

Article

Probabilistic Remaining Useful Life Estimation for Buried Pipeline Pitting Corrosion via Mechanics-Regularized Limit-State Learning

Haipeng Liu ^{1,2}, Yuntao Shi ², Long Chen ², Haotian Wei ³ , Shaohua Dong ¹ and Yinuo Chen ^{4,*}

¹ College of Safety and Ocean Engineering, China University of Petroleum (Beijing), Beijing 102249, China; shdong@cup.edu.cn (S.D.)

² China National Petroleum Corporation International Pipeline Company, Beijing 102206, China

³ College of Mechanical and Transportation Engineering, China University of Petroleum, Beijing 102249, China; weihaotian@cup.edu.cn

⁴ School of Engineering and Technology, China University of Geosciences (Beijing), Beijing 100083, China

* Correspondence: chenyn@cugb.edu.cn

Abstract

Buried steel pipelines are susceptible to external pitting corrosion whose spatially heterogeneous and time-dependent nature makes probabilistic remaining useful life (RUL) estimation both necessary and difficult. This study develops a three-stage framework comprising statistical pit depth characterization, a proxy corrosion-growth learner, and a mechanics-regularized limit-state RUL model anchored to the modified B31G burst criterion and an auxiliary conservative depth-screen rule, in which a corrosion-growth prior learned from sparse field measurements is embedded directly into the learning objective as a regularizer rather than merely used to construct training labels, and conformal calibration is applied at both the proxy and the limit-state stages to ensure honest empirical coverage, with the mechanics-regularized limit-state model consistently achieving coverage close to or above the nominal 90% target. Applied to the Velázquez excavation dataset and further examined on a 127-sample engineering database used as an auxiliary transferability check, the framework delivered stable point accuracy across all three wall thickness scenarios while maintaining empirical 90% coverage consistently above the nominal target. These results demonstrate that embedding mechanics-based growth constraints into the learning objective improves predictive consistency under sparse field data, while the resulting reliability outputs are best interpreted as scenario-based screening evidence for comparative integrity prioritization rather than as literal asset-life certification.

Keywords: pitting corrosion; buried pipelines; remaining useful life; probabilistic prediction; conformal calibration; modified B31G; reliability-based inspection



Academic Editor: Zhibin Lin

Received: 22 April 2026

Revised: 4 June 2026

Accepted: 8 June 2026

Published: 17 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Buried transmission pipelines are long-lived infrastructure assets whose safe operation depends on the ability to detect, quantify, and anticipate localized corrosion before leakage or rupture occurs. External pitting corrosion poses a particular challenge because its governing processes span soil chemistry, moisture transport, coating disbondment, cathodic protection effectiveness, and stochastic electrochemical heterogeneity [1–4]. The resulting defect morphology is highly variable in both space and time, and the engineering consequences are nonlinear: a modest change in pit depth or growth rate can materially

alter burst resistance, structural reliability, and the urgency of inspection. Effective integrity management therefore requires more than a snapshot assessment of current damage; it demands a probabilistic characterization of how rapidly defects may evolve and when failure-relevant thresholds are likely to be reached.

The buried pipeline corrosion literature provides an important foundation for this purpose. The field investigation of Velázquez et al. remains one of the most informative excavation-based datasets, as it links measured pit depth with service age, soil physico-chemical indicators, coating condition, and local pipe-to-soil characteristics [3]. Building on that foundation, subsequent studies have advanced the statistical description of underground pitting through Monte Carlo analysis of pit depth and pit rate distributions, Markov-chain representations of pit evolution, corrosion rate-based reliability calculations, and probabilistic modeling at inspection time [5–9]. These contributions clarified how soil and operational variables collectively shape corrosion severity, yet most of them concentrated on characterizing the present defect geometry or projecting future corrosion growth rather than directly forecasting remaining life against explicit structural failure criteria.

A separate and substantial body of work has focused on the remaining strength and reliability assessment of corroded pipelines. The modified B31G standard and related assessment models remain central to engineering practice, and numerous investigations have scrutinized their underlying assumptions, quantified model error, assessed analytical consistency, and benchmarked them against alternative burst-pressure formulations [10–15]. Reliability-oriented studies have gone further by incorporating time-dependent corrosion growth, Monte Carlo simulation, probabilistic failure-pressure analysis, and system reliability accounting for multiple defects [16–22]. Nevertheless, these mechanically oriented studies tend to treat corrosion growth prediction and structural failure assessment as independent modules, and many retain deterministic growth assumptions or scenario simplifications that prevent the full propagation of predictive uncertainty in a decision-ready format.

Recent pipeline corrosion studies also show a clear shift toward data-driven and hybrid modeling. Recent contributions include an interpretable ensemble-learning framework for internal corrosion rate prediction, a KPCA–PSO–ELM model for buried pipeline external corrosion under scarce data, a multi-stage remaining-life prediction framework for corrosion damage in oil and gas pipeline systems, and a time-variant reliability analysis highlighting the sensitivity of failure estimates to corrosion growth and defect aspect ratio assumptions [23–26]. These developments confirm the timeliness of probabilistic learning for pipeline integrity, but they also reinforce the need to state clearly which parts of a workflow are data-driven, which are mechanics-driven, and which are scenario reconstructions imposed because key field measurements are unavailable.

In parallel, the broader prognostics and health-management community has made rapid strides in remaining useful life (RUL) prediction, condition-based maintenance, and uncertainty-aware inference [27–30]. Recent contributions in *Processes* illustrate the range of current practice, including uncertainty-integrated aero-engine prognostics, autoencoder–GRU architectures, robust relevance vector machines, and hybrid deep-learning models for battery RUL estimation [31–34]. These studies demonstrate the value of probabilistic learning and composite architectures, but their degradation physics, operating conditions, and failure modes differ substantially from buried pipeline corrosion, where soil-driven pit growth must ultimately be translated into limit-state failure risk under specific structural and pressure constraints.

At the algorithmic level, modern uncertainty-aware learning offers a rich and well-developed toolbox for this translation. Quantile regression and its extensions enable tail-sensitive modeling [35,36]; ensemble trees and boosting methods provide flexible

nonlinear prediction [37–40]; Gaussian processes and natural-gradient boosting support fully probabilistic forecasting [41,42]; and conditional quantile ensembles, calibration theory, conformal prediction, and explainability methods improve interval validity and interpretability [43–47]. Despite the maturity of each component, these tools have rarely been assembled into a unified pipeline that is simultaneously data-driven, mechanics-constrained, probabilistically calibrated, and directly interpretable in terms of buried pipeline structural failure criteria.

Four methodological gaps persist in the current literature and collectively motivate the present work. First, statistical characterization, probabilistic corrosion prediction, and remaining-life assessment are commonly reported as separate studies rather than stages of a coherent integrated workflow. Second, many corrosion-learning studies focus on point prediction and do not assess predictive uncertainty in a calibration-aware manner. Third, remaining-life investigations typically adopt a single failure criterion, even though integrity decisions are sensitive to both burst-pressure loss and severe wall thickness depletion occurring on different timescales. Fourth, theoretical corrosion-growth models are frequently employed only to generate proxy training labels, rather than being incorporated into the learning objective to serve as a physics-based regularizer. The contributions of this study address each gap in turn.

1. A three-stage workflow is established that connects statistical pit depth characterization, probabilistic pit depth prediction, and reliability-oriented RUL forecasting on the Velázquez field dataset.
2. A systematic benchmark of modern probabilistic predictors is provided for pit depth estimation, with simultaneous comparison of point accuracy, interval calibration quality, and Shapley additive explanations (SHAP)-based feature attribution.
3. A dual-criterion limit-state target is formulated by combining the modified B31G burst-resistance model with an auxiliary conservative depth-screen rule, so that violation of either condition constitutes failure.
4. A mechanics-regularized probabilistic RUL model is developed by embedding a theoretical power-law growth prior directly into the learning objective, followed by conformal interval calibration and translation of the resulting predictive distributions into failure-probability trajectories and inspection trigger schedules.

Prior buried pipeline studies have typically treated statistical corrosion characterization, probabilistic pit depth prediction, and structural reliability assessment as separate research activities, each reported in isolation. Growth models are commonly applied only to generate training labels rather than to constrain the learning objective itself, and most remaining-life investigations adopt a single failure criterion without propagating predictive uncertainty in a calibration-aware format. The present work addresses these gaps without claiming to introduce a new corrosion-growth law or a new burst-pressure equation. Instead, it makes three specific methodological advances over the existing literature. First, a probabilistic pit depth benchmark is conducted directly on excavation data, with simultaneous evaluation of point accuracy, interval calibration, and feature attribution across multiple modern learners. Second, the fitted corrosion-growth kinetics are embedded directly into the learning objective as a mechanics-anchored regularizer, rather than merely being used to construct training labels as in prior studies, so that the learned response surface is constrained toward physically plausible trajectories even when the field dataset is small and heterogeneous. Third, a dual-criterion failure definition combining burst-pressure loss and wall thickness depletion replaces the single-criterion assessments common in the literature, and conformal calibration is applied to ensure that the resulting uncertainty intervals carry honest empirical coverage. The resulting workflow is therefore

new in its coupling strategy and decision translation rather than in any individual statistical or mechanical component viewed in isolation.

2. Dataset and Preprocessing

2.1. Vel Ázquez Field Data

The source data originate from the excavation campaign of Velázquez and coauthors [3], in which 259 locations along buried pipelines in southern Mexico were characterized through measured pit depth, service age, soil physicochemical indicators, and coating and soil class descriptors. The cleaned analysis set retained 250 samples after removing the nine outliers already flagged in the source study. Table 1 summarizes the core data inventory used throughout the analysis.

Throughout this paper, $d_{\max}$ denotes the measured maximum pit depth. The coating codes are NC (no coating), AEC (asphalt enamel coating), WTC (wrap tape coating), CTC (coal-tar coating), and FBE (fusion-bonded epoxy). The soil texture codes follow the parsed source taxonomy: C (clay), SCL (sandy clay loam), CL (clay loam), SYCL (silty clay loam), SC (silty clay), and SL (silt loam).

Table 1. Core data inventory after preprocessing.

Statistic	Value
Original records	259
Clean records	250
Outliers removed	9
Mean $d_{\max}$	1.917 mm
Median $d_{\max}$	1.190 mm
Maximum $d_{\max}$	10.41 mm
Mean pit rate	0.0882 mm/year
Median service age	21.0 years
Most common coating	CTC
Most common soil class	C

As shown in Table 2, the cleaned dataset exhibits a certain degree of imbalance in coating types and soil categories. Specifically, CTC and WTC account for a relatively large proportion of the coating samples, while the soil types are mainly concentrated as clay, sandy clay loam, and clay loam. In contrast, categories such as AEC, SC, and SL contain relatively few samples. To reduce the influence of categorical imbalance on model interpretation, coating type and soil type were incorporated into the model through pooled low-dimensional encodings, together with continuous physicochemical variables for prediction. In addition, during model evaluation, error analysis and prediction interval coverage analysis were further conducted to examine whether obvious bias existed in model errors and uncertainty representation across different coating and soil categories.

The numeric variables comprise maximum pit depth $d_{\max}$, pit rate, service age, pH, pipe-to-soil potential, soil resistivity, water content, bulk density, chloride, bicarbonate, sulfate, and redox potential. The categorical variables are coating type and USDA-style soil texture class. Detection limit values reported as “<1.00” in the original source were replaced by 0.5 and tracked throughout analysis with dedicated indicator flags. To stabilize the heavily right-skewed distributions, chloride, bicarbonate, sulfate, resistivity, and redox variables were log-transformed prior to modeling.

Table 2. Categorical composition of the cleaned dataset.

