Deep State-Space Modeling of Channel Dynamics for CSI Prediction in LEO Non-Terrestrial Networks

Iqra Hameed, Md Habibur Rahman, Muhammad Usman, Hyoung-Kyu Song, Muhammad Rafid, Christos Verikoukis

***Abstract*—Non-terrestrial networks (NTNs) are emerging as a key enabler for global wireless coverage, yet high mobility, long propagation delays, and pronounced Doppler effects complicate the timely and accurate acquisition of channel state information (CSI). Predictive CSI modeling provides a practical solution by forecasting future channel conditions from historical observations, reducing dependence on feedback, and improving beamforming performance. In this work, we propose an enhanced predictive CSI framework for NTN based on the structured state-space for sequences (S4) model. We extend the standard S4 by introducing Doppler-Regularized S4 (DR-S4), which incorporates domain knowledge of satellite-induced Doppler shifts into the eigenvalues of the state matrix, ensuring physically consistent oscillatory modes and stable long-term predictions. To further improve system-level performance, we integrate a task-aware objective that directly optimizes downlink rate via differentiable beamforming loss, and an uncertainty-gated predict-or-feedback mechanism that dynamically balances prediction and CSI feedback to reduce overhead. We evaluate the proposed framework in a 3GPP-compliant NTN simulation environment with realistic mobility, fading, and Doppler conditions. Comparative experiments with traditional sequence models demonstrate that DR-S4 achieves superior prediction accuracy, faster convergence, and improved throughput performance across various antenna array sizes and uncertainty thresholds. These results highlight the potential of the proposed DR-S4 framework for enhancing beamforming reliability and spectral efficiency in practical NTN deployments.**

***Index Terms*—Non-terrestrial networks (NTN), low earth orbit (LEO), sequential state-space structures (S4) modeling, channel prediction, beamforming, energy efficiency, task-aware objective, mobility robustness, Doppler regularization, feedback reduction**

Copyright (c) 2026 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2020R1A6A1A03038540), and by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2025-RS-2023-00254529) grant funded by the Korea government(MSIT), and by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2021-II211816), and by the EC funded projects under the SNS-JU program ADROIT6G (101095363) and SUNRISE-6G (101139257).

Iqra Hameed, Md Habibur Rahman, Muhammad Usman and Hyoung-Kyu Song are with Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea (e-mail: engriqrahameed056@gmail.com; habibur@sejong.ac.kr; m_usman@sju.ac.kr; songhk@sejong.ac.kr).

Muhammad Rafid is with ISI/ATHINA, Patras, Greece. (email: mrafid@isi.gr).

Christos Verikoukis is with ISI/ATHINA, Patras, Greece, also with CEID, University of Patras, 265 04 Patras, Greece, and also with Iquadrat Informatica S. L., 08006 Barcelona, Spain (e-mail: cveri@upatras.gr).

Hyoung-Kyu Song is the corresponding author.

I. Introduction

ACCESS to reliable and high-speed wireless communication remains a significant challenge in rural and remote areas across the globe. According to the international telecommunication union (ITU), nearly 2.6 billion people still lack regular internet access, with rural regions facing the greatest connectivity gap [1]. To bridge this divide, non-terrestrial networks (NTNs), particularly those based on low earth orbit (LEO) satellites are gaining global attention. With satellites orbiting at altitudes between $500\,\mathrm{km}$ and $2000\,\mathrm{km}$, LEO systems offer low-latency, wide-area coverage and global accessibility, making them ideal for extending connectivity to areas where terrestrial infrastructure is limited or infeasible [2].

In rural NTN environments, the wireless channel is typically dominated by a strong line-of-sight (LoS) component, especially when directional antennas are used. Unlike urban areas where rich scattering causes severe multipath effects, rural terrain introduces minimal reflection and diffraction. As a result, the channel state information (CSI) in such scenarios evolves in a structurally smooth and predictable manner over time. This behavior opens the door to applying machine learning for channel prediction. If the temporal behavior of the channel can be learned from past observations, it is possible to forecast future CSI without requiring frequent pilot or feedback-based estimation. This idea is particularly useful in LEO-based NTN systems, where the high orbital velocity of satellites (typically around $7.5\,\mathrm{km\,s^{-1}}$) results in rapidly changing channel conditions, including strong Doppler shifts and frequent beam realignment [3]. While the channel varies quickly, the line-of-sight (LoS) component under stable rural geometries often exhibits predictable dynamics when viewed over small time windows. With proper Doppler compensation and beam tracking, the residual channel variations become sufficiently smooth to support long-horizon CSI prediction. This enables the transmitter to anticipate the CSI ahead of time, thus reducing the frequency of CSI updates.

Reducing CSI update overhead is essential in NTNs. Frequent training and feedback transmissions consume bandwidth and battery power, particularly for energy-constrained terminals in remote areas. Therefore, predictive CSI models offer a compelling solution. By forecasting the CSI for multiple future time steps, these models can maintain beamforming performance while significantly lowering signaling and energy costs. Many researchers have explored deep learning techniques for CSI prediction [4]–[8]. Recurrent neural networks (RNNs) are designed for time-series modeling and can capture short-term

channel dynamics [9], [10]. Long short-term memory (LSTM) networks improve upon RNNs by addressing the vanishing gradient problem, allowing them to store longer dependencies [11]. Transformer models use attention mechanisms to learn temporal patterns without sequential processing delays, making them suitable for medium-range prediction [12]–[14]. However, these models have drawbacks, they require large datasets and significant computational resources. More importantly, they struggle with long-horizon prediction, which is essential for reducing CSI feedback frequency in fast-varying NTN scenarios.

Despite their success, existing deep-learning-based CSI predictors for NTNs are largely data-driven and implicitly assume that temporal channel dynamics can be learned solely from observed samples. These models do not explicitly incorporate the underlying physical mechanisms governing NTN channels, particularly satellite-induced Doppler effects. As a result, their learned latent dynamics may deviate from physically consistent frequency evolution, leading to instability and performance degradation when extrapolating over long prediction horizons. Furthermore, conventional recurrent and attention-based architectures are not specifically designed to preserve oscillatory dynamics induced by Doppler shifts, limiting their ability to model long-term phase evolution in high-mobility NTN environments.

Recently, a new class of models based on structured state-space representations has emerged, designed specifically for efficient and scalable learning over long sequences. Among these, the structured state-space sequence (S4) model introduced by Gu et al. [15] stands out. S4 addresses key limitations of traditional sequence models such as RNNs, LSTMs, and Transformers. Unlike RNNs, which process inputs sequentially and suffer from slow inference, or Transformers, which require quadratic memory for attention computations, S4 leverages control-theoretic principles to model sequences via internal hidden states that evolve according to structured system dynamics. At its core, S4 represents a sequence through a state-space formulation using diagonal matrices combined with low-rank updates, enabling efficient computation. This design allows the model to scale linearly with sequence length in both time and memory. Furthermore, by exploiting the convolution theorem and Fast Fourier transform (FFT) techniques, S4 performs fast, parallel computations over long sequences. Consequently, the model is capable of capturing both short-term transitions and long-range dependencies with low latency and minimal resource usage.

In LoS dominated rural NTN environments, the wireless channel varies continuously due to satellite motion and long propagation distances. Accurately modeling these dynamics requires capturing long-term temporal patterns, as the CSI often exhibits correlations spanning multiple time steps. S4's stateful architecture is well-suited for this scenario, allowing it to retain historical information over extended sequences. Unlike Transformers, whose computational and memory costs grow rapidly with sequence length, S4 maintains low resource requirements, making it ideal for real-time processing on satellite payloads or rural user terminals with limited computational capability.

A. Prior Works

Researchers have widely explored machine learning approaches to predict CSI in wireless systems. Among these, deep learning models have shown strong potential for modeling temporal patterns in time-varying channels. Initial approaches adopted deep neural networks (DNNs) to learn direct mappings from past channel states to future values. Turan and Coleri used multi-layer perceptrons (MLPs) to predict CSI in visible light communication (VLC) environments, demonstrating the feasibility of simple neural models in low-mobility scenarios [16]. Yuan et al. extended this idea by applying convolutional neural networks (CNNs) to predict CSI in high-mobility settings, where spatial filtering provided advantages [17]. These models offered fast inference but struggled to learn long-term dependencies effectively.

To address this, researchers introduced RNNs, which model time-series data through recursive structures. Ding and Hirose implemented RNNs for fading channel prediction and showed clear improvements over linear predictors [9]. To enhance the memory capabilities of RNNs, Peng et al. employed LSTM networks, which improved robustness to long sequences and noise in massive MIMO systems [11]. Building on this, Liu et al. proposed convolutional-LSTM (CLSTM) networks that combined spatial feature extraction with temporal memory, achieving improved CSI prediction under mobility [18]. Recently, attention-based models such as Transformers have gained popularity for predictive modeling. Jiang et al. applied a Transformer-based predictor to handle extended sequence modeling under mobile conditions [12]. Wang et al. leveraged Transformer networks in massive MIMO systems to jointly address spatial and temporal prediction challenges [13]. Guo et al. introduced a dual-path Transformer structure for neural beamforming, enabling fine-grained modeling of both signal direction and time correlation [19].

In parallel, many works have focused on prediction-assisted beamforming. Yuan et al. designed deep learning-based beamformers for UAV systems, accounting for challenges like jitter and Doppler shift [20], [21]. Liu et al. developed predictive beamforming techniques based on integrated sensing and communication (ISAC) to support vehicular networks [22]. Additional contributions include Liu et al.'s predictive precoder design for orthogonal time frequency space (OTFS)-enabled systems [23], and Du et al.'s dynamic beamforming methods using CSI prediction in vehicle-to-infrastructure networks [24]. These prior works highlight the progression from basic feedforward architectures to advanced sequence modeling techniques. While they show strong performance in terrestrial networks, their applicability to NTN environments remains limited due to increased CSI dynamics and sequence length requirements. This gap motivates our exploration of the S4 sequence model for NTN-specific CSI forecasting. For example, Ying et al. [25] recently proposed a deep learning framework combining CNN/LSTM-based CSI prediction with multibeam precoding in LEO satellite IoT systems. While effective, their approach still relies on conventional recurrent and convolutional architectures, which lack explicit mechanisms to capture long-horizon Doppler-induced dynamics inherent in

NTN channels.

