Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Frame-Level LLR Calibration and Risk-Controlled Blind Copies for Soft-Combined Underwater Acoustic HARQ: A Simulation Study

Mr. Sengol I

imsengol@gmail.com

St. Joseph college of engineering, Sriperpumbudur

Dr. Regilan S

Asan memorial college of engineering and technology

Mr. Pavunboss S

St. Joseph college of engineering, Sriperpumbudur

Mr. Savariappan L

St. Joseph college of engineering, Sriperpumbudur

Research Article

Keywords: underwater acoustic communication, learned receiver, calibration, hybrid ARQ, conformal risk control, log-likelihood ratio

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11591922/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Frame-Level LLR Calibration and Risk-Controlled Blind Copies for Soft-Combined Underwater Acoustic HARQ: A Simulation Study

Mr. I. Sengol [1], Assistant Professor, Department of Computer Science & Engineering, St. Joseph College of Engineering, Sriperumbudur, Chennai, India. imsengol@gmail.com*

Dr. S. Regilan [2] , Associate Professor, Department of Electronics and Communication Engineering, Asan Memorial College of Engineering and Technology, Chennai, India.* lssregilan@gmail.com

Mr. S. Pavun Boss [3], Assistant Professor, Department of Computer Science & Engineering, St. Joseph College of Engineering, Sriperumbudur, Chennai, India. pavunboss@gmail.com*

Mr. L. Savariappan [4], Assistant Professor, Department of Artificial Intelligence and Data Science, St. Joseph College of Engineering, Sriperumbudur, Chennai, India. savariappan@gmail.com*

Abstract. Underwater acoustic links have round-trip times of seconds, so a sender must decide how much redundancy to send before feedback arrives, and the confidence of a receiver's soft outputs matters when retransmitted copies are combined. In a simulated OFDM link with a classical receiver and two small learned receivers in five channel regimes, we show that a max-log Viterbi decoder on real-valued inputs is blind to a common scale, so a global or session-level temperature cannot change a combined decision; only the variation of confidence across copies matters. A temperature fitted per frame on known pad bits, which cost no airtime in this frame design, reduces the copies the classical receiver needs per delivery by 4.1% in the combined regime and 2.2% under high Doppler when the copies fade independently (confirmed on fresh sessions, Holm-adjusted $p \leq 0.016$). The gain is conditional: it falls to +0.037 copies in the combined regime at a correlation of 0.5 between the copies and is gone at 0.9, a simple pilot-variance estimator does not reproduce it, and it does not reliably carry over to quantised decoders. A network trained on several regimes saves more (10.3%). A conformal risk-control rule for the number of blind first-round copies met its deadline target in all but 2 of 200 cells when calibrated in the same regime, failed in 94 of 120 when calibrated in the wrong one, and became infeasible when the copies were strongly correlated; it raised goodput by 60% over stop-and-wait at a two-second round trip in the hardest regime, but not over a well-chosen constant number of copies. The effects are small in absolute terms, and the study is simulated.

Keywords: underwater acoustic communication; learned receiver; calibration; hybrid ARQ; conformal risk control; log-likelihood ratio

1. Introduction

Sound travels at about 1500 m/s, so a link of 1.5 km has a round-trip time of about two seconds, while a frame of a few hundred bits occupies the channel for a fraction of a second. A hybrid automatic-repeat-request (HARQ) scheme therefore pays for every retransmission round with an idle channel, and

designers send redundancy ahead of the acknowledgement instead of waiting for it [1, 2]. At the same time, receivers that learn to demodulate directly from the received OFDM signal have started to appear in underwater trials [3], following the terrestrial line of work on learned receivers [4].

A learned receiver is trained in one channel regime, but the underwater channel changes with the sea state, platform motion and the snapping-shrimp noise that dominates many shallow-water sites [5]. The receiver's soft outputs, the log-likelihood ratios (LLRs), feed the channel decoder and, in a HARQ scheme, the combiner that merges successive copies of a frame. If the LLRs of a copy are too confident or not confident enough relative to those of the other copies, the combiner weighs the copies wrongly. Neural networks are known to be miscalibrated [6]; a terrestrial 5G study has shown that a neural receiver can fail silently when the setting moves outside its training envelope, and recommended running a conventional receiver alongside it [7]; conformal prediction has been used to calibrate set predictions of AI demodulators [8, 9]. What has not been examined, to our knowledge, is what calibration does to soft-combined retransmission on an acoustic link, and what a sender with a slow feedback channel should do with it.

This paper studies that question with a simulated time-domain acoustic OFDM link, a classical receiver, two small learned receivers and a family of LLR calibrators. Our starting observation is mathematical and simple. A max-log Viterbi decoder is blind to a common scale factor on its input (Lemma 1), so the usual recipe of fitting one temperature on a validation set, or even re-fitting it over a session, cannot change any decision of a soft-combined HARQ receiver. Only the *relative* confidence of the copies matters, so calibration must act at the level of the frame. Each frame of our link carries a few known pad bits that cost no airtime, and these give a per-frame calibration signal.

The second half of the paper concerns the sender. Because feedback arrives a round trip late, the sender can only choose how many copies k to send blindly in the first round. We choose k by conformal risk control [10] so that the probability that delivery needs more than $R_{\max}$ rounds stays below a level α, and we test the guarantee, which needs exchangeability, against regime shifts, with a static, a rolling-window and an adaptive controller [11].

Contributions.

1. *Theory, restated for HARQ.* The scale invariance of the max-log decoder, a closed-form combining efficiency $\eta = 1/(1 + \mathrm{CV}^2)$ for Chase combining of copies with different overconfidence, and the corollary that a shared temperature (global or per session) cannot change combined decisions of an unquantised decoder (Section 4.2); the corollary is shown to fail for 3-bit and 4-bit inputs (Section 5.8). These are known ingredients of soft decoding; we state them in the form needed to decide where calibration can help.
2. *A per-frame calibrator and its comparison.* A temperature fitted on the known pad bits of each frame with a prior on its log, compared with no calibration, two decision-directed session-level calibrators, a pilot-variance demapper, and an oracle (Sections 4.1 and 5).
3. *A controlled study with a confirmatory set.* Five channel regimes (one never used in training), three receivers, ten sessions per regime in an evaluation set and ten in a confirmatory set whose primary comparison was fixed before the runs (Sections 4.3 and 5.1 to 5.5).
4. *Conditions under which the gain holds.* Stress tests on channel memory across copies, pad position, interleaving, quantisation, penalty and the type of trial that benefits (Section 5.8): the gain needs copies that fade (partly) independently, and it vanishes at a correlation of 0.9.

5. *Risk-controlled blind copies.* A conformal rule for the number of first-round copies with a finite-sample guarantee (Proposition 2), tested in and out of regime, under regime switching and against constant numbers of copies (Sections 5.6 and 5.7).

Findings in brief. (i) A learned receiver trained in one regime and a classical receiver are both overconfident out of regime (median oracle temperature up to 5.07 for the classical and 1.60 for the single-regime network in the combined regime DSN), while a network trained on several regimes stays close to calibrated (1.01). (ii) With independent fading across copies, frame-level calibration from the pad bits reduces the copies needed to deliver by 4.1% in DSN (confirmed on fresh sessions, Holm-adjusted $p = 0.008$) and by 2.2% in the high-Doppler regime, and it does nothing in the training regime or under impulsive noise; session-level calibration, global or decision-directed, changes no decision of an unquantised decoder, as the theory says. (iii) The gain depends on the copies seeing different channels: it is +0.037 copies in DSN at a correlation of 0.5 and +0.005 at 0.9, and the oracle shows the same decay. (iv) Training the network on several regimes saves more than calibration does (0.249 copies in DSN against 0.100). (v) Conformal control of the first-round copies kept the deadline-miss rate at or below the target in all but 2 of 200 cells when calibrated in the same regime, and was violated in 94 of 120 cells when calibrated in the wrong regime; it beats stop-and-wait but not a well-chosen constant number of copies in goodput. The effects of calibration are small in absolute terms, and the study is simulated.

2. Related Work

Learned receivers for underwater acoustic OFDM. A deep neural network receiver that maps the received OFDM signal to bits without separate channel estimation, equalisation and demapping has been tested in a river trial [3]; its authors note that a network trained on one channel may need retraining when the channel changes. DeepRx and its successors show gains of learned receivers over conventional ones on terrestrial links [4]. A recent benchmark of a public 5G neural receiver finds gains in a few scenarios, ties in most, and a silent, confident failure at one setting outside training; its detect-and-rollback layer chooses between a neural and a conventional receiver per slot and is evaluated without HARQ [7]. We differ in the medium, in studying what calibration does to HARQ combining, and in treating the sender's decision.

Calibration and conformal methods in wireless. Temperature scaling is a standard post-hoc fix for overconfident networks [6]. Conformal prediction, which gives finite-sample coverage under exchangeability [12], has been applied to symbol demodulation, modulation classification and channel prediction, including an online variant for time-varying data [8, 9]; those works calibrate set-valued predictions of the symbol, not the LLRs that feed a decoder, and have no retransmission. Conformal risk control extends the guarantee to any bounded monotone loss [10], and adaptive conformal inference tracks a target under distribution shift by feeding back the realised errors [11]. We use the first for the choice of blind copies and compare with the second.

HARQ underwater. Underwater ARQ and HARQ are limited by long propagation delay; cooperative HARQ with incremental redundancy has been proposed for sensor networks and shown to raise throughput and energy efficiency over stop-and-wait ARQ [1]. Chase combining adds the LLRs of copies and is optimal when the LLRs are consistent [13]. We keep the protocol simple (Chase combining, blind copies in the first round) and ask how the quality of the LLRs and of the sender's choice enters.

Reliability weighting, LLR scaling and soft-input quantisation. The idea that soft information must be scaled by the reliability of the copy that produced it is old. The maximum-likelihood combination of copies weights each by its reliability [13]; soft-combining HARQ with Chase combining and incremental redundancy has been analysed extensively for cellular systems [14]; data-aided SNR estimation from known symbols is a standard receiver function [15]. Scaling the soft values that enter a max-log decoder is a known practical fix in turbo decoding, where the max-log approximation overestimates the extrinsic information [16]. The scale of the soft input of a Viterbi decoder matters in fixed-point implementations, which quantise to a few bits [17]. The pad-bit calibrator of this paper is therefore best read as a per-frame, data-aided estimate of the reliability of a copy, obtained at no airtime cost in this frame design; our contribution is not the idea of reliability weighting but its evaluation for learned and classical underwater receivers, the identification of what a max-log decoder is and is not sensitive to, and the sender's decision. Section 5.8 compares the calibrator with a per-frame variance estimate taken from the pilots, which needs no known bits.

Impulsive ambient noise. Snapping shrimp produce impulsive noise that departs from the Gaussian assumption of conventional demappers [5]; we model it as sparse high-power impulses in the time domain.

Gap. We found no prior study that combines frame-level LLR calibration of a learned underwater receiver, soft-combined retransmission and conformal control of the sender's redundancy. Absence in our searches is not proof that none exists.

3. System Model

3.1 Link, frame and channel regimes

The link is a time-domain simulation of one transmitter and $M = 2$ receive antennas (half-wavelength spacing). The carrier is 12 kHz and the sampling rate 4 kHz. An OFDM symbol has 128 subcarriers (spacing 31.25 Hz), a 16 ms cyclic prefix and a total duration $T_{\mathrm{sym}} = 48$ ms. Every fourth subcarrier is a pilot (32 per symbol) and 96 carry QPSK data. A frame carries 512 information bits (including a 16-bit CRC, treated as an ideal error detector), encoded by the rate-1/2 constraint-length-7 convolutional code (171, 133 octal) with tail termination. The 1036 coded bits and 116 known *pad* bits fill $N_{\mathrm{sym}} = 6$ OFDM symbols, so a frame lasts $T_F = 0.288$ s and the pad costs no airtime. The pad bits are the unused tail of the last OFDM symbol, and in the main study the coded bits are not interleaved; Section 5.8 moves the pad and adds an interleaver.

The channel has five paths with an exponential power-delay profile (decay constant σ_τ, delays capped at $\tau_{\max}$), complex Gaussian gains, random arrival angles at the array and per-path Doppler shifts $f_d = f_c v \cos\varphi / c + \sigma_f \xi$ with speed $v \sim U(0, v_{\max})$ and Gaussian jitter ξ. The designed arrival of the first path is 4 ms after the symbol start. The channel is redrawn for each transmission (independent fading across attempts, which favours a per-copy calibrator; Section 5.8 replaces it by correlated gains). The per-subcarrier SNR of each attempt is drawn uniformly from 3 to 13 dB. Noise is complex Gaussian; in the impulsive regimes each time sample additionally receives, with a small probability, a Gaussian impulse 18 dB above the noise floor.

Table 1 defines five regimes. T is the training regime. D, S and N change one aspect at a time; DSN changes all three and is never used in training or in tuning.