Category	Code	Clean Count
Coating	CTC	110
Coating	WTC	95
Coating	NC	31
Coating	FBE	8
Coating	AEC	6
Soil class	C	107
Soil class	SCL	75
Soil class	CL	59
Soil class	SYCL	6
Soil class	SC	2
Soil class	SL	1

2.2. Engineered Features for Reliability-Oriented Modeling

Two progressively structured feature sets were employed. The first, used for pit depth and proxy-RUL learning, augmented the raw measurements with corrosion physics-inspired interaction terms :

$$\log t = \log(\text{age}), \quad (1)$$

$$\text{aggressiveness index} = \log(\text{chloride}) - \text{pH}, \quad (2)$$

$$\text{ionic strength proxy} = \log(1 + \text{chloride} + \text{bicarbonate} + \text{sulfate}), \quad (3)$$

$$\text{moisture-resistivity ratio} = \frac{\text{water content}}{\text{soil resistivity} + 10^{-6}}, \quad (4)$$

$$\text{coating-age interaction} = \text{age} (1 - \text{coating score}), \quad (5)$$

$$\text{depth-rate interaction} = d_{\max} \times \text{pit rate}. \quad (6)$$

The second feature set, reserved for the limit-state study, additionally incorporated wall thickness, pressure, and burst strength descriptors:

$$\text{remaining thickness} = t_w - d_{\max}, \quad (7)$$

$$\text{current margin} = P_{\text{burst},0} - P_{\text{op}}, \quad (8)$$

$$\text{current pressure ratio} = \frac{P_{\text{op}}}{P_{\text{burst},0}}, \quad (9)$$

$$\text{depth-to-thickness ratio} = \frac{d_{\max}}{t_w}, \quad (10)$$

$$\text{pressure-depth interaction} = P_{\text{op}} \frac{d_{\max}}{t_w}. \quad (11)$$

These engineered quantities are not merely cosmetic transformations of the raw inputs; they are designed to encode the physical relationships governing monotonic constraints and limit-state interpretation in the final mechanics-regularized model.

3. Statistical Characterization of Pit Depth

3.1. Distribution Fitting

The depth and pit rate distributions were screened against the lognormal, Fisk, gamma, and Weibull families. For each candidate family k with parameter vector θ_k and dimensionality p_k , maximum-likelihood estimation was applied as follows:

$$\hat{\theta}_k = \arg \max_{\theta_k} \ell_k(\theta_k; y_1, \dots, y_n), \quad (12)$$

and the information criteria used for ranking were

$$AIC_k = -2\ell_k(\hat{\theta}_k) + 2p_k, \quad (13)$$

$$BIC_k = -2\ell_k(\hat{\theta}_k) + p_k \log n. \quad (14)$$

Table 3 reports the ranking results. Among all candidate families, both $d_{\max}$ and pit rate are best described by the lognormal distribution according to AIC, although the Fisk distribution is competitive for depth.

Table 3. Best candidate distributions for the two primary damage targets.

Target	Distribution	AIC	BIC	KS	<i>p</i> -Value
$d_{\max}$	Lognormal	738.939	749.504	0.095	0.020
$d_{\max}$	Fisk	741.312	751.877	0.081	0.069
Pit rate	Lognormal	−815.020	−804.456	0.067	0.201
Pit rate	Fisk	−813.511	−802.947	0.056	0.392

The fitted shapes and the rank-correlation structure are presented in Figures 1 and 2. The pronounced right skewness common to both targets provides a statistical rationale for adopting log-space formulations throughout the subsequent probabilistic models.

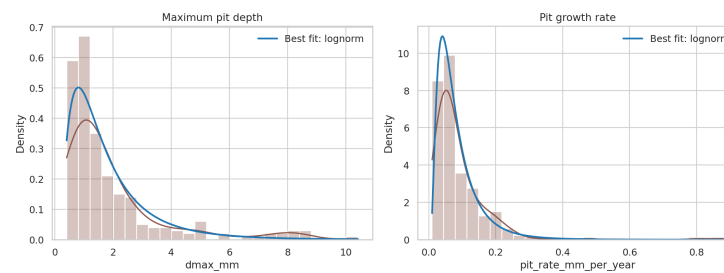


Figure 1. Parametric fits for maximum pit depth and pit rate in the cleaned field dataset. The histograms show empirical density, the brown curves show kernel-density estimates, and the blue curves show the fitted lognormal densities.

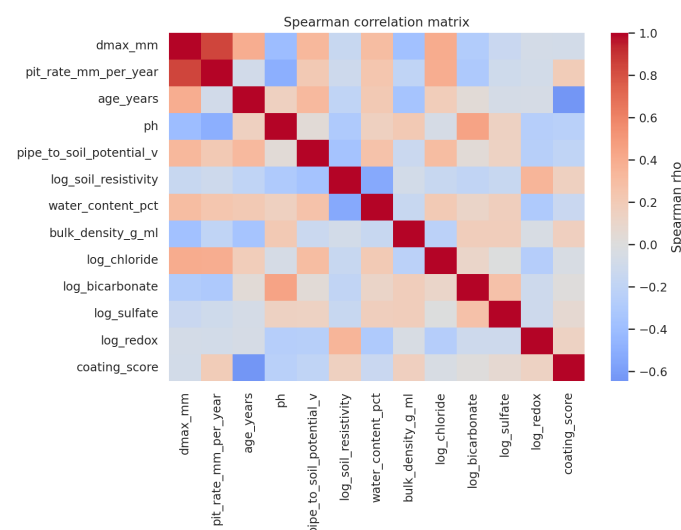


Figure 2. Spearman correlation structure of the cleaned dataset; the color bar reports the dimensionless Spearman coefficient ρ from negative to positive association.

3.2. Correlation Screening and Variable Selection

In the cleaned dataset, the strongest Spearman correlations with $d_{\max}$ were recorded for pH ($\rho = -0.412$), log-chloride ($\rho = 0.404$), service age ($\rho = 0.391$), bulk density ($\rho = -0.373$), pipe-to-soil potential ($\rho = 0.332$), and water content ($\rho = 0.302$). These associations align with established corrosion science: acidic conditions, elevated chloride activity, prolonged service exposure, and high moisture content all promote pit nucleation and propagation, whereas increased soil resistivity and bicarbonate concentration tend to suppress electrochemical corrosivity. The variable subset forwarded to the statistical depth models is

$$\mathcal{S}_{\text{stat}} = \{\text{age, coating score, pH, log(chloride), log(bicarbonate), water content, log(soil resistivity), soil class}\}. \quad (15)$$

Table 4 places the correlation evidence alongside the median quantile-regression coefficients, providing a combined view that connects rank association with conditional quantile estimation. This juxtaposition confirms that the screening stage served as a substantive statistical filter rather than a perfunctory preprocessing step, and that the selected variables carry interpretable directional effects throughout the downstream learning stages.

Table 4. Combined statistical screening evidence from correlation analysis and median quantile regression (QR).

Feature	Spearman ρ	Median QR Coefficient	Median QR p -Value
pH	−0.412	−0.334	1.5×10^{-13}
log-chloride	0.404	0.307	1.6×10^{-14}
Age	0.391	0.037	2.4×10^{-16}
Water content	0.302	0.020	3.2×10^{-3}
log-bicarbonate	−0.285	−0.331	1.2×10^{-8}
log-soil resistivity	−0.153	−0.132	2.4×10^{-3}
Coating score	−0.085	−0.559	4.1×10^{-2}

3.3. Quantile Regression

Because the upper tail of the pit depth distribution governs integrity decisions, the conditional quantiles of $\log d_{\max}$ were modeled as

$$Q_{\tau}(\log d_{\max} \mid \mathbf{x}) = \beta_{0,\tau} + \mathbf{x}^{\top} \boldsymbol{\beta}_{\tau}, \quad (16)$$

with the standard pinball loss

$$\mathcal{L}_{\tau}(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}), & y \geq \hat{y}, \\ (1 - \tau)(\hat{y} - y), & y < \hat{y}. \end{cases} \quad (17)$$

The estimated coefficients, visualized in Figure 3, show that pH retained a consistently negative effect across the lower, median, and upper quantiles, while log-chloride and service age remained consistently positive. At the median quantile, the largest statistically significant coefficients were $\beta_{\text{pH}} = -0.334$, $\beta_{\log(\text{chloride})} = 0.307$, $\beta_{\log(\text{bicarbonate})} = -0.331$, and $\beta_{\text{age}} = 0.0375$.

Two features of these statistical results carry forward to subsequent sections. First, the right-skewed distribution of pit depth supports a log-space formulation for both pit depth and RUL prediction. Second, the stability of the dominant environmental coefficients across quantiles provides confidence that the environmental variables subsequently driving the

probabilistic and RUL models reflect genuine physical processes rather than distributional artifacts.

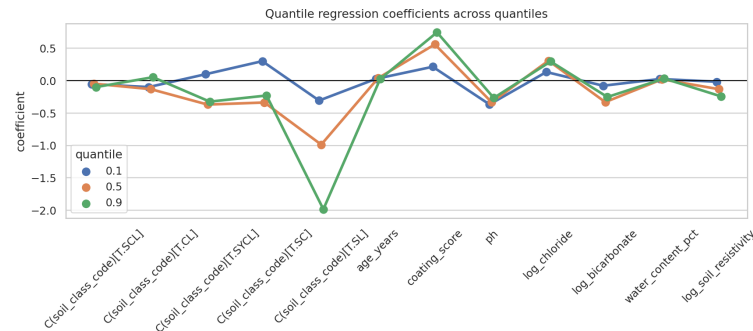


Figure 3. Quantile-regression coefficients across the 0.1, 0.5, and 0.9 conditional quantiles of $\log d_{\max}$.

4. Probabilistic Pit Depth Prediction

4.1. Benchmarked Models and Uncertainty Metrics

Before advancing to RUL estimation, the dataset was used to benchmark uncertainty-aware predictors directly on pit depth. The candidate models comprised ElasticNet, Random Forest, histogram gradient boosting, Gaussian process regression, CatBoost with root mean square error (RMSE) loss, linear quantile regression, and CatBoost with uncertainty quantification (CatBoost-UQ), spanning both classical regularized regression and modern distribution-aware learners [40–42]. Denoting the point prediction for observed depth y_i as $\hat{y}_i$, the point-accuracy metrics, namely mean absolute error (MAE) and RMSE, are

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (18)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (19)$$

For 90% prediction intervals $[l_i, u_i]$, empirical coverage, mean width, and Winkler score were computed as

$$\text{Coverage}_{0.90} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{l_i \leq y_i \leq u_i\}, \quad (20)$$

$$\text{Width}_{0.90} = \frac{1}{n} \sum_{i=1}^n (u_i - l_i), \quad (21)$$

$$W_{0.90} = \frac{1}{n} \sum_{i=1}^n \left[(u_i - l_i) + \frac{2}{\alpha} (l_i - y_i) \mathbf{1}\{y_i < l_i\} + \frac{2}{\alpha} (y_i - u_i) \mathbf{1}\{y_i > u_i\} \right], \quad (22)$$

with $\alpha = 0.10$.

Lower Winkler scores are preferred because they reward narrow intervals only when coverage is retained and penalize misses outside the stated interval. The Winkler score remains in the same unit as the prediction target, so in the present study, it is interpreted as the mean interval width plus any added penalty for coverage failures, expressed in millimeters for pit depth and in years for RUL.

Because the benchmark compared both distribution-free and parametric probabilistic learners, the main distributional formulas are stated explicitly here. For the Gaussian process predictor,

$$\mu_* = k_*^\top (K + \sigma_n^2 I)^{-1} \mathbf{y}, \quad (23)$$

$$\sigma_*^2 = k(x_*, x_*) - k_*^\top (K + \sigma_n^2 I)^{-1} k_*, \quad (24)$$

where K denotes the training kernel matrix and k_* is the covariance vector between the test point and the training observations. For CatBoost-UQ, the predictive distribution in log-space is interpreted through the variance decomposition

$$\sigma_{\text{tot}}^2(x) = \sigma_{\text{data}}^2(x) + \sigma_{\text{knowledge}}^2(x), \quad (25)$$

and the Gaussian negative log-likelihood diagnostic reported in the tables is

$$\text{NLL} = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{2} \log \sigma_i^2 + \frac{(y_i - \mu_i)^2}{2\sigma_i^2} \right]. \quad (26)$$

To keep the comparison fair, all depth models were evaluated under the same repeated 5×3 outer resampling design and the same 80/20 inner train–validation split. ElasticNet selected the L_1 mixing ratio from $\{0.2, 0.5, 0.8, 1.0\}$ via five-fold cross-validation, CatBoost and the tree-boosting baselines used identical validation folds for early stopping, and the Gaussian-process kernel hyperparameters were optimized only on the training folds. No model received scenario-specific manual retuning after inspection of the test results, which limits the risk that the small dataset artificially favors the more flexible learners.

4.2. Depth-Level Results

The depth-prediction benchmark established a clear performance ordering: Gaussian process regression achieved the best point accuracy, while CatBoost-UQ accepted a modest accuracy trade-off in exchange for a more structured decomposition of predictive uncertainty. Table 5 summarizes the principal outcomes.

