Physics-Grounded Deep Reinforcement Learning for Power Control in LEO Non-Terrestrial O-RAN: A QoS–Power Pareto and Cross-Orbit Robustness Study

Hsiu-Chi Tsai*, Chia-Tung Chung†
National Yang Ming Chiao Tung University, Hsinchu, Taiwan
*Department of Computer Science, ORCID 0000-0001-7421-8027, hctsai1006@cs.nctu.edu.tw
†Department of Photonics, ORCID 0009-0009-2596-8783, jun.514114.ee10@nycu.edu.tw

***Abstract*—We ask *when* learning helps for O-RAN non-terrestrial power control, and answer with a leakage-free fair-baseline protocol on a causal SGP4+3GPP-TR-38.811 LEO-NTN uplink environment. A converged PPO policy is measured, over 35 disjoint held-out SGP4 passes (hierarchical cluster bootstrap, Holm correction), against a *fair* classical family: open-loop fractional power control (OLPC), adaptive-PRB greedy control, and a TS 38.213 closed-loop TPC; and a perfect-model certainty-equivalence *oracle* (Kalman fade prediction + chance-constrained minimum power). Across the full operating envelope (two link margins $\times$ two HARQ depths), *PPO is the best controller nowhere*: a classical controller significantly exceeds it at three of the four corners and ties it at the fourth. The envelope has structure. At moderate margin ($G/T = 6.5$ dB/K) the perfect-model oracle is strongest, dominating PPO by up to $D = +0.063$ SLA satisfaction at *half* its transmit power, with the deployable greedy heuristic close behind (17–35% less power than PPO). Under severe starvation ($G/T = 1.1$ dB/K) the deployable greedy heuristic is best, dominating PPO with HARQ off and matching it at NR-default HARQ, where the starved regime levels every controller. This is the certainty-equivalence prediction for a known-model, correlated-fade problem: here, a converged model-free policy (PPO) does not beat a well-tuned formula. We contribute the causal, O-RAN-integrable environment and fair protocol as a reusable benchmark, an E2SM-NTN E2 service model exposing NTN link-budget state and a power-control action, and an honest map that scopes learning's value to the multi-cell-interference and bursty/URLLC regimes where no closed-form optimum exists.**

***Index Terms*—Non-terrestrial networks, O-RAN, RIC, deep reinforcement learning, power control, LEO satellite, link budget, domain randomization, E2 service model.**

I. Introduction

Low-Earth-orbit (LEO) non-terrestrial networks (NTN) are now a normative part of 5G New Radio: 3GPP Release-17 standardizes direct access from handheld UE to transparent-payload satellites in S-band, with a power class 3 (23 dBm) uplink [2], [4]. Operating such links through the O-RAN architecture (a Near-Real-Time RAN Intelligent Controller (Near-RT RIC) driving an E2 node through standardized E2 service models) would let learned control policies adapt the radio to the violently time-varying geometry of an overhead pass. Uplink transmit-power control is a natural target: the handheld is power-limited, the slant range and Doppler swing over tens of seconds, and meeting an SLA while not simply transmitting at full power is a genuine optimization.

Deep reinforcement learning (DRL) for NTN and O-RAN power and resource control is a crowded area [26]–[29], yet auditing this literature (and our own prior code) exposes recurring weaknesses this paper targets. **Gap 1: non-causal environments.** Many RAN-control "RL environments" sample the channel/KPI state *independently of the agent's action*. If the next SINR does not depend on the chosen power, no power-control policy can be learned and any violation reduction is a reward-shaping artifact. Restoring action-to-state causality through a physical link budget is a precondition we treat as a first-class contribution. **Gap 2: hidden power cost.** Most DRL-NTN papers report QoS gains but not the *transmit power* spent to obtain them. A policy that cuts violations by drifting toward max power has merely traded power for QoS. We make transmit power a first-class, swept cost and report the QoS–power Pareto frontier against pre-registered classical endpoints. **Gap 3: untested distribution shift.** NTN-RL policies are typically trained and tested on the same geometry. Whether a policy transfers across altitude/elevation (routine as different satellites serve a cell) is rarely measured. We benchmark this cross-orbit transfer.

Our contributions are: (C1) a physics-grounded, action-to-state-coupled LEO-NTN power-control RL environment (Section III); (C2) an outer-loop $[\text{power}, \text{PRB}]$ + inner-loop AMC formulation, evaluated under a leakage-free fair-baseline protocol, which shows that on the deployment satisfaction–power frontier a well-tuned classical controller *matches or exceeds* the learned policy at both tested link margins and HARQ depths, the learned policy being efficient only at its own matched high-power operating point (Sections IV–V); and (C3) an *operating envelope* across orbital geometry and link margin: cross-orbit OOS robustness together with a fair characterization across link margins in which the classical advantage holds at both the moderate and the severe-starvation regime (Sections IV–V). As a systems/standards contribution we specify E2SM-NTN, an E2 service model carrying the link-budget state and transmit-power action the stock E2 surface

lacks (Section VI). We frame the findings honestly: a fairly-benchmarked map of *when* learned NTN power control helps, and, in every single-user regime tested here, it does not beat a well-tuned classical, while scoping the multi-cell and bursty-traffic regimes where learning is more likely to help, not a universal-superiority claim.

II. Background and Related Work

A. NTN link budget and geometry

The 3GPP NTN study items define the system model we ground on. TR-38.811 gives the NTN channel and shadow-fading model [1]; TR-38.821 specifies the spherical-earth pass geometry, handheld EIRP, and satellite G/T figures for link-budget calibration [2]; TS-38.214 provides the CQI/MCS tables that map SINR to spectral efficiency [3]; and the Release-17 normative TS-38.101-5 fixes the handheld S-band (band n256) power class 3 uplink at 23 dBm [4]. Open-loop fractional power control follows TS-38.213 §7.1.1, $P = \min(P_{\max}, P_0 + 10\log_{10} M_{\mathrm{RB}} + \alpha \cdot \mathrm{PL})$ [5]. Our scenario (transparent payload, S-band 2 GHz, handheld 23 dBm uplink) is deliberately Release-17-aligned. Regenerative payloads, store-and-forward and Ka-band are out of scope.

