

RESEARCH

Open Access



Environment-aware indoor LoRaWAN path loss: parametric regression comparisons, shadow fading, and calibrated fade margins

Nahshon Mokuia Obiri^{1*} and Kristof Van Laerhoven¹

*Correspondence:
nahshon.obiri@student.uni-siegen.de

¹ Department of Electrical Engineering and Computer Science, University of Siegen, Siegen, Germany

Abstract

Indoor long-range wide area network (LoRaWAN) propagation is shaped by structural and time-varying environmental factors, which limit single-slope log-distance models and the standard log-normal shadowing assumption. We propose an environment-conditioned path loss framework that augments a log-distance multi-wall baseline with co-recorded environmental covariates (relative humidity, temperature, carbon dioxide, particulate matter, and barometric pressure) and receiver-reported signal-to-noise, and we validate both the mean and the residual law statistically. The approach is evaluated on a 12-month campaign in an eighth-floor office (240 m²) using time-blocked 5-fold cross-validation and a chronological hold-out. Across parametric regressors (regularized multiple linear regression (MLR), conjugate Bayesian linear regression, and a selective quadratic MLR extension on continuous predictors), the selective polynomial mean improves out-of-sample accuracy, reducing cross-validated root mean square error from 8.23 to 7.38 dB and increasing R^2 from 0.81 to 0.84. Out-of-fold (OOF) residuals are distinctly non-Gaussian and are best summarized by a compact 3-component Gaussian mixture with a sharp core and a light, broad tail. Finally, we translate prediction error to reliability by prescribing the fade margin as an upper-tail percentile of OOF errors, attaching moving-block bootstrap uncertainty, and validating the achieved outage on a held-out set. At 99% reliability, the polynomial model requires 25.73 dB versus 27.79 to 28.05 dB for linear baselines, enabling tighter indoor massive Internet of Things link budgets aligned with sixth-generation reliability targets under energy constraints.

Keywords: LoRaWAN, Indoor propagation, Path loss, Internet of Things, Environmental sensing, Analysis of variance, Shadow fading, Fade margin

1 Introduction

Indoor long range wide area network (LoRaWAN) technology deployments are challenging to model because propagation is shaped by building materials and layout, time-varying occupancy, and environmental dynamics such as humidity and temperature [1]. These factors induce rich multipath, attenuation discontinuities, and shadowing that break stationarity and yield heavy-tailed, heteroskedastic errors, violating the assumptions of classical log-distance path loss models (LDPLMs) and leading to underestimated

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

uncertainty and fade margins [2]. Moreover, statistical validation criteria must be met when developing reliable path loss models. Specifically, it is essential to assess the statistical significance of model parameters, such as the path loss exponent, using analysis of variance (ANOVA). Furthermore, many classical formulations commonly assume that the residuals (shadow fading) **(i)** follow an approximately Gaussian law in decibels (dB) (i.e., log-normal in linear power), **(ii)** exhibit approximately constant variance (homoskedasticity), and **(iii)** are approximately independent of the predictors (e.g., distance, structural composition, environmental parameters) and weakly autocorrelated [3]. In this work, we test these assumptions diagnostically rather than presupposing them.

Motivated by the stringent reliability and adaptability requirements envisioned for sixth-generation (6 G) wireless networks, such as ultra-reliable and low-latency communication (URLLC) and massive machine-type communication (mMTC) [4], this work extends conventional indoor LoRaWAN propagation modeling. We complement recent outdoor, weather-driven LoRaWAN predictability studies, such as [3] and [5], by bringing the same idea inside the building. We propose and validate an advanced statistical modeling approach that explicitly integrates environmental variables into parametric regression techniques, ensuring robust statistical diagnostics. Specifically, we build upon our previous empirical characterization of an extensive dataset [6], comprising a 12-month measurement dataset collected on the 8th floor of a single 240 m² operational office space at 250 m above sea level, as elaborated in Table 1 of research questions (RQs).

We extend our prior study [7] by moving beyond multiple linear regression (MLR) to a comparative analysis of three parametric families: **(i)** baseline MLR with ridge regression (least squares with an ℓ_2 (squared Euclidean norm) penalty, also known as Tikhonov regularization), least absolute shrinkage and selection operator (Lasso) with an ℓ_1 (absolute-value / sparsity) penalty, and elastic net regularization; **(ii)** POLY2, a selective quadratic extension MLR that adds second-order terms only for continuous predictors (distance, environment, the signal-to-noise ratio (SNR)), while keeping wall-count terms linear, and **(iii)** Bayesian linear regression (BLR) with conjugate priors, namely the normal inverse gamma prior and the Zellner g prior. This comparison addresses nonlinear propagation behavior and multicollinearity among environmental predictors (discussed in Sec. 2.3). Under a 5-fold cross-validation, the second-order polynomial emerged as superior, confirming nonlinear dependencies between environmental factors

Table 1 Research questions guiding this study

#	Research Question
RQ1	Do environment-aware covariates deliver statistically significant and practically meaningful improvements over structure-only baselines, as verified by heteroskedasticity-robust <i>Type II/III</i> analysis of variance (ANOVA) and nested partial- <i>F</i> tests?
RQ2	Which parametric mean model among regularized multiple linear regression (MLR), conjugate Bayesian linear regression (BLR), and a selective quadratic extension on continuous predictors for the MLR, achieves the best bias-variance trade-off under time-ordered <i>k</i> -fold cross-validation?
RQ3	What distribution best characterizes out-of-fold (OOF) shadow fading among single-component parametric baselines and Gaussian mixture models (GMMs), according to the goodness-of-fit criteria and residual diagnostics?
RQ4	How do the residual upper-tail quantiles map to calibrated fade margins at target outages of 1–5%, and what savings in dB do they enable relative to fixed margin heuristics?

and signal attenuation and substantially reducing prediction errors relative to the linear alternatives.

Integrating environmental sensing directly into the path loss modeling process is crucial for deploying context-aware, resilient indoor Internet of Things (IoT) infrastructures, as targeted by 6 G [4]. It enables accurate network planning, adaptive resource allocation, and power optimization in dynamic scenarios where environmental and occupancy conditions continuously fluctuate. As observed in [8], while continuous monitoring of these parameters adds moderate complexity and power overhead, deploying low-cost environmental sensors alongside LoRaWAN devices provides actionable, real-time insights that enable dynamic adjustment of path loss predictions, precise calibration of fade margins accounting for residual multimodality, and consequently, robust and reliable network performance.

Our analytical framework incorporates a detailed statistical validation procedure that combines comparative parametric regression with classical predictors of distance and wall counts, along with weights for five environmental factors: relative humidity, temperature, CO₂, particulate matter, and barometric pressure. We apply heteroskedasticity-robust *Type II* and *Type III* ANOVA, together with nested partial *F* tests, to validate predictor relevance and interactions and to quantify the reduction in unexplained variance attributable to the environmental terms. Shadow fading is examined using parametric distributions, including Normal, Skew–Normal, and GMMs, as well as non-parametric methods such as kernel density estimation (KDE), bootstrapping for confidence interval (CI) estimation, and Kruskal–Wallis tests for group-wise location shifts, and Brown–Forsythe/Levene-type tests for group-wise dispersion (variance heterogeneity). Finally, we treat fade margin as a reliability-controlled buffer inferred from the upper tail of cross-validated residuals, and we verify calibration on held-out data via the achieved outage and its complement, the achieved reliability (non-exceedance) on received packets. This calibration operationalizes our statistical gains into a deployable reliability control parameter (fade margin) for network planning with minimal overhead.

This integrated methodological toolkit provides deeper theoretical insights and directly addresses the practical deployment challenges envisioned for 6 G wireless systems. The outcomes of this research provide a framework for future investigations, including machine learning (ML) extensions, multi-site validations, and real-time adaptive propagation models, thereby contributing to the development of reliable, sustainable, and context-aware indoor IoT networks. The contributions of this work are:

- (i) A head-to-head statistical evaluation of physically grounded parametric regressors using time-blocked 5-fold cross-validation, including regularized linear MLR, conjugate BLR with shrinkage priors, and regularized selective second-order MLR. The best second-order model reduced the root mean square error (RMSE) from 8.23 to 7.38 dB and increased R^2 from 0.81 to 0.84 (Table 5).
- (ii) Explicit, heteroskedasticity-robust *Type II/III* ANOVA with nested partial-*F* tests that quantify predictor relevance. Relative to a structure-only baseline, adding environment-aware covariates and SNR reduces unexplained variance by 44% (Section 4.2, Table 7).
- (iii) A residual (shadow-fading) characterization from OOF errors. Goodness-of-fit criteria and diagnostics favor a 3-component Gaussian mixture model (GMM)

over single-component baselines. KDE and groupwise dispersion tests confirm a sharp core with a light, broad tail (Table 9).

- (iv) A quantile-based fade margin rule, $\widehat{M}_F(p) = Q_{1-p}(r)$, with p as the target outage probability, with bias-corrected and accelerated (BCa) moving-block bootstrap uncertainty and held-out validation. At $p = 1\%$, the selective polynomial mean requires 25.73 dB versus 27.79 to 28.05 dB for linear baselines ($\gtrsim 2.2$ dB lower; Eq. (15), Table 11, Fig. 6).

The remainder of this paper is organized as follows: Section 2 reviews indoor LoRaWAN propagation and motivates the incorporation of environmental context. Section 3 details the experimental setup, feature construction, model specification, validation protocol, and fade margin calibration procedure. Section 4 reports comparative predictive performance and ANOVA effect sizes, characterizes the residual (shadow fading) distribution, and translates residual upper-tail behavior into calibrated fade margins validated on a held-out set. Section 5 concludes with the main takeaways, and Section 6 outlines limitations and future directions.

2 Background

6 G roadmaps consistently foreground three constraints for massive IoT: **(i)** reliability under nonstationary conditions, **(ii)** energy efficiency within strict duty-cycle and battery limits, and **(iii)** context awareness so that networks sense and adapt to their environment rather than treating it as exogenous noise [4]. In this setting, LoRaWAN plays a complementary role to higher-throughput 6 G interfaces by offering sub-GHz penetration and multi-year device lifetimes for dense indoor sensing, as we established in our recent survey [9], provided that propagation models are site-specific and time-aware to support low-margin, reliable links. Conditioning path loss and shadowing on ambient state reduces unexplained variance and enables tighter fade margins with lower transmit energy in practice [8]. Our 6 G-aligned statistical evaluation similarly showed that incorporating environmental covariates reduces unexplained variance and reveals multi-modal residual structure that standard single-slope baselines miss [7]. Accordingly, this section connects these 6 G constraints to the case for environment-conditioned indoor LoRaWAN propagation models adopted in this work.

2.1 Indoor LoRaWAN signal propagation

Indoor LoRaWAN propagation is notoriously complex due to the rich scattering and attenuation effects caused by walls, floors, and clutter in buildings [10]. Most empirical wireless network propagation models, including LoRaWAN, assume a logarithmic increase in path loss with distance, as per the LDPLM. This simple model often serves as a valuable starting point, but it does not capture all the complexities of indoor environments. In practice, measured path loss exponents in buildings vary widely. Corridor measurement in [11] found an apparent path loss slope below 2.0 in one non-line-of-sight (NLoS) region and above 3.0 in a more obstructed section. Such variability reflects how indoor layout and materials either guide propagation, for example, along corridor waveguides, or impose excess loss through multiple brick walls. For instance, campus-scale studies in [12] report exponents from 1.37 (indoor line-of-sight (LoS)) to values

greater than 2.3 (indoor NLoS), highlighting that internal materials and layout dominate performance and that one-size exponents are rarely adequate. Moreover, it has also been observed in [12, 13] that uncalibrated models tend to either over- or under-predict coverage.

To address structural variability, refined models explicitly encode building features. Comparative evaluations against indoor LoRa measurements (e.g., ITU-R P.1238 [14], COST 231 multi-wall model (MWM) [15], Motley–Keenan [16]) show that multi-wall formulations often deliver the best accuracy when parameters are locally calibrated [13]. However, even with careful calibration, purely structural models fail to capture temporal influences. For example, multi-floor residential measurements in [17] show pronounced room- and time-dependent fading (standard deviation up to 18.4 dB in LoS), pointing to occupancy and heating, ventilation, and air conditioning (HVAC) cycles as additional covariates of variability.

Consequently, recent work has shifted toward hybrid or semi-empirical approaches, starting from a physically anchored path loss law, then learning data-driven corrections that reflect site materials, geometry, and temporal context [10, 18]. This motivates our stance of augmenting structural predictors with environmental variables (e.g., humidity, temperature, CO₂, pressure, particulate matter) and validating them using intensive statistics rather than assuming log-normal shadowing and homoscedasticity by default. We now operationalize this by comparing parametric regressors using cross-validation, quantifying effect sizes via ANOVA, and characterizing shadow fading with mixture models, thereby aligning with background theory and providing a deployment-ready, reliability-calibrated model for indoor LoRaWAN propagation.

2.2 Environment-aware LoRaWAN propagation

Beyond static architecture, indoor links are shaped by time-varying context: occupancy, HVAC cycles, and the microclimate. Long-duration deployments show that temperature and humidity fluctuations correlate with shadow fading over days [19]. On the other hand, crowded periods introduce human-body absorption and moving scatterers, increasing variance of the received signal strength indicator (RSSI) and reducing short-term stability [2, 20]. Even in office settings, Wi-Fi or Bluetooth Low Energy (BLE) links become erratic during busy hours and stabilize overnight [21]. These observations suggest that purely structural models cannot capture nonstationarity.

Outdoor long-term evidence shows diurnal weather structure in SNR and link-specific sensitivity to temperature and humidity [5]. From a physics perspective, microclimate mechanisms explain the measurable path loss drift. For instance, humidity modifies the effective permittivity of air and porous materials while temperature and air-density changes alter boundary conditions, which shape attenuation at sub-GHz [3, 22]. Empirically, higher indoor temperatures or associated ventilation states can coincide with slightly lower path loss, while elevated CO₂ tracks occupancy and richer multipath [7]. Thus, variables such as temperature, relative humidity, CO₂, pressure, and particulate matter serve as practical proxies for the environment's latent state, which governs propagation.

Motivated by this, recent work augments path loss models with co-located environmental sensing. In our prior study [6], injecting CO₂, humidity, temperature, pressure,

and particulate matter into an LDPLM and wall-loss baseline reduced unexplained variance by over 40%, because the model could condition on occupancy and ventilation rather than treating their effects as noise. Others similarly report gains when connectivity control becomes environmental-aware [3]. Therefore, environment-aware modeling, which conditions on occupancy and ventilation via co-located sensing, sharpens link budgets and makes fade margins situational rather than static, yielding more reliable indoor LoRaWAN deployments.

Concurrently, the field is moving beyond curve fitting toward statistically rigorous and adaptive models. Hybrid methods add learned corrections to physics-grounded baselines [23], and dynamic filters improve distance estimates by tempering fast fading and transient blockages [24]. In this paper, we take a principled approach: we compare parametric regressors that balance interpretability and nonlinearity, validate our contributions with ANOVA, and investigate shadow fading using mixture models and non-parametric diagnostics. The result is an environment-aware formulation grounded in testable assumptions and a reliability-calibrated interface for network planning via quantile-based fade margins.