In addition to channel dynamics, practical wireless systems are also affected by hardware impairments such as phase noise, amplifier nonlinearities, mutual coupling, and RF chain distortions, which can significantly limit performance in high-SNR regimes. Recent studies have investigated these effects in advanced architectures, including stacked intelligent metasurface (SIM)-based holographic MIMO systems for cell-free networks, where RF hardware impairments at access points and user terminals are shown to constrain achievable rates [26]. Hardware-aware beamforming has also been studied in holographic metasurface-assisted multi-altitude LEO satellite networks, where mutual coupling and practical non-idealities impact beamforming performance [27]. Moreover, the impact of phase shift errors and RF hardware impairments has been theoretically characterized for near-field reconfigurable holographic surfaces in cell-free systems, revealing fundamental spectral efficiency limits under non-ideal hardware assumptions [28].

In this work, we focus on predictive CSI modeling and Doppler-induced temporal dynamics in LoS-dominant NTN scenarios, assuming ideal radio-frequency hardware to isolate the impact of long-horizon sequence modeling. Incorporating hardware impairments into the proposed DR-S4 framework is a complementary and important direction for future research.

B. Motivation and Contributions

While prior learning-based CSI prediction methods have demonstrated significant progress, their deployment in NTNs remains challenging. The high mobility of satellites, long propagation delays, and Doppler-induced channel dynamics create substantial difficulties for sequence modeling. RNN-based approaches often fail to retain memory over long sequences, a limitation exacerbated under rapid fading and dynamic line-of-sight conditions. LSTM networks partially address this issue but still suffer from memory degradation over extended horizons. Transformer-based architectures, with their attention mechanisms, can capture long-range dependencies effectively; however, their quadratic complexity in sequence length results in prohibitive computational and memory costs, rendering them impractical for deployment on NTN user terminals or satellite payloads with limited resources.

These challenges underscore the need for a predictive modeling approach that efficiently captures long-range temporal dependencies while remaining computationally tractable. Motivated by this, we investigate S4 models for NTN channel prediction. To the best of the authors' knowledge, this is the first work to explore the use of S4 models in wireless systems, particularly for CSI prediction in NTNs. Unlike prior S4 variants, our Doppler-Regularized S4 (DR-S4) is explicitly physics-informed, embedding domain knowledge of LEO orbital dynamics into the state-space representation. By initializing and constraining eigenvalues around expected Doppler frequencies, DR-S4 transforms the purely data-driven S4 into a hybrid signal-processing and deep learning model. This design ensures that the learned latent dynamics remain consistent with physical channel oscillations, yielding stable long-horizon predictions that conventional DL models cannot sustain. To the best of our knowledge, this is the first NTN CSI predictor that integrates orbital Doppler physics into a deep sequence model.

To address the identified gaps, our main contributions are summarized as follows:

- We extend the standard S4 model [15] by developing a novel **Doppler-Regularized S4** framework for predictive

TABLE I
LIST OF NOTATIONS

Symbol	Description
t	Discrete time index
N_t	Number of transmit antennas at the satellite
$\mathbf{h}_t \in \mathbb{C}^{N_t \times 1}$	Downlink channel vector at time t
$\hat{\mathbf{h}}_t$	Predicted CSI vector at time t
$\mathbf{x}_t \in \mathbb{C}^{N_t \times 1}$	Transmitted beamformed signal
n_t	Additive white Gaussian noise
σ^2	Noise variance
f_c	Carrier frequency
c	Speed of light
$\mathbf{v}_{\text{sat}}(t)$	Satellite velocity vector at time t
$\theta(t)$	Satellite elevation angle
$f_{D,l}(t)$	Doppler frequency of the m-th path
M	Number of multipath components
$\alpha_m(t)$	Complex gain of the m-th path
ϕ_m	Initial phase of the m-th path
$\mathbf{a}(\theta)$	Array response vector
K	CSI history length (input sequence length)
$\mathcal{F}_\theta(\cdot)$	Learned CSI prediction function
$\mathcal{L}_{\text{MSE}}$	Mean squared error loss
$x(t) \in \mathbb{R}^N$	Continuous-time S4 hidden state
$u(t)$	Input signal to the S4 model
$y(t)$	Output of the S4 model
A, B, C	State-space matrices
N	State dimension of the S4 model
Λ	Diagonal eigenvalue matrix of A
P, Q	Low-rank matrices in NPLR parameterization
r	Rank of the low-rank decomposition
λ_i	Initialized eigenvalue of the state matrix
μ_i	Learned eigenvalue during training
γ_i	Stability parameter for real part of eigenvalues
Δt	Sampling interval
L_{dop}	Doppler regularization loss
α, β	Doppler regularization weights
w_t	Beamforming vector at time t
$R(t)$	Achievable downlink rate
$\mathcal{L}_{\text{beam}}$	Beamforming (rate-based) loss
$\mathcal{L}_{\text{total}}$	Total training loss
η_1, η_2	Loss weighting coefficients
ζ_t^2	Prediction uncertainty at time t
ζ_{th}	Uncertainty threshold
χ	Prediction horizon
B	Training batch size
T_{train}	Training complexity
T_{infer}	Inference complexity

CSI modeling. DR-S4 integrates domain knowledge of satellite-induced Doppler shifts by initializing the eigenvalues of the state matrix to align with expected Doppler frequencies. A Doppler regularization term during training constrains the learned eigenvalues to remain close to these frequencies, thereby enhancing long-horizon prediction stability and accuracy in highly dynamic NTN channels.

- We introduce a **task-aware objective function** that directly optimizes the predicted CSI for downlink rate performance, moving beyond traditional mean squared error (MSE) minimization and aligning model training with system-level communication objectives.
- We propose an **uncertainty-gated predict-or-feedback mechanism** that adaptively decides whether to rely on model predictions or request fresh CSI feedback. This strategy reduces feedback overhead while ensuring robust performance in scenarios with high channel variability.
- We implement DR-S4 within a practical S4-based architecture for CSI prediction in NTN systems. The model captures the temporal evolution of the downlink channel, leveraging the efficient memory representation of S4 to forecast future CSI values. Extensive simulations, conducted in a 3GPP-compliant LEO NTN environment, evaluate performance under realistic satellite mobility, Doppler, and fading conditions, and demonstrate the effectiveness of DR-S4 across varying antenna array sizes and uncertainty thresholds.

The remainder of this paper is organized as follows. Section II presents the system model and problem formulation for CSI prediction in LEO-based NTN scenarios. Section III introduces the proposed Doppler-Regularized S4 (DR-S4) framework, including the structured state-space formulation, Doppler-aware eigenvalue initialization and regularization, task-aware training objective, uncertainty-gated feedback mechanism, and computational complexity analysis. For clarity, the main notations used throughout Section III are summarized in Table I. Section IV describes the simulation setup and evaluation methodology and discusses the numerical results and provides insights into the performance gains and robustness of the proposed approach. Finally, Section V concludes the paper and outlines directions for future work.

II. System Model and Problem formulation

We consider a downlink NTN architecture where a single LEO satellite communicates with multiple ground users using a multiple-input single-output (MISO) configuration[1]. The satellite is equipped with a uniform linear array (ULA) consisting of N_t transmit antennas, while the user terminal has a single receive antenna. The system operates in time division duplexing (TDD) mode and uses frame-based transmission as shown in Figure 1. Let $\mathbf{h}_t \in \mathbb{C}^{N_t \times 1}$ denote the channel vector between the satellite and the user at discrete time t. The received signal at the user can be written as:

$$y_t = \mathbf{h}_t^H \mathbf{x}_t + n_t, \tag{1}$$

where $\mathbf{x}_t \in \mathbb{C}^{N_t \times 1}$ is the transmitted beamformed signal, and $n_t \sim \mathcal{CN}(0, \sigma^2)$ represents complex Gaussian noise.

Since the satellite transmits data using beamforming, accurate CSI is essential. However, the long round-trip delay and rapid motion in NTNs make it difficult to obtain instantaneous CSI feedback. To address this limitation, we focus on predicting future CSI based on previously observed channel states.

A. 3GPP-Compliant NTN Channel Model

To simulate the wireless environment in NTNs, we follow the 3GPP TR 38.811 channel model [29], which extends conventional terrestrial models by incorporating satellite mobility, altitude, and Doppler effects. In this model, the satellite follows a LEO trajectory, while the user is assumed to be fixed at ground level. The downlink channel between the satellite and the user is modeled as a sum of M multipath components:

$$\mathbf{h}_t = \sum_{m=1}^{M} \alpha_m(t) \mathbf{a}(\theta_m^{\text{AoD}}(t)), \tag{2}$$

where $\alpha_m(t)$ is the complex path gain of the l-th component, $\theta_m^{\text{AoD}}(t)$ is the angle of departure (AoD) at the satellite array, and $\mathbf{a}(\cdot)$ denotes the array response vector of the uniform linear array (ULA).

Each path gain $\alpha_m(t)$ includes Doppler rotation and is modeled as:

$$\alpha_m(t) = \sqrt{P_m} e^{j(2\pi f_{D,m}(t)t + \phi_m)}, \tag{3}$$

where P_m is the average path power, $f_{D,m}(t)$ is the Doppler frequency shift at time t, and ϕ_m is a random initial phase.