B. O-RAN RIC and E2 service models

In O-RAN, a Near-RT RIC controls an E2 node through E2 service models: E2SM-KPM for measurement [9], E2SM-RC for RAN control [8], among others. Crucially, although the current normative E2SM-RC [8] includes a generic uplink transmit-power control feature, it carries none of the NTN link-budget state (per-UE SINR/BLER, satellite geometry (elevation, slant range, Doppler), fade and ephemeris context) that a satellite power-control policy must observe, and no standardized E2SM-NTN exists. NTN is treated only by non-normative material [10], [11], and recent O-RAN-NTN architecture proposals [37] define the space–ground split but neither an E2SM-NTN nor an NTN-aware power-control mechanism. This shapes both our deployment story (the PRB half of the action rides stock E2SM-RC; the power half needs an NTN-aware extension) and our systems contribution (Section VI).

C. DRL for NTN/O-RAN power and resource control

Among 2025–2026 work, Shahabi *et al.* [26] optimize energy efficiency for an NTN O-RAN split but omit transmit-power cost and action-to-state causality. RSLAQ [27] is an SLA-aware xApp without a swept power–QoS frontier or cross-orbit test. Qazzaz *et al.* [38] manage resources and handovers for NTN users with an xApp, but without a transmit-power action, an NTN service model, or a released environment. Nearest in spirit, Tran *et al.* [28] apply DRL to *resilience* in integrated satellite–terrestrial networks and Xie *et al.* [29] to multi-satellite beam hopping and power allocation, but neither targets uplink transmit-power control on a causal link budget. Against this literature we add an explicit power–QoS Pareto, a cross-orbit OOS study, an SGP4-grounded causal environment, and the E2SM-NTN interface. Demirel *et al.* [30] apply domain randomization terrestrially, which our pass library supports across orbital regimes. Our outer/inner-loop split follows the classical separation of power control from link adaptation [12]–[14], since joint power+MCS learning is non-convex [15]. We claim no first DRL-NTN controller. The contribution is the fairly-benchmarked *operating-envelope map* (where learning helps for NTN power control and where a formula suffices) on a causal environment with honest power accounting, plus a deployable E2SM-NTN realization.

III. Physics-Coupled NTN-RL Environment (C1)

The environment makes the agent's action causally drive the reward through a closed-form, NumPy-only physics chain (CPU-light, vectorizable across cores):

$$\underbrace{\text{geometry}}_{\text{SGP4 pass}} \rightarrow \underbrace{\text{link budget}}_{\text{TR-38.811/821}} \rightarrow \text{SINR} \rightarrow \text{CQI/BLER} \rightarrow \text{thr} \rightarrow \text{SLA}.$$

A. Geometry

Each episode traverses one real rise–peak–set satellite pass. We pre-compute a pass library by propagating cached CelesTrak two-line elements (Starlink ∼550 km, OneWeb ∼1200 km) with SGP4 [7] from a single ground cell, transforming ECI→ECEF→topocentric to obtain per-second elevation, slant range and Doppler, resampling each pass to a fixed length, and binning by orbit shell and peak elevation to support domain randomization and OOS (C3). The episode thus exposes genuine ephemeris-derived geometry and per-step Doppler, rather than the static or quasi-static geometry common in NTN-RL.

B. Link budget and KPI mapping

The controlled link is the *uplink*: a handheld UE sets its transmit power ($\leq$ 23 dBm at 0 dBi, so EIRP = Tx power) and PRB allocation; the satellite is the receiver, characterized by G/T[1] [2]. Free-space path loss follows Friis. We add gaseous absorption, ionospheric scintillation, an AR(1) temporally correlated log-normal shadow-fading term, and a residual-Doppler inter-carrier-interference penalty after feeder-link CFO pre-compensation. The fixed UE power is spread over the allocated PRBs, so per-PRB SINR is

$$\mathrm{SINR}_{\mathrm{dB}} = \mathrm{EIRP}_{\mathrm{PRB}} + \tfrac{G}{T} - \mathrm{PL} - 10\log_{10}(k\,B_{\mathrm{PRB}}), \quad (1)$$

with $\mathrm{EIRP}_{\mathrm{PRB}} = \mathrm{EIRP} - 10\log_{10} N_{\mathrm{PRB}}$, capped by an aggregate co-channel C/I. Concentrating power into fewer PRBs raises SINR; spreading it lowers SINR—the PRB lever. SINR maps to a channel CQI and, given the chosen MCS, to a BLER through a logistic link curve centered so that $\mathrm{BLER} = 0.10$ at the CQI's nominal 10%-BLER SINR threshold. Throughput is the Shannon-capped spectral efficiency times bandwidth times $(1 - \mathrm{BLER})$. End-to-end latency is exact propagation $(2d/c)$ plus a load-dependent queueing term. An SLA *violation* is declared when $\mathrm{SINR} < -3$ dB *or* throughput < 0.5 Mbps *or* $\mathrm{BLER} > 0.1$.

[1] G/T is the receiver figure-of-merit (antenna gain over system noise temperature) and enters the per-PRB SINR once. RSRP, reported separately, uses the receive antenna gain and does not feed SINR, so antenna gain is not double-counted.

TABLE I
ENVIRONMENT AND LINK-BUDGET PARAMETERS.

