Robust Beamforming for Massive MIMO LEO Satellite Communications: A Risk-Aware Learning Framework

Madyan Alsenwi, *Member, IEEE*, Eva Lagunas, *Senior Member, IEEE*, Symeon Chatzinotas, *Fellow, IEEE*

***Abstract*—This paper proposes a robust beamforming algorithm for massive multiple-input multiple-output (MIMO) low earth-orbit (LEO) satellite communications under uncertain channel conditions. Specifically, a risk-aware optimization problem is formulated to optimize the hybrid digital and analog precoding aiming at maximizing the energy efficiency of the LEO satellite while considering the required quality-of-service (QoS) by each ground user. The Conditional Value at Risk (CVaR) is used as a risk measure of the downlink data rate to capture the high dynamic and random channel variations due to satellite movement, achieving the required QoS under the worst-case scenario. A deep reinforcement learning (DRL) based framework is developed to solve the formulated stochastic problem over time slots. Considering the limited computation capabilities of the LEO satellite, the training process of the proposed learning algorithm is performed offline at a central terrestrial server. The trained models are then sent periodically to the LEO satellite through ground stations to provide online executions on the transmit precoding based on the current network state. Simulation results demonstrate the efficacy of the proposed approach in achieving the QoS requirements under uncertain wireless channel conditions.**

***Index Terms*—6G, LEO satellites communication, NTN, massive MIMO, risk-aware learning, DRL, digital precoding, analog beamforming.**

I. Introduction

Future wireless networks are expected to provide ubiquitous uninterrupted connectivity to everyone, everywhere, and everything with extremely high reliability, ultrahigh data rate, and low latency. However, deploying terrestrial base stations (BSs) with wired backhaul infrastructure in remote areas with sparse users is expensive and unprofitable. These limitations in terrestrial wireless systems drive the vision towards non-terrestrial networks (NTNs) as an enabler technology in next-generation wireless communications [1], [2]. In particular, NTNs can overcome several limitations in traditional terrestrial networks, such as coverage holes and the surge in throughput demands. In this regard, the 3rd generation partnership project (3GPP) Release 15 and Release 16 have included several studies to support NTNs in 5G and beyond [3], [4]. Among NTNs technologies, Low earth orbit (LEO) satellite systems are a promising candidate to satisfy future wireless network requirements in terms of global coverage and ubiquitous connectivity. In particular, LEO satellites are typically deployed at $500-2000$ km from Earth, allowing more focused beams and achieving faster communications and less signal attenuation than the conventional geostationary (GSO) satellite systems. Several LEO satellite constellations, including Starlink, Telesat, and OneWeb, have been launched recently to provide seamless and high-capacity wireless services.

Multiple terrestrial communication technologies have been developed in recent years to enhance wireless connectivity. Massive multiple-input multiple-output (MIMO) transmission is a promising 5G technology that can provide high-performance gains and enhanced coverage. Numerous studies on massive MIMO in terrestrial wireless networks have been recently conducted [5]. However, massive MIMO in satellite communication systems is still in the relatively early stages of research. In this work, we exploit massive MIMO technology for LEO satellite wireless communication systems by equipping the LEO satellite with large antenna arrays. In particular, most LEO satellite systems are envisaged to operate in the Ku and Ka bands, i.e., the mmWave band in the terrestrial networks. Such high frequencies allow reducing antenna size significantly, packing more antennas in a small space.

The massive MIMO technique can provide high spectral efficiency due to the availability of multiple degrees of freedom in the spatial domain, making it an optimistic approach for future satellite communication systems [6]–[9]. However, the performance of the massive MIMO depends on the availability of up-to-date channel state information (CSI), which is hard and even infeasible to obtain in LEO satellite systems [10]. In practice, the fast movement of LEO satellites relative to ground users causes outdated CSI estimations[1]. Therefore, the inclusion of massive MIMO in LEO communication systems is a challenging research direction. More specifically, the fundamental challenge is how to design reliable precoding that can ensure link quality in uncertain and dynamic wireless environments with outdated CSI estimations.

This work has been supported by the project TRANTOR, which has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101081983. The working time of E. Lagunas has been supported by the Luxembourg National Research Fund (FNR) under the project SmartSpace (C21/IS/16193290).

Madyan Alsenwi, Eva Lagunas, and Symeon Chatzinotas are with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg, L-1855, Luxembourg (e-mail: madyan.alsenwi@uni.lu, eva.lagunas@uni.lu, symeon.chatzinotas@uni.lu).

[1]The issue of outdated channel estimations is particularly acute in LEO communication systems due to the high speed and mobility of the satellites. LEO satellites typically have a velocity of approximately 7.8 km/s, which means they complete one orbit every 90 minutes if they orbit at an altitude of 1000 km [11]. This fast movement can cause rapid variations in the channel conditions, which result in outdated channel estimations if the ground station is not able to keep up with the changes.

The conventional digital precoding approach leads to high power consumption as many radio frequency chains are required due to the numerous antennas in the massive MIMO design. To reduce power consumption in massive MIMO, the hybrid precoding architecture has been proposed by leveraging the low-dimension digital precoding and analog precoding technology [12]. The hybrid precoding design in terrestrial networks has been well investigated in several recent studies [13]–[16]. For instance, the work in [13] studied the hybrid beamforming design in millimeter wave communication systems. The authors in [14] proposed a hybrid beamforming framework for massive MIMO wireless systems considering accurate channel state information estimation. Another study in [15] proposed a hybrid beamforming and user-beam selection approach in massive MIMO networks based on machine learning aiming at improving the system's energy efficiency. The work in [16] studies the hybrid precoding in a mixed design of fully and partially connected architectures. Specifically, the authors proposed a sorted successive interference cancellation-based method for analog beamforming. Furthermore, a modified baseband block-diagonalization approach is proposed for digital beamforming to reduce information loss. In [17], the authors studied the hybrid beamforming in multi-user massive MIMO mmWave communication systems. The Zero-forcing (ZF) technique is used to design the transmit hyper-beamforming considering both fully and partially connected approaches. Another work in [18] proposed a two-stage hybrid analog-digital precoding approach for massive MIMO networks aiming to maximize energy efficiency and spectral efficiency.

Some recent works have studied the transmit design of massive MIMO LEO satellite communications [6]–[8], [19]–[21]. The work in [6] exploits statistical CSI for massive MIMO LEO satellite communications. Based on the statistical CSI, the authors developed a low-complexity uplink receiver and downlink precoder closed-form. In [8], the authors proposed different network architectures for distributed massive MIMO LEO satellite systems. A joint power and handover management approach was proposed to optimize the joint power allocation and handover. Furthermore, artificial intelligence technology was leveraged to solve the joint power allocation and handover optimization problem. The work in [9] analyzed the throughput of multi-user MIMO multibeam LEO satellite systems with four-color frequency reuse. A deterministic model was used for the spacecraft motion and channel state information generation. The study in [19] proposed a long short-term memory (LSTM) based approach to predict the CSI in massive MIMO LEO satellite systems by exploiting channel change correlation. Another work in [20] formulated an energy efficiency maximization problem to optimize the hybrid analog and digital precoding in massive LEO satellite communication systems. The authors used the iteratively weighted minimum mean-square error technique and the Dinkelbach approach to obtain the digital precoding, while an alternating optimization approach was used to calculate the hybrid precoders. In [21], the authors formulated an energy efficiency optimization problem to optimize the hybrid precoding, i.e., analog/digital, in massive MIMO LEO satellites. A closed-form tight upper-bound approach is applied to obtain an approximate data rate. The authors then proposed an algorithm to calculate the hybrid precoding considering fully and partially connected architecture scenarios.

The work in [22] focused on developing secrecy-energy efficient hybrid beamforming strategies for integrated satellite-terrestrial networks. The primary objective was the simultaneous formulation of hybrid beamforming at the BS and digital beamformers at the satellite, aiming to enhance secrecy-energy efficiency. Meanwhile, the study in [23] concentrated on elevating transmission security and lowering energy usage by optimizing system secrecy energy efficiency within the limits of a total transmit power budget. The authors introduced an alternating optimization approach to address the inherent non-convexity of the problem. In [24], the focus was on physical layer security within a satellite network that shares downlink spectral resources with a terrestrial cellular network. This study presented two beamforming techniques, namely hybrid zero-forcing and partial zero-forcing, to address the optimization challenge and ascertain the beamforming weight vectors. Additionally, the study evaluated the secrecy performance of the satellite network in two different scenarios, each based on varying assumptions about the eavesdroppers' channel state information. The authors of [25] examined the application of Reconfigurable Intelligent Surfaces (RIS) in creating a flexible radio environment and enhancing the received signal power through intelligent coordination of the phase shifts of passive elements at the RIS. The paper revolved around collaborative design and optimization of beamforming in RIS-assisted satellite-terrestrial relay networks, addressing scenarios where the connections from the satellite and BS to users are obstructed.