2.3 Statistical regression methods

Modeling indoor LoRaWAN propagation spans from transparent linear formulations to more flexible parametric and ML approaches. A recurring theme is the trade-off between interpretability, needed to attribute loss to distance, walls, or climate, and flexibility to capture mild nonlinearities and correlated predictors. Accordingly, studies commonly adopt linear baselines for physical clarity, then introduce controlled extensions, such as polynomial terms or regularization, and assess adequacy using cross-validation, information criteria, and residual diagnostics to ensure that out-of-sample gains warrant the added complexity. This section reviews representative linear baselines and polynomial extensions.

2.3.1 Multiple linear regression

MLR is a foundational regression technique that models path loss as a linear combination of multiple predictors, including distance, wall penetration, and environmental parameters (e.g., humidity, temperature). If V denotes the number of predictors, it takes the general form in Eq. (1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_V x_V + \psi, \quad (1)$$

where y represents the observed path loss, β_0 is the intercept, β_1, \dots, β_V denote regression coefficients, x_1, \dots, x_V are the predictors, and ψ is the residual error [25], typically assumed to be normally distributed ($\psi \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is the residual (noise) variance).

A primary advantage of MLR is its interpretability. Each coefficient corresponds to a physically meaningful impact, such as attenuation per wall or unit distance, making it particularly valuable for practical engineering applications and network planning. Beyond estimating these coefficients, an ANOVA can be performed to confirm the overall statistical significance of the regression and to determine whether each predictor makes a meaningful contribution. For example, Bertoldo et al. [13] used MLR to quantify

propagation losses tied to specific building materials and structures. However, reliability depends on assumptions of linearity, approximate normality, homoscedasticity of residuals, and weak autocorrelation. Empirical work therefore emphasizes residual diagnostics, Quantile–Quantile plots, and tests such as Jarque–Bera (normality), Breusch–Pagan (heteroskedasticity), and Durbin–Watson (autocorrelation), to verify these assumptions [8]. Violations typically motivate carefully constrained extensions (polynomial terms or regularization) rather than abandoning the physically anchored linear form.

2.3.2 Polynomial regression

Polynomial regression extends the capabilities of MLR by incorporating polynomial or interaction terms into the model to capture subtle nonlinearities commonly observed in indoor propagation environments. Let V denote the number of predictors. Equation (2) gives a common second-order specification.

$$y = \beta_0 + \sum_{i=1}^V \beta_i x_i + \sum_{i=1}^V \sum_{j=i}^V \beta_{ij} x_i x_j + \psi, \quad (2)$$

where β_{ij} are coefficients for second-order terms, yielding quadratic predictors when $i = j$ (i.e., x_i^2) and cross-product interaction predictors when $i < j$ (i.e., $x_i x_j$), with indices $i, j \in \{1, \dots, V\}$. The constraint $j \geq i$ avoids duplicating symmetric interaction terms (since $x_i x_j = x_j x_i$), enabling the model to capture curvature and pairwise interactions between variables (e.g., distance–humidity coupling). This extension is practical when the log-distance relationship departs slightly from strict linearity, since even a single quadratic term for distance can approximate multi-slope behavior and reduce error while preserving interpretability [2]. Studies typically select second-order terms via cross-validation or information criteria (Akaike information criterion (AIC) [26] or Bayesian information criterion (BIC) [27]), and re-check residual assumptions (normality, homoscedasticity, and independence) because added terms may increase complexity and multicollinearity. In summary, polynomial regression is a compact extension that maintains interpretability while improving fit over MLR.

2.3.3 Regression regularization methods

Regularization enhances regression by introducing penalty terms that regulate variance and improve generalization in the presence of multicollinearity and limited sample sizes. These conditions are common in indoor LoRaWAN studies where environmental variables often co-vary. The three standard approaches are Ridge (an ℓ_2 penalty), Lasso (an ℓ_1 penalty), and Elastic Net (a mixture of ℓ_1 and ℓ_2). They differ in the penalty imposed and, consequently, in how coefficients are shrunk and selected [28].

Here, \mathbb{R} denotes the real numbers. Let N denote the number of observations (samples) and let V denote the number of predictors after feature mapping. Let $\mathbf{y} \in \mathbb{R}^N$ collect the observed path loss values, and let $\mathbf{1} \in \mathbb{R}^N$ denote the all-ones vector (so the intercept is not penalized). Let $\beta_0 \in \mathbb{R}$ denote the intercept and let $\boldsymbol{\beta} \in \mathbb{R}^V$ denote the regression coefficient vector. Let $\Phi \in \mathbb{R}^{N \times V}$ denote the standardized design matrix obtained from the physics-guided second-order feature map (nonlinear predictors included), and let \mathbf{X} denote the linear design when no polynomial expansion is used. Estimators are

defined as solutions to penalized least-squares problems. The notation $\|\cdot\|_2^2$ denotes the Euclidean norm and $\|\cdot\|_1$ the ℓ_1 norm; thus $\frac{1}{2N}\|\cdot\|_2^2$ is the per-sample average of squared residuals, and the factor $\frac{1}{2}$ is conventional and does not affect the minimizer. For ordinary MLR without polynomial terms, the feature map is the identity, hence $\Phi \equiv \mathbf{X}$ and the problems below reduce to standard MLR. For Elastic Net (ENet), $\alpha \in [0, 1]$ controls the ℓ_1/ℓ_2 mix ($\alpha = 1$ yields Lasso; $\alpha = 0$ yields Ridge up to the conventional scaling of λ under this normalization) and $\lambda > 0$ is the overall penalty strength. Equivalently, the two-parameter form $\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2$ corresponds to $\lambda_1 = \lambda\alpha$ and $\lambda_2 = \lambda(1-\alpha)/2$ under this normalization. Then,

Let $\mathbf{y} \in \mathbb{R}^N$ denote the path loss observations and let $\Phi \in \mathbb{R}^{N \times V}$ be the standardized design matrix (intercept excluded). Ridge, Lasso, and Elastic Net estimate $(\beta_0, \boldsymbol{\beta})$ by minimizing the mean squared error plus an ℓ_2 and/or ℓ_1 penalty, with $\lambda > 0$ controlling overall shrinkage and $\alpha \in [0, 1]$ controlling the ℓ_1/ℓ_2 mix.

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\text{ols}}) = \arg \min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2N} \|\mathbf{y} - \beta_0 \mathbf{1} - \Phi \boldsymbol{\beta}\|_2^2, \quad (3)$$

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\text{ridge}}) = \arg \min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2N} \|\mathbf{y} - \beta_0 \mathbf{1} - \Phi \boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2, \quad (4)$$

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\text{lasso}}) = \arg \min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2N} \|\mathbf{y} - \beta_0 \mathbf{1} - \Phi \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1, \quad (5)$$

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\text{ENet}}) = \arg \min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2N} \|\mathbf{y} - \beta_0 \mathbf{1} - \Phi \boldsymbol{\beta}\|_2^2 + \lambda \left(\frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right). \quad \alpha \in [0, 1]. \quad (6)$$

The standard ordinary least squares (OLS) objective in Eq. (3) is augmented with penalty terms to improve stability, interpretability, and out-of-sample performance. Lasso, defined by Eq. (5), uses an ℓ_1 penalty that can drive some coefficients to zero and produce a compact model. In indoor Wi-Fi fingerprinting, such sparsity reduced localization error and, with a modest extension, enabled the identification of outlier access points [29]. Ridge, given in Eq. (4), uses an ℓ_2 penalty that shrinks all coefficients while keeping them nonzero, which is valuable when predictors are highly correlated. Comparative LoRa measurements report similar indoor behavior for Ridge and Lasso. In contrast, simple linear models degrade outdoors, suggesting the importance of feature design and mild nonlinearity over the specific penalty choice in that scenario [30]. Elastic Net, specified in Eq. (6), combines ℓ_1 and ℓ_2 penalties and performs feature selection while retaining groups of correlated variables. It has demonstrated strong performance with correlated received signal strength features in a wireless local area network [29] and has served effectively as a global regressor in visible-light positioning, remaining accurate with few training samples when paired with a lightweight residual-correction step [31]. Across these works, predictors are standardized, and penalty parameters are selected using cross-validation. Elastic Net offers a balanced default for strongly correlated feature sets, which are common in indoor sensing. Ridge preserves stability without removing variables, whereas Lasso yields a compact, interpretable subset of predictors.

2.3.4 Bayesian linear regression

BLR models the relationship between path loss and predictors in a fully probabilistic manner, treating both the coefficients and the noise variance as random. With a Gaussian likelihood and the conjugate Normal–Inverse-Gamma (NIG) prior over $(\boldsymbol{\beta}, \sigma^2)$, Eq. (7) specifies the BLR model:

$$\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_N), \quad \boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{\Sigma}_0), \quad \sigma^2 \sim \text{Inv-Gamma}(a_0, b_0), \quad (7)$$

$\mathbf{X} \in \mathbb{R}^{N \times V}$ stacks standardized predictors (e.g., log-distance, wall counts, humidity, temperature, CO₂, pressure, particulate matter); $\sigma^2 > 0$ is the noise (residual) variance and $\sigma > 0$ is the corresponding residual standard deviation; \mathbf{I}_N is the $N \times N$ identity; $(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ and (a_0, b_0) are prior mean/covariance and shape/scale hyperparameters, with $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{V \times V}$ and $\boldsymbol{\Sigma}_0 \succ 0$ (positive definite). Standardizing predictors (zero mean, unit variance) makes coefficient priors comparable across features and stabilizes inference. The posterior updates in closed form, and the posterior predictive distribution is Student- t , a practical benefit when residuals exhibit heavier tails than Gaussian [32].

Regularization appears naturally as a prior choice where a zero-mean Gaussian prior over coefficients yields ridge-like shrinkage (i.e., Maximum A Posteriori (MAP) estimation with a Gaussian prior), whereas a Laplace prior corresponds to Lasso-type sparsity. Both provide stability when environmental covariates are correlated, without abandoning interpretability [33]. A widely used alternative is Zellner's g -prior, which ties the coefficient prior covariance to the observed design and preserves conjugacy (including in common mixtures of g variants), providing analytic tractability and adaptive shrinkage in regression settings [32]. In particular, Eq. (8) defines the prior:

$$\boldsymbol{\beta} \mid \sigma^2 \sim \mathcal{N}(\mathbf{0}, g \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}), \quad (8)$$

where $g > 0$ is a scalar hyperparameter that controls the shrinkage strength and $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists under full column rank, i.e., invertible. Here, \mathbf{X} excludes the intercept column $\mathbf{1}$, while $\boldsymbol{\beta}_0$ is treated separately and is not assigned a g -prior.

Model assessment in modern Bayesian workflows typically relies on out-of-sample criteria computed from the full posterior distribution. Leave-one-out cross-validation (LOO) and the Widely Applicable Information Criterion (WAIC) are widely recommended. On the other hand, Pareto-smoothed importance sampling (PSIS-LOO) makes LOO efficient and diagnostically transparent (via shape-parameter checks), helping guard against overconfident fits [34]. When sparsity or group-wise shrinkage is desired (e.g., many correlated climate features), hierarchical priors such as the horseshoe concentrate mass near zero while preserving heavy tails for truly non-negligible effects, offering adaptive regularization without manual feature pruning.

2.3.5 Machine learning regressors

Beyond parametric approaches, indoor propagation studies increasingly employ ML models that learn nonlinear mappings from measurements to path loss with minimal functional assumptions [35]. Commonly used models include support vector regression (SVR), Gaussian process regression (GPR), tree ensembles (random forests and boosted

ensembles), and neural networks. Kernel SVR captures moderate nonlinearities through flexible similarity measures while retaining a convex training objective. Its performance depends on the choice of kernel and the SVR hyperparameters (C, ε) , where $C > 0$ is the penalty (regularization) strength controlling the trade-off between model flatness and training errors, and $\varepsilon > 0$ is the width of the ε -insensitive tube in the SVR loss. GPR provides a probabilistic surrogate with uncertainty quantification via kernel covariances, but its $\mathcal{O}(N^3)$ training complexity in the number of samples N often requires sparse or inducing-point approximations in long campaigns. Tree ensembles are robust to heterogeneous features and monotone transformations, handle interactions without requiring polynomial terms, and provide variable-importance profiles that facilitate interpretation. Neural networks (from shallow multilayer perceptrons (MLPs) to temporal convolutional neural networks (CNNs) or long short-term memory (LSTM) variants) can further reduce error when large and diverse datasets are available and when nonlinear couplings (e.g., distance \times occupancy proxies) are strong. However, they typically trade interpretability for accuracy and require careful regularization.

Reported performance for indoor LoRaWAN propagation is heterogeneous, with ML gains ranging from negligible to substantial, depending on the residual complexity remaining after accounting for dominant structural drivers and the available training data. When distance and obstruction proxies (e.g., wall counts) already capture much of the large-scale attenuation, incremental gains from purely data-driven models can be modest, and limited data further increases the risk of overfitting. Consequently, ML is frequently deployed in a complementary role, for example, as a correction layer on top of a physically grounded baseline [36] or as a context-aware component that adapts to occupancy and microclimate indicators. Best practices emphasize leakage-safe evaluation (e.g., grouped and time-aware splits, nested validation, and stability checks) so that improvements reflect genuine generalization rather than shortcut learning. Overall, ML regressors are most effective when used in hybrid strategies that balance physical interpretability with flexible, data-driven corrections.

2.4 Predictor statistical significance

A central question in indoor propagation studies is whether each predictor contributes uniquely to explaining path loss beyond correlated alternatives. Standard practice combines coefficient t -tests with partial- F tests via ANOVA to assess main effects and interactions under a linear modeling framework [37]. For nested models $\mathcal{M}_0 \subset \mathcal{M}_1$, the partial- F statistic compares the residual sum of squares, here denoted by $SS_{\text{res}}(\cdot)$, as given in Eq. (9):

$$F = \frac{(SS_{\text{res}}(\mathcal{M}_0) - SS_{\text{res}}(\mathcal{M}_1))/(V_1 - V_0)}{SS_{\text{res}}(\mathcal{M}_1)/(N - V_1 - 1)}, \quad (9)$$

where $i \in \{0, 1\}$ indexes the reduced and full models, N is the sample size, and V_i is the number of predictors in \mathcal{M}_i (excluding the intercept). Under \mathcal{M}_0 and standard regularity assumptions, F follows an F distribution with numerator and denominator degrees of freedom (df) $(V_1 - V_0, N - V_1 - 1)$. A large F indicates that the added terms (e.g., humidity or an interaction such as distance \times humidity) reduce unexplained variance beyond chance.

Since indoor datasets are often unbalanced and predictors can be correlated (e.g., temperature with humidity), the choice of ANOVA type matters [38]. *Type I* (sequential) ANOVA attributes sums of squares in the order variables enter the model and is order-dependent. It is mainly used for designed and balanced experiments. *Type II* ANOVA evaluates each main effect after adjusting for the other main effects (but not interactions) and is preferred when interactions are absent or excluded. *Type III* ANOVA tests each effect while adjusting for all other main effects and interactions, and is common when interaction terms are included. Across types, collinearity inflates uncertainty and can obscure effects. For this reason, studies often report variance inflation factors (VIFs), standardized predictors, and interpret *Type II* and *Type III* results with caution.