Parameter	Value
Carrier / SCS / PRB bandwidth	2 GHz (S-band) / 15 kHz / 180 kHz
UE max EIRP (PC3, TS-38.101-5)	23 dBm @ 0 dBi
Satellite G/T (default / sensitivity)	6.5 / 1.1 dB/K
Satellite Rx antenna gain	30 dBi
Atmospheric / scintillation loss	0.1 / 2.2 dB
Shadow fading σ / AR(1) ρ	4 dB / 0.95
PRB allocation range	$[1, 8]$
Tx-power range	$[-10, 23]$ dBm
Orbit altitudes	550 / 1200 km
Peak-elevation regimes	$20/40/60/85^\circ$
SLA thresholds	SINR ≥ -3 dB, thr ≥ 0.5 Mbps, BLER ≤ 0.1
Episode length	one SGP4 pass ($\sim$300 steps, 1 s)

C. Calibration and verification

The link budget is calibrated against TR-38.821: at zenith, max power and a single PRB it reproduces the uplink template SINR of $\approx +14.5$ dB under the conservative Set-1 $G/T = 1.1$ dB/K, falling to ≈ -35 dB at the 10° edge. Our main experiments adopt an *assumed* advanced-payload $G/T = 6.5$ dB/K (illustrative of a more capable receiver, *not* a 3GPP-specified value) which raises the SINR ceiling $\sim$5.4 dB above the conservative 1.1 dB/K. We sweep both and treat the resulting margin *boundary* (Section V), not any single G/T, as the contribution. Violation magnitudes are sensitive to this choice and reported explicitly. A 12-test machine-checked suite asserts the causal couplings: SINR monotone in power, falling with PRB-spreading, geometry coupling, MCS-driven $\mathrm{BLER} \to 1$, the 3GPP calibration point, and that more power cuts violations but strictly costs power—machine evidence for C1.

D. Observation and action interface

The observation is a 14-D vector: 11 one-step-delayed KPIs (throughput, PRB use, active UEs, CQI, RSRP/RSRQ, SINR, latency, BLER) plus three features of the *known upcoming* geometry (elevation, slant range, Doppler) that a real UE obtains from the SIB19 ephemeris and its GNSS fix, putting the agent on equal footing with the geometry-aware baselines. The action is two-dimensional, $[N_{\text{PRB}}, P_{\text{tx}}] \in [-1, 1]^2$, rescaled to $N_{\text{PRB}} \in [1, 8]$ ($\sim$1–4 PRBs per TR-38.821) and $P_{\text{tx}} \in [-10, 23]$ dBm. The MCS is set by the inner-loop AMC (Section IV), not the agent. Table I lists the parameters.

IV. METHOD: OUTER-LOOP RL, INNER-LOOP AMC, AND THE PARETO SWEEP (C2/C3)

A. Two-timescale formulation

We deliberately do *not* ask one agent to learn both power and modulation. Classical link adaptation (matching MCS to the channel SINR for a target BLER) is a solved sub-problem, and forcing PPO to relearn it alongside power destabilized training in our pilots (per-seed bimodal learn-or-collapse). Instead we adopt the standard two-timescale architecture: an *outer loop* (the RL agent) chooses $[\text{power}, \text{PRB}]$, and an *inner loop* (deterministic AMC) selects the MCS matched to the expected per-PRB SINR for a $\sim$10%-BLER target. This mirrors real power-control-plus-AMC separation and is supported by the action-as-power DRL paradigm [12], outer-loop link adaptation [13], and PHY abstraction [14]. The residual joint power+PRB problem remains non-convex, justifying RL [15]. We use PPO [22] via Stable-Baselines3 [23].

B. Reward and the power-penalty Pareto knob

The per-step reward is

$$r = w_{\text{tp}} \frac{\text{thr}}{\text{thr}_{\text{ref}}} + w_{\text{lat}} \frac{\text{lat}}{100} + w_{\text{bler}}\, \mathrm{BLER} + w_{\text{pow}}\, \frac{P_{\text{lin}}}{P_{\max,\text{lin}}} + w_{\text{viol}}\, g_{\text{viol}}, \tag{2}$$

with $(w_{\text{tp}}, w_{\text{lat}}, w_{\text{bler}}, w_{\text{viol}}) = (0.2, -0.05, -0.5, -2.0)$, $\text{thr}_{\text{ref}} = 3$ Mbps, and $g_{\text{viol}} \in [0, 1]$ a margin-based violation penalty (graded by the worst of the throughput/SINR/BLER deficits) so the agent is rewarded for getting close to the SLA even where the link cannot fully close, and cannot win the energy objective by surrendering in hopeless geometry. The violation term dominates: the agent's first job is to *meet* the SLA, with throughput a small bounded bonus. The transmit-power weight w_{pow} is the Pareto knob: sweeping it over $\{0, -0.25, -0.5, \dots\}$ traces the violation-vs-power frontier, with $w_{\text{pow}} = 0$ the vanilla-PPO point. As a constrained-RL cross-check (future work in this draft), the same frontier can be traced by a PID-Lagrangian power budget [16]–[18] that solves for the shadow price $\lambda(\text{budget})$ rather than a fixed weight. We note this recovers operating points a fixed-weight scalarization can skip.

C. Training stabilization

The single highest-impact stabilizer is VecNormalize (observation and return normalization, $\text{clip}_{\text{obs}} = 10$), which removes the per-seed gradient-scale variance from the heterogeneous-unit state that drove the bimodal collapse [20], [21]. We also initialize a small policy ($\log \sigma_0 = -0.7$, action std ≈ 0.5) off the tanh-saturation edges, use a fixed entropy floor, and cap the PRB action to its realistic band so the optimum sits near the action-box center. The policy is a $[128, 128]$ MLP trained for 10^6 steps on CPU (the environment, not the small network, is the bottleneck). We verify convergence to 2×10^6 steps (Fig. 2) and train up to 16 seeds per configuration. Evaluation is on disjoint held-out SGP4 passes (Sec. V).