The Doppler shift is determined by the relative motion between the satellite and the user. Assuming the user is stationary and the satellite moves in a circular orbit, the Doppler shift is computed as:

$$f_{D,m}(t) = \frac{f_c}{c} \mathbf{v}_{\text{sat}}(t) \cdot \hat{\mathbf{d}}_m(t), \tag{4}$$

where $\mathbf{v}_{\text{sat}}(t)$ is the satellite velocity vector, and $\hat{\mathbf{d}}_m(t)$ is the unit line-of-sight vector from the satellite to the user for the m-th path. This formulation captures the time-varying Doppler shifts due to satellite orbital motion. This channel model reflects key NTN propagation characteristics, including free-space path loss, dynamic Doppler effects, and time-varying LoS geometry.

B. Problem Formulation: Predictive CSI in NTN

Our objective is to forecast the future channel state at time t using a sequence of past channel states observed over a fixed prediction window. Let $\{\mathbf{h}_{t-K}, \ldots, \mathbf{h}_{t-1}\}$ represent the

[1]Although this work focuses on a narrowband MISO channel model to isolate and study temporal Doppler dynamics in NTN scenarios, the proposed DR-S4 framework is not inherently limited to flat-fading channels. In wideband OFDM-based NTN systems, each subcarrier experiences a narrowband channel with similar Doppler evolution. Consequently, DR-S4 can be applied on a per-subcarrier basis or extended to jointly model groups of subcarriers using a shared state-space representation. A full wideband extension is left for future work.

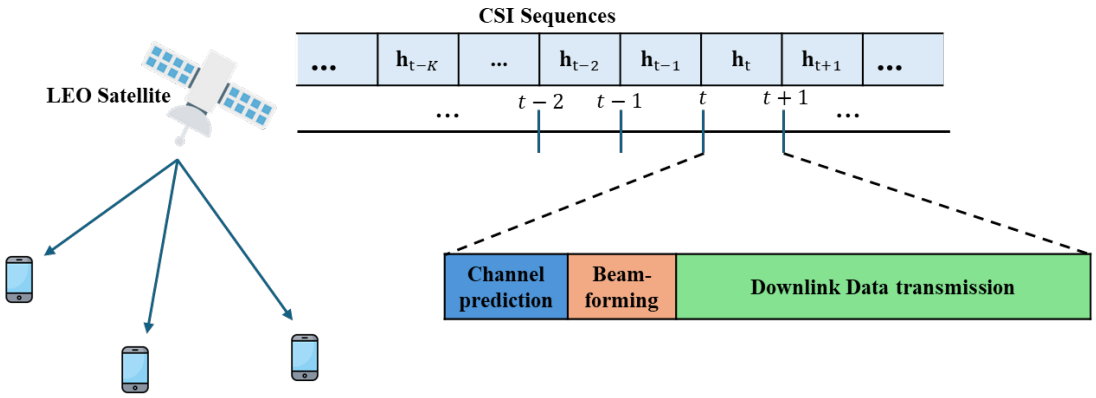

Fig. 1. System model of the proposed predictive beamforming in a non-terrestrial network (NTN) scenario.

known channel history of length K. The prediction task is to estimate $\mathbf{h}_t$ based on this past information [2].

We aim to learn a function $\mathcal{F}_\theta$, parameterized by trainable weights θ, such that the predicted channel vector $\hat{\mathbf{h}}_t$ satisfies:

$$\hat{\mathbf{h}}_t = \mathcal{F}_\theta(\mathbf{h}_{t-K}, \ldots, \mathbf{h}_{t-1}). \tag{5}$$

The training objective is to minimize the mean squared error between the predicted and actual channel vectors.

$$\mathcal{L}_{MSE} = \mathbb{E}\left[\|\hat{\mathbf{h}}_t - \mathbf{h}_t\|^2\right]. \tag{6}$$

The function $\mathcal{F}_\theta$ can be instantiated using any temporal modeling architecture, such as LSTM, Transformer, or the S4 model. In this work, we focus on the S4 model, which is designed to efficiently capture long-range temporal dependencies in CSI sequences while maintaining linear time and memory complexity. This capability makes it a strong candidate for deployment in NTN systems, where long prediction windows and resource constraints are common.

This prediction mechanism supports proactive beamforming and resource allocation without relying on real-time CSI feedback. As a result, it enhances link reliability and spectral efficiency in NTN environments that suffer from latency and mobility-induced challenges.

III. Structured State-Space Sequence Modeling for NTN Channel Prediction

Since structured state-space sequence models are relatively new to the wireless communications community, we briefly summarize the key S4 derivations required to understand the proposed Doppler-regularized formulation.

A. Structured State-Space Sequence (S4) Model

The S4 model is grounded in control theory and provides an efficient method for modeling long-range dependencies in sequential data. At its core, S4 begins with a continuous-time state-space model:

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t), \tag{7}$$

where $x(t) \in \mathbb{R}^N$ is the hidden state, $u(t)$ is the input signal, and $y(t)$ is the output. The matrices A, B, and C are learnable and govern the dynamics of the system. For time-discrete applications such as CSI prediction, this system is discretized, resulting in the recurrence:

$$x_k = Ax_{k-1} + Bu_k, \quad y_k = Cx_k, \tag{8}$$

which shares a structural similarity with RNNs, though it is mathematically more expressive and stable.

However, computing powers of the matrix A naively is computationally expensive, requiring $\mathcal{O}(N^2L)$ operations for a sequence of length L, where N is the state dimension. This approach also risks numerical instability, especially when A is ill-conditioned or non-normal.

a) Normal Plus Low-Rank (NPLR) Parameterization: A critical bottleneck in applying classical state-space models to long sequences lies in the cost of computing powers and inverses of the state matrix A. For arbitrary dense matrices, evaluating the state-transition convolution kernel involves $\mathcal{O}(N^2L)$ complexity for a sequence of length L and hidden state dimension N. Furthermore, if A is not normal (i.e., does not commute with its adjoint), the system may become numerically unstable.

To overcome these challenges, the S4 model adopts the *Normal Plus Low-Rank* (NPLR) form for the matrix A, written as:

$$A = \Lambda - PQ^*, \tag{9}$$

where $\Lambda \in \mathbb{C}^{N\times N}$ is a diagonal matrix with strictly negative real parts (ensuring stability), and $P, Q \in \mathbb{C}^{N\times r}$ are low-rank matrices, with $r \ll N$. This representation guarantees that A is unitarily diagonalizable and spectrally well-behaved. The use of diagonal Λ ensures that the eigenvalues are decoupled across the state dimensions, while the low-rank update PQ^* introduces global interactions in a structured way.

The convolutional representation of the state-space system

[2]In our framework, the CSI history used as input to the prediction model is assumed to be acquired through traditional pilot-based estimation methods, as supported by current 3GPP-compliant NTN specifications [29]. Specifically, we simulate CSI sampled at 1 ms intervals (1 kHz), consistent with the frame-based transmission structure used in satellite systems [29], [30]. This setting reflects realistic NTN architectures that support periodic CSI acquisition using downlink pilots. We do not use recursively predicted CSI to construct input sequences, thereby avoiding cumulative error.

in the frequency domain is given by the transfer function:

$$\hat{T}(z; A, B, C) = C^*(I - zA)^{-1}B, \tag{10}$$

where z is a point on the unit circle in the complex plane, and $\hat{T}(z)$ is the frequency-domain representation of the convolution kernel. This kernel describes the system's impulse response over time and is used to compute output sequences from input sequences efficiently via FFT methods.

A naive inversion of $(I - zA)$ for every frequency point is expensive. However, because A is in NPLR form, we can compute the inverse efficiently using the Woodbury matrix identity [31]. Applying it yields:

$$(I - z(\Lambda - PQ^*))^{-1} = D - DP(I + Q^*DP)^{-1}Q^*D, \tag{11}$$

where $D = (I - z\Lambda)^{-1}$ is easy to compute since Λ is diagonal. The term $(I + Q^*DP)$ is a small $r \times r$ matrix due to the low-rank structure, and its inverse can be computed in $\mathcal{O}(r^3)$ time, which is negligible when r is small (e.g., $r = 1, 2, 4$ in practice).

This structure reduces the computation of each frequency component of $\hat{T}(z)$ to $\mathcal{O}(N + r^2)$ time, and since FFT-based methods evaluate the kernel at a fixed grid of L points (roots of unity), the total cost to generate the convolution kernel becomes $\mathcal{O}(L(N + r^2))$. This is nearly linear in sequence length and drastically faster than conventional dense recurrent or attention-based mechanisms.

Finally, after computing $\hat{T}(z)$ across all z using FFT grids, the inverse FFT is applied to obtain the time-domain convolution kernel T, which is then used to convolve with the input signal. This process ensures both numerical stability and efficient sequence transformation. The NPLR parameterization enables S4 to model long-range dependencies efficiently, providing a solution that is both expressive and computationally practical for tasks like CSI prediction in NTN systems.

B. Doppler-Regularized S4 (DR-S4)

In the context of NTN systems, particularly with LEO satellites, the channel's time-varying dynamics are heavily influenced by Doppler shifts, which arise from the high orbital velocity of the satellite. These Doppler-induced shifts lead to predictable oscillatory behavior in the channel, which can be captured in the S4 model through careful eigenvalue initialization.

To this end, we propose the *Doppler-Regularized S4 (DR-S4) model*, which anchors the poles of the S4 model to the expected Doppler frequencies based on the satellite's velocity and the geometry of the channel. This modification ensures that the model captures the physical dynamics of the Doppler shifts, improving stability, convergence, and long-horizon prediction accuracy. By aligning the internal model states with the expected Doppler shifts, DR-S4 can predict long-term channel behavior more effectively than traditional S4.