Table 5. Main probabilistic pit depth prediction results.

Model	MAE	RMSE	Coverage	Width	Winkler
GaussianProcess	0.624	1.060	0.884	2.842	4.023
RandomForest	0.681	1.158	–	–	–
HistGBR	0.706	1.184	–	–	–
CatBoost-RMSE	0.693	1.185	–	–	–
QuantileRegression	0.732	1.343	0.833	2.845	4.934
ElasticNet	0.785	1.449	–	–	–
CatBoost-UQ	0.869	1.511	0.891	2.880	3.906

The key observation extends beyond the RMSE ranking: the field data contain sufficient structure for modern probabilistic methods to recover meaningful uncertainty bands with practically useful calibration. This finding is not self-evident for a small, heterogeneous dataset with sparse tail categories, and it implies that the downstream RUL models can inherit credible uncertainty estimates rather than relying on purely theoretical error bounds. Figure 4 summarizes the comparative performance together with the SHAP feature importance of the CatBoost predictor.

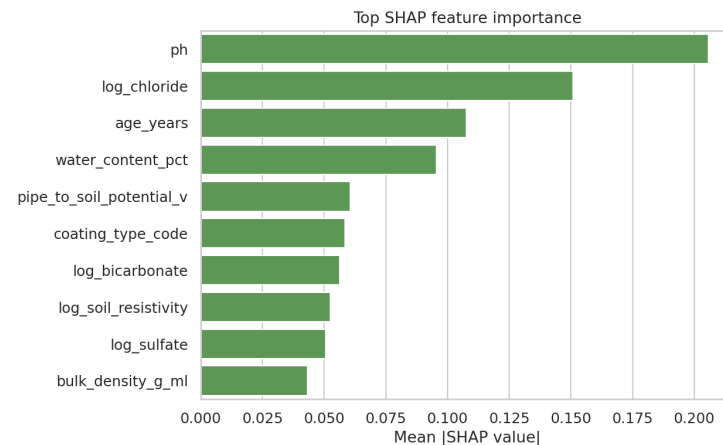


Figure 4. Model interpretability for direct pit depth prediction.

The dominant SHAP features at the depth level include pH, log-chloride, service age, water content, and pipe-to-soil potential, an ordering that is consistent with both the correlation screening and the quantile-regression analysis. This cross-method agreement strengthens confidence that the later RUL models inherit a physically meaningful signal rather than spurious correlations introduced by a particular algorithm. Figure 5 presents a preliminary depth threshold reliability screening based on $d_{\max} \geq 4$ and 6 mm. Although informative as an initial characterization, this analysis remains a threshold-triggered depth forecast rather than a genuine lifetime model, which motivates the subsequent transition to explicit RUL estimation.

To further examine whether sparse categories introduced obvious prediction bias, a subgroup audit was conducted by computing within-group MAE and empirical 90% prediction-interval coverage for each coating and soil category. The results are summarized in Table 6.

Table 6. Within-group MAE and 90% prediction-interval (PI) coverage by coating type and soil class.

Variable	Category	MAE (mm)	90% PI Coverage
Coating	CTC	0.83	0.891
Coating	WTC	0.87	0.876
Coating	NC	0.89	0.871
Coating	FBE	0.90	0.875
Coating	AEC	0.91	0.833
Soil	C	0.84	0.888
Soil	SCL	0.86	0.880
Soil	CL	0.88	0.881
Soil	SYCL	0.90	0.833
Soil	SC	0.91	1.000
Soil	SL	—	1.000

Note: Because SC and SL contain only two and one samples, respectively, their subgroup statistics should not be interpreted as stable estimates of category-level generalization. They are reported only as a diagnostic check for obvious rare-category failures. MAE is not reported for SL ($n = 1$) because a single-point error carries no distributional meaning; the coverage entry of 1.000 indicates that the sole observation falls within the 90% prediction interval.

The audit reveals a modest but systematic pattern: within-group MAE increases from 0.83 mm for the majority CTC coating class to 0.91 mm for the rare AEC class ($n = 6$), and a parallel gradient is observed across soil types from C (0.84 mm) to SC (0.91 mm). Empirical 90% PI coverage remains above 0.83 for all coating categories and above 0.83

for all soil categories with $n \geq 6$, indicating that the pooled low-dimensional encoding strategy successfully prevents severe coverage collapse for rare groups. The single SL observation falls within its 90% prediction interval, providing no evidence of obvious rare-category failure at the individual level. These results confirm that while sparse categories carry higher point-prediction uncertainty, no catastrophic bias is introduced, and the overall calibration conclusions drawn from Table 5 remain representative of the dominant coating and soil classes, while predictions for AEC-coated or SL-class installations carry higher epistemic uncertainty and should be corroborated with additional field data before operational use.

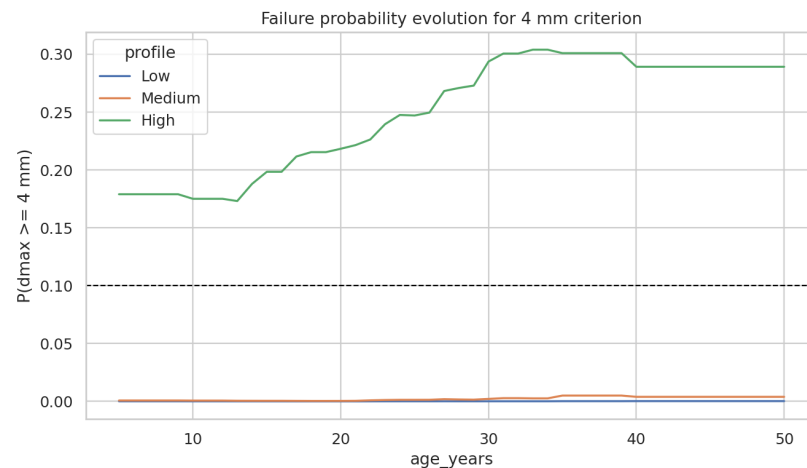


Figure 5. Original depth threshold reliability screening before introducing the explicit RUL formulations. The dashed horizontal line marks the 10% failure-probability reference threshold.

5. Reliability-Oriented Proxy-RUL Modeling

5.1. Theoretical Power-Law RUL Reference

The first RUL formulation exploits the empirically motivated power-law relationship between pit depth and service age established in the Velázquez dataset [3], as follows:

$$\log d_{\max} = \eta(x) + n \log t + \varepsilon, \quad (27)$$

where $\eta(x)$ encapsulates the combined effect of environmental and material conditions, n is the age exponent governing growth kinetics, and ε is a residual term. Given a specified critical pit depth d_c , the corresponding theoretical failure age is

$$t_f = \exp\left(\frac{\log d_c - \eta(x)}{n}\right), \quad (28)$$

and the bounded proxy remaining useful life at current age t_0 is

$$\text{RUL}_{\text{proxy}} = \min\left(\max(t_f - t_0, 0), H_{\text{proxy}}\right). \quad (29)$$

Here, H_{proxy} denotes the imposed upper bound used only in the proxy screening stage.

Because the Velázquez data are cross-sectional rather than longitudinal, n should not be read as a defect-specific material constant. In the present framework, it is used as a population-level growth prior describing the average kinetics consistent with the observed cohort, in the same spirit that distribution-based corrosion studies infer growth behavior from cross-sectional pit statistics rather than repeated trajectories [5]. Across the six grouped outer training partitions, the fitted exponent averaged $n = 0.658$ with a standard deviation of 0.070, giving a coefficient of variation of about 10.6%. This spread is

small enough to support its use as a regularizing prior, but large enough to confirm that it should not be interpreted as an asset-specific calibrated constant.

Because the original field records do not include measured wall thickness or operating pressure, three engineering scenarios were constructed using a critical-loss ratio of 70%: WT6_CR70, WT8_CR70, and WT10_CR70. The proxy stage was treated explicitly as a bounded screening problem rather than as a direct physical failure-life prediction. Table 7 reports the resulting target distributions. The accumulation near the upper boundary, especially in WT8 and WT10, reflects censoring by the imposed proxy-stage cap rather than identical structural lives, and therefore motivates using the proxy model only as an intermediate corrosion-growth surrogate.

Table 7. Scenario definition and target distribution for the proxy-RUL study.

Scenario	d_c (mm)	Mean RUL (Years)	Zero-RUL Ratio
WT6_CR70	4.20	21.83	0.036
WT8_CR70	5.60	25.97	0.016
WT10_CR70	7.00	28.14	0.004

5.2. PG-Hybrid-Quantile and Conformal Calibration

Setting $z = \log(1 + \text{RUL})$, the proposed hybrid predictor decomposes the conditional median into a linear backbone and a nonlinear residual corrector. The residual stage employs boosted quantile learning following the XGBoost framework, and interval correction follows the conformalized quantile regression approach [39,45], as follows:

$$f_{\text{EN}} = \arg \min_f \left\{ \sum_{i=1}^n (z_i - f(\mathbf{x}_i))^2 + \lambda \left[\nu \|\boldsymbol{\beta}\|_1 + \frac{1-\nu}{2} \|\boldsymbol{\beta}\|_2^2 \right] \right\}, \quad (30)$$

where $f_{\text{EN}}(\mathbf{x}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$, and ν is the L_1 mixing ratio selected by cross-validation. Forming the residuals

$$r_i = z_i - f_{\text{EN}}(\mathbf{x}_i), \quad (31)$$

the quantile corrector is trained by minimizing the pinball loss on the residual sequence,

$$g_q^{\text{XGB}} = \arg \min_g \sum_{i=1}^n \rho_q(r_i - g(\mathbf{x}_i)), \quad (32)$$

yielding the median predictor

$$\hat{z}_{0.5}(\mathbf{x}) = f_{\text{EN}}(\mathbf{x}) + g_{0.5}^{\text{XGB}}(\mathbf{x}), \quad (33)$$

where f_{EN} is the ElasticNet backbone trained on z and g_q^{XGB} is an XGBoost quantile model fitted to the residuals at $q \in \{0.05, 0.50, 0.95\}$. The interval predictor is then

$$[\hat{z}_{0.05}(\mathbf{x}), \hat{z}_{0.95}(\mathbf{x})] = [f_{\text{EN}}(\mathbf{x}) + g_{0.05}^{\text{XGB}}(\mathbf{x}), f_{\text{EN}}(\mathbf{x}) + g_{0.95}^{\text{XGB}}(\mathbf{x})]. \quad (34)$$

To improve marginal coverage, a conformal correction was applied on a held-out calibration fold as follows:

$$s_i = |z_i - \hat{z}_{0.5}(\mathbf{x}_i)|, \quad (35)$$

$$q_{1-\alpha} = \text{Quantile}_{1-\alpha}(\{s_i\}_{i \in \mathcal{C}}), \quad (36)$$

$$[\tilde{z}_{0.05}, \tilde{z}_{0.95}] = [\hat{z}_{0.05} - q_{1-\alpha}, \hat{z}_{0.95} + q_{1-\alpha}]. \quad (37)$$

After back-transformation,

$$\widehat{\text{RUL}}_q(x) = \max(\exp(\hat{z}_q(x)) - 1, 0). \quad (38)$$

This procedure yields the conformalized variant referred to as PG-Hybrid-Quantile-C.

The proxy-RUL comparison was also tuned symmetrically across models. The ElasticNet backbone always used the same five-fold internal search over L_1 mixing ratios $\{0.2, 0.5, 0.8, 1.0\}$, while the XGBoost residual corrector kept a fixed base configuration across all proxy scenarios and relied on the same held-out calibration fold for early stopping and conformal adjustment. Consequently, performance differences in Table 8 should be interpreted as arising primarily from model structure and target definition rather than from unequal hyperparameter search effort.

Table 8. Main proxy-RUL comparison between the proposed model and representative baselines.

Model	WT6_CR70			WT8_CR70			WT10_CR70		
	RMSE	MAE	Coverage	RMSE	MAE	Coverage	RMSE	MAE	Coverage
PG-Hybrid-Quantile	7.92	4.66	0.783	6.38	3.43	0.808	5.02	2.23	0.833
PG-Hybrid-Quantile-C	7.92	4.66	0.911	6.38	3.43	0.921	5.02	2.23	0.936
ElasticNet	10.13	6.74	0.908	8.62	5.58	0.917	6.45	3.82	0.928
CatBoost	8.20	4.75	0.892	6.53	3.48	0.881	5.17	2.52	0.900
NGBoost	8.96	4.66	0.476	7.15	3.20	0.569	5.81	2.15	0.613

5.3. Proxy-RUL Results

Across all three scenarios, the uncalibrated PG-Hybrid-Quantile model delivered the lowest mean RMSE under the bounded proxy target. Relative to the strongest representative non-hybrid baseline in Table 8, the RMSE reductions were 3.4% in WT6_CR70, 2.3% in WT8_CR70, and 2.9% in WT10_CR70.