D. Baselines (fair family)

All policies act on identical physics and identical per-seed fading sequences, so the comparison isolates the power/PRB decision, and each classical's operating point is selected on the training passes and evaluated on the disjoint test passes, keeping the comparison leakage-free. (i) **OLPC(prb4)**: 3GPP fractional open-loop PC [5] at a fixed four-PRB allocation. With the OCUDU/srsRAN defaults ($P_0 = -76$ dBm, $\alpha = 1.0$) full compensation saturates at $P_{\max}$ here. The smooth

classical frontier is recovered not by backing off α (which *cliffs*: the link collapses below $\alpha \approx 0.55$) but by sweeping the *target-power* parameter P_0 at $\alpha = 1$. (ii) **Adaptive-PRB greedy**: combines adaptive-P_0 power with per-step greedy PRB selection. (iii) **Closed-loop TPC**: a 3GPP TS-38.213 closed-loop transmit-power-control loop that tracks the measured SINR (the only *fading-aware* classical) which at a fixed PRB saturates near 148 mW. (iv) **Perfect-model oracle**: a certainty-equivalence MPC (Kalman fade prediction + chance-constrained minimum power) *granted* the exact link model and fade statistics—a strong informed *reference*, not a deployable controller, included to bound what perfect knowledge buys (Sec. V). The deployable classicals (i)–(iii), between them, hold every lever the learned agent has.

E. Cross-orbit robustness (C3)

We define orbital regimes by (peak elevation, altitude) and test generalization by evaluating the leo550/60°-trained policy (and the classical family) *unchanged* on held-out regimes (cross-altitude LEO-1200, cross-elevation 20°), deriving Doppler and path loss from the sampled geometry rather than sampling them independently [24], [25]. Because the policy observes the pass geometry (elevation, slant range, Doppler), we ask whether it generalizes *by construction*, and whether any such robustness is unique to learning or shared by the adaptive classical controllers.

V. Evaluation

A. Methodology

Motivated by the reproducibility critique of [19], [20] and recent fair-baseline benchmarks in adjacent control domains in which calibrated controllers rival deep RL [36], every reported difference is a *paired hierarchical (cluster-by-pass) bootstrap* (10^4 resamples) with Holm correction within each baseline family, rather than a bare point estimate. A gap is significant when its 95% CI excludes zero after correction. The primary metric is the SLA-violation rate (equivalently, satisfaction $= 1 -$ violation). We also report mean transmit power, throughput, and energy per delivered bit. Crucially, operating-point selection and reporting use *disjoint* orbital passes: each controller's knob (and the learned policy's power-weight) is selected on 30 training passes, and *every* reported number is estimated on 35 held-out test passes (each under four independent fading realizations)—a leakage-free generalization test rather than an in-distribution one. All numbers are recomputed from the per-seed result files by the released analysis scripts. Nothing is hand-entered.

B. A well-tuned classical controller matches or beats the learned policy

Figure 1 shows each controller's satisfaction–power frontier over 35 disjoint held-out passes. Because the deployable operating point must be chosen without the test data, we select each classical's knob on the *training* passes and estimate the difference on the disjoint test passes (paired hierarchical bootstrap, Holm correction). In the moderate-margin regime ($G/T = 6.5$ dB/K) a well-tuned classical Pareto-dominates the learned policy at *both* HARQ depths. With HARQ disabled ($h = 1$), the train-selected greedy point is 0.833 at 123 mW versus PPO's 0.799 at 188 mW ($D = +0.034$, $p_{\mathrm{Holm}} < 0.001$), higher satisfaction at $\approx 35\%$ lower power, and greedy also exceeds PPO at adjacent lower-power knobs (0.801 at 115 mW, 0.807 at 158 mW), so the dominance is not a lone peak. Open-loop OLPC likewise dominates (0.817 at 115 mW, $D = +0.018$, $p_{\mathrm{Holm}} = 0.002$). With NR-default HARQ ($h = 4$) the train-selected greedy point is 0.927 at 158 mW versus PPO's 0.922 at 190 mW ($D = +0.005$, $p_{\mathrm{Holm}} < 0.001$), equal satisfaction at $\approx 17\%$ lower power (its test-frontier peak reaches 0.929 at 123 mW). Here OLPC does not dominate (0.906 at 170 mW; neither Pareto-dominates the other). The advantage is confined to this lower-power region: at PPO's own ≈ 190 mW operating point PPO is the more efficient of the two (greedy there falls to 0.760 at $h = 1$ and 0.916 at $h = 4$). Accordingly, at PPO's own high power PPO significantly exceeds the fixed-PRB TPC (by ≈ 0.11; the TPC saturates near 148 mW and cannot reach PPO's point) and, with the classicals dragged past their own optima, greedy at $h = 1$ ($D = -0.039$) and OLPC at $h = 4$ ($D = -0.022$). It significantly but negligibly exceeds greedy at $h = 4$ ($D = -0.011$, TOST-equivalent within ± 0.02) and is indistinguishable from OLPC at $h = 1$ ($p_{\mathrm{Holm}} = 0.103$), efficient use of a fixed high budget, not frontier superiority. The decisive lever is the joint power–PRB trade, which the adaptive-PRB greedy controller exploits and the fixed-PRB TPC cannot. Open-loop OLPC dominates at $h = 1$ through its target-power setting alone. The frontier comparison is the deployment-relevant lens but rests on PPO's single train-selected anchor and is reported as exploratory. The paired matched-power test is the strict inferential comparison. Holm is applied within each baseline family.