Inference assumes residuals are approximately Gaussian, variance-stable, and weakly autocorrelated. When diagnostics suggest deviations, several robustifications appear in the literature. Heteroskedasticity-consistent (HC) covariance estimators for t - and F -tests, permutation (randomization) ANOVA in small or non-normal samples, and block-bootstrap CIs when short-memory temporal dependence is present [38]. Non-parametric rank tests (e.g., Kruskal–Wallis) are used for group-wise comparisons when normality is doubtful, complementing rather than replacing parametric ANOVA.

Empirical indoor work illustrates these points through varied designs. Two-way ANOVA has been used to evaluate the effects of categorized temperature and humidity on RSSI in mote-based testbeds, finding significant main effects [39]. Full-factorial analyses have quantified interactions among channel, link path, and transmit power, with interaction terms explaining sizable portions of variability [40]. ANOVA has also appeared in network-level studies, for example, to identify redundant sensing clusters, demonstrating statistically grounded routes to energy savings [41]. Collectively, these practices establish a procedure that begins with a physically motivated linear specification, tests main effects with *Type II* or *Type III* ANOVA as design dictates, inspects residual assumptions, and deploys robust or rank-based alternatives when diagnostics warrant.

2.5 Residual distribution diagnostics

Residual distribution (shadow fading) refers to the residual variability that remains after accounting for distance, structural losses, and auxiliary predictors, such as the indoor microclimate [8]. In indoor settings, these residuals often deviate from the Gaussian assumption. Specifically, skewness and heavy tails arise due to dynamic blockage and hardware heterogeneity, while multimodality reflects regime changes, such as occupancy cycles or HVAC operation. A defensible characterization, therefore, considers distributional shape, variance stability, and potential temporal or group dependence using complementary parametric and nonparametric tools, as summarized in Table 2. In practice, to avoid optimistic bias, all diagnostic fits and tail summaries are computed on OOF residuals.

Regarding distributional shape, studies fit unimodal families such as the Normal, Student's t , and Skew–Normal (asymmetry) [55], and occasionally Cauchy (very heavy tails) [56], and inspect Q–Q plots alongside omnibus tests (e.g., Kolmogorov–Smirnov (KS) [57]) to gage tail thickness and asymmetry. When a single law is inadequate, GMMs are used to represent heterogeneous regimes, with the component count selected by AIC

Table 2 Residual diagnostic methods used in indoor path loss modeling. In the Type column, a circle ○ marks parametric methods, whereas a square □ marks nonparametric methods; i.i.d. means independent and identically distributed

Method	Type	Purpose and typical use
Single-family goodness-of-fit	○	Screen for Gaussianity versus heavy tails/asymmetry using Q–Q plots plus omnibus tests; establishes whether a unimodal law suffices [42]
Gaussian Mixture Models (GMMs)	○	Compact representation of heterogeneous regimes (e.g., occupancy/HVAC states); select components by Akaike information criterion (AIC) or Bayesian information criterion (BIC) and validate on held-out data (e.g., likelihood) [43]
Kernel Density Estimation (KDE)	□	Distribution visualization without parametric constraints; reveals secondary modes or mild asymmetry that guide model choice [44]
Heteroskedasticity tests	○	Detect variance changes with fitted values or covariates; motivates weighting or robust standard errors when variance is not constant [45]
Group-wise location/dispersion tests	□	Kruskal–Wallis for median/location differences; Brown–Forsythe (median-centered Levene) or Fligner–Killeen for dispersion. [46]
Autocorrelation diagnostics	○	Assess short-memory dependence that affects uncertainty estimation and motivates blocked validation protocols [47]
Bootstrap quantiles (i.i.d. and block)	□	Empirical confidence intervals (CIs) for tail quantiles used in fade margin budgeting; block variants preserve dependence [48]

or BIC and validated on held-out data [58]. Nonparametric KDE often reveals subsidiary modes or mild asymmetry that inform model choice [44]. Variance heterogeneity is probed with scale–location plots and classical tests (Breusch–Pagan/White). Grouped comparisons by device, location, or time-of-day use Levene/Brown–Forsythe or rank-based tests (Kruskal–Wallis) to detect structure-induced dispersion shifts [46]. Temporal diagnostics such as the autocorrelation function (ACF) and partial autocorrelation function (PACF), the Durbin–Watson statistic, and the Ljung–Box test check short-memory effects that can bias uncertainty estimates if ignored [47]. For uncertainty in tail summaries, bootstrap resampling provides empirical CIs for error quantiles [48], and block variants preserve short-range dependence when it is present.

In practice, the aim is not maximal flexibility but the simplest residual model consistent with diagnostics and calibrated tails. Indoors, this often means starting from a Normal fit, escalating to *t* or Skew–Normal when tails or asymmetry dominate, and adopting a low-order GMM only when multimodality is persistent and interpretable. The resulting upper-tail quantiles directly support fade margin budgeting in reliability analyses, provided they are derived from OOF residuals.

2.6 Design goals and methodological positioning

Across prior work (see Table 3), persistent gaps include the limited integration of environmental context, scarce leakage-safe evaluation with calibrated uncertainty, and residual shapes that are seldom scrutinized beyond simple normality checks. Consequently, fade margin prescriptions are rarely validated on held-out data. Our study addresses these gaps with a physics-grounded mean, selective nonlinear terms, time-blocked cross-validation, and residual diagnostics to obtain interpretable effects and calibrated reliability margins. Guided by these gaps, our design goals are to: (i) preserve physical

Table 3 A comparative summary of existing indoor LoRaWAN propagation studies. We compare prior work by assessing whether environmental factors (EF), advanced modeling (AM), and residual analysis (RA) were addressed. For **EF**, [51] indicates included, I indicates indirect treatment, and [55] indicates not included; for **AM**, [51] indicates performed, P indicates partially performed, and [55] indicates not performed; for **RA**, [51] indicates performed and [55] indicates not performed

Ref.	Focus and summary of findings	EF ^a	AM ^b	RA ^c
[49] (2017)	Single-floor office tests (868 MHz) measuring RSSI and packet delivery vs. spreading factor (SF) 7–12 across four locations; higher SF did not uniformly improve delivery (long time-on-air effects), offering practical configuration guidance	I	X	X
[36] (2017)	Outdoor–indoor office study: adjusted COST 231 MWM with a neural network residual corrector; test error improves (mean squared error from 21.0 to 11.23 dBm ²) when generalizing from the 8th to the 7th floor	X	✓	X
[50] (2017)	Campus-scale indoor trials (868 MHz) with a Kerlink gateway (≈ 24 m) and static/wearable nodes; reliable room-level coverage at SF12, 14 dBm with packet delivery $\approx 96.7\%$; descriptive path-loss statistics	X	X	X
[51] (2018)	Indoor multi-floor measurements (Duisburg): same-floor RSSI/SNR largely SF-independent; basement delivery ranged 62% (SF7) to 100% (SF10/SF12); larger payloads reduced delivery	X	X	X
[52] (2019)	Indoor empirical path loss modeling at 868 MHz: COST 231 MWM and floor with $n = 2.85$, $PL_0 = 120.4$ dB, per-wall $L_w = 1.41$ dB, per-floor $L_f = 10$ dB, and $b = 0.47$; reports shadowing $\sigma \approx 8$ – 9.7 dB	X	P	X
[53] (2019)	Indoor multi-floor study (gateway on Level 9; nodes on Levels 5–9): measured RSSI and packet delivery vs. SF7–SF12 (125 kHz) and payload; delivery drops with floor separation and larger payloads, time-on-air rises at high SF	X	X	X
[19] (2021)	Three-story office mapping ($N = 89$ points): hallway RSSI–distance shows anomalies at $\sim 40, 80, 120$ m from window reflections; brick walls cause strong short-range loss	X	X	X
[54] (2021)	Indoor multi-block building (SF12, 14 dBm): near-perfect delivery within the gateway block; first losses at ~ 40 m, steep drop beyond ~ 75 m; recommends $\lesssim 70$ m links	X	X	X
[20] (2022)	Live indoor multi-gateway office (EU-868, 26 days): first-attempt success 99.95% indoors vs 95.7% outdoors; city-site scans show interference up to 97.3% (uplink) and 54% (downlink)	X	X	X
[11] (2023)	Indoor corridor at 868 MHz: fits the close-in reference and floating-intercept forms to LoRa (plus Zigbee/5 G) in two NLoS zones; finds $n < 2$ (waveguide-like) in NLoS-1 and $n > 3$ in NLoS-2; floating-intercept forms outperforms close-in reference in NLoS-2 and is used to relate received power to bit error rate thresholds	X	P	X
[18] (2023)	Indoor 8-story building with rooftop end device, gateway moved floor-by-floor; fixed SF7, 14 dBm, 125 kHz, 868 MHz; 50 packets/floor at 5-s intervals; RSSI/SNR rise with proximity (e.g., Floor 1 RSSI ≈ -110 dBm, and Floor 8 ≈ -71 dBm); near-zero loss on Floors 2–8	X	X	X
[23] (2024)	For an 18-floor indoor building with a rooftop gateway, compares end device logs (1 day) vs Adeunis Field Test Device (FTD) (100 samples/floor) at Floors 1/6/12/18; reports RSSI/SNR trends with distance and SF allocations; links operate at negative SNR down to -13.8 dB	X	X	X
[24] (2024)	For an indoor lab: proposes a dynamic log-distance model with an additive noise term T and Kalman filtering; estimates indoor path loss exponent $n = 2.103$; mean distance error 0.565 m with 90% of errors < 1.08 m over 3–7 m	X	✓	X
[6] (2025)	Our dataset descriptor for the deployment (indoor office, 8th floor), described in this work. We compare a structure-only multi-wall log-distance baseline against an environment-augmented variant (environmental sensing and SNR), improving 5-fold cross-validation RMSE from 10.577 to 8.034 dB and R^2 from 0.691 to 0.822.	✓	P	X
[7] (2025)	We perform early statistical evaluation on the campaign in [6] using MLR with ANOVA, and residual distribution analysis. Adding environment-aware covariates materially reduced unexplained variance and motivated multimodal shadowing models beyond Gaussian residual assumptions.	✓	P	✓
<i>This Work</i>	Using a 12-month dataset with multivariate path loss with distance, walls/floors, and environment variables, we (i) compare MLR vs. second-order polynomial and regularized/Bayesian variants via cross-validation, (ii) perform residual diagnostics (GMM, KDE/bootstrapping, Kruskal–Wallis), and (iii) pursue quantile-based fade margin calibration on OOF residuals	✓	✓	✓

interpretability so that distance, walls, and environmental contributions remain attributable, **(ii)** incorporate environmental context to reduce unexplained variance rather than absorbing it into shadowing, **(iii)** enforce leakage-safe evaluation so reported gains reflect true generalization, and **(iv)** quantify predictive uncertainty for reliability budgeting, not just point accuracy.

These goals are operationalized with a mean specification anchored in Eq. (12) and a selective quadratic extension in Eq. (13) applied to the continuous predictors (distance, environmental factors, and the SNR, while wall counts and weights remain additive and linear. Polynomial expansion precedes standardization, while all preprocessing (feature mapping and scaling) is confined within cross-validation folds (time-blocked) to prevent leakage, and diagnostics and fade margin quantiles are computed on OOF residuals. Three model families are compared: a linear baseline MLR, the second-order polynomial design, and a BLR model (conjugate NIG and Zellner g -prior) on the linear design for calibrated predictive uncertainty. Penalized variants (Ridge, Lasso, and Elastic Net) are tuned by cross-validation with parsimony-oriented selection.

3 Methodology

3.1 Experimental setup

In this section, we describe the deployment geometry, sensing hardware, radio configuration, and the backend pipeline used to acquire the dataset for our path loss modeling and analysis.

3.1.1 Site and Layout

This data measurement campaign builds on the indoor LoRaWAN deployment reported in our EU 868 MHz dataset descriptor [6]. We summarize the elements required to reproduce the measurement conditions and the end-to-end data logging pipeline. Measurements were collected on the 8th floor of an academic building (240 m² and approx. 250 m above sea level). The floor follows a corridor-centric office layout with a semi-open central walkway (Fig. 1(a)). Structural walls are brick and reinforced concrete, while internal separations are primarily lightweight wooden partitions and wooden doors. This mixture yields links spanning LoS and NLoS, with ED0 being the only LoS device to the gateway. End device to gateway distances span $d \in [8, 40]$ m, and each link is tagged

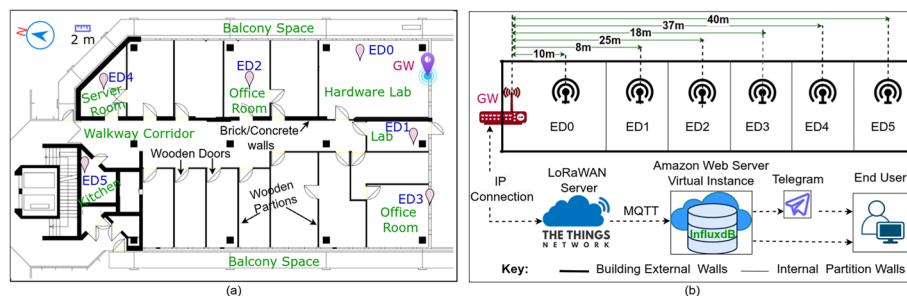


Fig. 1 Overview of the indoor LoRaWAN measurement campaign. (a) Floor plan with the gateway (GW) and six end devices (ED0–ED5), annotated obstructions distinguish brick/reinforced-concrete walls and wooden partitions/doors. Scale bar: 2 m; icons not to scale. (b) Data acquisition chain, forwarding uplinks from The Things Network (TTN) to an InfluxDB time-series database via Message Queuing Telemetry Transport (MQTT)

in Fig. 1(a) with its wall-count tuple $\mathbf{w} = (W_{\text{brick}}, W_{\text{wood}})$: ED0 (0, 0), ED1 (1, 0), ED2 (0, 2), ED3 (1, 2), ED4 (0, 5), ED5 (2, 2).

Measurements were collected during normal use of offices, laboratories, and shared spaces (uncontrolled human presence and mobility), spanning both working-hour and off-hour regimes with intermittent corridor traffic. Human presence and movement were not scripted, and no fixed headcount is assumed. We therefore treat occupancy as an uncontrolled driver of indoor non-stationarity and capture its effect indirectly through the co-recorded environmental time series. In practice, indoor CO_2 exhibits strong diurnal structure with peaks that plausibly track human activity. At the same time, $\text{PM}_{2.5}$ and other measurements can also be shaped by operational factors such as heating, time of day, and window/door opening [2]. Since headcount was not instrumented, these variables serve as proxy signals for activity-driven dynamics that may covary with LoRaWAN link reliability in situ.

3.1.2 End devices

We deployed six identical end devices (EDs, ED0–ED5) at fixed locations for the duration of the campaign (Fig. 1(a)). Each end device was built around an Arduino MKR WAN 1310 (Murata LoRa, EU 868 MHz) and housed in a ventilated 3D-printed enclosure (Fig. 2(a) and (b)). The platform exposes standard peripheral buses, namely Inter-Integrated Circuit (I2C), Serial Peripheral Interface (SPI), and Universal Asynchronous Receiver/Transmitter (UART), which simplifies clean sensor integration and debugging without additional interface hardware [3]. In this deployment, all sensors were connected to a shared I2C bus to minimize wiring and ensure consistent sampling (Table 4). For radio transmission, each ED uses an external omnidirectional antenna (peak gain $G_{\text{tx}} \approx 0.4$ dBi) connected via a short pigtail ($L_{\text{tx}} \approx 0.14$ dB) and transmits at $P_{\text{tx}} \approx 14$ dBm. Antennas were mounted vertically at ~ 0.8 m above floor level to reflect typical indoor IoT installations.