Unlike the previous works, this paper proposes a risk-ware robust precoding approach for massive MIMO LEO communication systems that considers the random and dynamic behavior of wireless channels. Specifically, we formulate a stochastic optimization problem based on the Conditional Value at Risk (CVaR) to optimize the digital and analog precoding at the LEO satellite side. The formulated problem aims to improve the energy efficiency of the LEO satellite considering the required transmission reliability. In other words, the provided CVaR-based optimization problem can ensure achieving the required data rate by each ground user under the worst-case scenario of the wireless channels. To solve the formulated problem, we propose a practical online learning framework that considers the limited computation capability of the LEO satellite by providing an offline training process at a terrestrial cloud server. The central server broadcasts the trained models to the associated ground stations, which forward them to the LEO satellite to perform online decisions based on the current network states. To our knowledge, this work is the first to propose a risk-aware two-stage online learning framework based on the CVaR and DRL for massive MIMO LEO satellite communication. To sum up, the main contributions of this work are summarized as follows:

- We formulate a risk-aware stochastic optimization problem to obtain robust beamforming for massive MIMO LEO satellite communications under dynamic wireless

variations and outdated CSI estimations. The formulated problem aims to maximize the energy efficiency of the LEO satellite while meeting the transmission reliability required by ground users. The CVaR is used as a risk measure to ensure meeting the minimum data rate requirements in the worst-case conditions of wireless channels. Unlike the average-based formulation, CVaR can quantify the potential loss in the tail of a distribution of possible data rate returns.

- We design a two-stage learning framework based on the DRL technique to obtain dynamic solutions over time slots to the optimization problem according to the instantaneous network states. The proposed learning framework composes of a training stage and an execution stage. The training process is performed offline at a central terrestrial server due to the limited processing capability of the LEO satellite. The central server periodically transmits the trained models to the LEO satellite through ground stations located in different geographical areas. The LEO satellite observes the current network information and makes online decisions on the digital and analog precoding matrices based on the received trained models. Then, the LEO satellite sends the network information to the ground sever to improve learning accuracy further.
- We conduct extensive simulation and computation complexity analysis to evaluate the performance of the proposed scheme. The obtained results verify the ability of the proposed algorithm to provide online decisions on the digital and analog beamforming while ensuring the required QoS by the ground users in the presence of uncertain channel changes.

A. Structure of the Paper

The subsequent sections of this paper are structured as follows: Section II outlines the system model under consideration and the formulation of the problem, focusing on the introduced CVaR-based risk-aware methodology. Section III details the proposed two-stage learning framework and presents the computational complexity of the algorithm. Section V delves into assessing the performance of the proposed methodology and conducts comparative evaluations with analogous methods.

II. System Model and Problem Formulation

We consider a LEO satellite serving a set $\mathcal{K} = \{1, 2, \ldots, K\}$ of ground users over the Ka-band, as depicted in Fig. 1. The LEO satellite connects with a central terrestrial data server via ground gateway stations distributed in different locations due to satellite movement. The satellite covers a region of interest using a set of potential beams following the scenarios in the 3GPP technical reports on solutions for NR to support non-terrestrial networks [3]. Uniform planar arrays (UPAs) of active digital antennas are installed at the satellite and ground users, where the phase of each antenna element can be digitally controlled [26]. Specifically, the LEO satellite is equipped with a large-scale UPA composed of $N_s = N_s^x \times N_s^y$ antenna elements where N_s^x is the number of elements on the x-axis, and N_s^y is the number of elements on the y-axis.

Table I: Summary of Abbreviations

Abbreviation	Definition
LEO	Low Earth-Orbit.
DRL	Deep Reinforcement Learning.
NTN	Non-Terrestrial Networks.
3GPP	3rd Generation Partnership Project.
MIMO	Multiple-Input Multiple-Output.
UPA	Uniform Planar Array.
CSI	Channel State Information.
QoS	Quality-of-Service.
GSO	Geostationary.
BS	Base Station.
VaR	Value-at-Risk.
CVaR	Conditional Value-at-Risk.
CDF	Cumulative Distribution Function.
PG-DRL	Policy Gradient DRL.
DNN	Deep Neural Network.
ZF	Zero-Forcing.
LSTM	Long Short-Term Memory.

Table II: Summary of Notations

Notation	Definition
$\mathcal{K}$	Set of ground users.
N_s^x, N_s^y	Number of antenna elements in $x-$axis and $y-$axis at the satellite, respectively.
$N_u^{x'}, N_u^{y'}$	Number of antenna elements in $x'-$axis and $y'-$axis at users side, respectively
$\boldsymbol{H}_k(t,f)$	Channel model between satellite and user k at time slot t over frequency f.
$g_{k,l}(t)$	Channel gain between the satellite and user k.
$\boldsymbol{y}_k(t)$	Received signal at user k at time slot t.
$\gamma_k(t)$	data rate of user k at time slot t.
$\eta(t)$	Energy efficiency of the LEO satellite.
$\boldsymbol{s}(t)$	State set at time slot t
$\boldsymbol{a}(t)$	Action set at time slot t.
$R(t)$	The instantaneous reward at time slot t.
$Q(s,a)$	State-action function.
$\boldsymbol{\mu}_{k,l}$	Array response vector of user k.
M	Number of RF chains in the LEO satellite.
$\boldsymbol{x}_k$	Transmit symbols to user k.
$\boldsymbol{W}$	Digital precoding matrix.
$\boldsymbol{V}$	Analog precoding matrix.
B	Downlink bandwidth.
τ	Outdated CSI period.
$\gamma_k^{\min}$	Minimum required data rate by user k.
$P_{\max}$	Maximum transmit power by the satellite.
β	Weighting parameter.
ζ	Discount factor.
$\boldsymbol{\pi}$	Beamforming policy.

In the considered scenario, the x-axis is the direction of the satellite movement, while the y-axis is the orthogonal direction of the movement. A UPA consists of $N_u = N_u^{x'} \times N_u^{y'}$ omnidirectional elements is installed at each ground user in the x$'$-axis and y$'$-axis. Table I and Table II summarize the key abbreviations and notations used in this paper.

A. Communication Model

We adopt the multi-path channel model for the downlink LEO communication system. Using the ray-tracing approach, the downlink channel response $\boldsymbol{H}_k(t,f)$ between the satellite

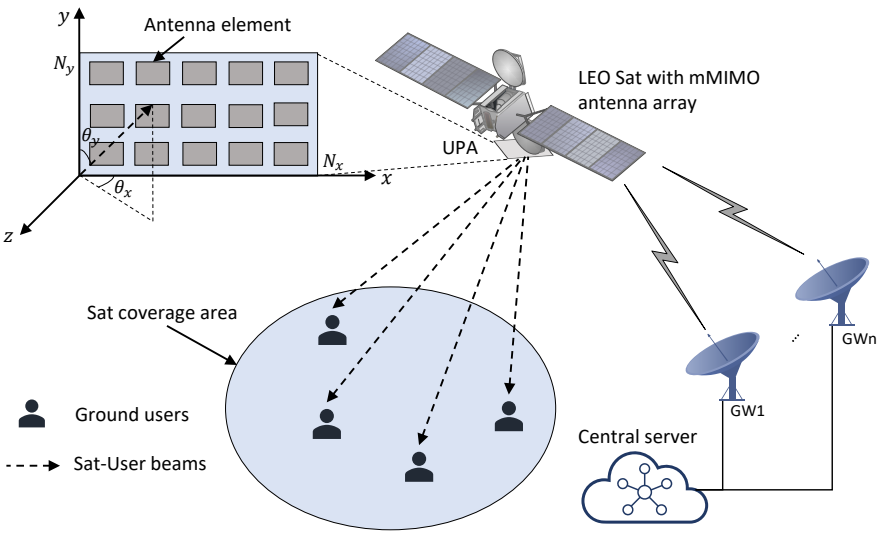

Figure 1: Considered system model

and the k^{th} ground user at time t and frequency f can be defined as [6], [27]

$$\mathbf{H}_k(t,f) = \sum_{l=0}^{L_k-1} g_{k,l}(t,f)\cdot\exp\left\{\bar{\jmath}2\pi\left[\nu_{k,l} - f\tau_{k,l}\right]\right\}\cdot\boldsymbol{\mu}_{k,l}, \quad (1)$$

where L_k denotes the number of propagation paths of the ground user k, and $g_{k,l}$ denotes the channel gain between user k and the LEO satellite. $\nu_{k,l}$ and $\tau_{k,l}$ are the Doppler shift and the propagation delay, respectively. $\boldsymbol{\mu}_{k,l}$ represents the array response vector associated with the l^{th} path of the k^{th} user which is expressed as follows:

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^{\text{x}} \otimes \boldsymbol{\mu}_k^{\text{y}} = \boldsymbol{\mu}_{\text{x}}\left(\phi_k^{\text{x}}\right) \otimes \boldsymbol{\mu}_{\mathbf{y}}\left(\phi_k^{\text{y}}\right), \quad (2)$$

where $\boldsymbol{\mu}_k^{\text{x}}$ and $\boldsymbol{\mu}_k^{\text{y}}$ denote the response vector with angles measured from the x-axis and y-axis, and defined as

$$\boldsymbol{\mu}_k^{\text{x}} = \frac{1}{\sqrt{N_{\text{s}}^x}}\left[1\exp\left\{-\jmath\pi\phi_k^{\text{x}}\right\}\cdots\exp\left\{-\jmath\pi\left(N_{\text{s}}^x-1\right)\phi_k^{\text{x}}\right\}\right]^T, \quad (3)$$

$$\boldsymbol{\mu}_k^{\text{y}} = \frac{1}{\sqrt{N_{\text{s}}^y}}\left[1\exp\left\{-\jmath\pi\phi_k^{\text{y}}\right\}\cdots\exp\left\{-\jmath\pi\left(N_{\text{s}}^y-1\right)\phi_k^{\text{y}}\right\}\right]^T. \quad (4)$$

In the spatial domain, the wireless channel's propagation characteristics hinge on the parameters ϕ_k^{x} and ϕ_k^{y}, defined as $\phi_k^{\text{x}} = \sin\theta_k^{\text{y}}\cos\theta_k^{\text{x}}$ and $\phi_k^{\text{y}} = \cos\theta_k^{\text{x}}$, where θ_k^{x} and θ_k^{y} represent the angles relative to the x-axis and y-axis for the propagation paths of user k, respectively. It is important to note that, due to the significantly higher altitude of LEO satellites compared to the scatterers near the ground users, the propagation paths from the satellite to an individual user can be assumed to be parallel. This assumption allows for the simplification that all propagation paths for a given user have the same physical angles, i.e., $\theta_{k,l}^{\text{x}} = \theta_k^{\text{x}}$ and $\theta_{k,l}^{\text{y}} = \theta_k^{\text{y}}$ [6]. Additionally, LEO satellite systems predominantly operate under Line-of-Sight (LoS) conditions [28]. Therefore, the term $g_k(t,f)$ in this context can be modeled using a Rician fading distribution, characterized by the factor υ_k and power $\mathbb{E}\left\{\left|g_k(t,f)\right|^2\right\}$. In this work, we consider the network dynamic, i.e., the movement of the LEO satellite. Therefore, the channel parameters in (1) will be updated over time. Additionally, we assume that an effective synchronization mechanism is in place. This is often achieved using synchronization signals or pilots transmitted by the satellite, allowing the ground users to estimate the time delay and frequency offset [29]. Note that the proposed DRL-based method is able to learn and build knowledge about the wireless channels without knowing the channel model by observing the rewards and finding solutions to the optimization problem.

In the considered network model, the transmitter of the LEO satellite employs the hybrid beamforming architecture with M RF chains, such that $K \leq M \leq N$. Let the vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_K]^T \in \mathbb{C}^{K\times 1}$ be the transmit symbols. The vector $\boldsymbol{x}$ is first precoded with a digital precoder and then processed by an analog precoder. Let $\boldsymbol{W} \in \mathbb{C}^{M\times K}$ be the digital precoding matrix and $\mathbf{V} \in \mathbb{C}^{N\times M}$ denote the analog precoding matrix. We consider that the analog precoder is designed with phase shifters that adjust the signal phase only, i.e., $|\mathbf{V}_{i,j}| = 1$. The phase shifters are implemented by a uniform quantizer with c-bits resolution and step size $\Delta = 2\pi/2^c$. Furthermore, a discrete phase shifter with a finite set of resolutions is considered in this work, i.e., each reflection element can obtain a phase shift value from 2^c discrete values within the interval $[0, 2\pi)$. To this end, the set of values of the analog precoder $\mathbf{V}$ satisfy

$$\mathcal{V} = \left\{\mathbf{V} \mid \mathbf{V}_{n,m} = \exp\left\{\jmath\left(\frac{2\pi}{2^c}\iota + \frac{\pi}{2^c}\right)\right\}, \quad \forall n, m, \iota = 0, 1, \ldots, 2^c - 1\right\} \quad (5)$$

Accordingly, the received signal at the ground user k at time

slot t can be modeled as

$$y_k(t) = \mathbf{H}_k^H(t) \sum_{k\in\mathcal{K}} \boldsymbol{V}(t)\boldsymbol{w}_k(t)x_k(t) + \varsigma_k(t)$$
$$= \overbrace{\boldsymbol{H}_k^H(t)\boldsymbol{V}(t)\boldsymbol{w}_k(t)x_k(t)}^{\text{Desired signal}} + \overbrace{\boldsymbol{H}_k^H(t)\sum_{k'\neq k}\boldsymbol{V}\boldsymbol{w}_{k'}x_{k'}}^{\text{inter-beam interference}} + \varsigma_k, \tag{6}$$

where $\varsigma_k \sim \mathcal{CN}\left(0, \sigma^2\right)$ is the additive Gaussian white noise variable of the k^{th} ground user with a zero mean and variance σ^2. In this paper, we opted for a fully connected architecture due to the increased flexibility it provides in beamforming, especially for LEO satellite networks. In a fully connected architecture, each RF chain can control an individual antenna, resulting in a higher degree of freedom for beamforming, potentially improving system performance. While this architecture could be more power-intensive compared to the partially connected architectures approaches due to the presence of more phase shifters, it can be managed efficiently by optimizing the LEO satellite energy efficiency. Thus, the SINR of the k^{th} user can be given by

$$\text{SINR}_k(t) = \frac{\left|\boldsymbol{V}(t)\boldsymbol{w}_k^H(t)\boldsymbol{H}_k(t)\right|^2}{\sum_{k'\neq k}\left|\boldsymbol{V}(t)\boldsymbol{w}_{k'}^H(t)\boldsymbol{H}_{k'}(t)\right|^2 + \sigma^2}, \tag{7}$$

and the obtained downlink data rate of the user k is defined as $\gamma_k(t) = B\log\left(1+\text{SINR}_k(t)\right)$, where B is the downlink bandwidth. Considering the outdated CSIs, the data rate at time slot t is obtained based on the estimated channel $\boldsymbol{H}(t)$, which is, in practice, the experienced wireless channel at $t-\tau$, i.e., $\boldsymbol{H}(t-\tau)$, where τ is the outdated period. The estimated channel may no longer accurately reflect the current state of the wireless channel if the outdated period is long, impacting network performance. In the next section, we discuss how the proposed approach can reduce the impact of outdated CSIs by learning dependencies between the estimated channels over time.

B. Problem Formulation

The objective is to design a robust precoding technique considering the dynamic channel variations and energy consumption of the LEO satellite. To achieve that, we formulate an optimization problem that maximizes the energy efficiency of the LEO satellite while considering transmission reliability. In this work, we define transmission reliability in terms of the minimum data rate satisfaction of ground users. In particular, uncertainty in channel variations due to the dynamic nature of wireless systems exacerbates the obtained throughput by each user over time slots, impacting transmission reliability. In such a case, formulating an optimization problem based on the average quantity of data rate without considering the uncertainty of channel conditions over time may violate the minimum required data rate by each user over time slots. Thus, the LEO satellite needs to construct appropriate beamforming preferences for users with bad channel conditions. In other words, If some users are experiencing poor channel conditions, the satellite might have to "invest" more resources (beam direction and power) to maintain a good connection. To this end, in this work, we use the CVaR as a risk measure to capture the dynamic characteristics of wireless channels, as it efficiently characterizes risk on investments in modern portfolio theory [30]. In such a case, incorporating CVaR in the data rate constraint during beamforming optimization effectively characterizes the potential loss in data rate due to poor channel conditions, as CVaR captures the risk in the tail distribution of the data rate by defining the average potential loss that exceeds the Value-at-Risk (VaR), allowing the system to prepare for extreme conditions and enhancing transmission reliability. In general, the CVaR of a random variable Z is defined as [31]

$$\text{CVaR}_\alpha(Z) := \inf_{\delta\in\mathbb{R}}\left[\delta + \frac{1}{1-\alpha}\mathbb{E}\left[(Z-\delta)^+\right]\right], \tag{8}$$

where $\alpha \in (0,1)$. Thus, the data rate constraint of each ground user can be formulated based on the CVaR as follows

$$\text{CVaR}_\alpha[\gamma_k(t)] \geq \gamma_k^{\min},\ \forall k \in \mathcal{K}, \tag{9}$$

where $\gamma_k^{\min}$ is the minimum required data rate by the ground use k.