Table 1. Channel regimes. f_d per path is $12000\,v\cos\varphi/1500$ Hz plus jitter.

Regime	max speed (m/s)	Doppler jitter (Hz)	delay decay (ms)	max excess delay (ms)	impulse prob.
T: training regime	0.5	0.2	1.5	6	0
D: high Doppler	2.0	1.0	1.5	6	0
S: longer delay spread	0.5	0.2	3.5	7	0
N: impulsive noise	0.5	0.2	1.5	6	0.4%
DSN: all three (never in training)	2.0	1.0	3.5	7	0.4%

3.2 Receivers

All receivers estimate the channel by least squares at the pilots of each symbol, after removing the known designed delay, and interpolate linearly across frequency.

Classical. Maximal-ratio combining of the two antennas followed by a Gaussian soft demapper that assumes the known noise variance (unit noise floor); the pilot-based estimate carries the signal scale, so the receiver needs no SNR estimate. The LLR of a bit on the in-phase (quadrature) component of a Gray-mapped QPSK symbol is $2\sqrt{2}\,\mathrm{Re}(z)/\sigma^2$ ($2\sqrt{2}\,\mathrm{Im}(z)/\sigma^2$), with $z = \sum_m \hat{h}_m^* y_m$. The classical receiver does not model channel-estimation error, intercarrier interference or impulses, which makes its LLRs overconfident.

Learned. A multilayer perceptron with two hidden layers of 64 ReLU units maps 24 features per data resource element to the posterior over the four QPSK symbols; the bit LLRs are the log-ratios of the marginals. The features are the received values and channel estimates of both antennas, the classical equalised symbol and its gain, the energy at the element and at its two neighbours, the roughness of the pilot channel estimate (a Doppler and delay indicator) and the per-frame SNR, which we supply as an estimate (Appendix B perturbs it by 2 and 4 dB). We train it with Adam on 600 000 elements (cross-entropy). *NN-T* is trained on 1600 frames of regime T only; *NN-mix* is trained on 400 frames of each of T, D, S and N. Comparing the two separates the effect of calibration from the effect of broader training.

3.3 HARQ with a long round trip

A packet is sent as copies of the same codeword. The receiver adds the LLRs of the copies received so far (Chase combining), decodes after every arrival and sends an acknowledgement. Let c be the number of copies that must be combined before the CRC passes ($c = \infty$ if not within $J = 5$). The sender transmits k copies back to back, waits one round-trip time, and then sends one copy per round until acknowledged, so that delivery takes $1 + \max(0, c - k)$ rounds. The sender's channel occupancy per delivered packet is

$$t(c,k) = \Big(\max\big(k, \min(c,J)\big)\Big)\, T_F + \big(1 + \max(0, c-k)\big)\,\mathrm{RTT}, \qquad (1)$$

and the goodput is the 496 payload bits times the delivery ratio divided by the mean occupancy. We use $\mathrm{RTT} = 2$ s (a 1.5 km link) and vary it from 0.5 to 8 s.

Table 2. Notation.

Symbol	Meaning
$\ell, \tilde{\ell}$	LLR of a bit from a receiver; calibrated LLR $\ell/\hat{\kappa}$
$\kappa, \hat{\kappa}$	temperature (overconfidence factor, > 1 means overconfident) and its estimate
c, k, J	copies needed to decode; blind first-round copies; maximum copies (5)
$R_{\max}, \alpha$	deadline in rounds (2); target deadline-miss probability
T_F, RTT	frame duration (0.288 s); round-trip time
μ_i, η	calibrated mean LLR of copy i; combining efficiency
ECE	expected calibration error of the bit probabilities

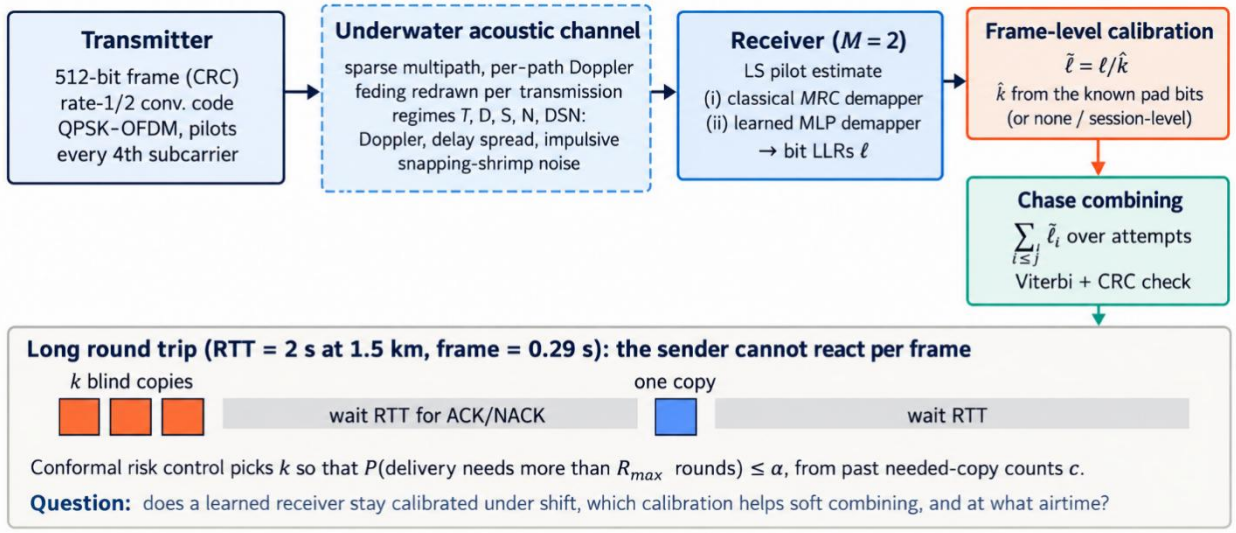


Figure 1. *System model. A frame passes through a time-varying acoustic channel to a receiver (classical or learned) that outputs bit LLRs; a calibrator rescales them per frame; Chase combining adds the calibrated LLRs of the copies, followed by Viterbi decoding and a CRC check. Feedback arrives a round trip later, so the sender chooses k blind first-round copies by conformal risk control.*

4. Method

4.1 LLR calibration

A receiver output ℓ is *calibrated* if $P(b = 0 \mid \ell) = \sigma(\ell)$ with σ the logistic function. We calibrate by a temperature: $\tilde{\ell} = \ell/\hat{\kappa}$, where $\kappa > 1$ means the receiver is overconfident. For a set of bits with known values b_j the negative log-likelihood is $\sum_j \log\left(1 + \exp\left(-s_j \ell_j/\kappa\right)\right)$ with $s_j = 1 - 2b_j$, minimised over a grid of 60 log-spaced values from 0.15 to 25. We compare five calibrators:

- **none**: $\hat{\kappa} = 1$.
- **pad**: $\hat{\kappa}$ minimises the negative log-likelihood on the 116 known pad bits of the frame itself, plus a Gaussian prior on $\log\kappa$ with centre $\log\kappa_g$ (the pooled temperature fitted on an in-distribution validation set: 2.21 for the classical receiver and 1.01 for the networks) and strength λ pseudo-bits ($\lambda = 20.0$, chosen on pilot sessions). It costs no airtime and no extra transmission in this frame design: the 116 bits are the unused tail of the last OFDM symbol, which a 1036-bit codeword does not fill (Section 3.1). Because they sit in one place, the temperature is a local estimate; Section 5.8 moves them to the head of the frame and spreads them through it.
- **dd** (decision-directed, session level): $\hat{\kappa}$ is the pooled temperature of the last $W = 100$ frames that decoded on their own (CRC passed), whose bits are then known. It costs nothing, but the labelled frames are the easy ones.

- **ddreg** (decision-directed, frame level): $\log\hat{\kappa} = a + b\,(\phi - \bar{\phi})$ with ϕ = log of the mean $|\ell|$ of the frame, fitted on the same labelled frames.
- **oracle**: $\hat{\kappa}$ minimises the negative log-likelihood on all bits of the frame (uses the transmitted bits; an upper bound).

A plain *global* temperature, the common recipe, is not listed because by Corollary 1 it gives the same decisions as **none** in every cell of the study; it enters only through the prior of **pad** and through the calibration diagnostics. We report the expected calibration error (ECE) of the bit probabilities in 15 bins and the median oracle temperature as a measure of overconfidence.

4.2 Theory

Let $\ell \in \mathbb{R}^{2T}$ be the LLRs of one codeword. The max-log Viterbi decoder returns the sequence $\hat{u} = \operatorname{argmax}_{u\in\mathcal{C}} \sum_j \left(1 - 2c_j(u)\right) \ell_j/2$ over the code $\mathcal{C}$ [18].

Lemma 1 (scale invariance). For every $\gamma > 0$ the decoder output for $\gamma\ell$ equals that for ℓ.

Proof. The path metric of every codeword is multiplied by γ, which preserves the order. ▫

Corollary 1 (shared temperatures do nothing). Chase combining of copies $i = 1, \ldots, j$ forms $\sum_i \ell_i / \hat{\kappa}_i$. If $\hat{\kappa}_i = \hat{\kappa}$ for all i, the combined LLR is $\hat{\kappa}^{-1} \sum_i \ell_i$ and, by Lemma 1, every decision is identical to that of no calibration. In particular a global temperature, and any temperature that is constant within a trial, cannot change a combined decision or a CRC outcome.

Remark 1 (what the corollary does not say). Lemma 1 holds for a decoder that works on real-valued soft inputs. A fixed-point Viterbi decoder quantises the soft input to a few bits with a fixed step and clips it [17]; the scale of the input then decides how much of the useful range is used, and a global scale does change decisions. The statement "a global temperature cannot help" is therefore a property of unquantised max-log decoding. Section 5.8 measures the effect for 3-bit and 4-bit inputs; there a good global scale matters, a per-frame scale matters in addition, and the two should not be confused.

Proposition 1 (combining efficiency). Suppose the calibrated LLR of copy i is Gaussian conditional on the bit with mean $s\mu_i$ and variance $2\mu_i$ ($s = \pm 1$; this is the consistency condition of a calibrated Gaussian LLR), and the receiver outputs $\ell_i = \kappa_i \tilde{\ell}_i$. Let $p_i = \mu_i / \sum_m \mu_m$. The signal-to-noise ratio of the Chase-combined statistic $\sum_i \ell_i$ is the one of the ideal combination $\sum_i \tilde{\ell}_i$ multiplied by

$$\eta = \frac{(\sum_i \kappa_i \mu_i)^2}{(\sum_i \mu_i)(\sum_i \kappa_i^2 \mu_i)} = \frac{1}{1 + \mathrm{CV}_p^2(\kappa)}, \qquad \mathrm{CV}_p^2(\kappa) = \frac{\mathrm{Var}_p(\kappa)}{\mathbb{E}_p[\kappa]^2} \le 1/4\,(\kappa_{\max}/\kappa_{\min} - 1)^2. \qquad (2)$$

Thus $\eta \le 1$ with equality iff κ_i is constant over the copies with $\mu_i > 0$. With a calibrator, κ_i is replaced by $\kappa_i/\hat{\kappa}_i$.

Proof. The statistic has mean $s \sum_i \kappa_i \mu_i$ and variance $2 \sum_i \kappa_i^2 \mu_i$, so its SNR is $(\sum \kappa_i \mu_i)^2 / (2 \sum \kappa_i^2 \mu_i)$, while the ideal statistic has SNR $\sum \mu_i / 2$. The ratio is $\mathbb{E}_p[\kappa]^2 / \mathbb{E}_p[\kappa^2] = 1/(1 + \mathrm{CV}_p^2)$. The bound on CV^2 is Popoviciu's inequality. ▫

Proposition 1 is the familiar loss of maximal-ratio combining with mismatched weights, written for LLRs; we state it because it says what calibration has to fix. It is what makes calibration matter and says where: not in the level of κ but in how much it varies from copy to copy, weighted by the reliability of the copies. A receiver that is overconfident by a fixed factor is harmless; one whose overconfidence

changes with the fading realisation of each copy wastes a share $1-\eta$ of the combining gain. We test equation (2) in Section 5.9.

Proposition 2 (risk-controlled blind copies). Let $c_1, \dots, c_n, c_{n+1}$ be exchangeable copy counts and $\ell_k(c) = \mathbf{1}\{c > k + R_{\max} - 1\}$, the indicator that delivery needs more than $R_{\max}$ rounds with k first-round copies; ℓ_k is non-increasing in k with values in $\{0,1\}$. Let

$$\hat{k} = \min\left\{k \in \{1, \dots, K_{\max}\}: \frac{1 + \sum_{i \le n} \ell_k(c_i)}{n+1} \le \alpha\right\}. \qquad (3)$$

If the set is non-empty, then $\mathbb{E}\,\ell_{\hat{k}}(c_{n+1}) \le \alpha$.