a) Severe starvation ($G/T = 1.1$ dB/K).: The ordering degrades gracefully at the harsher margin. With HARQ off ($h = 1$) greedy still Pareto-dominates PPO: 0.517 at 170 mW versus 0.478 at 199 mW ($D = +0.039$, 95% CI $[+0.037, +0.042]$, $\approx 15\%$ lower power), and the perfect-model oracle likewise (0.499 at 165 mW, $D = +0.021$). At NR-default HARQ ($h = 4$) the starved regime compresses every controller toward maximum power and the gap closes to a statistical tie: greedy (0.539 at 190 mW) and the oracle (0.526) sit numerically above PPO (0.523 at 196 mW) but neither F4-dominates it ($p_{\mathrm{Holm}} > 0.2$). At matched (≈ 199 mW) power PPO exceeds OLPC ($D = -0.128/-0.140$) and the fixed-PRB TPC (which falls into outage far below PPO's power) at both depths, and exceeds greedy at $h = 1$ ($D = -0.033$) while tying it at $h = 4$. Under severe starvation, then, no controller (learned, heuristic, or perfect-model) separates at $h = 4$, and the classical controllers lead at $h = 1$. The learned policy is never ahead on the satisfaction–power frontier. Across all four corners, in sum, PPO is the best controller nowhere: a classical controller significantly exceeds it at three of the four (margin × HARQ) corners and ties it at the fourth.

b) The perfect-model benchmark.: As a strong model-based *informed reference* we add a standard certainty-equivalence oracle: a per-step controller that Kalman-predicts the (AR(1), $\rho = 0.95$) shadow fade from the delayed CSI (the *same* observation the learned policy receives) and solves an outage-constrained minimum-power×PRB allocation at a swept target outage ε (the reliability-constrained transmit-power control of [31], [33] with Gudmundson-correlated shadowing [32]). We *grant* it the exact link model and fade statistics, so it is not a contribution nor a deployable controller but a strong informed benchmark bounding what perfect knowledge buys. So endowed it is strongest at moderate margin: at $G/T = 6.5$ dB/K, $h = 1$, its 0.863 at 95 mW F4-dominates PPO's 0.799 at 188 mW by $D = +0.063$ (95% CI $[+0.055, +0.074]$), half the power, and at $h = 4$ it ties the greedy heuristic at the top (0.928 vs 0.929). Under severe starvation its chance-constraint turns conservative near the outage cliff: it still dominates PPO at $h = 1$ ($D = +0.021$) but only ties it at $h = 4$, where the deployable model-free greedy heuristic is best. Its frontier is smooth and its optimum sits at the lowest power of any controller (95–165 mW versus PPO's 188–199 mW). That the informed optimum is a Kalman/chance-constraint *formula* which model-free PPO (seeing the same CSI) does not match is exactly what certainty-equivalence theory predicts for a known-model, correlated-fade problem [34], [35]: the value of learning must lie where the model breaks, not here.

C. Cross-orbit robustness

We test generalization by evaluating the leo550/60°-trained policy, and the classical family, *unchanged* on held-out orbits: cross-elevation (20° peak) and cross-altitude (LEO-1200), at $G/T = 6.5$ dB/K, $h = 4$ (in-distribution reference: satisfaction $0.922/0.929$ for PPO/greedy). Cross-elevation is mild: satisfaction holds near 0.90 (PPO 0.900, greedy 0.906, OLPC 0.868). Cross-altitude is severe (the ≈ 6.7 dB extra path loss roughly halves satisfaction) and it *separates adaptive-PRB from fixed-PRB control*: the learned PPO (0.503) and the adaptive-PRB greedy heuristic (0.511) both generalize and are statistically tied (paired $D = +0.006$, $p_{\mathrm{Holm}} = 0.49$) by adapting their PRB allocation to the harder link, whereas the fixed-four-PRB OLPC (0.335) and closed-loop TPC (0.334) collapse toward outage. The learned policy is thus robust to the orbital shift—but so is the adaptive-PRB heuristic that matches it: robustness here is a property of *adaptive resource control*, not of learning. The naive geometry-conditioned policy generalizes by construction (its observation carries elevation, slant range and Doppler), without any domain randomization. This cautions against crediting *learning* for what is in fact adaptive PRB control: in our own pipeline the once-apparent catastrophic LEO-1200 floor vanished once residual HARQ BLER was modeled, and the surviving robustness is shared by the adaptive-PRB heuristic. These two orbital shifts (one elevation, one altitude) are an initial out-of-sample probe, not a full cross-orbit characterization.

TABLE II
CAPABILITIES A SATELLITE UPLINK POWER-CONTROL XAPP REQUIRES VERSUS WHAT STOCK E2 SERVICE MODELS EXPOSE (Y: PROVIDED; –: ABSENT). ONLY E2SM-NTN CARRIES THE NTN LINK-BUDGET STATE; E2SM-RC'S GENERIC POWER FEATURE DOES NOT.

Capability for NTN uplink power control	KPM	RC	CCC	**NTN**
Per-UE SINR and BLER report	–	–	–	Y
Satellite geometry (elev., slant, Doppler)	–	–	–	Y
NTN impairments (Doppler, delay, path loss)	–	–	–	Y
Link-budget state (Tx/Rx power, margin)	–	–	–	Y
Uplink transmit-power control action	–	gen.†	–	Y
PRB / slice-quota control action	–	Y	ratio	Y

†E2SM-RC (from v7.00) adds a *generic* uplink-power-control feature that carries none of the NTN link-budget state above.