We begin by computing the expected Doppler frequencies from the satellite's motion and the angle between the satellite and the user. Specifically, we use the following expression for the Doppler shift:

$$f_{D,m}(t) = \frac{f_c}{c}\mathbf{v}_{\text{sat}}(t) \cdot \hat{\mathbf{d}}_m(t),$$

where f_c is the carrier frequency, c is the speed of light, $\mathbf{v}_{\text{sat}}(t)$ is the satellite's velocity vector, and $\hat{\mathbf{d}}_l(t)$ is the unit line-of-sight (LoS) vector from the satellite to the user terminal. In practice, for LoS-dominated NTN channels, this expression can be approximated as:

$$f_D(t) \approx \frac{v_{\text{sat}}(t)\cos\theta(t)}{\lambda_c}, \tag{12}$$

where $v_{\text{sat}}(t)$ is the satellite's velocity, $\theta(t)$ is the elevation angle between the satellite and the user terminal, and λ_c is the wavelength corresponding to the carrier frequency f_c.

The Doppler shift $f_D(t)$ forms the basis for initializing the eigenvalues of the S4 model. Each eigenvalue is initialized such that its imaginary part corresponds to the Doppler-induced angular frequency:

$$\text{Im}(\lambda_i) = 2\pi f_i \Delta t, \tag{13}$$

where f_i is selected from a set of Doppler frequencies and Δt is the system's sampling interval. We introduce small perturbations around the dominant Doppler shift to account for slight variations due to geometry changes or residual multipath effects. This approach ensures that the model accurately reflects both the dominant and the finer-scale Doppler shifts.

The real part of each eigenvalue is set to a small negative value $-\gamma_i$, where γ_i is chosen to be a small constant (typically between 0.01 and 0.1). This ensures that the system remains stable by placing the poles of the state matrix in the stable region of the complex plane. This initialization helps prevent numerical instability, ensuring that the state-space system has decaying impulse responses and stabilizes quickly during training.

As the model trains, the low-rank term PQ^* adjusts the eigenvalues, allowing the model to learn from the data. However, without regularization, these eigenvalues might deviate too far from the expected Doppler frequencies. To address this, we introduce a Doppler regularization term that penalizes large deviations from the expected Doppler frequencies. The regularizer is given by:

$$\mathcal{L}_{\text{dop}} = \frac{1}{N}\sum_{i=1}^{N}\left[\alpha(\text{Im}(\mu_i) - 2\pi f_i)^2 + \beta\max(0, |\mu_i| - 1)^2\right] \tag{14}$$

where μ_i are the learned eigenvalues of the state matrix. The first term penalizes deviations in the imaginary part of the eigenvalues, ensuring that the learned frequencies align closely with the expected Doppler frequencies. The second term prevents the magnitude of the eigenvalues from exceeding the stable region of the complex plane, ensuring system stability.

The regularization weights α and β are crucial for controlling the balance between physical fidelity to the Doppler shifts and flexibility for the model to adapt to the data. A larger value of α enforces a stronger penalty on the deviations from the expected Doppler shifts, while a larger β provides more

stringent stability constraints. These weights were selected through empirical validation to achieve the best trade-off between prediction accuracy and stability.

a) Sensitivity and Practical Considerations: The proposed Doppler-regularized formulation is designed for LoS-dominant NTN scenarios, as discussed in the Section I, where satellite-induced Doppler effects constitute the temporal channel variation. In this context, the Doppler frequencies used for eigenvalue initialization serve as informative priors rather than exact constraints. The regularization term is applied softly, allowing the learned low-rank dynamics to adapt the eigenvalues during training when moderate inaccuracies in the Doppler prior arise due to geometry estimation errors or residual motion uncertainty. It is important to note that the regularizer does not enforce strict frequency locking; instead, it biases the model toward physically plausible oscillatory behavior while preserving flexibility to fit the observed CSI data. In scenarios where additional channel components beyond the dominant LoS path are present, their influence is implicitly captured through the data-driven low-rank corrections rather than being explicitly modeled. The regularization weights α and β control the strength of this bias, where larger values emphasize physical consistency and stability, and smaller values allow greater adaptability. In practice, these parameters are selected via validation to balance robustness and long-horizon prediction stability within the intended LoS-dominant operating regime.

C. Task-Aware Objective and CSI Prediction

To further enhance the performance of the DR-S4 model in the context of predictive CSI for NTN systems, we introduce a task-aware objective that optimizes the model's performance for the downstream task of beamforming (or throughput) rather than just minimizing MSE. Standard MSE loss does not directly correspond to system-level metrics such as throughput or rate. By incorporating a differentiable beamforming loss, we ensure that the model is optimizing for link performance, i.e., the downlink rate.

For beamforming, we compute the beamforming vector $\mathbf{w}(\hat{\mathbf{h}}_t)$ from the predicted CSI $\hat{\mathbf{h}}_t$ as:

$$\mathbf{w}(\hat{\mathbf{h}}_t) = \frac{\hat{\mathbf{h}}_t}{\|\hat{\mathbf{h}}_t\|} \tag{15}$$

where $\mathbf{w}(\hat{\mathbf{h}}_t)$ is the beamforming vector and $\hat{\mathbf{h}}_t$ is the predicted CSI. The spectral efficiency or rate is then calculated as:

$$\mathrm{R}(t) = \log_2\left(1 + \frac{p_t|\mathbf{h}_t^H \mathbf{w}(\hat{\mathbf{h}}_t)|^2}{\sigma^2}\right), \tag{16}$$

where $\mathbf{h}_t$ is the true channel. The beamforming loss over a prediction horizon χ is then the average rate across χ time steps:

$$\mathcal{L}_{\text{beam}} = \frac{1}{\chi}\sum_{\tau=1}^{\chi} \mathrm{R}(t+\tau). \tag{17}$$

Hence, the overall final loss for training becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \eta_1 \mathcal{L}_{\text{dop}} + \eta_2 \mathcal{L}_{\text{beam}} \tag{18}$$

where $\mathcal{L}_{\text{MSE}}$ is the standard MSE loss in Eq. 6 and L_{dop} is the Doppler regularize term in Eq. 14 . The weights η_1 and η_2 control the importance of the beamforming and Doppler terms, ensuring that both prediction accuracy and system performance objectives are optimized.

The task-aware loss function is integrated into the end-to-end training process by combining both components: the standard MSE loss for prediction accuracy and the task-specific loss to optimize for real-world deployment metrics. The total loss is minimized during training, and the backpropagation algorithm updates the model parameters to reduce both prediction error and task-specific performance loss simultaneously. In our implementation, we use the total loss as the guiding objective for training the model, allowing it to jointly optimize for both accuracy and system-level performance. This ensures that the model not only predicts correctly but also performs well on the specific task (e.g., maximizing throughput, reducing energy usage, etc.).This task-aware loss function allows DR-S4 to be trained end-to-end for real-time CSI prediction and beamforming tasks, optimizing not only for accurate CSI estimation but also for system-level objectives such as the downlink rate.

D. Uncertainty-Gated Predict-or-Feedback Controller

In NTN systems, predictive CSI modeling reduces the need for frequent feedback, but blindly using model predictions can lead to performance degradation under rapidly changing channel conditions. To address this, we introduce an uncertainty-gated predict-or-feedback controller that dynamically decides whether to rely on the predicted CSI of the DR-S4 model or to request new feedback. The key idea is to quantify the uncertainty in the predicted channel and to make feedback decisions based on this confidence level.

The uncertainty of the DR-S4 prediction is estimated using either *Monte Carlo (MC) dropout* or a *model ensemble*. In MC dropout, dropout is applied at inference time, and the model is executed multiple times to generate a set of predictions $\hat{\mathbf{h}}_t^{(m)}$, for $m = 1, \ldots, M$. The mean prediction is computed as

$$\bar{\mathbf{h}}_t = \frac{1}{M}\sum_{m=1}^{M} \hat{\mathbf{h}}_t^{(m)}, \tag{19}$$

and the variance across these predictions provides a measure of the model's uncertainty:

$$\zeta_t^2 = \frac{1}{M}\sum_{m=1}^{M} \left\|\hat{\mathbf{h}}_t^{(m)} - \bar{\mathbf{h}}_t\right\|_2^2. \tag{20}$$

A high variance ζ_t^2 indicates that the model is uncertain about the predicted CSI, while a low variance reflects high confidence.

Once the uncertainty is calculated, it is compared with a predefined threshold ζ_{th}. The threshold for uncertainty (ζ_{th}) was selected empirically based on the validation data, balancing the need for accurate predictions and feedback efficiency. A lower threshold would trigger uncertainty gating more frequently, leading to higher prediction accuracy but reduced throughput, while a higher threshold would allow

for faster inference but potentially reduce the accuracy of uncertain predictions. In our experiments, we determined that the optimal threshold minimized both False positive rate and ensured the system could maintain high throughput, even with uncertain predictions.

If the uncertainty is below this threshold, the model is considered confident, and the predicted CSI $\hat{\mathbf{h}}_t$ is used directly to form the beam without asking for feedback. In contrast, if the uncertainty exceeds the threshold, feedback is requested and the true CSI $\mathbf{h}_t$ is provided, either to update the model or directly for beamforming. This decision can be formally expressed as follows.

$$\text{Feedback}_t = \begin{cases} 0, & \text{if } \zeta_t^2 < \zeta_{th}, \\ 1, & \text{if } \zeta_t^2 \geq \zeta_{th}, \end{cases} \tag{21}$$

ensuring that feedback is requested only when necessary. This approach significantly reduces the number of feedback transmissions, saving energy and bandwidth, which is especially important in NTN systems with high satellite mobility.

For multistep predictions, the controller can consider the *cumulative uncertainty* over a prediction horizon of χ steps. The total uncertainty can be weighted across future predictions as follows.

$$\zeta_{t:t+\chi}^2 = \sum_{\tau=1}^{\chi} w_\tau \, \zeta_{t+\tau}^2, \tag{22}$$

where $w_\tau \in [0, 1]$ are weights that can prioritize near-term predictions over distant ones. Feedback is requested if the cumulative uncertainty exceeds the threshold ζ_{th}, allowing the system to maintain reliable performance over longer horizons.