Proof. This is the conformal risk control theorem [10] for a monotone loss; we give the short argument. Let $k^\star$ be the smallest k with $\sum_{i \le n+1} \ell_k(c_i)/(n+1) \le \alpha$. Since $\ell \le 1$, the left side of (3) is at least that average, so $\hat{k} \ge k^\star$ and, by monotonicity, $\ell_{\hat{k}}(c_{n+1}) \le \ell_{k^\star}(c_{n+1})$. Because $k^\star$ is a symmetric function of $c_1, \dots, c_{n+1}$, exchangeability gives $\mathbb{E}\,\ell_{k^\star}(c_{n+1}) = \mathbb{E}\,\frac{1}{n+1}\sum_{i \le n+1} \ell_{k^\star}(c_i) \le \alpha$. ▫

The guarantee is over the draw of the calibration set and the test packet, and it needs exchangeability. A regime shift breaks it; we therefore compare three controllers in Section 5.6: **static** (calibrated once, in regime T), **rolling** (the last $W_r = 100$ observed counts) and **ACI**, which in addition updates its level by $\alpha_{t+1} = \alpha_t + \gamma(\alpha - e_t)$ with e_t the realised miss [11]. The long-run guarantee of adaptive conformal inference relies on a prediction-set family that can reach zero error; with a finite $K_{\max}$ this does not hold here, so we claim only empirical behaviour for ACI. The counts c are fully observable: the receiver decodes after every arrival, so the first arrival after which the CRC passes is known (up to J), and the loss of every candidate k can be evaluated in hindsight.

4.3 Experimental protocol

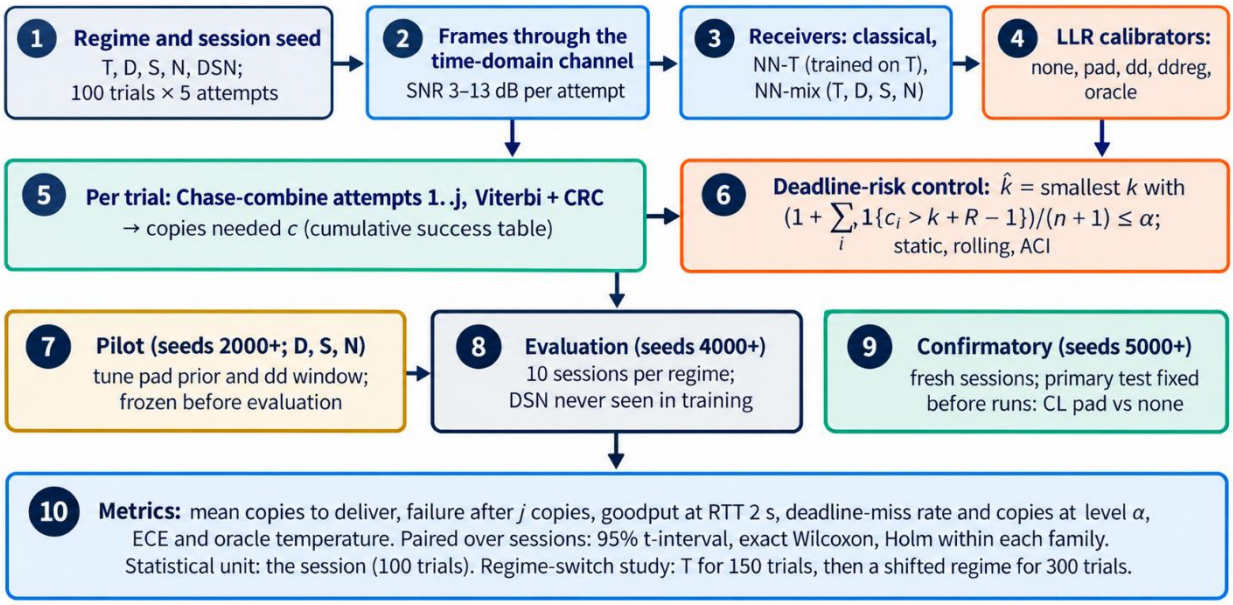


Figure 2. *Methodology. Sessions in a regime pass through the channel; three receivers and five calibrators give LLRs; Chase combining and Viterbi decoding give the number of copies c needed per trial; conformal control picks the blind copies k. Hyper-parameters are tuned on pilot sessions, the comparisons run on evaluation sessions, and the primary test is repeated on confirmatory sessions fixed in advance.*

Worlds and sets. A *session* is a run of 100 HARQ trials of 5 attempts each in one regime; its seed fixes the channel draws, noise, data and per-attempt SNRs. Five seed sets are disjoint: training (1000s), validation (3000s, 3100s; to fit κ_g and the global temperature), pilot (2000s; regimes D, S, N; four sessions each), evaluation (4000s; all five regimes; ten sessions each) and confirmatory (5000s; all five regimes; ten fresh sessions each). The statistical unit is the session.

Tuning. The prior strength $\lambda \in \{0,20,60,200\}$ and the decision-directed window $W \in \{10,30,100\}$ were chosen on the pilot sessions by the largest reduction in summed failures over $j = 1, \ldots, 5$ combined copies (Appendix B) and frozen before the evaluation and confirmatory runs. The held-out regime DSN and the evaluation seeds were not used for tuning. The neural networks were trained once, before any evaluation run, with a fixed budget of 20 epochs and no model selection.

Primary comparison. Fixed after tuning and before the evaluation and confirmatory runs: the classical receiver with the **pad** calibrator against the classical receiver with no calibrator, on the mean number of copies to deliver, in each of the four shifted regimes D, S, N and DSN, with Holm correction over the four. It is evaluated on the confirmatory sessions. Every other comparison is secondary and labelled as such.

Statistics. Differences are paired over sessions. We report the mean difference with a 95% t-interval ($n = 10$) and the exact two-sided Wilcoxon signed-rank p-value (smallest possible 0.002, which is reached when all ten sessions have the same sign; after Holm correction over four tests the smallest adjusted value is 0.008, and that value therefore means "all ten sessions agree", not a very small probability). p-values are Holm-adjusted within each family (the four primary tests; the 15 receiver-by-regime tests of pad against none; the 15 of oracle against none; the six of the decision-directed calibrators; the receiver comparisons). A difference is *established* if the interval excludes zero and the Holm-adjusted $p < 0.05$, *suggested* if the interval excludes zero or the unadjusted $p < 0.05$ but the adjusted p fails, and *not distinguishable* (n.d. in the tables) otherwise. The mean number of copies to deliver counts an undelivered packet as $J + 1 = 6$ copies and is reported together with the delivery ratio.

What was decided after seeing data. The list of calibrators was extended after the first test session with the frame-level decision-directed regression (ddreg), and the global temperature was dropped from the calibrator set after Corollary 1 showed it to be redundant; both happened before the pilot tuning. The regime-switch study, the goodput analysis and the theory check were designed after the evaluation sessions had been generated, using the same cached sessions; we mark them as exploratory. After the main results had been reviewed, we added the stress tests of Section 5.8 (channel memory, pilot-variance demapper, pad position, interleaver, quantised input, trial-type split, penalty), the fixed-k baselines of Section 5.7 and the feasibility check of Section 5.6; the results of these were reported whatever they showed, and they changed the claims of the paper (the gain is conditional on independent fading; the controller's goodput gain is relative to stop-and-wait only).

Compute. About 90 000 simulated frames for the main study (roughly 1.5 hours on two CPU cores) and about 130 000 for the stress tests of Section 5.8; the code is in the archive.

5. Results

All results come from simulation. Unless stated otherwise they are means over the ten evaluation sessions of a regime (100 trials of up to five copies each, SNR 3 to 13 dB per copy, RTT $= 2$ s). "Copies

to deliver" is the mean number of copies that must be combined before the CRC passes, counting an undelivered packet as six; positive differences are copies saved. Every numeric table has a matching figure.

5.1 Receivers and calibration diagnostics

Table 3. Calibration diagnostics per receiver and regime (evaluation sessions). ECE: expected calibration error of the bit probabilities; the oracle temperature κ is fitted per frame on all bits, so $\kappa > 1$ means overconfident.

Regime	Receiver	Uncoded BER	ECE	Oracle κ (median)	Oracle κ (geo. mean)	FER after 1 copy
T	Classical	0.052	0.025	2.02	1.98	0.304
T	NN-T	0.050	0.001	0.93	0.89	0.289
T	NN-mix	0.051	0.008	0.85	0.78	0.295
D	Classical	0.094	0.062	3.09	3.49	0.564
D	NN-T	0.093	0.025	1.20	1.25	0.552
D	NN-mix	0.092	0.002	0.91	0.90	0.547
S	Classical	0.101	0.067	2.93	3.27	0.449
S	NN-T	0.090	0.022	1.10	1.17	0.413
S	NN-mix	0.086	0.007	0.86	0.91	0.399
N	Classical	0.065	0.037	2.62	2.60	0.398
N	NN-T	0.063	0.010	1.10	1.08	0.390
N	NN-mix	0.064	0.003	0.93	0.91	0.392
DSN	Classical	0.156	0.118	5.07	5.68	0.739
DSN	NN-T	0.147	0.063	1.60	1.74	0.724
DSN	NN-mix	0.145	0.018	1.01	1.08	0.713

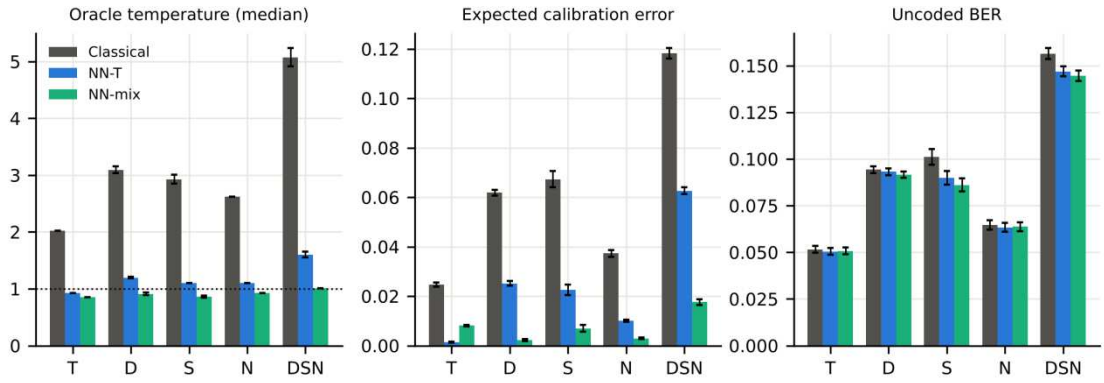


Figure 3. *Oracle temperature (median over frames), expected calibration error and uncoded bit-error rate per regime and receiver (bars: mean over ten sessions with 95% intervals). The dotted line marks a calibrated receiver,* $\kappa = 1$.

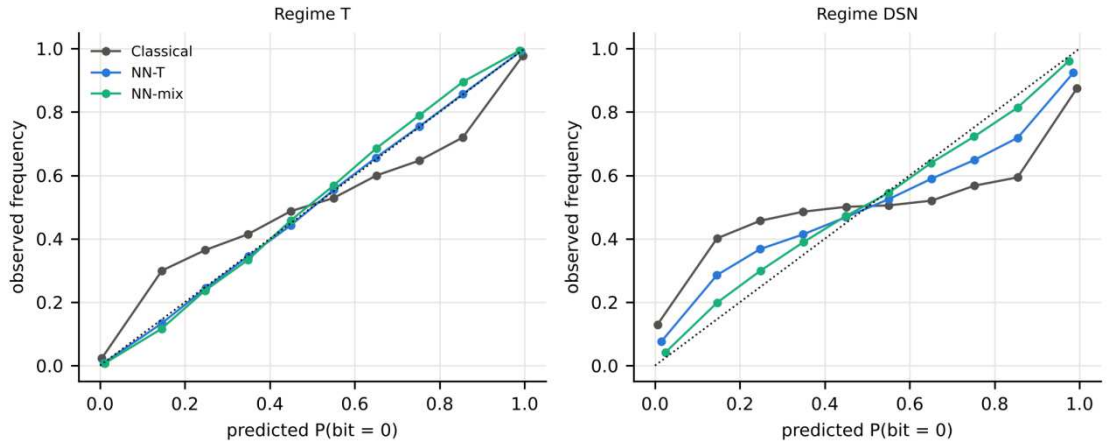


Figure 4. *Reliability diagrams of the bit probabilities in the training regime (left) and in the unseen combined regime DSN (right). The classical receiver is overconfident in both; the network trained on one regime is calibrated in T and overconfident in DSN.*

The classical receiver is overconfident even in the training regime (median $\kappa = 2.02$, because the Gaussian demapper ignores channel-estimation error) and increasingly so under shift, up to 5.07 in DSN (Table 3, Figure 3). The network trained on regime T alone is calibrated there (0.93) and becomes overconfident under shift: 1.20, 1.10 and 1.10 in D, S and N and 1.60 in DSN, the combination it never saw. The network trained on four regimes stays between 0.85 and 1.01 everywhere, including DSN. This is the silent-failure pattern reported for terrestrial receivers [7], seen here as a loss of calibration rather than a collapse of accuracy: the uncoded bit-error rates of the two networks are at or below the classical one in every regime (for example 0.101 classical, 0.090 NN-T and 0.086 NN-mix in S). The failure rate after a single copy is 30.4% (classical) and 28.9% (NN-T) in T and rises to 73.9% and 72.4% in DSN, and, by Lemma 1, no calibrator changes it.