VI. E2SM-NTN: An E2 Service Model for NTN State and Power Control

A learned power-control policy is deployable only if the RAN interface can both *report* the state it conditions on and *carry* the action it emits. We audited a stock, current-generation open RAN stack (OCUDU 26.04, the BSD-3 srsRAN successor) at the code level. The DU-side E2 agent wires exactly three service models: KPM, RC, CCC. Its entire control surface is E2SM-RC Style 2 / Action 6 "slice-level PRB quota" (PRB ratios only) plus a CCC RRM-policy ratio. Its entire measurement surface (E2SM-KPM) is CQI, RSRP/RSRQ, the PRB counters, UE throughput and delay families, *but neither SINR nor BLER, and no transmit power or MCS*. There is no transmit-power, MCS, or power-headroom control action *implemented* anywhere in the stack. The current normative E2SM-RC [8] does define a generic uplink-power-control feature, but the audited open stack does not expose it, and that feature carries none of the NTN link-budget state (per-UE SINR/BLER, satellite geometry) a satellite policy must observe. A stock E2 node can thus neither tell a power-control xApp the SINR/BLER it needs nor accept an NTN-aware power command: a purpose-built service model is a precondition, not a convenience. Table II summarizes this gap: none of the stock service models (including E2SM-RC's generic uplink-power feature) exposes the NTN link-budget state a satellite policy must observe.

Design. E2SM-NTN is an ASN.1 extension inside the O-RAN E2AP/E2SM framework, announced at E2 Setup as RAN function ID 10. Its Format-1 indication composes six measurement groups that render the satellite link legible to the RIC: SGP4-derivable geometry (elevation, azimuth, slant range, velocity), channel quality *including SINR and BLER* (the quantities stock KPM cannot report), NTN impairments (Doppler shift/rate, delay, path loss, attenuation), link budget (Tx/Rx power, margin, achieved/required SNR), and handover/performance metrics. The control side adds the missing action: an NTN-PowerControlAction (target UE, `target-tx-power-dbm` typed INTEGER$(-2000..500)$,

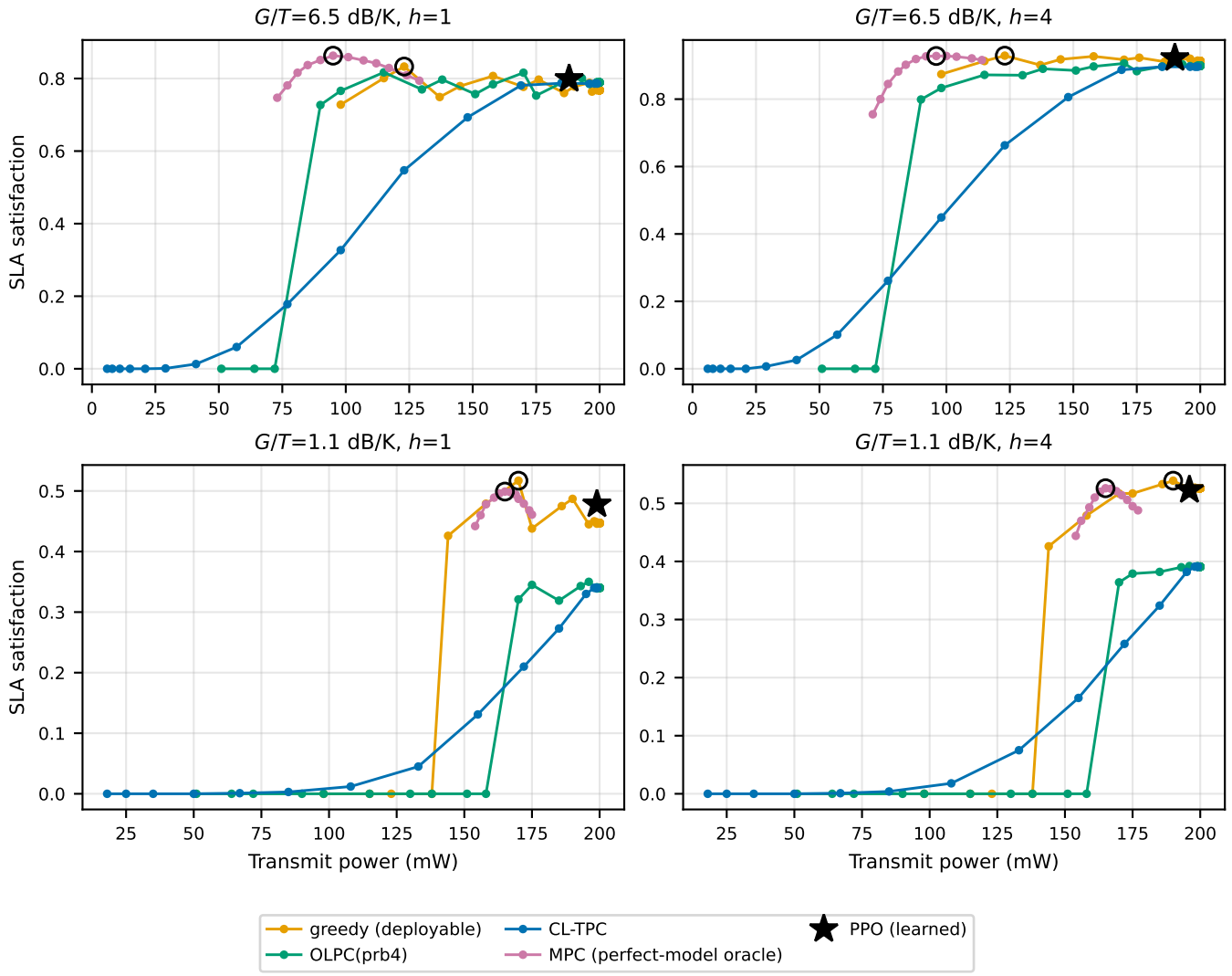

Fig. 1. Each controller's held-out satisfaction–power frontier at the four G/T×HARQ corners. A classical controller matches or exceeds PPO's operating point at every corner—Pareto-dominating at three and tying at the fourth (severe starvation, NR-default HARQ); the perfect-model oracle (MPC) is strongest at moderate margin, at roughly half PPO's power.

i.e. dBm×10, a signed adjustment, and a coded reason) and a sibling link-adaptation action for joint power+PRB/MCS.