3.1.3 Gateway and Backhaul

We used a Kerlink Wirnet iFemtoCell indoor gateway (GW, Fig. 2(c)), wall-mounted at 1.0 m above floor level. The GW was equipped with a vertically polarized omnidirectional rubber-dipole antenna (peak gain $G_{\text{rx}} \approx 3$ dBi) connected directly at the RF port, hence receiver-side cable losses L_{rx} were negligible. In its EU868 configuration, the gateway provides a receiver sensitivity (S) up to -141 dBm at SF12 (125 kHz bandwidth

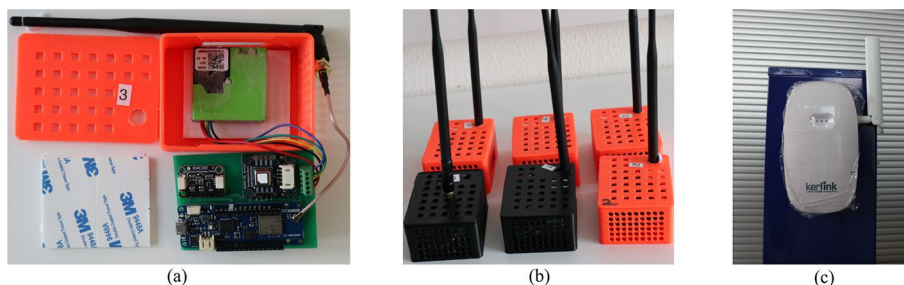


Fig. 2 LoRaWAN end devices and indoor gateway used in the deployment. (a) End device internals: Arduino MKR WAN 1310 (Murata LoRa, EU 868 MHz) with BME280 (pressure), SCD41 (CO_2 , temperature, humidity), and SPS30 ($\text{PM}_{2.5}$). (b) Assembled end devices in 3D-printed enclosures. (c) Kerlink Wirnet iFemtoCell indoor gateway (wall-mounted)

Table 4 Sensor suite integrated into each end device. The table summarizes the sensing modules, their operating interfaces or measurement principles, and the variables captured with their reported sensitivity. I2C refers to Inter- Integrated Circuit, SPI to Serial Peripheral Interface, and UART to Universal Asynchronous Receiver/Transmitter

Sensor	Interface / principle	Measured variables and sensitivity
Sensirion SCD41	I2C (used); UART is supported; includes photoacoustic CO ₂ sensing with integrated temperature and relative humidity channels; automatic self-calibration	Measures, (i) carbon dioxide concentration: 400ppm-to5000ppm; accuracy $\pm(40ppm + 5\%$ of reading), (ii) temperature: -10°C to 60°C ; accuracy $\pm 1.5^{\circ}\text{C}$ (typ. $\pm 0.8^{\circ}\text{C}$ for 15°C to 35°C), and (iii) relative humidity: 0%RHto100%RH; accuracy $\pm 9\%$ RH (typ. $\pm 6\%$ RH for 15°C to 35°C and 20%RHto65%RH)
Adafruit BME280	I2C (used); SPI supported; pressure sensing based on Bosch technology	Measures barometric pressure in the range of 300–1100 hPa; absolute accuracy ± 1.0 hPa
Sensirion SPS30	I2C (used); UART supported; laser-scattering particulate sensing with contamination-resistance; self cleaning fan	Measures particulate matter mass concentration ($PM^{2.5}$) in the range of 0–1000 $\mu\text{g}/\text{m}^3$; its reported accuracy is $\pm 10\%$ for the range within 100–1000 $\mu\text{g}/\text{m}^3$ size of particles

(BW)), enabling robust packet capture on heavily obstructed links. As was deployed in [59], uplinks were forwarded over an Ethernet backhaul to The Things Network (TTN) stack (v3.35.2). The end devices operated under the LoRaWAN MAC specification v1.0.2, and TTN exposed per-packet metadata required for link analysis (timestamps, SF, frequency, RSSI, and SNR).

3.1.4 Radio configuration and measurement loop

Each end device followed a deterministic one-minute cycle: **(i)** it sampled the sensor suite, **(ii)** packed the readings into a compact 18-byte binary frame, and **(iii)** transmitted a single uplink. The frame carried temperature (14 bit), relative humidity (14 bit), barometric pressure (17 bit), carbon dioxide concentration (13 bit), particulate matter ($PM_{2.5}$, 17 bit), and a monotonic packet counter (32 bit), totaling 107 bit (≈ 14 bytes), with the remaining 37 bit reserved as padding and for future extensions. Uplinks were sent at a transmit power of 14 dBm in the EU868 band. We cycled the spreading factor (SF) from SF7 to SF10 to probe a practical robustness–airtime trade-off, since higher SF increases robustness at the cost of longer transmissions [2]. With a 125 kHz bandwidth (coding rate set to 4/5) and an 18-byte payload, using the standard LoRa time-on-air model with an header, cyclic redundancy check (CRC) enabled, and an 8-symbol preamble, the time-on-air remains bounded (approximately 51 ms at SF7 and 330 ms at SF10). This supports the 60 s reporting interval under EU868 duty-cycle constraints [60], and reduces self-induced congestion.

3.1.5 Data pipeline and storage

As shown in Fig. 1(b) (adopted from [6, 59]), uplinks arriving at the TTN server were decoded using an application-level JavaScript payload formatter and streamed via Message Queuing Telemetry Transport (MQTT) (v3.1.1). A subscriber running on an Amazon Web Services Elastic Compute Cloud (AWS EC2) instance ingested the MQTT

stream and persisted both decoded sensor fields and TTN metadata into an InfluxDB (v1.8.9) time-series database. The pipeline operated continuously from October 2024 to September 2025, yielding a dense, time-aligned record of environmental conditions and link state for subsequent modeling and analysis.

3.2 Path loss modeling

Path loss modeling in LoRaWAN is a link-budget exercise, because meeting a target reliability means the transmit setting (e.g., SF) must cover the mean propagation loss plus a safety headroom. We index the end device to the gateway link by ℓ and the received packets by i . Let $L_\ell(d, f, \dots)$ denote the (random) path loss on link ℓ under propagation conditions (d, f, \dots) . In link-budget form, successful reception requires that Eq. (10) holds:

$$P_{\text{tx}} \geq L_{\text{tx}} - G_{\text{tx}} + L_\ell(d, f, \dots) - G_{\text{rx}} + L_{\text{rx}} + M_{\text{L}} + S(\text{SF}, \text{BW}), \quad (10)$$

where P_{tx} is the transmit power, G_{tx} and G_{rx} are antenna gains, L_{tx} and L_{rx} are feeder/connector losses, $S(\text{SF}, \text{BW})$ is the receiver sensitivity, and M_{L} is the total link margin. [8]. All power units are in dBm, while all losses and margins are in dB. In LoRaWAN's adaptive data rate (ADR), this headroom is implemented via the SNR-margin rule, with a fixed fade margin constant (often set to 10 dB). We keep the same principle but replace the fixed hedge by a data-driven fade margin inferred from the upper tail of the OOF residuals after fitting $\hat{L}(\cdot)$ (formalized in Sec. 3.4). Note that in this work, we interpret the link margin as $M_{\text{L}} = M_{\text{F}} + M_{\text{I}}$, where M_{F} is the calibrated fade margin and M_{I} collects any additional implementation or interference margins. In our evaluation, we set $M_{\text{I}} = 0$ to isolate the propagation-driven component. To anchor the analysis to our dataset, we construct the experimental path loss used as the regression response directly from TTN metadata by treating the gateway-reported RSSI as a received-power estimate. Let $P_{\text{rx},i}$ denote the gateway-reported received power for packet i (reported as RSSI in TTN metadata). Equation (11) gives the per-packet path loss:

$$L_{\ell,i} = P_{\text{tx}} - L_{\text{tx}} + G_{\text{tx}} + G_{\text{rx}} - L_{\text{rx}} - P_{\text{rx},i}. \quad (11)$$

While COST 231 MWM (log-distance) [15] captures geometry and obstruction losses, it does not represent the occupied-building dynamics that drive indoor non-stationarity, so leaving these effects in the residual inflates the margin required for high reliability. We therefore augment it with an environmental state vector and SNR as a link-state indicator to capture indoor non-stationarity. Let $\hat{L}_{\ell,i}$ denote a model-based mean prediction for $L_{\ell,i}$ for packet i at distance d (m) relative to a reference distance $d_0 = 1$ m, and let f denote the carrier frequency (MHz). Let W_k be the number of penetrated obstructions of wall type $k \in \{1, \dots, K\}$ (with loss ω_k in dB per wall), so that $\mathbf{w} = (W_1, \dots, W_K)^\top$ (in our site, $\mathbf{w} = (W_{\text{brick}}, W_{\text{wood}})^\top$). Let E_j be the j -th environmental covariate ($j \in \{1, \dots, P\}$), stacked as $\mathbf{e} = (E_1, \dots, E_P)^\top$, and let γ denote the gateway-reported SNR (in dB). The additive linear mean model is given in Eq. (12) as,

$$L_{\ell,i} = \underbrace{\beta_0 + 10n \log_{10}\left(\frac{d_i}{d_0}\right) + 20 \log_{10}(f_i) + \sum_{k=1}^K \omega_k W_{k,i} + \sum_{j=1}^P \varepsilon_j E_{j,i} + k_\gamma \gamma_i + \psi_i}_{\hat{L}_{\ell,i}}, \quad (12)$$

where β_0 is the intercept and n is the path loss exponent, while the term $20 \log_{10}(f)$ accounts for frequency-dependent free-space loss [61]. In our single-band EU868 deployment, it is constant and is absorbed into β_0 during fitting. The coefficient ε_j weights the environmental covariates E_j (for readability in later tables, we use the following mnemonic subscripts for the coefficients: ε_C for CO₂, ε_{RH} for relative humidity, ε_{PM} for PM_{2.5}, ε_{BP} for barometric pressure, and ε_T for temperature). The factor k_γ scales the link-state term.

To allow mild curvature and context coupling while preserving the additive multi-wallstructure, we extend Eq. (12) with a restricted second-order polynomial applied only to the continuous predictors. Define $z_d \equiv 10 \log_{10}(d/d_0)$ and $\mathbf{u} \equiv [z_d, \mathbf{e}^\top, \gamma]^\top \in \mathbb{R}^q$, where $q = P + 2$. Walls remain purely additive (no wall squares and no wall interactions). Under this restriction, the polynomial model (Eq. (13)) is written as:

$$L_{\ell,i} = \underbrace{\widehat{L}_{\ell,i}}_{\widehat{L}_{\ell,\text{poly},i}} + \Delta_{\text{poly}}(\mathbf{u}_i) + \psi_i, \quad \Delta_{\text{poly}}(\mathbf{u}_i) = \sum_{a=1}^q \tilde{h}_{aa} u_{a,i}^2 + \sum_{a=1}^{q-1} \sum_{b=a+1}^q \tilde{h}_{ab} u_{a,i} u_{b,i}, \quad (13)$$

where $\{\tilde{h}_{ab}\}_{1 \leq a < b \leq q}$ are unique quadratic coefficients in one-to-one correspondence with the monomials $\{u_a^2\}$ and $\{u_a u_b\}_{a < b}$ among the continuous predictors, and where u_a denotes the a -th component of \mathbf{u} with indices $a, b \in \{1, \dots, q\}$. Equivalently, $\Delta_{\text{poly}}(\mathbf{u}) = \mathbf{u}^\top H \mathbf{u}$ with $H = H^\top$ (a symmetric coefficient matrix), where $h_{aa} = \tilde{h}_{aa}$ and $h_{ab} = \tilde{h}_{ab}/2$ for $a < b$. This introduces quadratic terms among the continuous predictors (e.g., z_d^2 , E_j^2 , γ^2 , $z_d E_j$, $z_d \gamma$, $E_j \gamma$, and $E_j E_k$ for $j \neq k$), while wall terms remain additive. For $q = 7$ continuous predictors (distance, five environmental variables, and γ), the quadratic correction contributes $q(q+1)/2 = 28$ additional regressors. Together with the q linear continuous terms already present in Eq. (12), this yields $q(q+3)/2 = 35$ continuous regressors (excluding the intercept term β_0). Adding the two wall-count terms gives 37 non-intercept regressors in total.

In the data preparation pipeline shown in Fig. 3, raw measurements, comprising LoRaWAN-reported metadata and environmental variables, are retrieved from the InfluxDB database, cleaned upstream (including de-duplication, SF filtering, and outlier handling), and then sorted globally by timestamp. We retain SF7–SF10 to balance sensitivity and airtime [2]. We remove measurement outliers using an Isolation Forest (contamination of 0.01) trained on predictor-space features (excluding the response) as a data-quality screening step targeting sensor/ingestion faults. Because this procedure is unsupervised and does not use the response, it is applied prior to modeling and does not tune any model parameters. For model fitting, we use a chronological fixed 80:20 train/test split as deployed in [3]. Within the training subset, we use 5-fold time-ordered cross-validation using scikit-learn's `TimeSeriesSplit` (`gap=0`). Each validation fold is a contiguous time window that follows its corresponding training window (i.e., a forward-chaining/expanding-window protocol), ensuring that no future samples appear in training for any fold. We reuse identical fold assignments across all model families, so comparisons are made on the same validation windows. Predictors include distance d , frequency f , wall predictors \mathbf{W} , environmental covariates \mathbf{E} , and SNR γ . We linearize

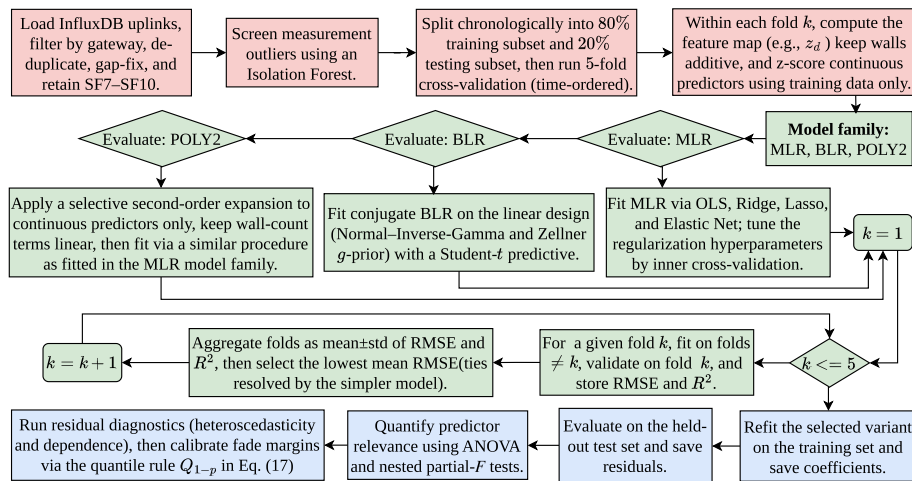


Fig. 3 Model fitting pipeline. In three phases: (i) database retrieval and cleaning, (ii) time-blocked cross-validation with within-fold preprocessing and model fitting, and (iii) model selection and hold-out evaluation, followed by analysis of variance (ANOVA), residual diagnostics, and fade margin calibration

distance as $10 \log_{10}(d/d_0)$ and treat $20 \log_{10}(f)$ as a fixed offset by subtracting it from the response during training (and adding it back for evaluation in the path loss domain).

