–6.

[19] A. Aminjavaheri and B. Farhang-Boroujeny, "UWA massive MIMO communications," in *Proc. OCEANS-MTS/IEEE Washington*, Oct. 2015, pp. 1–6.

[20] K. Pelekanakis and A. B. Baggeroer, "Exploiting space–time–frequency diversity with MIMO–OFDM for underwater acoustic communications," *IEEE J. Ocean. Eng.*, vol. 36, no. 4, pp. 502–513, Oct. 2011.

[21] L. Shen, B. Henson, Y. Zakharov, and P. Mitchell, "Digital self-interference cancellation for full-duplex underwater acoustic systems," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 1, pp. 192–196, Jan. 2020.

[22] M. Stojanovic, J. G. Proarkis, J. A. Rice, and M. D. Green, "Spread spectrum underwater acoustic telemetry," in *Proc. IEEE Ocean. Eng. Society. Conf.*, vol. 2, Jun. 1998, pp. 650–654.

[23] Y. Zhou, J. Shi, J. Zhang, and N. Chi, "Spectral scrambling for high-security PAM-8 underwater visible light communication system," in *Proc. Asia Commun. Photon. Conf. (ACP)*, Oct. 2018, pp. 1–3.

[24] C. Wang and Z. Wang, "Signal alignment for secure underwater coordinated multipoint transmissions," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6360–6374, Dec. 2016.

[25] A. Zhao, Y. Cheng, T. An, and J. Hui, "Covert underwater acoustic communication system using parametric array," *Mar. Technol. Soc. J.*, vol. 53, no. 1, pp. 20–26, Jan. 2019.

[26] H. Yan, Z. Shi, and J. Cui, "DBR: Depth-based routing for underwater sensor networks," in *Proc. Netw. Ad Hoc Sensor Netw.*, A. Das, H. K. Pung, F. B. S. Lee, and L. W. C. Wong, Eds., Berlin, Germany: Springer, 2008, pp. 72–86.

[27] P. Xie, J. Cui, and L. Lao, "VBF: Vector-based forwarding protocol for underwater sensor networks," in *NETWORKING 2006. Networking Technologies, Services, Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems*, F. Boavida, T. Plagemann, B. Stiller, C. Westphal, and E. Monteiro, Eds., Berlin, Germany: Springer, 2006, pp. 1216–1221.

[28] I. Romdhane and G. Kaddoum, "A reinforcement-learning-based beam adaptation for underwater optical wireless communications," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 20270–20281, Oct. 2022.

[29] Z. Liu, F. Yang, S. Sun, J. Song, and Z. Han, "Physical layer security in NOMA-based VLC systems with optical intelligent reflecting surface: A max-min secrecy data rate perspective," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 7180–7194, Mar. 2025.