5.2 Frame-level calibration and soft combining

Table 4. Mean copies to deliver by receiver, regime and calibrator (evaluation sessions). Lower is better; dd is the session-level decision-directed calibrator and ddreg its frame-level regression.

Regime	Receiver	none	pad	dd	ddreg	oracle
T	Classical	1.312	1.313	1.312	1.311	1.312
T	NN-T	1.298	1.298	1.298	1.297	1.297
T	NN-mix	1.304	1.304	1.304	1.304	1.302
D	Classical	1.803	1.764	1.803	1.788	1.719
D	NN-T	1.727	1.706	1.727	1.727	1.693
D	NN-mix	1.716	1.711	1.716	1.720	1.696
S	Classical	1.635	1.618	1.635	1.641	1.600
S	NN-T	1.526	1.513	1.526	1.526	1.502
S	NN-mix	1.491	1.486	1.491	1.494	1.476
N	Classical	1.429	1.430	1.429	1.427	1.420
N	NN-T	1.422	1.422	1.422	1.424	1.417
N	NN-mix	1.423	1.424	1.423	1.424	1.416
DSN	Classical	2.422	2.322	2.422	2.433	2.257
DSN	NN-T	2.236	2.203	2.236	2.233	2.174
DSN	NN-mix	2.173	2.142	2.173	2.202	2.127

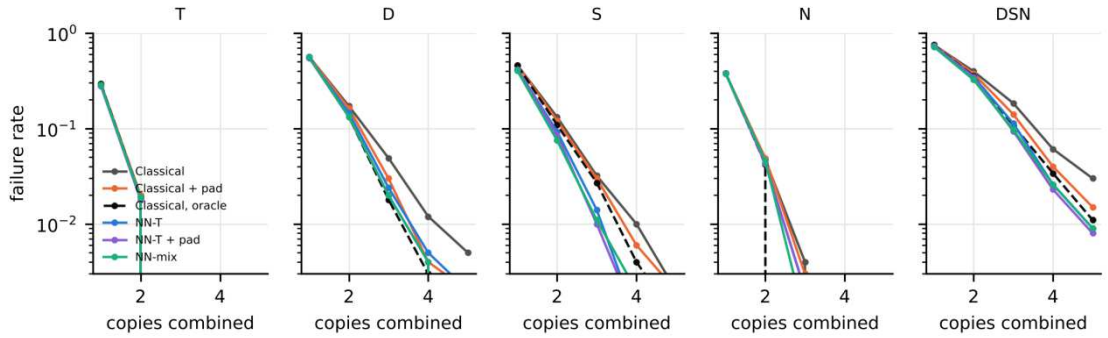


Figure 5. *Failure rate after combining j copies, per regime, for the classical receiver (none, pad, oracle), the network trained on T (none, pad) and the network trained on four regimes (none). Calibration lowers the failure of the combined copies in D, S and DSN and not in T and N.*

Table 5. Copies saved per delivery by the pad calibrator relative to no calibration, by receiver and regime (evaluation sessions; Holm correction over the 15 comparisons).

Receiver	Regime	Δ copies	95% t-interval	Wilcoxon p	Holm p	Sessions won	Label
Classical	T	-0.001	[-0.003, +0.001]	1.000	1.000	0/10	n.d.
Classical	D	+0.039	[+0.026, +0.052]	0.002	0.029	10/10	established
Classical	S	+0.017	[-0.001, +0.035]	0.078	0.703	7/10	n.d.
Classical	N	-0.001	[-0.006, +0.004]	1.000	1.000	1/10	n.d.
Classical	DSN	+0.100	[+0.087, +0.113]	0.002	0.029	10/10	established
NN-T	T	+0.000	[-0.003, +0.003]	1.000	1.000	1/10	n.d.
NN-T	D	+0.021	[+0.005, +0.037]	0.023	0.258	7/10	suggested
NN-T	S	+0.013	[+0.004, +0.022]	0.031	0.312	6/10	suggested
NN-T	N	+0.000	[-0.005, +0.005]	1.000	1.000	2/10	n.d.
NN-T	DSN	+0.033	[+0.018, +0.048]	0.004	0.051	9/10	suggested
NN-mix	T	+0.000	[+0.000, +0.000]	1.000	1.000	0/10	n.d.
NN-mix	D	+0.005	[-0.002, +0.012]	0.234	1.000	6/10	n.d.
NN-mix	S	+0.005	[-0.004, +0.014]	0.375	1.000	5/10	n.d.
NN-mix	N	-0.001	[-0.003, +0.001]	1.000	1.000	0/10	n.d.
NN-mix	DSN	+0.031	[+0.015, +0.047]	0.008	0.094	8/10	suggested

Calibration from the pad bits does nothing in the training regime T and under impulsive noise N (differences of -0.001 and -0.001 copies for the classical receiver, not distinguishable). In the high-Doppler regime D it saves +0.039 copies (2.2%, interval [+0.026, +0.052], established) and in DSN +0.100 copies (4.1%, [+0.087, +0.113], established), against +0.165 for the oracle; the pad bits therefore capture about 61% of what a perfect per-frame temperature could give in DSN and 46% in D. For the network trained on T the gains are smaller and, after correction, suggested only (+0.021, +0.013 and +0.033 in D, S and DSN); for the network trained on four regimes only DSN shows a suggested gain (+0.031), as one would expect from a receiver that is already close to calibrated (Figure 5, Figure 6, left).

Proposition 1 explains the pattern qualitatively but not completely. Calibration can only help trials in which combining is needed: 30.4% of single copies fail in T and 73.9% in DSN. It also needs the overconfidence to vary from copy to copy, and the spread of the per-frame log temperature is not much larger in DSN (0.64 for the classical receiver) than in T (0.60); the benefit in DSN comes from the trials that need three or more copies, not from the marginal trials that need a second one (Section 5.8). Under impulsive noise even the oracle temperature saves only +0.009 copies for the classical receiver: impulses corrupt a few symbols, which a single temperature per frame cannot repair. We have not tested a symbol-level calibrator; we state this as a hypothesis.

5.3 Session-level calibrators change nothing

The decision-directed session-level calibrator dd gives exactly the same copies to deliver as no calibration in every regime (Table 4; difference 0.000, not distinguishable), as Corollary 1 requires: its temperature changes only slowly, so it is almost constant within a trial. The frame-level regression ddreg, which could in principle follow the frame, saves +0.015 copies in D and loses in S and DSN (-0.006 and -0.011); none of its six comparisons is distinguishable from zero (Table 6). We attribute this to a selection effect: only frames that decoded on their own provide labels, and these are the frames with the highest reliability, so the fitted relation between reliability and temperature is extrapolated to the frames that need help. This is an interpretation and was not tested separately.

Table 6. Copies saved by the decision-directed calibrators relative to no calibration, classical receiver (evaluation sessions; Holm correction over six comparisons).

Calibrator	Regime	Δ copies	95% t-interval	Wilcoxon p	Holm p	Sessions won	Label
dd	T	+0.000	[+0.000, +0.000]	1.000	1.000	0/10	n.d.
dd	D	+0.000	[+0.000, +0.000]	1.000	1.000	0/10	n.d.
dd	S	+0.000	[+0.000, +0.000]	1.000	1.000	0/10	n.d.
dd	N	+0.000	[+0.000, +0.000]	1.000	1.000	0/10	n.d.
dd	DSN	+0.000	[+0.000, +0.000]	1.000	1.000	0/10	n.d.
ddreg	T	+0.001	[-0.003, +0.005]	1.000	1.000	2/10	n.d.
ddreg	D	+0.015	[-0.006, +0.036]	0.195	1.000	6/10	n.d.
ddreg	S	-0.006	[-0.023, +0.011]	0.578	1.000	3/10	n.d.
ddreg	N	+0.002	[-0.007, +0.011]	0.574	1.000	5/10	n.d.
ddreg	DSN	-0.011	[-0.034, +0.012]	0.289	1.000	2/10	n.d.

5.4 Training coverage versus calibration

Table 7. Receivers compared without calibration, and the pad-calibrated classical receiver against the network trained on four regimes (evaluation sessions; copies saved by the second receiver relative to the first; Holm correction over the 15 comparisons).

Receiver pair	Regime	Δ copies	95% t-interval	Wilcoxon p	Holm p	Sessions won	Label
NN-T	T	+0.014	[-0.007, +0.035]	0.152	0.914	6/10	n.d.
NN-T	D	+0.076	[+0.051, +0.101]	0.002	0.029	10/10	established
NN-T	S	+0.109	[+0.087, +0.131]	0.002	0.029	10/10	established
NN-T	N	+0.007	[-0.012, +0.026]	0.484	1.000	5/10	n.d.
NN-T	DSN	+0.186	[+0.144, +0.228]	0.002	0.029	10/10	established
NN-mix	T	+0.008	[-0.012, +0.028]	0.422	1.000	5/10	n.d.
NN-mix	D	+0.087	[+0.058, +0.116]	0.002	0.029	10/10	established
NN-mix	S	+0.144	[+0.105, +0.183]	0.002	0.029	10/10	established
NN-mix	N	+0.006	[-0.013, +0.025]	0.686	1.000	6/10	n.d.
NN-mix	DSN	+0.249	[+0.195, +0.303]	0.002	0.029	10/10	established
Classical+pad vs NN-mix	T	+0.009	[-0.011, +0.029]	0.320	1.000	6/10	n.d.
Classical+pad vs NN-mix	D	+0.048	[+0.027, +0.069]	0.008	0.055	8/10	suggested
Classical+pad vs NN-mix	S	+0.127	[+0.093, +0.161]	0.002	0.029	10/10	established
Classical+pad vs NN-mix	N	+0.007	[-0.014, +0.028]	0.674	1.000	6/10	n.d.
Classical+pad vs NN-mix	DSN	+0.149	[+0.090, +0.208]	0.002	0.029	10/10	established

The networks need fewer copies than the classical receiver in D, S and DSN, and the same number in T and N (Table 7). The single-regime network saves +0.186 copies in DSN and the four-regime network +0.249, against +0.100 for calibrating the classical receiver. The four-regime network still beats the pad-calibrated classical receiver in S and DSN (+0.127 and +0.149 copies saved; the first row of the pair is the calibrated classical receiver). Broad training is therefore the larger lever in our setting, and calibration of the classical receiver recovers part of the gap at no training cost. Since DSN was not in the training

mixture of any network, the gain of the four-regime network is generalisation to a combination of seen shifts rather than memorisation of the test regime; whether it would transfer to shifts that differ in kind is untested.

5.5 Confirmatory test

Table 8. Primary test on fresh confirmatory sessions: copies saved per delivery by the pad calibrator for the classical receiver, in each shifted regime (Holm correction over the four).

Receiver	Regime	Δ copies	95% t-interval	Wilcoxon p	Holm p	Sessions won	Label
Classical	D	+0.047	[+0.029, +0.065]	0.004	0.012	9/10	established
Classical	S	+0.032	[+0.013, +0.051]	0.008	0.016	8/10	established
Classical	N	-0.003	[-0.006, +0.000]	0.250	0.250	0/10	n.d.
Classical	DSN	+0.110	[+0.083, +0.137]	0.002	0.008	10/10	established

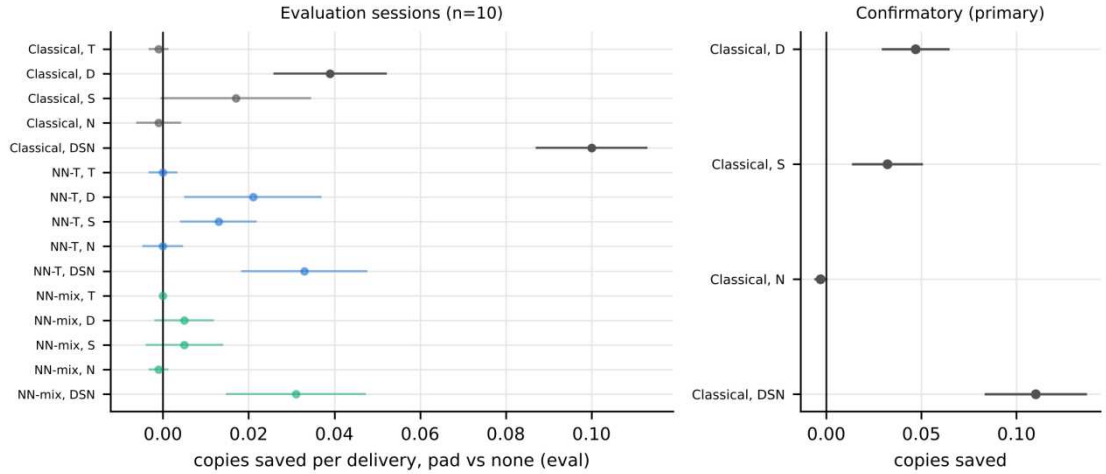


Figure 6. *Copies saved per delivery by the pad calibrator, with 95% intervals. Left: all receiver-regime pairs on the evaluation sessions (full colour: established after Holm correction). Right: the prespecified primary test on confirmatory sessions.*

The primary hypothesis is confirmed in three of four regimes: the pad calibrator saves +0.047 copies in D ([+0.029, +0.065], Holm $p = 0.012$), +0.032 in S ([+0.013, +0.051], 0.016) and +0.110 in DSN ([+0.083, +0.137], 0.008), and it does not help in N (-0.003, [-0.006, +0.000]; the interval is slightly below zero, that is, no gain). In D and S the evaluation sessions gave +0.039 and +0.017, so the effects replicate in sign and size; in S the evaluation interval included zero and the confirmatory interval does not. All effects are small in absolute terms: DSN is the largest at about 4.1% of the copies.