Model comparison follows three families under the same time-ordered cross-validation protocol: (i) a linear MLR in Eq. (12) and its regularized variants fit on the linearized design (distance as $10 \log_{10}(d/d_0)$ and the frequency term $20 \log_{10}(f)$). Since all links are at 868 MHz, the frequency term is a constant and is absorbed via the intercept. In multi-frequency settings, one would reintroduce the $20 \log_{10}(f)$ dependence; (ii) POLY2, a second-order polynomial in Eq. (13) applied to the continuous predictors $\{z_d, E_1, \dots, E_p, \gamma\}$ while wall counts W_k remain additive and linear (also fitted with the regularization specifications); (iii) a BLR on the linear design, reported for both a conjugate NIG prior and a Zellner g -prior, providing calibrated predictive uncertainty (Student- t predictive) without changing the mean feature set. Out-of-sample performance is evaluated using RMSE and R^2 on the held-out test set. In contrast, residuals are evaluated with partial F tests and *Type II* and *III* ANOVA, conditional on residual diagnostics (normality, independence, and variance stability).

Estimation and leakage-safe preprocessing are as follows: for penalized MLR (and for the BLR workflow, where applicable), standardization is fit within each fold using only the fold’s training indices; for POLY2, the selective quadratic expansion is applied first, followed by within-fold standardization, so that scaling reflects only the training data. Hyperparameters are selected by grid search on the same $K = 5$ cross-validation folds, choosing the configuration that minimizes mean validation RMSE. We sweep Ridge $\lambda \in \text{logspace}(-4, 3)$ (15 points), Lasso $\lambda \in \text{logspace}(-4, 1)$ (15 points), and ElasticNet $\lambda \in \text{logspace}(-4, 1)$ (10 points) with mix $\alpha \in \{0.2, 0.5, 0.8\}$. For BLR, we select prior settings by the same cross-validation criterion (NIG prior scale $\Sigma_0 \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$ with $a_0 = b_0 = 10^{-2}$, and Zellner g -prior mode in `uip` (unit-information, $g = N$), `eb`

(empirical Bayes that selects g by maximizing the marginal likelihood on the training fold.) with $a_0 = b_0 = 10^{-2}$) [25]. After tuning, each selected model is refit on the full training split and evaluated once on the held-out test set. OOF residuals from the cross-validation stage are exported and used downstream for residual diagnostics and fade margin calibration (Section 3.4).

3.3 Shadow fading diagnostics

Our prior study on the same site reported clear deviations from Normal residuals and heavier right tails [7] (based on MLR). Motivated by those findings, without presupposing outcomes, we treat the residual law as an empirical object to be validated and repeat the same procedures. All steps below use the OOF residuals from the final mean specification (the second-order polynomial; Section 3.2), so distributional checks and any tail modeling reflect true generalization.

We fit five families to the OOF residuals: Normal, Student's t , Skew-Normal, Cauchy, and finite GMM with $K \in \{1, \dots, 5\}$ components. Parameters are estimated by maximum likelihood. For GMMs, we use the Expectation–Maximization algorithm with multiple random initializations, an identifiability constraint that orders component means for reporting, and a small variance floor to ensure numerical stability. Model choice follows a pre-specified rule to avoid hindsight bias: **(i)** compute BIC for each fitted family (and K for GMMs); **(ii)** select the minimum-BIC model; **(iii)** break practical ties using the KS distance; **(iv)** when BIC or the KS distance are essentially indistinguishable, prefer the simpler family. Classical omnibus tests (D'Agostino–Pearson; Jarque–Bera) are used only as diagnostics to motivate heavier-tailed/asymmetric families when warranted, and hence they do not drive selection.

For residuals $\{r_i\}_{i=1}^N$ and a candidate parametric density $f(r; \theta)$, we fit θ by maximum likelihood by maximizing the log-likelihood $\mathcal{L}(\theta)$ in Eq. (14). We then compute AIC and BIC from the maximized log-likelihood using the standard complexity penalties ($2m$ for AIC and $m \ln(N)$ for BIC), where m is the number of estimated distribution parameters and N is the sample size. Finally, we apply the KS test by comparing each fitted CDF with the empirical CDF of the residuals, where a smaller KS statistic indicates closer agreement over the full support.

$$\mathcal{L}(\theta) = \sum_{i=1}^N \log(f(r_i; \theta)). \quad (14)$$

Since parametric fits can mask structure when the family is misspecified, we add a model-agnostic KDE to visualize shape (core vs. tails) and to check for latent modality without imposing a functional form. We compute a Gaussian-kernel KDE on a fixed grid via Fast Fourier Transform (FFT) convolution (exact on the grid and computationally stable) and overlay two bandwidths: Silverman's rule of thumb and a cross-validated log-likelihood bandwidth. We also run Hartigan's dip test and Silverman's critical-bandwidth test to summarize modality [44]. However, the KDE is not used to produce tail quantiles; it is strictly a diagnostic to corroborate (or challenge) the parametric choice.

To check stability, mixture fits use multiple random initializations and a small variance floor, and spurious tiny-weight components are merged or discarded by the

Expectation–Maximization stopping rule. Seeds are fixed for reproducibility. In the results, we report the log-likelihood, AIC, BIC, and KS test distance for each candidate, along with Q–Q plots and KDE overlays. The distribution selected by the rule above is carried forward to the tail-quantile step in Sec. 3.4, where the construction of fade margins and their uncertainty is defined.

3.4 Fade margin calibration

We calibrate a fade margin that maps predictive uncertainty into a reliability buffer in the link budget of Eq. (10). For each sample i , let $\widehat{L}_{\ell,i}$ denote the OOF mean path loss prediction by the regressor trained without sample i (equivalently, within each fold $\widehat{\mathbf{L}} = \widehat{\beta}_0 \mathbf{1} + \Phi \widehat{\beta}$). The corresponding OOF error is $r_i = L_{\ell,i} - \widehat{L}_{\ell,i}$. With a fade margin M_F , an outage occurs when the realized loss exceeds the predicted loss plus buffer ($L_{\ell,i} > \widehat{L}_{\ell,i} + M_F$), i.e., $r_i > M_F$. Thus, prescribing a target outage probability p reduces to selecting M_F as an upper-tail quantile of the error distribution, consistent with classical outage-based margin setting (Q-function formulation under log-normal shadowing) [63]. This yields the empirical fade margin estimator in Eq. (15):

Define the (conditional) non-exceedance probability as $\rho(M_F) \triangleq \Pr(r \leq M_F)$ and the corresponding outage (exceedance) probability as $p(M_F) \triangleq \Pr(r > M_F) = 1 - \rho(M_F)$. In the held-out validation, we report \widehat{p} and its complement $\widehat{\rho} = 1 - \widehat{p}$ computed on the held-out residuals.

$$\widehat{M}_{F,\text{emp}}(p) = Q_{1-p}(\{r_i\}_{i=1}^N), \quad p \in (0, 1) \quad (15)$$

where $Q_{1-p}(\cdot)$ is the $(1-p)$ -quantile of the cross-validated residuals. When residuals are heavy-tailed or multimodal, we model the far tail with a 3-component GMM, consistent with the residual diagnostics in Sec. 4.3.1. To guard against under-budgeting in the far tail, we define the conservative fade margin estimator $\widehat{M}_F(p)$ as the maximum of the empirical and mixture-based tail quantiles. For $p \leq 0.02$ we also evaluate the $(1-p)$ -quantile of the fitted 3-component GMM and prescribe the conservative margin via Eq. (16):

$$\widehat{M}_F(p) = \max\{\widehat{M}_{F,\text{emp}}(p), \widehat{M}_{F,\text{GMM}}(p)\}. \quad (16)$$

This conservative estimator preserves the quantile-based reliability mapping of Eq. (15), while using a mixture tail model to protect the far tail [64]. Uncertainty of the empirical estimator is quantified using BCa bootstrap intervals. When short-memory dependence is detected, we use a moving-block bootstrap with the block length selected by ACF/PACF diagnostics (aggregated across devices). For mixture-based margins, we report parametric-bootstrap CIs obtained by sampling from the fitted 3-component GMM and recomputing Q_{1-p} . We also summarize fold-to-fold dispersion across cross-validation splits.

Calibration is validated on a held-out test set. Let N_{test} be the number of held-out test samples, and let $\mathbb{1}\{\cdot\}$ denote the indicator function (equal to 1 if its condition holds and 0 otherwise), with index $i \in \{1, \dots, N_{\text{test}}\}$. For a prescribed $\widehat{M}_F(p)$, we compute the achieved outage using Eq. (17):

$$\hat{p} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}\{L_{\ell,i} > \hat{L}_{\ell,i} + \hat{M}_F(p)\}. \quad (17)$$

Since $L_{\ell,i}$ is observed only for received packets (Eq. (11)), the reported achieved reliability is $\hat{\rho} = 1 - \hat{p}$ ($\Pr(r \leq \hat{M}_F(p) \mid \text{received})$), i.e., a conditional non-exceedance on received packets. Sweeping $p \in \{0.05, 0.02, 0.01\}$, we plot the achieved reliability $\hat{\rho}$ against the prescribed margin and overlay target iso-lines, where alignment indicates reliability-correct calibration on unseen data. Finally, we report $M_{F,99} \equiv M_F(0.01)$ with 95% CIs, compare against a 10 dB fixed heuristic margin, adopted from [65], and include achieved conditional reliability on received packets $\hat{\rho}$ versus the fade margin calibration visualization.

4 Results and analysis

This section follows the pipeline in Fig. 3 to: (i) report fitting predictive performance, (ii) interpret the fitted mean, (iii) summarize residual diagnostics, and (iv) use them to support fade margin calibration with held-out validation.

4.1 Regression

From Table 5, the empirical COST 231 MWM, included as a structural-only reference (distance and wall obstructions), exhibits substantially weaker generalization (RMSE ≈ 12.07 dB, $R^2 \approx 0.59$), indicating that structural descriptors alone underfit the variability present in the measurements. The extended linear baselines (MLR and BLR) are essentially indistinguishable under the time-blocked protocol, achieving a cross-validated RMSE of ≈ 8.24 dB with a test RMSE of ≈ 8.47 dB and $R^2 \approx 0.80$ for both cases. This stability suggests the linear regime is well-conditioned for the available predictors, and that any multicollinearity among environmental covariates does not materially affect predictive accuracy [28]. While BLR does not improve point-error metrics relative to MLR, it retains practical value by providing calibrated posterior uncertainty (e.g., coefficient credible intervals), which is useful when link variability is modulated by occupancy and ventilation dynamics. In contrast, the second-order polynomial substantially improves generalization, reducing the cross-validated RMSE to 7.3761 dB ($\approx 10.4\%$) and, respectively, improving R^2 to 0.8439, with consistent gains on the held-out test. This

Table 5 Comparative performance across the model families (RMSE and coefficient of determination (R^2)). We report the best configuration for each model family (OLS/Ridge/Lasso/ElasticNet), selected by the OOF mean RMSE with \pm denoting standard deviation across folds. Within-family OOF RMSE ranges are as follows: (i) COST 231 MWM ≤ 0.01 dB, (ii) MLR ≤ 0.01 dB, (iii) BLR ≈ 0 dB, and (iv) POLY2 ≤ 0.09 dB

Model Family	RMSE (dB)		R^2	
	Cross-validated (OOF)	Test (held-out)	Cross-validated (OOF)	Test (held-out)
COST 231 [15]	10.9739 \pm 1.0518	12.0704	0.6533	0.5890
MLR	8.2333 \pm 0.5577	8.4845	0.8061	0.7969
BLR	8.2416 \pm 0.5827	8.4534	0.8055	0.7984
POLY2	7.3761 \pm 0.6048	7.7708	0.8439	0.8296

indicates that second-order interaction terms capture nonlinear dependencies not captured by purely linear models, thereby improving on the MLR baseline.

For Table 6, we report the OLS coefficients as a representative linear fit together with 95% confidence intervals computed using the MacKinnon–White heteroskedasticity-consistent covariance estimator (HC3; leverage-adjusted) to quantify coefficient uncertainty under potential heteroskedasticity. Across the frequentist linear variants, Ridge and Lasso yield near-identical estimates to OLS, with differences at most 0.01 for the dominant geometric terms (path loss exponent and wall losses) and around 10^{-3} or smaller for the environmental slopes. Likewise, the two BLR specifications yield the same point estimates to numerical precision, consistent with a data-dominant likelihood. Their practical advantage is calibrated posterior uncertainty (e.g., coefficient credible intervals) rather than lower point-error. ElasticNet can induce stronger shrinkage on the correlated environmental block, effectively redistributing weights among covariates without materially changing predictive performance (Table 5). For interpretability, we therefore summarize the linear regime using the OLS coefficients.

The fitted intercept reflects fixed offsets (including free space path loss ($f, 1$ m) at 868 MHz and centering), so we do not interpret β_0 physically. The path loss exponent ($n \approx 3.87$) is consistent with established indoor LoRaWAN ranges and reflects rapid decay in a multi-obstruction office environment [10, 63]. The estimated obstruction losses show the expected material contrast, with brick inducing about 6.83 dB per wall compared with about 1.98 dB for wooden partitions, in line with reported attenuation differences for building materials around 868 MHz [66]. Moreover, the environmental terms retain the same sign pattern, whereas the SNR scaling was strongly negative ($k_\gamma \approx -2.03$). In contrast to typical outdoor regression settings, where meteorological variables are often treated as direct propagation modifiers [3], indoors, these quantities should be interpreted primarily as conditional state proxies rather than causal attenuation terms. Elevated CO₂ and humidity co-vary with occupancy and HVAC scheduling, i.e., regimes in which the indoor channel is time-varying due to human motion, door usage, and airflow control. Consequently, their fitted signs reflect a correlation with

Table 6 Predictor coefficients for the linear model (OLS) with HC3 confidence intervals (CIs). HC3 refers to the MacKinnon–White heteroskedasticity-consistent covariance estimator

Variable	Unit	Coefficient	Estimate	95% CIs (HC3)(dB)	
				lower	upper
Intercept	dB	β_0	2.31	1.80	2.81
Path loss exponent	–	n	3.87	3.86	3.88
Brick wall loss	dB	ω_{brick}	6.83	6.81	6.85
Wood partition loss	dB	ω_{wood}	1.98	1.96	1.99
Carbon dioxide	dB/ppm	ε_C	–0.0024	–0.0025	–0.0022
Relative humidity	dB/%	ε_{RH}	–0.0917	–0.0938	–0.0897
Particulate matter	dB/($\mu\text{g}/\text{m}^3$)	ε_{PM}	–0.0953	–0.1011	–0.0895
Barometric pressure	dB/hPa	ε_{BP}	–0.0080	–0.0094	–0.0066
Temperature	dB/°C	ε_T	–0.1410	–0.1443	–0.1378
SNR scaling	–	k_γ	–2.0344	–2.0426	–2.0262

building-operation states that shift the conditional link budget after controlling for distance and wall counts, rather than radiative absorption in the strict physical sense [2]. The SNR term is expected to be strongly informative because it acts as an instantaneous demodulation-margin indicator tied to received power and noise; including it captures fast channel and interference conditions and reduces residual variance, but it should not be interpreted as an independent physical loss mechanism [67]. Collectively, the environmental block functions as an implicit low-dimensional channel-state indicator, supporting environment-aware reliability modeling for indoor IoT operation at 868 MHz.