This uncertainty-gated predict-or-feedback strategy provides a principled mechanism to balance prediction and feedback, enabling energy-efficient and reliable CSI acquisition. By dynamically adapting feedback requests based on the model's confidence, the controller ensures accurate beamforming while minimizing unnecessary feedback transmissions, which is critical in NTN scenarios with high satellite mobility and strict energy constraints.

E. DR-S4 Architecture and Its Application to Predictive CSI in NTNs

The DR-S4 transforms classical state-space systems into a trainable deep learning framework capable of modeling long temporal sequences with linear computational complexity. The deep S4 network builds upon the core S4 layer by stacking multiple instances of state-space convolutional blocks, interleaved with non-linear activations and projection layers. This section outlines the full network architecture and its integration into the task of CSI prediction for NTNs.

Each S4 layer applies a global convolution operation to the input sequence. The convolution kernel $\mathcal{K}$ is derived from the diagonal-plus-low-rank state matrix A, as described previously. Given a sequence of input embeddings $\mathbf{u} = \{u_0, u_1, ..., u_{L-1}\}$ of length L, the output of the S4 layer is computed via a 1D convolution:

$$\mathbf{y} = \mathcal{K} * \mathbf{u}, \tag{23}$$

where $\mathcal{K} \in \mathbb{C}^L$ is the learned convolution kernel derived analytically from the system parameters (A, B, C) via FFT-based computation.

This convolution is global every output token depends on the entire input history and its structure encodes temporal dependencies across all scales. Unlike standard CNNs with limited kernel width or recurrent models with local state transitions, the S4 layer can model both short-term transitions and long-range memory with a single operation, making it highly suitable for CSI modeling in NTN environments where the signal dynamics span both fast and slow time scales.

Each layer operates on multiple channels (or feature dimensions), similar to convolutional layers in deep CNNs. To support this, S4 uses depth-wise separable convolutions, where each channel has its own S4 kernel. Following the convolution, a pointwise linear projection (typically a fully connected layer) mixes information across channels. Non-linear activation functions such as ReLU are applied between layers to enable non-linear representation learning. Layer normalization or dropout may also be inserted to aid convergence and regularization during training.

Let $\mathbf{H}_{\text{in}} \in \mathbb{R}^{L \times d}$ be the input to the S4 block, where L is the sequence length and d is the channel dimension. A single S4 block performs:

$$\mathbf{H}_1 = \text{S4Conv}(\mathbf{H}_{\text{in}}), \tag{24}$$

$$\mathbf{H}_2 = \text{Activation}(\mathbf{H}_1), \tag{25}$$

$$\mathbf{H}_3 = \text{Linear}(\mathbf{H}_2), \tag{26}$$

$$\mathbf{H}_{\text{out}} = \text{LayerNorm}(\mathbf{H}_{\text{in}} + \mathbf{H}_3), \tag{27}$$

where the residual connection and layer normalization stabilize training and help with gradient flow.

a) Integration with CSI Prediction in NTNs: We formulate the CSI prediction task as a supervised sequence-to-one learning problem. Given a window of K past CSI observations $\{\mathbf{h}_{t-K}, \ldots, \mathbf{h}_{t-1}\}$, the model is trained to predict the next channel state $\mathbf{h}_t$:

$$\hat{\mathbf{h}}_t = \mathcal{F}_\theta(\mathbf{h}_{t-K}, \ldots, \mathbf{h}_{t-1}), \tag{28}$$

where $\mathcal{F}_\theta$ denotes the full deep S4 model, parameterized by θ.

To prepare the input, we first embed each CSI vector into a higher-dimensional representation space using a learnable input projection:

$$\mathbf{u}_k = W_{\text{in}}\mathbf{h}_k + b_{\text{in}}, \quad k = t-K, \ldots, t-1, \tag{29}$$

where $W_{\text{in}} \in \mathbb{R}^{d \times N_t}$ and $b_{\text{in}} \in \mathbb{R}^d$ are the projection weights and bias, and N_t is the number of transmit antennas (i.e., the dimension of $\mathbf{h}_k$). This results in an input sequence $\mathbf{u} = \{\mathbf{u}_{t-K}, \ldots, \mathbf{u}_{t-1}\}$ of shape $K \times d$.

This embedded sequence is passed through multiple S4 blocks. The final output sequence is pooled using a global average or final-timestep extractor, and passed through an output projection to produce the predicted CSI vector:

$$\hat{\mathbf{h}}_t = W_{\text{out}} \cdot \text{Pool}(\mathbf{H}_{\text{final}}) + b_{\text{out}}, \tag{30}$$

where $W_{\text{out}} \in \mathbb{R}^{N_t \times d}$, and $\mathbf{H}_{\text{final}}$ is the sequence representation after the last S4 layer. The model is trained to minimize $\mathcal{L}_{\text{total}}$

in Eq.

b) Advantages for NTN: The deep S4 model offers several advantages in this context. Its linear scaling in both time and memory with respect to the input sequence length makes it suitable for long-term modeling, even with large windows of historical CSI. Unlike Transformer models, which suffer from quadratic attention overhead, or LSTMs, which degrade over long horizons, the S4 architecture maintains stable gradients and efficient inference. Moreover, because the convolutional kernels in S4 are derived analytically from the state-space system, the model is both expressive and interpretable. This fits well with the structured and physics-informed nature of channel modeling in NTNs, where signals evolve under predictable, though complex, dynamics. By leveraging this architecture, our proposed approach enables accurate and efficient CSI prediction for proactive beamforming in NTN systems, overcoming the limitations of both feedback delay and computational resource constraints.

Algorithm 1 Proposed DR-S4 Predictive CSI Framework

Input: CSI history $\{\mathbf{h}_{t-K}, \ldots, \mathbf{h}_{t-1}\}$, Doppler set $\{f_i\}$, thresholds ζ_{th}, training data $\mathcal{D}$
Training Phase:
  Initialize diagonal matrix Λ with eigenvalues $\lambda_i = -\gamma_i + j2\pi f_i \Delta t$, ensuring $\Re(\lambda_i) < 0$.
  Construct state matrix $A = \Lambda - PQ^*$ with low-rank factors P, Q.
  For each minibatch $(\{h_{t-K}, \ldots, h_{t-1}\}, h_t) \in \mathcal{D}$:
    a) Embed input sequence: $\mathbf{u}_k = W_{in}\mathbf{h}_k + b_{in}$.
    b) Pass through stacked S4 blocks: $H = \text{S4Stack}(U; \Theta)$.
    c) Pool and project: $\hat{h}_t = W_{out} \cdot \text{Pool}(H) + b_{out}$.
    d) Compute losses:
      – $\mathcal{L}_{MSE} = \|\hat{\mathbf{h}}_t - \mathbf{h}_t\|^2$
      – $\mathcal{L}_{dop}$: Doppler regularization on eigenvalues μ_i of A
      – $\mathcal{L}_{beam}$: negative downlink rate via MRT beamforming
    e) Update parameters: $L_{total} = L_{MSE} + \eta_1 L_{dop} + \eta_2 L_{beam}$
Inference Phase:
Embed recent CSI history and predict $\hat{\mathbf{h}}_t$ with DR-S4.
Estimate uncertainty using MC-dropout (M samples): $\bar{h}_t = \frac{1}{M}\sum_m \hat{\mathbf{h}}_t^{(m)}, \quad \zeta_t^2 = \frac{1}{M}\sum_m \|\hat{\mathbf{h}}_t^{(m)} - \bar{\mathbf{h}}_t\|^2$.
  If $\zeta_t^2 < \zeta_{th}$, set $Feedback_t = 0$, compute MRT beamformer: $\mathbf{w}_t = \hat{\mathbf{h}}_t / \|\hat{\mathbf{h}}_t\|$.
  Else, request feedback ($Feedback_t = 1$), set $\mathbf{w}_t = \mathbf{h}_t / \|\mathbf{h}_t\|$. Transmit using $\mathbf{w}_t$.
Output: Predicted CSI $\hat{h}_t$, beamforming vector $\mathbf{w}_t$, feedback decision $Feedback_t$

F. Complexity Analysis

The core computational cost of DR-S4 arises from its sequence processing, where the state-space model handles sequences of length L and a state dimension N. For standard S4, the computational complexity per sequence step is $\mathcal{O}(LN^2)$, which arises from the matrix multiplication involved in the recurrence step. In DR-S4, we add low-rank updates to the model parameters, which introduces an additional cost of $\mathcal{O}(LNr)$, where r is the rank of the low-rank decomposition. Since r is typically much smaller than N (e.g., $r \sim 8$ or 16), the low-rank term contributes only modestly to the overall complexity.

The regularization term (penalizing deviations from the expected Doppler frequencies) incurs an additional computational cost of $\mathcal{O}(LN)$, but it is negligible compared to the dominant recurrence and low-rank update terms. Therefore, the total computational complexity for DR-S4 during training is:

$$T_{\text{DR-S4, train}} = \mathcal{O}(BLN^2 + BLNr)$$

where: - B is the batch size, - L is the sequence length, - N is the state dimension, - r is the rank of the low-rank updates.

At inference time, the complexity reduces to:

$$T_{\text{DR-S4, infer}} = \mathcal{O}(L(N^2 + Nr))$$

Thus, while the introduction of the low-rank term increases complexity marginally, the overall time complexity remains manageable and close to that of the standard S4 model.

The memory footprint of DR-S4 is influenced by the number of model parameters and the activation memory required for sequence processing. The model parameters consist of the weights for the state transition matrix, the low-rank terms P and Q, and the task-aware loss parameters (if applicable). Specifically, the memory for DR-S4's parameters is given by:

$$\text{Memory}_{\text{params}} = N^2 + N + 2Nr$$

For example, with $N = 128$ and $r = 16$, the number of parameters is $\approx 25,000$, which is modest and suitable for deployment on edge devices.