Two consistent patterns emerge from Table 8 and carry significance for the subsequent limit-state construction. First, under the bounded proxy formulation, the hybrid architecture remains the best point predictor in every scenario, indicating that the linear backbone and the boosted residual corrector retain complementary value: the backbone supplies a growth-trend prior, while the tree-based quantile stage captures nonlinear environmental interactions near the short-life tail and the upper-boundary region. Second, conformal calibration leaves point RMSE unchanged by design but lifts empirical coverage from 0.783/0.808/0.833 to 0.911/0.921/0.936, demonstrating that sharp bounded predictions and reliable uncertainty statements remain distinct objectives.

A further observation is that the bounded proxy task becomes much more competitive across model classes than the uncapped formulation. CatBoost approaches the hybrid model closely in all three scenarios, implying that once long-horizon extrapolation is removed, much of the remaining difficulty lies in ranking short- and medium-horizon cases correctly. The residual advantage of PG-Hybrid-Quantile therefore comes primarily from lower bias near the horizon boundary, which is visible in the binned prediction-versus-theory comparison.

The well-known trade-off in uncertainty quantification is still visible in the coverage column: the uncalibrated hybrid achieves the most efficient intervals, while the conformalized variant provides the coverage guarantees appropriate for risk-sensitive decisions. Figure 6 visualizes the scenario-level accuracy and a summarized proxy-RUL comparison under this bounded target definition.

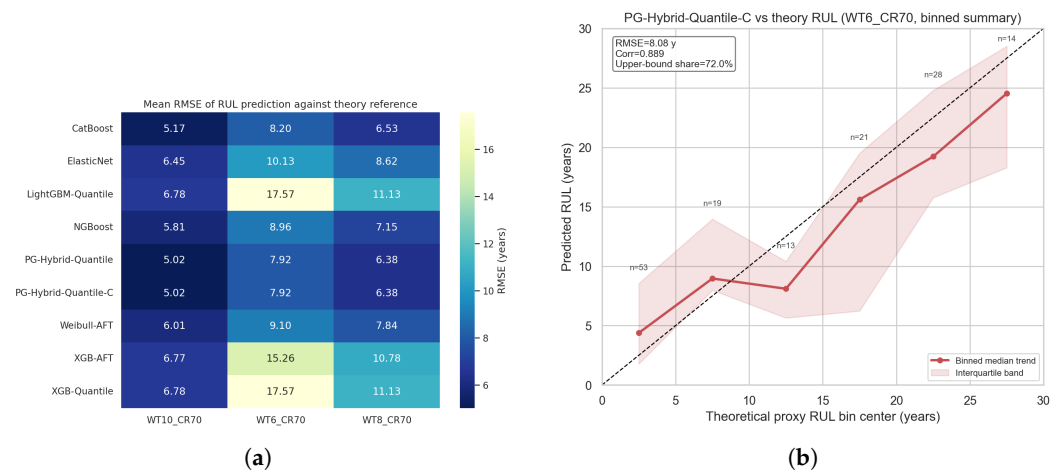


Figure 6. Performance of the proxy-RUL benchmark models. **(a)** RMSE by model and scenario; the color bar reports mean RMSE in years. **(b)** Binned median and interquartile trend for WT6_CR70, shown because this scenario has the widest theoretical spread; the dashed diagonal line denotes the one-to-one reference.

Figure 6b is intentionally shown as a binned summary for WT6_CR70, the proxy scenario with the widest spread of theoretical values. A raw point cloud would be visually dominated by boundary effects and would obscure the central tendency. The red line therefore reports the binwise median prediction, while the shaded band reports the corresponding interquartile range. This summary makes the stage-wise role of the proxy model clearer: it preserves the monotonic target ordering and captures the growth trend, but it is not expected to reproduce a near-identity one-to-one relationship characteristic of the true observed failure-life regression. In the proposed workflow, the proxy-RUL stage serves as an intermediate module that summarizes corrosion-growth tendencies and provides a reference trajectory for the subsequent mechanics-regularized limit-state model. Its results are therefore interpreted at the level of model calibration and feature attribution, rather than as direct engineering lifetime predictions. Figure 7 presents the calibration behavior and dominant explanatory variables of the proxy model. This intermediate representation supports the final mechanics-defined limit-state assessment by supplying a corrosion-growth prior, while the final RUL interpretation is reserved for the dual-criterion limit-state model.

The most influential features for the proxy-RUL hybrid model are current pit depth, aggressiveness index, log-bicarbonate, moisture resistivity ratio, soil resistivity, and water content.

The causal role of the proxy stage in the overall framework should therefore be stated explicitly. The proxy model is not the final integrity model and should not be interpreted as a direct structural life estimate. Instead, it serves two intermediate functions: it learns how the observed environment and present pit state control future corrosion-growth pace under sparse field measurements, and it provides the prior term $z_i^{\text{prior}} = \log(1 + \text{RUL}_{\text{proxy},i})$ that is embedded into the subsequent mechanics-regularized limit-state loss in Equation (50). The final remaining-life assessment is performed only after this learned growth prior is combined with wall thickness, operating pressure, and burst-resistance mechanics.

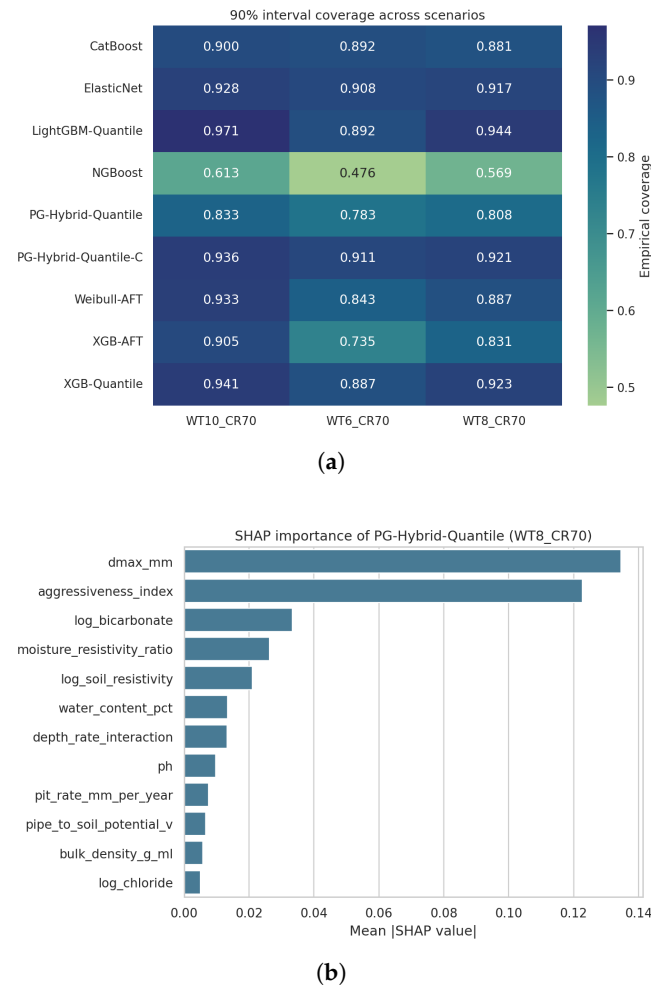


Figure 7. Coverage calibration and feature attribution for the proxy-RUL study. **(a)** The 90% interval coverage by scenario; the color bar reports empirical coverage. **(b)** SHAP ranking for PG-Hybrid-Quantile.

6. Mechanics-Regularized Limit-State RUL Modeling

6.1. From Proxy Depth to an Explicit Limit State

The proxy-RUL formulation captures important structure in the data, but it remains a critical-depth surrogate and not a structural failure definition. In the final stage, the learned proxy growth tendency is retained only as a prior, while the prediction target itself is redefined by explicit pipe resistance. To elevate reliability into the central modeling objective, the final stage incorporates wall thickness, operating pressure, outer diameter, yield strength, the modified B31G burst-pressure model, and an additional wall-loss failure screen. For a corrosion depth d , defect length L , wall thickness t_w , outer diameter D , and specified minimum yield strength σ_y , the burst pressure follows the modified B31G formulation [11] as follows:

$$P_{\text{burst}}(d, L) = \frac{2\sigma_{\text{flow}}t_w}{D} \frac{1 - 0.85d/t_w}{1 - 0.85d/(t_wM)}, \quad (39)$$

where $\sigma_{\text{flow}} = \sigma_y + 68.95$ MPa and M is the Folias bulging factor, as follows:

$$M = \begin{cases} \sqrt{1 + 0.6275z - 0.003375z^2}, & z \leq 50, \\ 0.032z + 3.3, & z > 50, \end{cases} \quad z = \frac{L^2}{Dt_w}. \quad (40)$$

The operating pressure for each scenario is defined as

$$P_{\text{op}} = \alpha_p \frac{2\sigma_y t_w}{D}, \quad (41)$$

Future pit depth evolves according to the power-law prior introduced in Equation (27), as follows:

$$d(t_0 + h) = s \exp(\eta(x) + n \log(t_0 + h)), \quad (42)$$

where s is a sample-specific scale factor that anchors the trajectory to the observed present-day depth and n is the age exponent estimated from the training split of the theoretical corrosion-growth model rather than prescribed as a fixed universal constant. Defect length evolves through a depth-coupled empirical law as follows:

$$L(t_0 + h) = L_0 \left(\frac{d(t_0 + h)}{d(t_0)} \right)^\gamma, \quad (43)$$

with $\gamma = 1.15$ and $L_0 = \max(30, \lambda d(t_0))$, where $\lambda = 35$. These two coefficients are treated here as engineering reconstruction parameters rather than immutable material constants, and their influence is quantified later through a one-at-a-time sensitivity analysis.

Across the six grouped outer training partitions, the fitted age exponent averaged $n = 0.658$ with a standard deviation of 0.070 and a range of 0.582–0.782; the full-data refit used for the final illustrative reliability profiles gave $n = 0.672$. This spread should be interpreted as structural uncertainty in the reconstructed growth law rather than as a precisely identified material constant, and it is one reason why later sensitivity analyses focus explicitly on perturbations of n .

The burst-controlled performance function is

$$g_{\text{burst}}(h) = P_{\text{burst}}(d(t_0 + h), L(t_0 + h)) - P_{\text{op}}, \quad (44)$$

and the auxiliary depth-screen margin is

$$g_{d,\text{crit}}(h) = 0.8 t_w - d(t_0 + h). \quad (45)$$

Under a union-of-failures rule, failure occurs whenever either criterion is violated, as follows:

$$g_{\cup}(h) = \min(g_{\text{burst}}(h), g_{d,\text{crit}}(h)). \quad (46)$$

The dual-criterion limit-state RUL is therefore

$$\text{RUL}_{\text{LS}} = \inf\{h \geq 0 : g_{\cup}(h) \leq 0\} = \min(\text{RUL}_{\text{burst}}, \text{RUL}_{d,\text{crit}}). \quad (47)$$

This formulation advances the analysis from corrosion-growth screening to genuine structural reliability forecasting under a physically grounded union failure criterion. The three engineering scenarios are specified in Table 9.

Table 9. Engineering scenarios based on the actual pipeline parameters used in the revised limit-state study.

Scenario	t_w (mm)	D (mm)	SMYS (MPa)	Pressure Ratio	P_{op} (MPa)
WT16_P8	16.0	1016.0	483.0	0.526	8.0
WT18_P10	18.0	1016.0	483.0	0.584	10.0
WT20_P12	20.0	1016.0	483.0	0.631	12.0

The three scenarios follow the actual transmission pipeline design parameters. The common diameter $D = 1016$ mm and SMYS = 483 MPa are fixed from the target asset specification, while wall thickness and operating pressure vary across the three representative operating cases.