5.6 Risk-controlled blind copies

Table 9. Deadline-miss rate of the conformal controller ($R_{\max} = 2$ rounds), calibrated on the other nine evaluation sessions of the same regime ($n = 900$ trials; 100 test trials per session; mean over ten test sessions). The target is the column level.

Receiver + calibrator	Regime	$\alpha = 0.02$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
Classical	T	0.016	0.019	0.019	0.019
Classical	D	0.012	0.038	0.049	0.172
Classical	S	0.010	0.032	0.032	0.132
Classical	N	0.004	0.045	0.047	0.047
Classical	DSN	0.030	0.030	0.061	0.184
Classical+pad	T	0.010	0.020	0.020	0.020
Classical+pad	D	0.004	0.030	0.030	0.163

Receiver + calibrator	Regime	$\alpha = 0.02$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
Classical+pad	S	0.006	0.030	0.030	0.121
Classical+pad	N	0.003	0.038	0.049	0.049
Classical+pad	DSN	0.015	0.040	0.040	0.141
NN-mix	T	0.017	0.019	0.019	0.019
NN-mix	D	0.015	0.020	0.020	0.133
NN-mix	S	0.011	0.011	0.075	0.075
NN-mix	N	0.001	0.046	0.046	0.046
NN-mix	DSN	0.009	0.026	0.092	0.097

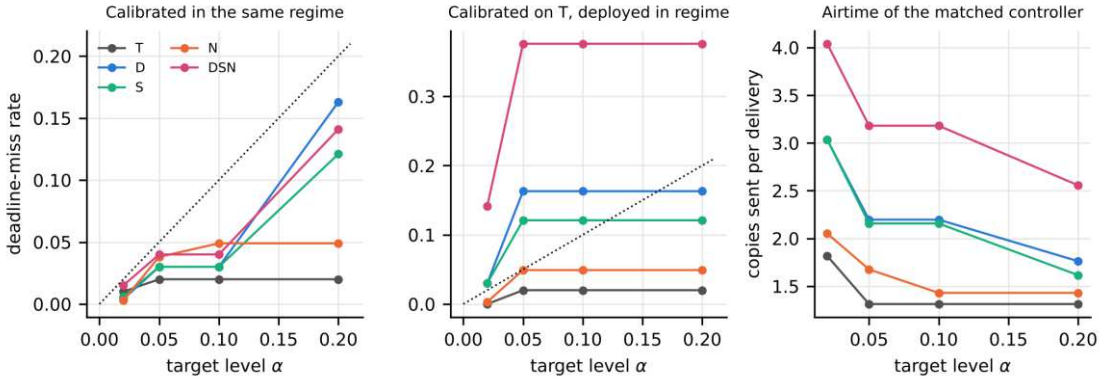


Figure 7. *Left: miss rate of the controller calibrated in the same regime against the target level (dotted line: equality), classical receiver with pad calibration. Middle: the same when calibrated on regime T and deployed in the regime. Right: copies sent per delivery of the matched controller.*

A guarantee is also met by always sending $K_{\max} = 4$ copies, so the airtime has to be compared. At $\alpha = 0.10$ the calibrated rule sends 3.18 copies per delivery in DSN for the pad-calibrated classical receiver, against 3.11 for the smallest constant k that meets the target on the test session itself (known only in hindsight) and 4.04 for $K_{\max}$; in D and T it sends the same number as the hindsight rule (2.20 and 1.31). The rule is therefore not wasteful in our regimes; it also cannot do better than a constant k tuned in hindsight, which no sender knows.

Channel memory. When the copies share the channel (Section 5.8) the target may be infeasible. With $K_{\max} = 4$ and $\alpha = 0.10$, the rule finds a feasible k in D for $\rho = 0.5$ (100% of the calibration sets) but not in DSN (0%), and in neither D nor DSN for $\rho \geq 0.9$; it then returns $K_{\max}$ and the miss rate is 0.177 in D and 0.296 in DSN at $\rho = 0.9$ (0.428 in DSN at $\rho = 0.99$). This is the case "set empty" of Proposition 2, where no guarantee is given, and the infeasibility is visible to the sender, which could then lengthen the deadline or send more than four copies.

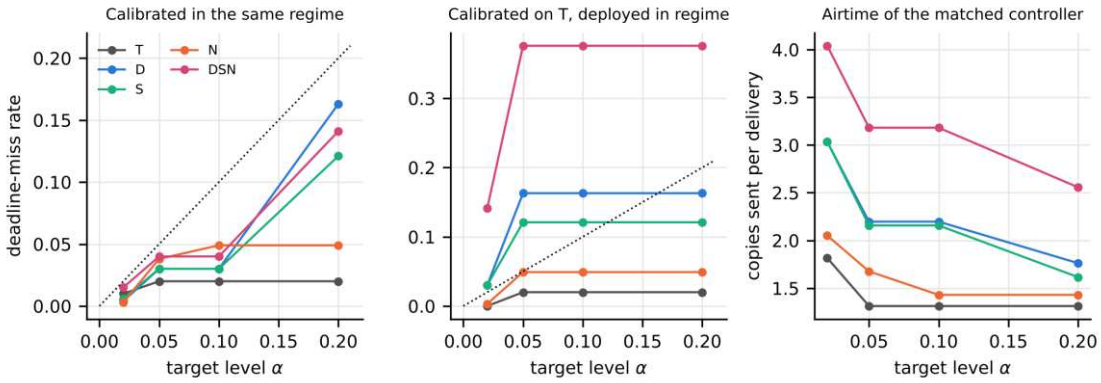


Figure 7. *Left: miss rate of the controller calibrated in the same regime against the target level (dotted line: equality), classical receiver with pad calibration. Middle: the same when calibrated on regime T and deployed in the regime. Right: copies sent per delivery of the matched controller.*

When the calibration sessions come from the same regime the guarantee of Proposition 2 holds in practice: the measured miss rate exceeds the target in 2 of 200 cells (five configurations, five regimes, four levels, evaluation and confirmatory sets), by at most 0.010, which is of the order of the sampling error of 1000 test trials per cell (standard error 0.004 to 0.009 at these levels; Tables 9 and 18). With a calibration set of a single session ($n = 100$) the same holds in all but 2 of 200 cells, by at most 0.010. The levels are often not met tightly, because the loss takes few values: k is an integer and the miss rate jumps between levels. When the controller is calibrated in regime T and deployed in a shifted regime, the guarantee fails, as it must: the miss rate exceeds the target in 94 of 120 cells of D, S and DSN, by up to 0.377, for example 0.397 against 0.10 in DSN (Figure 7, middle). Meeting the target costs airtime that grows with the severity of the regime (Figure 7, right): at $\alpha = 0.05$ the matched controller sends about 3.2 copies per delivery in DSN and 1.3 in T.

Table 10. Regime-switch study (exploratory): 150 trials in T followed by 300 trials in the shifted regime; classical receiver with pad calibration; $\alpha = 0.10$. "Static" is calibrated on T; "rolling window" uses the last 100 observed counts; ACI adds the level update with step γ; "matched oracle" is calibrated on the shifted regime. Mean over ten sequences.

Shifted regime	Controller	Miss, first 50 trials	Miss, later trials	Miss, all	Copies per delivery
D	static (T)	0.140	0.168	0.163	1.76
D	rolling	0.132	0.040	0.055	2.11
D	ACI 0.02	0.140	0.127	0.129	1.88
D	ACI 0.05	0.140	0.135	0.136	1.86
D	oracle	0.028	0.030	0.030	2.20
S	static (T)	0.106	0.124	0.121	1.62
S	rolling	0.106	0.057	0.065	1.93
S	ACI 0.02	0.106	0.122	0.119	1.63
S	ACI 0.05	0.106	0.124	0.121	1.62
S	oracle	0.024	0.031	0.030	2.16
N	static (T)	0.046	0.050	0.049	1.43
N	rolling	0.046	0.050	0.049	1.43
N	ACI 0.02	0.046	0.050	0.049	1.43
N	ACI 0.05	0.046	0.050	0.049	1.43
N	oracle	0.046	0.050	0.049	1.43
DSN	static (T)	0.380	0.375	0.376	2.31
DSN	rolling	0.258	0.050	0.085	2.98
DSN	ACI 0.02	0.310	0.095	0.131	2.74
DSN	ACI 0.05	0.324	0.100	0.137	2.76

Shifted regime	Controller	Miss, first 50 trials	Miss, later trials	Miss, all	Copies per delivery
DSN	oracle	0.044	0.039	0.040	3.18

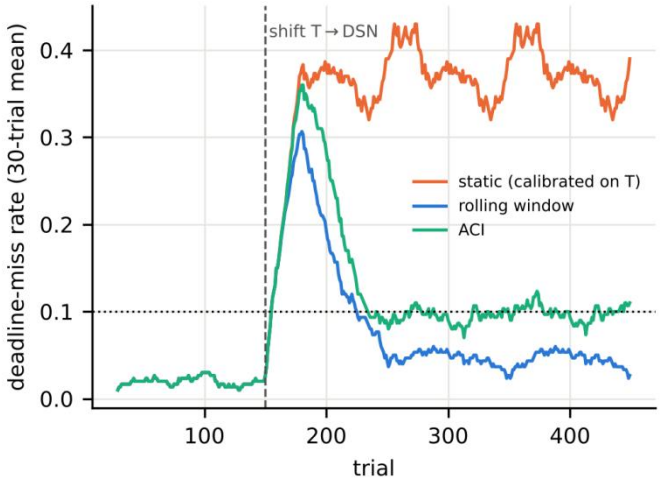


Figure 9. *Deadline-miss rate (30-trial moving average, mean over ten sequences) for a switch from T to DSN at trial 150 with* $\alpha = 0.10$.

The regime-switch study (Table 10, Figure 9) shows what the controllers do after a shift. The static controller stays at a miss rate of 0.376 in DSN. The rolling window needs roughly 50 to 80 trials to recover (Figure 9); its miss rate is 0.258 in the first 50 trials after the shift and 0.050 later, below the target, at 2.98 copies per delivery. ACI with $\gamma = 0.05$ recovers more slowly (0.324 in the first 50 trials) but then tracks the target (0.100) at 2.76 copies: it spends less airtime than the rolling window and is closer to the target, but it does not meet the target in the transient. In D and S the rolling window again meets the target after the transient (0.040 and 0.057) while ACI stays above it (0.135 and 0.124). Under impulsive noise the regime does not change the required k and all controllers behave alike. We make no claim that ACI is the better choice: its theoretical guarantee does not apply with a finite $K_{\max}$ (Section 4.2), and in D and S it was worse than the rolling window in our sequences.

5.7 Goodput under long round-trip times

Table 11. Goodput (bit/s) at RTT $= 2$ s and 8 s (exploratory; $\alpha = 0.10$; the risk-controlled $\hat{k}$ is calibrated on the other nine evaluation sessions of the regime) for fixed numbers of first-round copies, the risk-controlled $\hat{k}$ and the best constant k chosen in hindsight.