[30] H. Jin, Z. Li, J. Yuan, and Y. Xia, "Improper Gaussian signaling for secrecy transmission in RIS-aided downlink NOMA," *IEEE Commun. Lett.*, vol. 29, no. 1, pp. 50–54, Jan. 2025.
[31] A. B. Sarawar, A. S. M. Badrudduza, M. Ibrahim, I. S. Ansari, and H. Yu, "Secrecy performance analysis of integrated RF-UOWC IoT networks enabled by UAV and underwater RIS," *IEEE Internet Things J.*, vol. 12, no. 3, pp. 2592–2608, Feb. 2025.
[32] N. Adam, M. Ali, F. Naeem, A. S. Ghazy, and G. Kaddoum, "State-of-the-art security schemes for the Internet of Underwater Things: A holistic survey," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 6561–6592, 2024.
[33] A. S. Ghazy, G. Kaddoum, and S. Satinder, "IRS-aided secure reliable underwater acoustic communications," *IEEE Trans. Veh. Technol.*, vol. 73, no. 11, pp. 16861–16875, Nov. 2024.
[34] T. Hu and Y. Fei, "An adaptive and energy-efficient routing protocol based on machine learning for underwater delay tolerant networks," in *Proc. IEEE Int. Symp. Model., Anal. Simul. Comput. Telecommun. Syst.*, Aug. 2010, pp. 381–384.
[35] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[36] Z. Huang and S. Wang, "Multilink and AUV-assisted energy-efficient underwater emergency communications," *IEEE Internet Things J.*, vol. 10, no. 9, pp. 8068–8082, May 2023.
[37] M. Zhang, X. Ding, Y. Tang, S. Wu, and K. Xu, "STAR-RIS assisted secrecy communication with deep reinforcement learning," *IEEE Trans. Green Commun. Netw.*, vol. 9, no. 2, pp. 739–753, Jun. 2025.
[38] Z. Sun, H. Guo, P. Wang, and I. F. Akyildiz, "Acoustic intelligent surface system for reliable and efficient underwater communications," in *Proc. 15th Int. Conf. Underwater Netw. Syst.*, New York, NY, USA, Nov. 2021, pp. 1–8.
[39] Z. Sun, H. Guo, and I. F. Akyildiz, "High-data-rate long-range underwater communications via acoustic reconfigurable intelligent surfaces," *IEEE Commun. Mag.*, vol. 60, no. 10, pp. 96–102, Oct. 2022.
[40] R. P. Naik and W.-Y. Chung, "Evaluation of reconfigurable intelligent surface-assisted underwater wireless optical communication system," *J. Lightw. Technol.*, vol. 40, no. 13, pp. 4257–4267, Jul. 15, 2022.
[41] Y. Ata, H. Abumarshoud, L. Bariah, S. Muhaidat, and M. A. Imran, "Intelligent reflecting surfaces for underwater visible light communications," *IEEE Photon. J.*, vol. 15, no. 1, pp. 1–10, Feb. 2023.
[42] R. Salam, A. Srivastava, V. A. Bohara, and A. Ashok, "An optical intelligent reflecting surface-assisted underwater wireless communication system," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1774–1786, 2023.
[43] Y. Ata, M. C. Gökçe, and Y. Baykal, "Intelligent reflecting surface aided vehicular optical wireless communication systems using higher-order mode in underwater channel," *IEEE Trans. Veh. Technol.*, vol. 73, no. 8, pp. 1–13, Aug. 2024.
[44] M. Shahwar, M. Ahmed, T. Hussain, S. Ahmad, W. Ullah Khan, M. Sheraz, and T. Chee Chuah, "Terahertz-based IRS-assisted secure symbiotic radio communication: A DRL approach," *IEEE Access*, vol. 13, pp. 24014–24027, 2025.
[45] J. Smith, "NOAA tide predictions," National Oceanic and Atmospheric Administration, Washington, DC, USA, Tech. Rep., Jun. 2023.
[46] R. H. Stewart, *Introduction to Physical Oceanography*. College Station, TX, USA: Texas A&M University, 2008.
[47] B. Assouar, B. Liang, Y. Wu, Y. Li, J. Cheng, and Y. Jing, "Acoustic metasurfaces," *Nature Rev. Mater.*, vol. 3, no. 12, pp. 460–472, 2018.
[48] H. Wang, Z. Sun, H. Guo, P. Wang, and I. F. Akyildiz, "Designing acoustic reconfigurable intelligent surface for underwater communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 8934–8948, Dec. 2023.
[49] A. S. Ghazy, S. Hranilovic, and M.-A. Khalighi, "Angular MIMO for underwater wireless optical communications: Link modeling and tracking," *IEEE J. Ocean. Eng.*, vol. 46, no. 4, pp. 1391–1407, Oct. 2021.
[50] H. George, P. Brian, B. Roger, and P. Mirko, "Methodology for estimating tidal current energy resources and power production by tidal in-stream energy conversion (TISEC) devices," EPRI, Oslo, Norway, Tech. Rep., 2006.
[51] Y. Tajima, T. Hiraguri, T. Matsuda, T. Imai, J. Hirokawa, H. Shimizu, T. Kimura, and K. Maruta, "Analysis of wind effect on drone relay communications," *Drones*, vol. 7, no. 3, p. 182, Mar. 2023.
[52] C. Cox and W. Munk, "Measurement of the roughness of the sea surface from photographs of the Sun's glitter," *J. Opt. Soc. Amer.*, vol. 44, no. 11, p. 838, Nov. 1954.
[53] *Self-elevating Units*, document DNV-RP-C104, DET NORSKE VERITAS (DNV), Oslo, Norway, Nov. 2012.
[54] M. Sailing. (2023). *The Fastest Tidal Current in the World*. Accessed: Jun. 11, 2023. [Online]. Available: https://www.youtube.com/watch?v=x8oRRFfP9Ic
[55] S. K. Moorthy and Z. Guan, "Beam learning in mmWave/THz-band drone networks under in-flight mobility uncertainties," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 1945–1957, Jun. 2022.
[56] J. Smith, "Meteorological conversions and calculations," National Oceanic and Atmospheric Administration, Washington, DC, USA, Tech. Rep. EPRI-TP-001 NA RV 2, 2021.
[57] L. Wu, M. Shao, and E. Sahlée, "Impact of air–wave–sea coupling on the simulation of offshore wind and wave energy potentials," *Atmosphere*, vol. 11, no. 4, p. 327, Mar. 2020. [Online]. Available: https://www.mdpi.com/2073-4433/11/4/327
[58] M. Stojanovic, "On the relationship between capacity and distance in an underwater acoustic communication channel," in *Proc. 1st ACM Int. workshop Underwater Netw.*, New York, NY, USA, 2006, p. 41, doi: 10.1145/1161039.1161049.
[59] L. Brekhovskikh and Y. Lysanov, *Fundamentals of Ocean Acoustics* (Springer Series in Electronics and Photonics). Heidelberg, Germany: Springer, 1982.
[60] A. A. Farid and S. Hranilovic, "Outage capacity optimization for free-space optical links with pointing errors," *J. Lightw. Technol.*, vol. 25, no. 7, pp. 1702–1710, Jul. 15, 2007.
[61] R. Coates, *Underwater Acoustic Systems*. Cham, Switzerland: Springer, 1989.
[62] T. Hu and Y. Fei, "QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 6, pp. 796–809, Jun. 2010.
[63] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1057–1063.
[64] M. Chitre, S. Shahabudeen, and M. Stojanovic, "Underwater acoustic communications and networking: Recent advances and future challenges," *Mar. Technol. Soc. J.*, vol. 42, no. 1, pp. 103–116, Mar. 2008.
[65] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.
[66] J. Zhang, Z. Wang, J. Li, Q. Wu, W. Chen, F. Shu, and S. Jin, "How often channel estimation is required for adaptive IRS beamforming: A bilevel deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8744–8759, Aug. 2024.
[67] N. Mastronarde and M. van der Schaar, "Fast reinforcement learning for energy-efficient wireless communication," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6262–6266, Dec. 2011.
[68] L. Freitag, M. Grund, S. Singh, J. Partan, P. Koski, and K. Ball, "The WHOI micro-modem: An acoustic communications and navigation system for multiple platforms," in *Proc. OCEANS MTS/IEEE*, vol. 2, Jul. 2005, pp. 1086–1092.