During inference, the memory required for activations is primarily driven by the sequence length L and the state dimension N. For a given sequence length, the memory needed to store activations is $\mathcal{O}(LN)$, which scales linearly with both the sequence length and the state dimension.

In training, the memory requirement increases due to the need to store both the forward and backward passes. However, the low-rank updates and Doppler regularization add only a modest amount of additional memory overhead, making DR-S4 feasible for real-time deployment, even on resource-constrained devices.

Given the memory and computational complexity of DR-S4, its deployment on real-time, resource-constrained devices such as satellite payloads, edge processors, or mobile user terminals is feasible, especially when optimized for fixed-point arithmetic or when low-rank approximations are used. At inference time, DR-S4 is able to process sequences efficiently, and its low memory footprint ensures that it can be used in devices with limited storage and compute capabilities.

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Carrier frequency f_c	14 GHz
Speed of light c	3×10^8 m/s
Wavelength λ_c	c/f_c
Sampling frequency f_s	1000 Hz
Simulation duration	10 s
Satellite altitude h_{sat}	600 km
Satellite init. position	$(0, h_{\text{sat}}, 0)$
Tx antenna configs N_t	{4, 8, 16, 32, 64, 128}

IV. SIMULATION SETUP AND RESULTS

A. Simulation Setup

To evaluate the performance of the proposed DR-S4 predictive CSI framework, we conducted extensive simulations based on NTN downlink MISO scenario. The channel dynamics are generated using a 3GPP-compliant NTN model, considering realistic Doppler and propagation effects due to satellite mobility.

The key simulation parameters are summarized in Table II. The carrier frequency is set at 14 GHz, corresponding to Ka-band satellite communications. A LEO satellite at an altitude of 600 km is modeled, positioned initially at the coordinate $(0, h_{\text{sat}}, 0)$. Ground users are located at the origin. We assume frame-based transmissions sampled at 1000 Hz over a total duration of 10 seconds. A variety of antenna configurations are considered, ranging from small to large array sizes.

The dataset is constructed by simulating the downlink channel vectors over time, including realistic Doppler shifts and fading dynamics. Each sample sequence consists of a fixed-length history window used for training and testing the predictive models. In the following sections, we present quantitative evaluation results comparing the performance of the DR-S4 model with baseline architectures under various configurations.

B. Model Training and Hyperparameters

The predictive DR-S4 model is trained to minimize $\mathcal{L}_{\text{total}}$ between the predicted and ground-truth channel vectors. The training objective is optimized using the Adam optimizer, which adaptively adjusts learning rates for each parameter. An early stopping criterion with a patience of 5 epochs is applied to prevent overfitting, monitoring the validation loss during training. The dataset is constructed by simulating the dynamic downlink channel, generating a total of 10,000 CSI samples under the 3GPP-compliant NTN mobility model. Each sample sequence consists of 50 consecutive past CSI observations, used to predict the future channel state at the next time step.

The main training and model configuration parameters are summarized in Table III.

Each input sample to the network represents a sequence of complex CSI vectors, where real and imaginary parts are separated, effectively doubling the input dimension. Training is performed using GPU acceleration to ensure efficient convergence. Unless otherwise stated, all experimental results in the subsequent sections are based on this setup.

TABLE III
TRAINING AND MODEL HYPER-PARAMETERS

Parameter	Value
Loss function	Mean Squared Error (MSE)
Optimizer	Adam
Batch size	128
Training window size(s)	[50, 2*Transmit Antennae]
Input dimension	$2 \times N_t$ (real and imaginary parts)
Number of CSI samples	10,000
Maximum training epochs	500
Early stopping patience	5 epochs
Device	GPU (CUDA-enabled)

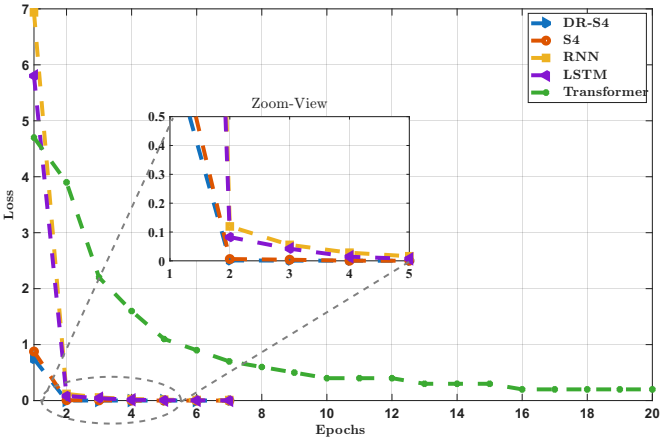

Fig. 2. Model loss vs. epochs for different architectures.

C. Visualization of CSI Prediction Accuracy

Figure 2 illustrates the training loss as a function of epochs for the compared architectures. It is evident that both DR-S4 and S4 achieve rapid and stable convergence, reaching the lowest loss values among all models. This performance reflects the advantage of structured state-space formulations, which directly encode long-range temporal dependencies through well-conditioned recurrent dynamics. By doing so, they alleviate optimization difficulties commonly faced in sequential models and make learning more data-efficient. In contrast, the Transformer converges more slowly and stabilizes at a higher loss level. This slower progress can be attributed to its heavy reliance on self-attention, which introduces a large parameter space and requires extensive training iterations to generalize effectively. The RNN shows an initial sharp drop in loss but quickly saturates, a behavior consistent with vanishing gradient issues that limit its capacity for longer sequences. The LSTM partially mitigates this through gating mechanisms, allowing for better retention of temporal information, yet its final convergence remains inferior to S4-based models. Overall, the results clearly indicate that state-space architectures are better suited for predictive channel modeling, as their structural design not only accelerates convergence but also achieves lower final training loss. This demonstrates their potential to provide more accurate and stable predictions in practical wireless communication scenarios.

Figure 3 shows the training loss curves for DR-S4 and S4 under different transmitter configurations (TX4, and TX16).

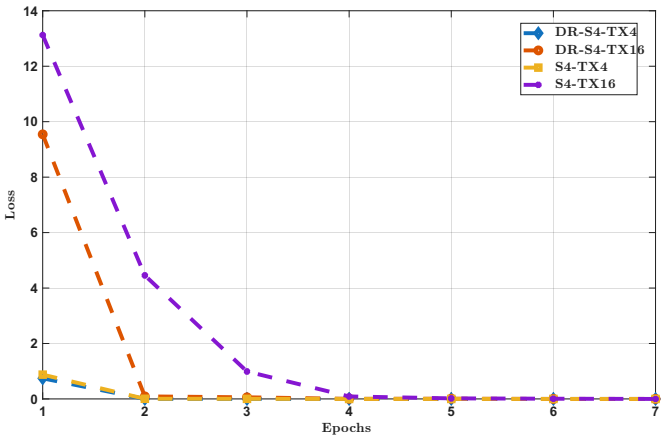

Fig. 3. Model loss comparison across different numbers of TX antennas.

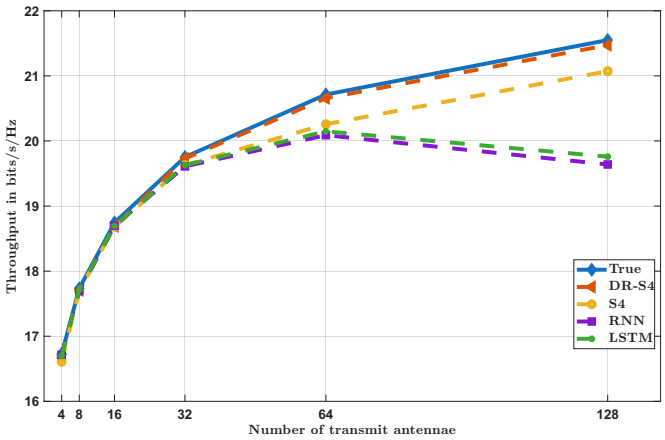

Fig. 5. Predicted vs. true throughput across different TX antennas.

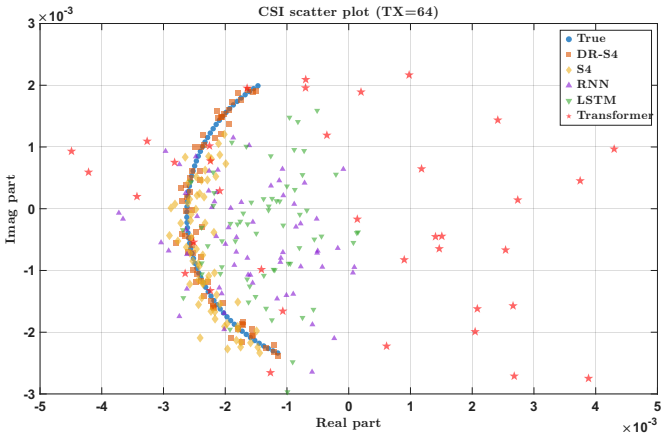

Fig. 4. CSI scatter plot comparing true and predicted outputs.

Across all cases, both models converge rapidly, confirming their robustness in handling varying system dimensions. For lower antenna settings (TX4), the loss drops sharply within the first epoch, reaching near-zero values by the second epoch, which highlights the efficiency of state-space models in capturing channel dynamics even in low-dimensional scenarios. For the higher-dimensional setting (TX16), convergence is slower during the first few epochs, as indicated by the elevated initial loss, but the models still achieve stable and low loss values after sufficient iterations. This behavior is consistent with the increased complexity of the prediction task as the number of antennas grows, requiring the model to learn a larger parameter space. Importantly, DR-S4 demonstrates slightly faster convergence than S4 across all configurations, reflecting its enhanced ability to stabilize long-sequence dynamics through diagonal plus low-rank parameterization. These results emphasize that S4-based architectures not only scale effectively with system size but also maintain strong convergence properties, making them suitable for large-scale MIMO prediction tasks in NTN scenarios.