Regime	Receiver + calibrator	RTT (s)	Fixed $k = 1$	Fixed $k = 2$	Fixed $k = 3$	Risk-controlled $\hat{k}$	Best fixed k (hindsight)
T	Classical	2	166	189	173	166	189
T	Classical	8	46	57	56	46	57
T	Classical+pad	2	165	189	173	165	189
T	Classical+pad	8	46	57	56	46	57
T	NN-mix	2	166	189	173	166	189
T	NN-mix	8	46	57	56	46	57
D	Classical	2	120	159	164	159	164
D	Classical	8	33	47	53	47	53
D	Classical+pad	2	123	164	168	164	168
D	Classical+pad	8	34	49	54	49	54
D	NN-mix	2	126	169	170	169	171
D	NN-mix	8	35	50	55	50	55
DSN	Classical	2	88	119	141	141	146
DSN	Classical	8	24	35	44	44	50

Regime	Receiver + calibrator	RTT (s)	Fixed $k = 1$	Fixed $k = 2$	Fixed $k = 3$	Risk-controlled $\hat{k}$	Best fixed k (hindsight)
DSN	Classical+pad	2	93	127	149	149	151
DSN	Classical+pad	8	26	37	47	47	52
DSN	NN-mix	2	99	137	156	139	157
DSN	NN-mix	8	27	40	50	41	52

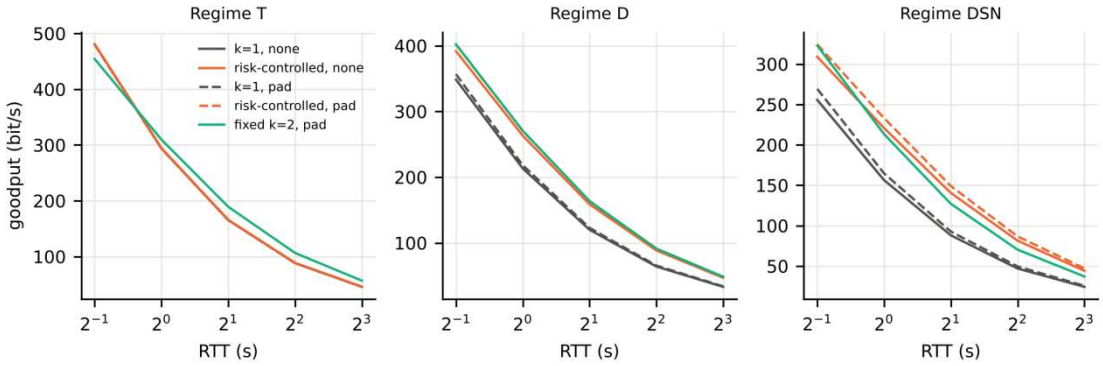


Figure 10. *Goodput against round-trip time in regimes T, D and DSN for stop-and-wait ($k = 1$), the risk-controlled number of blind copies and a fixed $k = 2$ (green; pad-calibrated classical receiver), classical receiver without and with pad calibration.*

Sending redundancy ahead of the acknowledgement raises goodput over stop-and-wait: at RTT $= 2$ s the risk-controlled copies take the classical receiver from 120 to 159 bit/s in D (32%) and from 88 to 141 bit/s in DSN (60%). This comparison is against the weakest policy, and the stronger comparison is a constant k chosen once. No constant is best everywhere: $k = 2$ is the best of the constants in T (189 bit/s) and nearly so in D, $k = 3$ is best in DSN (141). Against $k = 2$ the controller gives the same goodput in D, +21 bit/s (95% interval [+20, +23]) in DSN and -24 bit/s (95% interval [-29, -19]) in T, where it keeps $k = 1$ because the deadline target is already met (166 against 189): a deadline guarantee is not the goodput optimum, and a sender that wants goodput should set the target accordingly. In D and DSN the controller is within about 4% of the best constant in hindsight (141 against 146 in DSN). Its value is therefore the deadline guarantee together with automatic adaptation to the regime, not a goodput gain over a well-chosen constant. Pad calibration adds a further 149 against 141 bit/s in DSN (about 6%), and the four-regime network gives 139.

5.8 Where the gain comes from and what it depends on

The main results leave four questions open. Where in a trial do the saved copies come from? Does the gain survive when the copies share the channel? Does a different free per-frame estimate give the same gain? And does it depend on details of the frame design (where the pad sits, whether the code is interleaved) or of the decoder (quantisation)? This section answers them with additional runs on the evaluation seeds (same sessions, same calibrator settings, nothing re-tuned). They were designed after the main results had been reviewed and are exploratory; Holm correction is applied within each variant.

Where the saved copies come from. Table 12 splits the trials of the classical receiver by what they needed without calibration. In DSN, pad calibration saves 0.221 copies per trial in the 37% of trials that need three to five copies, and 0.608 per trial in the 3% that would not be delivered within five copies, while trials that need exactly two copies lose 0.015 per trial. The gain therefore comes from the hard trials, in which several copies are combined; in T only 2% of the trials need three or more copies and there is nothing to gain (Figure 11, right).

Table 12. Pad calibration, classical receiver: copies saved by the type of trial (type defined by the copies needed without calibration; evaluation seeds; positive is better).

Regime	Trial type (uncalibrated)	Share of trials	Copies saved per trial of this type	Contribution to total
T	one copy suffices	70.7%	+0.000	+0.000
T	two copies	27.4%	-0.005	-0.001
T	three to five copies	1.9%	+0.000	+0.000
T	not delivered	0.0%	+0.000	+0.000
D	one copy suffices	43.5%	+0.000	+0.000
D	two copies	39.3%	-0.020	-0.008
D	three to five copies	16.7%	+0.228	+0.037
D	not delivered	0.5%	+0.800	+0.010
DSN	one copy suffices	25.0%	+0.000	+0.000
DSN	two copies	35.3%	-0.015	-0.005
DSN	three to five copies	36.7%	+0.221	+0.081
DSN	not delivered	3.0%	+0.608	+0.024

Channel memory across copies. The main study redraws the channel for every attempt, which is the assumption that favours a per-copy calibrator most. We replaced it by path gains that follow a first-order autoregressive process across the five attempts of a trial with correlation ρ, keeping the delays, arrival angles and Doppler shifts fixed within the trial. Acoustic channels can stay correlated for seconds [2], so for blind copies sent 0.288 s apart a correlation near the upper end of the range is plausible. Table 13 and Figure 11 (left) show that the gain of the pad calibrator is not robust to this. In DSN it falls from +0.100 (independent) to +0.037 at $\rho = 0.5$, +0.005 at $\rho = 0.9$ and +0.001 at $\rho = 0.99$, and in D from +0.039 to +0.035, +0.008 and -0.005; at $\rho \geq 0.9$ it is not distinguishable from zero. The oracle shows the same decay (DSN: +0.165, +0.086, +0.033, +0.011), so the cause is not a failure of the pad estimate: when the copies see the same channel their overconfidence is nearly equal, the factor η of Proposition 1 is close to one and there is nothing left to correct. The network trained on regime T behaves in the same way (gain in DSN +0.033, +0.012, +0.007, +0.000). Channel memory also raises the copies needed without calibration, from 2.422 to 3.689 in DSN, because combining loses its diversity gain. The per-frame calibrator is therefore a tool for links whose retransmissions see (partly) independent fading, for example when the copies are separated in time by more than the coherence time or sent on different frequencies; it is not a general improvement of acoustic HARQ.

Table 13. Pad calibration for the classical receiver when the path gains of the copies of a trial are correlated (evaluation seeds; "independent" is the main study).

Regime	Correlation ρ between copies	none	pad	oracle	Δ copies (pad)	95% interval	Label
T	independent	1.312	1.313	1.312	-0.001	[-0.003,+0.001]	n.d.
T	0.5	1.345	1.345	1.344	+0.000	[+0.000,+0.000]	n.d.
T	0.9	1.410	1.411	1.410	-0.001	[-0.003,+0.001]	n.d.
T	0.99	1.559	1.559	1.561	-0.000	[-0.003,+0.003]	n.d.
D	independent	1.803	1.764	1.719	+0.039	[+0.026,+0.052]	established
D	0.5	2.151	2.116	2.085	+0.035	[+0.012,+0.058]	suggested
D	0.9	2.533	2.525	2.524	+0.008	[-0.007,+0.023]	n.d.
D	0.99	2.817	2.822	2.811	-0.005	[-0.021,+0.011]	n.d.
DSN	independent	2.422	2.322	2.257	+0.100	[+0.087,+0.113]	established
DSN	0.5	2.791	2.754	2.705	+0.037	[+0.018,+0.056]	suggested
DSN	0.9	3.331	3.326	3.298	+0.005	[-0.009,+0.019]	n.d.
DSN	0.99	3.689	3.688	3.678	+0.001	[-0.013,+0.015]	n.d.

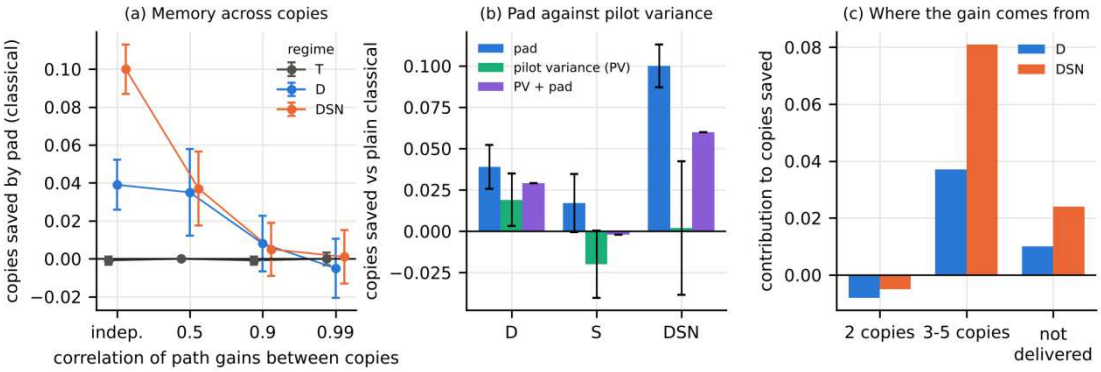


Figure 11. *Left: copies saved by the pad calibrator (classical receiver, 95% intervals) against the correlation of the path gains between copies. Middle: copies saved relative to the plain classical receiver by the pad calibrator, by a demapper scaled with a variance measured on the pilots (PV) and by both (no interval for the combination). Right: contribution of each type of trial to the copies saved by pad calibration.*

A free estimate from the pilots. The pad bits cost nothing only in this frame design; the obvious alternative is a per-frame estimate that needs no known bits. We built one: the classical demapper with a per-frame estimate of the channel-estimation error variance $\hat{v}$, taken from the second difference of the least-squares estimates at the pilots (white noise of variance σ^2 gives a second difference of variance $6\sigma^2$), with the matched-filter output scaled by $g/(g(1+\hat{v})+\hat{v}M)$, which reduces to the error-aware demapper of Appendix B for $\hat{v}=1$. Table 14 and Figure 11 (middle) compare it with the plain demapper and with pad calibration. This pilot-variance (PV) demapper does not reproduce the gain: relative to the plain demapper it saves -0.019 ([-0.037, -0.001], suggested) copies in T, +0.019 ([+0.003, +0.035], suggested) in D and +0.002 ([-0.038, +0.042], n.d.) in DSN (negative: it needs more), and the pad calibrator still saves +0.058 ([+0.037, +0.079], established) on top of it in DSN. With channel memory ($\rho \geq 0.9$) it needs 0.03 to 0.05 more copies than the plain demapper. We read this as a negative result for one simple estimator, not as proof that no pilot-based estimate can do the job: the roughness of the pilot estimates mixes noise and channel curvature, and a better estimator may exist.

Table 14. Classical receiver (CL) and the pilot-variance demapper (PV): copies to deliver without and with pad calibration, and paired differences (positive: first-named needs fewer copies; est. = established, sugg. = suggested, n.d. = not distinguishable).

Regime	CL none	CL pad	PV none	PV pad	PV vs CL none	PV vs CL pad	pad on top of PV
T	1.312	1.313	1.331	1.329	-0.019 (sugg.)	-0.018 (sugg.)	+0.002 (n.d.)
D	1.803	1.764	1.784	1.774	+0.019 (sugg.)	-0.020 (sugg.)	+0.010 (sugg.)
S	1.635	1.618	1.655	1.637	-0.020 (sugg.)	-0.037 (est.)	+0.018 (sugg.)
N	1.429	1.430	1.451	1.449	-0.022 (sugg.)	-0.021 (sugg.)	+0.002 (n.d.)
DSN	2.422	2.322	2.420	2.362	+0.002 (n.d.)	-0.098 (est.)	+0.058 (est.)

Pad position and interleaving. The 116 pad bits of the main study are the unused tail of the last OFDM symbol, and the code is not interleaved. Table 15 and Figure 12 (left) move the pad to the head of the frame and spread it evenly through the frame, and add a fixed random bit interleaver over the coded bits. The position of the pad does not matter: the gain in DSN is +0.100 (tail), +0.093 (head) and +0.100 (spread), and in D +0.039, +0.046 and +0.042. Interleaving makes the link much better (copies to deliver in DSN fall from 2.422 to 1.974, in T from 1.312 to 1.135) and reduces the gain of the pad calibrator by about a third in DSN and by half in D, to +0.068 in DSN and +0.020 in D, both still established, and still zero in T. The absence of an interleaver therefore does not create the gain, but it enlarges it.