ABDALLAH S. GHAZY was born in Giza, Egypt, in 1983. He received the B.Sc. degree in electrical engineering from Al-Azhar University, Egypt, in 2007, the M.Sc. degree in electrical engineering from the Electrical and Computer Engineering School, Egypt-Japan University for Science and Technology, Alexandria, Egypt, in 2016, and the Ph.D. degree in electrical engineering from the Electrical and Computer Engineering School, McMaster University, ON, Canada, in 2021. From 2008 to 2017, he worked at Unified Communication Systems and Networks (UCSN), Telecom, Egypt, and AVAYA Inc., Cairo, Egypt. In 2012, he joined the Electrical Engineering School, Al-Azhar University. His research interests include communication systems and networks, signal processing, and machine learning.

GEORGES KADDOUM (Senior Member, IEEE) received the bachelor's degree in electrical engineering from the École Nationale Supérieure de Techniques Avancées (ENSTA Bretagne), Brest, France, the M.S. degree in telecommunications and signal processing (circuits, systems, and signal processing) from the Université de Bretagne Occidentale and Telecom Bretagne (ENSTB), Brest, in 2005, and the Ph.D. degree (Hons.) in signal processing and telecommunications from the National Institute of Applied Sciences (INSA), University of Toulouse, Toulouse, France, in 2009. He is currently an Associate Professor and a Tier 2 Canada Research Chair with the École de Technologie Supérieure (ÉTS), Université du Québec, Montréal, QC, Canada. Since 2010, he has been a Scientific Consultant in the field of space and wireless telecommunications for several U.S. and Canadian companies. He has published more than 200 journals and conference papers and has two pending patents. His research interests include mobile communication systems, modulations, security, and space communications and navigation. He received the Best Papers Award at the 2014 IEEE International Conference on Wireless and Mobile Computing, Networking, Communications (WIMOB), with three co-authors, and at the 2017 IEEE International Symposium on Personal Indoor and Mobile Radio Communications, with four co-authors. In 2014, he was awarded the ÉTS Research Chair in physical-layer security for wireless networks. Moreover, he received the IEEE Transactions on Communications Exemplary Reviewer Award for the years 2015, 2017, and 2019. In addition, he received the Research Excellence Award of the Université du Québec, in 2018. In 2019, he received the Research Excellence Award from the ÉTS in recognition of his outstanding research outcomes. He is serving as an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE COMMUNICATIONS LETTERS.

CHAMESEDDINE TALHI received the Ph.D. degree in computer science from Laval University, Québec City, QC, Canada, in 2007. He is currently an Associate Professor with the Department of Software Engineering and IT, ÉTS, University of Quebec, Montreal, QC, Canada. He is leading a research group that investigates smartphones, embedded systems, and the IoT security. His research interests include cloud security and secure sharing of embedded systems.

NAVEED IQBAL (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, and the Ph.D. degree from the King Fahd University of Petroleum and Minerals, Saudi Arabia. He is currently an Assistant Professor with the King Fahd University of Petroleum and Minerals. His research interests include adaptive algorithms, compressive sensing, heuristic algorithms, signal/image processing, seismic processing, machine learning, and data acquisition networks.

ALI HUSSEIN MUQAIBEL (Senior Member, IEEE) received the Ph.D. degree from Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, in 2003. He is currently a Professor with the Electrical Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM). He is also the Director of the Interdisciplinary Research Center for Communication Systems and Sensing (IRC-CSS). During his study at Virginia Tech, he was with the Time Domain and RF Measurements Laboratory and the Mobile and Portable Radio Research Group. He was a Visiting Associate Professor with the Center of Advanced Communications, Villanova University, Villanova, PA, USA, in 2013; a Visiting Professor with the Georgia Institute of Technology, in 2015; and a Visiting Scholar with the King Abdullah University for Science and Technology (KAUST), Thuwal, Saudi Arabia, in 2018 and 2019. He has authored three book chapters and over 170 articles. His research interests include communications and sensing applications, including the direction of arrival estimation, through-wall imaging, localization, channel characterization, and ultra-wideband signal processing. He was a recipient of many awards for excellence in teaching, advising, and instructional technology.