Figure 4 depicts the scatter plot of the predicted CSI for TX=64 in the complex plane, comparing several learning architectures against the true CSI. The true channel coefficients serve as the reference distribution. Predictions from DR-S4 and S4 align most closely with this true distribution, showing dense clustering around the ground-truth trajectory. This reflects their superior capability to capture both the amplitude and phase variations of the wireless channel. In contrast, the RNN and LSTM predictions exhibit larger dispersion, with noticeable deviations from the reference, indicating residual errors caused by their limited ability to model long-term dependencies. The Transformer produces even wider scattering, with many points spread away from the true curve, which highlights its difficulty in learning stable CSI patterns under the given training regime. The comparative spread of points illustrates an important scientific insight: while conventional recurrent and attention-based models struggle to accurately reconstruct the channel in higher-dimensional settings, state-space architectures such as S4 and DR-S4 maintain precise and robust modeling of the underlying dynamics. This validates their effectiveness for predictive beamforming tasks, where small deviations in CSI estimates can significantly impact system performance.

Figure 5 illustrates the achievable throughput as a function of the number of transmit antennas for different models, with the true CSI serving as the upper bound. As expected, throughput increases monotonically with antenna count due to enhanced spatial diversity and beamforming gains. However, the performance gap between models reflects their CSI prediction accuracy. DR-S4 tracks the true CSI curve most closely, consistently achieving near-optimal throughput across all antenna configurations. S4 also performs well, though it shows a slightly larger deviation at higher antenna counts, suggesting some degradation in modeling precision as system dimensionality grows. In contrast, RNN and LSTM underperform relative to the state-space models, with their throughput saturating earlier and diverging more from the true CSI baseline, which reflects the cumulative effect of their residual prediction errors. These results demonstrate that inaccuracies in predicted CSI directly translate into reduced spectral efficiency, highlighting the critical role of precise sequence modeling. The superior alignment of DR-S4 with the true throughput underscores its suitability for large-scale NTN systems, where even small CSI errors can significantly degrade end-to-end performance.

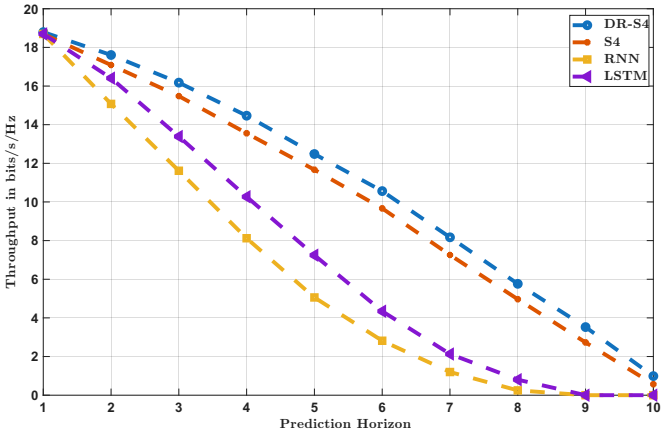

Fig. 6. Predicted throughput vs prediction horizon (χ) for TX=16.

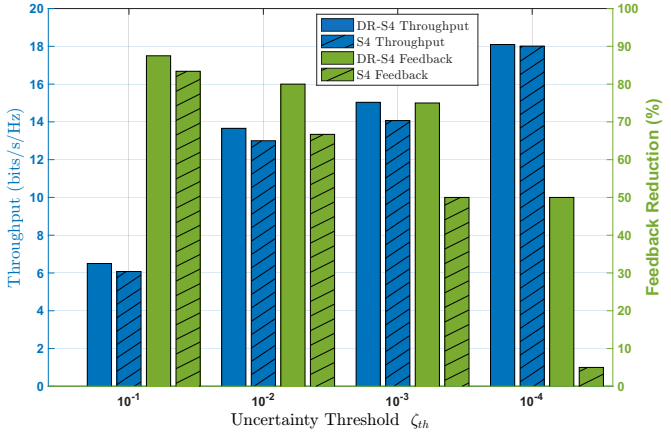

Fig. 7. The trade off between throughput and feedback reduction for changing uncertainty threshold

TABLE IV
STATISTICALLY SIGNIFICANCE TEST: P-VALUES FOR COMPARISONS BETWEEN DR-S4 AND BASELINE MODELS

Model	S4	LSTM	RNN	Transformer
DR-S4	0.00426	0.0027	0.0010	0.00035

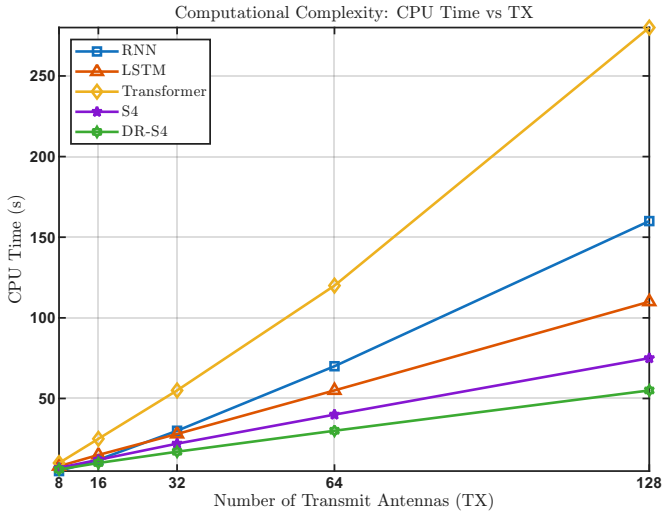

Fig. 8. The time complexity analysis across different numbers of transmit antennae

Figure 6 presents the throughput performance as a function of the prediction horizon, comparing DR-S4, S4, RNN, and LSTM. As expected, throughput decreases with longer horizons due to the accumulation of prediction errors, which gradually degrade the fidelity of the estimated CSI. However, the rate of degradation differs across models, reflecting their ability to capture long-term channel dynamics. DR-S4 exhibits the most graceful decay, maintaining relatively high throughput for up to 6–7 prediction steps before dropping more sharply. This highlights its strength in stabilizing long-sequence predictions through diagonal-plus-low-rank parameterization. S4 also shows strong short-horizon accuracy but degrades faster than DR-S4 as the horizon extends, suggesting that it is less robust when modeling deeper temporal dependencies. By contrast, RNN and LSTM experience much steeper declines, with throughput collapsing rapidly beyond horizon lengths of 3–4, consistent with their limitations in retaining long-term information. These results demonstrate the practical importance of robust long-sequence modeling: in non-terrestrial networks where frequent feedback is costly, architectures such as DR-S4 extend the useful prediction window, enabling more reliable and efficient beamforming decisions over longer time intervals.

To validate the observed performance differences between DR-S4 and the baseline models, we conducted two-sample t-tests for each pairwise comparison. The results, shown in Table IV, indicate that the differences in throughput between DR-S4 and all baseline models are statistically significant. Specifically, the p-values are all below the 0.05 threshold, suggesting that the improvements observed in DR-S4 are not due to random chance, but represent genuine performance gains. This statistical analysis provides strong evidence supporting the superiority of DR-S4 in throughput performance across different transmission configurations.

Figure 7 illustrates the trade-off between throughput and feedback reduction under different uncertainty thresholds ζ_{th} for DR-S4 and S4. As the threshold decreases, the controller becomes more conservative, requesting feedback more frequently. This leads to higher throughput, since CSI prediction errors are corrected more often, but at the cost of reduced feedback savings. Conversely, with larger thresholds, the system tolerates higher prediction uncertainty, which significantly reduces feedback but lowers throughput due to occasional prediction errors. Comparing the two models, DR-S4 consistently achieves slightly higher throughput across all thresholds, reflecting its superior prediction stability. At the same time, DR-S4 also sustains effective feedback reduction, showing a better balance between accuracy and efficiency compared to S4. These results highlight an important system-level insight: uncertainty-aware controllers allow flexible adaptation between spectral efficiency and signaling overhead, and the choice of prediction model directly impacts where this trade-off is achieved.

Figure 8 compares the computational complexity of dif-

ferent architectures in terms of CPU time as the number of transmit antennas increases. As expected, all models exhibit increasing runtime with system dimension, but the rate of growth differs significantly. The Transformer shows the steepest scaling, with CPU time rising sharply and exceeding 250 seconds at TX = 128, reflecting the quadratic complexity of self-attention with respect to sequence length. RNN and LSTM scale more moderately but still demand considerable computation, as recurrent updates become costly in high-dimensional settings. In contrast, S4 and DR-S4 demonstrate substantially lower computational overhead, with DR-S4 consistently achieving the lowest runtime across all antenna configurations. This efficiency arises from the structured state-space formulation, which reduces the per-step complexity of sequence modeling while retaining long-range modeling capability. These findings emphasize a key practical insight: state-space architectures not only deliver superior predictive accuracy but also offer computational scalability, making them more suitable for real-time NTN deployments where both accuracy and latency are critical.

V. Conclusion

In this work, we addressed the problem of predictive CSI modeling in LEO non-terrestrial networks, where high mobility, Doppler shifts, and long propagation delays make conventional feedback-based estimation impractical. We proposed the DR-S4 framework, which integrates orbital Doppler knowledge into the state-space formulation to produce stable and physically consistent long-horizon predictions. Alongside this, we introduced a task-aware training objective that aligns learning with throughput performance, and an uncertainty-gated predict-or-feedback controller that reduces signaling overhead while safeguarding reliability. Our 3GPP-compliant NTN simulations revealed several key insights. DR-S4 achieves faster and more stable convergence than recurrent and attention-based baselines, and it scales effectively with antenna array size while maintaining efficiency. Importantly, it sustains higher throughput over longer prediction horizons and consistently balances accuracy with computational complexity, making it suitable for deployment in resource-constrained terminals and satellite payloads. These results confirm that structured, physics-informed sequence models can offer practical advantages over purely data-driven architectures for NTN prediction tasks. Looking ahead, this work opens multiple research directions. Extending the framework frequency-selective fading and OFDM-based wideband channels will make the predictor directly applicable to next-generation satellite systems. Furthermore, exploring online and adaptive learning strategies can enable the model to continuously update under dynamic constellations, handovers, and real mobility.