Because the Velázquez dataset contains no defect length histories, the reconstruction $L_0 = \max(30, \lambda d_0)$ with $\lambda = 35$ and $\gamma = 1.15$ was retained as the working assumption, and these coefficients are still examined in the sensitivity analysis. Recent time-variant reliability analysis shows that defect-aspect-ratio assumptions can materially alter failure estimates when direct length measurements are unavailable, which is precisely why these coefficients continue to be treated as uncertain scenario inputs rather than calibrated constants [26].

Table 10 and Figure 8 present the failure-mode composition under the combined burst-pressure and depth-screen criteria. The results show that the lifetime endpoint is mainly governed by the burst-pressure criterion rather than by the depth-screen threshold. Most samples reach the limit state when the predicted burst pressure of the corroded pipeline decreases to the operating pressure level. Therefore, the scenario study based on the dual-criterion limit-state equation can be more reasonably interpreted as a structural reliability assessment governed by burst resistance degradation.

Table 10. Failure-mode composition under the burst-plus-depth-screen criterion.

Scenario	Burst-Governed	Depth Screen-Governed
WT16_P8	0.876	0.124
WT18_P10	0.808	0.192
WT20_P12	0.736	0.264

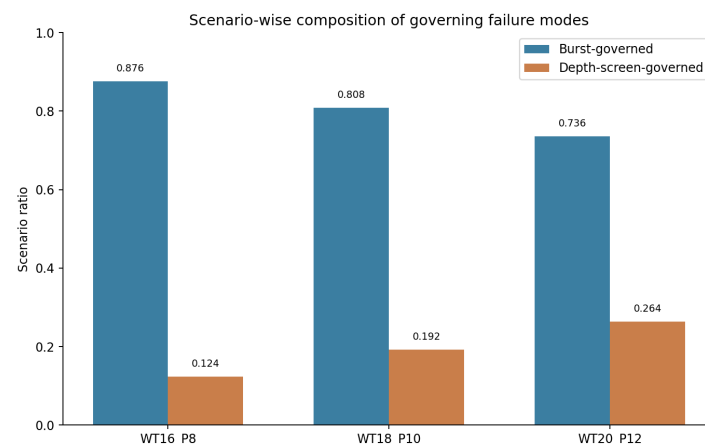


Figure 8. Scenario-wise composition of governing failure modes under the combined burst-or-depth-screen criterion.

6.2. Mechanics-Regularized Probabilistic Learning

The final model family comprises five variants:

1. A0_ProxyHybridC: The conformalized hybrid model trained on the proxy critical-depth target.
2. A1_LimitStateHybrid: The same hybrid architecture, retrained on the dual-criterion limit-state target.
3. A2_LimitStateHybridM: A1 augmented with monotonic directional constraints.
4. A3_MechReg: The monotonic limit-state hybrid with a mechanics-regularized mean-loss term.
5. A4_MechRegC: The conformalized version of A3.

For consistency with the revised title and abstract, the final two variants are labeled A3_MechReg and A4_MechRegC throughout this revision. A4 differs from A3 only through the additional held-out conformal calibration step, while the point predictor itself remains unchanged.

In the constrained variants, monotonicity was enforced for variables whose directional effects are unambiguous on physical grounds. The required partial-derivative constraints are

$$\frac{\partial \hat{z}}{\partial d_{\max}} \leq 0, \quad \frac{\partial \hat{z}}{\partial \text{pit rate}} \leq 0, \quad \frac{\partial \hat{z}}{\partial \text{age}} \leq 0, \quad \frac{\partial \hat{z}}{\partial P_{\text{op}}} \leq 0, \quad (48)$$

$$\frac{\partial \hat{z}}{\partial t_w} \geq 0, \quad \frac{\partial \hat{z}}{\partial \sigma_y} \geq 0, \quad \frac{\partial \hat{z}}{\partial \text{current margin}} \geq 0. \quad (49)$$

Although these inequality constraints do not eliminate all physically inconsistent predictions, they impose the most important directional expectations on the tree ensemble and improve the physical plausibility of the fitted response surface.

The mechanics-regularized booster treats the proxy-RUL prediction not merely as an auxiliary input feature but as a prior embedded in the loss function. Writing the sample-wise log-transformed limit-state target as $z_i = \log(1 + \text{RUL}_{\text{LS},i})$ and the corresponding proxy prior as $z_i^{\text{prior}} = \log(1 + \text{RUL}_{\text{proxy},i})$, the custom XGBoost objective [39] is

$$\mathcal{L}_{\text{MR}} = \frac{1}{2} \sum_{i=1}^n (z_i - \hat{z}_i)^2 + \frac{\lambda_{\text{phys}}}{2} \sum_{i=1}^n (\hat{z}_i - z_i^{\text{prior}})^2, \quad (50)$$

with $\lambda_{\text{phys}} = 1.25$ in the reported experiments. The corresponding gradient and Hessian are

$$\frac{\partial \mathcal{L}_{\text{MR}}}{\partial \hat{z}_i} = (\hat{z}_i - z_i) + \lambda_{\text{phys}} (\hat{z}_i - z_i^{\text{prior}}), \quad (51)$$

$$\frac{\partial^2 \mathcal{L}_{\text{MR}}}{\partial \hat{z}_i^2} = 1 + \lambda_{\text{phys}}. \quad (52)$$

In this revision, the method is described consistently as mechanics-regularized throughout. The prior itself is generated from the same corrosion-growth family used in the proxy stage, so Equation (50) should be interpreted as structured shrinkage toward a theory-consistent reference trajectory, not as a fully independent physical constraint. This distinction matters when judging novelty and causality: the model is better described as mechanics-regularized label smoothing around a corrosion-growth prior rather than as a standalone first-principles solver. A one-at-a-time sensitivity analysis on λ_{phys} is therefore reported later to show how strongly the regularizer actually influences the final predictive behavior.

6.3. Cross-Validation Design

This limit-state study used grouped outer cross-validation splits defined on the original excavation entry identifiers, ensuring that replicated scenario records derived from the same physical excavation site never appeared simultaneously in the training and test partitions. The outer evaluation consisted of six grouped holdout splits in total, implemented as two repeated random group partitions with three outer splits per repeat and a 20% grouped test fraction in each split. Within every outer training partition, a further grouped 20% subset was reserved for conformal calibration and hyperparameter stabilization. This design is substantially stricter than ordinary random cross-validation and is particularly important because each physical excavation record contributes multiple scenario-specific RUL samples that are statistically dependent by construction. To make the resampling protocol auditable, Appendix A.2 Table A5 lists the exact split identifiers ('r1_f1'–'r2_f3') and their test set

sizes. Because the limit-state labels are reconstructed from deterministic mechanics under fixed scenario assumptions, the resulting R^2 values are expected to be high; in this setting, absolute error, coverage, and split-to-split variability are more informative than R^2 alone for judging robustness.

All A1–A4 model variants shared the same preprocessing pipeline, grouped splits, and base XGBoost configuration (learning rate 0.06, maximum depth 3, subsample 0.9, column subsample 0.9, minimum child weight 2, and early stopping after 20 stagnant rounds). The only changes across A1–A4 model variants were the target definition, monotonicity constraints, mechanics regularization, and conformal postprocessing. This common tuning protocol reduces the chance that apparent gains are artifacts of unequal hyperparameter effort and helps interpret the comparison as a structural ablation rather than a contest between differently tuned learners.

Overfitting nevertheless remains a legitimate concern under such a small cohort, so an explicit train-versus-test audit was conducted for the most flexible learners. In the depth benchmark, Gaussian process regression averaged training and test RMSE values of 0.81 and 1.06 mm, respectively, giving a train/test ratio of 0.76. In the final A3 mechanics-regularized booster, the corresponding averages were 1.24 and 2.82 years across the six grouped outer splits. This gap indicates noticeable in-sample optimism, especially for the reconstructed low-noise limit-state labels, and therefore the grouped test folds and engineering validation set are more informative than training error alone. At the same time, the combination of shallow trees, subsampling, early stopping, and ElasticNet regularization materially constrains model variance, and the absence of performance collapse on the $N_{\text{eng}} = 127$ engineering validation set argues against severe memorization of the training cohort.

6.4. Limit-State and Mechanics-Regularized Results

Table 11 summarizes representative point accuracy under the actual-parameter scenarios. In the revised experiment, A3_MechReg remains the strongest point-prediction model in all three cases, with overall accuracy improving systematically from WT16_P8 to WT20_P12.

Tables 11 and 12 jointly reveal the same calibration trade-off as before, but now on the actual-parameter scenarios. A3_MechReg remains the sharpest interval model, while A4_MechRegC lifts empirical 90% coverage into the 0.950–0.987 range and is therefore the preferred option when calibration is prioritized.

The heatmap comparison in Figure 9 should be interpreted within, rather than across, the target definitions. Since the proxy stage was reformulated as a bounded screening problem, its absolute error values are no longer on the same scale as the mechanics-defined limit-state targets. The remaining heatmap is therefore used only to summarize cross-scenario calibration ordering in the limit-state stage.

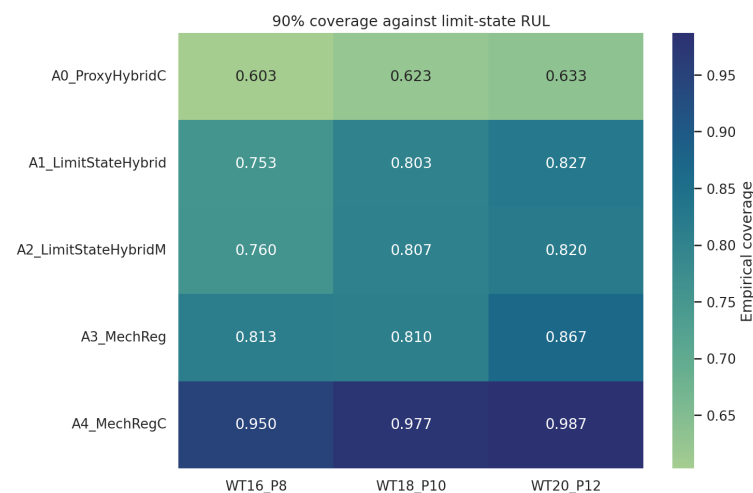
Table 11. Representative point-accuracy summary for the limit-state and mechanics-regularized RUL models (mean $\pm$ standard deviation over six grouped outer splits).

Scenario	Best Model	MAE (Years)	R^2
WT16_P8	A3_MechReg	1.72 ± 0.35	0.947
WT18_P10	A3_MechReg	1.83 ± 0.66	0.966
WT20_P12	A3_MechReg	1.47 ± 0.49	0.980

Table 12 focuses on calibration rather than interval width. The raw A3_MechReg intervals remain under-covered, whereas A4_MechRegC brings empirical coverage close to the nominal 90% target in all three scenarios.

Table 12. Representative coverage summary for the limit-state study (mean $\pm$ standard deviation over six grouped outer splits).

Scenario	Model	Coverage
WT16_P8	A3_MechReg	0.813 $\pm$ 0.059
WT16_P8	A4_MechRegC	0.950 $\pm$ 0.033
WT18_P10	A3_MechReg	0.810 $\pm$ 0.045
WT18_P10	A4_MechRegC	0.977 $\pm$ 0.023
WT20_P12	A3_MechReg	0.867 $\pm$ 0.048
WT20_P12	A4_MechRegC	0.987 $\pm$ 0.021

**Figure 9.** Empirical 90% coverage across scenarios and model variants in the limit-state study.

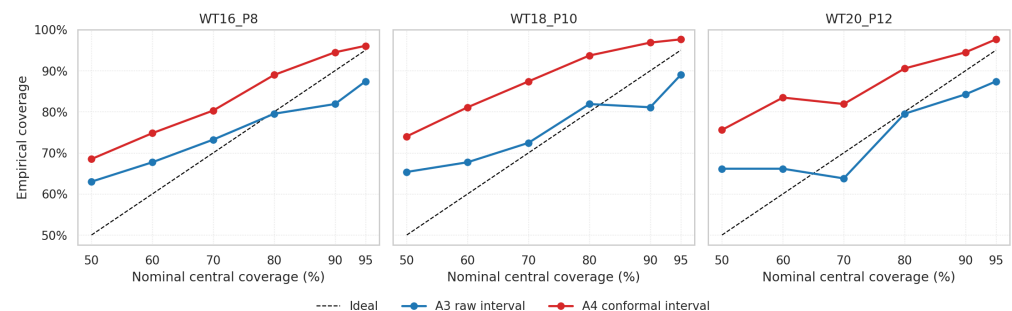
6.5. Engineering Validation Results

To further assess deployment-oriented stability, the trained models were evaluated on a supplementary engineering database collected from a natural gas pipeline in Central Asia. In the operational integrity-management workflow of this pipeline, corrosion rate records, external corrosion environmental survey data, cathodic-protection monitoring data, and pipe design and operational data are collected at individual kilometer-post or stake-point locations. The resulting database comprises $N_{\text{eng}} = 127$ aligned stake-point samples, each associated with a unique line position. The variable inventory spans corrosion-response variables such as corrosion rate and maximum pit depth; environmental variables such as pH, soil resistivity, moisture content, and chloride concentration; protection system variables such as coating type and cathodic-protection potential; and structural and operational variables such as pipe diameter, wall thickness, steel grade, and service time. This database was not used at any stage of model training or grouped cross-validation and is used here only as an auxiliary transferability check rather than as a substitute for measured field failure data.