Measured result. Messages serialize with ASN.1 Unaligned Packed Encoding Rules (UPER), the family O-RAN mandates for E2AP. Over a 1000-message Monte-Carlo of randomized Format-1 indications, the UPER encoding is a constant 92 B versus a mean 1358.99 B for a readable JSON rendering of the same fields, a 93.2% **smaller message** (an NTN-PowerControlAction control message is 12 B). JSON is a didactic reference for schema compactness, not a deployed E2AP alternative (O-RAN mandates UPER). This is an *illustrative bytes-on-wire* figure (*not* a packet error rate) from dropping repeated ASCII keys and bounding each field to a fixed-width integer; independently reproduced (asn1tools 0.167.0). The PRB half of the action rides stock E2SM-RC, so a stock-RIC baseline is unaffected. Only the power half needs the extension. We claim, bounded: to our knowledge E2SM-NTN is the *first publicly-released NTN-specific E2 service model to expose satellite link-budget state (SGP4 geometry with TR-38.811 SINR/CQI/BLER) and bind it to a transmit-power control action* (E2SM-RC's generic power control carries none of this state), not "the first O-RAN NTN" work. This contribution is a design, schema and offline benchmark. The service model is specified and plumbed but not yet integrated as a live xApp loop (Section VII).

VII. Discussion and Limitations

We are explicit about scope. The inner-loop AMC assumes zero-delay, fading-free CSI known from ephemeris. A real NTN uplink acts roughly one round-trip after that CSI, and a one-slot delay plus an OLLA BLER offset is left as a refinement. HARQ is abstracted as an instantaneous BLER (soft-buffer lifetime and retransmission limits under the long

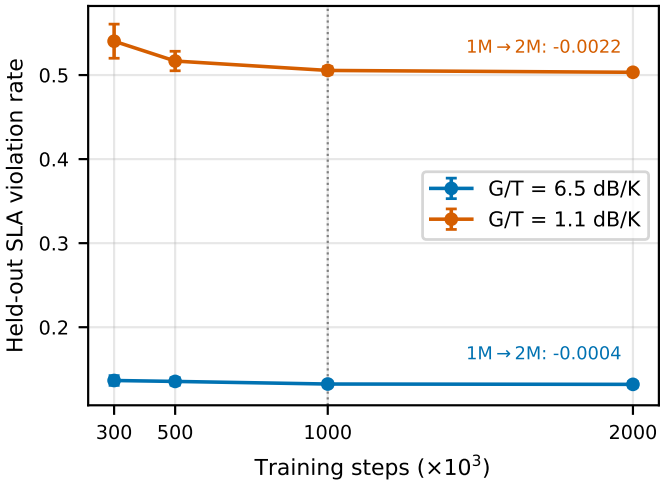

Fig. 2. Held-out SLA-violation rate versus training budget at both link margins; the curves are flat by 1M steps.

RTT are unmodeled), and the outer/inner power–MCS separation, though principled, can trail a fully joint controller. The learned arm is a single on-policy algorithm (PPO). Whether off-policy (SAC/TD3) or constrained-RL learners narrow the gap is untested, so the reality-check is scoped to PPO under this formulation. The environment is single-UE with saturated traffic and a fixed co-channel C/I cap (the regime a formula handles well) so it does not yet exercise the multi-cell-interference or bursty/URLLC coupling where learning is most likely to help, and the cross-orbit study spans two orbital shifts rather than a full sweep. Absolute violation magnitudes depend on the satellite G/T and the SLA thresholds, so we emphasize relative comparisons on a fixed, explicitly reported configuration. Finally, E2SM-NTN is designed, ASN.1-specified and offline-benchmarked but not yet a live xApp loop: the deployment path (containerized policy on the OSC Near-RT RIC via `oran-sc-ric`, no live RF) is plumbed, and the self-assigned RAN function ID/OID would need O-RAN allocation for interoperability.

VIII. Conclusion

Using a causal SGP4+TR-38.811 NTN uplink environment, an E2SM-NTN service model, and a leakage-free fair-baseline protocol, we mapped the operating envelope of learned uplink power control and found the learned policy to be the best controller *nowhere*: a classical controller significantly exceeds converged PPO at three of the four (link-margin $\times$ HARQ) corners and ties it at the fourth. This is not an artifact of under-training or weak tuning: at both link margins the held-out violation rate is already flat between one and two million training steps. The 1M $\rightarrow$ 2M change is essentially zero (<0.001 at $G/T = 6.5$ dB/K and -0.002 at $G/T = 1.1$ dB/K, eight seeds each; Fig. 2), and a best-on-train hyperparameter search does not lift it past the classical. It follows instead from the problem structure: a known link model, a short-horizon objective, and a slowly-varying (AR(1)) shadow fade a linear predictor tracks well, a regime in which certainty-equivalence model-based control is a strong reference and the tested model-free RL (PPO) can, at best, match it [34], [35]. The envelope makes this concrete: at moderate margin a perfect-model oracle is strongest, dominating PPO at half its transmit power, with a deployable greedy heuristic close behind. Under severe starvation the greedy heuristic is best, the oracle's chance-constraint turning conservative near the outage cliff. We also report the fair-baseline pitfalls we had to correct: an informed baseline initially excluded by a knob-grid mismatch, a per-PRB control lever silently cancelled by the power-law coupling, frontier peaks lost to coarse sampling, and a matched-power lens that flattered the learned policy, as a cautionary illustration of how readily deep-RL-for-wireless comparisons overclaim, and as motivation for the causal environment and fair protocol as a reusable yardstick. The value of learning in NTN radio control is more likely to lie elsewhere: multi-cell interference coordination, non-convex joint scheduling, and bursty or URLLC traffic with temporal coupling, where no closed-form optimum exists, which we leave, on the same causal and fairly-benchmarked footing, as the natural next step.