References

[1] International Telecommunication Union, "Measuring Digital Development: Facts and Figures 2023," International Telecommunication Union, Tech. Rep., 2023, accessed: 2024-05-27. [Online]. Available: https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

[2] 3GPP, "Study on New Radio Access Technology: Channel Model for Frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Report TR 38.811 V14.0.0, September 2017, accessed: 2024-05-27. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3161

[3] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond," *IEEE Access*, vol. 7, pp. 78 729–78 757, 2019.

[4] J. Yuan, H. Q. Ngo, and M. Matthaiou, "Machine learning-based channel prediction in massive mimo with channel aging," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 2960–2973, 2020.

[5] O. Stenhammar, G. Fodor, and C. Fischione, "A comparison of neural networks for wireless channel prediction," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 235–241, 2024.

[6] A. Alkhateeb, G. Huang, and R. W. Heath Jr, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 7, pp. 68 551–68 565, 2019.

[7] J. Zhang, G. Zheng, Y. Zhang, I. Krikidis, and K.-K. Wong, "Deep learning based predictive beamforming design," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 8122–8127, 2023.

[8] C. Liu, W. Yuan, S. Li, X. Liu, H. Li, D. W. K. Ng, and Y. Li, "Learning-based predictive beamforming for integrated sensing and communication in vehicular networks," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 8, pp. 2317–2334, 2022.

[9] L. Ding and A. Hirose, "Fading channel prediction based on recurrent neural network," *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 862–870, 2014.

[10] W. Jiang and H. D. Schotten, "Recurrent neural networks with long short-term memory for fading channel prediction," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[11] B. Peng, H. Li, and X. Gao, "Deep LSTM for channel prediction in Massive MIMO under imperfect CSI," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 658–662, 2020.

[12] W. Jiang and H. Schotten, "Transformer-based predictive beamforming in high mobility environments," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8124–8136, 2022.

[13] J. Wang, L. Wang, and Q. Zhang, "Transformer-based CSI prediction and beamforming for massive MIMO," *IEEE Transactions on Communications*, vol. 71, no. 4, pp. 2378–2392, 2023.

[14] S. Zhang, S. Zhang, and W. Yuan, "Transformer-empowered predictive beamforming for rate-splitting multiple access in non-terrestrial networks," *IEEE Transactions on Communications*, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10741218

[15] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2022. [Online]. Available: https://arxiv.org/abs/2111.00396

[16] B. Turan and S. Coleri, "Prediction of VLC channel using multilayer perceptron," *IEEE Access*, vol. 9, pp. 16 778–16 788, 2021.

[17] Y. Yuan, H. Ngo, and M. Matthaiou, "Learning-based predictive beamforming for high mobility scenarios," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 688–13 693, 2020.

[18] Y. Liu, Y. He, and S. Wang, "CLSTM for predictive beamforming in vehicular networks," in *2022 IEEE International Conference on Communications (ICC)*. IEEE, 2022, pp. 1–6.

[19] H. Guo, Y. Lin, and X. Chen, "Dual-path transformer for neural beamforming in wireless systems," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1123–1134, 2023.

[20] Y. Yuan, H. Ngo, and M. Matthaiou, "Deep learning-based beam prediction for UAV communications with platform jittering," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6786–6799, 2020.

[21] ——, "Bayesian learning for predictive beamforming in V2I communication systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 3521–3535, 2021.

[22] H. Liu, J. Zhang, and X. Wang, "Predictive beamforming using ISAC signals in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 12 345–12 356, 2022.

[23] Z. Liu, L. Zhang, and C. Shen, "Predictive precoder design for OTFS-enabled URLLC systems," *IEEE Transactions on Wireless Communications*, vol. 22, no. 7, pp. 4532–4544, 2023.

[24] K. Du, T. Sun, and R. Wang, "Dynamic beamforming using CSI prediction in V2I networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5678–5689, 2023.

[25] M. Ying, X. Chen, Q. Qi, and W. Gerstacker, "Deep learning-based joint channel prediction and multibeam precoding for leo satellite internet of

things," *IEEE Transactions on Wireless Communications*, vol. 23, no. 10, pp. 13 946–13 960, 2024.
[26] Q. Li, M. El-Hajjar, C. Xu, J. An, C. Yuen, and L. Hanzo, "Stacked intelligent metasurfaces for holographic MIMO-aided cell-free networks," *IEEE Transactions on Communications*, vol. 72, no. 11, pp. 7139–7151, 2024.
[27] Q. Li, M. El-Hajjar, K. Cao, C. Xu, H. Haas, and L. Hanzo, "Holographic metasurface-based beamforming for multi-altitude LEO satellite networks," *IEEE Transactions on Wireless Communications*, vol. 24, no. 4, pp. 3103–3116, 2025.
[28] Q. Li, M. El-Hajjar, Y. Sun, and L. Hanzo, "Performance analysis of reconfigurable holographic surfaces in the Near-Field scenario of cell-free networks under hardware impairments," *IEEE Transactions on Wireless Communications*, vol. 23, no. 9, pp. 11 972–11 984, 2024.
[29] 3GPP, "Study on New Radio (NR) to support Non-Terrestrial Networks (NTN)," 3rd Generation Partnership Project (3GPP), 3GPP Technical Report TR 38.821, 2020, release 16, version 16.0.0. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.821/38821-g00.zip
[30] H. Lee, I. P. Roberts, J. Heo, J. Son, H. Kim, Y. Lee, and D. Hong, "Can TDD be employed in LEO SatCom systems? challenges and potential approaches," 2025. [Online]. Available: https://arxiv.org/abs/2502.08179
[31] S. Ameli and S. C. Shadden, "A singular woodbury and pseudo-determinant matrix identities and application to gaussian process regression," *Applied Mathematics and Computation*, vol. 452, p. 128032, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0096300323002011

Iqra Hameed Iqra Hameed received the B.Sc. and M.Sc. degrees from the University of Engineering and Technology (UET) Lahore, Pakistan, in 2013 and 2017, respectively, and the Ph.D. degree from the Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, South Korea, in 2024. From 2013 to 2018, she was a Lecturer at FAST–National University of Computer and Emerging Sciences, Pakistan. She is currently a Postdoctoral Researcher at Sejong University, Seoul, South Korea. Her research interests include 5G and B5G networks, wireless power transfer, intelligent reflecting surfaces, sensing-assisted communications, and AI-enabled wireless networks.

Md Habibur Rahman received the B.S. degree in Electrical and Electronic Engineering from Islamic University, Kushtia, Bangladesh, in 2016, and the M.S. degree in Electronic Engineering from Pukyong National University, Busan, Republic of Korea, in August 2020. He completed the Ph.D. degree in August 2025 with the Department of Information and Communication Engineering and the Department of Convergence Engineering for Intelligent Drone, Sejong University, Seoul, Republic of Korea. He is currently working as a postdoctoral researcher in the same department at Sejong University. He actively serves as a reviewer for various international journals and conferences. His research interests include machine learning and deep learning applications in wireless communication systems, with a focus on channel estimation, signal detection, and resource optimization for intelligent reflecting surface (IRS)-assisted networks.

Muhammad Usman received a B.Sc. from the University of Agriculture Faisalabad (UAF), Faisalabad, Pakistan, in 2023. Currently, he is pursuing MS-PhD degree at Sejong University, Seoul, South Korea. His research interests include deep learning/ artificial intelligence-aided channel prediction and channel estimation of wireless networks.

Hyoung-Kyu Song received the B.S., M.S., and Ph.D. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1990, 1992, and 1996, respectively. From 1996 to 2000, he was a Managerial Engineer at the Korea Electronics Technology Institute (KETI), South Korea. Since 2000, he has been a Professor with the Department of Information and Communication Engineering, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul. His research interests include digital and data communications, information theory, and their applications with an emphasis on mobile communications.

Muhammad Rafid received the B.Sc. degree in Electrical Engineering from the University of Science and Technology, Bannu, Pakistan, in 2019, and the M.S. degree in Electrical Engineering from the National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2022. He served as a Researcher at the School of Electrical Engineering and Computer Science (SEECS), NUST. He is currently pursuing the Ph.D. degree with the Signal Theory and Communications Department, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. In parallel, he is working as a Researcher with the Industrial Systems Institute, Athena Research Center (ISI/Athena RC), Patras, Greece, as a Marie Skłodowska-Curie Doctoral Fellow. His research interests include self-managed and resilient cloud-to-edge infrastructures for 6G networks, intelligent and decentralized orchestration, CR-VANETs, V2X communications, AI-enabled wireless systems, and resource management.

Christos Verikoukis (Senior Member, IEEE) received the Ph.D. degree from the Technical University of Catalonia (UPC), Barcelona, Spain, in 2000. He is currently an Associate Professor with the University of Patras, Patras, Greece, while he is an affiliated Faculty Member with ISI/ATHENA, Greece. He has authored or coauthored 180 journal papers and more than 250 conference papers. He is also a coauthor of 6 books, 20 chapters in other books, and 4 patents. He has participated in more than 40 competitive projects, and was a project coordinator of several funded projects from the European Commission and of national projects in Greece and Spain. He is currently the IEEE ComSoc Director for Conferences Operation.