Lemma 1. *Constraint* (9) *can guarantee the minimum required data rate with a probability higher than* $1-\alpha$.

Proof. The α-percentile (Value at Risk) of a random variable, i.e., the value for which the likelihood of a random variable is less than or equal to it is at least α, is given by [31]:

$$\text{VaR}_\alpha(Z) = \arg\inf_\delta\{\delta : \Pr(Z > Z_{\min}) \leq \alpha\}, \tag{10}$$

where $\Pr(\cdot)$ defines the probability. Thus, $\text{VaR}_\alpha[\gamma_k(t)] \leq \gamma_k^{\min}$ is equivalent to following probabioity constraint:

$$\Pr[\gamma_k(t) \geq \gamma_k^{\min}] \geq 1-\alpha. \tag{11}$$

Since the VaR of distribution of γ is a minimizer of the right-hand side in (8), $\text{CVaR}_\alpha[\gamma_k(t)] \geq \text{VaR}_\alpha[\gamma_k(t)]$ always holds. Therefore, the constraint (9) defines an approximation of the chance constraint (11). ■

The energy efficiency of the LEO satellite at time slot t is given by

$$\eta(t) = \frac{\sum_{k\in\mathcal{K}}\gamma_k(t)}{\sum_{k\in\mathcal{K}}||\boldsymbol{V}\boldsymbol{w}_k||^2 + P_0}, \tag{12}$$

where the term $\sum_{k\in\mathcal{K}}||\boldsymbol{V}\boldsymbol{w}_k||^2$ is the total transmit power of the LEO satellite and P_0 is the power consumption of the circuit hardware. In general, the hardware power consumption of the LEO satellite transmitter can be approximated as [32]:

$$P_0 = M \times N_s \times P_{\text{ps}} + M \times P_{\text{RFC}} + P_{\text{LO}} + P_{\text{BB}}, \tag{13}$$

where P_{ps} denotes the power consumption of the phase shifter, while P_{LO} and P_{BB} indicate the power consumption of the local oscillator and the baseband digital precoder, respectively. The power utilized by each RF chain, represented as P_{RFC}, is calculated as $P_{\text{RFC}} = P_{\text{DAC}} + P_{\text{mixer}} + P_{\text{LPF}} + P_{\text{BBA}}$. Here, P_{DAC} refers to the power consumption of the digital-to-analog converter, P_{mixer} is the power consumed by the mixer, P_{LPF} pertains to the low pass filter's power consumption, and P_{BBA} signifies the power used by the baseband amplifier.

Accordingly, we formulate the following stochastic optimization problem

$$\underset{\boldsymbol{V},\boldsymbol{W}}{\text{maximize}} \quad \frac{1}{T}\sum_{t=1}^{T}\eta(t) \tag{14a}$$

$$\text{subject to} \quad \text{CVaR}_{\alpha}[\gamma_k(t)] \geq \gamma_k^{\min},\ \forall k \in \mathcal{K}, \tag{14b}$$

$$\sum_{k\in\mathcal{K}} ||\boldsymbol{V}\boldsymbol{w}_k||^2 \leq P_{\max}, \tag{14c}$$

$$\boldsymbol{V}_{m,n} = \exp\left\{\jmath\left(\frac{2\pi}{2^c}\iota + \frac{\pi}{2^c}\right)\right\},$$
$$\iota = \{0,1,\ldots,2^c-1\},\ \forall n \in \mathcal{N}, m \in \mathcal{M} \tag{14d}$$

where $P_{\max}$ represents the upper limit of the transmission power permissible for the LEO satellite. The objective of Problem (14) is to determine the optimal analog and digital precoding matrices, denoted as $\boldsymbol{V}^*$ and $\boldsymbol{W}^*$ respectively, in order to maximize the LEO satellite's energy efficiency while maintaining the required transmission reliability. The constraint (14b) is set to achieve the data rate required by each ground user regardless of the channel variations. Constraints (14c) and (14d) define the feasibility regions of the optimization variables. We reformulate problem (14) by using the CVaR definition in (8) as follows

$$\underset{\boldsymbol{V},\boldsymbol{W},\delta\in\mathbb{R}}{\text{maximize}} \quad \frac{1}{T}\sum_{t=1}^{T}\eta(t) \tag{15a}$$

$$\text{subject to} \quad \delta + \frac{\mathbb{E}[(\gamma_k(t)-\delta)^+]}{(1-\alpha)} \leq \gamma_k^{\min},\ \forall k \in \mathcal{K}, \tag{15b}$$

$$\sum_{k\in\mathcal{K}} ||\boldsymbol{V}\boldsymbol{w}_k||^2 \leq P_{\max}, \tag{15c}$$

$$\boldsymbol{V}_{m,n} = \exp\left\{\jmath\left(\frac{2\pi}{2^c}\iota + \frac{\pi}{2^c}\right)\right\},$$
$$\iota = \{0,1,\ldots,2^c-1\},\ \forall n \in \mathcal{N}, m \in \mathcal{M}. \tag{15d}$$

The formulated risk-aware stochastic problem (15) is a mixed-integer and non-convex optimization problem making obtaining a closed-form solution using the classical optimization techniques infeasible. Therefore, we propose in the following section a dynamic online learning framework to solve the formulated problem.

III. Proposed Learning-based Solution

The random time-varying nature of wireless channels in satellite networks increases the complexity of solving problem (15). In particular, the fast movement of LEO satellites may lead to outdated CSI estimations, making obtaining an optimal solution more challenging. Classical optimization methods are deterministic and mathematically rigorous and can provide a guaranteed optimal solution if the optimization problem is well-defined and convex. However, classical optimization methods can be computationally expensive and require knowledge of the exact network dynamics, which may not be available in the presence of outdated channel estimations. In contrast, DRL-based approaches are model-free and can learn directly from experience without the need for explicit knowledge of the system dynamics [33]. Specifically, DRL algorithms can update decision policies to reach optimal system performance through feedback based on previously performed actions, even without up-to-date information.

To this end, we propose a DRL-based approach that can learn and build knowledge about wireless channels by observing the rewards from the wireless environment and finding out solutions to (15). The proposed DRL-based approach can learn nonlinear dependencies between past and present channel values., i.e., dependencies between CSI estimations over time, and adapt to statistics channel changes. Precisely. the proposed algorithm uses the recent CSI estimations as an input to obtain the optimal beamforming instead of considering only the instantaneous CSI at each time step. In the following subsections, we first model the formulated optimization problem as a Markov process and present a sliding window-based mechanism to store the recent CSI estimations and incorporate them in the state space to reduce the impact of outdated channel estimations. Then, we propose a two-stage policy gradient-based learning algorithm to solve the problem.

A. Markov Process Model

The Markov model is characterized by state set $\mathcal{S}$, action set $\mathcal{A}$, and instantaneous reward $R(t)$. The network state $\mathcal{S}$ contains all possible network configurations, including channel information of all ground users. At a time slot t, the agent observes the network state $s(t)$ which belongs to the state space $\mathcal{S}$, and selects an action $a(t)$ from the action space $\mathcal{A}$ based on the policy $\boldsymbol{\pi}(t)$, where $\boldsymbol{\pi}(t)$ is a function mapping from $\mathcal{S}$ to $\mathcal{A}$. In the proposed model, the four fundamental components: agents, states, actions, and reward, are characterized as follows:

Agent: The LEO satellite is regarded as an agent. The agent can get network states that include the information of all ground users within its coverage and acts as an action selection policy. Thus, the agent observes the states and then chooses an action from the action space. After that, the agent receives a reward from the environment and moves to the next state.

Action Space: The action space contains all possible values of the decision variables in (14). Therefore, we define the action space as $\mathcal{A} = \{\boldsymbol{V}, \boldsymbol{W}\}$. Specifically, the action set at time slot t contains the analog and digital beamforming matrices.