Code and Data Availability

The causal environment, the fair baseline family, the leakage-free evaluation protocol, and the E2SM-NTN service model (ASN.1 module, PER codec, and size benchmark) are released as an open-source benchmark at https://github.com/thc1006/ntn-power-control-benchmark (Apache-2.0), including the SGP4 pass library and the regression tests that assert action-to-state causality.

References

[1] 3GPP, "Study on New Radio (NR) to support non-terrestrial networks," TR 38.811, V15.4.0, 2020.
[2] 3GPP, "Solutions for NR to support non-terrestrial networks (NTN)," TR 38.821, V16.2.0, 2023.
[3] 3GPP, "NR; Physical layer procedures for data," TS 38.214 (Rel-18).
[4] 3GPP, "NR; User Equipment (UE) radio transmission and reception; Part 5: Satellite access Radio Frequency (RF) and performance," TS 38.101-5 (Rel-17/19).
[5] 3GPP, "NR; Physical layer procedures for control," TS 38.213, §7.1.1 (fractional open-loop power control).
[6] ITU-R, "Propagation data and prediction methods required for the design of Earth-space telecommunication systems," Rec. P.618.
[7] D. A. Vallado, *Fundamentals of Astrodynamics and Applications*, 4th ed. Microcosm Press, 2013.
[8] O-RAN Alliance WG3, "E2 Service Model (E2SM), RAN Function Network Interface Controller (RC)," O-RAN.WG3.E2SM-RC, v9.00.
[9] O-RAN Alliance WG3, "E2 Service Model (E2SM), Key Performance Measurement (KPM)," O-RAN.WG3.E2SM-KPM, v7.00.
[10] O-RAN Alliance, "O-RAN Non-Terrestrial Networks (NTN) Deployments," White Paper O-RAN-2025.04.02, v08.4, 2025.
[11] O-RAN Alliance WG3, "RIC Enabling NTN Deployments," Technical Study, v1.00.
[12] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, 2019.
[13] S. K. Pulliyakode and S. Kalyani, "Reinforcement learning techniques for outer loop link adaptation in 4G/5G systems," arXiv:1708.00994, 2017.
[14] S. Lagen *et al.*, "New radio physical layer abstraction for system-level simulations of 5G networks," arXiv:2001.10309, 2020.

[15] Z. Wang *et al.*, "Deep reinforcement learning-aided transmission design for energy-efficient link optimization in vehicular communications," arXiv:2404.12595, 2024.
[16] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by PID Lagrangian methods," in *ICML*, 2020.
[17] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *ICLR*, 2019.
[18] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *ICML*, 2017.
[19] R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in *NeurIPS*, 2021.
[20] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *AAAI*, 2018.
[21] M. Andrychowicz *et al.*, "What matters in on-policy reinforcement learning? A large-scale empirical study," 2020.
[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.
[23] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-Baselines3: Reliable reinforcement learning implementations," *J. Mach. Learn. Res.*, vol. 22, no. 268, pp. 1–8, 2021.
[24] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IROS*, 2017.
[25] R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel, "A survey of zero-shot generalisation in deep reinforcement learning," *J. Artif. Intell. Res.*, vol. 76, 2023.
[26] S. M. M. Shahabi *et al.*, "Energy-efficient deep reinforcement learning-based network function disaggregation in hybrid non-terrestrial open radio access networks," arXiv:2506.06876, 2025.
[27] N. M. Yungaicela-Naula, V. Sharma, and S. Scott-Hayward, "RSLAQ: A robust SLA-driven 6G O-RAN QoS xApp using deep reinforcement learning," arXiv:2504.09187, 2025.
[28] D.-H. Tran *et al.*, "Resilience optimization in 6G and beyond integrated satellite–terrestrial networks: A deep reinforcement learning approach," arXiv:2602.01102, 2026.
[29] X. Xie *et al.*, "Multi-satellite beam hopping and power allocation using deep reinforcement learning," arXiv:2501.02309, 2025.
[30] B. Demirel *et al.*, "Generalization in reinforcement learning for radio access networks," arXiv:2507.06602, 2025.
[31] S. Kandukuri and S. Boyd, "Optimal power control in interference-limited fading wireless channels with outage-probability specifications," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 46–55, 2002.
[32] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electron. Lett.*, vol. 27, no. 23, pp. 2145–2146, 1991.
[33] W. Sun, H. Yu, Y. Yang, Q. Li, D. Mu, and X. Xu, "Confidence interval based model predictive control of transmit power with reliability constraint," *Wireless Networks*, vol. 26, no. 5, pp. 3245–3256, 2020.
[34] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 2, pp. 253–279, 2019.
[35] S. Tu and B. Recht, "The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint," in *Proc. COLT*, 2019.
[36] G. Zhang *et al.*, "When does deep reinforcement learning beat calibrated baselines? A benchmark study on adaptive resource control," arXiv:2605.26418, 2026.
[37] E. Baena, P. Testolina, M. Polese, D. Koutsonikolas, J. Jornet, and T. Melodia, "Space-O-RAN: Enabling intelligent, open, and interoperable non-terrestrial networks in 6G," *IEEE Commun. Mag.*, vol. 64, no. 2, pp. 112–118, Feb. 2026.
[38] M. M. H. Qazzaz *et al.*, "xApp empowered resource management for non-terrestrial users in 5G O-RAN networks," arXiv:2605.10704, 2026.