For the best-performing nonlinear model (POLY2, Lasso-selected), the intercept absorbs fixed offsets as well as centering and standardization effects induced by the polynomial expansion; accordingly, we do not attach direct physical meaning to it. The distance dependence is captured primarily by curvature in the distance term, while the linear distance contribution is strongly regularized. We verified numerically that the net gradient $\partial \hat{\mathcal{L}}_{\ell, \text{poly}} / \partial z_d$ remains positive across the evaluated distance range when conditioned on the observed covariates. Wall-loss effects remain physically consistent with material-dependent attenuation, in line with sub-GHz measurement evidence [68]. Most first-order environmental slopes (e.g., ε_C , ε_{RH}) are shrunk toward zero under regularization, whereas a small set of interaction terms involving distance, humidity, and link-state indicators remains nonzero. This pattern supports a conditional channel-state interpretation that the retained environment–distance and environment–SNR couplings likely act as proxies for occupancy/HVAC-regime changes and other latent dynamics rather than direct atmospheric absorption at 868 MHz. A first-order term with mild curvature dominates link-state dependence. Finally, removing the SNR predictor increases RMSE while preserving the qualitative role of the retained environmental interactions.

4.2 Analysis of variance (ANOVA)

For consistency across models, we report *Type II* and *Type III* ANOVA-style F tests for the additive MLR baseline using the same HC3-robust covariance estimator as in Table 6, and compare nested predictor blocks via classical OLS residual sum of squares partial- F tests (structure, then +environmental variables, then +SNR). Since the sample size is very large ($N \approx 1.66 \times 10^6$), the associated p -values become numerically saturated (often effectively 0). Thus, statistical significance here should be read as detectability rather than practical magnitude [38]. We therefore interpret effects primarily via partial η^2 (an F -derived partial effect-size index that coincides with the usual ANOVA partial-variance share under homoskedastic OLS), which is bounded and comparable across predictors. The resulting relative ordering is summarized in Table 7, from which we conclude that: **(i)** Per-predictor (Panel A): Structural terms dominate, with the path loss exponent contributing most (partial $\eta^2 \approx 0.21$), followed by brick-wall loss. The SNR term is of comparable magnitude (partial $\eta^2 \approx 0.12$) and exceeds the wood-partition term, while the environmental covariates are individually modest but directionally consistent once structure is controlled. The ranking and signs are stable under *Type III* (discussed next). **(ii)** Block gains (Panel B): In Panel B, adding the five environmental covariates after structure yields a statistically detectable but modest reduction in residual sum of squares (partial $\eta^2 \approx 0.0098$), whereas adding SNR afterward delivers the dominant improvement (partial $\eta^2 \approx 0.4307$), indicating that SNR carries substantial

Table 7 Additive MLR significance tests. Residual degrees of freedom (df) = 1663617; partial $\eta^2 = \frac{F df_1}{F df_1 + df_2}$, where df_1 and df_2 are the numerator and denominator df. (Panel A: $df_1 = 1$). HC3: MacKinnon-White heteroskedasticity-consistent covariance estimator.

Panel A: HC3-robust Type II ANOVA-style per-predictor tests						
Variable	Unit	Coefficient	F	p	partial η^2	Sign
Path loss exponent	–	n	439732.636	$< 10^{-16}$	0.2091	+
Brick wall loss	dB	ω_{brick}	318053.020	$< 10^{-16}$	0.1605	+
Wood partition loss	dB	ω_{wood}	55431.275	$< 10^{-16}$	0.0322	+
Carbon dioxide	dB/ppm	ϵ_C	1841.148	$< 10^{-16}$	0.0011	–
Relative humidity	dB/%	ϵ_{RH}	7743.697	$< 10^{-16}$	0.0046	–
Particulate matter	dB/($\mu\text{g}/\text{m}^3$)	ϵ_{PM}	1025.772	$< 10^{-16}$	0.0006	–
Barometric pressure	dB/hPa	ϵ_{BP}	126.342	$< 10^{-16}$	0.0001	–
Temperature	dB/°C	ϵ_T	7228.963	$< 10^{-16}$	0.0043	–
SNR	–	k_γ	236234.977	$< 10^{-16}$	0.1243	–

Panel B: Nested block comparisons (OLS residual sum of squares partial-F)				
Comparison	df ₁ , df ₂	F	p	partial η^2
Structure → +Environmentals	(5, 1663618)	3298.886	$< 10^{-16}$	0.0098
+Environmentals → +SNR	(1, 1663617)	1258588.250	$< 10^{-16}$	0.4307

non-redundant link-state information beyond structure and microclimate. This is consistent with the gap between COST 231 MWM and MLR results in Table 5. Since the MLR specification includes only main effects (no interactions), *Type III* ANOVA yields the same tests as *Type II* for all predictors, and hence the effect-size ordering is insensitive to ANOVA type. *Type III* ANOVA additionally reports an intercept test, which we do not interpret physically because it absorbs fixed offsets (e.g., the constant frequency term).

To dissect predictor contributions in the polynomial model, we conducted partial-*F* tests on nested blocks denoted as: *W* (walls), *D* (distance), *E* (environment-only), *S* (SNR), and *X* (all interactions of z_d , γ and \mathbf{e} , excluding squares) (Table 8). In nested additions, distance provides a strong initial gain (partial $\eta^2 = 0.181$), SNR contributes the largest incremental improvement (partial $\eta^2 = 0.514$), and interactions add a further non-trivial component (partial $\eta^2 = 0.102$). In contrast, the environment-only block is modest in isolation (partial $\eta^2 = 0.008$). In the drop-one tests, we remove the indicated term block and compare the reduced model to the full model. Removing walls or SNR incurs the steepest penalties (partial $\eta^2 = 0.193$ and 0.147 , respectively), confirming the primacy of structural attenuation and instantaneous link state. Omitting interactions also causes a measurable loss (partial $\eta^2 = 0.102$), indicating that nonlinear couplings refine fidelity beyond additive effects. Therefore, the environment block contributes most meaningfully when expressed through interactions rather than as purely additive slopes, supporting environment-aware reliability tuning without sacrificing coverage.

Table 8 Partial-F tests. Left: nested block additions, and right: drop-one (second-order polynomial; train set)

Block	Nested additions				Drop-one versus full			
	Δdf	F	p	η^2	Dropped	Δdf	F	η^2
$D(z_d, z_d^2)$	2	184141	$< 10^{-16}$	0.181	$W(\mathbf{w})$	2	198617	0.193
$E(\mathbf{e})$	10	1346	$< 10^{-16}$	0.008	$D(z_d, z_d^2)$	2	21039	0.025
$S(\gamma, \gamma^2)$	2	878961	$< 10^{-16}$	0.514	$E(\mathbf{e})$	10	307	0.002
X (cross-terms in \mathbf{u})	21	8959	$< 10^{-16}$	0.102	$S(\gamma, \gamma^2)$	2	143520	0.147
					X (cross-terms in \mathbf{u})	21	8959	0.102

Table 9 Distributional fit diagnostics on the second-order polynomial regression residuals

Distribution	Log-Likelihood	AIC	BIC	KS test
Normal	-5669116.41	11338236.83	11338261.48	0.0603
Skew-Normal	-5627994.19	11255994.38	11256031.35	0.0483
Gaussian Mixture (K=3)	-5513071.13	11026158.25	11026256.85	0.0108
Cauchy	-5729734.65	11459473.30	11459497.95	0.0816
t -Distribution	-5537405.70	11074817.41	11074854.38	0.0279

4.3 Shadow fading analysis

4.3.1 Model Fit Diagnostics

Residual diagnostics for the second-order polynomial model reject Normality decisively (Omnibus/D’Agostino statistic $\approx 3.49 \times 10^5$ and Jarque–Bera statistic $\approx 3.08 \times 10^6$, both $p < 10^{-16}$), with a right-skewed and heavy-tailed shape (skewness ≈ 0.765 , excess kurtosis ≈ 6.488). Autocorrelation is not a dominant concern at this resolution (Durbin–Watson ≈ 2.022). Even with a strong predictive fit (OOF $R^2 \approx 0.844$), a single Gaussian law is inadequate, consistent with our earlier baseline MLR analysis [7]. As summarized in Table 9, a three-component Gaussian mixture provides the best residual description, achieving the lowest BIC and the smallest KS statistic. Skew-Normal improves over Normal but still under-represents the right tail, while the t distribution offers an intermediate fit. This residual-tail characterization forms the basis of the fade margin calibration in Sec. 3.4.

4.3.2 Parametric Residual Distributions

Table 9 reports the KS statistics and information criteria for the candidate families. Due to the large value of N , KS p values saturate, so we compare KS statistics and BIC. Table 9 and Fig. 4 support the following observations: (i) Normal underestimates tail mass, with $KS \approx 0.0603$ (Fig. 4(a)). (ii) Skew–Normal captures asymmetry better (shape ≈ 1.60) but still deviates in the extreme quantiles ($KS \approx 0.0483$; Fig. 4(b)). (iii) A Gaussian mixture with $K=3$ achieves the best parsimony–fit trade-off (lowest BIC $\approx 1.1026 \times 10^7$; $KS \approx 0.0108$) and closely tracks both the core and the tails (Fig. 4(c)). The fitted mixture

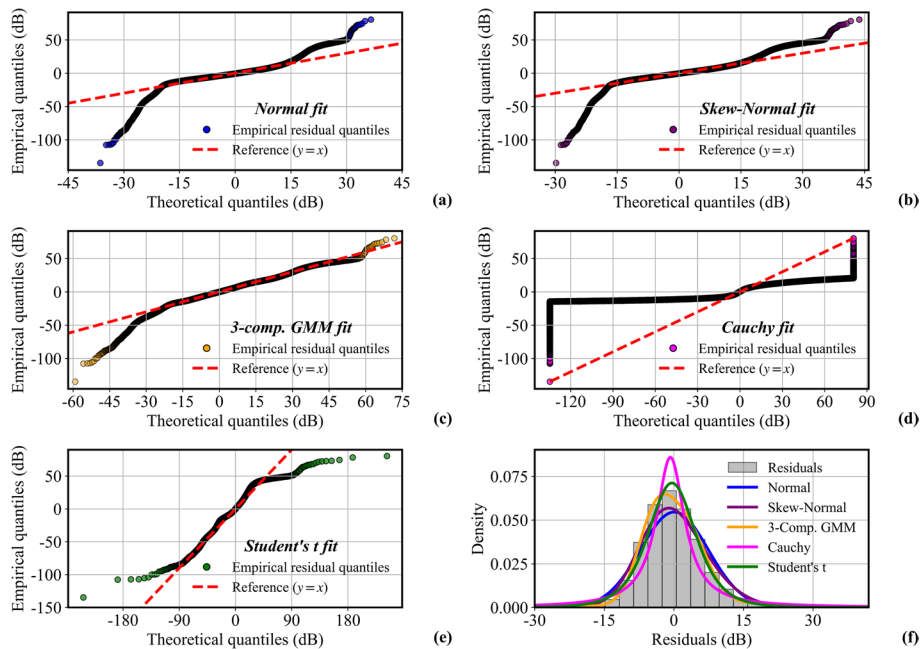


Fig. 4 Parametric residual distribution diagnostics. Q-Q plots of the second-order polynomial OOF residuals against candidate parametric families in (a)–(e), and the residual histogram with fitted densities in (f). The selected three-component GMM matches both the sharp core and heavy tails, with two dominant narrow components and a lower-weight broad component

comprises two dominant narrow components and a lower-weight broad component, consistent with heterogeneous indoor shadowing states [69]. (iv) Cauchy over-weights extremes and misrepresents the bulk ($KS \approx 0.0816$; Fig. 4(d)). (v) Student’s t thickens tails (degrees of freedom $\nu \approx 4.29$) and improves fit relative to Normal, but remains intermediate overall ($KS \approx 0.0279$; Fig. 4(e)). Compared with MLR-based site analysis in [7] that favored a higher mixture order, the polynomial mean reduces residual heterogeneity to $K=3$, indicating that mild nonlinear couplings are absorbed. At the same time, a compact multimodal shadow-fading structure remains. This selected residual law is carried forward to the fade margin calibration in Sec. 3.4.

4.3.3 Nonparametric density and modality

As shown in Fig. 5(a), both KDE views indicate a sharp central mass with a mild shoulder and a slightly heavier right tail. The narrower bandwidth ($h \approx 0.307$ dB) reveals local structure, whereas the broader bandwidth ($h \approx 0.975$ dB) merges minor bumps into an effectively unimodal core [44]. Formal modality diagnostics are consistent with this picture. Hartigan’s dip test rejects strict unimodality at the native scale ($p < 10^{-16}$), but the bandwidth required to enforce unimodality is modest since Silverman’s critical bandwidth is $h^* \approx 0.466$ dB (smoothed-bootstrap $p \approx 1.00$ for $H_0: \leq 1$ mode). The mode-count curve in Fig. 5(b) mirrors this behavior, with many spurious peaks at very small h that rapidly collapse toward a single mode as smoothing increases, reaching unimodality around $h \sim 0.5$ dB. Together with the BIC-selected $K=3$ GMM (Table 9), the evidence supports a near-unimodal core with a low-probability heavy-tail regime, motivating

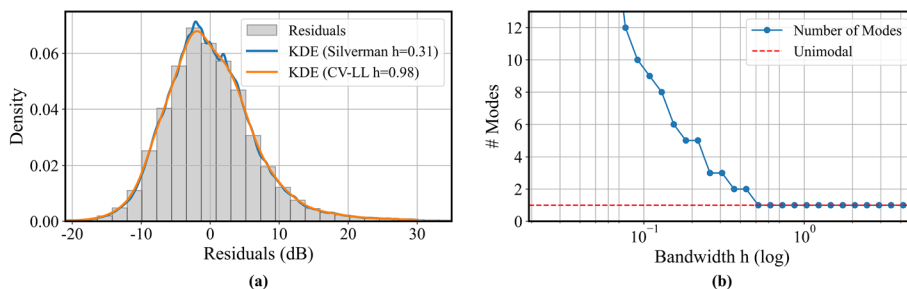


Fig. 5 Nonparametric residual density and modality diagnostics. (a) Residual density with kernel density estimation (KDE) overlays using Silverman’s rule ($h \approx 0.307$ dB) and a cross-validated log-likelihood (CV-LL) bandwidth ($h \approx 0.975$ dB). (b) Estimated mode count versus bandwidth on a logarithmic sweep using prominence-based peak detection

Table 10 Group-wise OOF residual location and dispersion diagnostics. MAD: median absolute deviation ($\sigma \approx 1.4826$ MAD under Gaussian residuals) [70]. ϵ_{KW}^2 : Kruskal-Wallis effect size (location/ranks). η_{BF}^2 : Brown-Forsythe effect size (dispersion; median-centered Levene)

Partition	Group	N_{grp}	Median (dB)	MAD (dB)	ϵ_{KW}^2	η_{BF}^2
LoS vs NLoS	LoS	277295	-1.98	4.14	0.0111	0.0007
	NLoS	1386332	-0.27	3.97		
CO ₂ terciles	Low	557015	-0.82	3.95	0.0008	0.0001
	Mid	543064	-0.30	4.06		
	High	563548	-0.57	4.06		

mixture-based tail calibration for fade margins while using KDE strictly as a diagnostic view of residual shape.