State Space: We consider the state space as tuples of the channel state of each ground user and the selected actions during the previous time slot $t-1$. *Furthermore, we consider the dependencies between past and present channel information to mitigate the impact of outdated CSI by incorporating historical channel measurements into the state representation.* We use a fixed-size buffer to store the most recent CSIs and update it at each time slot. At time step t, the CSI buffer stores the latest received channel information and discards the oldest one in the buffer, creating a rolling window of the most recent CSIs and ensuring that the state set is always up-to-date with the latest available channel information. This sliding window mechanism provides a snapshot of the most recent states of

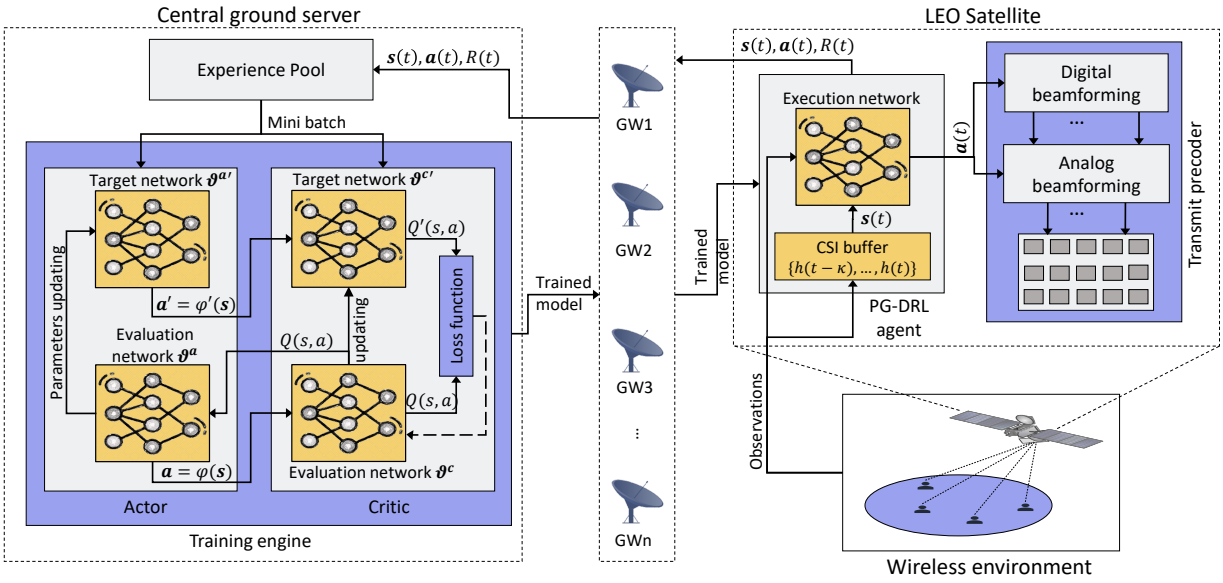

Figure 2: Block diagram of the proposed learning algorithm.

the channel, allowing the DRL algorithm to base its decisions on the most relevant data. Specifically, this approach enables the agent to learn dependencies between CSIs and provides more effective beamforming strategies that can adapt to the changing channel conditions in real-time.

Therefore, the state set at time slot t can be defined as $\boldsymbol{s}(t) = \{\hat{\boldsymbol{h}}(t), \boldsymbol{V}(t-1), \boldsymbol{W}(t-1)\}$, where $\hat{\boldsymbol{h}}(t) = \{\boldsymbol{h}(t-\kappa), \boldsymbol{h}(t-\kappa+1), \ldots, \boldsymbol{h}(t-1), \boldsymbol{h}(t)\}$, κ is the size of the CSI buffer.

Reward: The reward function is a measure of the selected action at a given network state. The agent updates the policy $\boldsymbol{\pi}$ towards an optimal one, i.e., the policy with higher rewards. To this end, the reward function must meet the objective of the optimization problem without violating its constraints. Our problem aims to improve the energy efficiency of the LEO satellite while keeping the QoS constraints of all ground users. Thus, we formulate a reward function as a tradeoff of two main parts, including the energy efficiency and the minimum data rate constraint, as follows

$$R(t) = \eta(t) + \beta(t)\left[\delta + \frac{\mathbb{E}[(\gamma(t)-\delta)^{+}]}{(1-\alpha)} - \gamma_k^{\min}\right]. \quad (16)$$

where $\beta(t)$ represents a weighting parameter that varies with time. Note that the values of $\beta(t)$ change across different time slots. This variation is to ensure fulfilling the minimum QoS requirements while simultaneously enhancing the energy efficiency of the system.

Algorithm 1 : Offline Training Algorithm.

1: Initialize the target and evaluation networks at both critic and actor units;
2: Set the size of the experience buffer;
3: **for each training epoch:**
4: **repeat**
5: Set the parameters within the critic's evaluation network by minimizing the loss function as defined in (18);
6: Adjust the parameters in the actor's evaluation network by maximizing (20);
7: Adjust the parameters in the critic's target network in accordance with the method outlined in (19);
8: Update the parameters in the actor's target network following (22);
9: **until** Convergence
10: **if** Replay buffer memory is full
11: Sample a mini-batch of the recent data in the replay memory;
12: Start a new training epoch;
13: **else**
14: Wait until the buffer size is full;
15: Sample a mini-batch of the recent data in the replay memory;
16: Start a new training epoch;
17: **end if**
18: **end for**

B. Proposed Two-Stage Learning Architecture

As depicted in Fig 2, the proposed framework consists of a training unit located at a central terrestrial server and an online execution unit installed at the LEO satellite. The training unit is designed as an Actor-Critic with evaluation and target networks to improve the training performance. At each training epoch, the training engine gets a mini-batch of training data from the experience pool. After convergence, the trained parameters of the actor unit are sent to the execution unit for beamforming executions. The LEO satellite feeds the network states as an input to the execution neural network and receives the output, which is the decision on the hybrid beamforming. The LEO satellite obtains the reward and forwards all information, including the reward, selected action,

and the network state, to the experience pool at the central data server through ground stations to use as training data for the next training epoch.

The ϵ-greedy mechanism is used to balance the exploration and exploitation of action selection during the training process. In practice, the agent chooses random actions with $1-\epsilon$ probability during training, which may lead to entering unpredictable system states. To avoid this, the digital-twins technology [34] can be leveraged to evaluate the selected actions during the training phase. Furthermore, pre-trained neural network models can be used by the execution unit at the initial stage to ensure providing real-time decisions. Then, the training unit will train new network models at each training epoch based on the collected data during the execution process over time, adapting to network dynamics and improving action selection accuracy. Note that during the training process and transmission through ground stations in each training epoch, the execution unit continues performing action selection based on the received model in the previous training epoch. Algorithms 1 and 2 illustrate step-by-step the training and execution processes, respectively.

Algorithm 2 : Online Execution Algorithm.

1: Initialize the execution network parameters according to the received trained model from the training unit;
2: **for each time step:**
3: Observe the current network state $s(t)$;
4: Choose an action from the action space according to the policy π;
5: Execute the selected action and receive the reward $R(t)$;
6: Observe the new network state $s(t+1)$;
7: Send the $a(t), s(t), R(t), s(t+1)$ to the experience memory at the ground station;
8: **if** New trained model received
9: Update the execution network parameters;
10: **end if**
11: Start new time step;
12: **end for**

C. Policy Gradient DRL Algorithm

In the Policy Gradient DRL (PG-DRL) approach, the agent aims to obtain the optimal action selection policy π that maximizes the long-term cumulative reward. In this paper, we consider a PG-DRL algorithm in the form of actor-critic networks and deploy it at the ground station for the training process while another actor network is located at the LEO satellite to perform online executions. The actor network performs the action selection process, while the purpose of the critic network is to evaluate the selected action based on the network observations and the obtained reward during the training process. We use the experience replay technology and target networks to accelerate training time and enhance the stabilization of the algorithm. The network observations, the selected action, and the obtained reward are stored in the replay memory at each time step. Then, the training unit samples data from the replay memory (mini-batch) at each training epoch to train both the actor network and critic network, improving the action decision accuracy.