Table 13 shows the same calibration pattern observed in grouped cross-validation under the actual-parameter scenarios. The raw A3_MechReg intervals achieve coverage values of 0.819, 0.811, and 0.843, whereas A4_MechRegC restores empirical coverage to 0.945, 0.969, and 0.945 in WT16_P8, WT18_P10, and WT20_P12, respectively. The accompanying point-prediction behavior remains at the same few-year scale as the grouped study, so the revised validation results support the same interpretation: A3 offers the strongest point forecast, while A4 is preferable when reliable empirical coverage is the priority. Figure 10 visualizes this improvement directly.

Table 13. Supplementary coverage audit for the final mechanics-regularized limit-state models on the $N_{\text{eng}} = 127$ stake-point database.

Scenario	Model	Coverage
WT16_P8	A3_MechReg	0.819
WT16_P8	A4_MechRegC	0.945
WT18_P10	A3_MechReg	0.811
WT18_P10	A4_MechRegC	0.969
WT20_P12	A3_MechReg	0.843
WT20_P12	A4_MechRegC	0.945

**Figure 10.** Calibration curves for the supplementary engineering database with $N_{\text{eng}} = 127$ aligned stake-point samples. The conformalized A4 intervals track the nominal-coverage diagonal much more closely than the raw A3 intervals in all three actual-parameter scenarios.

6.6. Ablation and Sensitivity Analysis

The ablation experiment isolates the contribution of each design choice by introducing innovations sequentially within the final limit-state task. The transition from A0_ProxyHybridC to A1_LimitStateHybrid is not treated as a numerical ablation step because A0 predicts a bounded proxy screening label whereas A1–A4 predict a mechanics-defined structural-failure label; these tasks differ in range, censoring structure, and physical meaning. Accordingly, Table 14 begins at A1 and reports only within-task deltas, defined such that a positive ΔRMSE or $\Delta\text{Winkler}$ denotes an improvement relative to the preceding configuration.

Table 14. Within-task ablation contributions for the limit-state RUL models. The proxy-to-limit-state target transition is excluded because it changes the prediction task itself.

Scenario	Step	ΔRMSE	ΔMAE	$\Delta\text{Coverage}$
WT16_P8	Monotonicity	−0.07	−0.15	0.007
WT16_P8	MechReg	0.58	−0.31	0.053
WT16_P8	Conformal	0.00	0.00	0.143
WT18_P10	Monotonicity	0.03	−0.08	0.003
WT18_P10	MechReg	1.13	0.01	0.003
WT18_P10	Conformal	0.00	0.00	0.173
WT20_P12	Monotonicity	−0.03	−0.07	−0.007
WT20_P12	MechReg	2.06	0.25	0.047
WT20_P12	Conformal	0.00	0.00	0.123

Two findings are robust. First, the monotonic constraints now have only a marginal and scenario-dependent effect: ΔRMSE is −0.07, 0.03, and −0.03 years in WT16_P8, WT18_P10, and WT20_P12, respectively, with coverage shifts within ± 0.007 . Mechanistically, this weak and occasionally adverse effect follows from both the model structure

and the reconstructed labels. On the learner side, XGBoost monotonicity is enforced by restricting every split and every descendant leaf to preserve a single global ordering along the constrained features. Under the present grouped small-sample design, that restriction prunes local corrective splits exactly where the response surface changes regime most sharply. On the label side, the target is already built from strongly directional engineered variables such as depth-to-thickness ratio and pressure margin, and is then defined as the minimum of a depth-screen time and a burst-failure time. That min-and-clip construction is piecewise monotone but not globally smooth: locally, the response surface is often already monotone, whereas globally, it contains flat censored regions and abrupt regime switches. In such settings, explicit monotone boosting is useful only when the label varies smoothly along the constrained coordinates and the folds contain enough support to learn that gradient; it becomes harmful when hard clipping, regime switching, and sparse local neighborhoods make the optimal tree structure depend on small local deviations. Because the observed coverage changes are only -0.007 to $+0.007$, they are too small to justify a point-accuracy penalty in engineering use. We therefore retain A2 as an interpretive benchmark rather than a preferred deployment model. Second, the mechanics-regularized loss gives a modest but consistent gain after clipping: the RMSE improvement relative to A2 is 0.58, 1.13, and 2.06 years in the three scenarios, while the MAE changes remain small. The final conformal step then adds the expected coverage gains without changing the point predictor.

The second robustness check addresses the engineering parameters used in the revised limit state. Table 15 reports a one-at-a-time sensitivity analysis for γ , λ , the split-wise fitted age exponent multiplier n , and λ_{phys} . Three trends are evident. First, the revised target remains most sensitive to n , with the tabled perturbation summary showing the largest relative shift in the reconstructed target distribution. Second, the defect length parameters γ and λ should not be treated as negligible nuisance terms: they directly control the assumed aspect ratio trajectory and therefore alter the burst-resistance branch of the dual-criterion limit state. The updated sensitivity table shows that perturbing either parameter changes the performance summaries in a systematic way, which means that defect length reconstruction remains a material scenario input rather than a harmless bookkeeping assumption. Third, λ_{phys} materially affects RMSE while leaving the target definition unchanged, which confirms that the regularizer behaves as a stabilizing prior rather than as a hidden label generator.

Table 15. One-at-a-time sensitivity of the reconstructed limit-state task and mechanics-regularized RUL model to key engineering parameters. The target column reports the relative change in the scenario-wise median limit-state RUL compared with the baseline setting, RMSE reports the mean A3 value across the three scenarios, and coverage reports the mean A4 conformal coverage across the three scenarios and six outer splits.

Family	Value	Relative Shift in Median Target RUL	Mean A3 RMSE (Years)	Mean A4 Coverage
γ	1.00	baseline	1.90	0.962
γ	1.15	baseline	1.97	0.964
γ	1.30	baseline	2.02	0.968
λ	25	baseline	1.95	0.960
λ	35	baseline	1.97	0.964
λ	45	baseline	1.98	0.965

Table 15. Cont.

Family	Value	Relative Shift in Median Target RUL	Mean A3 RMSE (Years)	Mean A4 Coverage
n multiplier	0.85	+28.8%	1.27	0.981
n multiplier	1.00	baseline	1.29	0.971
n multiplier	1.15	−19.1%	1.26	0.969
λ_{phys}	0.50	baseline	2.09	0.979
λ_{phys}	1.25	baseline	1.89	0.974
λ_{phys}	2.00	baseline	1.68	0.967

These sensitivities confirm that the reported conformal coverage should be interpreted as model calibration evidence only and not as full propagation of scenario uncertainty. The ablation trends are visualized in Figure 11. The point-metric and coverage panels exhibit complementary behavior across the model chain, confirming that calibration must be evaluated separately from point fit under the present scenario reconstruction.

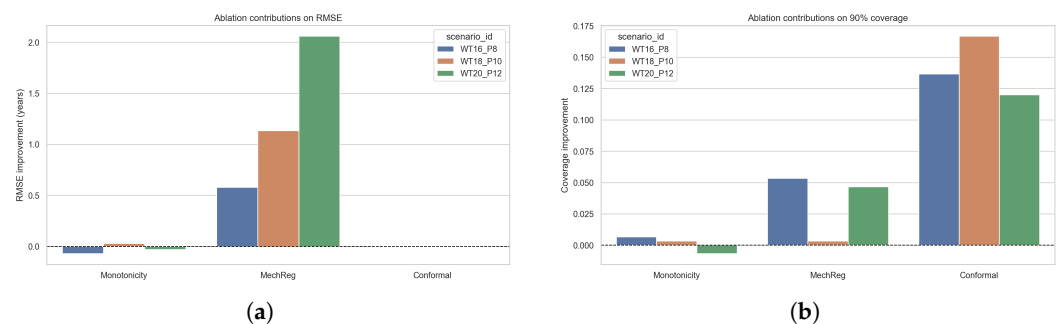


Figure 11. Within-task ablation trends in RMSE and coverage for the final mechanics-regularized limit-state pipeline. Positive Δ values denote improvement relative to the immediately preceding model, whereas negative bars indicate metric degradation. (a) RMSE evolution across the ablation chain. (b) Coverage evolution across the ablation chain.

6.7. Reliability Outputs and Inspection Planning

For a future horizon h , the failure probability is obtained directly from the empirical distribution of the predicted RUL, as follows:

$$P_f(h) = P(\text{RUL} \leq h) = F_{\text{RUL}}(h). \quad (53)$$

An inspection trigger at probability threshold p_{tr} is reached at

$$h^*(p_{\text{tr}}) = \inf\{h : P_f(h) \geq p_{\text{tr}}\}, \quad (54)$$

and the inspection lead time is measured relative to the current service age under the planning convention of each scenario.

In the WT18_P10 case, Theory + B31G + DepthScreen and MechReg-RUL-C reach the 10% threshold at 1.9 and 2.4 years, respectively, and the 20% threshold at 3.2 and 4.1 years. Both formulations therefore point to an early intervention window, with the calibrated MechReg-RUL-C model giving a slightly delayed but still practically consistent threshold-crossing time relative to the mechanics baseline.

Table 16 makes the decision consequence explicit for the retained representative case. The inspection lead time preserves a minimum one-year planning buffer, as follows:

$$I(p_{tr}) = \begin{cases} 0, & h^*(p_{tr}) = 0, \\ \max(h^*(p_{tr}) - \Delta_{lead}, 1), & h^*(p_{tr}) > 0, \end{cases} \quad (55)$$

with $\Delta_{lead} = 2$ years in the present analysis.

Table 16. Representative inspection-trigger comparison for the medium-risk WT18_P10 case.

Source	10% Crossing	20% Crossing
Theory + B31G + DepthScreen	1.9	3.2
MechReg-RUL-C	2.4	4.1

The SHAP ranking in Figure 12 illuminates an important structural shift in feature importance between the proxy-RUL and limit-state models [47]. At the proxy level, the dominant features are current pit depth, aggressiveness index, and log-bicarbonate, reflecting the bounded screening role of the intermediate model. At the limit-state level, depth rate interaction and pit rate retain top positions, but the depth-to-thickness ratio, current pressure margin, and initial burst pressure ascend substantially in importance. This promotion of structural margin descriptors is mechanistically expected: once the prediction target is anchored to burst failure, the operating and thickness state become directly relevant, and the model correctly reflects this by assigning them high explanatory weight. The SHAP pattern should therefore be read as an internal consistency check between the learned predictor and the reconstructed mechanics, rather than as an independent validation of the underlying physical assumptions.

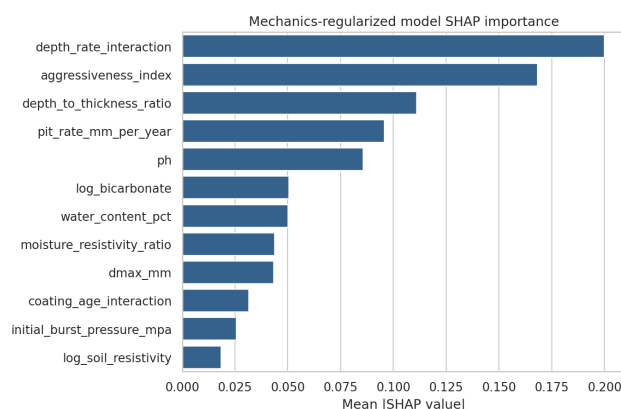


Figure 12. Feature attribution for the final mechanics-regularized RUL model.