Table 15. Pad calibration for the classical receiver with the pad at the tail (main study), at the head, spread through the frame, and with a bit interleaver (evaluation seeds).

Pad position / design	Regime	none	pad	oracle	Δ copies (pad)	95% interval	Label
tail (main study)	D	1.803	1.764	1.719	+0.039	[+0.026,+0.052]	established
tail (main study)	DSN	2.422	2.322	2.257	+0.100	[+0.087,+0.113]	established
head	D	1.820	1.774	1.734	+0.046	[+0.027,+0.065]	established
head	DSN	2.393	2.300	2.223	+0.093	[+0.074,+0.112]	established
spread	D	1.792	1.750	1.697	+0.042	[+0.023,+0.061]	established
spread	DSN	2.385	2.285	2.215	+0.100	[+0.078,+0.122]	established
tail + bit interleaver	T	1.135	1.135	1.135	+0.000	[+0.000,+0.000]	n.d.
tail + bit interleaver	D	1.424	1.404	1.393	+0.020	[+0.010,+0.030]	established
tail + bit interleaver	DSN	1.974	1.906	1.862	+0.068	[+0.053,+0.083]	established

Quantised soft inputs. Lemma 1 assumes real-valued inputs. Table 16 and Figure 12 (right) decode with the LLR of each copy divided by a global scale, quantised to a signed integer with a fixed step (4 bits: step 2, levels $-7..7$; 3 bits: step 4, levels $-3..3$) and clipped, and the integers added. With unquantised input every scale gives 2.422 copies in DSN, as Lemma 1 says. With 3-bit input the copies range from 1.342 to 2.549 in T and from 2.461 to 3.387 in DSN, depending on the scale, and with 4-bit input from 1.324 to 1.714 in T. The statement that a global temperature cannot help is therefore a property of unquantised decoding, and in a fixed-point receiver the scale has to be set. The per-frame pad temperature is not a general remedy: with 3-bit input it needs more copies than the scale 1 in T (1.434 against 1.342) and D (1.900 against 1.815), because dividing by a temperature above one moves the values into the coarse part of the quantiser, and it is better only in DSN (2.428 against 2.461; at 4 bits 2.350 against 2.429). A calibrator for a fixed-point decoder has to be designed together with the quantiser step; we have not done this.

Table 16. Copies to deliver for the classical receiver in regimes T, D and DSN with a global scale, the per-frame pad temperature and the oracle, for unquantised, 4-bit and 3-bit soft input (evaluation seeds).

Regime	Soft input	$\hat{\kappa} = 0.5$	1	2.2	4	8	pad (per frame)	oracle (per frame)
T	unquantised	1.312	1.312	1.312	1.312	1.312	1.313	1.312
T	4 bit	1.335	1.324	1.354	1.417	1.714	1.356	1.345
T	3 bit	1.351	1.342	1.438	1.713	2.549	1.434	1.432
D	unquantised	1.803	1.803	1.803	1.803	1.803	1.764	1.719
D	4 bit	1.817	1.791	1.812	1.900	2.194	1.793	1.783
D	3 bit	1.842	1.815	1.905	2.181	3.056	1.900	1.989
DSN	unquantised	2.422	2.422	2.422	2.422	2.422	2.322	2.257
DSN	4 bit	2.530	2.429	2.423	2.467	2.677	2.350	2.353
DSN	3 bit	2.573	2.461	2.468	2.660	3.387	2.428	2.588

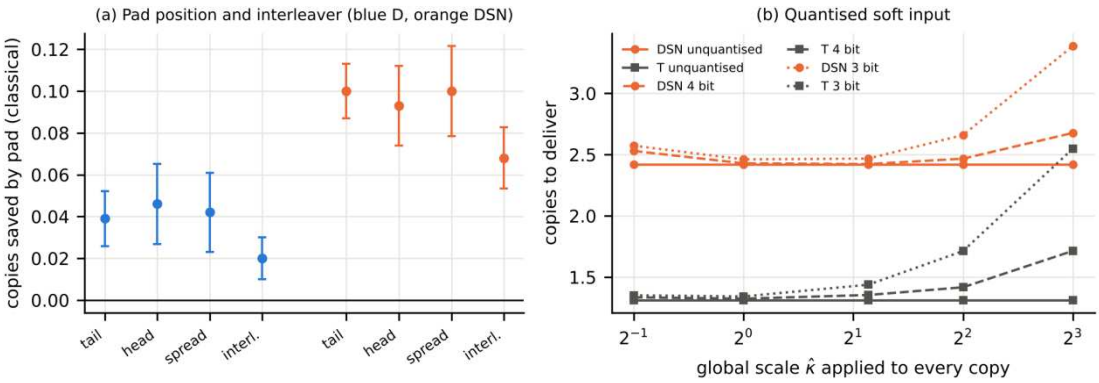


Figure 12. *Left: copies saved by the pad calibrator (classical receiver, 95% intervals) with the pad at the tail (main study), head or spread through the frame, and with a bit interleaver, in D (blue) and DSN (orange). Right: copies to deliver against a global scale applied to every copy, for unquantised, 4-bit and 3-bit soft input, in T (grey squares) and DSN (orange circles).*

Penalty for undelivered packets. The copies to deliver count an undelivered packet as six. With a penalty of 5, 6 or 10 the gain in DSN is +0.085, +0.100 and +0.160, and in D +0.036, +0.039 and +0.051: the conclusion does not depend on the penalty.

5.9 Checks of the theory

Scale invariance. Scaling the LLRs of 300 classical frames in DSN by 3.7 changed the decoded bits of 0 of them, as Lemma 1 requires.

Combining efficiency. Table 17 compares the combining loss $-10\log_{10}\eta$ predicted by equation (2) from the measured per-frame oracle temperatures and calibrated mean LLRs of the first two copies of each trial with the loss measured on the combined LLRs (the ratio of the deflection of the signed combined LLR with and without the oracle temperature). The prediction ranks the regimes in the right order for the classical receiver (Pearson correlation 0.66 over the ten cells, Spearman 0.64), but it overstates the loss where it is small: 0.30 dB predicted against 0.03 dB measured in T and 0.29 against 0.01 in N, and in D and S by a factor of about two. The reason is that equation (2) assumes that a copy has one reliability μ_i, while the reliability varies strongly from bit to bit within a frame. We therefore read (2) as a qualitative explanation and as an upper estimate, not as a predictor.

Table 17. Predicted and measured combining loss of uncalibrated Chase combining of two copies (dB; 200 trials per regime from fresh seeds).

Receiver	Regime	Predicted loss (dB)	Measured loss (dB)
Classical	T	0.30	0.03
Classical	D	0.58	0.26
Classical	S	0.43	0.26
Classical	N	0.29	0.01
Classical	DSN	0.51	0.40
NN-T	T	0.16	0.01
NN-T	D	0.29	0.23
NN-T	S	0.21	0.18
NN-T	N	0.15	0.17
NN-T	DSN	0.41	0.40

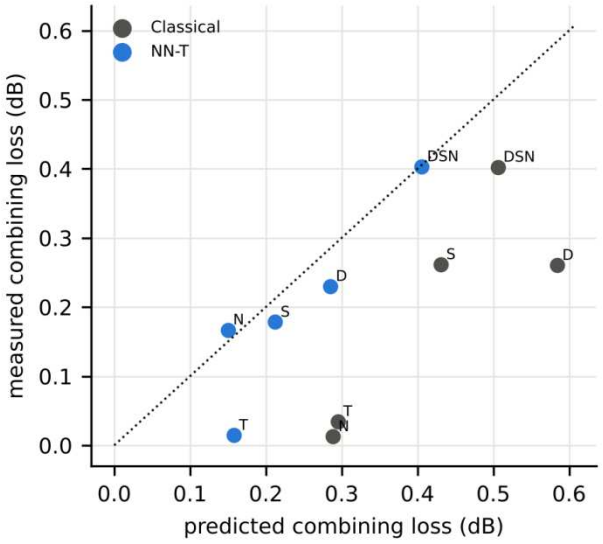


Figure 13. *Predicted (equation (2)) against measured combining loss; dotted line: equality.*

6. Discussion and Limitations

What the evidence supports. A max-log Viterbi decoder on real-valued inputs is insensitive to how a receiver scales its LLRs, except through the *variation* of that scale across the combined copies. When the copies see independent fading, a per-frame temperature fitted on known pad bits lets the classical receiver deliver with a few percent fewer copies in the regimes where combining is often needed (4% in DSN), a result confirmed on fresh sessions and not dependent on where the pad sits or on whether the code is interleaved. The larger gains in this study come from elsewhere: training the network on several regimes, and sending blind copies instead of waiting. The conformal rule gives what it promises when its calibration data come from the current regime, flags itself as infeasible when the deadline cannot be met with the allowed copies, and gives nothing when its data come from another regime.

What the evidence does not support. The gain of the calibrator disappears when the copies are strongly correlated (Section 5.8), which is the likely case for back-to-back blind copies in a real channel; we do not claim that calibration improves acoustic HARQ in general, only that it does when retransmissions see sufficiently independent channels. We do not show that frame-level calibration helps under impulsive noise; it does not, and even the oracle gains little. We do not show that a free pilot-based estimate can replace the known bits (one simple estimator failed to), nor that the pad is the best use of the spare bits. We do not show that ACI is preferable to a rolling window, nor that the conformal rule improves goodput over a well-chosen constant number of copies; its value is the deadline guarantee and the adaptation to the regime.

Limitations.

- *Simulation only.* The channel is a statistical model with five paths and a fixed Doppler and delay structure; no sea trial or recorded channel was used. Channel memory was tested by an autoregressive model of the path gains, not by recorded channels, and the delay spreads (at most 7 ms) stay inside the 16 ms cyclic prefix, so there is no inter-symbol interference.
- *Small learned receivers.* The networks are small multilayer perceptrons on hand-built features (including a supplied SNR estimate) trained with scikit-learn for a fixed budget, not the convolutional receivers of the

literature [3, 4]. Their overconfidence under shift may differ from that of larger models, and their calibration was not tuned.

- *A simple classical baseline.* The classical receiver uses linear interpolation and a Gaussian demapper that ignores estimation error. An error-aware variant (Appendix B) and a pilot-variance variant (Appendix B) did not change the conclusions, but we did not try interpolation across time, MMSE channel estimation or iterative receivers, which would be better calibrated.
- *Perfect CRC and idealised feedback.* The CRC is an exact error detector; feedback is error-free and arrives after exactly one round trip; the number of copies needed is observed without error.
- *The code and the decoder.* A single rate-1/2 convolutional code with a max-log decoder. Lemma 1 holds for real-valued inputs; with 3-bit or 4-bit soft inputs the scale matters (Section 5.8), and sum-product decoders with clipping or offsets also respond to the level of the LLRs. The calibrator was not designed for a quantised decoder.
- *The pad is a design choice.* The bits are free because 1036 coded bits do not fill six OFDM symbols; a different code rate or frame length would remove the free bits, and the calibrator would then cost airtime.
- *Statistical design.* Ten sessions of 100 trials per regime give wide intervals; the Wilcoxon test cannot give a p-value below 0.002. The primary hypothesis was fixed after the pilot tuning and before the evaluation and confirmatory runs, but not registered anywhere. The risk-control, goodput, theory-check and all robustness analyses (Sections 5.7 to 5.9, Appendix B) were designed after the evaluation sessions had been generated and are exploratory; the confirmatory set was used only for the primary test and for the repeated risk tables.
- *Post-hoc choices.* Listed in Section 4.3.
- *Dependence between regimes.* The regime DSN is a combination of shifts seen separately by the four-regime network; it is "unseen" in the sense of never occurring in training, not in the sense of being unrelated.

7. Conclusions

For soft-combined HARQ over an acoustic link with an unquantised max-log decoder, only the relative confidence of the copies matters, so a global or session-level temperature cannot help and calibration has to act per frame. A temperature fitted on the known pad bits of each frame lets the classical receiver deliver with 4.1% fewer copies in the combined shift DSN and 2.2% fewer under high Doppler, confirmed on fresh sessions, when the copies fade independently; the gain shrinks with the correlation of the copies and is gone at a correlation of 0.9, it is unchanged by the position of the pad and reduced by a third (DSN) to a half (D) when the code is interleaved, and it is not reproduced by the simple pilot-variance estimator we tried. Training a network on several regimes is the larger lever (+0.249 copies in DSN). Choosing the number of blind first-round copies by conformal risk control keeps the deadline-miss rate at the target when the calibration comes from the current regime, signals infeasibility when the allowed copies cannot meet it, and raises goodput at a two-second round trip by 60% over stop-and-wait in DSN, though not over a well-chosen constant number of copies; after a regime change a rolling window restores the target after roughly 50 to 80 trials. Future work should test recorded channels with memory across copies, calibrators designed for quantised decoders, better free estimates of frame reliability, symbol-level calibration for impulsive noise, larger learned receivers, incremental redundancy instead of Chase combining, and a regime-change detector in front of the risk controller.