1) Critic Deep Neural Networks Design: The critic network assesses the selected actions using the state-action function defined as [35]

$$Q^{\boldsymbol{\pi}}\left(\boldsymbol{s},\boldsymbol{a}\right)=\mathbb{E}_{\pi}\left[\sum_{\tau=0}^{\infty}\xi^{\tau}R(t+\tau)\mid\boldsymbol{s}(t)=\boldsymbol{s},\boldsymbol{a}(t)=\boldsymbol{a}\right], \tag{17}$$

where ξ^{τ}, which lies within the range $(0,1]$, serves as a discount factor. The critic network employs two Deep Neural Networks (DNNs), designated as the target and evaluation networks. The parameters of the critic evaluation and target networks are denoted by $\boldsymbol{\vartheta}^{c}$ and $\boldsymbol{\vartheta}^{c\prime}$, respectively. The critic network acquires the parameter $\boldsymbol{\vartheta}^{c}$ through the minimization of the state-action function loss, represented as $L(\boldsymbol{\vartheta}^{c})$, which is defined as follows:

$$L\left(\boldsymbol{\vartheta}^{c}\right)=\mathbb{E}\left[\left[Q\left(s(t),a(t)\right)-\left(R(t)+\xi Q'\left(s,a\right)\right)\right]^{2}\right] \tag{18}$$

where $Q'(\cdot)$ defines the target network state-action function. The loss function $L\left(\boldsymbol{\vartheta}^{c}\right)$ is typically based on the Temporal Difference (TD) error [36]. The square of the difference between these two terms forms the TD error, which is the loss function for the critic network. Through the minimization of this loss, the training of the critic network converges towards a more precise approximation of the actual Q-value function. As a result, this leads to enhanced accuracy in the estimation of action-values, thereby enabling the DRL agent to make more effective decisions. The adjustment of the parameters $\boldsymbol{\vartheta}^{c}$ is accomplished by employing the gradient of the loss function $L(\boldsymbol{\vartheta}^{c})$. The parameters of the target network are updated as

$$\boldsymbol{\vartheta}^{c\prime}=\zeta^{c}\boldsymbol{\vartheta}^{c}+\left(1-\zeta^{c}\right)\boldsymbol{\vartheta}^{c\prime}. \tag{19}$$

The parameters of the target network $\boldsymbol{\vartheta}^{c\prime}$ are revised to reflect a combination of their existing values and the recently updated parameters from the main critic network $\boldsymbol{\vartheta}^{c}$. The degree of mixing is controlled by the parameter ζ^{c}, which is typically set to a small value $\zeta^{c}\ll 1$, so the target network parameters change slowly over time [37].

2) Actor Deep Neural Networks Design: As the objective of the actor network is to perform action selection for each network state with the aim of maximizing the cumulative reward, the parameters of the evaluation network at the actor are adjusted by maximizing the following

$$J\left(\boldsymbol{\vartheta}^{a}\right)=\mathbb{E}\left[Q\left(s,a\right)\mid a=\varphi\left(s\right)\right], \tag{20}$$

where $\varphi(\cdot)$ is the actor's evaluation network function and $\boldsymbol{\vartheta}^{a}$ is the parameters of the actor evaluation network. In particular, the function $\varphi(\cdot)$ represents the policy $\boldsymbol{\pi}: s\mapsto a$. Using the gradient of (20), the parameter $\boldsymbol{\vartheta}^{a}$ can be updated in the direction $\nabla_{\boldsymbol{\vartheta}}J\left(\boldsymbol{\vartheta}^{a}\right)$ as follows

$$\boldsymbol{\vartheta}^{a}(t+1)=\boldsymbol{\vartheta}^{a}(t)+\rho_{a}\nabla_{\boldsymbol{\vartheta}}J(\boldsymbol{\vartheta}^{a}), \tag{21}$$

where ρ_{a} is the learning rate. The parameters of the target network are updated as follows

$$\boldsymbol{\vartheta}^{a\prime}=\zeta^{a}\boldsymbol{\vartheta}^{a}+\left(1-\zeta^{a}\right)\boldsymbol{\vartheta}^{a\prime}, \tag{22}$$

where $\zeta^{a}\ll 1$.

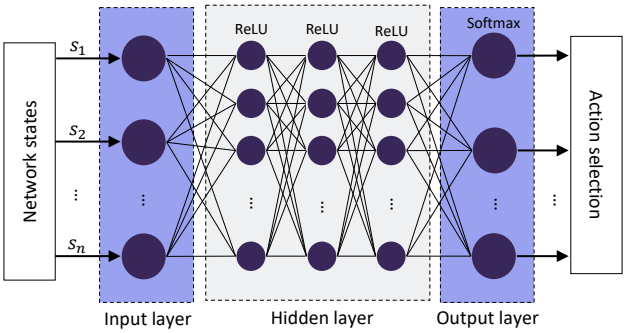

Figure 3: Architecture of the actor neural network.

D. Complexity and Convergence Analysis

1) Computation Complexity Analysis: Let L, Z_o and Z_l denote the training layers, the size of the input layer and the number of neurons in the l^{th} layer of the DNN, respectively. Thus, at each time step, the computational complexity of the agent is $\mathcal{O}(Z_o Z_l + \sum_{l=1}^{L-1} Z_l Z_{l+1})$. Let η_{epi} be the number of episodes of each mini-batch with T time steps in each episode. Thus, the total computation complexity of the DNN is $\mathcal{O}\Big(\eta_{epi} T(Z_o Z_l + \sum_{l=1}^{L-1} Z_l Z_{l+1})\Big)$. In particular, due to the high computation complexity, the training phase of the DNN is performed offline at the ground station.

The complexity of choosing an action $a \in \mathcal{A}$ at the agent depends on the dimensions of the action and state spaces. Thus, the computation complexity can be defined as $\mathcal{O}(|\mathcal{S}|^2 \times |\mathcal{A}|)$. Accordingly, the total complexity of the proposed DRL-based algorithm is $\mathcal{O}\Big(\eta_{epi} T(Z_o Z_l + \sum_{l=1}^{L-1} Z_l Z_{l+1}) + |\mathcal{S}|^2 \times |\mathcal{A}|\Big)$.

2) Convergence Analysis: We conclude the convergence of the proposed algorithm in the following proposition:

Proposition 1. *The convergence of the proposed two-stage DRL-based algorithm depends on the gradient-based learning procedure at the training unit.*

Proof. The proposed two-stage learning framework trains a DRL agent at the central ground server based on the policy gradient approach. At the LEO satellite, the DRL agent performs decisions according to the received network information using the trained parameters at the ground sever. Thus, the convergence of the proposed algorithm depends on the convergence of the training of the gradient-based DRL at the training engine. ■

The convergence of gradient-based DRL can be proven by showing the convergence of the objective function in a mean-square sense using the stochastic approximation theory, which can be found in [38].

IV. Performance Evaluation

A. Simulation Settings

We provide simulation analysis in this section to evaluate the proposed approach. In the simulation environment, we consider a LEO satellite orbiting at 1300 km and operating at the Ka-band with a carrier frequency of 20 GHz. At the LEO satellite side, a UPA antenna is installed composed of N_s^x antenna elements in the x-axis and N_s^y antennas in the y-axis. Unless stated otherwise, the values of N_s^x and N_s^y are set as $N_s^x = N_s^y = 20$. Thus, the total number of antenna elements at the satellite is $N_s = N_s^x \times N_s^y = 400$. To consider the network dynamics, the ground users within the satellite coverage area are generated at each time slot following the Poisson process with a time-varying arrival rate $\lambda(t)$. We generate synthetic data for training by adjusting channel parameters over time slots. The Rician factor values are generated randomly within the interval $[0.1, 3]$ to capture the dynamic nature of the wireless environment. During training, the agent selects actions randomly from the action space or samples actions based on its policy. The ϵ-greedy method is used to balance exploration and exploitation during action selection. The ϵ value is initialized to 0.99 at the beginning of the training process to encourage exploration. Then, the epsilon value is gradually decreased over time slots as the agent becomes more confident in its action decisions.

We use a DNN model with three hidden layers, where the first hidden layer contains 600 neurons, the second hidden layer consists of 300 neurons, and the third hidden layer has 250 neurons. The number of neurons in the input and output layers is set as the same as the number of network states and actions, respectively. The ReLU activation function is used in the hidden layers while the Softmax function is used as an activation function for the output layer to get the probability distribution of the actions and the sampled action. Fig. 3 demonstrates the used DNN architecture. Furthermore, the discount factor and learning rate are adjusted at 0.95 and 0.001, respectively. The size of the experience buffer memory and the size of each mini-batch are 2000 and 32, respectively. We generate synthetic data to simulate a massive MIMO LEO satellite communication system. Specifically, the data is generated using Python based on the provided channel models relevant to LEO satellite communication. We simulate different environmental conditions, including various distances between users and the LEO satellite and the number of users. The dataset was generated over 10000 time steps under various configurations to encapsulate a realistic and dynamic network environment. During each time step, the generated data was represented as a matrix. The dimensions of this matrix correspond to the number of users and the number of antennas at the LEO satellite. For each user and the corresponding antenna element, the dataset includes various state variables, including channel gain, transmit power, and phase shift values. The parameters for the LEO satellite operation, such as altitude, speed, and orbital inclination, were chosen based on typical values used in real-world LEO satellite systems [39]. While we start with initial channel estimations, our scheme continuously updates these estimations based on the system's dynamic behavior. This represents a more realistic scenario in LEO satellite networks, where perfect CSI is typically not available due to the inherent dynamic nature of the system.