We also tested whether coarse context partitions explain the residual heterogeneity using full-sample Kruskal–Wallis (location) and Brown–Forsythe (scale) tests (Table 10). Although differences are statistically detectable at this sample size, the corresponding effect sizes are small: LoS/NLoS accounts for only about 1% of rank-based location variability and well below 0.1% of dispersion, while CO₂ terciles are smaller still. These partitions, therefore, explain little of the residual geometry, so the mixture-protected tail treatment remains justified for far-tail fade margin calibration rather than being an artifact of stratifying by LoS/NLoS or CO₂ alone.

4.4 Fade margin calibration

Following the procedure in Sec. 3.4, we estimate $M_F(p)$ from the residuals and validate calibration on held-out data via the achieved reliability (computed on received packets only (conditional non-exceedance)) at the prescribed $\hat{M}_F(p)$. Table 11 reports the calibrated margins, bootstrap 95% confidence intervals, and held-out achieved. As shown in Fig. 6 (a), the calibrated operating points lie close to the target iso-lines, and the held-out achieved reliability is near the nominal levels ($\leq \sim 1.1\%$ absolute across all targets including 95%, and $\leq \sim 0.5\%$ for the stricter outage targets ($p \leq 2\%$)), which is a considerably reliable calibration under our proposed protocol. Fig. 6 (b) summarizes the comparison as fade margin differences $\Delta M_F(p)$, where positive values mean a model requires more margin. The linear baselines (BLR/MLR) are essentially tied (within ≈ 0.3 dB), while

Table 11 Fade margin calibration from out-of-fold (OOF) residuals and held-out validation.

The table summarizes the calibrated fade margin M_F for target outage levels 100p(%), based on either empirical quantiles or GMM-tail estimation, alongside 95% confidence intervals and the achieved held-out reliability $\hat{\rho}$

Model	100p(%)	Estimator	M_F (dB)	95% CIs (dB)		Achieved reliability $\hat{\rho}$
				lower	upper	
COST 231 MWM	5	Empirical	12.09	12.08	12.15	0.9544
COST 231 MWM	2	Empirical	44.61	44.40	44.61	0.9705
COST 231 MWM	1	Empirical	60.76	60.75	61.07	0.9898
BLR	5	Empirical	13.18	13.12	13.24	0.9565
BLR	2	GMM tail	21.86	21.76	21.96	0.9810
BLR	1	Empirical	27.79	27.40	28.13	0.9866
MLR	5	Empirical	12.99	12.93	13.05	0.9559
MLR	2	GMM tail	21.59	21.48	21.70	0.9801
MLR	1	Empirical	28.05	27.65	28.44	0.9865
POLY2	5	Empirical	11.96	11.89	12.01	0.9608
POLY2	2	GMM tail	19.65	19.55	19.76	0.9793
POLY2	1	GMM tail	25.73	25.61	25.85	0.9856

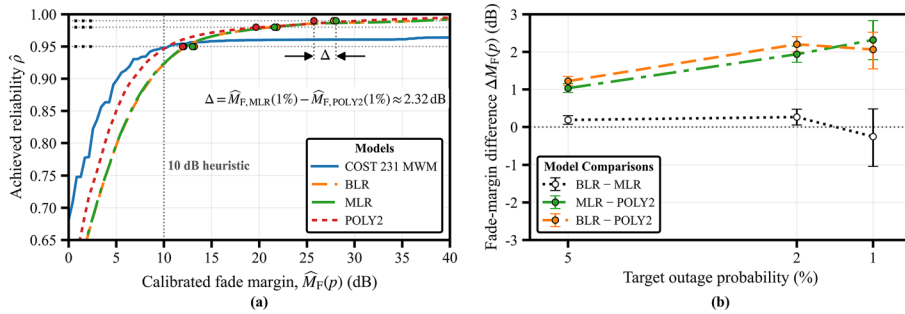


Fig. 6 Fade margin calibration on held-out test set. (a) Held-out achieved reliability versus fade margin with calibrated operating points $\hat{M}_F(p)$ for $p \in \{5, 2, 1\}$ % and 95% confidence intervals. The COST 231 MWM is shown for context, but its calibrated margins for $p \leq 2\%$ exceed the plotted range (Table 11). (b) Fade margin differences $\Delta M_F(p)$ ($M_{F,A}(p) - M_{F,B}(p)$) among BLR/MLR/POLY2 models at matched outage targets (COST 231 MWM omitted for scale)

POLY2 steadily lowers the required margin, with the benefit growing as p tightens. At $p = 1\%$, the reduction is ≈ 2.32 dB, roughly one SF step (≈ 3 dB transmit power). Operationally, this saving is comparable to the link-budget difference between adjacent SFs (order of a few dB), and thus meaningful for airtime/energy trade-offs [60]. A fixed 10 dB heuristic undershoots in this environment (held-out achieved reliability ≈ 0.923 – 0.947), motivating data-driven margins for reliability planning. Note that all fade margins here are calibrated from residuals of models that include SNR. If SNR is omitted at planning time, the same pipeline applies, but the calibrated $M_F(p)$ increases slightly, an ablation that we treat as part of our future work.

4.5 Discussion

A summary of answers to our research questions (RQ1–RQ4) (listed in Table 1) is as follows. First, with respect to RQ1, environment-aware covariates and the

receiver-reported SNR contribute beyond geometry and structure. HC3 *Type II/III* ANOVA yields a coherent effect-size ordering in which distance and wall losses dominate (e.g., partial $\eta^2 \approx 0.21$ for n and ≈ 0.16 for L_{brick}), while SNR remains substantial (partial $\eta^2 \approx 0.12$). Environmental terms are individually smaller yet directionally consistent. Nested block tests further confirm that adding the environmental block after structure and adding SNR afterward each deliver large, non-redundant gains (Table 7). Second, addressing **RQ2**, model comparison under the same chronological 80 : 20 hold-out and 5-fold time-ordered cross-validation shows that linear MLR (including regularized variants) and BLR are effectively tied in point accuracy, whereas the selective second-order polynomial applied only to continuous predictors provides the best bias–variance trade-off, improving generalization (RMSE from ≈ 8.24 dB to 7.38 dB and R^2 from 0.81 to 0.84, with consistent held-out performance; Table 5). Third, in response to **RQ3**, OOF residuals depart strongly from Gaussianity, and a compact $K=3$ Gaussian mixture yields the best fit among the tested families by BIC/KS (Table 9, Fig. 4), consistent with heterogeneous indoor regimes. Finally, for **RQ4**, converting residual tails into a deployment control, prescribing fade margins from upper-tail OOF quantiles calibrates reliably on the held-out set and yields an approximately 2.1 – 2.3 dB margin saving at 99% reliability for the selective quadratic mean relative to linear baselines (Table 11, Fig. 6), translating improved mean modeling and residual characterization into a concrete link-budget reduction.

5 Conclusion

This work develops an environment-aware indoor LoRaWAN path loss workflow that remains interpretable (distance and wall losses remain explicit) while accounting for indoor non-stationarity using measured environmental covariates and a receiver-side link indicator (SNR). Using a 12-month office deployment at 868 MHz, we evaluate linear MLR (with regularized variants), Bayesian linear regression, and a selective second-order polynomial extension applied only to continuous predictors, under a chronological 80 : 20 hold-out and 5-fold time-ordered cross-validation with within-fold preprocessing. The selective polynomial model improves generalization relative to linear baselines (cross-validated RMSE drops from about 8.24 dB to about 7.38 dB and cross-validated R^2 rises from about 0.81 to about 0.84, with consistent held-out performance). OOF residual diagnostics show clear departures from Gaussianity, and a compact three-component Gaussian mixture captures the sharp core plus a light broad tail. Turning this residual behavior into a deployment control, we prescribe fade margins from the upper-tail quantiles of OOF residuals and validate on the held-out test set. At 99% delivery reliability, the polynomial model requires about 25.7 dB versus roughly 27.8–28.1 dB for linear baselines, giving an approximately 2.3 dB margin saving at fixed reliability.

6 Limitations and future directions

This study is site-calibrated: measurements were collected on a single floor within one building, using six fixed end device to the gateway links, so fitted wall-loss parameters and regression coefficients should be interpreted as deployment-specific rather than universally transferable. Environmental covariates are co-located with the end devices and

serve as practical proxies for indoor state and thus their generalization across building classes, sensor placements, and HVAC/occupancy regimes remains to be established. Finally, path loss is reconstructed from gateway-reported received-power metadata for successfully received packets. Therefore, packets that are not received do not provide received-power metadata and are therefore not directly represented in the path loss residual analysis or in the tail-based margin calibration.

Next, we will prioritize external validity and measurable state context rather than adding more model families. We will replicate the campaign across multiple buildings and floor-plan types, then use hierarchical (multi-level) analyses to separate within-site effects from between-site variability in wall losses and environmental coefficients, enabling portable priors. We will also link residual regimes to observable building state (e.g., HVAC telemetry and coarse occupancy proxies) instead of treating mixture components as purely phenomenological. On the deployment side, we will extend the same pipeline to planning-mode operation, where receiver-side link indicators may be unavailable, by refitting without such terms and recalibrating fade margins. We will then integrate the calibrated margins into LoRaWAN ADR and transmit-power control to quantify reliability–energy trade-offs, and evaluate lightweight geometry-aware features from floor plans (e.g., corridor-path descriptors) that reduce tail risk without sacrificing interpretability. Finally, we will standardize reporting via shared splits, consistent residual-law selection rules, and bootstrap uncertainty for tail quantiles to support fair, reproducible comparisons across indoor LoRaWAN studies.

Abbreviations

AIC	Akaike information criterion
ANOVA	Analysis of variance
BCa	Bias-corrected and accelerated (bootstrap)
BIC	Bayesian information criterion
BLR	Bayesian linear regression
CI	Confidence interval
df	Degrees of freedom
GMM	Gaussian mixture model
HC3	Heteroskedasticity-consistent covariance estimator (MacKinnon–White HC3)
HVAC	Heating, ventilation, and air conditioning
IoT	Internet of Things
KDE	Kernel density estimation
KS	Kolmogorov–Smirnov (test)
Lasso	Least absolute shrinkage and selection operator
LoRaWAN	Long-range wide area network
LoS	Line-of-sight
MLR	Multiple linear regression
MWM	Multi-wall model
NIG	Normal–inverse-gamma (prior)
NLoS	Non-line-of-sight
OOF	Out-of-fold
OLS	Ordinary least squares
POLY2	Selective second-order polynomial mean model (continuous predictors only)
RMSE	Root mean square error
RSSI	Received signal strength indicator
SF	Spreading factor
SNR	Signal-to-noise ratio

Acknowledgements

The authors thank the Ubiquitous Computing Group at the University of Siegen for support with the indoor deployment and infrastructure, and for helpful discussions during the year-long campaign.

Author Contributions

N.M.O. designed and executed the measurement campaign; implemented data ingestion, cleaning, and feature engineering; developed and evaluated the regression models (MLR, regularized variants, Bayesian linear regression,

and second-order polynomial); performed ANOVA and residual diagnostics; carried out the fade-margin calibration; prepared figures and tables; and drafted the manuscript. **K.V.L.** conceived the study; supervised the methodology and experimental design; contributed to the interpretation of results and to writing and critical revision of the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was partly supported by the German Academic Exchange Service (DAAD) through the Kenyan–German Postgraduate Training Programme 2023/2024 under Grant 57652455.

Additional funding was provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant 425,868,829 as part of Priority Program SPP2199: Scalable Interaction Paradigms for Pervasive Computing Environments.

Data Availability

The dataset and analysis scripts supporting this study are publicly available at: <https://github.com/nahshonmokua/LoRaWAN-Indoor-PL-parametrics>. The repository includes code to reproduce the data preparation, model fitting, cross-validation, residual diagnostics, and fade-margin calibration reported in the manuscript. Any additional information is available from the corresponding author upon reasonable request.

Declarations

Ethical Approval and Consent to Participate

Not applicable. The study involves environmental sensing and wireless link measurements. No personal or identifiable data, headcounts, or movement traces were collected.

Consent for Publication

Not applicable.

Conflict of interest

The authors declare that they have no conflict of interest.

Materials Availability

Not applicable. No new unique biological or chemical materials were generated. Commercial, off-the-shelf hardware is listed in the manuscript; build details can be provided on request.

Code Availability

The analysis and plotting code are available in the same public repository listed above under “Availability of data and materials.”