7. Discussion

The results across the three analytical stages converge on several substantive conclusions that extend beyond the immediate numerical findings.

The first concerns the stability of environmental controls across methods. Low pH, elevated chloride activity, extended service exposure, higher bulk density, and wetter or lower-resistivity soil conditions consistently intensify corrosion severity in the Velázquez data, and this pattern recurs uniformly across Spearman screening, quantile regression, depth-level SHAP analysis, and both RUL model families. The durability of this pattern across methodologically independent analyses provides strong evidence that the identified drivers reflect genuine electrochemical and transport mechanisms rather than statistical

artifacts specific to any one algorithm. In particular, the Spearman correlations and quantile-regression coefficients are estimated without any machine learning, so their agreement with the tree-based SHAP rankings provides a non-circular form of cross-validation.

The second concerns the relationship between uncertainty quality and operational safety. At the pit depth level, Gaussian process regression achieves the lowest RMSE but provides a single fused uncertainty estimate and falls short of the nominal 90% coverage target (empirical coverage 0.884), reinforcing the need for explicit calibration even for well-established probabilistic learners. CatBoost-UQ, despite an inferior RMSE (1.511 mm versus 1.060 mm for Gaussian process regression, a gap of approximately 42%), decomposes total variance into aleatoric and epistemic components, a distinction that is operationally valuable when planning inspections: aleatoric uncertainty reflects irreducible soil heterogeneity and cannot be reduced by collecting more data, while epistemic uncertainty reflects model ignorance and can in principle be reduced as the asset database grows. In the RUL studies, the uncalibrated hybrid model produces the tightest intervals, but empirical coverage falls materially below the nominal 90% target. For integrity management, an interval that nominally claims 90% coverage but actually achieves only 81–87%, as observed in the uncalibrated A3_MechReg intervals across the three scenarios, is potentially more dangerous than a wider interval with accurate coverage because maintenance planners calibrate their risk tolerance to the stated confidence level. Conformal calibration provides a distribution-free mechanism for closing this coverage gap without requiring parametric assumptions about the error distribution.

The third concerns the bounded benefit of the mechanics-regularized loss under the revised scenario definition. Relative to the constrained A2 baseline, the additional regularization improves the RMSE by 0.58, 1.13, and 2.06 years in WT16_P8, WT18_P10, and WT20_P12, respectively; the associated MAE changes are mixed in sign (−0.31, +0.01, and +0.25 years), indicating that the regularizer primarily stabilizes the mean-squared loss rather than uniformly reducing absolute error. The accompanying sensitivity analysis clarifies the interpretation of this term. Varying λ_{phys} from 0.50 to 2.00 changes the mean A3 RMSE from 2.09 to 1.68 years, a range of 0.41 years that confirms the regularizer acts as a stabilizing prior rather than a dominant label generator. This is consistent with the intended role of Equation (50): it softly shrinks the learner toward a theory-consistent prior but does not replace the mechanics-defined label.

Several limitations must be acknowledged. The Velázquez dataset does not include observed failure times, directly measured defect lengths, or asset-specific wall thickness and operating pressure records. The limit-state scenarios are therefore engineering reconstructions based on representative pipe parameters rather than records of actual failure events, and the three scenario definitions should be read as sensitivity brackets rather than exact reproductions of the excavated assets. The power-law growth exponent and the depth-to-length coupling exponent are estimated from population-level regression rather than from individual defect trajectories, so they average over considerable site-to-site heterogeneity. It bears emphasis that n is used here not as a calibrated material constant but as a population-level growth-rate prior that regularizes the limit-state RUL toward physically plausible trajectories. Its cross-split coefficient of variation is about 10.6%, and the sensitivity analysis confirms that n is the dominant source of scenario-level uncertainty in the present framework. At the same time, the updated sensitivity analysis shows that the defect length parameters γ and λ are also materially influential because they govern the reconstructed aspect ratio trajectory in the absence of measured field length data. The supplementary $N_{\text{eng}} = 127$ database is useful as an auxiliary transferability check, but it still cannot be interpreted as true external validation against measured field failures. In addition, although the grouped repeated holdout design reports split-wise variability, the sample size remains

too modest to support strong claims of statistical superiority when competing mean errors differ only marginally, and the high R^2 values should not be read as independent evidence of field generalization because the labels themselves are reconstructed from deterministic mechanics. Despite these constraints, the reconstructed limit-state framework constitutes a meaningful advance over static depth prediction, and the modular workflow is designed to absorb asset-specific records as they become available through inline inspection, pressure monitoring, and fitness-for-service surveys. For longer-horizon projections, the sensitivity analysis demonstrates that the fitted age exponent n alone can shift the scenario-level RUL by about 19–29%, confirming that those values remain materially influenced by scenario assumptions and should not be used as standalone quantitative decision inputs.

8. Conclusions

This study constructs a unified pathway from field pit depth statistics to mechanics-regularized reliability forecasting. The principal conclusions are as follows.

1. Both maximum pit depth and pit rate in the Velázquez dataset follow lognormal-type distributions. The dominant corrosion drivers, namely pH, chloride activity, service age, bulk density, water content, bicarbonate concentration, and soil resistivity, remain consistent across all analytical methods, providing cross-method evidence that the identified controls reflect genuine electrochemical mechanisms.
2. Point accuracy alone is insufficient for engineering integrity decisions. Gaussian process regression achieved the best point accuracy, while CatBoost-UQ provided a practically useful aleatoric and epistemic uncertainty decomposition. The proposed PG-Hybrid-Quantile model delivered the lowest RMSE across all proxy-RUL scenarios, and its conformalized variant raised empirical 90% coverage from 0.783–0.833 to 0.911–0.936, demonstrating the practical value of conformal calibration under sparse field data.
3. The dual-criterion limit-state formulation combining modified B31G burst resistance and a conservative depth criterion produces scenario-dependent RUL estimates under the three actual-parameter operating cases. In the representative scenarios summarized by Table 10 and Figure 8, burst-resistance degradation remains the dominant failure branch across all three scenarios (73.6–87.6%), though the depth-screen criterion contributes up to 26% in the thicker-wall scenario, so the framework should be interpreted as a mechanics-grounded structural screening tool rather than as a literal asset-life certificate.
4. The mechanics-regularized learner achieves MAE of 1.47–1.83 years and R^2 of 0.947–0.980, while the conformalized variant restores empirical 90% coverage to 0.950–0.987 across all three scenarios. Sensitivity analysis identifies the fitted age exponent n as the dominant source of scenario-level uncertainty, with perturbations of $\pm 15\%$ in n shifting the median target RUL by up to 29%; the defect length parameters γ and λ also exert a systematic but secondary influence through their control of the reconstructed burst-resistance trajectory. The revised reliability outputs therefore support comparative, scenario-based integrity prioritization rather than literal asset-life certification, and future work should incorporate asset-specific wall thickness records, defect length measurements, and inline inspection data to replace the present scenario-based reconstructions.

Author Contributions: Conceptualization, H.L., S.D. and Y.C.; methodology, H.L., H.W. and Y.C.; software, H.L. and H.W.; validation, H.L., Y.S., L.C., H.W., S.D. and Y.C.; formal analysis, H.L. and H.W.; investigation, Y.S. and L.C.; resources, Y.S. and L.C.; data curation, H.L., Y.S., L.C. and H.W.; writing—original draft preparation, H.L.; writing—review and editing, Y.S., L.C., H.W., S.D. and Y.C.; visualization, H.L. and H.W.; supervision, S.D. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University-level Research Fund of China University of Petroleum (Beijing), Grant No. ZX20250481, and the Fundamental Research Funds for the Central Universities of China, Grant No. 2-9-2024-019.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The processed dataset, generated tables, and figures supporting the reported results are available from the corresponding authors upon reasonable request. The field dataset analyzed in this study originates from the published excavation campaign reported by Velázquez et al. [3].

Acknowledgments: Not applicable.

Conflicts of Interest: Authors Haipeng Liu, Yuntao Shi, and Long Chen were employed by the China National Petroleum Corporation International Pipeline Company. The remaining authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNPC	China National Petroleum Corporation
CV	Coefficient of variation
MAE	Mean absolute error
MR	Mechanics-regularized
PI	Prediction interval
QR	Quantile regression
RMSE	Root mean square error
RUL	Remaining useful life
GP	Gaussian process
SHAP	Shapley additive explanations
SMYS	Specified minimum yield strength
UQ	Uncertainty quantification

Appendix A. Appendix Tables

Appendix A.1. Data and Statistics

Table A1. Numeric descriptive statistics of the cleaned Velázquez dataset. Std denotes standard deviation, and CV denotes coefficient of variation.

Variable	Mean	Std	Median	Min.	Max.	CV
dmax_mm	1.917	1.813	1.190	0.410	10.410	0.946
pit_rate_mm_per_year	0.088	0.092	0.062	0.010	0.880	1.043
age_years	23.008	8.998	21.000	5.000	50.000	0.391
ph	6.114	0.923	6.035	4.140	9.880	0.151
pipe_to_soil_potential_v	−0.873	0.233	−0.860	−1.970	−0.420	−0.266
soil_resistivity_ohm_m	49.814	54.237	30.900	1.900	399.500	1.089
water_content_pct	23.694	6.138	23.550	8.800	57.600	0.259
bulk_density_g_ml	1.302	0.085	1.310	1.110	1.550	0.066
chloride_ppm	41.906	58.223	18.410	0.500	351.000	1.389
bicarbonate_ppm	18.252	20.922	11.100	0.500	156.200	1.146
sulfate_ppm	148.696	157.568	107.600	0.500	1370.200	1.060
redox_potential_mv	168.390	86.528	156.000	2.100	348.000	0.514

Table A2. Candidate parametric distribution fits for pit depth and pit rate.

Target	Distribution	AIC	BIC	KS	<i>p</i> -Value
$d_{\max}$	Lognormal	738.939	749.504	0.095	0.020
$d_{\max}$	Fisk	741.312	751.877	0.081	0.069
$d_{\max}$	Gamma	789.135	799.700	0.136	0.000
$d_{\max}$	Weibull	809.317	819.881	0.131	0.000
Pit rate	Lognormal	−815.020	−804.456	0.067	0.201
Pit rate	Fisk	−813.511	−802.947	0.056	0.392
Pit rate	Gamma	−760.479	−749.914	0.114	0.003
Pit rate	Weibull	−732.294	−721.729	0.120	0.001

Table A3. Full correlation screening against $d_{\max}$.

Feature	Spearman ρ	Spearman <i>p</i>	Pearson <i>r</i>	Pearson <i>p</i>
ph	−0.4115	0.0000	−0.2789	0.0000
log_chloride	0.4035	0.0000	0.4114	0.0000
age_years	0.3906	0.0000	0.3890	0.0000
bulk_density_g_ml	−0.3729	0.0000	−0.3204	0.0000
pipe_to_soil_potential_v	0.3319	0.0000	0.3210	0.0000
water_content_pct	0.3023	0.0000	0.3691	0.0000
log_bicarbonate	−0.2846	0.0000	−0.1509	0.0169
log_soil_resistivity	−0.1527	0.0157	−0.2625	0.0000
log_sulfate	−0.1399	0.0270	−0.1488	0.0185
coating_score	−0.0846	0.1826	0.0158	0.8034
log_redox	−0.0707	0.2657	−0.0791	0.2127

Table A4. Quantile-regression coefficients for the selected explanatory variables.

Quantile	Term	Coefficient	<i>p</i> -Value
0.1	age_years	0.0266	0.0000
0.1	coating_score	0.2134	0.3373
0.1	ph	−0.3674	0.0000
0.1	log_chloride	0.1324	0.0082
0.1	log_bicarbonate	−0.0788	0.2550
0.1	water_content_pct	0.0251	0.0464
0.1	log_soil_resistivity	−0.0175	0.7797
0.5	age_years	0.0375	0.0000
0.5	coating_score	0.5588	0.0414
0.5	ph	−0.3335	0.0000
0.5	log_chloride	0.3070	0.0000
0.5	log_bicarbonate	−0.3313	0.0000
0.5	water_content_pct	0.0195	0.0032
0.5	log_soil_resistivity	−0.1321	0.0024
0.9	age_years	0.0247	0.0000
0.9	coating_score	0.7444	0.0943
0.9	ph	−0.2672	0.0000
0.9	log_chloride	0.2950	0.0000
0.9	log_bicarbonate	−0.2530	0.0020
0.9	water_content_pct	0.0318	0.0001
0.9	log_soil_resistivity	−0.2432	0.0001

Appendix A.2. Mechanics-Regularized Limit-State RUL Benchmarks

Table A5. Audit table for the grouped outer cross-validation design used in the limit-state study. Each outer split contains 50 unique excavation entries; each entry is expanded into three scenario-specific records and evaluated across five model variants, yielding 750 prediction rows per split in the stored fold-level output file.