The effectiveness of the proposed algorithm is evaluated through comparative analysis with the following baseline methodologies:

1) Classical-Average: [32]: This approach focuses on maximizing energy efficiency while considering the minimum

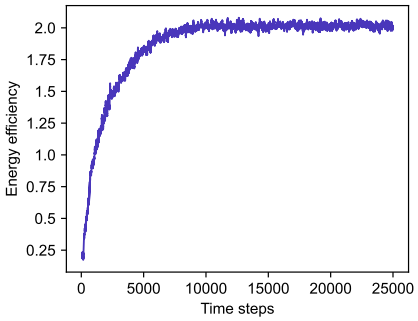

Figure 4: Obtained energy efficiency (Mbits/Joule) over time during training.

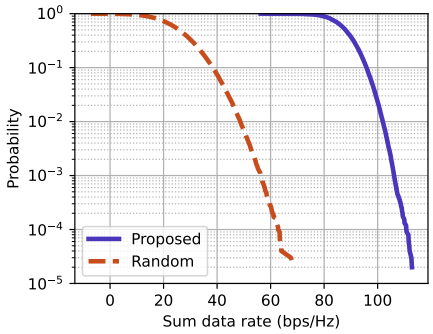

Figure 5: Sum data rate obtained over multiple time slots.

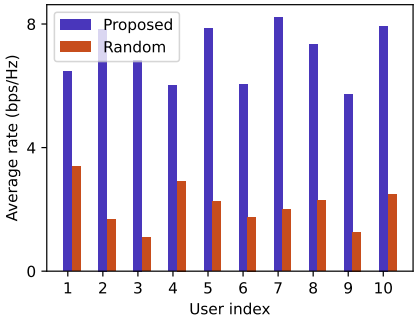

Figure 6: Average per-user downlink data rate.

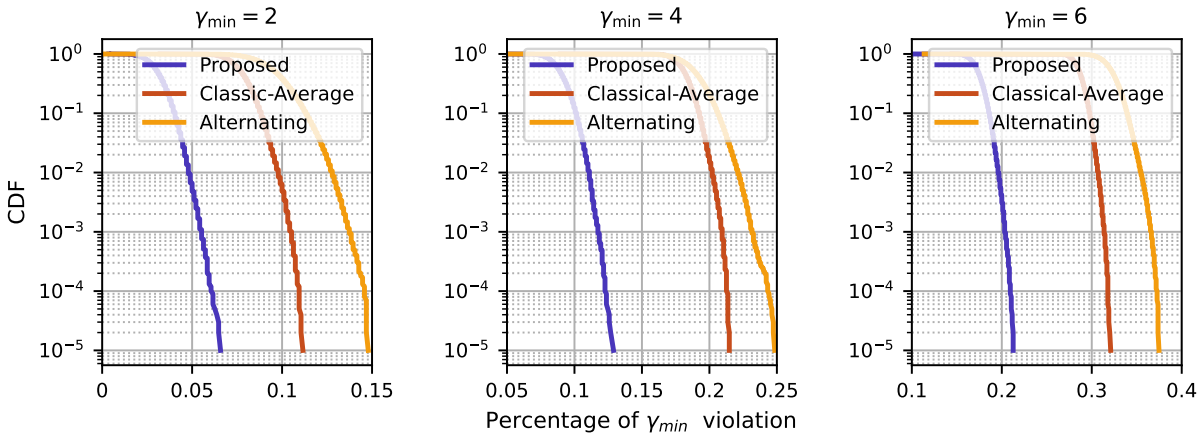

Figure 7: CDF of the QoS violation probability.

average SINR (without using the CVaR). The proposed DRL-based algorithm is used to solve this problem.

2) Alternating-Optimization [20]: An alternative optimization-based algorithm is used to obtain the hybrid precoding in massive MIMO LEO satellite systems.

3) Random: Where analog and digital precoding are set randomly regardless of the channel state of the users.

B. Results Discussion

We first show the convergence of the energy efficiency during the training process in Fig. 4. During the training time steps, the critic and actor update the parameters of the target and evaluation networks to obtain the optimal selection policies and the corresponding state-action function, achieving higher and more stable reward values over time steps. We can notice that the obtained energy efficiency values improve over training steps and convergences to optimal values of around 2 Mbits/Joule. As shown in Fig. 4, the proposed algorithm converges within 10000 time steps. During each training epoch, the execution unit performs decisions based on the received neural network models in the previous training epoch, ensuring real-time executions.

Fig. 5 presents the cumulative distribution function (CDF) for the aggregate data rate experienced by all ground users over various time slots. The results indicate that our proposed method achieves a higher data rate compared to the *Random* baseline. As depicted in Fig. 5, the range of the sum data rate for our method fluctuates from 80 to 110 bps/Hz, averaging around 100 bps/Hz. In contrast, the sum data rate for the *Random* approach spans from 10 bps/Hz to 65 bps/Hz, with an average data rate of 45 Mbps/Hz. This is because the proposed algorithm adjusts the analog and digital precoding matrices over time to improve the channel quality of each ground user, while in the *Random* method, the beamforming is set randomly without considering the network data rate. In Fig. 6, we discuss the average per-user data rate to show the achieved data rate at each ground user. Specifically, we plot the average over-time data rate of 10 randomly chosen ground users. As demonstrated in Fig. 6, the proposed algorithm gives a better data rate as it ensures that the data rate of each user is higher than a minimum threshold. Furthermore, Fig. 6 shows that the *Random* method has a lower data rate than the proposed algorithm. In fact, the *Random* approach sets the beamforming randomly without considering the channel state, resulting in a poor data rate.

In Fig. 7, the performance of the proposed algorithm is evaluated in terms of the probability of violating the minimum data rate requirement (γ^{min}) across various time slots and settings of γ^{min}. Additionally, we benchmark the outcomes achieved by our approach against the *Classic-Average* and *Alternating-Optimization* baselines. This comparison is vital

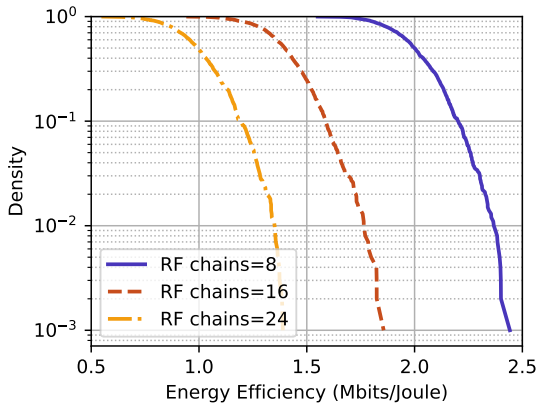

Figure 8: CDF of energy efficiency for a different number of RF chains.

to ascertain the efficacy of the proposed risk-aware strategy in enhancing the reliability of the system. The findings reveal that the highest violation probability for the proposed risk-aware method is 0.065, with an average probability of 0.05, observed when γ^{min} is set to 2. However, the violation probability of the *Classic-Average* and *Alternating-Optimization* methods are 0.11 and 0.13, respectively. When setting $\gamma^{min} = 4$, the violation probability of the proposed mechanism is 0.12 on average and 0.135 at worst, while the QoS average violation probability of the *Classic-Average* and *ALternating-Optimization* baselines are 0.2 and 0.225, respectively. Finally, setting the minimum data rate threshold at 6 increases the violation probability of the proposed approach, *Classic-Average*, and *Alternating-Optimization* to 0.2, 0.31, and 0.35, respectively. We can conclude that the proposed risk-aware method enhances transmission reliability by reducing the QoS violation probability with a dynamic wireless environment over time. Specifically, the obtained results illustrate that the proposed risk-aware algorithm performs better than other baselines. In particular, the proposed CVaR-based approach considers the variance of the data rate over time slots in addition to the average data rate, improving transmission reliability.

Finally, we study the relation between energy efficiency and the number of RF chains in Fig. 8. Specifically, we plot the CDF of energy efficiency across configurations with 8, 16, and 24 RF chains. Notably, a reduced number of RF chains yields better energy efficiency. However, increasing the RF chains (higher phase shift resolution) leads to a noticeable deterioration in energy efficiency performance. Specifically, the obtained results show that we can achieve energy efficiency up to 2.4 when setting the number of the RF chains to 8. However, increasing the number of RF chains to 24 reduces energy efficiency to around 1.4. This is because the increase in the number of RF chains, i.e., increase in the phase shift resolution, results in a significant increase in power consumption while providing a small improvement in the data rate. This consequently leads to a drop in energy efficiency. Conversely, with fewer RF chains, the rate becomes the dominant factor influencing energy efficiency, leading to better energy efficiency.