Declarations

Conflict of interest: The authors declare that they have no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability: No experimental datasets were generated or analysed in this study.

Code availability: Not applicable.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Acknowledgement: None.

Appendix A. Reproducibility

The study runs in numpy and scikit-learn on two CPU cores in roughly 3 hours. Run in order: the network training (train_nn.py), the global temperature (exp_cnr.py tau), the pilot tuning (tune_cnr.py), the evaluation and confirmatory sets (exp_cnr.py eval, exp_cnr.py conf), the exploratory variants (exp_extra.py, exp_ea.py), the revision experiments (exp_rev.py for each of the variants base, rho50, rho90, rho99, head, spread, intl; exp_quant.py), the analysis (run_analysis.sh), the figures (figs_cnr.py) and the manuscript (build_paper.py). Seeds: training 1000s, pilot 2000s, validation 3000s and 3100s, evaluation 4000s, confirmatory 5000s, theory checks 6000s to 8000s; the revision experiments re-simulate the evaluation sessions, and the base variant reproduces the main-study numbers exactly. Results are cached per session in results/cache.

Appendix B. Further sensitivity analyses

These analyses support Section 5 and are exploratory unless stated otherwise.

Risk control on the confirmatory sessions. The miss rates of the conformal controller on the confirmatory sessions are in Table 18 (same protocol as Table 9).

Table 18. As Table 9 on the confirmatory sessions.

Receiver + calibrator	Regime	$\alpha = 0.02$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.20$
Classical	T	0.015	0.015	0.015	0.015
Classical	D	0.015	0.015	0.055	0.177
Classical	S	0.004	0.029	0.029	0.122
Classical	N	0.004	0.004	0.064	0.064
Classical	DSN	0.027	0.027	0.058	0.157
Classical+pad	T	0.015	0.015	0.015	0.015
Classical+pad	D	0.003	0.041	0.041	0.161
Classical+pad	S	0.013	0.013	0.013	0.111
Classical+pad	N	0.006	0.006	0.065	0.065
Classical+pad	DSN	0.013	0.034	0.034	0.117
NN-mix	T	0.015	0.015	0.015	0.015
NN-mix	D	0.000	0.027	0.027	0.129
NN-mix	S	0.006	0.006	0.075	0.075
NN-mix	N	0.005	0.028	0.050	0.050
NN-mix	DSN	0.007	0.023	0.082	0.097

Number of known bits. The gain of the pad calibrator grows with the number of known bits (Table 19): in DSN the classical receiver saves 0.028 copies with 16 bits and 0.100 with all 116, against +0.165 for

the oracle; a short training field of 32 to 64 known bits recovers most of it. The estimate is limited by its variance, not by bias.

Table 19. Copies to deliver with the pad calibrator using the first 16, 32, 64 or all 116 known bits (exploratory; evaluation seeds; in brackets: copies saved relative to none).

Receiver	Regime	none	16 bits	32 bits	64 bits	116 bits (all)	oracle
Classical	D	1.803	1.790 (+0.013)	1.781 (+0.022)	1.768 (+0.035)	1.764 (+0.039)	1.719
Classical	DSN	2.422	2.394 (+0.028)	2.372 (+0.050)	2.346 (+0.076)	2.322 (+0.100)	2.257
NN-T	D	1.727	1.715 (+0.012)	1.711 (+0.016)	1.709 (+0.018)	1.706 (+0.021)	1.693
NN-T	DSN	2.236	2.228 (+0.008)	2.224 (+0.012)	2.209 (+0.027)	2.203 (+0.033)	2.174

Error in the SNR estimate. The classical receiver does not use an SNR estimate. Perturbing the SNR input of the network by a Gaussian error of 2 or 4 dB changes the copies to deliver of the network by at most a few thousandths and leaves the gain of the pad calibrator unchanged (Table 20).

Table 20. Effect of an error in the per-frame SNR given to the network and of the pad calibrator (exploratory; Holm correction within each error level). The classical receiver ignores the SNR estimate, so its rows reproduce the main results.

SNR error (σ)	Receiver	Regime	none	pad	oracle	Δ copies (pad)	Label
2 dB	Classical	T	1.312	1.313	1.312	-0.001	n.d.
2 dB	Classical	DSN	2.422	2.322	2.257	+0.100	established
2 dB	NN-T	T	1.293	1.295	1.295	-0.002	n.d.
2 dB	NN-T	DSN	2.231	2.196	2.170	+0.035	established
4 dB	Classical	T	1.312	1.313	1.312	-0.001	n.d.
4 dB	Classical	DSN	2.422	2.322	2.257	+0.100	established
4 dB	NN-T	T	1.294	1.295	1.294	-0.001	n.d.
4 dB	NN-T	DSN	2.234	2.196	2.171	+0.038	established

Pilot tuning. Table 21 and Figure 14 show the pilot results behind the choices $\lambda = 20$ and $W = 100$. The prior matters little (the pad calibrator changes the summed failure rate by 0.006 to 0.027 over the four prior strengths tried) and the decision-directed calibrators are indistinguishable from none.

Table 21. Pilot tuning (regimes D, S, N; four sessions each): change in the failures summed over $j = 1, \ldots, 5$ combined copies relative to no calibration (negative is better).

Receiver	Calibrator, parameter	Value	Change in summed failures
Classical	ddreg window (frames)	10	+0.0033
Classical	ddreg window (frames)	100	-0.0008
Classical	ddreg window (frames)	30	+0.0042
Classical	dd window (frames)	10	+0.0000
Classical	dd window (frames)	100	+0.0000
Classical	dd window (frames)	30	+0.0000
Classical	pad prior (pseudo-bits)	0.0	-0.0217
Classical	pad prior (pseudo-bits)	20.0	-0.0267
Classical	pad prior (pseudo-bits)	200.0	-0.0058
Classical	pad prior (pseudo-bits)	60.0	-0.0158
NN-T	ddreg window (frames)	10	+0.0008
NN-T	ddreg window (frames)	100	-0.0017
NN-T	ddreg window (frames)	30	+0.0008
NN-T	dd window (frames)	10	+0.0000
NN-T	dd window (frames)	100	+0.0000
NN-T	dd window (frames)	30	+0.0000
NN-T	pad prior (pseudo-bits)	0.0	-0.0083
NN-T	pad prior (pseudo-bits)	20.0	-0.0108
NN-T	pad prior (pseudo-bits)	200.0	-0.0050

Receiver	Calibrator, parameter	Value	Change in summed failures
NN-T	pad prior (pseudo-bits)	60.0	-0.0108

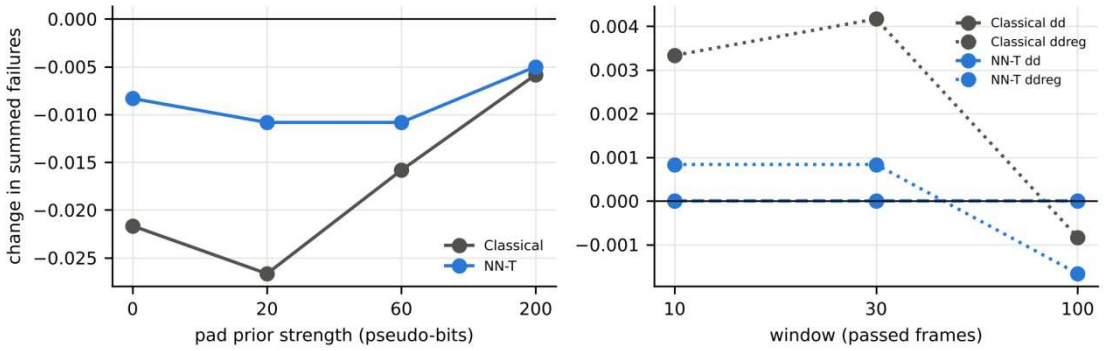


Figure 14. *Pilot tuning: change in summed failures against the strength of the prior of the pad calibrator (left) and against the window of the decision-directed calibrators (right).*

A stronger classical demapper. A reviewer may ask whether the gain of the pad calibrator is an artefact of a weak classical baseline. The unit-variance least-squares channel estimate makes the matched-filter output z noisier than the plain demapper assumes: its variance is about $2g + M$ rather than g, with $g = \sum_m |\hat{h}_m|^2$, which explains the median oracle temperature of 2 in the training regime. We therefore ran an error-aware demapper with LLR $= 2\sqrt{2}\, g\, \mathrm{Re}(z)/(2g + M)$ on the evaluation seeds (exploratory; Table 22). It is better calibrated in the training regime (median $\kappa = 0.79$) but still overconfident in DSN (2.23), it needs slightly more copies than the plain demapper (-0.045 copies saved in DSN, i.e. it is worse), and the gain of the pad calibrator stays as large as before: +0.042, +0.035 and +0.100 copies in D, S and DSN, and none in T and N. The finding does not depend on the particular weakness of the plain demapper; we have not tried other error-aware demappers.

Table 22. Error-aware classical demapper (exploratory; evaluation seeds): oracle temperature, copies to deliver and the gain of the pad calibrator; the last column compares the uncalibrated error-aware demapper with the plain classical one (positive: fewer copies).

Regime	Oracle κ	none	pad	oracle	Δ copies (pad), 95% interval	Label	Saved vs plain classical (none)
T	0.79	1.330	1.329	1.329	+0.001 [-0.003, +0.005]	n.d.	-0.018
D	1.30	1.835	1.793	1.748	+0.042 [+0.023, +0.061]	established	-0.032
S	1.20	1.666	1.631	1.611	+0.035 [+0.015, +0.055]	established	-0.031
N	1.10	1.444	1.444	1.438	+0.000 [-0.003, +0.003]	n.d.	-0.015
DSN	2.23	2.467	2.367	2.277	+0.100 [+0.077, +0.123]	established	-0.045

References

[1] A. Ghosh, J.-W. Lee, H.-S. Cho, Throughput and energy efficiency of a cooperative hybrid ARQ protocol for underwater acoustic sensor networks, Sensors 13(11) (2013) 15385–15408.

[2] M. Stojanovic, J. Preisig, Underwater acoustic communication channels: propagation models and statistical characterization, IEEE Commun. Mag. 47(1) (2009) 84–89.

[3] S. T. Hassan, P. Chen, Y. Rong, K. Y. Chan, Underwater acoustic orthogonal frequency-division multiplexing communication using deep neural network-based receiver: river trial results, Sensors 24(18) (2024) 5995.

[4] M. Honkala, D. Korpi, J. M. J. Huttunen, DeepRx: Fully convolutional deep learning receiver, IEEE Trans. Wireless Commun. 20(6) (2021) 3925–3940. arXiv:2005.01494.

[5] M. Chitre, J. Potter, S.-H. Ong, Optimal and near-optimal signal detection in snapping shrimp dominated ambient noise, IEEE J. Oceanic Eng. 31(2) (2006) 497–503.

[6] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proc. Int. Conf. Machine Learning (ICML), 2017. arXiv:1706.04599.

[7] A. Elnashar, When does a neural receiver help? Calibration-drift benchmarking and detect-and-rollback for 5G/6G NR, arXiv:2605.26157, 2026.

[8] K. M. Cohen, S. Park, O. Simeone, S. Shamai (Shitz), Calibrating AI models for wireless communications via conformal prediction, arXiv:2212.07775, 2022.

[9] K. M. Cohen, S. Park, O. Simeone, S. Shamai (Shitz), Calibrating AI models for few-shot demodulation via conformal prediction, in: IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[10] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster, Conformal risk control, in: Int. Conf. Learning Representations (ICLR), 2024. arXiv:2208.02814.

[11] I. Gibbs, E. Candès, Adaptive conformal inference under distribution shift, in: Advances in Neural Information Processing Systems (NeurIPS), 2021. arXiv:2106.00170.

[12] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, 2nd ed., Springer, Cham, 2022.

[13] D. Chase, Code combining: a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets, IEEE Trans. Commun. 33(5) (1985) 385–393.

[14] P. Frenger, S. Parkvall, E. Dahlman, Performance comparison of HARQ with Chase combining and incremental redundancy for HSDPA, in: Proc. IEEE Vehicular Technology Conf. (VTC Fall), vol. 3, 2001, pp. 1829–1833.

[15] D. R. Pauluzzi, N. C. Beaulieu, A comparison of SNR estimation techniques for the AWGN channel, IEEE Trans. Commun. 48(10) (2000) 1681–1691.

[16] J. Vogt, A. Finger, Improving the max-log-MAP turbo decoder, Electron. Lett. 36(23) (2000) 1937–1939.

[17] J. A. Heller, I. M. Jacobs, Viterbi decoding for satellite and space communication, IEEE Trans. Commun. Technol. 19(5) (1971) 835–848.

[18] A. J. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE Trans. Inf. Theory 13(2) (1967) 260–269.