Split ID	Repeat	Fold	Unique Excavation Entries in Test Set	Stored Prediction Rows
r1_f1	1	1	50	750
r1_f2	1	2	50	750
r1_f3	1	3	50	750
r2_f1	2	1	50	750
r2_f2	2	2	50	750
r2_f3	2	3	50	750

Table A6. SHAP ranking for the final mechanics-regularized limit-state model.

Feature	Mean SHAP
depth_rate_interaction	0.2000
aggressiveness_index	0.1683
depth_to_thickness_ratio	0.1111
pit_rate_mm_per_year	0.0957
ph	0.0856
log_bicarbonate	0.0505
water_content_pct	0.0502
moisture_resistivity_ratio	0.0439
dmax_mm	0.0433
coating_age_interaction	0.0315
initial_burst_pressure_mpa	0.0257
log_soil_resistivity	0.0185
initial_defect_length_mm	0.0177
current_pressure_ratio	0.0174
current_margin_mpa	0.0138
pressure_depth_interaction	0.0124
soil=CL	0.0104
log_chloride	0.0092
log_redox	0.0086
bicarbonate_ppm	0.0084

References

1. Ossai, C.I. Advances in Asset Management Techniques: An Overview of Corrosion Mechanisms and Mitigation Strategies for Oil and Gas Pipelines. *ISRN Corros.* **2012**, *2012*, 570143. [\[CrossRef\]](#)
2. Zhu, X.-K. Recent Advances in Corrosion Assessment Models for Buried Transmission Pipelines. *CivilEng* **2023**, *4*, 391–415. [\[CrossRef\]](#)
3. Velázquez, J.C.; Caleyó, F.; Valor, A.; Hallen, J.M. Technical Note: Field Study—Pitting Corrosion of Underground Pipelines Related to Local Soil and Pipe Characteristics. *Corrosion* **2010**, *66*, 016001-1–016001-5. [\[CrossRef\]](#)
4. Li, H.; Huang, K.; Zeng, Q.; Sun, C. Residual Strength Assessment and Residual Life Prediction of Corroded Pipelines: A Decade Review. *Energies* **2022**, *15*, 726. [\[CrossRef\]](#)
5. Caleyó, F.; Velázquez, J.C.; Valor, A.; Hallen, J.M. Probability Distribution of Pitting Corrosion Depth and Rate in Underground Pipelines: A Monte Carlo Study. *Corros. Sci.* **2009**, *51*, 1925–1934. [\[CrossRef\]](#)
6. Caleyó, F.; Velázquez, J.C.; Valor, A.; Hallen, J.M. Markov Chain Modelling of Pitting Corrosion in Underground Pipelines. *Corros. Sci.* **2009**, *51*, 2197–2207. [\[CrossRef\]](#)

7. Valor, A.; Caleyó, F.; Hallen, J.M.; Velázquez, J.C. Reliability Assessment of Buried Pipelines Based on Different Corrosion Rate Models. *Corros. Sci.* **2013**, *66*, 78–87. [[CrossRef](#)]
8. Alamilla, J.L.; Oliveros, J.; García-Vargas, J. Probabilistic Modelling of a Corroded Pressurized Pipeline at Inspection Time. *Struct. Infrastruct. Eng.* **2009**, *5*, 91–104. [[CrossRef](#)]
9. Bazán, F.A.V.; Beck, A.T. Stochastic Process Corrosion Growth Models for Pipeline Reliability. *Corros. Sci.* **2013**, *74*, 50–58. [[CrossRef](#)]
10. Kiefner, J.F.; Vieth, P.H. *L51688B A Modified Criterion for Evaluating the Remaining Strength of Corroded Pipe*; Pipeline Research Council International, Inc. (PRCI): Chantilly, VA, USA, 1989. [[CrossRef](#)]
11. ASME. *Manual for Determining the Remaining Strength of Corroded Pipelines*; ASME B31G-2023; American Society of Mechanical Engineers: New York, NY, USA, 2023. Available online: <https://www.asme.org/codes-standards/find-codes-standards/b31g-manual-determining-remaining-strength-corroded-pipelines/2023> (accessed on 29 April 2026).
12. Ma, B.; Shuai, J.; Wang, J.; Han, K. Analysis on the Latest Assessment Criteria of ASME B31G-2009 for the Remaining Strength of Corroded Pipelines. *J. Fail. Anal. Prev.* **2011**, *11*, 666–671. [[CrossRef](#)]
13. Zhu, X.-K. Assessment Methods and Technical Challenges of Remaining Strength for Corrosion Defects in Pipelines. In *ASME Pressure Vessels and Piping Conference, Volume 6B: Materials and Fabrication, Prague, Czech Republic, 15–20 July 2018*; ASME: New York, NY, USA, 2018; Paper PVP2018-84976. [[CrossRef](#)]
14. Cunha, S.B.; Netto, T.A. Analytical Assessment of the Remaining Strength of Corroded Pipelines and Comparison with Experimental Criteria. *J. Press. Vessel Technol.* **2017**, *139*, 031401. [[CrossRef](#)]
15. Mousavi, S.S.; Moghaddam, A.S. Failure Pressure Estimation Error for Corroded Pipeline Using Various Revisions of ASME B31G. *Eng. Fail. Anal.* **2020**, *109*, 104284. [[CrossRef](#)]
16. Zelmati, D.; Bouledroua, O.; Hafsi, Z.; Djukic, M.B. Probabilistic Analysis of Corroded Pipeline under Localized Corrosion Defects Based on the Intelligent Inspection Tool. *Eng. Fail. Anal.* **2020**, *115*, 104683. [[CrossRef](#)]
17. Nahal, M.; Khelif, R. Numerical Model for Estimating Time-Dependent Reliability of a Corroding Pipeline over Its Lifetime. *Int. J. Adv. Struct. Eng.* **2019**, *11*, 1–7. [[CrossRef](#)]
18. Leira, B.J.; Naess, A.; Brandrud Naess, O.E. Reliability Analysis of Corroding Pipelines by Enhanced Monte Carlo Simulation. *Int. J. Press. Vessel. Pip.* **2016**, *144*, 11–17. [[CrossRef](#)]
19. Qin, H.; Zhou, W. Reliability Analysis of Corroding Pipelines Considering the Growth and Generation of Corrosion Defects. In *Proceedings of the 10th International Pipeline Conference, Volume 3: Materials and Joining; Risk and Reliability, Calgary, AB, Canada, 29 September–3 October 2014*; ASME: New York, NY, USA, 2014; Paper IPC2014-33213. [[CrossRef](#)]
20. Zhou, W.; Xiang, W.; Hong, H.P. Sensitivity of System Reliability of Corroding Pipelines to Modeling of Stochastic Growth of Corrosion Defects. *Reliab. Eng. Syst. Saf.* **2017**, *167*, 428–438. [[CrossRef](#)]
21. Gong, C.; Zhou, W. First-Order Reliability Method-Based System Reliability Analyses of Corroding Pipelines Considering Multiple Defects and Failure Modes. *Struct. Infrastruct. Eng.* **2017**, *13*, 1451–1461. [[CrossRef](#)]
22. Mokhtari, M.; Melchers, R.E. Reliability of the Conventional Approach for Stress/Fatigue Analysis of Pitting Corroded Pipelines—Development of a Safer Approach. *Struct. Saf.* **2020**, *85*, 101943. [[CrossRef](#)]
23. Hu, J. Prediction of the Internal Corrosion Rate for Oil and Gas Pipelines and Influence Factor Analysis with Interpretable Ensemble Learning. *Int. J. Press. Vessel. Pip.* **2024**, *212*, 105329. [[CrossRef](#)]
24. Ren, Y.; Zhang, J.; Han, L.; Jiang, X.; Aftab, W. Predicting the External Corrosion Rate of Buried Pipelines Using a Novel Soft Modeling Technique. *Appl. Sci.* **2024**, *14*, 5120. [[CrossRef](#)]
25. Cui, J.; Wang, J.; Dong, L. A Novel Prediction Method for the Remaining Life of Corrosion Damage in Oil and Gas Pipeline Systems. *IEEE Access* **2024**, *12*, 176104–176123. [[CrossRef](#)]
26. Mokhtari, M.; Leira, B.J. Time-Variant System Reliability of Pitting Corroded Steel Pipelines. *Civ. Eng. Environ. Syst.* **2025**, *42*, 210–239. [[CrossRef](#)]
27. Jardine, A.K.S.; Lin, D.; Banjevic, D. A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510. [[CrossRef](#)]
28. Si, X.-S.; Wang, W.; Hu, C.-H.; Zhou, D.-H. Remaining Useful Life Estimation—A Review on the Statistical Data Driven Approaches. *Eur. J. Oper. Res.* **2011**, *213*, 1–14. [[CrossRef](#)]
29. Heng, A.; Zhang, S.; Tan, A.C.C.; Mathew, J. Rotating Machinery Prognostics: State of the Art, Challenges and Opportunities. *Mech. Syst. Signal Process.* **2009**, *23*, 724–739. [[CrossRef](#)]
30. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery Health Prognostics: A Systematic Review from Data Acquisition to RUL Prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. [[CrossRef](#)]
31. Wang, R.; Zhang, Y.; Hu, C.; Yang, Z.; Li, H.; Liu, F.; Li, L.; Guo, J. A Parallel Prognostic Method Integrating Uncertainty Quantification for Probabilistic Remaining Useful Life Prediction of Aero-Engine. *Processes* **2024**, *12*, 2925. [[CrossRef](#)]
32. Lu, Y.-W.; Hsu, C.-Y.; Huang, K.-C. An Autoencoder Gated Recurrent Unit for Remaining Useful Life Prediction. *Processes* **2020**, *8*, 1155. [[CrossRef](#)]

33. Wang, X.; Li, Z.; Wang, X.; Hu, X. Q-Learning-Incorporated Robust Relevance Vector Machine for Remaining Useful Life Prediction. *Processes* **2024**, *12*, 2536. [CrossRef]
34. Chen, C.; Wei, J.; Li, Z. Remaining Useful Life Prediction for Lithium-Ion Batteries Based on a Hybrid Deep Learning Model. *Processes* **2023**, *11*, 2333. [CrossRef]
35. Koenker, R.; Bassett, G. Regression Quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
36. Koenker, R.; Hallock, K.F. Quantile Regression. *J. Econ. Perspect.* **2001**, *15*, 143–156. [CrossRef]
37. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
39. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
40. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31, Montréal, QC, Canada, 3–8 December 2018*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2018; pp. 6638–6648. Available online: <https://papers.nips.cc/paper/7898-catboost-unbiased-boosting-with-categorical-features> (accessed on 29 April 2026).
41. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2005. [CrossRef] [PubMed]
42. Duan, T.; Anand, A.; Ding, D.Y.; Thai, K.K.; Basu, S.; Ng, A.; Schuler, A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. In *Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020*; PMLR: Brookline, MA, USA, 2020; Volume 119, pp. 2690–2700. Available online: <https://proceedings.mlr.press/v119/duan20a.html> (accessed on 29 April 2026).
43. Meinshausen, N. Quantile Regression Forests. *J. Mach. Learn. Res.* **2006**, *7*, 983–999. Available online: <https://jmlr.org/papers/v7/meinshausen06a.html> (accessed on 29 April 2026).
44. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017*; PMLR: Brookline, MA, USA, 2017; Volume 70, pp. 1321–1330. Available online: <https://proceedings.mlr.press/v70/guo17a.html> (accessed on 29 April 2026).
45. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized Quantile Regression. In *Proceedings of the Advances in Neural Information Processing Systems 32, Vancouver, BC, Canada, 8–14 December 2019*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2019; pp. 3543–3553. Available online: <https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html> (accessed on 29 April 2026).
46. Gneiting, T.; Raftery, A.E.; Westveld, A.H.; Goldman, T. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Mon. Weather Rev.* **2005**, *133*, 1098–1118. [CrossRef]
47. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2017; pp. 4765–4774. Available online: <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions> (accessed on 29 April 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.