Received: 7 October 2025 Accepted: 9 March 2026

Published online: 13 May 2026

References

1. M. Cattani, C.A. Boano, K. Römer, An experimental evaluation of the reliability of LoRa long-range low-power wireless communication. *J. Sens. Actuator Netw.* **6**(2), 7 (2017). <https://doi.org/10.3390/jsan6020007>
2. J. Grübel, T. Thrash, L. Aguilar, M. Gath-Morad, D. Hélaï, R.W. Sumner, C. Hölscher, V.R. Schinazi, Dense indoor sensor networks: Towards passively sensing human presence with LoRaWAN. *Pervasive Mob. Comput.* **84**, 101640 (2022). <https://doi.org/10.1016/j.pmcj.2022.101640>
3. M. González-Palacio, D. Tobón-Vallejo, L.M. Sepúlveda-Cano, S. Rúa, G. Pau, L.B. Le, LoRaWAN path loss measurements in an urban scenario including environmental effects. *Data* **8**(1), 4 (2023). <https://doi.org/10.3390/data8010004>
4. M.N.A. Siddiky, M.E. Rahman, M.S. Uzza, H.M.D. Kabir, A comprehensive exploration of 6G wireless communication technologies. *Computers* **14**(1), 15 (2025). <https://doi.org/10.3390/computers14010015>
5. D. Szafranski, Predictability of LoRaWAN Link Quality based on Weather Data: Insights from a Long-Term Study. In: 2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 249–258 (2024). <https://doi.org/10.1109/WoWMoM60985.2024.00048>
6. N.M. Obiri, K. Van Laerhoven, A comprehensive data description for LoRaWAN path loss measurements in an indoor office setting: effects of environmental factors. *IEEE Access* **13**, 83148–83170 (2025). <https://doi.org/10.1109/ACCESS.2025.3569164>
7. N.M. Obiri, K. Van Laerhoven, A Statistical Evaluation of Indoor LoRaWAN Environment-Aware Propagation for 6G: MLR, ANOVA, and Residual Distribution Analysis. In: 2025 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), pp. 494–499 (2025). <https://doi.org/10.1109/EuCNC/6GSummit63408.2025.11037092>
8. M. González-Palacio, D. Tobón-Vallejo, L.M. Sepúlveda-Cano, S. Rúa, L.B. Le, Machine-learning-based combined path loss and shadowing model in LoRaWAN for energy efficiency enhancement. *IEEE Internet Things J.* **10**(12), 10725–10739 (2023). <https://doi.org/10.1109/JIOT.2023.3239827>
9. N.M. Obiri, K. Van Laerhoven, A survey of LoRaWAN-integrated wearable sensor networks for human activity recognition: applications, challenges and possible solutions. *IEEE Open J. Commun. Soc.* **5**, 6713–6735 (2024). <https://doi.org/10.1109/OJCOMS.2024.3484002>

10. J.A. Azevedo, F. Mendonça, A critical review of the propagation models employed in LoRa systems. *Sensors* **24**(12), 3877 (2024). <https://doi.org/10.3390/s24123877>
11. R. Robles-Enciso, I.P. Morales-Aragón, A. Serna-Sabater, M.T. Martínez-Inglés, A. Mateo-Aroca, J.-M. Molina-García-Pardo, L. Juan-Llácer, LoRa, Zigbee and 5G propagation and transmission performance in an indoor environment at 868 MHz. *Sensors* **23**(6), 3283 (2023). <https://doi.org/10.3390/s23063283>
12. I. Anisah, Wirawan, Suwadi, M. Yuliana, Experimental Results of LoRa Network Radio Propagation Modeling in Campus Area. In: 2023 6th International Seminar on Research of Information Technology and Intelligent Systems (ISRIT), pp. 427–432 (2023). <https://doi.org/10.1109/ISRIT60336.2023.10467989>
13. S. Bertoldo, M. Paredes, L. Carosso, M. Allegretti, P. Savi, Empirical indoor propagation models for LoRa radio link in an office environment. In: 2019 13th European Conference on Antennas and Propagation (EuCAP), pp. 1–5 (2019)
14. International Telecommunication Union, *Radiocommunication Sector (ITU-R): Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 450 GHz* (Technical report, International Telecommunication Union, Geneva, Switzerland, 2021)
15. European Commission: Digital mobile radio towards future generation systems. Final Report, European Commission, Directorate-General for the Information Society and Media, Luxembourg (1999)
16. A.G.M. Lima, L.F. Menezes, Motley-Keenan model adjusted to the thickness of the wall. In: SBMO/IEEE MTT-S International Conference on Microwave and Optoelectronics, 2005., pp. 180–182 (2005). <https://doi.org/10.1109/MOC.2005.1580040>
17. C. Zhong, Measurement and modeling of LoRa signal in multi-floor home environment. *Commun. Netw.* **17**(1), 1–19 (2025). <https://doi.org/10.4236/cn.2025.171001>
18. E.H. Alkhamzi, S.M. Elkawafi, A.A. Aldarrat, M.A. Abbas, H. Abubakr, H.A. Shamatah, Analysis of Real-World LoRaWAN Network Performance Across Outdoor and Indoor Scenarios. In: 2023 IEEE 11th International Conference on Systems and Control (ICSC), pp. 329–334 (2023). <https://doi.org/10.1109/ICSC58660.2023.10449775>
19. R. Muppala, A. Navnit, S. Poondla, A.M. Hussain, Investigation of Indoor LoRaWAN Signal Propagation for Real-World Applications. In: 2021 6th International Conference for Convergence in Technology (I2CT), pp. 1–5 (2021). <https://doi.org/10.1109/I2CT51068.2021.9418173>
20. E. Harinda, A.J. Wixted, A.-U.-H. Qureshi, H. Larjani, R.M. Gibson, Performance of a live multi-gateway LoRaWAN and interference measurement across indoor and outdoor localities. *Computers* **11**(2), 25 (2022). <https://doi.org/10.3390/computers11020025>
21. S. Sadowski, P. Spachos, RSSI-based indoor localization with the internet of things. *IEEE Access* **6**, 30149–30161 (2018). <https://doi.org/10.1109/ACCESS.2018.2843325>
22. R.R. Guerra, A. Vizziello, P. Savazzi, E. Goldoni, P. Gamba, Forecasting LoRaWAN RSSI using weather parameters: A comparative study of ARIMA, artificial intelligence and hybrid approaches. *Comput. Netw.* **243**, 110258 (2024). <https://doi.org/10.1016/j.comnet.2024.110258>
23. A. Aksoy, Ö. Yildiz, S.E. Karlık, Comparative analysis of end device and field test device measurements for RSSI, SNR and SF performance parameters in an indoor LoRaWAN network. *Wireless Pers. Commun.* **134**(1), 339–360 (2024). <https://doi.org/10.1007/s11277-024-10911-z>
24. H. Vo, V.H.L. Nguyen, V.L. Tran, F. Ferrero, F.-Y. Lee, M.-H. Tsai, Advance path loss model for distance estimation using LoRaWAN network's received signal strength indicator (RSSI). *IEEE Access* **12**, 83205–83216 (2024). <https://doi.org/10.1109/ACCESS.2024.3412849>
25. J.J. Faraway, *Practical Regression and Anova Using R* (University of Bath, Bath, UK, 2002)
26. H. Akaike, A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**(6), 716–723 (1974). <https://doi.org/10.1109/TAC.1974.1100705>
27. G. Schwarz, Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978). <https://doi.org/10.1214/aos/1176344136>
28. A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, And TensorFlow, 2nd edition (updated for tensorflow 2) edn* (O'Reilly Media Inc, Sebastopol, CA, 2019)
29. A. Khalajmehrabadi, N. Gatsis, D.J. Pack, D. Akopian, A joint indoor WLAN localization and outlier detection scheme using LASSO and elastic-net optimization techniques. *IEEE Trans. Mob. Comput.* **16**(8), 2079–2092 (2017). <https://doi.org/10.1109/TMC.2016.2616465>
30. B.P.R. Bhavanam, P. Ragam, Exploring LoRa Signal Propagation in Indoor and Outdoor Environments: A Comparative Study. In: 2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE), pp. 1–6 (2024). <https://doi.org/10.1109/ICSPCRE62303.2024.10675046>
31. B. Lin, J. Chen, B. Xu, J. Chao, B. Zheng, G. Pang, J. Luo, Z. Ghassemlooy, Indoor NLOS-VLP system based on image sensor and pixel coordinate fingerprinting. *IEEE Internet Things J.* **12**(12), 20251–20260 (2025). <https://doi.org/10.1109/IJOT.2025.3542370>
32. F. Liang, R. Paulo, G. Molina, M.A. Clyde, J.O. Berger, Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.* **103**(481), 410–423 (2008). <https://doi.org/10.1198/016214507000001337>
33. A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, *Bayesian Data Analysis*, 3rd edn. (CRC Press, New York, 2013)
34. A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**(5), 1413–1432 (2017). <https://doi.org/10.1007/s11222-016-9696-4>
35. M.K. Elmezughi, O. Salih, T.J. Afullo, K.J. Duffy, Comparative analysis of major machine-learning-based path loss models for enclosed indoor channels. *Sensors* **22**(13), 4967 (2022). <https://doi.org/10.3390/s22134967>
36. S. Hosseinzadeh, M. Almoathen, H. Larjani, K. Curtis, A neural network propagation model for LoRaWAN and critical analysis with real-world measurements. *Big Data Cogn. Comput.* **1**(1), 7 (2017). <https://doi.org/10.3390/bdcc1010007>
37. D.C. Montgomery, *Design and Analysis of Experiments* (John Wiley & Sons, Hoboken, New Jersey, 2017)

38. Ø. Langsrud, ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Stat. Comput.* **13**(2), 163–167 (2003). <https://doi.org/10.1023/A:1023260610025>
39. S.A. Chelloug, Impact of the temperature and humidity variations on link quality of xm1000 mote sensors. *Int. J. Ad hoc, Sensor Ubiquitous Comput.* **5**(6), 21–29 (2014). <https://doi.org/10.5121/ijasuc.2014.5603>. arXiv:1501.01073 [cs]
40. D. Christmann, I. Martinovic, Experimental design and analysis of transmission properties in an indoor wireless sensor network. In: 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, And Wireless Networks, pp. 342–347 (2010)
41. H. Harb, A. Makhoul, R. Couturier, An enhanced K-means and ANOVA-based clustering approach for similarity aggregation in underwater wireless sensor networks. *IEEE Sens. J.* **15**(10), 5483–5493 (2015). <https://doi.org/10.1109/JSEN.2015.2443380>
42. B. Allen, S. Mahato, Y. Gao, S. Salous, Indoor-to-outdoor empirical path loss modelling for femtocell networks at 0.9, 2, 2.5 and 3.5 GHz using singular value decomposition. *IET Microw. Antennas Propag.* **11**(9), 1203–1211 (2017). <https://doi.org/10.1049/iet-map.2016.0416>
43. Y. Wang, W. Ren, L. Cheng, J. Zou, A grey model and mixture Gaussian residual analysis-based position estimator in an indoor environment. *Sensors* **20**(14), 3941 (2020). <https://doi.org/10.3390/s20143941>
44. B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. Routledge, New York (2018). <https://doi.org/10.1201/9781315140919>
45. O.L.O. Astivia, B.D. Zumbo, Heteroskedasticity in Multiple Regression Analysis: What it is, How to Detect it and How to Solve it with Applications in R and SPSS. *Practical Assessment, Research & Evaluation* **24**(1) (2019)
46. Y. Chan, R.P. Walmsley, Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Phys. Ther.* **77**(12), 1755–1761 (1997). <https://doi.org/10.1093/ptj/77.12.1755>
47. L. Zhang, S.L. Cotton, S.K. Yoo, H.Q. Ngo, M. Fernández, W.G. Scanlon, A time series-based study of correlation, channel power imbalance, and diversity gain in indoor distributed antenna systems at 60 GHz. *IEEE Trans. Antennas Propag.* **69**(11), 7782–7794 (2021). <https://doi.org/10.1109/TAP.2021.3076171>
48. B. Efron, Bootstrap methods: another look at the Jackknife. *Ann. Stat.* **7**(1), 1–26 (1979). <https://doi.org/10.1214/aos/1176344552>
49. E.D. Ayele, C. Hakkenberg, J.P. Meijers, K. Zhang, N. Meratnia, P.J.M. Havinga, Performance analysis of LoRa radio for an indoor IoT applications. In: 2017 International Conference on Internet of Things for the Global Community (IoTGC), pp. 1–8 (2017). <https://doi.org/10.1109/IoTGC.2017.8008973>
50. J. Petäjärvi, K. Mikhaylov, R. Yasmin, M. Hämäläinen, J. Iinatti, Evaluation of LoRa LPWAN technology for indoor remote health and wellbeing monitoring. *Int. J. Wireless Inf. Netw.* **24**(2), 153–165 (2017). <https://doi.org/10.1007/s10776-017-0341-8>
51. M.M. Erbaty, G. Schiele, G. Batke, Analysis of LoRaWAN technology in an Outdoor and an Indoor Scenario in Duisburg-Germany. In: 2018 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 273–277 (2018). <https://doi.org/10.1109/CCOMS.2018.8463224>
52. R. El Chall, S. Lahoud, M. El Helou, LoRaWAN network: radio propagation models and performance evaluation in various environments in Lebanon. *IEEE Internet Things J.* **6**(2), 2366–2378 (2019). <https://doi.org/10.1109/JIOT.2019.2906838>
53. M.I. Muzammir, H.Z. Abidin, S.A.C. Abdullah, F.H.K. Zaman, Performance Analysis of LoRaWAN for Indoor Application. In: 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE), pp. 156–159 (2019). <https://doi.org/10.1109/ISCAIE.2019.8743982>
54. M. Saban, O. Aghzout, L.D. Medus, A. Rosado, Experimental analysis of IoT networks based on LoRa/LoRaWAN under indoor and outdoor environments: performance and limitations. *IFAC-PapersOnLine* **54**(4), 159–164 (2021). <https://doi.org/10.1016/j.ifacol.2021.10.027>
55. A. Azzalini, A. Capitanio, *The Skew-Normal and Related Families* (Cambridge University Press, Cambridge, 2014)
56. N.L. Johnson, S. Kotz, N. Balakrishnan, *Continuous Univariate Distributions*, vol. 2 (John Wiley & Sons, New York, USA, 1995)
57. F.J. Massey Jr., The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**(253), 68–78 (1951). <https://doi.org/10.1080/01621459.1951.10500769>
58. D. Reynolds, Gaussian Mixture Models. In: Li, S.Z., Jain, A. (eds.) *Encyclopedia of Biometrics*, pp. 659–663. Springer, Boston, MA (2009). https://doi.org/10.1007/978-0-387-73003-5_196
59. N.M. Obiri, H. Shikunzi, Long-range wide area network (LoRa-WAN) connectivity and range evaluation in a rural setting. *Int. J. Computer Appl.* **185**(3), 61–67 (2023). <https://doi.org/10.5120/ijca2023922699>
60. B.S. Chaudhari, M. Zennaro, (eds.): *LPWAN Technologies for IoT and M2M Applications*, Cambridge, MA, United States (2020). <https://doi.org/10.1016/C2018-0-04787-8>
61. H.T. Friis, A note on a simple transmission formula. *Proc. IRE* **34**(5), 254–256 (1946). <https://doi.org/10.1109/JRPROC.1946.234568>
62. G.M. Bianco, R. Giuliano, G. Marrocco, F. Mazzenga, A. Mejia-Aguilar, LoRa system for search and rescue: path-loss models and procedures in mountain scenarios. *IEEE Internet Things J.* **8**(3), 1985–1999 (2021). <https://doi.org/10.1109/JIOT.2020.3017044>
63. A. Goldsmith, *Wireless Communications* (Cambridge University Press, Cambridge, United Kingdom, 2005)
64. E.N. Papsotiriou, A.-A.A. Boulageorgos, A. Alexiou, Outdoor THz fading modeling by means of gaussian and gamma mixture distributions. *Sci. Rep.* **13**(1), 6385 (2023). <https://doi.org/10.1038/s41598-023-33598-x>
65. C.H. Barriquello, F.E.S. Silva, D.P. Bernardon, L.N. Canha, M.J.D.S. Ramos, D.S. Porto, Fundamentals of Wireless Communication Link Design for Networked Robotics. In: Neves, António J. R. (ed.) *Service Robots*, pp. 127–142. IntechOpen, London, United Kingdom (2018). <https://doi.org/10.5772/intechopen.69873>
66. R. Rudd, K. Craig, M. Ganley, R. Hartless, *Building Materials and Propagation*. Final Report 2604/BMEM/R/3/2.0, Ofcom, United Kingdom (2014)
67. T.S. Rappaport, *Wireless Communications: Principles and Practice* (Prentice Hall Communications Engineering and Emerging Technologies Series. Prentice Hall, Upper Saddle River, N.J., 2002)

68. M.T. Sebastian, *Dielectric Materials for Wireless Communication* (Elsevier, Amsterdam, Netherlands, 2010)
69. Z. Yao, P. Chang, Q. Zhu, W. Sun, A hierarchical positioning model for WiFi-based indoor localization in large-scale complex environments. *Intell. Robot.* **5**(3), 745–63 (2025). <https://doi.org/10.20517/ir.2025.38>
70. P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* **88**(424), 1273–1283 (1993). <https://doi.org/10.1080/01621459.1993.10476408>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.