V. Conclusion

This paper has discussed the precoding optimization in massive MIMO LEO satellite communication systems. A risk-sensitive stochastic optimization problem has been formulated to optimize the hybrid digital/analog beamforming aiming at maximizing the energy efficiency of the satellite while considering the QoS requirements of all ground users. The CVaR has been proposed as a risk measure to guarantee the required data rate by each ground user under uncertain channel variations. To solve the formulated problem, we have proposed a two-stage learning framework based on the PG-DRL technique, which can provide real-time decisions on digital and analog beamforming. The obtained results show the efficacy of the proposed approach in improving the energy efficiency of the system while keeping the QoS requirements under dynamic channel variations. The proposed algorithm can be extended beyond the scope of a single LEO satellite to a more realistic scenario of a constellation of LEO satellites. In such a scenario, a multi-agent DRL-based approach is a promising approach that allows each satellite to act as an individual agent, learning and evolving with its peers. Inter-Satellite Links (ISL) can be leveraged to provide a platform for cooperative learning among satellites.

References

[1] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kodheli, H. Sallouha, A. Colpaert, J. F. Mendoza Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani, E. Lagunas, and B. Ottersten, "Evolution of Non-Terrestrial Networks from 5G to 6G: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2633–2672, 2022.

[2] A. Sattarzadeh, Y. Liu, A. Mohamed, R. Song, P. Xiao, Z. Song, H. Zhang, R. Tafazolli, and C. Niu, "Satellite-based Non-Terrestrial Networks in 5G: Insights and Challenges," *IEEE Access*, vol. 10, pp. 11 274–11 283, 2022.

[3] 3GPP, "Solutions for NR to support Non-Terrestrial Networks (NTN)," 2021.

[4] ——, "Study on New Radio (NR) to support Non-Terrestrial Networks," 2019.

[5] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 836–869, 2018.

[6] L. You, K.-X. Li, J. Wang, X. Gao, X.-G. Xia, and B. Ottersten, "Massive MIMO transmission for LEO satellite communications," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1851–1865, 2020.

[7] K.-X. Li, L. You, J. Wang, X. Gao, C. G. Tsinos, S. Chatzinotas, and B. Ottersten, "Downlink transmit design for massive MIMO LEO satellite communications," *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 1014–1028, 2022.

[8] M. Y. Abdelsadek, G. K. Kurt, and H. Yanikomeroglu, "Distributed massive MIMO for LEO satellite networks," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 2162–2177, 2022.

[9] H. Cheporniuk, R. T. Schwarz, T. Delamotte, and A. Knopp, "MIMO throughput performance analysis in LEO communication scenario," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, 2021, pp. 01–06.

[10] M. A. Vazquez, M. R. B. Shankar, C. I. Kourogiorgas, P.-D. Arapoglou, V. Icolari, S. Chatzinotas, A. D. Panagopoulos, and A. I. Pérez-Neira, "Precoding, scheduling, and link adaptation in mobile interactive multibeam satellite systems," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 5, pp. 971–980, 2018.

[11] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband leo satellite communications: Architectures and key technologies," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 55–61, 2019.

[12] L. Liang, W. Xu, and X. Dong, "Low-complexity hybrid precoding in massive multiuser MIMO systems," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 653–656, 2014.

[13] A. Kaushik, E. Vlachos, C. Tsinos, J. Thompson, and S. Chatzinotas, "Joint bit allocation and hybrid beamforming optimization for energy efficient millimeter wave MIMO systems," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 1, pp. 119–132, 2021.

[14] R. Dilli, "Performance analysis of multi user massive MIMO hybrid beamforming systems at millimeter wave frequency bands," *Wireless Networks*, vol. 27, no. 3, pp. 1925–1939, 2021.

[15] I. Ahmed, M. K. Shahid, H. Khammari, and M. Masud, "Machine learning based beam selection with low complexity hybrid beamforming design for 5G massive MIMO systems," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 4, pp. 2160–2173, 2021.

[16] M. Han, J. Du, Y. Zhang, X. Li, K. M. Rabie, and G. Nauryzbayev, "Efficient hybrid beamforming design in mmWave massive MU-MIMO DF relay systems with the mixed-structure," *IEEE Access*, vol. 9, pp. 66 141–66 153, 2021.

[17] S. Gherekhloo, K. Ardah, and M. Haardt, "Hybrid beamforming design for downlink MU-MIMO-OFDM millimeter-wave systems," in *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2020, pp. 1–5.

[18] K. Ardah, G. Fodor, Y. C. B. Silva, W. C. Freitas, and A. L. F. de Almeida, "Hybrid analog-digital beamforming design for SE and EE maximization in massive MIMO networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 377–389, 2020.

[19] Y. Zhang, Y. Wu, A. Liu, X. Xia, T. Pan, and X. Liu, "Deep learning-based channel prediction for LEO satellite massive MIMO communication system," *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1835–1839, 2021.

[20] X. Qiang, L. You, K.-X. Li, C. G. Tsinos, W. Wang, X. Gao, and B. Ottersten, "Hybrid A/D precoding for downlink massive MIMO in LEO satellite communications," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021, pp. 1–6.

[21] L. You, X. Qiang, K.-X. Li, C. G. Tsinos, W. Wang, X. Gao, and B. Ottersten, "Hybrid analog/digital precoding for downlink massive MIMO LEO satellite communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 5962–5976, 2022.

[22] Z. Lin, M. Lin, B. Champagne, W.-P. Zhu, and N. Al-Dhahir, "Secrecy-energy efficient hybrid beamforming for satellite-terrestrial integrated networks," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 6345–6360, 2021.

[23] Z. Lin, K. An, H. Niu, Y. Hu, S. Chatzinotas, G. Zheng, and J. Wang, "Slnr-based secure energy efficient beamforming in multibeam satellite systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 2, pp. 2085–2088, 2023.

[24] K. An, M. Lin, J. Ouyang, and W.-P. Zhu, "Secure transmission in cognitive satellite terrestrial networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 11, pp. 3025–3037, 2016.

[25] Z. Lin, H. Niu, K. An, Y. Wang, G. Zheng, S. Chatzinotas, and Y. Hu, "Refracting ris-aided hybrid satellite-terrestrial relay networks: Joint beamforming design and optimization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 4, pp. 3717–3724, 2022.

[26] W. Hong, Z. H. Jiang, C. Yu, J. Zhou, P. Chen, Z. Yu, H. Zhang, B. Yang, X. Pang, M. Jiang, Y. Cheng, M. K. T. Al-Nuaimi, Y. Zhang, J. Chen, and S. He, "Multibeam antenna technologies for 5G wireless communications," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6231–6249, 2017.

[27] A. G. Kanatas and A. D. Panagopoulos, "Radio wave propagation and channel modeling for earth-space systems," *New York, NY, USA: CRC Press*, 2016.

[28] N. Letzepis and A. J. Grant, "Capacity of the multiple spot beam satellite channel with rician fading," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5210–5222, 2008.

[29] M. Huang, J. Chen, and S. Feng, "Synchronization for OFDM-Based satellite communication system," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 5693–5702, 2021.

[30] H. M. Markowitz and G. P. Todd, *Mean-variance analysis in portfolio choice and capital markets*. John Wiley & Sons, 2000, vol. 66.

[31] R. T. Rockafellar, S. Uryasev *et al.*, "Optimization of Conditional Value-at-Risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[32] Y. Liu, C. Li, J. Li, and L. Feng, "Robust energy-efficient hybrid beamforming design for massive mimo leo satellite communication systems," *IEEE Access*, vol. 10, pp. 63 085–63 099, 2022.

[33] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[34] H. X. Nguyen, R. Trestian, D. To, and M. Tatipamula, "Digital twin for 5G and beyond," *IEEE Communications Magazine*, vol. 59, no. 2, pp. 10–15, 2021.

[35] Y. Yuan, L. Lei, T. X. Vu, Z. Chang, S. Chatzinotas, and S. Sun, "Adapting to dynamic LEO-B5G systems: Meta-critic learning based efficient resource scheduling," *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 9582–9595, 2022.

[36] N. Vlassis, M. Ghavamzadeh, S. Mannor, and P. Poupart, "Reinforcement learning: Adaptation, learning and optimization," 2021.

[37] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[38] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[39] P.-D. Arapoglou, S. Cioni, E. Re, and A. Ginesi, "Direct access to 5G New Radio user equipment from NGSO satellites in millimeter waves," in *2020 10th Advanced Satellite Multimedia Systems Conference and the 16th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, 2020, pp. 1–8.